

**Universidad Don Bosco**  
**Maestría en Ciencia de Datos**  
**Aprendizaje No Supervisado**  
**Ciclo 02 - 2025**



**Reporte Ejecutivo:** Análisis Avanzado de ML No Supervisado para Control de Calidad

**Periodo Analizado:** 2025

**Duración de análisis:** 11 meses (06/01/2025 al 15/11/2025)

**Autores**

AT172560 – Leonardo Antonio Aguilera Torres

BM252943 – Brenda Carolina Barillas Marroquín

GM131896 – Josué Elías Granados Martínez

RH252948 – Heedy Eufemia Ramírez Hernández

**Docente Titular** Joel Orellana

San Salvador, noviembre de 2025

## Introducción

La globalización de servicios y manufactura han creado un nuevo nivel de competitividad en los servicios de manufactura de forma que todos los procesos involucrados deberán ser empujados a tener su máxima eficiencia y enfocados a la experiencia del cliente donde se considera como base principal la investigación y la innovación en el mercado de plásticos de alta densidad. Ganar eficiencia de forma que la reputación y expansión del negocio depende de cómo se sientan todos los clientes al recibir y utilizar los productos de cada empresa siendo aquella la más tecnológicamente eficaz, intuitiva e innovadora al manejar y monitorear cada una de las fases del proceso de manufactura permitiéndole así detectar más rápidamente todo aquello que genere desperdicio o retrabajo y solo aquellas con suficiente visión podrán salir adelante.

Reconociendo que para la empresa participante de este estudio de análisis de datos de control de calidad es una gran oportunidad de ganar eficacia es sus procesos de manufactura al aplicar algoritmos de aprendizaje automático no supervisado. El conocer su misión, visión y valores es una parte muy importante de la cultura organizacional.

- Visión

SOMOS LA AUTORIDAD LÍDER Y FABRICANTE DE PLÁSTICOS DE INGENIERÍA EN AMÉRICA DEL NORTE.

- Misión

CREAMOS PLÁSTICOS DE INGENIERÍA INNOVADORES SOLUCIONES PARA CONSTRUIR UN MUNDO MÁS SEGURO Y MÁS FUTURO SOSTENIBLE.

- Valores

APOYAR A LOS CLIENTES, PROTEGER LA SALUD HUMANA, CONSERVAR RECURSOS VITALES Y MODERNIZAR INFRAESTRUCTURA PARA CUMPLIR NUESTRA MISIÓN.

## Sumario

El departamento de tecnologías de la información de XXXX América es responsable de gestionar todos los servicios necesarios que generen valor y ayuden a la digitalización y automatización del procesamiento de datos en sus procesos, siendo uno de los más importantes el de gestionar los datos generados tanto en las oficinas, laboratorios de calidad como en las áreas de producción utilizando los diferentes sistemas de la empresa tales como su ERP y sistemas de calidad. Los servicios de TI incluyen visitas a todas las plantas de manufactura donde se actualizan y se da soporte en todas las áreas donde se necesite estandarizar y aplicar las políticas que ayuden al alcanzar los objetivos de la empresa.

## Objetivo del análisis

El objetivo de este análisis es evaluar los resultados de las pruebas de laboratorio de forma que se puedan encontrar agrupaciones o algún tipo de correlaciones que permitan definir y monitorear cuando los datos están dentro de lo esperado o no. Para ello, se busca:

- Reducir los grupos dimensionales de forma que permita mantener los datos comprimidos sin perder más información para posterior análisis.
- Identificar cuando existe alguna desviación durante la producción.
- Analizar las tendencias históricas de los datos basados en ítems corredores.
- Detectar inconsistencias en los datos de las pruebas de forma que esto permita definir donde se requiere el ajuste ya sea en la máquina de producción o en las áreas de prueba

Este análisis permitirá tomar decisiones informadas para mejorar la gestión de los procesos de manufactura y pruebas de calidad que permitan conjuntamente monitorear los resultados y al detectar la mínima desviación mandar algún tipo de alarma a los operadores o personal del laboratorio para que se tomen acciones.

## Metodología

Para llevar a cabo este análisis, partimos de un dataset proporcionado por la empresa por medio de una extracción con varios SQL scripts.

Este análisis empleó técnicas de aprendizaje automático no supervisado para identificar patrones, defectos y optimizar procesos de control de calidad para lotes de productos en 2025. La metodología consistió en cuatro etapas principales:

1. Análisis Exploratorio de Datos (EDA) y Preprocesamiento: Exploración inicial de datos, limpieza y estandarización.
2. Reducción de Dimensionalidad: Análisis comparativo de PCA, t-SNE y UMAP para reducir características de control de calidad de alta dimensionalidad a representaciones 2D.
3. Análisis de Agrupamiento: Aplicación de K-means, DBSCAN y Agrupamiento Espectral para identificar patrones de calidad y anomalías.
4. Interpretación y Conclusiones: Integración de los mejores métodos para proporcionar insights accionables para el control de calidad en fabricación.

El conjunto de datos contenía resultados de pruebas de control de calidad en múltiples características, incluyendo mediciones de espesor, aspereza, contenido de carbono, densidad y varias métricas de resistencia. Todos los análisis se realizaron en características numéricas estandarizadas después de la eliminación de valores nulos.

## **EDA y Preprocesamiento**

### **Resumen del Conjunto de Datos:**

22 características numéricas más identificadores (ItemID, LotNumber, manufacturedDate).

### **Calidad de Datos:**

No se detectaron valores faltantes; valores atípicos eliminados usando el método IQR ( $Q1 - 1.5 \cdot IQR$  a  $Q3 + 1.5 \cdot IQR$ ).

### **Preprocesamiento:**

Características estandarizadas usando StandardScaler para normalización.

### **Insights Clave:**

Se observaron correlaciones fuertes entre mediciones relacionadas (ej. métricas de espesor, propiedades de resistencia); las distribuciones mostraron sesgo variable en las características.

## **Reducción de Dimensionalidad**

**PCA:** Explicó ~70% de varianza en los primeros 5 componentes; computación rápida pero limitada a estructuras lineales.

**t-SNE:** Excelente preservación de estructura local (confiabilidad/continuidad ~0.8-0.9); más lento e inestable en ejecuciones.

**UMAP:** Rendimiento equilibrado con preservación de vecindad ~0.85-0.95; estable y computacionalmente eficiente.

**Selección:** UMAP elegido por la mejor preservación de variaciones no lineales de lotes mientras mantiene eficiencia computacional.

### **Análisis de Agrupamiento (Clustering)**

**K-means:** k óptimo=3 basado en método del codo y análisis de silueta; bueno para clusters esféricos pero asume convexidad.

**DBSCAN:** Ajustado con  $\text{eps}=0.5$ ,  $\text{min\_samples}=10$ ; identificó 3-5 clusters más puntos de ruido como anomalías.

**Agrupamiento Espectral:** Funcionó bien en estructuras no convexas pero intensivo computacionalmente.

**Selección:** DBSCAN seleccionado por sus superiores capacidades de detección de anomalías cruciales para control de calidad.

## **Hallazgos**

Durante el análisis, Se encontraron varios puntos importantes que ayudaron a entender mejor el desempeño y la rentabilidad del servicio que ofrece la consultora.

Hallazgos más relevantes:

### **Mejores Métodos:**

UMAP para reducción de dimensionalidad combinado con DBSCAN para agrupamiento proporcionó resultados óptimos para análisis de control de calidad.

### **Patrones de Calidad:**

Los clusters corresponden a diferentes niveles de calidad de lotes de productos, con lotes de alta calidad mostrando resultados consistentes en todas las características.

### **Detección de Anomalías:**

Los puntos de ruido en DBSCAN representan potenciales defectos o valores atípicos en procesos de fabricación, permitiendo control de calidad proactivo.

### **Insights de Fabricación:**

Los lotes anómalos a menudo exhiben desviaciones en parámetros críticos como resistencia a ruptura y resistencia al desgarro.

## **Conclusiones:**

- UMAP redujo efectivamente la dimensionalidad preservando estructuras no lineales.
- DBSCAN identificó exitosamente clústeres de calidad y anomalías.
- Las anomalías corresponden a defectos potenciales en lotes de fabricación.
- Este análisis proporciona una base para control de calidad predictivo.

## **Limitaciones:**

- Los parámetros de DBSCAN son sensibles y pueden necesitar ajuste para nuevos datos.
- Asume que las anomalías son defectos, lo cual no siempre es cierto.
- El rendimiento del modelo depende de la calidad de los datos y el preprocesamiento.

## **Reproducibilidad:**

- Usar `random_state=42` para resultados consistentes.
- Documentar elecciones de parámetros y versiones de datos.
- Validar en conjuntos de retención para generalización.

## **Recomendaciones**

1. Monitorear de cerca los lotes identificados como anomalías.
2. Configurar alertas automatizadas para clústeres con alta variabilidad.
3. Investigar causas raíz en producción para lotes defectuosos.
4. Usar este modelo de agrupamiento para control de calidad continuo.
5. Reentrenar el modelo periódicamente con nuevos datos.