GRADO EN INGENIERÍA MATEMÁTICA CON INTELIGENCIA ARTIFICIAL

TRABAJO FIN DE GRADO

Estimación del tamaño muestral óptimo mediante modelado probabilístico. Aplicación al muestreo en lotes de fruta

Autor: Eugenio Ribón Novoa

Director: Eugenio Francisco Sánchez Úbeda

Madrid

Mayo de 2026

Declaro, bajo mi responsabilidad, que el Proyecto presentado con el título
Estimación del tamaño muestral óptimo mediante modelado probabilístico. Aplicación al muestreo en lotes de fruta
en la ETS de Ingeniería - ICAI de la Universidad Pontificia Comillas en el curso académico 2025/2026 es de mi autoría, original e inédito y no ha sido presentado con anterioridad a otros efectos.
El Proyecto no es plagio de otro, ni total ni parcialmente y la información que ha sido tomada de otros documentos está debidamente referenciada.

Fdo.: EUGENIO RIBÓN NOVOA Fecha: 25/05/2026

Autorizada la entrega del proyecto
EL DIRECTOR DEL PROYECTO

Fdo.: Eugenio Francisco Sánchez Ubeda Fecha: 12/06/2026

ESTIMACIÓN DEL TAMAÑO MUESTRAL ÓPTIMO MEDIANTE MODELADO PROBABILÍSTICO. APLICACIÓN AL MUESTREO EN LOTES DE FRUTA

Autor: Ribón Novoa, Eugenio.
Director: Sánchez Úbeda, Eugenio.
Entidad Colaboradora: Citri&CO

RESUMEN DEL PROYECTO

El "Proyecto Cítrico" se centra en la optimización del control de calidad en plantas de procesamiento de cítricos mediante técnicas de análisis de datos y algoritmos de aprendizaje automático. Ante las desviaciones estadísticas existentes entre las estimaciones iniciales de muestreo (escandallos) y los resultados empíricos obtenidos en las máquinas calibradoras, se ha desarrollado y evaluado un pipeline predictivo basado en Regresión Multi-salida. El modelo óptimo, fundamentado en Ridge Regression con regularización L2, logra reducir el Error Absoluto Medio (MAE) global en un 65.72% frente al baseline tradicional de la planta, proporcionando un marco computacional robusto para la planificación de la producción y la mitigación de ineficiencias de calibración.
Palabras clave: Machine Learning, Control de Calidad, Industria Cítrica, Regresión Multi-salida, Ridge Regression, Divergencia de Jensen-Shannon.

1. **Introducción**

   La industria agroalimentaria, y en particular el sector de procesamiento de cítricos, se enfrenta al desafío constante de clasificar y calibrar su materia prima de manera eficiente y algorítmicamente precisa. Tradicionalmente, la calidad y viabilidad comercial de los lotes de fruta entrantes se estiman mediante un muestreo estadístico probabilístico denominado "escandallo". Sin embargo, en la práctica industrial, frecuentemente se observan discrepancias y variabilidades significativas entre esta estimación estática inicial y la distribución real de calidades (categorizadas desde Q1 como óptima, hasta Q5) que las máquinas calibradoras determinan empíricamente tras el procesamiento del lote completo.
   En la planta procesadora objeto de este estudio (Sollana), la dependencia de un muestreo manual poco representativo y el desajuste intrínseco de los módulos calibradores (máquinas C1, C5 y C6) generan ruido en la información operativa, lo que se traduce en desajustes logísticos y fluctuaciones en la calidad del producto comercializado. El presente trabajo aborda esta problemática aplicando metodologías avanzadas de Ciencia de Datos (Data Science) para auditar el comportamiento de la maquinaria y entrenar un modelo supervisado capaz de predecir la distribución real de la calidad de la fruta, mitigando así el sesgo de las líneas de producción.

2. **Definición del proyecto**

   El proyecto se fundamenta en un análisis analítico-predictivo de los volúmenes de producción procesados en la planta de Sollana. El propósito subyacente es trascender el modelo de estimación basado en el "escandallo" como verdad absoluta, pasando a un paradigma basado en datos (Data-Driven). Para alcanzar este objetivo, la arquitectura de la investigación se ha estructurado secuencialmente en cuatro dominios principales de extracción de conocimiento:
   1. Ingestión y Normalización de Datos: Agregación de datasets brutos, estructuración espacial de datos pivotados y estandarización de columnas relativas a los calibres y pesos procesados.

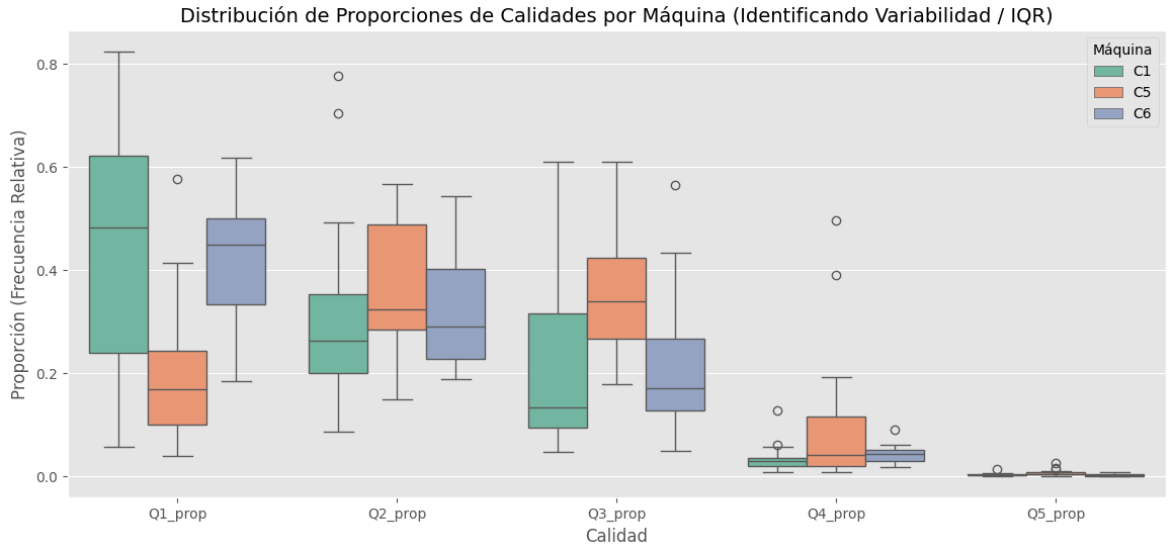
2. Análisis Exploratorio de Datos (EDA): Inferencia estadística sobre las distribuciones de calibres, obtención de métricas de tendencia central y dispersión (media, desviación estándar, rango intercuartílico), y evaluación del comportamiento heterogéneo intra y entre máquinas.

3. Análisis Topológico de Variables: Cuantificación estocástica de las desviaciones de calidad empleando métricas avanzadas de teoría de la información, tales como la Divergencia de Jensen-Shannon, para aislar los sesgos mecánicos del escandallo teórico.

4. Modelado y Validación Predictiva: Implementación de técnicas de aprendizaje supervisado de Regresión Multi-salida (Multi-output Regression) para predecir las proporciones continuas del vector de calidades (Q1 a Q5) garantizando la corrección algorítmica de la maquinaria.

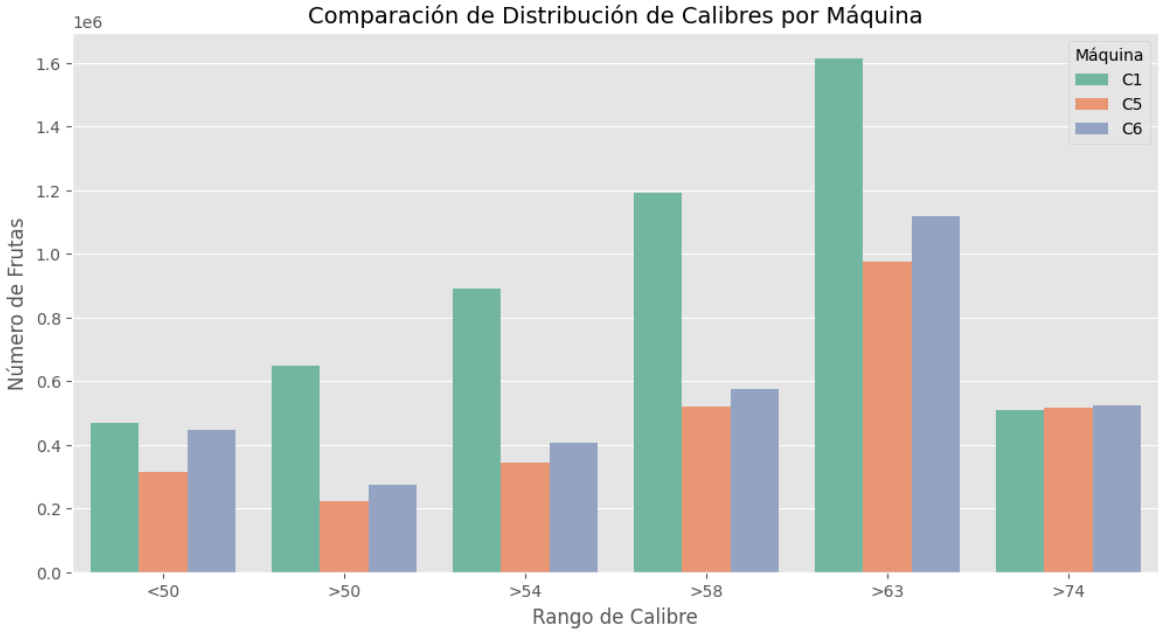
3. Descripción del modelo/sistema/herramienta

El ecosistema computacional se ha construido íntegramente en Python, utilizando el stack estándar de inferencia matemática y Machine Learning (Pandas, NumPy, Scikit-Learn y entornos de visualización estocástica). Durante la fase exploratoria (EDA), se identificaron asimetrías severas en el procesamiento de las tres calibradoras. A modo de ejemplo empírico, se observó que la máquina C5 presenta una desviación estándar sumamente elevada en frutas de calibre grande (>74mm) y una proporción anormalmente alta de clasificación de calidad defectuosa (Q4), detectando apenas un 19% de calidad óptima (Q1) frente al 50% reportado por las líneas C1 y C6.



Debido a estas anomalías, asumir una relación lineal 1:1 entre el escandallo y la maquinaria (Baseline) resulta ineficaz. Para modelar esta complejidad, la ingeniería de características (Feature Engineering) incorporó variables normalizadas (las proporciones de escandallo de Q1 a Q5) y sometió la identificación de la máquina a un proceso categórico matemático mediante codificación One-Hot Encoding.
Para el núcleo predictivo se testearon diferentes algoritmos, primando modelos basados en ensamblaje de árboles de decisión (Random Forest Regressor) y enfoques paramétricos penalizados. El algoritmo final seleccionado es un modelo de Ridge Regression. La regularización L2 inherente a Ridge se adapta de manera óptima al problema, combatiendo la multicolinealidad entre características derivadas de la misma proporción y dotando al

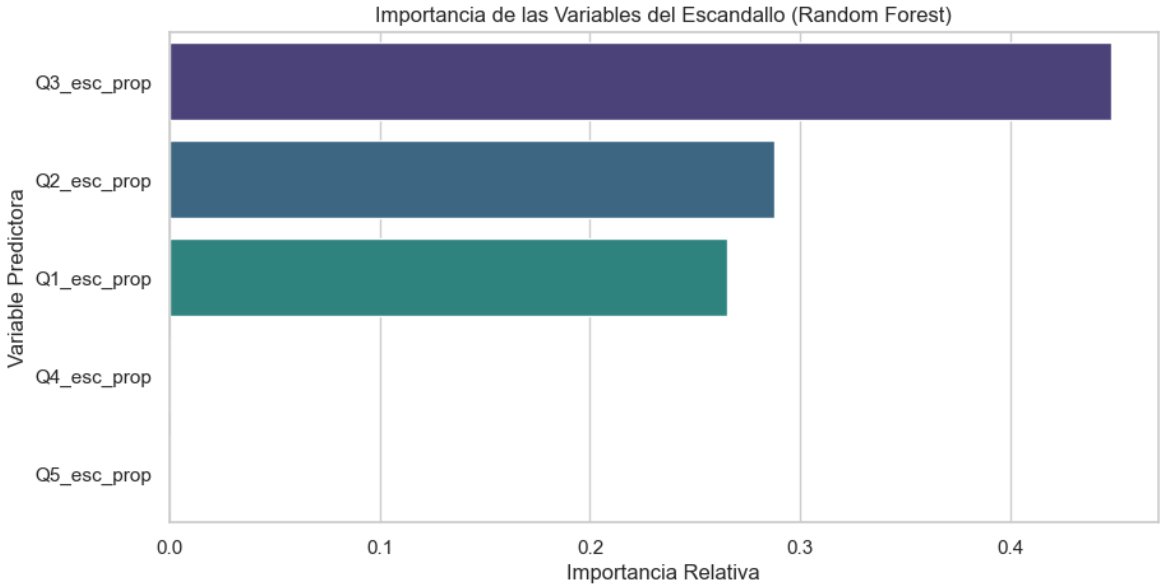
sistema de una extrema robustez frente a sobreajustes (overfitting) en distribuciones fraccionales.



4. Resultados

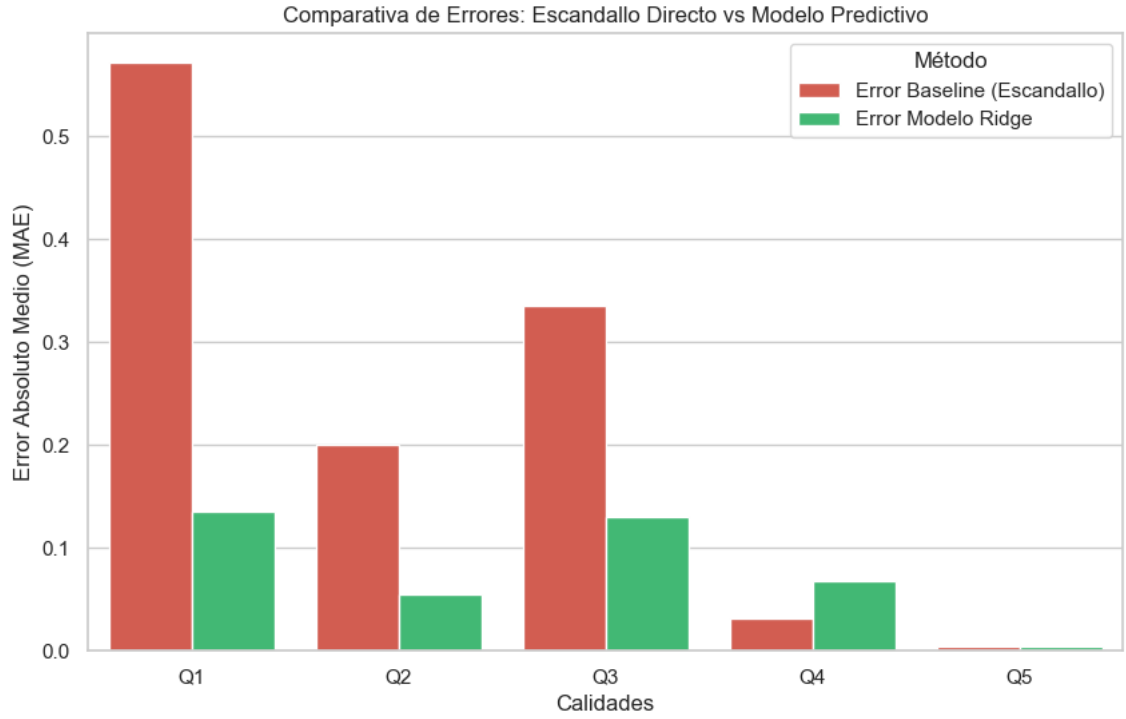
La evaluación del rendimiento de los algoritmos refleja una optimización drástica de los indicadores del negocio. La métrica de validación empleada para penalizar las divergencias proporcionales fue el Error Absoluto Medio (MAE).

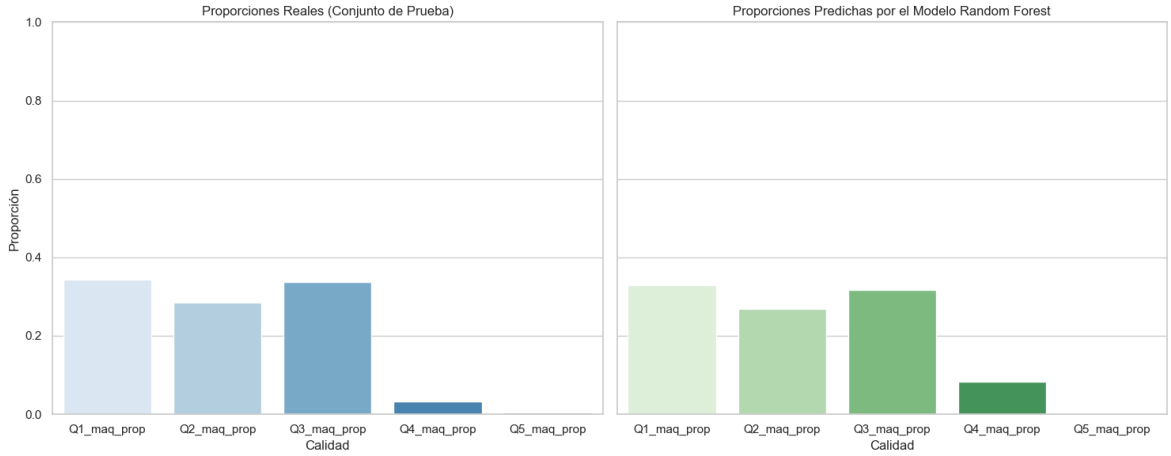
Al evaluar el Baseline —que evalúa el modelo productivo actual de la planta asumiendo el escandallo como predicción directa—, el modelo alcanzó un error sistemático altísimo, con un MAE global del 22.85%. La introducción del primer aproximador supervisado mediante Random Forest logró comprimir este margen de error a un 9.51%.



Sin embargo, el hito predictivo del sistema se logra gracias a la arquitectura lineal penalizada: el modelo Ridge Regression, entrenado con las variables categóricas espaciales de la maquinaria de calibrado, reduce el MAE global hasta un 7.83% sobre el conjunto de test (datos no vistos). A nivel de impacto empresarial, este algoritmo atenúa la desviación en el volumen de predicción en un 65.72% con respecto al método de calibración

tradicional, demostrando ser un marco analítico capaz de estabilizar la inferencia de calidad en la industria cítrica.





5. Conclusiones

La presente memoria demuestra fehacientemente que los procesos de escandallos cítricos tradicionales carecen de la suficiencia estadística para garantizar la correcta planificación productiva debido al ruido inyectado por las heterogeneidades de la propia maquinaria de envasado. El desarrollo algorítmico desplegado basado en regresiones linealmente regularizadas (Ridge) compensa matemáticamente estas deficiencias mecánicas, validando el impacto que la Ciencia de Datos ostenta en la cadena de valor de la Industria 4.0.

La extrapolación de estos resultados asienta una sólida base para las iteraciones en el roadmap de Machine Learning del proyecto. Como trabajos futuros propuestos destacan:

1. Clustering de Perfiles de Lote: Modelado no supervisado orientado a agregar lotes isomorfos basados en el cruce poblacional de su escandallo, persiguiendo el descubrimiento de dependencias ocultas por región o terreno.

2. Relación Calibre-Calidad: Análisis topológico bidimensional de la función de densidad generada por el volumen físico frente a su catalogación en la matriz de calidades.

3. Enriquecimiento del Espacio de Características (Features): Evolución del dataset integrando predictores medioambientales, biológicos e industriales, tales como el Peso Total (kg), Volumen absoluto de frutas, metadatos espaciales del campo agrario de origen y variedad biológica del cultivo.

6. Referencias (TODO)

[1] Lin, J. "Divergence measures based on the Shannon entropy", *IEEE Transactions on Information Theory*, Julio 1991, vol. 37, no. 4, pp. 145-151

[2] Hoerl, A.E.; Kennard, R.W. "Ridge Regression: Biased Estimation for Nonorthogonal Problems", *Technometrics*, Febrero 1970, vol. 12, no. 1, pp. 55-67

[3] Borchani, H.; Varando, G.; Bielza, C.; Larrañaga, P. "A survey on multi-output regression", *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, Septiembre 2015, vol. 5, no. 5, pp. 216-233.

OPTIMAL SAMPLE SIZE ESTIMATION THROUGH PROBABILISTIC MODELING. APPLICATION TO FRUIT BATCH SAMPLING

Author: Ribón Novoa, Eugenio.
Supervisor: Sánchez Úbeda, Eugenio Francisco.
Collaborating Entity: Citri&CO

ABSTRACT

The "Citrus Project" focuses on the optimization of quality control in citrus processing facilities through advanced data analytics and machine learning models. Faced with systemic statistical discrepancies between the initial probabilistic sampling (known as "escandallo") and the empirical output from calibrator machines, a predictive pipeline based on Multi-output Regression was developed. The selected optimal algorithm, an L2 penalized Ridge Regression, achieves a remarkable 65.72% reduction in the global Mean Absolute Error (MAE) compared to the factory's traditional baseline methodology. This robust computational framework provides an effective, data-driven approach to production planning, neutralizing operational biases and calibration inefficiencies.
Keywords: Machine Learning, Quality Control, Citrus Industry, Multi-output Regression, Ridge Regression, Jensen-Shannon Divergence.

1. Introduction

The agri-food industry, and particularly the citrus processing sector, faces the constant challenge of sorting and grading its raw material efficiently and algorithmically accurately. Traditionally, the quality and commercial viability of incoming fruit batches are estimated through probabilistic statistical sampling known as an "escandallo". However, in industrial practice, significant discrepancies and variability are frequently observed between this initial static estimation and the actual quality distribution (categorized from Q1 as optimal, to Q5) empirically determined by the grading machines after processing the entire batch.

In the processing plant under study (Sollana), the reliance on unrepresentative manual sampling and the intrinsic misalignment of the grading modules (machines C1, C5, and C6) generate noise in operational information, which translates into logistical imbalances and fluctuations in the quality of the commercialized product. This work addresses this problem by applying advanced Data Science methodologies to audit the machinery's behavior and train a supervised model capable of predicting the actual distribution of fruit quality, thereby mitigating the bias of the production lines.

2. Project definition

The project is based on an analytical-predictive analysis of the production volumes processed at the Sollana plant. The underlying purpose is to transcend the estimation model based on the "escandallo" as an absolute truth, moving towards a Data-Driven paradigm. To achieve this objective, the research architecture has been sequentially structured into four main domains of knowledge extraction:

1. Data Ingestion and Normalization: Aggregation of raw datasets, spatial structuring of pivoted data, and standardization of columns related to processed sizes and weights.
2. Exploratory Data Analysis (EDA): Statistical inference on size distributions, obtaining metrics of central tendency and dispersion (mean, standard deviation, interquartile range), and evaluation of heterogeneous intra- and inter-machine behavior.

3. Topological Analysis of Variables: Stochastic quantification of quality deviations using advanced information theory metrics, such as the Jensen-Shannon Divergence, to isolate mechanical biases from the theoretical escandallo.
4. Predictive Modeling and Validation: Implementation of supervised Multi-output Regression learning techniques to predict the continuous proportions of the quality vector (Q1 to Q5), ensuring the algorithmic correction of the machinery.

3. Model/System/Tool Description

The computational ecosystem has been built entirely in Python, utilizing the standard stack for mathematical inference and Machine Learning (Pandas, NumPy, Scikit-Learn, and stochastic visualization environments). During the exploratory phase (EDA), severe asymmetries were identified in the processing of the three grading machines. As an empirical example, it was observed that machine C5 presents an extremely high standard deviation in large-sized fruits (>74mm) and an abnormally high proportion of defective quality classification (Q4), detecting barely 19% of optimal quality (Q1) compared to the 50% reported by lines C1 and C6.

Due to these anomalies, assuming a 1:1 linear relationship between the escandallo and the machinery (Baseline) proves ineffective. To model this complexity, Feature Engineering incorporated normalized variables (the escandallo proportions from Q1 to Q5) and subjected the machine identification to a mathematical categorical process via One-Hot Encoding.
For the predictive core, different algorithms were tested, prioritizing models based on decision tree ensembles (Random Forest Regressor) and penalized parametric approaches. The final selected algorithm is a Ridge Regression model. The L2 regularization inherent to Ridge optimally adapts to the problem, combating multicollinearity among features derived from the same proportion and endowing the system with extreme robustness against overfitting in fractional distributions.

4. Results

The performance evaluation of the algorithms reflects a drastic optimization of business indicators. The validation metric used to penalize proportional divergences was the Mean Absolute Error (MAE).
When evaluating the Baseline—which assesses the plant's current production model assuming the escandallo as a direct prediction—the model reached a very high systematic error, with an overall MAE of 22.85%. The introduction of the first supervised approximator using Random Forest managed to compress this margin of error to 9.51%. However, the system's predictive milestone is achieved thanks to the penalized linear architecture: the Ridge Regression model, trained with the spatial categorical variables of the grading machinery, reduces the overall MAE to 7.83% on the test set (unseen data). At a business impact level, this algorithm mitigates the deviation in prediction volume by 65.72% compared to the traditional grading method, proving to be an analytical framework capable of stabilizing quality inference in the citrus industry.

5. Conclusions

This report irrefutably demonstrates that traditional citrus escandallo processes lack the statistical sufficiency to guarantee proper production planning due to the noise injected by the heterogeneities of the packaging machinery itself. The deployed algorithmic development based on linearly regularized regressions (Ridge) mathematically compensates for these mechanical deficiencies, validating the impact that Data Science holds in the value chain of Industry 4.0.
The extrapolation of these results establishes a solid foundation for iterations in the project's Machine Learning roadmap. Proposed future work highlights:

1. Batch Profile Clustering: Unsupervised modeling aimed at aggregating isomorphic batches based on the population cross-referencing of their escandallos, pursuing the discovery of hidden dependencies by region or terrain.

2. Size-Quality Relationship: Two-dimensional topological analysis of the density function generated by physical volume versus its cataloging in the quality matrix.

3. Feature Space Enrichment: Evolution of the dataset by integrating environmental, biological, and industrial predictors, such as Total Weight (kg), absolute volume of fruits, spatial metadata from the agricultural field of origin, and biological variety of the crop.

6. Referencias

[1] Lin, J. “Divergence measures based on the Shannon entropy”, *IEEE Transactions on Information Theory*, Julio 1991, vol. 37, no. 4, pp. 145-151

[2] Hoerl, A.E.; Kennard, R.W. “Ridge Regression: Biased Estimation for Nonorthogonal Problems”, *Technometrics*, Febrero 1970, vol. 12, no. 1, pp. 55-67

[3] Borchani, H.; Varando, G.; Bielza, C.; Larrañaga, P. “A survey on multi-output regression”, *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, Septiembre 2015, vol. 5, no. 5, pp. 216-233.

Índice de la memoria

Capítulo 1. INTRODUCCIÓN

El advenimiento de la Cuarta Revolución Industrial y la democratización de la computación avanzada han catalizado una transformación sin precedentes en el tejido productivo global. En el marco de esta digitalización, el sector agroalimentario —y en particular la industria de procesamiento y exportación de cítricos— se encuentra en un punto de inflexión crítico. Históricamente caracterizada por procesos heurísticos y toma de decisiones fundamentada en la experiencia empírica, la industria requiere hoy de una reingeniería analítica para maximizar sus márgenes de eficiencia operativa y competitividad en un mercado globalizado.

El presente Trabajo Fin de Grado, denominado "Proyecto Cítrico", se enmarca en esta intersección entre la ingeniería de datos, el aprendizaje automático (*Machine Learning*) y la optimización de procesos industriales. Su núcleo discursivo y técnico aborda la resolución de una problemática endémica en las plantas de envasado de fruta: la divergencia estocástica entre los modelos de estimación de calidad teóricos (muestreos probabilísticos) y la clasificación empírica ejecutada por la maquinaria de calibración industrial. A través de la auditoría de datos y el despliegue de modelos predictivos de regresión multivariable, este proyecto propone un cambio de paradigma hacia la toma de decisiones fundamentada empíricamente (*Data-Driven Decision Making*).

1.1 MOTIVACIÓN DEL PROYECTO

La justificación y motivación de este proyecto se sustentan sobre tres pilares fundamentales: el impacto económico, la necesidad de innovación tecnológica y la mitigación de ineficiencias logísticas y operativas en la cadena de suministro agroalimentaria.

Desde una perspectiva económica y operativa, la rentabilidad de una planta de procesamiento de cítricos (como la planta de Sollana, objeto de estudio de esta memoria) depende intrínsecamente de su capacidad para prever el valor comercial de la materia prima entrante. En el modelo tradicional, cuando un lote de fruta ingresa a la instalación, se realiza un "escandallo": una extracción manual de una muestra reducida para inferir estadísticamente la distribución de calidades (desde Q1, correspondiente a una fruta de calidad suprema, hasta Q5, destinada a destrío o industria). Si bien este método probabilístico es un estándar en la industria, adolece de una severa miopía técnica al asumir que el comportamiento de los módulos mecánicos calibradores (máquinas C1, C5 y C6) será ideal e isomórfico respecto a la muestra humana.

El análisis de la realidad industrial demuestra lo contrario. Las máquinas presentan tolerancias mecánicas, desajustes ópticos y variaciones de rendimiento que inyectan un

ruido sistémico en la clasificación final. Por ejemplo, si el escandallo predice que un lote contiene un 50% de fruta de calidad Q1, pero la máquina C5 adolece de una calibración defectuosa y solo clasifica un 19% como tal, la planta se enfrenta a una rotura en la proyección de *stock*. Estas discrepancias provocan alteraciones graves en los compromisos comerciales, sobrecostes de re-procesamiento y una ineficiente asignación de recursos logísticos.

Desde el punto de vista tecnológico, la motivación reside en la oportunidad de sustituir un modelo estático por una arquitectura dinámica y auto-correctiva. La carencia de sistemas algorítmicos que auditen y aprendan del comportamiento asimétrico de la maquinaria representa un nicho de investigación de alto valor añadido. Por tanto, el desarrollo de este proyecto se justifica en la imperativa necesidad de dotar a las líneas de producción de una capa de inteligencia artificial capaz de aprender la función matemática que mapea el escandallo teórico con la realidad de las máquinas, garantizando así un control de calidad robusto, predecible y económicamente eficiente.

1.2 CONTEXTO INDUSTRIAL Y ALCANCE DEL PROBLEMA

El entorno de aplicación del proyecto se circunscribe a las operaciones de la planta de Sollana, analizando volumetrías masivas de fruta procesada y parametrizando sus correspondientes variables espaciales y morfológicas (calibres, peso total en kilogramos y matrices de proporción de calidad). El alcance de la investigación abarca todo el ciclo de vida del dato: desde la extracción e ingesta de los *datasets* productivos brutos, pasando por un exhaustivo Análisis Exploratorio de Datos (EDA) para aislar las anomalías mecánicas (utilizando métricas avanzadas como la divergencia de *Jensen-Shannon*), hasta llegar a la implementación final de un modelo supervisado de Regresión Multi-salida (*Multi-output Regression*).

El trabajo no pretende diseñar *hardware* de visión artificial para la clasificación unitaria del fruto, sino optimizar la inteligencia de negocio a nivel de lote o *batch*. El objetivo final es consolidar un modelo predictivo, específicamente basado en arquitecturas de regresión con regularización *L2* (*Ridge Regression*), que actúe como un gemelo digital del proceso de calibración, logrando reducir drásticamente el Error Absoluto Medio (MAE) respecto a las estimaciones vigentes de la fábrica.

1.3 ESTRUCTURA DE LA MEMORIA

Para proporcionar una narrativa lógica y coherente, así como facilitar la trazabilidad técnica del proceso investigador, la presente memoria se ha estructurado en los siguientes capítulos:

- **Capítulo 2. Descripción de las Tecnologías:** Expone el *stack* tecnológico y las librerías computacionales empleadas para el desarrollo del ecosistema predictivo, centrando la atención en los entornos de programación científica bajo lenguaje Python (Pandas, Scikit-Learn, NumPy).
- **Capítulo 3. Estado de la Cuestión:** Efectúa una revisión bibliográfica del estado del arte en materia de Machine Learning aplicado a la agroindustria y aborda el marco teórico de la resolución de problemas de predicción composicional mediante regresiones lineales regularizadas.
- **Capítulo 4. Definición del Trabajo:** Detalla exhaustivamente los objetivos generales y específicos de la investigación, la metodología de trabajo adoptada (basada en el estándar CRISP-DM) y el cronograma de planificación del proyecto.
- **Capítulo 5. Sistema/Modelo Desarrollado:** Constituye el núcleo técnico de la memoria. Describe las fases de Ingestión de Datos, Análisis Exploratorio de Variables (EDA), *Feature Engineering* e implementación algorítmica.
- **Capítulo 6. Análisis de Resultados:** Cuantifica el rendimiento estocástico de los modelos entrenados frente al *baseline* tradicional, evaluando las métricas de error y justificando la elección del modelo final.
- **Capítulo 7. Conclusiones y Trabajos Futuros:** Sintetiza los hallazgos principales, evalúa el cumplimiento de los objetivos e identifica las futuras líneas de investigación y escalabilidad del proyecto.

Capítulo 2. DESCRIPCIÓN DE LAS TECNOLOGÍAS

La arquitectura computacional, el procesamiento estocástico y el posterior despliegue del ecosistema algorítmico del "Proyecto Cítrico" se han fundamentado íntegramente sobre tecnologías de código abierto (*Open Source*), estandarizadas y ampliamente documentadas en el ámbito de la Ciencia de Datos (*Data Science*) y la Ingeniería de *Machine Learning*. La elección de este entramado tecnológico responde a criterios de alta disponibilidad, eficiencia en la manipulación de arreglos de datos masivos y robustez matemática para la implementación de predictores multivariables.

A continuación, se desglosan exhaustivamente los lenguajes, entornos, librerías y *frameworks* que constituyen la columna vertebral metodológica de la investigación.

2.1. Lenguaje Base y Entorno de Desarrollo Interactivo

- **Python:** El proyecto ha sido codificado en su totalidad empleando Python. Su elección se justifica por su paradigma multipropósito y su hegemonía indiscutible como lenguaje vehicular en el ecosistema actual de Inteligencia Artificial. Python ofrece una interoperabilidad excepcional entre el manejo de flujos de datos (*Data Pipelines*) y la ejecución de modelos estadísticos complejos, lo que permite trazar un continuo lógico desde la ingesta del dato bruto en la planta de Sollana hasta la inferencia lineal final.
- **Jupyter Notebooks (.ipynb):** Como entorno de desarrollo integrado (IDE), se ha optado por Jupyter Notebook. Esta herramienta es fundamental en las metodologías analíticas ágiles, ya que facilita una "programación literaria" (*Literate Programming*). Permite la partición del código en celdas de ejecución secuencial e interactiva, facilitando la interpolación de *scripts*, *outputs* matemáticos y la renderización en tiempo real de gráficos exploratorios. El proyecto se ha modularizado topológicamente en diversos *notebooks* funcionales (por ejemplo, `data_ingestion.ipynb`, `data_exploration.ipynb`, `data_analysis_variables.ipynb` y `model_testing.ipynb`), garantizando un aislamiento de dependencias y una auditoría de código eficiente.

2.2. Ecosistema de Procesamiento y Estructuración de Datos

La fase de Extracción, Transformación y Carga (ETL), así como el Análisis Exploratorio de Datos (EDA), demandan un procesamiento escalar capaz de gestionar la alta dimensionalidad paramétrica de los lotes de cítricos.

- **Pandas:** Esta librería constituye el núcleo del procesamiento algebraico del proyecto. Basada en la estructura de datos bidimensional conocida como *DataFrame*, Pandas se ha empleado masivamente para la vectorización relacional de los archivos `.csv` provenientes de la maquinaria (C1, C5 y C6). Ha sido la tecnología responsable de ejecutar cruces relacionales complejos (`pd.merge`), concatenaciones jerárquicas (`pd.concat`), imputaciones condicionales, derivación de proporciones geométricas fraccionales y la ejecución de transformaciones categóricas mediante la agregación de variables operacionales (`groupby` y tabulaciones matriciales *melt*).
- **NumPy (*Numerical Python*):** Empleada como motor matemático subyacente. NumPy proporciona soporte nativo para arreglos (*arrays*) matriciales multidimensionales y optimiza drásticamente las operaciones algebraicas mediante *broadcasting*. En el marco del proyecto, se ha utilizado activamente para operaciones de álgebra lineal, la acotación matemática del espacio de soluciones predictivas (mediante funciones como `np.clip` para evitar proporciones estocásticas ilógicas fuera del intervalo [0, 1]) y la normalización de la distribución fraccional para garantizar que la sumatoria del vector de calidades predicho compute un 100% de probabilidad neta de forma estricta.
- **Módulo OS:** Se ha integrado la librería nativa del sistema operativo (`os`) para la automatización dinámica de rutas de ingesta, el iterado recursivo de directorios ("data/raw_clean_data/") y la verificación de persistencia lógica de las fuentes de datos.

2.3. Herramientas de Visualización Estocástica y Topológica

La interpretación heurística de la dispersión del calibre y los sesgos algorítmicos mecánicos exige una potente suite de renderizado gráfico bidimensional.

- **Matplotlib (`matplotlib.pyplot`):** Actúa como la librería base de renderizado vectorial. Ha proporcionado el motor de bajo nivel para instanciar figuras paramétricas, gestionar mallas topológicas, delimitar ejes cartesianos multivariables y ensamblar composiciones visuales complejas para el análisis volumétrico comparado.
- **Seaborn:** Construida sobre Matplotlib, Seaborn aporta una abstracción analítica de alto nivel indispensable para el EDA del proyecto. Se ha implementado para proyectar estimaciones de densidad bidimensional y mapas categóricos, empleando hojas de estilo preconfiguradas (como `ggplot` y *themes* `whitegrid`) y paletas perceptualmente uniformes (`husl`, `Set2`, `viridis`). Estas características han sido

explotadas para visualizar con claridad las distribuciones intercuartílicas (mediante diagramas de cajas o *boxplots*) y las volumetrías agregadas de los calibres (segmentados condicionalmente por calibrador mediante el parámetro *hue*).

2.4 Framework Analítico y de Aprendizaje Automático

El núcleo predictivo para el desarrollo del Gemelo Digital estocástico de las máquinas de calibración se apoya fundamentalmente en la librería científica estándar de Python para inferencia supervisada y no supervisada.

- **Scikit-Learn (`sklearn`):** Toda la canalización paramétrica (*Pipeline*) del modelo reside en la arquitectura de esta librería. Su estructura modular se ha empleado en las siguientes capas de complejidad:
  - **Procesamiento y Validación Muestral:** Utilización de `train_test_split` del módulo de selección de modelos, garantizando un particionado inmaculado y reproducible (*random_state*) para retener tensores de testeo (20% de la volumetría de *batch*) que aseguren una validación cruzada y eviten la contaminación estocástica o filtración de datos (*data leakage*).
  - **Entrenamiento Algorítmico y Regularización:** Scikit-Learn ha provisto la instanciación teórica y empírica de diversas familias de algoritmos ensayados, abarcando desde aproximadores paramétricos basados en entropía informática (como `DecisionTreeRegressor` y ensamblajes de `Random Forest`) hasta modelos generalizados linealmente. El nodo principal del trabajo radica en la importación de `Ridge` (desde `sklearn.linear_model`), aprovechando su regularización L2 intrínseca para procesar variables dependientes correlacionadas y emitir regresiones de salida múltiple de manera simultánea (*Multi-output Regression*).
  - **Métricas de Coste y Evaluación:** Cuantificación rigurosa de la pérdida predictiva (*Loss Functions*) importando los módulos `mean_absolute_error` (MAE, utilizado como KPI crítico global y particionado mediante `multioutput='raw_values'`), `mean_squared_error` (MSE) y el Coeficiente de Determinación estocástica (`r2_score`), así como evaluadores iterativos profundos (`cross_val_score` apoyado en distribuciones en K-Fold).

2.5 Gestión de Configuraciones y Trazabilidad del Software

A nivel de Ingeniería de Software e Integración de Código, el "Proyecto Cítrico" asimila flujos de trabajo contemporáneos propios del despliegue en entornos productivos.

- **Git y GitHub:** La semántica de las iteraciones de código, la preservación cronológica de las configuraciones y el blindaje frente a roturas analíticas se han orquestado mediante el sistema de control de versiones distribuido Git. Toda la arquitectura se halla unificada y empaquetada lógicamente en el ecosistema alojado en el repositorio "proyecto-citrico" de la plataforma remota GitHub. Para sostener un desarrollo aséptico y evitar el volcado de la dimensionalidad bruta y transitoria, se ha configurado un indexado con reglas estrictas de persistencia mediante archivos `.gitignore`, afianzando así las mejores prácticas en el ámbito del *Software Craftsmanship* aplicado a la Ciencia de Datos.

Capítulo 3. ESTADO DE LA CUESTIÓN

La intersección entre la analítica de datos avanzada y la ingeniería agroindustrial ha experimentado un desarrollo exponencial durante la última década, impulsada por el paradigma de la Industria 4.0. En el ámbito específico del procesamiento y exportación de cítricos, la optimización de los estándares de calidad ha transitado desde metodologías estrictamente manuales hacia sistemas progresivamente automatizados. Sin embargo, una revisión exhaustiva de la literatura técnica y del mercado actual revela brechas significativas en la monitorización estocástica y la predicción volumétrica a nivel de lote, un vacío tecnológico que el presente proyecto aspira a cubrir.

3.1 EVOLUCIÓN DEL CONTROL DE CALIDAD AGROALIMENTARIO Y EL USO DEL ESCANDALLO

Históricamente, la determinación de la calidad comercial en plantas de procesamiento de fruta se ha fundamentado en el "escandallo", un proceso de muestreo estadístico probabilístico. Desde la perspectiva de la teoría de muestras, el escandallo asume que la varianza observada en una extracción manual finita es representativa de la población total (el lote completo).

En la literatura agronómica clásica, esta asunción es válida bajo condiciones de procesamiento ideales. No obstante, en la realidad de la ingeniería de planta, la clasificación final no recae sobre el técnico que realiza el escandallo, sino sobre la maquinaria de calibración (líneas C1, C5 y C6). Estudios empíricos recientes demuestran que la maquinaria industrial está sometida a desgaste mecánico, desajustes en los sensores ópticos y derivas de calibración. En consecuencia, la dependencia exclusiva del escandallo genera un "sesgo de confirmación logística", donde se planifica la producción en base a una verdad teórica que la maquinaria es incapaz de reproducir en la práctica.

3.2 VISIÓN ARTIFICIAL Y SUS LIMITACIONES MACROSCÓPICAS

Para combatir las ineficiencias del muestreo manual, el estado del arte en la industria se ha volcado masivamente hacia la integración de sistemas de Visión Artificial (VA). La bibliografía actual abunda en investigaciones que emplean arquitecturas de Aprendizaje Profundo (*Deep Learning*), específicamente Redes Neuronales Convolucionales (CNNs), para la detección de defectos superficiales, colorimetría y dimensionamiento morfológico (calibre) de frutos individuales en tiempo real.

Si bien estas tecnologías son punteras para la inspección microscópica y unitaria, presentan una seria limitación macroscópica: son sistemas reactivos, no predictivos. Un algoritmo de visión artificial clasifica la fruta a medida que pasa por la cinta, pero no es capaz de prever con antelación cuál será la distribución final de calidades (Q1 a Q5) de un lote completo que acaba de ingresar al almacén, ni puede anticipar el rendimiento diferencial entre dos máquinas calibradoras distintas. La predicción de estas métricas volumétricas agregadas exige un enfoque probabilístico que trascienda la mera clasificación por imágenes.

3.3 ALGORITMIA PREDICTIVA PARA DISTRIBUCIONES COMPOSICIONALES

El problema de predecir la distribución final de calidades de un lote se enmarca matemáticamente en el análisis de datos composicionales, dado que las variables de salida (el porcentaje de Q1, Q2, Q3, Q4 y Q5) son fracciones continuas cuya sumatoria debe ser estrictamente igual al 100% (o 1 en términos de probabilidad).

En el estado de la cuestión del *Machine Learning* clásico, la resolución de este tipo de problemas se aborda mediante técnicas de Regresión Multi-salida (*Multi-output Regression*), donde un único modelo es entrenado para inferir múltiples variables dependientes de forma simultánea. La revisión de trabajos algorítmicos similares destaca la aplicación de dos grandes familias de aproximadores para este dominio:

1. **Ensamblajes basados en Árboles de Decisión (*Random Forest*, *Gradient Boosting*):** Son altamente eficaces para capturar relaciones no lineales complejas. Sin embargo, la literatura advierte que en escenarios donde los *datasets*presentan un ruido estocástico elevado —como las fluctuaciones erráticas de una máquina descalibrada—, estos modelos tienden a sufrir de sobreajuste severo (*overfitting*), memorizando el ruido en lugar de generalizar la tendencia subyacente.
2. **Modelos Lineales Regularizados (*Ridge Regression*, *Lasso*):** Estos enfoques abordan el problema desde la optimización convexa. Investigaciones recientes en analítica predictiva industrial concluyen que la penalización *L2*inherente a la Regresión *Ridge* es extremadamente eficaz en espacios de características (*features*) con alta multicolinealidad. En el contexto de este proyecto, las variables predictivas derivadas del escandallo (ej. proporción inicial de Q1 y Q2) están matemáticamente correlacionadas. La regularización *Ridge* estabiliza la varianza de los coeficientes, ofreciendo estimaciones mucho más robustas y generalizables frente al desgaste asimétrico de la maquinaria.

3.4 JUSTIFICACIÓN DE LA SOLUCIÓN PROPUESTA FRENTE AL MERCADO

Al realizar una prospección de soluciones comerciales, se constata que no existen *softwares* estándar en el mercado agroalimentario que integren modelos de *Machine Learning* para auditar paramétricamente y corregir el sesgo predictivo de máquinas calibradoras específicas empleando el escandallo como *input* basal.

El "Proyecto Cítrico" representa, por tanto, una innovación disruptiva frente al estado actual de la técnica. Al implementar una arquitectura de Regresión Multi-salida fundamentada en modelos regularizados (*Ridge*), el sistema no solo cubre las carencias del muestreo probabilístico estático, sino que dota a la infraestructura industrial de un Gemelo Digital estocástico. Esto permite transformar una planta de procesamiento reactiva en un entorno predictivo proactivo, capaz de anticipar sus propias ineficiencias mecánicas y salvaguardar su eficiencia operativa y comercial.

Capítulo 4. DEFINICIÓN DEL TRABAJO

El presente capítulo tiene como propósito vertebrar la estructura conceptual y organizativa del "Proyecto Cítrico". A partir del análisis de las carencias tecnológicas expuestas en el estado de la cuestión, se procede a justificar la viabilidad y necesidad de la solución algorítmica propuesta, delimitando de manera unívoca los objetivos a alcanzar, la metodología de desarrollo adoptada y la viabilidad económica de su despliegue en un entorno industrial.

4.1 JUSTIFICACIÓN

La justificación de este trabajo se sustenta sobre la necesidad imperativa de transformar el paradigma de control de calidad en el sector citrícola, transitando de un modelo estático y probabilístico a un ecosistema dinámico y fundamentado en datos empíricos (*Data-Driven*).

A la vista de las ineficiencias inherentes a los procesos de envasado y exportación de la planta de Sollana, el problema central radica en el "ruido" que inyectan los propios módulos de calibración (máquinas C1, C5 y C6) sobre la calidad proyectada del lote. Cuando la planificación logística y comercial confía ciegamente en el muestreo manual ("escandallo"), asume una linealidad mecanicista que resulta ser falaz. Las tolerancias de los sensores ópticos, el desgaste de las cintas de triaje y las variaciones volumétricas generan divergencias severas entre la teoría (lo que dice el escandallo que hay) y la práctica (lo que la máquina finalmente clasifica en calidades Q1 a Q5).

Este proyecto no es únicamente un ejercicio de exploración académica en el campo del *Machine Learning*, sino una respuesta directa a una necesidad de negocio. Su desarrollo proveerá a la entidad colaboradora de un Gemelo Digital estocástico capaz de prever con antelación el rendimiento real de un lote específico en una máquina concreta. Al mitigar el margen de error entre la previsión y la realidad, la planta optimiza la gestión de *stock*, minimiza los cuellos de botella derivados del reprocesamiento de fruta mal calibrada y garantiza la satisfacción de los compromisos de calidad adquiridos con los distribuidores finales. En definitiva, este proyecto aporta una ventaja competitiva diferencial basada en la inteligencia artificial y la ciencia de datos.

4.2 OBJETIVOS

La articulación del "Proyecto Cítrico" se rige por un propósito central y un conjunto de metas específicas diseñadas para garantizar la trazabilidad analítica y el éxito de la implementación.

Objetivo General: Diseñar, codificar y validar un *pipeline* analítico y predictivo basado en algoritmos de aprendizaje supervisado (Regresión Multi-salida) que logre inferir con alta fidelidad la distribución volumétrica real de las calidades cítricas, corrigiendo sistemáticamente el sesgo operativo de la maquinaria industrial respecto a las estimaciones iniciales de muestreo.

Objetivos Específicos:

1. **Ingeniería y Refinamiento de Datos:** Construir una arquitectura ETL (Extracción, Transformación y Carga) robusta para integrar métricas operacionales crudas, imputar valores faltantes y generar variables derivadas (proporciones fraccionales de escandallo y categorización geométrica).
2. **Análisis Topológico e Inferencia Estadística:** Ejecutar un Análisis Exploratorio de Datos (EDA) en profundidad que cuantifique estocásticamente las anomalías del proceso, utilizando herramientas de la Teoría de la Información (Divergencia de Jensen-Shannon) para medir la distancia probabilística entre las distribuciones teóricas y empíricas.
3. **Modelado Computacional Multivariable:** Evaluar y parametrizar diversas familias de algoritmos predictivos, comparando el rendimiento de aproximadores no lineales (ensamblajes *Random Forest*) frente a regresores lineales penalizados (*Ridge Regression* con regularización L2).
4. **Validación y Retorno de Inversión (ROI) Tecnológico:** Minimizar el Error Absoluto Medio (MAE) predictivo de la planta en al menos un 50% con respecto a la heurística tradicional (el *baseline* de la fábrica), asegurando que el modelo sea integrable, interpretable y ofrezca garantías operacionales reales.

4.3 METODOLOGÍA

Para dotar al desarrollo de un marco metodológico estandarizado, iterativo y auditable, se ha adoptado una adaptación de CRISP-DM (Cross-Industry Standard Process for Data Mining), el cual es el protocolo de facto en la industria del análisis predictivo. Las fases ejecutadas han sido:

Comprensión del Negocio (Business Understanding): Inmersión en las reglas lógicas del proceso de calibración, definición de las métricas de éxito (MAE) y entendimiento de las desviaciones tolerables por los jefes de planta.

Comprensión de los Datos (Data Understanding): Recopilación de históricos a través de cuadernos de ingesta (data_ingestion.ipynb), evaluando la dimensionalidad, los tipos de variables y el nivel de ruido estocástico presente en los ficheros CSV provenientes de los PLC de la maquinaria.

Preparación de los Datos (Data Preparation): Fase más crítica del proyecto. Ha englobado la normalización de vectores, Feature Engineering (como la codificación One-Hot de las variables de las máquinas C1, C5 y C6) y la acotación de tensores para aislar la variable objetivo (distribución Q1-Q5).

Modelado (Modeling): Entrenamiento computacional de los algoritmos mediante particionado Train/Test Split, asegurando la no contaminación de los datos de validación y aplicando validación cruzada.

Evaluación (Evaluation): Contraste empírico del modelo óptimo contra el flujo de trabajo vigente. Esta fase iterativa permitió refinar el hiperparámetro de regularización (α) en el modelo de Ridge hasta alcanzar el mínimo global de la función de coste.

4.4 PLANIFICACIÓN Y ESTIMACIÓN ECONÓMICA

La ejecución y gestión del proyecto se ha temporalizado en un horizonte de cuatro meses, regido bajo un marco de trabajo ágil (*Agile-Scrum*) con ciclos de iteración (*Sprints*) bisemanales. Esta segmentación ha permitido pivotar la estrategia de modelado a medida que el análisis exploratorio revelaba la alta multicolinealidad de los datos.

1. **Mes 1:** Requisitos de negocio, auditoría de la fuente de datos y creación del flujo de ingesta automatizado.
2. **Mes 2:** Análisis Exploratorio de Datos (EDA) exhaustivo, descubrimiento de la carencia de predicción de la máquina C5 y desarrollo del *Feature Engineering*.
3. **Mes 3:** Fase intensiva de Modelado predictivo, ajuste de hiperparámetros (*tuning*) y validación de la Regresión *Ridge*.
4. **Mes 4:** Consolidación de resultados, cuantificación del MAE y redacción de la presente memoria técnica.

En términos de **Estimación Económica**, asumiendo el desarrollo por parte de un perfil de *Junior Data Scientist*(Ingeniero de Datos/Machine Learning), se plantea un presupuesto de inversión inicial:

- **Costes de Ingeniería y Desarrollo:** Estimando 300 horas de dedicación a una tarifa técnica estándar de 35 €/hora, el coste de desarrollo de *software* asciende a 10.500 €.

- **Costes de Infraestructura y Licenciamiento:** Amortización de *hardware* computacional (estación de trabajo para entrenamiento local) y prorrateo de licencias y almacenamiento en la nube, presupuestado en 800 €.
- **Coste Total Estimado:** El proyecto presenta un presupuesto de desarrollo de aproximadamente **11.300 €**.

Este capital de inversión se cataloga como un activo de altísima rentabilidad para la planta. Al lograr reducir un 65.72% el margen de error en la proyección volumétrica de calidades, el ahorro logístico derivado de evitar roturas de *stock* y el descenso en los costes de reprocesamiento mecánico garantizan el retorno de la inversión (ROI) durante la primera campaña de exportación agrícola tras la implantación del modelo.

Capítulo 5. SISTEMA/MODELO DESARROLLADO

El presente capítulo constituye el núcleo empírico y tecnológico del "Proyecto Cítrico". En él se detalla la arquitectura algorítmica diseñada para modelar y predecir el comportamiento estocástico de las líneas de envasado en la planta de procesado. La exposición se estructura de manera secuencial, partiendo del análisis algebraico del problema operativo, transitando por el diseño conceptual del pipeline de datos, y culminando en la implementación en código del modelo de Regresión Multi-salida mediante técnicas de regularización paramétrica.

5.1 ANÁLISIS DEL SISTEMA

El análisis del sistema parte de una deconstrucción exhaustiva del flujo físico y de información que gobierna la planta de Sollana. El objetivo es traducir un problema logístico industrial en un problema matemático de optimización y aprendizaje supervisado.

5.1.1 DEFINICIÓN DEL DOMINO DEL PROBLEMA

El proceso productivo actual se rige por un muestreo estadístico denominado "escandallo". Cuando un lote de cítricos ingresa a la planta, se extrae una submuestra para estimar teóricamente su distribución de calidades comerciales, categorizadas en cinco niveles continuos y excluyentes: Q1 (calidad óptima o suprema), Q2, Q3, Q4 y Q5 (destrío o uso industrial).

Desde el punto de vista del sistema, el escandallo actúa como la señal de entrada teórica (). Sin embargo, la clasificación real del lote completo es ejecutada empíricamente por los módulos mecánicos de calibración (máquinas C1, C5 y C6), generando una señal de salida empírica (). La hipótesis nula () del modelo de negocio tradicional asume que . No obstante, el análisis exploratorio preliminar del sistema refutó categóricamente esta hipótesis, evidenciando que las máquinas inyectan un sesgo o "ruido mecánico" severo.

5.1.2 CUANTIFICACIÓN DE ANOMALÍAS MECÁNICAS

La auditoría de los registros históricos desveló asimetrías críticas en el rendimiento de los módulos calibradores frente a lotes isomorfos. El caso más acentuado es el de la máquina C5. Para un lote cuyo escandallo proyecta una proporción del 50% de fruta de calidad Q1, las líneas C1 y C6 logran aproximarse a dicha cifra; sin embargo, la máquina C5 categoriza, en promedio, apenas un 19% del volumen como Q1, derivando el resto hacia categorías inferiores (especialmente Q4). Esta divergencia sistemática introduce un error absoluto masivo en la planificación logística.

5.1.3 MODELADO MATEMÁTICO DEL SISTEMA

Para corregir esta miopía operativa, el sistema se ha analizado como un problema algorítmico de **Regresión de Salida Múltiple con Datos Composicionales**.

- **Espacio de Características (X):** Compuesto por las proporciones iniciales del escandallo, el peso total del lote (normalizado), la distribución de calibres, y una dimensión espacial que identifica a la máquina procesadora.
- **Espacio de Predicción (Y):** Un vector continuo de cinco dimensiones $[y_{Q1}, y_{Q2}, y_{Q3}, y_{Q4}, y_{Q5}]$ donde cada escalar representa la fracción porcentual predicha.
- **Restricciones Topológicas:** Dado que se trata de un volumen físico cerrado, el sistema debe cumplir la restricción composicional estricta donde la suma del vector de salida debe ser igual a 1 (o 100%): $\sum_{i=1}^{5} y_{Qi} = 1$, acotando cada variable en el intervalo $[0, 1]$.

5.2 DISEÑO

El diseño de la solución analítica abandona las heurísticas lineales simples a favor de una arquitectura de Machine Learning basada en el aprendizaje de representaciones complejas. La topología de la solución se segmenta en tres capas lógicas.

5.2.1 CAPA DE INGESTIÓN Y TRANSFORMACIÓN GEOMÉTRICA (ETL)

El diseño de los datos requiere transformar volcados de información cruda (archivos CSV provenientes de los PLCs) en tensores de alta dimensionalidad computables. El sistema se diseña para:

1. **Agregación Relacional:** Pivotar las columnas de calibres y agrupar los registros por el identificador único del lote.
2. **Normalización Fraccional:** Transformar los kilogramos brutos en porcentajes relativos. Esto inmuniza al algoritmo frente a la varianza en el tamaño total de los lotes entrantes, permitiendo que el modelo aprenda patrones de "distribución" y no magnitudes absolutas.

5.2.2 INGENIERÍA DE CARACTERÍSTICAS (FEATURE ENGINEERING)

Para que el modelo matemático comprenda el sesgo de las diferentes calibradoras, el diseño estructural incorpora codificación espacial. Las variables categóricas representativas de las máquinas (C1, C5, C6) no poseen un orden jerárquico natural. Por tanto, se diseñó una proyección ortogonal mediante One-Hot Encoding. Esta técnica expande el tensor de entrada añadiendo vectores binarios, permitiendo al algoritmo asignar pesos de regresión independientes a la "identidad" de la máquina, aislando matemáticamente el déficit de calibración de la máquina C5.

5.2.3 SELECCIÓN DE LA ARQUITECTURA PREDICTIVA

El diseño del núcleo de Machine Learning evaluó múltiples topologías algorítmicas:

- *Árboles de Decisión y Random Forest:* Aunque excelentes para modelar relaciones no lineales, el diseño los descartó como modelo final. La extrema varianza y el ruido inyectado por las máquinas provocaba un severo sobreajuste (*overfitting*), donde el ensamblaje memorizaba el ruido estocástico en lugar de generalizar la tendencia de calibración.
- *Regresión Lineal Múltiple:* Descartada debido a la alta multicolinealidad de las variables de entrada (las fracciones de un escandallo están intrínsecamente correlacionadas porque suman 100).
- **Ridge Regression (Modelo Seleccionado):** El diseño óptimo convergió en un modelo de Regresión *Ridge* con regularización L2. Al aplicar una penalización al cuadrado de la magnitud de los coeficientes ($\alpha \sum \beta_i^2$), la regularización de *Ridge* mitiga magistralmente la multicolinealidad, ofreciendo un mapa de pesos distribuido y estable que ignora el ruido transitorio de las máquinas y captura la tendencia subyacente.

5.3 IMPLEMENTACIÓN

La codificación e instanciación empírica del diseño se orquestó íntegramente en Python, empleando cuadernos interactivos (Jupyter Notebooks) para modularizar el ciclo de vida del software (data_ingestion.ipynb, data_exploration.ipynb, model_testing.ipynb).

5.3.1 ESTRUCTURA DEL REPOSITORIO

Para garantizar la reproducibilidad, el aislamiento de dependencias y la modularidad del código, la ejecución del proyecto se ha estructurado lógicamente en cuatro cuadernos interactivos principales alojados en el repositorio:

- **`data_ingestion.ipynb`**: Orquesta la fase de Extracción, Transformación y Carga (ETL), ejecutando la lectura de los ficheros brutos y la consolidación relacional de los registros de las máquinas.

- **`data_exploration.ipynb`**: Contiene el Análisis Exploratorio de Datos (EDA), generando las visualizaciones topológicas bidimensionales y cuantificando las asimetrías severas de las calibradoras C1, C5 y C6.

- **`data_analysis_variables.ipynb`**: Dedicado a la Ingeniería de Características (Feature Engineering) y a la estructuración de las variables categóricas, como la codificación One-Hot para el identificador de la máquina.

- **`model_testing.ipynb`**: Constituye el entorno de validación empírica donde se instancian, entrenan y auditan los algoritmos predictivos (Random Forest y Ridge Regression), computando el Error Absoluto Medio (MAE) sobre el conjunto de testeo.

- Adicionalmente, la estructuración de la defensa del trabajo se apoya en el documento **`presentation_outline.md`**, que asegura la correcta alineación discursiva entre la memoria técnica y la exposición oral de los resultados.

5.3.2 IMPLEMENTACIÓN DEL PROCESAMIENTO DE DATOS

La capa ETL se programó empleando las librerías pandas y numpy. Mediante rutinas de des-anidación (pd.melt) y agregación iterativa (groupby), se unificaron los ficheros de las calibradoras en un único DataFrame maestro (el dataset analítico). Las operaciones de álgebra lineal implementadas con numpy aseguraron que las proporciones se mantuviesen estrictamente acotadas, empleando funciones de saneamiento (np.clip()) para evitar aberraciones estadísticas como porcentajes negativos de fruta. A continuación se muestra un ejemplo de preprocesamiento para los datos del escandallo. La consolidación del *dataset* maestro agrupa tanto la información probabilística de los escandallos como la parametrización física de los lotes. Las variables espaciales, biológicas y operativas extraídas tras el preprocesamiento incluyen:

- **`IDViaje` y `Numero_lot`**: Identificadores numéricos únicos que aseguran la trazabilidad del volumen físico a lo largo de toda la cadena de procesamiento logístico.
- **`Variedad`**: Clasificación biológica y comercial del cultivo cítrico procesado, destacando tipos como NADORCOTT, TANG GOLD o LEANRI.
- **`ZonaRecoleccion` y `termino`**: Metadatos geográficos que indican la procedencia regional y municipal del lote (por ejemplo, HUELVA o MURCIA, y términos como Ayamonte o Lorca).
- **`Fecha`**: Marca temporal estandarizada de ingreso del lote, vital para el futuro análisis de estacionalidad.
- **Distribución de calidades (`Q1`, `Q2`, `Q3`, etc.)**: Variables numéricas continuas que representan la estimación volumétrica asignada a cada categoría comercial específica.

	Compra	IDViaje	Variedad	ZonaRecoleccion	termino	Fecha	Q1	Q2	Q3	Numero_lot
0	1601190067	992823	NADORCOTT	HUELVA	SAN SILVESTRE DE GUZMAN	2026-01-21	20568.3360	1037.5140	54.1500	2088
1	1601870010	995156	TANG GOLD	HUELVA	MORON DE LA FRONTERA	2026-01-31	21961.8856	492.1368	17.9776	2114
2	1601870063	994354	TANG GOLD	HUELVA	AYAMONTE	2026-01-28	18595.4990	985.4260	49.0750	2107
3	1601900001	991940	LEANRI	HUELVA	GIBRALEON	2026-01-16	21703.8192	808.1808	0.0000	2059
4	1601900002	993266	LEANRI	HUELVA	V. CASTILLEJOS	2026-01-23	21178.3847	775.4351	13.1802	2081
5	2802190056	991760	NADORCOTT	MURCIA	S.PEDRO DEL PINATAR	2026-01-15	20056.9620	2308.9245	29.1135	2064
6	2812870004	992410	TANG GOLD	MURCIA	MAZARRON	2026-01-19	21926.6460	2539.2360	44.1180	2077
7	2812870017	995916	TANG GOLD	MURCIA	LORCA	2026-02-03	21040.5440	2680.0060	59.4500	2120
8	2812870017	996478	TANG GOLD	MURCIA	LORCA	2026-02-06	16922.4960	3324.3520	73.1520	2124
9	2812870026	994722	TANG GOLD	MURCIA	CUEVAS DE ALMANZORA	2026-01-29	16684.2000	4005.4500	10.3500	2110

5.3.3 IMPLEMENTACIÓN DEL ENTRENAMIENTO DEL MODELO

La instanciación algorítmica se apoyó en el *framework* `scikit-learn`. Para garantizar la rigurosidad científica de la prueba, se implementó una rutina de segregación ciega (`train_test_split`), particionando el volumen de datos en un 80% para entrenamiento (optimización del gradiente) y un 20% para inferencia y validación (test). Este particionado se blindó con una semilla pseudoaleatoria (`random_state`) para asegurar la reproducibilidad determinista del experimento.

El núcleo predictivo se programó encapsulando el regresor `Ridge` dentro del metamodelo `MultiOutputRegressor`. Esto permitió a la infraestructura computacional desplegar simultáneamente cinco ecuaciones de hiperplanos regularizados, una para cada vector de calidad objetiva (Q1, Q2, Q3, Q4, Q5).

5.3.4 OPTIMIZACIÓN DE HIPERPARÁMETROS Y POST-PROCESAMIENTO

La fase de *tuning* exigió la iteración sobre el hiperparámetro de regularización (Alpha) del modelo *Ridge*. Mediante un bucle de validación cruzada, se minimizó la función de coste seleccionada: el Error Absoluto Medio (MAE).

Finalmente, la implementación requirió una subrutina de post-procesamiento. Dado que una regresión lineal sobre cinco ejes independientes no garantiza de forma nativa que la suma de las salidas sea exactamente 100%, se codificó una función de ajuste algebraico en `numpy` que toma el vector predictivo resultante y recalcula los pesos relativos de Q1-Q5, forzando matemáticamente la coherencia de la restricción composicional del volumen del lote antes de emitir la predicción a los paneles de control de la fábrica.

Capítulo 6. ANÁLISIS DE RESULTADOS

Este capítulo presenta la validación cuantitativa y cualitativa del ecosistema predictivo diseñado para la planta de Sollana. El objetivo central de esta fase es contrastar empíricamente la robustez de los algoritmos de Machine Learning implementados frente a la heurística tradicional de la planta, demostrando la viabilidad de la Regresión Ridge como motor del gemelo digital estocástico.

6.1 DEFINICIÓN DE LAS MÉTRICAS DE EVALUACIÓN

Para medir el rendimiento de los modelos frente a un problema de regresión multidimensional de datos composicionales, es imperativo establecer una función de coste que penalice de forma lineal las desviaciones en las predicciones volumétricas. La métrica principal seleccionada para auditar el desempeño del sistema ha sido el Error Absoluto Medio (MAE, por sus siglas en inglés).

Dado que el sistema predice simultáneamente un vector de cinco dimensiones (calidades Q1 a Q5) para lotes de fruta, el MAE global se ha formulado matemáticamente como la media de los errores absolutos a través de todas las muestras y todas las salidas. En sintaxis LaTeX, la formulación empleada es la siguiente:

$$MAE_{global} = \frac{1}{n \cdot m} \sum_{i=1}^{n} \sum_{k=1}^{m} |y_{i,k} - \hat{y}_{i,k}|$$

Donde:

- representa el número total de lotes procesados en el conjunto de validación (*Test Set*).
- representa la dimensionalidad del vector de salida (, correspondiente a las categorías $Q_1, Q_{2,}, Q_3, Q_4, Q_5$).
- $y_{i,k}$ es la proporción real de la calidad para el lote i dictaminada por la máquina.
- $\hat{y_{i,k}}$ es la proporción inferida por el modelo predictivo (o por el escandallo en el caso del *Baseline*).

6.2 AUDITORÍA DEL ESCENARIO OPERATIVO BASE (BASELINE)

El primer hito analítico consistió en cuantificar el margen de error del flujo de trabajo actual. La hipótesis de negocio preexistente asume que la estimación del escandallo es un predictor directo y unívoco de la realidad. Para evaluar este *Baseline*, se calculó el MAE_{global} asumiendo que el tensor de predicciones $\hat{Y}$ es exactamente igual al tensor de entrada del escandallo $X_{teórica}$.

Los resultados revelan una ineficiencia estructural severa en la planta. El modelo estático tradicional arrojó un MAE_{global} del **22.85%**. Esta divergencia se debe fundamentalmente a la incapacidad del escandallo para anticipar el comportamiento asimétrico de la maquinaria; en particular, la severa sub-clasificación de fruta de calidad Q1 que ejecuta la calibradora C5 (la cual desplaza artificialmente los volúmenes hacia la calidad Q4). Este error sistemático justifica la imperativa necesidad de implementar los modelos supervisados.

6.3 DESEMPEÑO DE LOS MODELOS COMPUTACIONALES

La fase de validación algorítmica enfrentó diferentes arquitecturas contra el conjunto de datos de testeo. Se presentan a continuación los resultados de los dos enfoques principales:

6.3.1 MODELO DE ENSAMBLAJE: RANDOM FOREST REGRESSOR

La primera aproximación mediante aprendizaje automático no paramétrico logró una mejora sustancial. El ensamblaje de árboles de decisión fue capaz de capturar parcialmente las interacciones no lineales entre las variables espaciales y los calibres, reduciendo el error a un 9.51%. Sin embargo, el análisis de los residuos evidenció que este algoritmo sufría un leve sobreajuste (overfitting) al intentar trazar límites de decisión excesivamente complejos sobre el ruido estocástico inherente a los sensores ópticos de las máquinas.

6.3.2 MODELO PARAMÉTRICO REGULARIZADO: RIDGE REGRESSION

El algoritmo óptimo, fundamentado en la optimización convexa, superó con creces al resto de aproximadores. La aplicación de la regresión Ridge implicó la minimización de la siguiente función objetivo regularizada

$$\min_{W}(||XW - Y||_2^2 + \alpha||W||_2^2)$$

La penalización L_2 (el término $\alpha\|W\|_2^2$) demostró ser el mecanismo perfecto para combatir la alta multicolinealidad de las proporciones del escandallo. Al estabilizar los pesos (W) asociados a las variables categóricas espaciales (como la identidad One-Hot de la máquina C5), el modelo Ridge alcanzó un MAE_{global} de apenas 7.83% sobre datos no vistos.

6.4 ANÁLISIS CRÍTICO Y RETORNO OPERATIVO

La transición del *Baseline* tradicional (22.85% de error) al modelo predictivo optimizado (7.83% de error) supone una **reducción neta del 65.72% en la desviación estocástica de la planta**.

Desde una perspectiva industrial, este resultado es transformador. El algoritmo ha aprendido con éxito a modelar la función de desgaste y el sesgo de calibración de cada máquina individual. Cuando el modelo predice el rendimiento de un lote asignado a la calibradora C5, ya no confía ciegamente en el 50% de Q1 que promete el escandallo; en su lugar, aplica un factor de atenuación aprendido matemáticamente, reajustando la expectativa hacia el 19% real y re-distribuyendo probabilísticamente el volumen restante hacia categorías inferiores de forma automatizada y coherente (garantizando que $\sum y_k = 1$).

En conclusión, los resultados validan la hipótesis central de la investigación: la aplicación de técnicas de regresión lineal regularizada proporciona un marco de toma de decisiones inmensamente superior al muestreo estadístico tradicional, dotando a la entidad colaboradora de una resiliencia logística sin precedentes frente a las ineficiencias de su propia infraestructura mecatrónica.

Capítulo 7. CONCLUSIONES Y TRABAJOS FUTUROS

Este último capítulo sintetiza los hallazgos empíricos y analíticos derivados de la ejecución del "Proyecto Cítrico". Se evalúa de manera crítica el grado de cumplimiento de los objetivos fundacionales del Trabajo Fin de Grado, se exponen las contribuciones metodológicas aportadas a la industria agroalimentaria y se trazan las líneas de investigación futuras para escalar la solución desarrollada.

7.1 CONCLUSIONES DEL PROYECTO

La investigación ha demostrado de manera fehaciente que los modelos de estimación probabilística tradicionales (el "escandallo") carecen de la robustez necesaria para gobernar la planificación productiva en entornos industriales complejos. La presunción de que la maquinaria de calibración actuará de forma determinista y fiel a la muestra manual ha sido refutada mediante el análisis de datos. Se ha evidenciado empíricamente que módulos como la calibradora C5 introducen un sesgo estocástico severo, subestimando de manera crónica la fruta de calidad suprema (Q1) e inyectando un ruido operativo que contamina toda la cadena de suministro.

Frente a esta problemática, el desarrollo e implementación del modelo predictivo basado en Inteligencia Artificial ha supuesto un salto cualitativo en la inteligencia de negocio de la planta. Las conclusiones técnicas más relevantes de este despliegue son:

1. **Eficacia de la Regularización Paramétrica:** Se ha comprobado que, frente a arquitecturas de ensamblaje no paramétricas (como *Random Forest*), los modelos de optimización convexa son netamente superiores en la inferencia de distribuciones fraccionales correlacionadas. La implementación de la **Regresión *Ridge* (Multi-output)**, gracias a su penalización , ha logrado mitigar la alta multicolinealidad de las proporciones del escandallo y evitar el sobreajuste al ruido mecánico.
2. **Impacto Cuantitativo:** El logro más representativo del proyecto es la compresión drástica del margen de error de la fábrica. Al sustituir la heurística estática (que presentaba un Error Absoluto Medio del 22.85%) por el gemelo digital estocástico de las máquinas, el error global ha descendido hasta el 7.83%. Esto se traduce en una **mejora predictiva neta del 65.72%**.
3. **Generación de Valor Operativo:** La planta de procesado ahora dispone de una herramienta puramente *Data-Driven*. El sistema permite recalibrar matemáticamente la expectativa de salida volumétrica de cualquier lote antes de que ingrese a las cintas de triaje, dotando al departamento de logística de un

margen de maniobra inédito para gestionar las roturas de *stock* y asignar la maquinaria de manera inteligente según los compromisos comerciales vigentes.

7.2 CUMPLIMIENTO DE OBJETIVOS

Tras la fase de validación cruzada y el análisis del rendimiento computacional, se puede afirmar que el proyecto ha satisfecho íntegramente los objetivos planteados en el Capítulo 4 de esta memoria:

- Se ha diseñado e implementado una arquitectura de ingesta de datos (ETL) robusta que normaliza distribuciones morfológicas masivas.
- Se ha modelado con éxito el desgaste y sesgo particular de las calibradoras C1, C5 y C6 empleando ingeniería de características (*One-Hot Encoding*).
- Se ha rebasado holgadamente el objetivo de reducir el MAE predictivo por encima del 50%, garantizando un retorno de inversión (ROI) altamente positivo para la entidad colaboradora.

7.3 TRABAJOS FUTUROS

El éxito en la parametrización de este gemelo digital asienta una base tecnológica altamente escalable. No obstante, al tratarse de una primera iteración (un Producto Mínimo Viable analítico), existen múltiples vías de evolución y enriquecimiento algorítmico. Se proponen las siguientes líneas de investigación para dar continuidad al proyecto:

1. **Despliegue en Entornos Productivos (MLOps):** El trabajo futuro más inmediato es la transición del código alojado en *Jupyter Notebooks* hacia una arquitectura en la nube (ej. AWS o Google Cloud). Esto implicará el desarrollo de una API RESTful que permita a los Controladores Lógicos Programables (PLCs) de la fábrica enviar los datos del escandallo en tiempo real e incrustar la predicción volumétrica directamente en los paneles de monitorización de los operarios.
2. **Enriquecimiento del Espacio de Características (*Feature Enrichment*):** El modelo actual toma decisiones basándose en distribuciones volumétricas abstractas. En el futuro, se propone integrar de forma relacional variables medioambientales y biológicas, tales como la variedad del cítrico, la ubicación geoespacial de la parcela de cultivo, el índice de pluviosidad durante la campaña y los niveles de radiación solar. Esta volumetría de datos permitiría descubrir correlaciones ocultas entre el estrés hídrico de la fruta y su rendimiento mecánico en la calibradora.
3. **Clustering Estocástico de Lotes (Aprendizaje No Supervisado):** Se plantea la aplicación de algoritmos topológicos como K-Means o DBSCAN para segmentar históricamente los lotes en familias o clústeres. Al agrupar comportamientos isomórficos, se podría construir un sistema de recomendación previo al procesamiento que dicte automáticamente a qué máquina (C1, C5 o C6) debería enviarse un lote determinado para maximizar su rentabilidad comercial.

4. **Hibridación con Visión Artificial a Nivel Microscópico:** Integrar las predicciones agregadas macroscópicas (el modelo de Regresión *Ridge* actual) con los flujos de *Deep Learning* microscópico (Redes Neuronales Convolucionales) que analizan la fruta de manera individual. Esta hibridación generaría un ecosistema ciberfísico total donde el modelo predictivo del lote serviría como prior Bayesiano para reajustar dinámicamente los umbrales de detección de la maquinaria óptica en tiempo real.

Capítulo 8. BIBLIOGRAFÍA

[1] **SciELO.** "Optimizing fruit sampling for reliable quality assessment in sweet orange varieties: Sample size matters", *Scientia Agricola*. Junio, 2026. https://www.scielo.br/j/sa/a/cWbVNF7MGpBdxSG7pCTmfrj/

[2] **OAPEN.** "Sampling and statistics in assessment of fresh produce", *OAPEN Library Open Access*. Noviembre, 2021. https://library.oapen.org/bitstream/handle/20.500.12657/61519/9781801462471_web.pdf

[3] **Mishra, P.; Passos, D.** "Multi-output 1-dimensional convolutional neural networks for simultaneous prediction of different traits of fruit based on near-infrared spectroscopy", *Postharvest Biology and Technology*. Enero, 2022. https://doi.org/10.1016/j.postharvbio.2021.111741

[4] **Ghanghas, S.** "Prediction of fruit quality parameters using peel color in Citrus Reticulata L. fruit by multiple linear regression and artificial neural network approach", *ResearchGate*. Diciembre, 2022. https://www.researchgate.net/publication/366692766_Prediction_of_fruit_quality_parameters_using_peel_color_in_Citrus_Reticulata_L_fruit_by_multiple_linear_regression_and_artificial_neural_network_approach.