Predicción de resultados de permeabilidad en la toma de muestras de un laboratorio de análisis, 2022-2023

Laura Isabel Barrera Echeverri
Jorge Ignacio Morales

Especialistas en Analítica y Ciencia de Datos

Asesora
Daniela Serna Buitrago, Especialista en Analítica

Universidad de Antioquia
Facultad de ingeniería
Posgrado
Medellín
2023

Cita	Barrera Echeverri y Morales [1]
Referencia Estilo IEEE (2020)	[1] L. I. Barrera Echeverri y J. I. Morales, “Predicción de resultados de permeabilidad en la toma de muestras de un laboratorio de análisis, 2022 - 2023”, Especialistas en Analítica y Ciencia de Datos, Posgrado UdeA, Universidad de Antioquia, Medellín, 2023.

Especialización en analítica y ciencia de datos, Cohorte IV.

Dedicatoria

A nuestras familias por la paciencia y acompañamiento en este proceso de formación.

Agradecimientos

A cada uno de los profesores que nos acompañaron y formaron en esta carrera.

TABLA DE CONTENIDO

LISTA DE TABLAS

LISTA DE FIGURAS

SIGLAS, ACRÓNIMOS Y ABREVIATURAS

Esp.	Especialista
UdeA	Universidad de Antioquia
°C Lab.	Temperatura en grados centígrados del laboratorio
H.R LAB(%).	Porcentaje de humedad del empaque

RESUMEN

Este estudio se centra en la predicción de la permeabilidad de los materiales utilizados en la fabricación de empaques, con el fin de mejorar el proceso actual que se lleva a cabo de forma manual y depende de la evaluación subjetiva de los instrumentistas para determinar si un empaque cumple con los estándares de calidad adecuados. En la actualidad, este proceso se realiza mediante la recopilación manual de información en una hoja de cálculo de Excel que no está conectada a ningún dispositivo de medición de calidad del laboratorio. El objetivo de este estudio es minimizar los errores en la evaluación y garantizar la eficacia y durabilidad del producto final. Debido a la falta de conocimiento preciso sobre la permeabilidad de cada uno de los tipos de materiales disponibles en el mercado, existe un riesgo significativo de aprobar un producto defectuoso.

Para abordar esta problemática, se realizó un análisis inicial y exploratorio de los datos, seguido de una limpieza que esté afectando el análisis de estos. Posteriormente, se aplicaron diferentes modelos de predicción de datos, como la regresión lineal, máquinas de soporte vectorial y el Random Forest, para definir cuál de ellos se ajusta mejor a los datos entrenados vs predichos, con diferentes iteraciones y la medición del error cuadrático medio.

Los resultados obtenidos muestran que la aplicación del Random Forest, y las máquinas de soporte vectorial permiten predecir la permeabilidad de los materiales con alta precisión, mientras que la regresión lineal resultó menos eficiente. A partir de esto, se concluye que los modelos implementados pueden ser utilizados para predecir la permeabilidad de los materiales con gran precisión.

Este estudio proporciona una herramienta valiosa para predecir la permeabilidad de los materiales utilizados en la fabricación de empaques, lo que permitirá mejorar la calidad del producto y reducir el riesgo de aprobar productos defectuosos.

Palabras clave — **Permeabilidad, predicción, datos, análisis.**

ABSTRACT

This study focuses on predicting the permeability of materials used in packaging manufacturing to improve the current manual process that relies on subjective evaluation by instrument operators to determine if a package meets the appropriate quality standards. Currently, the process involves manually collecting information in an Excel spreadsheet that is not connected to any laboratory quality measuring device. The objective of this study is to minimize evaluation errors and ensure the effectiveness and durability of the final product. Due to the lack of precise knowledge about the permeability of each type of material available in the market, there is a significant risk of approving a defective product.

To address this issue, an initial exploratory data analysis was conducted, followed by data cleaning to eliminate any factors affecting the analysis. Various data prediction models, such as linear regression, support vector machines, and Random Forest, were then applied to determine which one best fit the trained vs. predicted data, with different iterations and measurement of mean squared error.

The results show that the implementation of Random Forest and support vector machines allows for highly accurate prediction of material permeability, while linear regression was less efficient. From this, it is concluded that the implemented models can be used to predict material permeability with high accuracy.

This study provides a valuable tool for predicting the permeability of materials used in packaging manufacturing, which will improve product quality and reduce the risk of approving defective products.

Keywords — **Permeability, prediction, data, analysis.**

I. INTRODUCCIÓN

Se pretende predecir la permeabilidad de los empaques tomando en consideración las variables de estructura, calibre, porcentaje de humedad y temperatura, con el fin de garantizar la eficacia y durabilidad del producto sin necesidad de intervención humana. El problema radica en la falta de una herramienta de apoyo que permita predecir valores de manera precisa para ser utilizados como guía en la medición de calidad de cualquier tipo de empaque, lo cual puede llevar a cometer muchos errores.

El objetivo de este estudio es proporcionar una herramienta que permita predecir la permeabilidad de los materiales en función de sus características, como el calibre, estructura, temperatura, humedad, entre otros. Para ello, se realizará un análisis inicial y exploratorio de los datos, seguido de una limpieza de la contaminación que esté afectando el análisis de los mismos. Posteriormente, se aplicarán diferentes modelos que permitan predecir los datos, como la regresión lineal, las máquinas de soporte vectorial y el Random Forest. Se definirá cuál de estos modelos se ajusta mejor a los datos entrenados vs predichos, con diferentes iteraciones y la medición del error cuadrático medio.

La justificación de este estudio radica en la importancia de garantizar la calidad del producto envasado mediante la predicción de la permeabilidad de los materiales utilizados en la fabricación de empaques. Una herramienta que permita predecir la permeabilidad de los materiales en función de sus características permitirá reducir el riesgo de aprobar productos defectuosos y mejorar la satisfacción del consumidor final.

II. PLANTEAMIENTO DEL PROBLEMA

La permeabilidad de un material es fundamental para la calidad de un empaque y, por ende, del producto envasado. Sin embargo, actualmente no se dispone de información precisa y detallada sobre la permeabilidad de los diversos tipos de materiales disponibles en el mercado. Esta falta de conocimiento dificulta el trabajo de los instrumentistas encargados de evaluar la calidad de los envases y aumenta el riesgo de aprobar productos defectuosos. En consecuencia, resulta necesario contar con valores de referencia que permitan a los instrumentistas obtener resultados precisos y fiables en sus mediciones, lo que mejoraría la calidad del producto y reduciría el riesgo de posibles consecuencias negativas para la empresa y el consumidor final.

A. Antecedentes

En relación al tema de la predicción de la permeabilidad de los materiales utilizados en la fabricación de envases y empaques, no se encontraron antecedentes de investigaciones previas que aborden específicamente esta problemática. Sin embargo, se ha recolectado información sobre las características necesarias para definir el resultado de permeabilidad, pero esta información se ha almacenado sin ser utilizada para abordar la problemática en cuestión.

III. JUSTIFICACIÓN

Dado que la permeabilidad es una característica crítica que debe ser controlada para garantizar la eficacia y durabilidad del producto envasado, resulta fundamental conocer en profundidad las propiedades de los diferentes tipos de materiales disponibles en el mercado.

En este sentido, se seleccionó la predicción de la permeabilidad de los materiales como tema de investigación, debido a la importancia que tiene en la industria de los empaques, y a la falta de estudios previos que aborden específicamente esta problemática.

El aporte que tendrá este texto a la ciencia radica en la implementación de diferentes aplicaciones en la analítica de datos, como modelos de aprendizaje supervisado para predecir la permeabilidad de los materiales con alta precisión.

- **Características:** estructura, calibre 1, calibre 2, calibre 3, calibre 4, calibre 5, calibre prom, muestra, °C Lab, H.R LAB(%), permeabilidad.

IV. OBJETIVOS

A. Objetivo general

Predecir la permeabilidad de un empaque del mercado, a través de técnicas de aprendizaje automático, para proporcionar valores de referencia a los instrumentistas al momento de tomar muestras aleatorias de análisis, y decidir sobre la calidad de los productos de forma más precisa e informada sobre aprobar un producto o no.

VI. MARCO TEÓRICO

El análisis exploratorio de datos (EDA) ha sido fundamental en el desarrollo de modelos de predicción de permeabilidad. La aplicación de diferentes técnicas de análisis estadístico ha permitido establecer relaciones matemáticas entre las variables que influyen en la permeabilidad y analizar su comportamiento ante diferentes condiciones. A continuación, se presentan los fundamentos teóricos que describen los conceptos clave relacionados con la problemática a investigar:

EDA - Análisis Exploratorio de Datos

El análisis exploratorio de datos (EDA) se utiliza para examinar y resumir los datos clave en una investigación. Proporciona una comprensión básica de los datos, incluyendo su distribución, valores faltantes y otras características relevantes [1]. Se pueden utilizar herramientas de EDA, ya sea a través de gráficos o de funciones específicas. Algunas de las funciones comunes incluyen forma, resumen, descripción, identificación de datos nulos, información sobre variables, tipos de datos, entre otras. Por otro lado, también se pueden utilizar visualizaciones como gráficos de dispersión, caja, barra, densidad y correlación como parte del enfoque gráfico del EDA [2].

Al considerar el análisis exploratorio de datos como parte integral del proceso, facilitará la identificación de variables clave y la selección de técnicas adecuadas para la predicción de la

permeabilidad. De esta manera, se logrará un enfoque más efectivo que permite evitar errores en la fabricación de empaques y garantizar la calidad del producto empacado, a través de una comprensión más profunda y precisa de los factores determinantes en la permeabilidad de los materiales.

Descripción de variables

- **Año:** variable de tiempo
- **Mes:** variable de tiempo
- **Día:** variable de tiempo
- **Consecutivo:** identificación única del registro
- **Referencia**: descripción del registro a medir
- **Estructura:** tipo de material del empaque
- **Calibre1:** primera medición de grosor de material
- **Calibre2:** segunda medición de grosor de material
- **Calibre3**: tercera medición de grosor de material
- **Calibre4:** cuarta medición de grosor de material
- **Calibre5:** quinta medición de grosor de material
- **Calibre prom:** promedio de medición de grosor de las cinco medidas
- **Muestra:** cantidad de empaques a medir
- **°C Lab:** temperatura ambiente
- **H.R LAB(%):** porcentaje de humedad del empaque
- **Resultado permeabilidad:** medición de capacidad de un empaque para permitir que un fluido lo atraviese sin alterar su estructura interna
- **Instrumentista:** sujeto que realiza la medición

La base de datos cruda contiene 1.119 registros, los cuales presentan las siguientes características:

```
Data columns (total 20 columns):
 #   Column                  Non-Null Count  Dtype
---  ------                  --------------  -----
 0   AÑO                     1119 non-null   int64
 1   MES                     1119 non-null   int64
 2   DÍA                     1119 non-null   int64
 3   CONSECUTIVO             1082 non-null   object
 4   REFERENCIA              1119 non-null   object
 5   ESTRUCTURA              1105 non-null   object
 6   CALIBRE 1               1102 non-null   object
 7   CALIBRE 2               1102 non-null   object
 8   CALIBRE 3               1102 non-null   object
 9   CALIBRE 4               1102 non-null   object
 10  CALIBRE 5               1102 non-null   object
 11  CALIBRE PROM            1118 non-null   object
 12  MUESTRA                 1118 non-null   float64
 13  °C LAB                  1118 non-null   float64
 14  H.R LAB(%)              1118 non-null   float64
 15  Resultado permeabilidad 1119 non-null   object
 16  INSTRUMENTISTA          1119 non-null   object
 17  Unnamed: 17             23 non-null     object
 18  Unnamed: 18             9 non-null      object
 19  Unnamed: 19             1 non-null      object
```

Figura 1. Descripción de las variables, base de datos cruda

Se presentan a continuación las estadísticas descriptivas de la base de datos cruda.

	count	mean	std	min	25%	50%	75%	max
AÑO	1119.0	2020.900804	0.819043	2020.0	2020.0	2021.0	2022.0	2022.0
MES	1119.0	6.508490	3.372210	1.0	3.0	7.0	9.0	12.0
DÍA	1119.0	14.991063	8.419667	1.0	8.0	14.0	22.0	31.0
MUESTRA	1118.0	1.500000	0.565733	1.0	1.0	1.0	2.0	4.0
°C LAB	1118.0	22.370358	6.168055	20.1	21.3	22.2	23.3	226.0
H.R LAB(%)	1118.0	51.986064	24.605519	38.0	47.3	49.5	54.0	563.0

Figura 2. Análisis descriptivo inicial de los datos obtenidos

Se han identificado varias columnas numéricas que están siendo clasificadas como objetos debido a la presencia de datos atípicos. Por consiguiente, se realizará la preparación de la fuente de datos con el objetivo de poder llevar a cabo un análisis adecuado.

Evaluación de registros en variables

Para determinar la calidad de los registros y la confiabilidad de las mediciones realizadas, se tuvo en cuenta lo siguiente:

- Se eliminaron columnas que no aportan valor, como el consecutivo y columnas en blanco.
- Se renombraron las columnas para hacerlas más legibles.
- Se evaluaron las columnas para identificar valores que no corresponden.
- Se identificaron datos inconsistentes en las mediciones de las variables Calibre 1 a Calibre 5. Y se eliminaron los caracteres especiales que no eran numéricos para tener valores únicos.
- Se reemplazaron los datos inconsistentes en las variables Calibre.

A pesar de la eliminación de variables que no aportan valor y la identificación de inconsistencias en la exploración inicial de los datos, no fue posible realizar la conversión de las variables numéricas a formato float.

- Se realizó una limpieza y preparación exhaustiva de las variables.
- Se identificaron datos inconsistentes en la variable Permeabilidad.
- Todas las variables numéricas se convirtieron a formato float y se procedió al análisis exploratorio de los datos.

	count	mean	std	min	25%	50%	75%	max
Calibre1	1065.0	80.056338	41.413473	9.0000	46.00000	77.0000	107.0000	201.0000
Calibre2	1065.0	80.079812	41.334963	8.0000	46.00000	76.0000	107.0000	201.0000
Calibre3	1065.0	80.698592	46.056318	5.0000	45.00000	77.0000	107.0000	736.0000
Calibre4	1063.0	80.757291	43.102813	8.0000	46.00000	77.0000	107.0000	321.0000
Calibre5	1065.0	80.014085	41.372492	7.0000	46.00000	77.0000	107.0000	200.0000
Calibre_Prom	1065.0	80.165556	41.542798	8.4000	45.60000	76.6000	106.2000	206.8000
Muestra	1118.0	1.500000	0.565733	1.0000	1.00000	1.0000	2.0000	4.0000
Grados_C	1118.0	22.370358	6.168055	20.1000	21.30000	22.2000	23.3000	226.0000
Porc_Humedad	1118.0	51.986064	24.605519	38.0000	47.30000	49.5000	54.0000	563.0000
Permeabilidad	1098.0	1.440807	2.472366	-0.0068	0.11685	0.4647	1.7104	14.5698

Figura 3. Análisis descriptivo de variables evaluados

Análisis de distribuciones en variables continuas

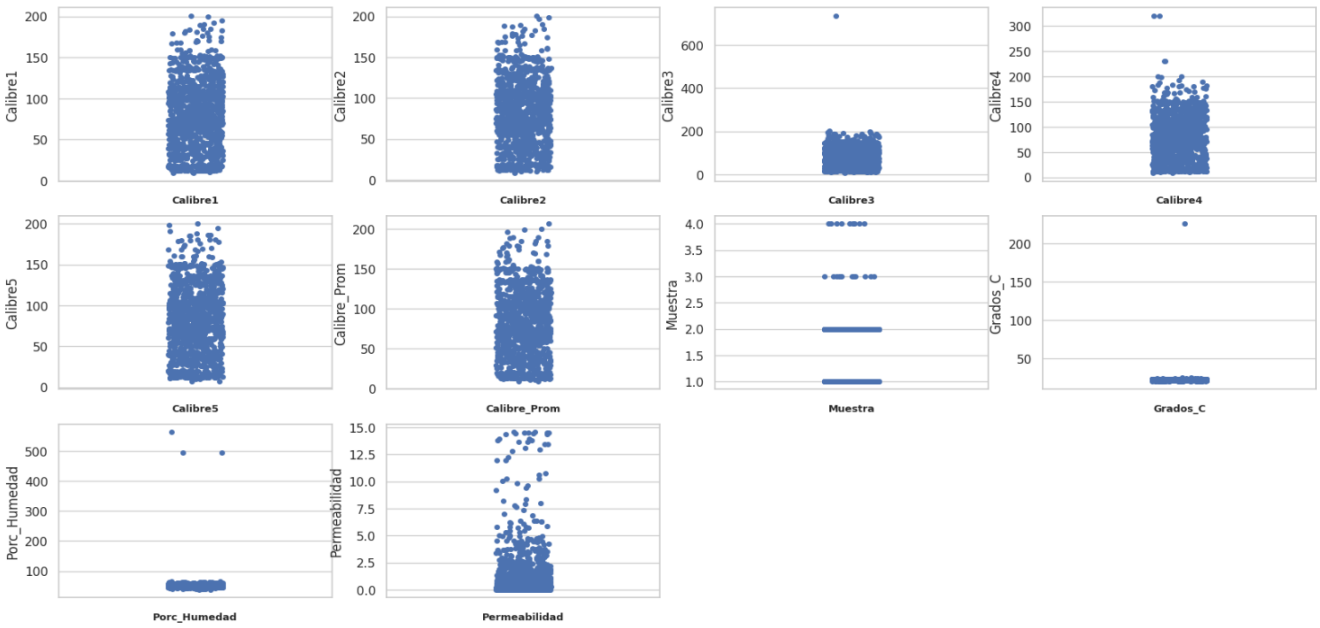

Figura 4. Diagrama strip plot de variables continuas

Según el análisis descriptivo de los datos evaluados y la visualización de los datos, se observa una alta desviación estándar en las variables de calibre1, calibre2, calibre3, calibre4, calibre5 y calibre_prom, ya que sus valores se encuentran en un rango aproximado de 5 a 320.

Todas estas variables muestran un patrón similar, pero se destaca que calibre 1 y calibre 5 presentan una menor dispersión, lo que sugiere una menor presencia de valores atípicos.

En contraste, la variable calibre 3 muestra la mayor dispersión, posiblemente debido a que su valor máximo es 736, el cual podría considerarse un valor atípico.

En cuanto a las demás variables, se observa que la variable "muestra" tiene la menor dispersión entre todas. Por otro lado, la variable "Grados_C" presenta un valor máximo de 226, el cual podría ser considerado atípico, ya que en un contexto de temperatura de laboratorio, resultaría poco realista y difícil de trabajar para los seres humanos.

En relación al porcentaje de humedad, se registra un valor máximo de 564, que también podría ser un valor atípico. Por último, la variable "permeabilidad" muestra valores mínimos negativos, lo cual no es posible en la medición de esta variable dentro del contexto de permeabilidad, lo que podría indicar la presencia de otro valor atípico.

Análisis de distribuciones en variables categóricas

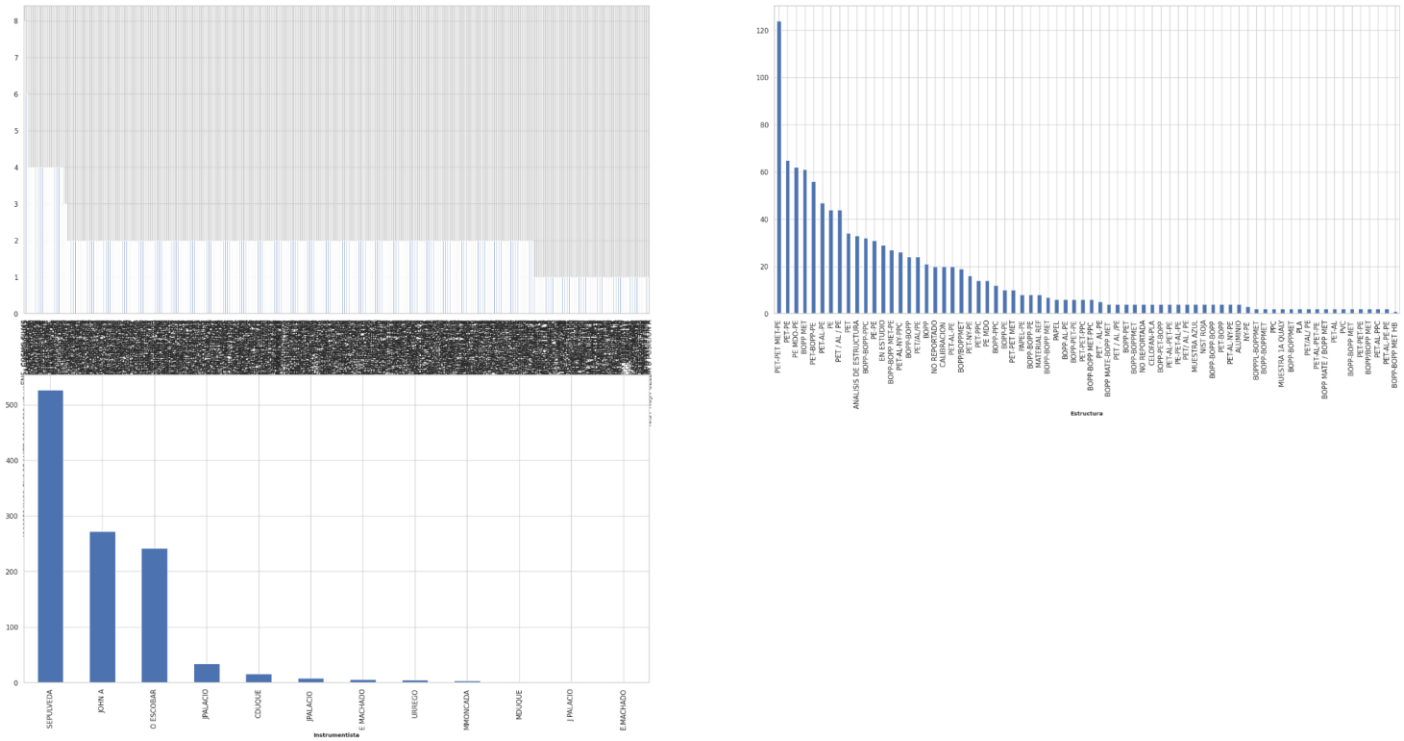

Figura 5. Histogramas de frecuencias con variables categóricas

En el análisis de las variables categóricas, no se observa un patrón representativo en la variable "referencia". Esta variable contiene numerosas categorías y visualmente no se percibe una participación destacada en ninguna de ellas.

En cuanto al comportamiento de la variable "estructura", también se registran numerosas categorías. Sin embargo, se destaca que la estructura "PET-PET MET-PE" presenta una frecuencia mayor que las demás en los datos analizados.

Con respecto a la variable "instrumentista", se observa que los sujetos que más frecuentemente realizan las mediciones son SEPULVEDA, JHON y O ESCOBAR. Sin embargo, esta información no aporta valor al objetivo del proyecto en cuestión.

Correlación entre variables evaluadas

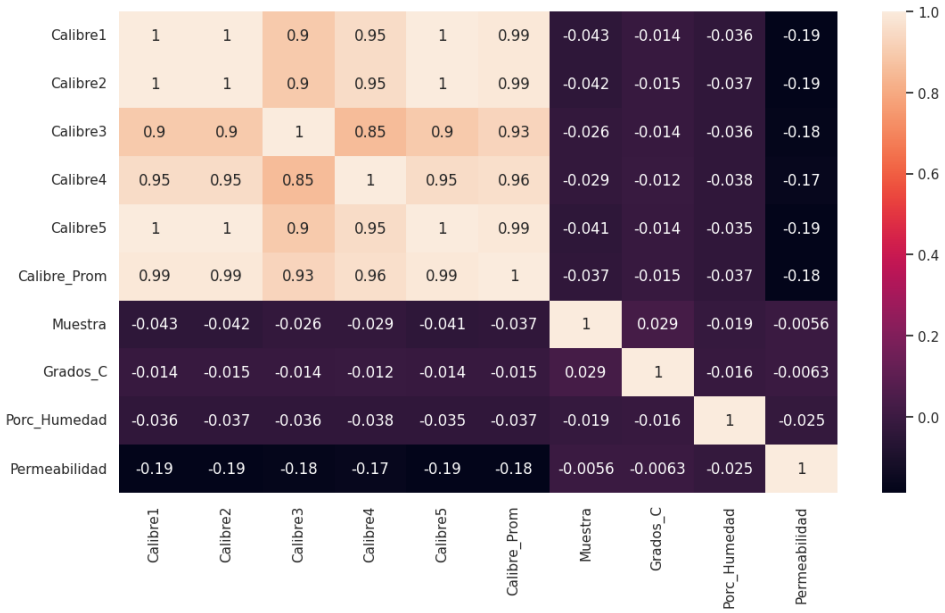

Figura 6. Matriz de correlación entre variables numéricas

Codificación de variables categóricas a números

Se utilizó el método One Hot Encoding, debido a que son numerosas categorías en la variables estructura y dicha variable es de alta importancia y dependencia en el análisis de la permeabilidad de un empaque.

Valores faltantes

El indicador de pérdida (Missing Indicator) es un método para buscar identificar presencia de valores faltantes en la base que se procesa:

Porcentaje de Datos Faltantes (%): 1.8642533936651582

Imputación (Algoritmo KNN)

Este método de rellenar los valores faltantes se utilizó con 5 vecinos y promedio ponderado, donde los vecinos más cercanos de un valor tendrán una mayor influencia que los vecinos que están más lejos, aplicando una distancia euclidiana.

Valores atípicos

Se utiliza el algoritmo LOF para calcular el porcentaje de datos atípicos para la estructura de datos numérica. Luego de aplicar el método LOF, se identificaron 29 filas, las cuales fueron eliminadas del dataset.

Normalización (método Min-Max)

Se aplica este método de normalización para que los datos estén en un sólo plano vectorial y los modelos de aprendizaje automático puedan calcularse de forma correcta, evitando datos atípicos en su comportamiento.

Luego de la depuración el dataset resultante tiene **4** Variables Cuantitativas y **49** variables Dummies resultantes de la aplicación del método One Hot Encoding para la variable Estructura

A Continuación podemos visualizar el comportamiento de las variables cuantitativas y categóricas antes y después de la imputación y el la normalización de las mismas.

Variables cuantitativas antes de detección de atípicos y escalamiento

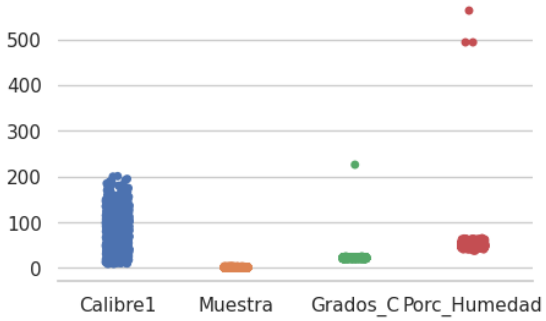

Figura 7. Escenario actual variables cuantitativas sin escalamiento

Variables cuantitativas después de detección de atípicos y escalamiento

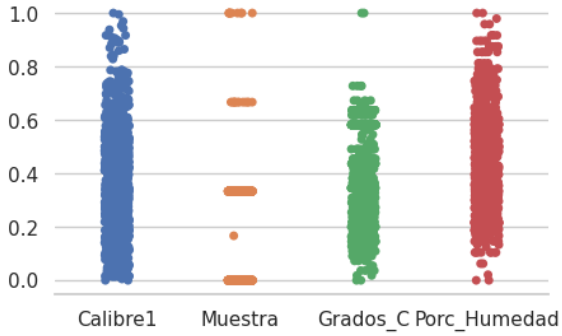

Figura 8. Escenario variables cuantitativas con escalamiento

Variables categóricas Antes y después Imputación y Escalamiento

En la tabla presentada, se pueden observar las columnas dummies que representan todos los valores únicos de la variable "Estructura", mientras que las cantidades correspondientes a "1" indican las frecuencias para cada columna respectiva.

Columnas	Total Registros Antes	Total Registros Después
Estructura_PET-AL-PE	160	157
Estructura_PET-PET-MET-PE	124	121
Estructura_BOPP-BOPP-MET	72	70
Estructura_PET-PE	65	63
Estructura_PE-MDO-PE	62	59
Estructura_BOPP-MET	61	60
Estructura_PET-BOPP-PE	56	56
Estructura_PE	44	44
Estructura_PET	34	31
Estructura_ANALISIS-DE-ESTRUCTURA	33	30
Estructura_BOPP-BOPP-PPC	32	32
Estructura_PE-PE	31	31
Estructura_EN-ESTUDIO	29	29
Estructura_PET-AL-NY-PPC	25	25
Estructura_BOPP-BOPP	24	24
Estructura_NO-REPORTADA	24	22
Estructura_BOPP	21	21
Estructura_CALIBRACION	20	20
Estructura_PET-NY-PE	16	16
Estructura_PE-MDO	14	13
Estructura_PET-PPC	14	14
Estructura_BOPP-PPC	12	11
Estructura_BOPP-PE	10	10
Estructura_PET-PET-MET	10	10
Estructura_BOPP-BOPP-PE	8	8
Estructura_MATERIAL-REF	8	6
Estructura_PAPEL-PE	8	8
Estructura_BOPP-MATE-BOPP-MET	6	6
Estructura_BOPP-PET-PE	6	6
Estructura_PAPEL	6	6
Estructura_PET-PET-PPC	6	6
Estructura_BOPP-AL-PE	4	4
Estructura_BOPP-BOPP-BOPP	4	4
Estructura_BOPP-PET	4	3
Estructura_BOPP-PET-BOPP	4	4
Estructura_CELOFAN-PLA	4	4
Estructura_MUESTRA-AZUL	4	4
Estructura_NIST-ROJA	4	3
Estructura_PET-AL-NY-PE	4	3
Estructura_PET-BOPP	4	4
Estructura_NY-PE	3	3
Estructura_BOPPPL-BOPPMET	2	2
Estructura_MUESTRA-1A-QUALY	2	2
Estructura_PET-AL	2	2
Estructura_PET-AL-PPC	2	2
Estructura_PET-PET-PE	2	2
Estructura_PLA	2	2
Estructura_PPC	2	2
Estructura_PVC	2	2

Figura 9. Escenarios imputación variables categóricas

Comportamiento de variable de salida: "Permeabilidad"

La distribución de los datos exhibe una similitud con la distribución de Poisson, lo cual es relevante en el contexto de la permeabilidad de empaques y envases. La visualización muestra que

los datos relacionados con la permeabilidad pueden seguir un patrón de comportamiento similar a eventos aleatorios y proporciona información valiosa para su análisis y predicción.

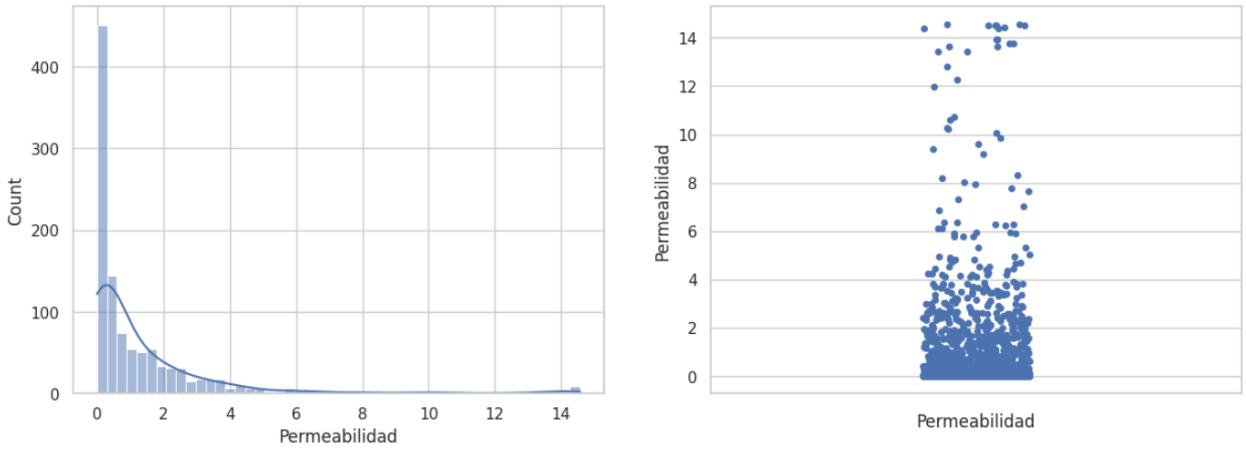

Figura 10. Comportamiento de variable de salida

VII. METODOLOGÍA

El estudio descrito emplea la metodología CRISP DM, la cual se basa en técnicas de análisis de datos numéricos para predecir la permeabilidad del material del empaque en función de sus características. Este enfoque se ve reforzado por la aplicación de diversos modelos analíticos que buscan encontrar el mejor ajuste y reducir el error cuadrático medio, asegurando así la precisión y fiabilidad de los resultados obtenidos.

Además, se destaca la realización de un EDA (análisis exploratorio) y la limpieza de datos, lo que sugiere un enfoque de investigación mixto en el estudio. Aunque se puede apreciar que el enfoque cuantitativo predomina en la metodología general del estudio, se considera importante destacar que se han considerado diversas perspectivas y enfoques para abordar la problemática en cuestión. Este enfoque integral y riguroso garantiza la validez de los hallazgos obtenidos y su aplicabilidad en el mundo real.

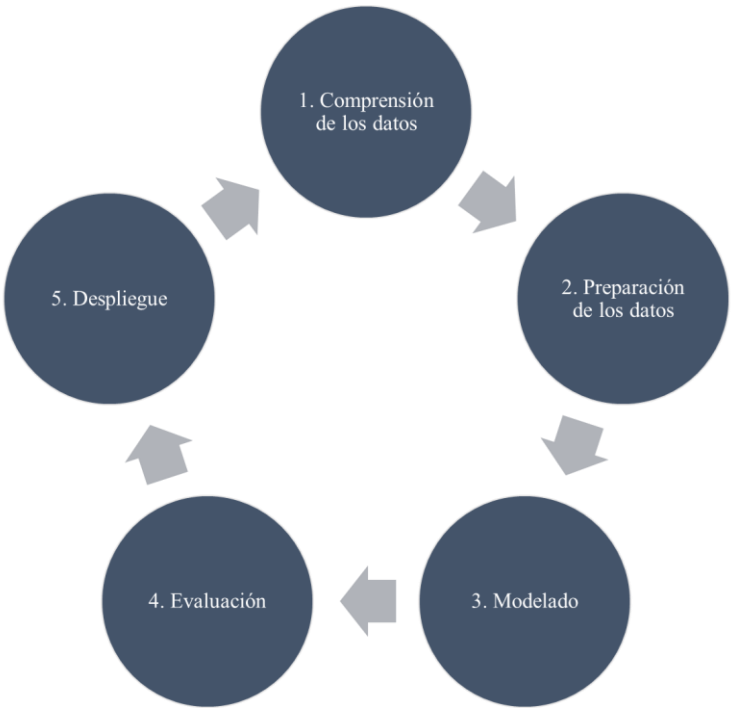

Figura 11. Etapas de metodología CRISP DM

La metodología CRISP-DM es un modelo de proceso estándar utilizado en el análisis de datos. Esta metodología se divide en seis etapas [3]:

- Comprensión del problema: en esta etapa, se investigó la literatura existente sobre estudios previos relacionados con la permeabilidad de empaques. Se revisaron investigaciones sobre los diferentes tipos de materiales utilizados en la fabricación de empaques y cómo afectan la permeabilidad. Además, se analizó la importancia de la permeabilidad en la calidad del producto final y los riesgos asociados con la aprobación de empaques defectuosos.

  Para obtener los datos, se estableció la recopilación de un histórico de datos exhaustivo que abarca una amplia variedad de empaques utilizados en diferentes contextos y líneas de productos. Esto permite tener una visión más completa y representativa de la permeabilidad en diversos escenarios y garantiza la fiabilidad de los análisis y predicciones.

  Además, se ha considerado trabajar de manera específica en todas las líneas de productos. Esto implica que se han tenido en cuenta las características y peculiaridades propias de cada línea, como los materiales utilizados, el proceso de fabricación y las condiciones de almacenamiento y transporte. Este enfoque estratégico brinda una comprensión más profunda de los factores que influyen en la permeabilidad de los empaques.

  Para entender cuando los valores de permeabilidad son considerados altos o bajos, se deben establecer criterios basados en las necesidades y estándares de la industria. Estos criterios pueden variar según el tipo de producto, las expectativas de calidad y las regulaciones aplicables. [4]

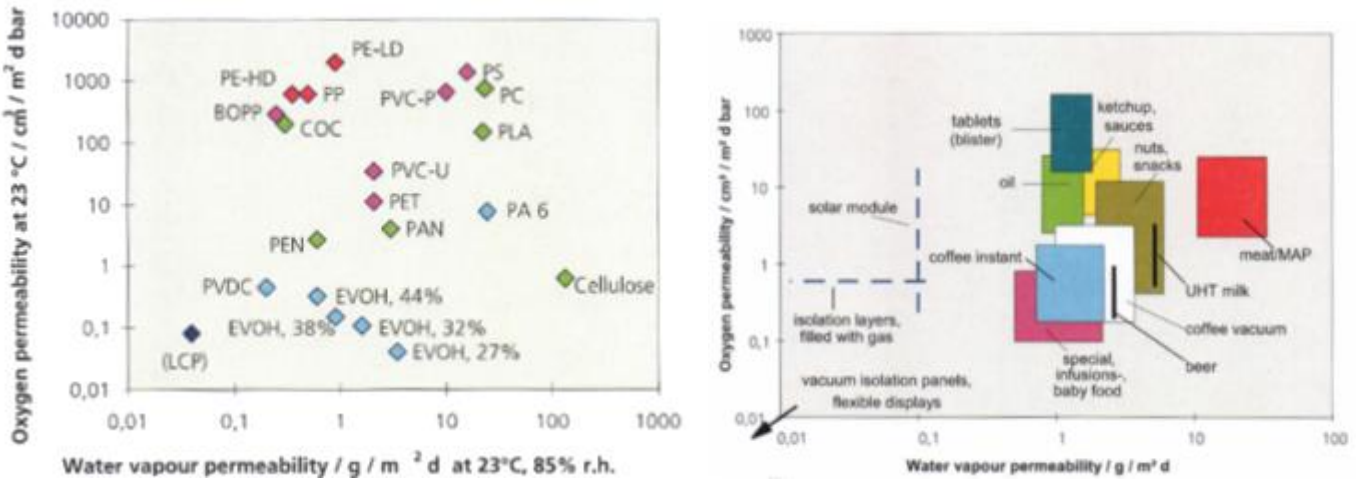

Figura 12. Criterios estándares de permeabilidad en la industria según tipo de material y producto

Al comparar los valores medidos de permeabilidad con los límites establecidos, es posible determinar si la permeabilidad se considera alta o baja. Si un valor de permeabilidad supera el límite máximo establecido, se considera alto y puede indicar que el empaque tiene una mayor probabilidad de permitir la entrada de agentes externos no deseados. Por otro lado, si un valor de permeabilidad se encuentra por debajo del límite mínimo establecido, se considera bajo y sugiere que el empaque presenta una menor capacidad de protección contra la permeabilidad.

La comprensión de los niveles de permeabilidad como altos o bajos depende de la comparación entre los valores medidos y los límites establecidos según los criterios definidos. Este enfoque proporciona una base objetiva para evaluar la calidad de los empaques en términos de su capacidad para resistir la permeabilidad y garantizar la preservación adecuada de los productos envasados.

- Comprensión de los datos: se recopilaron los datos disponibles sobre la permeabilidad de los diferentes tipos de empaques utilizados en la industria. Durante esta etapa, se identificaron algunos problemas en los datos, como la falta de homogeneidad en los nombres de las estructuras que corresponden a la combinación de diferentes tipos de

materiales. Para abordar este problema, se realizó una homologación de los nombres de las estructuras para garantizar la consistencia en los datos.

Basándose en las condiciones identificadas durante la etapa de limpieza de datos, se reconoce la importancia de las condiciones identificadas durante la etapa de limpieza de datos para garantizar que los empaques cumplan con las especificaciones requeridas por el negocio.

Una de las condiciones fundamentales es que la permeabilidad de los materiales no sea negativa. Esto se debe a que una permeabilidad negativa indicaría un comportamiento anómalo y no deseado en los empaques, lo cual comprometería la calidad del producto envasado. Por lo tanto, durante el proceso de limpieza de datos, se realiza una revisión exhaustiva para detectar y corregir cualquier valor de permeabilidad negativo, asegurando así que los empaques cumplan con los estándares de permeabilidad establecidos.

Además, en el caso específico de la línea de empaques plásticos, se establece la regla de negocio de utilizar estructuras únicas y avaladas por el equipo de laboratorio. Esto implica que se ha llevado a cabo un riguroso proceso de validación de las estructuras utilizadas en los empaques plásticos para asegurar su idoneidad y eficacia en términos de permeabilidad. Este proceso de validación táctico garantiza que los empaques plásticos utilizados en la línea cumplan con los estándares de calidad y sean capaces de proporcionar una protección adecuada al producto envasado.

- Preparación de los datos: en esta fase, se realizó una limpieza exhaustiva de los datos para garantizar la calidad y confiabilidad del análisis. Se identificaron datos negativos en la base de datos, lo cual no es posible en términos de permeabilidad.

Para solucionar esto, se aplicó una técnica de imputación utilizando el algoritmo k-Nearest Neighbors (k-NN) para reemplazar los valores negativos por estimaciones basadas en los vecinos más cercanos.

Además, se depuraron las variables, teniendo en cuenta la correlación entre ellas. Aquellas variables altamente correlacionadas se evaluaron y se eliminó una de ellas para evitar la redundancia. También se eliminaron columnas que no aportan valor significativo al análisis, como variables de tiempo, nombres y columnas en blanco.

- Modelado: en esta etapa, se aplicaron diferentes modelos de machine learning, como la regresión lineal, máquinas de soporte vectorial y Random Forest, para predecir la permeabilidad de los materiales de empaque. Se realizaron múltiples iteraciones, ajustando los parámetros de cada modelo y evaluando su rendimiento en términos de métricas de evaluación, como el error cuadrático medio (mse) y error absoluto medio (mae) .

- Evaluación: después de generar los modelos, se evaluó su rendimiento dividiendo las bases de en 80% de entrenamiento y el 20% para testeo y validamos con las métricas de MAE y MSE. Se compararon los resultados de los modelos y se seleccionó el modelo final “Random Forest” que mostró el mejor desempeño en la predicción de la permeabilidad de los materiales de empaque.

  Los siguientes fueron los resultados de los experimentos realizados con los diferentes modelos que implementamos

  - **Regresión lineal:** se selecciona este método de regresión lineal para realizar el análisis de datos que predice el valor de datos desconocidos mediante el uso de otro valor de datos relacionado y conocido.

	MAE	MSE
TRAIN	0.725368	1.935316

	MAE	MSE
TEST	1.048757	3.965027

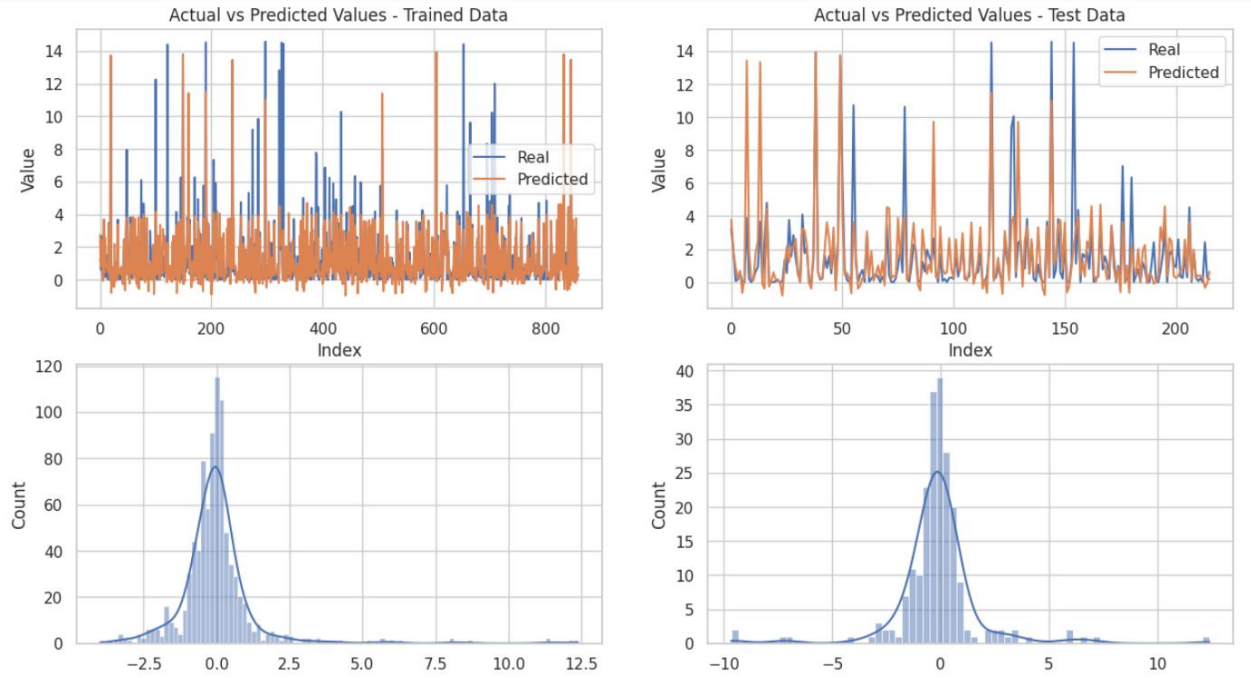

Figura 13. Resultado valores reales vs predichos modelo regresión lineal

- **Máquinas de Soporte Vectorial:**

	MAE	MSE
TRAIN	1.151115	5.751138

	MAE	MSE
TEST	1.350651	7.92231

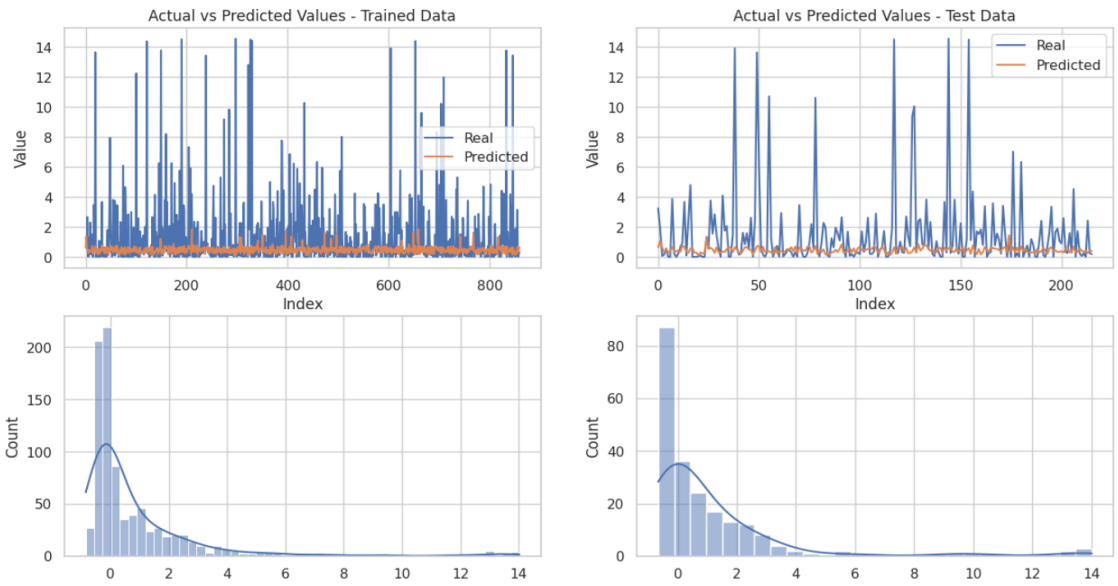

Figura 14. Resultado valores reales vs predichos modelo máquinas de Soporte Vectorial

- **Random Forest**

	MAE	MSE
TRAIN	0.187896	0.219317

	MAE	MSE
TEST	0.716245	3.216306

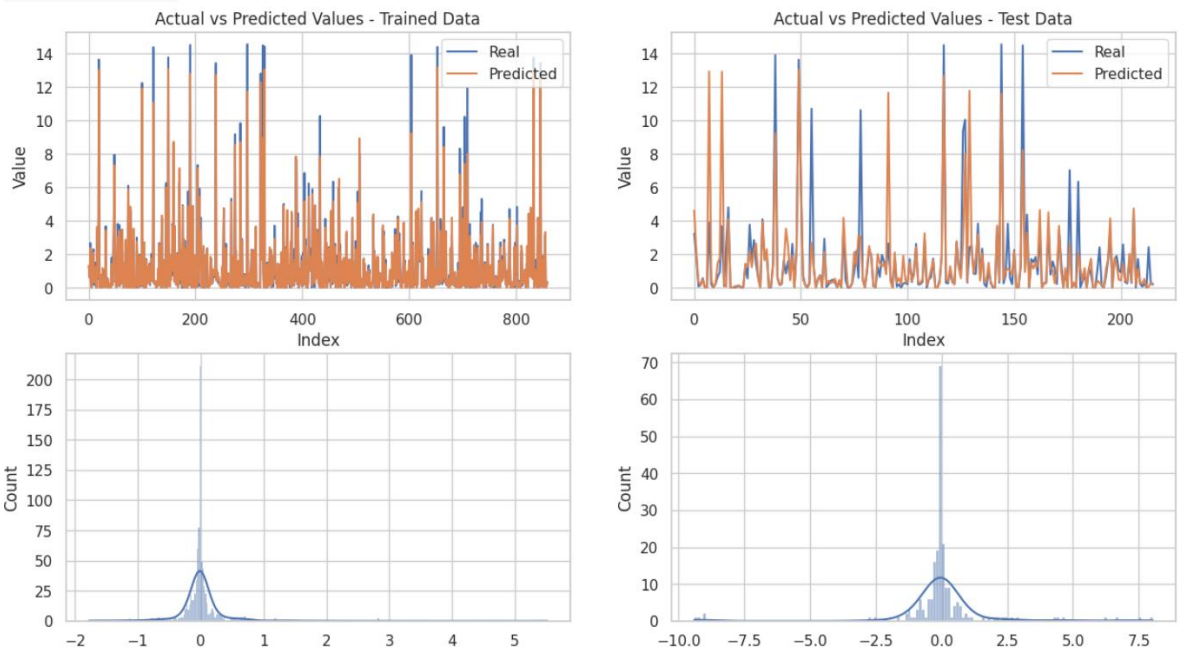

Figura 15. Resultado valores reales vs predichos Random Forest

Al analizar los resultados de los tres modelos que predicen la permeabilidad, teniendo en cuenta que la variable permeabilidad tiene una distribución de Poisson, se puede evaluar el rendimiento de los modelos utilizando las métricas de MAE (Error Absoluto Medio) y MSE (Error Cuadrático Medio).

En los datos de entrenamiento, el Modelo de Regresión Lineal muestra un MAE de 0.725 y un MSE de 1.93. El Modelo de Máquinas de Soporte Vectorial tiene un MAE de 1.151 y un MSE de 5.75, mientras que el Modelo de Random Forest muestra un MAE de 0.187 y un MSE de 0.219.

En los datos de prueba, el Modelo de Regresión Lineal obtiene un MAE de 1.048 y un MSE de 3.96. El Modelo de Máquina de Soporte Vectorial presenta un MAE de 1.35 y un MSE de 7.92, y el Modelo de Random Forest muestra un MAE de 0.716 y un MSE de 3.21.

Considerando estas métricas, se puede concluir que el mejor modelo para predecir la permeabilidad del empaque es el **Modelo de Random Forest**. Este modelo muestra el menor MAE tanto en los datos de entrenamiento (0.187) como en los datos de prueba (0.716). Además, presenta el MSE más bajo en los datos de entrenamiento (0.219) y un MSE competitivo en los datos de prueba (3.21). Estos resultados sugieren que el Modelo de Random Forest tiene un mejor ajuste y capacidad predictiva en comparación con los otros modelos evaluados.

- Despliegue: en la utilización del modelo Random Forest para el análisis de la permeabilidad en empaques, es importante tener en cuenta la medición del MAE (Mean Absolute Error) y el MSE (Mean Squared Error), los cuales deben ser menores a 1. Esto indica que los errores de predicción del modelo son relativamente bajos y cercanos a los valores reales.

  Por lo tanto, se recomienda que al evaluar el rendimiento del modelo Random Forest, se verifique que tanto el MAE como el MSE se encuentren por debajo de este umbral. Un MAE y MSE inferiores a 1 indican una mayor precisión en las predicciones y una menor variabilidad en los errores de estimación.

  Si los valores de MAE y MSE se encuentran en un rango recomendable, se puede afirmar que el modelo Random Forest está proporcionando resultados confiables y precisos para la predicción de la permeabilidad de los empaques. Sin embargo, es importante tener en cuenta que el rango óptimo puede variar dependiendo del tipo de estructura del empaque.

Para garantizar la efectividad continua del modelo, se estableció un proceso de monitoreo para evaluar su desempeño en producción. Esto implica el seguimiento regular de las predicciones del modelo, la evaluación de su precisión y la identificación de posibles desviaciones o necesidades de ajuste.

CONCLUSIONES

Evaluar de forma subjetiva la permeabilidad de los materiales utilizados en los empaques, es propenso a errores y resultados inconsistentes. Esto desata la necesidad de desarrollar métodos más precisos y confiables para predecir la permeabilidad como el uso de técnicas de análisis de datos para comprender las características de los mismos y seleccionar las variables relevantes en la predicción de la permeabilidad.

El análisis exploratorio de datos permitió comprender la naturaleza de los datos utilizados, proporcionar una visión general de la distribución de las variables, identificar posibles valores atípicos o anómalos, y revelar patrones o relaciones entre las variables. Este análisis permitió obtener conocimientos iniciales, lo que ayudó a tomar decisiones informadas en cuanto a qué métodos y técnicas eran más adecuados para el análisis posterior, y proporcionar una visión de la naturaleza de los datos y de su idoneidad para los objetivos planteados.

La limpieza y preparación exhaustiva de los datos, incluyendo el tratamiento de valores atípicos y la imputación de datos faltantes, fueron importantes en el análisis de datos para obtener resultados más precisos y confiables en la predicción de la permeabilidad, ya que casi todo conjunto de datos presenta anomalías en sus datos, lo que puede influir en el resultado final de la toma de decisiones.

La importancia de utilizar técnicas de análisis de datos avanzadas y modelos predictivos para mejorar la evaluación de la permeabilidad en los materiales de los empaques, influye en aplicar enfoques que pueden lograr mejoras significativas en la calidad y durabilidad de los productos, así como una reducción en el riesgo de productos defectuosos. Los modelos predictivos, como Random Forest y máquinas de soporte vectorial, calculan una precisión alta en la predicción de la permeabilidad de los materiales de los empaques. Estos modelos muestran que podrían ser

utilizados como herramientas eficientes en la industria para mejorar la calidad y durabilidad de los empaques.

Los métodos supervisados pueden manejar tanto variables continuas como categóricas, como en el conjunto de datos analizado. Por lo tanto, los métodos supervisados como la regresión lineal, las SVM y el random forest podrían adaptarse mejor a esta variabilidad y ofrecer una representación más precisa de la relación entre las variables y la permeabilidad.

Con la metodología CRISP-DM se proporcionó una estructura sistemática y bien definida para llevar a cabo el proyecto de analítica de datos, y permitió definir claramente los objetivos y las preguntas de investigación, estableciendo así el marco para el análisis de datos.

Este proyecto no se limitó únicamente a aplicar técnicas de modelado, sino que se enfocó en utilizar los datos para respaldar la toma de decisiones, lo cual implicó considerar el contexto del problema, e involucrar a las partes del negocio interesadas en esta solución, para garantizar que los resultados fueran prácticos y aplicables en el contexto real.

REFERENCIAS

[1] P. CN. "EDA - Exploratory Data Analysis: Using Python Functions". DigitalOcean | The Cloud for Builders. https://www.digitalocean.com/community/tutorials/exploratory-data-analysis-python (accedido el 5 de mayo de 2023).

[2] "7 Análisis exploratorio de datos (EDA) | _main". Bienvenida | _main. https://es.r4ds.hadley.nz/análisis-exploratorio-de-datos-eda.html (accedido el 5 de mayo de 2023)

[3] "Crisp DM". PowerPoint Templates, Graphics and Themes - PPT Slides | SketchBubble. https://www.sketchbubble.com/en/presentation-crisp-dm.html (accedido el 5 de mayo de 2023).

[4] Fraunhofer, "IPI Expert Seminar", 9 de diciembre de 2011, IPI PASIVE BARRIER, Google Drive - Carpeta Monografía.

ANEXOS

Notebook de Google Colab:

https://colab.research.google.com/drive/1sXDAQ5dS6urfTO8cwSLeTpBT_gb9-2gf?usp=sharing