Técnicas para el análisis y visualización de datos en el campo agropecuario

Carlos Farfán Almeida, Reynaldo Anchundia Macías,
Luis Cedeño-Valarezo.
Escuela Superior Politécnica Agropecuaria de Manabí
Manuel Félix López ESPAM MFL,
Calceta, Manabí,
Carrera de Computación,
Carrera de Agrícola.
carlos.farfan@espam.edu.ec,
reynaldo.anchundia@espam.edu.ec,
lcedeno@espam.edu.ec

DOI: https://doi.org/10.56124/encriptar.v8i16.004

Resumen

El estudio tuvo como objetivo la limpieza y preparación de datos relacionados con el peso de los lechones en diferentes etapas desde su nacimiento hasta el destete. Se abordan técnicas para el análisis y visualización de datos en el sector agropecuario mediante la aplicación parcial de la metodología CRISP-DM. En primer lugar, se recopilaron datos desde 2022 al 2024 de la ESPAM MFL referentes al hato porcino los mismos que estaban en tablas de Excel. Posteriormente, se realiza una revisión exhaustiva de la información, se determina que, de 1477 registros, únicamente el 6.23% estaban completos. Luego, se limpió y preparó los datos empleando un enfoque estructurado que incluyó la identificación y eliminación de datos atípicos mediante el método del rango intercuartílico, el filtrado de registros completos y el cálculo de tasas de crecimiento promedio entre columnas, las mismas que se emplearon para imputar valores faltantes, obteniendo un conjunto de datos limpio y consistente. El resultado más relevante incluye la generación de un conjunto de datos apto para modelos de aprendizaje automático, validado mediante diversas visualizaciones, como diagramas de cajas, histogramas y gráficos de dispersión. Este enfoque proporciona un marco reproducible y adaptable para la preparación de datos en el sector agropecuario.

Palabras-clave: Metodología CRISP-DM; procesamiento de datos; análisis de peso; análisis de datos.

Techniques for Data Analysis and Visualization in the Agricultural Field

Abstract

The study aims to clean and prepare data related to piglet weight at different stages from birth to weaning. It explores techniques for data analysis and visualization in the agricultural sector through the partial application of the CRISP-DM methodology.First, data from 2022 to 2024 related to the swine herd at ESPAM MFL was collected, originally stored in Excel tables. Subsequently, a thorough review of the information was conducted, revealing that only 6.18% of the 1,489 records were complete. The data was then cleaned and prepared using a structured approach that included identifying and removing outliers through the interquartile range method, filtering complete records, and calculating average growth rates across columns. These growth rates were used to impute missing values, resulting in a clean and consistent dataset. The most significant outcome is the generation of a dataset suitable for machine learning models, validated through various visualizations such as box plots, histograms, and scatter plots. This approach provides a reproducible and adaptable framework for data preparation in the agricultural sector.

Keywords: CRISP-DM methodology; data processing; weight analysis; data mining.

1. Introducción

Según Campos (2023), la incorporación de la inteligencia artificial en el ámbito agropecuario ha revolucionado los procesos de cultivo, producción y gestión, mejorando la eficiencia y productividad. Frackiewicz (2023) destaca que el uso de algoritmos avanzados y sistemas de aprendizaje automático permite a agricultores y ganaderos tomar decisiones más informadas, optimizando recursos como agua y fertilizantes, y contribuyendo a la sostenibilidad ambiental. Asimismo, Cuevas et al. (2021) subrayan que estas tecnologías facilitan el monitoreo en tiempo real del crecimiento de plantas, permitiendo la detección temprana de enfermedades y deficiencias de nutrientes. En el ámbito ganadero, Agrospray (2022) resalta el uso de la inteligencia artificial para supervisar la salud del ganado y predecir su comportamiento reproductivo.

La calidad de los datos utilizados para entrenar modelos de inteligencia artificial es crucial para garantizar su precisión y éxito. Según el Instituto de Ingeniería del Conocimiento (2021), los datos constituyen la base sobre la cual la inteligencia artificial aprende, identifica patrones y realiza predicciones. A medida que los modelos reciben mayores volúmenes de datos representativos y de alta calidad, su capacidad para adaptarse y generalizar mejora significativamente (González, 2022). Sin embargo, Corvalán (2018) y Dutta (2023) señalan que para evitar sesgos y garantizar la confiabilidad de los modelos, es fundamental que los datos sean diversos y estén correctamente gestionados. Esto abarca procesos de limpieza para eliminar inconsistencias, así como la adopción de un enfoque ético en su recopilación, como enfatizan Pérez (2021) y Paz (2023).

Un conjunto de datos destinado al entrenamiento de modelos debe reflejar con precisión la variabilidad del problema abordado. BigDataAnalytics (2023) indica que un dataset representativo y equilibrado asegura que los

modelos puedan generalizar correctamente y reconocer patrones en nuevas situaciones. Además, Bagnato (2020) añade que mantener un balance en las clases evita sesgos hacia categorías mayoritarias y mejora el desempeño en la clasificación de clases minoritarias.

El análisis y la visualización de datos son herramientas esenciales en disciplinas como la agricultura, la ingeniería y las ciencias sociales. Según Cavaller et al. (2018), estos procesos permiten descubrir patrones y relaciones significativas en los datos a través de métodos estadísticos y técnicas de aprendizaje automático. Eckert (2022) y Sunil (2023) destacan que estas técnicas no solo responden preguntas clave, sino que también apoyan la toma de decisiones informadas. Reyes (2018) y Himeur (2022) enfatizan que la representación gráfica de los datos facilita la identificación de tendencias, anomalías y correlaciones, lo que resulta especialmente útil en el ámbito agropecuario para mejorar la eficiencia y reducir riesgos.

En esta investigación, cuyo principal objetivo es construir un conjunto de datos para el aprendizaje automático basado en información agropecuaria, se aplican técnicas de análisis y visualización de datos. Estas introducen análisis de correlación, regresión y componentes principales (PCA) para identificar variables relevantes y explorar relaciones entre características agrícolas y ganaderas (profesorDATA, 2022; Coursera, 2023). Estas técnicas no solo optimizan la estructura del conjunto de datos, sino que también facilitan la identificación de patrones y tendencias clave para mejorar la eficiencia y sostenibilidad del sector agropecuario, como concluyen Bajić y Job (2023).

2. Materiales y métodos

En este estudio, se empleó a CRISP-DM como una herramienta metodológica. Ésta es una metodología estándar utilizada en el campo de la

minería de datos. Consta de seis etapas principales, de las cuales sólo se emplearon las tres primeras (IBM, 2023).

Para fundamentar el estudio, se llevó a cabo una revisión de la literatura en bases de datos científicas como Google Scholar, Scopus e IEEE Xplore, utilizando palabras clave relacionadas con "análisis de datos agropecuarios", "limpieza de datos", "metodología CRISP-DM" y "visualización de datos". Se seleccionaron artículos publicados en los últimos cinco años, priorizando aquellos con aplicaciones en el sector agropecuario. Se excluyeron referencias sin respaldo científico o que no estuvieran alineadas con los objetivos del estudio. Además, se analizaron revisiones previas sobre técnicas de procesamiento de datos aplicadas en la cría porcina, garantizando una base teórica actualizada y relevante para el desarrollo de la investigación.

A continuación, se presentan las etapas descritas en ésta metodología: Comprensión del negocio, Comprensión de datos, Preparación de los datos, Modelado, Evaluación y Despliegue.

2.1 Comprensión del negocio

Para completar esta fase, se estableció comunicación directa con la Carrera de Ingeniería Agrícola, alcanzando un acuerdo para realizar una visita al hato porcino y coordinar la entrega formal de los datos por parte del responsable de su gestión. Durante esta interacción, se expuso la naturaleza de la información requerida y se efectuó una revisión inicial de los archivos suministrados.

Posteriormente, tras la recopilación de los datos del hato porcino, se llevó a cabo un análisis preliminar en el que se identificaron diversos factores relevantes. La información se encontraba estructurada en fichas de parto individuales por cerda y en registros en formato bruto, lo que requirió un

proceso de organización y depuración para su adecuada integración y limpieza.

2.2 Estudio y comprensión de los datos

Luego de la recopilación de los datos provenientes del hato porcino, estos fueron organizados en filas y columnas dentro de una hoja de Excel, con el fin de facilitar su manipulación y procesamiento. Posteriormente, se importaron a Google Colab, donde se emplearon librerías especializadas para su limpieza y transformación.

En esta etapa, se realizó un análisis exhaustivo del conjunto de datos, verificando su estructura y calidad. Se identificaron inconsistencias, valores nulos y atípicos en los registros, asegurando que la información estuviera en condiciones óptimas para su uso. Finalmente, con base en este análisis, se seleccionaron los métodos y técnicas adecuados para la construcción del conjunto de datos completo, garantizando su coherencia y fiabilidad.

2.3 Preparación de los datos

Durante la fase de preparación de los datos, se implementaron diversas técnicas y métodos para garantizar la calidad, coherencia y completitud del conjunto de datos. A continuación, se describen los procedimientos aplicados.

Para eliminar los valores atípicos se utilizó el método IQR (*Interquartile Range*) para identificar y eliminar valores atípicos en las columnas de peso. Este método es una técnica estadística que define como valores atípicos aquellos que se encuentran fuera del rango determinado por los cuartiles Q1 (percentil 25%) y Q3 (percentil 75%), ajustando un multiplicador específico para cada variable.

Fórmula aplicada es $IQR = Q3 - Q1$, donde:

- Q1: Primer cuartil (25% de los datos ordenados).

- Q3: Tercer cuartil (75% de los datos ordenados).

A partir del IQR, se definieron los límites superior e inferior para identificar los valores atípicos según las fórmulas:

$$Límite\ Inferior = Q1 - (Multiplicador \times IQR)$$

$$Límite\ Superior = Q3 + (Multiplicador \times IQR)$$

Luego, para evitar sesgos en el análisis, se aplicó la técnica de filtrado por datos completos, que consiste en seleccionar únicamente aquellas filas donde todas las variables críticas de peso contaban con valores registrados. Esto garantiza que los cálculos posteriores se realizan sobre datos sin registros incompletos. Para ello, se utilizó la función dropna() de la librería pandas, la cual permite eliminar filas con valores nulos en las columnas especificadas, asegurando que solo se trabaje con datos completos y estructurados.

Posteriormente, se implementó el cálculo de la tasa de crecimiento como una métrica clave para evaluar la evolución del peso a lo largo del tiempo. Esta tasa mide el incremento porcentual del peso en distintos intervalos de días. La fórmula aplicada fue:

$$Tasa\ de\ Crecimiento = ((Peso\ Final - Peso\ Inicial)/Peso\ Inicial) \times 100$$

El cálculo de estas tasas permite analizar la dinámica de crecimiento y detectar posibles anomalías en los datos.

Por último, dado que varias columnas de peso contaban con valores faltantes, se aplicó la imputación de datos utilizando la tasa de crecimiento inter-etapas. Este método permite estimar valores ausentes basándose en el crecimiento promedio observado en los datos completos. Se aplicó la fórmula:

$$Valor\ Imputado = Peso\ Anterior \times (1 + Tasa\ de\ Crecimiento/100)$$

3. Resultados y discusión

3.1 Comprensión del negocio

La problemática radica en la falta de conjuntos de datos completos y de alta calidad en el sector agropecuario, lo que limita la implementación de tecnologías avanzadas, como el aprendizaje automático, en actividades agrícolas y ganaderas. Esta carencia dificulta la identificación de patrones y tendencias relevantes que podrían optimizar procesos productivos, mejorar la sostenibilidad y apoyar decisiones estratégicas basadas en datos. Por esta razón, en este trabajo se incluye el desarrollo de un conjunto de datos agropecuarios diseñado para garantizar su aplicabilidad en modelos de aprendizaje automático. Este conjunto de datos tiene como finalidad proporcionar información clave para analizar parámetros como el peso de los animales en diferentes etapas, identificar tendencias de crecimiento, y contribuir al desarrollo de prácticas más eficientes y sostenibles en el sector.

La información utilizada en este proyecto fue recopilada del hato porcino de la carrera de Ingeniería Agrícola de la ESPAM MFL. Estos datos fueron proporcionados en fichas de partos en formato Excel y contienen información detallada (en ciertas columnas, ya que otras existen valores nulos) sobre el peso al nacimiento, peso a los 8 días, peso a los 15 días, peso a los 22 días y peso al destete (30 días). Este registro es representativo de las condiciones productivas del sector y ha sido fundamental para garantizar la calidad y relevancia del análisis. La base de datos generada se podrá utilizar para mejorar la toma de decisiones en el ámbito agropecuario, fomentar la adopción de tecnologías innovadoras y sentar las bases para investigaciones futuras orientadas a la sostenibilidad y productividad del sector.

3.2. Estudio y Comprensión de los datos

Después de recopilar los datos y organizarlos en filas y columnas dentro

de una nueva hoja excel, se realizó un análisis del conjunto utilizando *Google Colab* como entorno de trabajo y empleando librerías como *pandas* y *matplotlib.pyplot*. Este análisis incluyó una revisión detallada de la estructura de la data, evaluando su calidad y contenido para identificar posibles inconsistencias o problemas. La descripción inicial del conjunto de datos (Tabla 1) abarcó las variables presentes, especificando su nombre, tipo y una breve explicación de su significado.

Tabla 1. *Descripción de Variables de la Data Set original.*

Categoría	Variable	Tipo	Descripción
Identificación de porcinos	CODIGO_LECHON	Categórico	Identificador único de cada lechón.
	CODIGO_MADRE	Categórico	Identificador de la madre del lechón.
	CODIGO_PADRE	Categórico	Identificador del padre del lechón.
Características básicas	SEXO	Categórico	Género del lechón.
	COLOR	Categórico	Color del lechón.
Pesos en diferentes etapas	PESO_AL_NACIMIENTO	Numérico	Peso al nacer del lechón.
	PESO_8_DIAS	Numérico	Peso del lechón a los 8 días de vida.
	PESO_15_DIAS	Numérico	Peso del lechón a los 15 días de vida.
	PESO_22_DIAS	Numérico	Peso del lechón a los 22 días de vida.

	PESO_30_DIAS	Numérico	Peso del lechón al destete (30 días de vida).
Información temporal y promedios	FECHA_DE_NACIMIENTO	Categórico	Fecha de nacimiento del lechón.
	PP_NACIMIENTO	Numérico	Peso promedio al nacer.
	PP_DESTETE	Numérico	Peso promedio al destete.

Fuente: Los autores (2024)

El análisis inicial permitió detectar ciertos problemas, como valores nulos en algunas columnas de peso y variaciones en el formato de las fechas. Estos hallazgos permitieron diseñar estrategias de limpieza y preparación de datos que garantizaran la integridad del análisis posterior. Este paso fue esencial para establecer una comprensión clara y detallada del conjunto de datos, sentando las bases para su transformación y análisis.

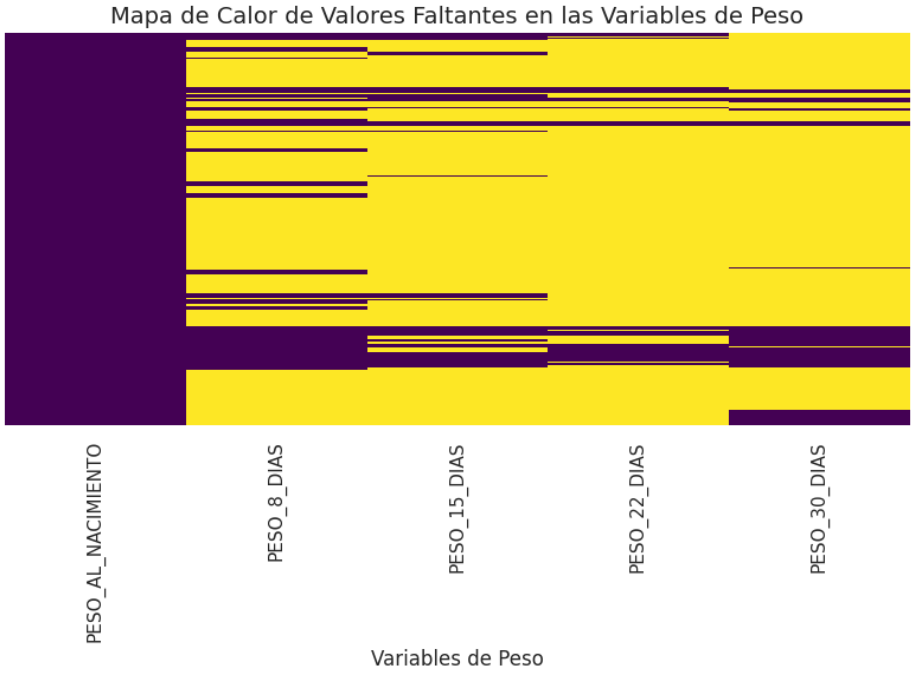
3.3. Preparación de los datos

La etapa de preparación de los datos integró un conjunto de técnicas orientadas a garantizar la calidad, integridad y completitud del conjunto de datos utilizado en el análisis. Este proceso fue dividido en varias fases consecutivas que incluyeron la detección y eliminación de valores atípicos, el filtrado por datos completos, el cálculo de tasas de crecimiento, la imputación de datos faltantes y, finalmente, la generación de gráficas representativas de la data procesada.

La figura 1 muestra un mapa de calor que visualiza los valores faltantes (color amarillo) y valores completos (color morado) en el conjunto de datos original, permitiendo identificar de manera clara las columnas de peso con registros incompletos.

La Tabla 2 complementa esta información al presentar la cantidad de valores faltantes y completos en cada una de las columnas de peso, proporcionando una visión detallada del estado inicial de los datos. En ella se observa que la variable PESO_AL_NACIMIENTO no presenta valores faltantes, mientras que el resto de las variables de peso muestran distintos niveles de ausencia de datos, siendo PESO_22_DIAS la columna con el mayor porcentaje de valores faltantes (74.27%). Además, se incluye una columna que expresa el porcentaje de datos ausentes en cada variable, lo que permite evaluar cuantitativamente el nivel de integridad de la información y su impacto en el análisis.

Figura 1. Mapa de calor de valores faltantes en las columnas de peso.



Fuente: Los autores (2024)

Tabla 2. *Distribución de valores ausentes por columna.*

Columnas de Peso	Valores Faltantes	Valores Completos	Porcentajes de Valores Faltantes
PESO_AL_NACIMIENTO	0	1477	00.00%
PESO_8_DIAS	842		57.00%
PESO_15_DIAS	1001		67.77%
PESO_22_DIAS	1097		74.27%
PESO_30_DIAS	1049		71.02%

Fuente: Los autores (2024)

3.3.1 Detección y eliminación de valores atípicos

El primer paso consistió en identificar valores atípicos en las columnas relacionadas con el peso (PESO_AL_NACIMIENTO, PESO_8_DIAS, PESO_15_DIAS, PESO_22_DIAS, PESO_30_DIAS), utilizando el método IQR.

Para controlar la sensibilidad en la detección de datos atípicos, se ajustaron los multiplicadores en la fórmula de los límites superiores e inferiores de cada columna. Este ajuste permitió mantener un equilibrio entre la eliminación de valores extremos y la conservación de datos representativos.

En la Tabla 3, se presentan los valores de los multiplicadores inferior y superior aplicados a cada columna de peso, junto con los límites inferior y superior resultantes tras su cálculo.

La Figura 2 muestra los diagramas de caja correspondientes a las columnas de peso del conjunto de datos: PESO_AL_NACIMIENTO, PESO_8_DIAS, PESO_15_DIAS, PESO_22_DIAS y PESO_30_DIAS,

evidenciando la presencia de valores atípicos en cada una de ellas. En la columna PESO_AL_NACIMIENTO, se identifican 79 valores atípicos, algunos superiores a 15 kg. En PESO_8_DIAS, destacan 136 valores atípico, unos están por encima de 50 kg, mientras que en PESO_15_DIAS se observan 94 valores atípicos entre ellos unos superiores a 7 kg. La columna PESO_22_DIAS presenta 82 valores atípicos, unos están por encima de 60 kg, mientras que en PESO_30_DIAS se obtienen 55 valores atípicos, siendo los más altos ligeramente superiores a 12 kg.

Estos resultados evidencian la presencia de datos atípicos en estas columnas, los cuales fueron identificados y eliminados mediante el método IQR, lo que permitió obtener un conjunto de datos más coherente y representativo de las tendencias de crecimiento en las diferentes etapas.

Tabla 3. *Distribución de valores ausentes por columna.*

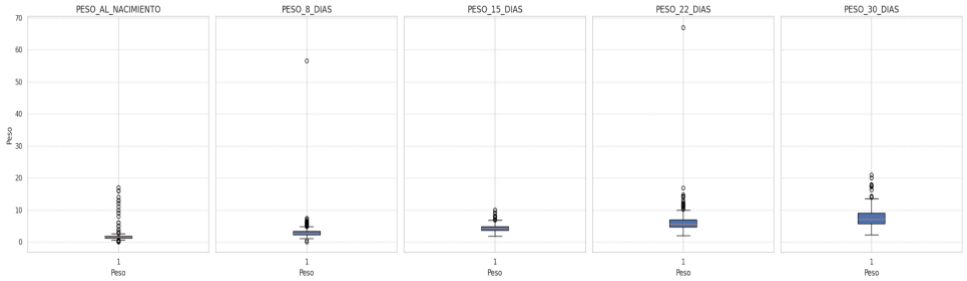
Columna	Multiplicador Inferior	Multiplicador Superior	Límite Inferior	Límite Superior
PESO_AL_NACIMIENTO	1	1	0.78	2.34
PESO_8_DIAS	0.25	0.50	2.11	3.77
PESO_15_DIAS	0.20	0.50	3.56	5.40
PESO_22_DIAS	0.25	0.25	4.77	6.90
PESO_30_DIAS	0.10	1.50	5.73	12.15

Fuente: Los autores (2024)

En la Figura 3, los diagramas de cajas muestran los datos de las columnas de peso en distintos días después de la eliminación de valores atípicos mediante el método IQR, asegurando una distribución más limpia y representativa. En el peso al nacimiento y en los primeros días (8, 15 y 22

días), los valores están concentrados en rangos estrechos, lo que indica baja variabilidad y mayor homogeneidad en los datos.

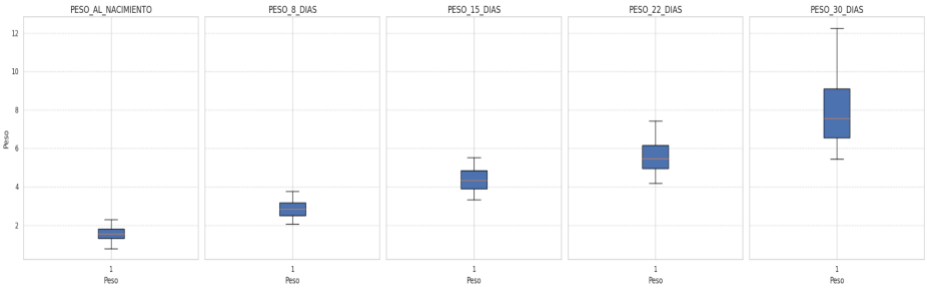
Figura 2. Diagramas de caja con valores atípicos en columnas de peso.



Fuente: Los autores (2024)

En el peso a los 30 días, se observa una mayor dispersión de valores, coherente con el impacto del crecimiento, pero sin presencia de valores extremos que afecten la distribución. Esto confirma que la aplicación del método IQR permitió obtener un conjunto de datos más preciso y consistente para el análisis de las tendencias de crecimiento.

Figura 3. Diagramas de caja sin valores atípicos en columnas de peso.



Fuente: Los autores (2024)

3.3.2 Filtrado por datos completos

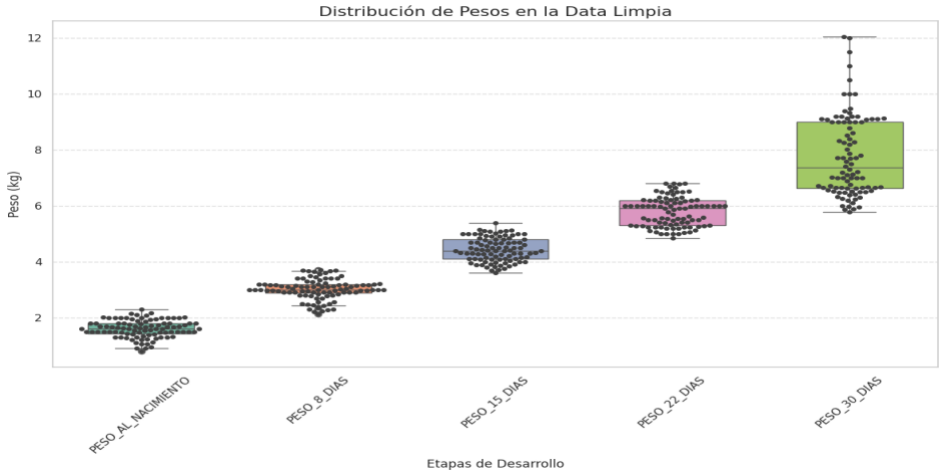
Esta sección describe el procedimiento para la depuración de un conjunto de datos relacionado con el crecimiento de lechones, con el propósito

de derivar un subconjunto compuesto exclusivamente por registros completos en las columnas que reflejan el peso en distintas etapas del desarrollo: nacimiento, 8, 15, 22 y 30 días. El proceso comienza con la creación de una copia del conjunto de datos original, garantizando así la preservación de su integridad y habilitando el trabajo en una versión independiente.

Posteriormente, se define un arreglo con las columnas relevantes y se aplica el método `dropna` para eliminar las filas que contengan valores nulos (NaN) en dichas columnas. Como resultado, se genera un nuevo subconjunto denominado `PORCINO_PESO_COMPLETO`. Este procedimiento asegura la eliminación total de datos incompletos, lo que resulta crucial para mantener la precisión y la consistencia en los análisis subsecuentes.

Finalmente, el código computa y presenta el número total de registros con datos completos, junto con un resumen preliminar de las primeras 92 filas del subconjunto filtrado. En la figura 4 se muestra la evolución del peso de los lechones mediante un *boxplot* combinado con un gráfico de violín, evidenciando un aumento progresivo en el peso promedio y una mayor dispersión en las etapas avanzadas, especialmente a los 30 días.

Figura 4. *Distribución de Peso en la Data Limpia.*



Fuente: Los autores (2024)

Este enfoque integral no solo mejora la calidad de los datos para análisis avanzados, sino que también reduce significativamente el riesgo de errores y aumenta la confiabilidad de los resultados obtenidos mediante técnicas analíticas más complejas.

3.3.3 Tasas de crecimiento

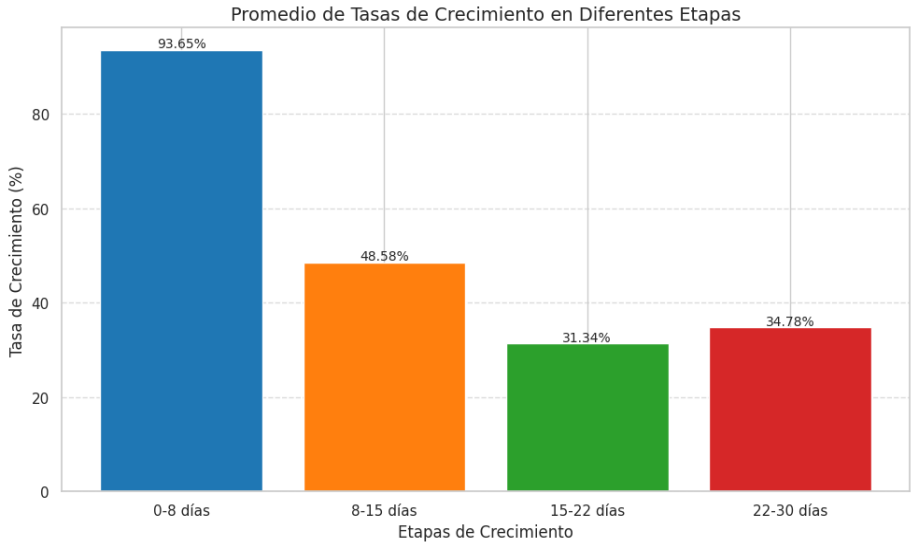
En esta parte se implementa el cálculo de las tasas de crecimiento porcentual entre diferentes etapas de desarrollo de lechones: de 0 a 8 días, de 8 a 15 días, de 15 a 22 días y de 22 a 30 días, utilizando los pesos registrados en cada intervalo. Estas etapas fueron seleccionadas porque forman parte de las variables analizadas en el modelo, al representar fases críticas del ciclo de desarrollo de los lechones que incluyen crecimiento rápido. Estas tasas se derivan utilizando la fórmula: , donde representa el peso en la etapa actual y el peso en la etapa previa.

Esta expresión permite cuantificar el crecimiento relativo en términos porcentuales, proporcionando una métrica clara y consistente para el análisis. Posteriormente, se determinan los valores promedio de dichas tasas para cada intervalo, ajustándose a dos decimales. Este procedimiento analítico facilita la identificación de patrones de desarrollo y la evaluación cuantitativa del rendimiento en cada fase, lo que tiene implicaciones prácticas significativas, como la optimización de estrategias de alimentación y manejo para maximizar el crecimiento en etapas críticas. Además, estos patrones pueden proporcionar una base para investigaciones científicas adicionales, como el análisis de la influencia de factores genéticos o ambientales en el desarrollo de los lechones, contribuyendo a un diseño más eficiente de programas de producción porcina.

La figura 5 muestra un gráfico de barras asociado que ilustra los promedios calculados, destacando un crecimiento inicial significativo del

93.65% entre 0-8 días, seguido por una disminución gradual en los intervalos subsiguientes: 48.58% entre 8-15 días, 31.34% entre 15-22 días, y un incremento marginal a 34.78% entre 22-30 días. Este comportamiento refleja un desarrollo progresivo influenciado por factores como la adaptación temprana y la estabilización fisiológica en etapas posteriores.

Figura 5. *Promedio de Tasas de Crecimiento en Diferentes Etapas.*



Fuente: Los autores (2024)

3.3.4 Imputación de Datos Faltantes

Para completar los valores faltantes en las columnas relacionadas con el peso (PESO_8_DIAS, PESO_15_DIAS, PESO_22_DIAS y PESO_30_DIAS), se aplicó un método de imputación basado en el promedio de las tasas de crecimiento entre etapas consecutivas. Este enfoque garantizó que los datos estimados respetaran las tendencias de crecimiento observadas en los registros completos y mantuvieran la consistencia lógica entre las columnas de peso.

Condiciones aplicadas durante la imputación:

A. Validación con el peso de la etapa siguiente (si disponible): Si el peso de la etapa siguiente estaba presente, se verificó que el valor imputado no fuera mayor o igual que el peso de dicha etapa. En caso de incumplir esta condición, la tasa de crecimiento promedio se redujo iterativamente en un 10% (tasa_reducción = 0.9), ajustando el cálculo del valor imputado.

B. Verificación de límites lógicos: El peso imputado debía ser mayor al peso de la etapa anterior y, si existía un dato posterior, menor al peso de la etapa siguiente. Esta regla evitó estimaciones incoherentes y garantizó una progresión de crecimiento lógica y realista.

C. Manejo de casos sin dato siguiente: Si no existía un valor registrado para la etapa siguiente, la validación se realizó únicamente en relación con el peso de la etapa anterior.

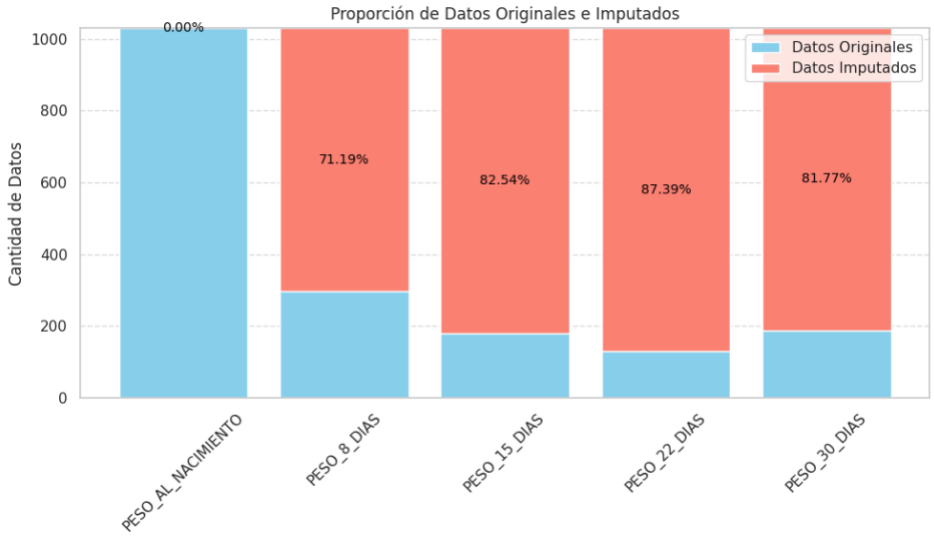
La Figura 6 muestra la distribución de datos originales e imputados en distintas etapas del crecimiento de los lechones. PESO_AL_NACIMIENTO cuenta con 100% de datos originales, reflejando una captura de datos completa en esta fase. Sin embargo, a medida que avanza el crecimiento, se incrementa la proporción de datos imputados, alcanzando 71.19% en PESO_8_DIAS, 82.54% en PESO_15_DIAS, 87.39% en PESO_22_DIAS y 81.77% en PESO_30_DIAS, lo que indica una pérdida significativa de información en etapas posteriores.

Para mitigar este problema, se aplicó el IQR para eliminar valores aberrantes y mejorar la calidad del conjunto de datos. Posteriormente, se realizó la imputación utilizando la tasa de crecimiento inter-etapas, asegurando que los valores generados fueran coherentes con la tendencia de desarrollo

natural de los lechones. Este método permitió mantener la continuidad de los registros sin alterar su estructura.

El alto porcentaje de imputación en etapas avanzadas puede deberse a problemas en la recolección de datos o inconsistencias operativas. Aunque la imputación es una estrategia válida para completar series de datos incompletas, es fundamental mejorar los procesos de captura para reducir la dependencia de este método. Además, en la columna PP_DESTETE, se aplicó una imputación basada en la media de PESO_30_DIAS dentro de cada grupo familiar, garantizando estimaciones coherentes con los patrones de crecimiento observados.

Figura 6. *Proporción de Datos Originales e Imputados.*



Fuente: Los autores (2024)

4. Conclusiones

El proceso de recopilación y análisis de datos agropecuarios, junto con la aplicación de métodos específicos para su limpieza, imputación y visualización, ha demostrado ser esencial para optimizar la gestión de datos en la cría porcina. La implementación de técnicas como la detección y

eliminación de valores atípicos mediante el IQR, el filtrado por datos completos y la imputación basada en tasas de crecimiento ha permitido construir un conjunto de datos completo, consistente y de alta calidad. Este conjunto de datos proporciona una base sólida para mejorar la toma de decisiones y la eficiencia operativa en el ámbito agropecuario.

La exploración detallada de las variables de peso, respaldada por técnicas estadísticas y representaciones gráficas, ha permitido identificar patrones clave en las diferentes etapas de crecimiento de los lechones, desde el nacimiento hasta los 30 días. Los ajustes realizados en el conjunto de datos aseguran que las tendencias observadas sean representativas y confiables, minimizando el impacto de valores extremos o datos faltantes que podrían comprometer la calidad del análisis.

En la etapa final, la creación de un conjunto refinado de datos con técnicas avanzadas de imputación ha destacado la relevancia de características específicas, como la relación entre el peso y las condiciones reproductivas. Este conjunto de datos se erige como una herramienta robusta y adaptada a las necesidades específicas del sector porcino, proporcionando *insights* valiosos para la toma de decisiones informadas.

Estos resultados convergen en un enfoque integral que no solo mejora la calidad y precisión de los datos en la cría porcina, sino que también impulsa la eficiencia en la gestión de procesos clave en este ámbito, consolidando el papel de las técnicas de análisis y visualización como herramientas esenciales en el sector agropecuario.

5. Referencias

@BigDataAnalytics. (2023). Dataset: Qué es y principales características. The Data Schools. https://thedataschools.com/que-es/data-set/

AgroSpray. (2022, January 10). Inteligencia artificial en el agro: más productivo y sustentable -. AgroSpray Blog. https://agrospray.com.ar/blog/inteligencia-artificial-en-el-agro/

Ambrósio, L. M., Linhares, H. M., David, J., Braga, R., Arbex, W., Campos, M. M., & Capilla, R. (2021). Enhancing the reuse of scientific experiments for agricultural software ecosystems. Journal of Grid Computing, 19(4). https://doi.org/10.1007/s10723-021-09583-x

Bagnato. (2020). Clasificación con datos desbalanceados | Aprende Machine Learning. Aprende Machine Learning. https://www.aprendemachinelearning.com/clasificacion-con-datos-desbalanceados/

Bajić, F., & Job, J. (2023). Review of chart image detection and classification. International Journal on Document Analysis and Recognition (IJDAR), 26(4), 453–474. https://doi.org/10.1007/s10032-022-00424-5

Cavaller, V., Ruiz, D., Terrádez, M., Martín, A., & García, R. (JUNIO de 2018). Técnicas avanzadas de análisis y visualización de datos. COMEIN(78). https://doi.org/https://doi.org/10.7238/issn.2014-2226

Campos, A. (2023). Inteligencia artificial en agricultura: la nueva revolución del sector. Increnta. https://increnta.com/insights/inteligencia-artificial-en-agricultura/

Chen, J., Xiangliang, C., Fu, H., & Lam, A. (2018). Agricultural product monitoring system supported by cloud computing. Cluster Computing, 22(S4), 8929–8938. https://doi.org/10.1007/s10586-018-2022-5

Corvalán, J. G. (2018). Inteligencia artificial: retos, desafíos y oportunidades - Prometea: la primera inteligencia artificial de Latinoamérica al servicio

de la Justicia*. https://www.redalyc.org/journal/5340/534057837015/html/

Coursera. (2023, June 15). Visualización de datos: Definición, ventajas y ejemplos. Coursera. https://www.coursera.org/mx/articles/data-visualization

Cuevas, M. F., Olguin, G. M., Noriega, A. L., & Ortiz, R. (2021, 07 15). Agricultura de precisión, análisis de imágenes aplicada a la producción agrícola. inecol. https://www.inecol.mx/inecol/index.php/es/ct-menu-item-25/ct-menu-item-27/17-ciencia-hoy/1431-agricultura-de-precision-analisis-de-imagenes-aplicada-a-la-produccion-agricola

De Catalunya, U. O. (2018). Técnicas avanzadas de análisis y visualización de datos. https://comein.uoc.edu/divulgacio/comein/es/numero78/articles/tecnicas-avanzadas-analisis-visualizacion-datos.html

De Oliveira, W., Braga, R., David, J., Ströele, V., Campos, F., & Castro, G. (2022). Visionary: a framework for analysis and visualization of provenance data. Knowledge and Information Systems, 64(2), 381–413. https://doi.org/10.1007/s10115-021-01645-6

Deeken, H., Wiemann, T., & Hertzberg, J. (2019). A spatio-semantic approach to reasoning about agricultural processes. Applied Intelligence, 49(11), 3821–3833. https://doi.org/10.1007/s10489-019-01451-2

Deniz, M., De Sousa, K. T., Gomes, I. C., Fabro, J. A., Vale, M. M. D., & Dittrich, J. R. (2021). Development and application of an autonomous data logger to measure environmental variables in livestock farming. International Journal of Environmental Science and Technology, 19(9), 8699–8712. https://doi.org/10.1007/s13762-021-03734-z

Dutta, S. (2023). Modelado de datos: Un pilar fundamental de su futura tecnología de IA. CAS. https://www.cas.org/es-

es/resources/blog/modelado-de-datos-un-pilar-fundamental-de-su-futura-tecnologia-de-ia

Espinosa-Zúñiga, J. J. (2020). *Aplicación de metodología CRISP-DM para segmentación geográfica de una base de datos pública*. Ingeniería, Investigación y Tecnología, 21(1), e00008. https://doi.org/10.22201/fi.25940732e.2020.21n1.008

Frąckiewicz, M. (2023, August 19). El papel de la computación espacial en la agricultura y la ganadería. TS2 SPACE. https://ts2.space/es/el-papel-de-la-computacion-espacial-en-la-agricultura-y-la-ganaderia/#gsc.tab=0

González, R. (2022). ¿Calidad o cantidad? Lo que más importa en el Big Data. Big Data Magazine. https://bigdatamagazine.es/calidad-o-cantidad-lo-que-mas-importa-en-el-big-data

GRAP. (2023). *Digitalización y análisis Big Data en granjas porcinas*. Recuperado de https://www.grap.udl.cat/es/recerca/linies-de-recerca/digitalitzacio-i-analisi-big-data-en-granges-porcines/

Himeur, Y., Elnour, M., Fadli, F., Meskin, N., Petri, I., Rezgui, Y., Bensaali, F., & Amira, A. (2022). AI-big data analytics for building automation and management systems: a survey, actual challenges and future perspectives. Artificial Intelligence Review, 56(6), 4929–5021. https://doi.org/10.1007/s10462-022-10286-2

IBM. (2023). CRISP-DM: Cross-Industry Standard Process for Data Mining. Guía para la Minería de Datos. Disponible en https://www.ibm.com/docs/es/spss-modeler/saas?topic=dm-crisp-help-overview.

Instituto de Ingeniería del Conocimiento. (2021, January 21). Machine Learning y Deep Learning - Expertos en IIC. Instituto De Ingeniería Del Conocimiento. https://www.iic.uam.es/inteligencia-artificial/machine-learning-deep-learning/

Paz, A. (2023, March 29). Calidad de Datos: Cómo Asegurar la Precisión y Confiabilidad en tu Información - Estrategia Uno. Estrategia Uno. https://estrategia.uno/calidad-de-datos-precision-y-confiabilidad/

Pérez, A. (2021, September 20). Inteligencia artificial y gestión de calidad. OBS Business School. https://www.obsbusiness.school/blog/inteligencia-artificial-y-gestion-de-calidad

profesorDATA. (2022, January 21). Reducción de la dimensionalidad: Análisis de Componentes Principales (PCA). profesorDATA.com. https://profesordata.com/2020/09/01/reduccion-de-la-dimensionalidad-analisis-de-componentes-principales-pca/

Ramírez-Morales, I., Mazon-Olivo, B., & Pan, A. (2018). Ciencia de datos en el sector agropecuario. En *Análisis de datos agropecuarios* (pp. 1-44). Machala, Ecuador: Editorial UTMACH. https://doi.org/10.48190/9789942241207.1

Reyes, V. (2018). Revista de los Estudios de Ciencias de la Información y de la Comunicación. Dianet. https://dialnet.unirioja.es/servlet/articulo?codigo=7472358

Sunil, C. K., Jaidhar, C. D., & Patil, N. (2023). Systematic study on deep learning-based plant disease detection or classification. Artificial Intelligence Review, 56(12), 14955–15052. https://doi.org/10.1007/s10462-023-10517-0