

**Universidad Peruana de Ciencias Aplicadas**

**Inteligencia Artificial**

**Ciclo:** 2024-2

**Sección:** CC63

**Docente:** Flor Cagniy Cardenas Mariño

**Informe Trabajo Final**

**Integrantes:**

* Jhonny Elias Ruiz Santos - U201910128
* Hector Jesus Quintana Robatti - U20201B280

2024

# **ÍNDICE**

1. Introducción…………………………………………………………………………...3
2. Descripción del problema……………………………………………………………..3
3. Descripción y visualización del dataset……………………………………………….4
4. Propuesta……………………………………………………………………………..11
5. Diseño de aplicativo………………………………………………………………….12
6. Validación de resultados y pruebas………………………………………………......15
7. Conclusiones………………………………………………………………………....17
8. Referencias Bibliográficas..……….…………………………….……………..….…17

# **Introducción**

# El presente trabajo se basa en la aplicación de diferentes técnicas avanzadas de inteligencia artificial con el objetivo de anticipar el desempeño futuro de una tienda y evaluar su viabilidad a largo plazo. A través del análisis de datos y la implementación de algoritmos predictivos, se busca determinar la probabilidad de que la tienda mantenga su viabilidad en el mercado. Para lograr esto se creará un sistema de predicción de rentabilidad. La solución hará uso de una interfaz web para mayor comodidad del usuario. El objetivo principal es ofrecer una perspectiva más precisa del posible éxito o fracaso del negocio en un determinado tiempo, proporcionando así información valiosa para la toma de decisiones estratégicas que puedan asegurar su continuidad.

# **Descripción del problema**

En la actualidad los supermercados enfrentan diversos problemas, siendo uno de los principales la sostenibilidad económica del negocio. Factores externos, como atención al cliente deficiente o falta de stock por un mal registro de ventas.Todo esto puede impactar negativamente en las ventas y la fidelidad de los clientes, lo que podría traducirse en una baja o incluso nula rentabilidad.

Ahora bien, ¿qué pasaría si fuera posible anticipar la rentabilidad futura de una tienda? Esta propuesta busca mejorar la gestión y viabilidad de los supermercados al proporcionar a los encargados ciertos datos para tener una visión predictiva del desempeño financiero próximo, permitiéndoles implementar estrategias preventivas y mitigar el riesgo de cierre.

# **Descripción y visualización del dataset**

**Metodología**

Para el desarrollo del trabajo, se seguirá una metodología estructurada en varias etapas:

1. Obtención de los datos:

El dataset empleado en este proyecto fue obtenido de la plataforma Kaggle, bajo el título “Supermarket Store Branches Sales Analysis”. Este conjunto de datos incluye información detallada sobre 896 sucursales de supermercados, incorporando variables como el tamaño del área de cada tienda, el número de clientes atendidos diariamente, el inventario disponible y las ventas totales. Dicha información resulta adecuada para realizar un análisis enfocado en la viabilidad y rentabilidad de las tiendas comerciales.

1. Preparación de los datos:

Además de limpiar y estandarizar el dataset, se generaron nuevas variables derivadas con el objetivo de enriquecer el análisis y capturar mejor las relaciones entre las variables existentes. Estas nuevas características se diseñaron utilizando cálculos específicos basados en las variables originales:

1. **Items Per Customer:** Proporción entre los artículos disponibles en la tienda y el número diario de clientes, representando la disponibilidad promedio por cliente.
2. **Items Per Area:** Relación entre los artículos disponibles y el área de la tienda, lo que permite analizar la densidad de inventario en función del espacio.
3. **Items Area Interaction:** Interacción entre el inventario disponible y el área de la tienda, para explorar cómo estos factores combinados influyen en las ventas.
4. **Sales Per Customer:** Promedio de ventas por cliente diario, reflejando el gasto promedio por cliente.
5. **Sales Per Item:** Relación entre las ventas totales y los artículos disponibles, para evaluar la eficiencia del inventario en la generación de ingresos.

Estas variables adicionales se seleccionaron para complementar las características originales del dataset y proporcionar mayor información al modelo predictivo. La variable objetivo (*target*) utilizada en el análisis fue **Store Sales**, mientras que las variables predictoras incluyeron tanto las originales como las generadas:

* Variables originales: Items Available, Daily Customer Count, y Store Area.
* Variables derivadas: Items Per Customer, Items Per Area, Items Area Interaction, Sales Per Customer, y Sales Per Item.

Este enfoque permite que el modelo de machine learning tenga acceso a información más rica y útil para realizar predicciones precisas.

1. División de los datos:

Para garantizar un entrenamiento adecuado del modelo y evaluar su capacidad de generalización, los datos fueron divididos en dos conjuntos:

* **Conjunto de entrenamiento (80%):** Utilizado para entrenar el modelo de aprendizaje automático.
* **Conjunto de prueba (20%):** Reservado para evaluar el desempeño del modelo en datos no previamente analizados.

La división de los datos se realizó de manera aleatoria, preservando la distribución de la variable objetivo (*Store Sales*). Esto asegura que el modelo sea capaz de aprender patrones representativos del conjunto completo y que los resultados obtenidos durante las evaluaciones sean fiables.

Adicionalmente, todas las variables predictoras se escalaron utilizando el método de **escalado estándar (StandardScaler)**. Este paso fue crucial para normalizar las magnitudes de las variables y evitar que aquellas con valores más grandes dominaran el entrenamiento del modelo.

1. Selección del modelo:

Inicialmente se propuso el uso de un modelo de regresión lineal para realizar las predicciones debido a su facilidad de interpretación y por ser la mejor opción para hacer predicciones. Sin embargo, al evaluar el desempeño del modelo, se encontró que el coeficiente de determinación obtenía valores negativos, indicando que el modelo no ajustaba correctamente los datos y era inadecuado para capturar las relaciones lineales en el conjunto de datos.   
Como solución, se optó por utilizar el algoritmo de Random Forest, un método basado en árboles de decisión que es robusto frente a relaciones complejas y no lineales. Hacer este cambio permitió obtener predicciones más precisas y un mejor desempeño general al aprovechar su capacidad para manejar grandes volúmenes de datos. Random Forest fue elegido porque:

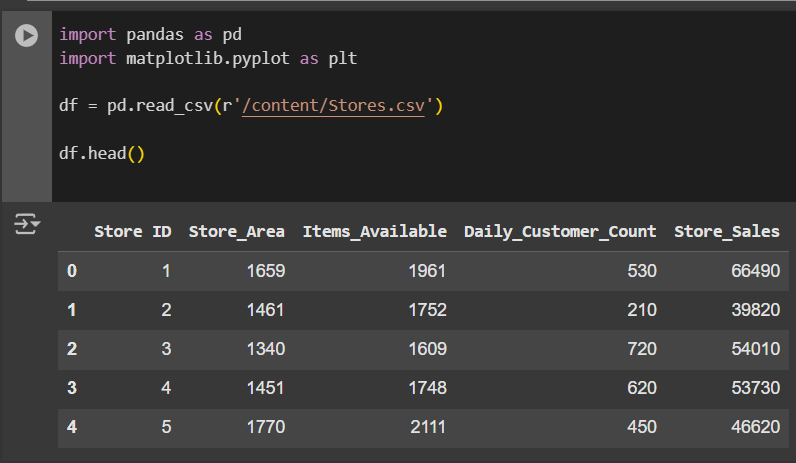
* Es eficaz para capturar relaciones no lineales entre las variables.
* Maneja bien las interacciones entre variables derivadas.
* Es menos susceptible al sobreajuste, especialmente cuando se ajustan correctamente los hiperparámetros como el número de árboles y la profundidad máxima.

El desempeño del modelo fue evaluado utilizando las métricas de **Error Cuadrático Medio (MSE)** y el coeficiente de determinación (R2R^2R2), mostrando una mejora significativa en comparación con la regresión lineal.

**Estructura del dataset**

| **variable** | **tipo** | **descripción** |
| --- | --- | --- |
| Store\_ID | int64 (Enteros de 64 bits) | Identificador único de cada tienda |
| Store\_Area | int64 (Enteros de 64 bits) | Espacio total de la tienda |
| Items\_Available | int64 (Enteros de 64 bits) | Cantidad de items disponibles en la tienda |
| DailyCustomerCount | int64 (Enteros de 64 bits) | Cantidad diaria de clientes que visitan la tienda |
| Store\_Sales | int64 (Enteros de 64 bits) | Cantidad de ventas totales de la tienda |

**Analisis Exploratoria de Datos (EDA)**

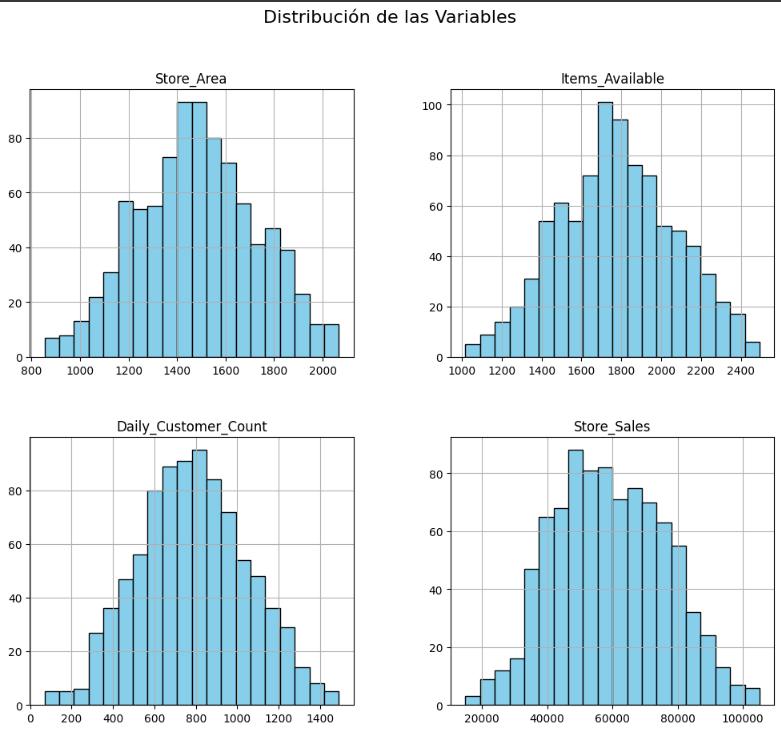
\*Verificar valores nulos

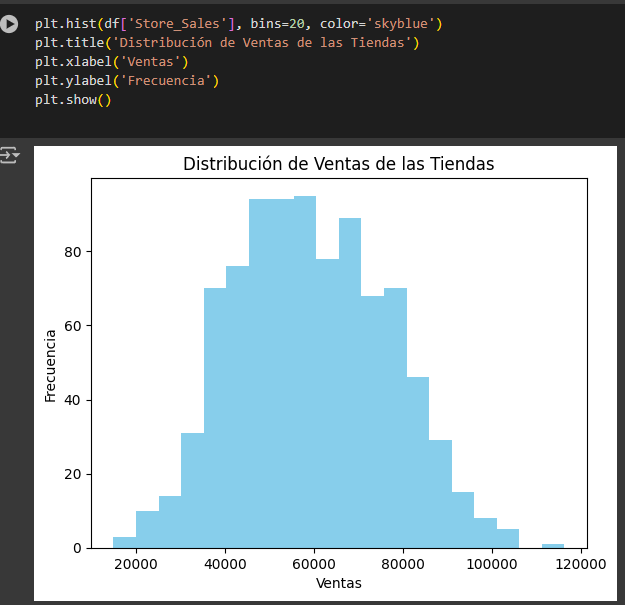
# 

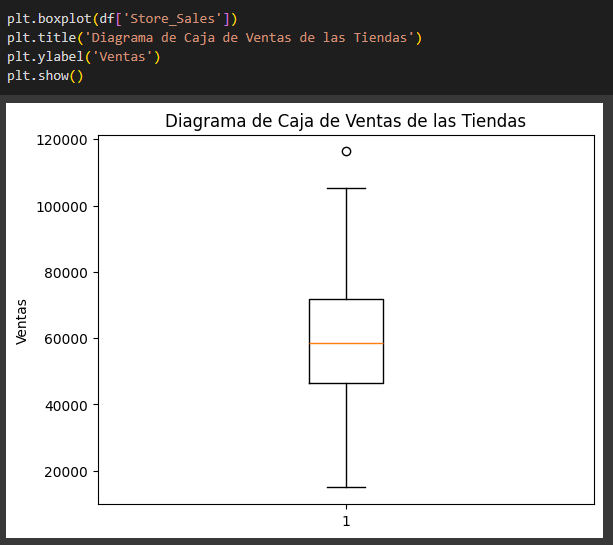
# 

# 

# \*Gráficos e histograma de ventas







# 

# 

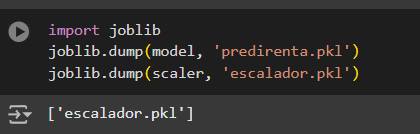
# 

# 

# 

# 

# 



# **Propuesta**

La solución propuesta consiste en desarrollar un sistema predictivo basado en un modelo de aprendizaje automático, utilizando inicialmente **regresión lineal** como punto de partida. Sin embargo, debido a limitaciones en el desempeño del modelo lineal, evidenciadas por un coeficiente de determinación negativo, así que se ajustó posteriormente a un modelo de **Random Forest Regressor**. Este enfoque avanzado permite capturar relaciones no lineales entre las variables, lo que resulta clave para un análisis más preciso.

El sistema utilizará un conjunto de datos históricos que incluye información como:

* **Variables originales**: Número de artículos disponibles (*Items Available*), cantidad diaria de clientes (*Daily Customer Count*), área de la tienda (*Store Area*) y las ventas totales (*Store Sales*).
* **Nuevas variables derivadas**: Se introdujeron variables adicionales que enriquecen el modelo, tales como:
  + Artículos por cliente (*Items Per Customer*).
  + Artículos por unidad de área (*Items Per Area*).
  + Interacción entre artículos disponibles y área de la tienda (*Items Area Interaction*).
  + Ventas por cliente (*Sales Per Customer*).
  + Ventas por artículo (*Sales Per Item*).

Estas variables derivadas ayudan a capturar patrones más complejos en los datos, permitiendo al modelo entender mejor las interacciones y dinámicas entre los diferentes factores que afectan la rentabilidad.

El sistema predictivo será capaz de estimar con mayor precisión la viabilidad y rentabilidad de futuras tiendas comerciales.

# **Diseño de aplicativo**

* **Comprensión del negocio:**

Nuestro objetivo es crear una aplicación web que permita subir a los usuario un csv con los datos de su tienda, estos serán: Area, stock, clientes diarios, ventas mensuales. Una vez cargado este archivo, se procede a aplicar el algoritmo Random Forest que nos permitirá predecir si el negocio está siendo o no rentable. Para mayor confiabilidad, se decidió crear un login, asegurando que los datos de cada tienda sean únicos por cada usuario, a la vez, se podría implementar un historial de predicciones, para que cada usuario vea si el negocio está o bajando su rentabilidad.

* **Comprensión de datos:**

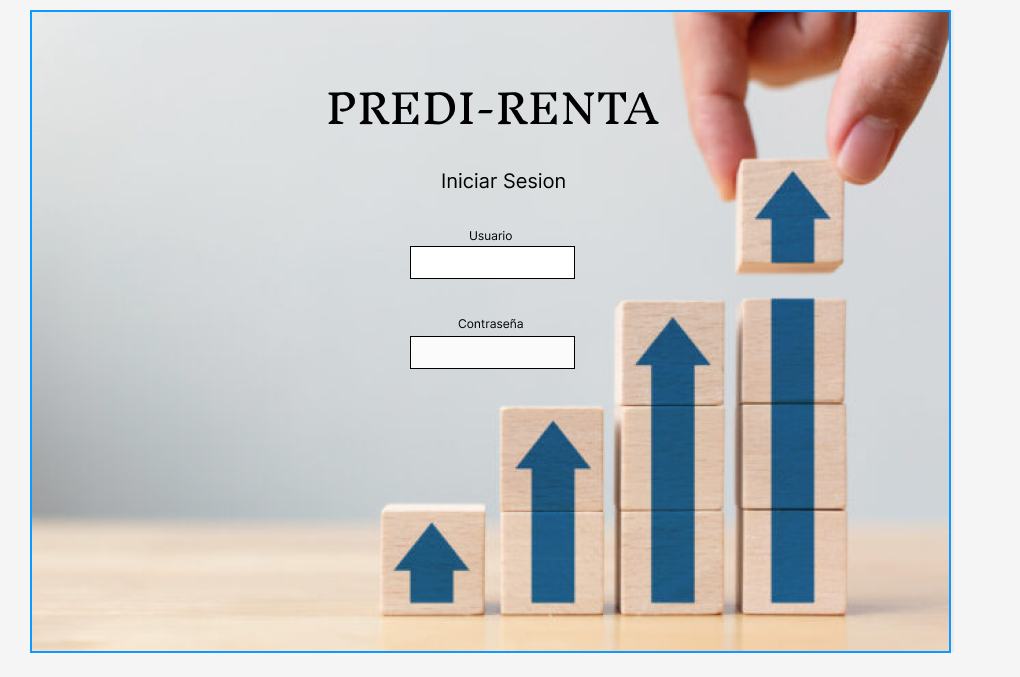
Una vez tengamos los datos necesarios, se procede a hacer una validación, para ello, el formulario solo permitirá subir archivos en formato csv.

* **Preparación de datos:**

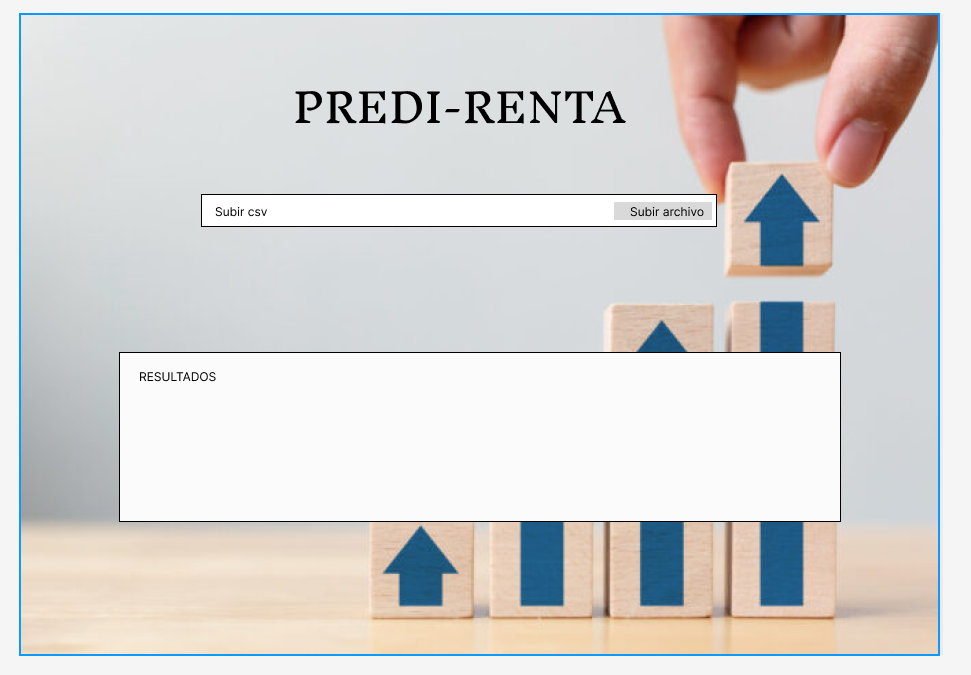
En esta etapa, el aplicativo podrá hacer la etapa de limpieza de datos de manera automática. Debido a que no se procesa una cantidad excesiva de datos en el caso de una tienda independiente, solo se incluye este pre-procesamiento de datos.

* **Prototipo:**

A continuación, mostramos el prototipo de nuestra aplicación, la cual tendrá por nombre “PREDI-RENTA”.

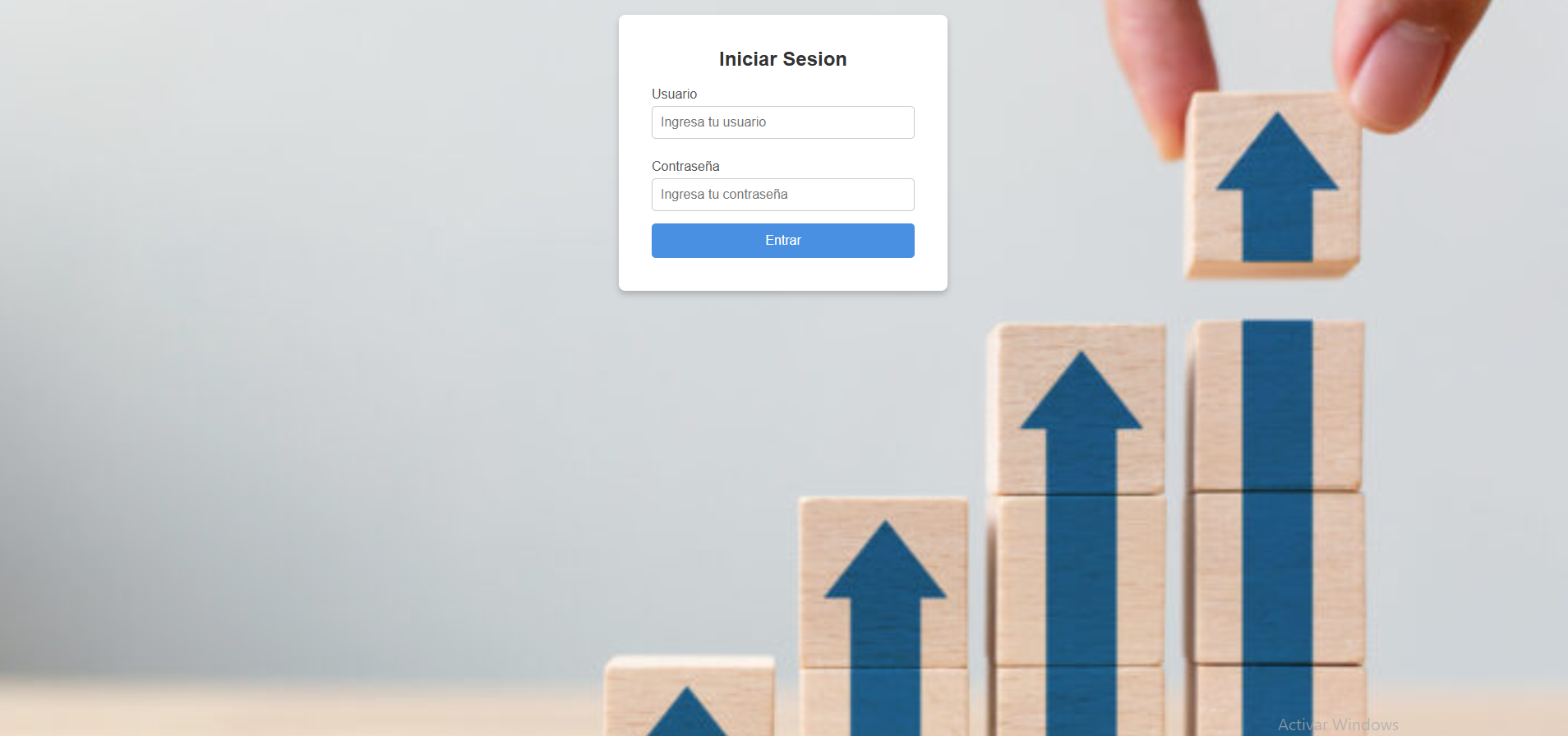
-Login:

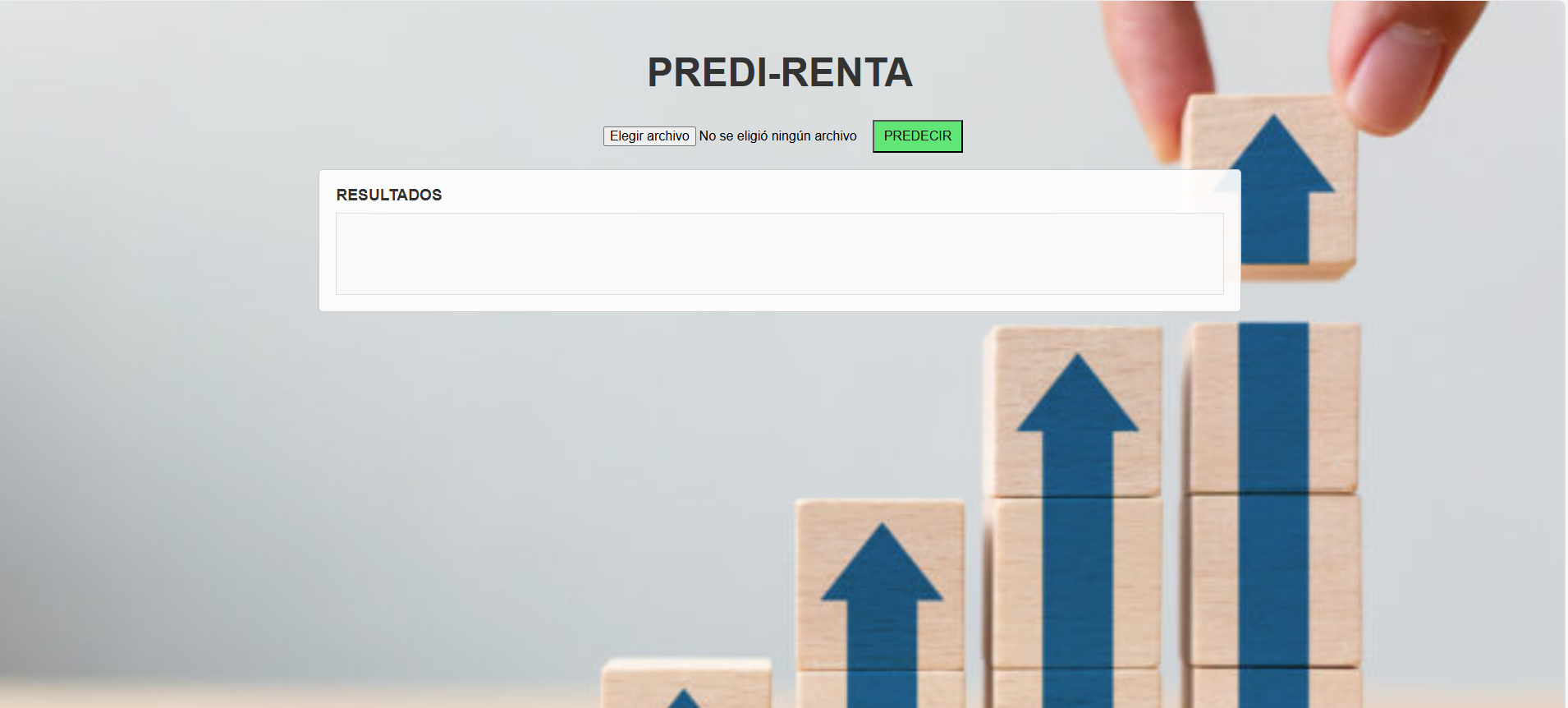
-Aplicación principal



* **Desarrollo:**

En esta sección, mostraremos el diseño de la aplicación ya desarrollado:

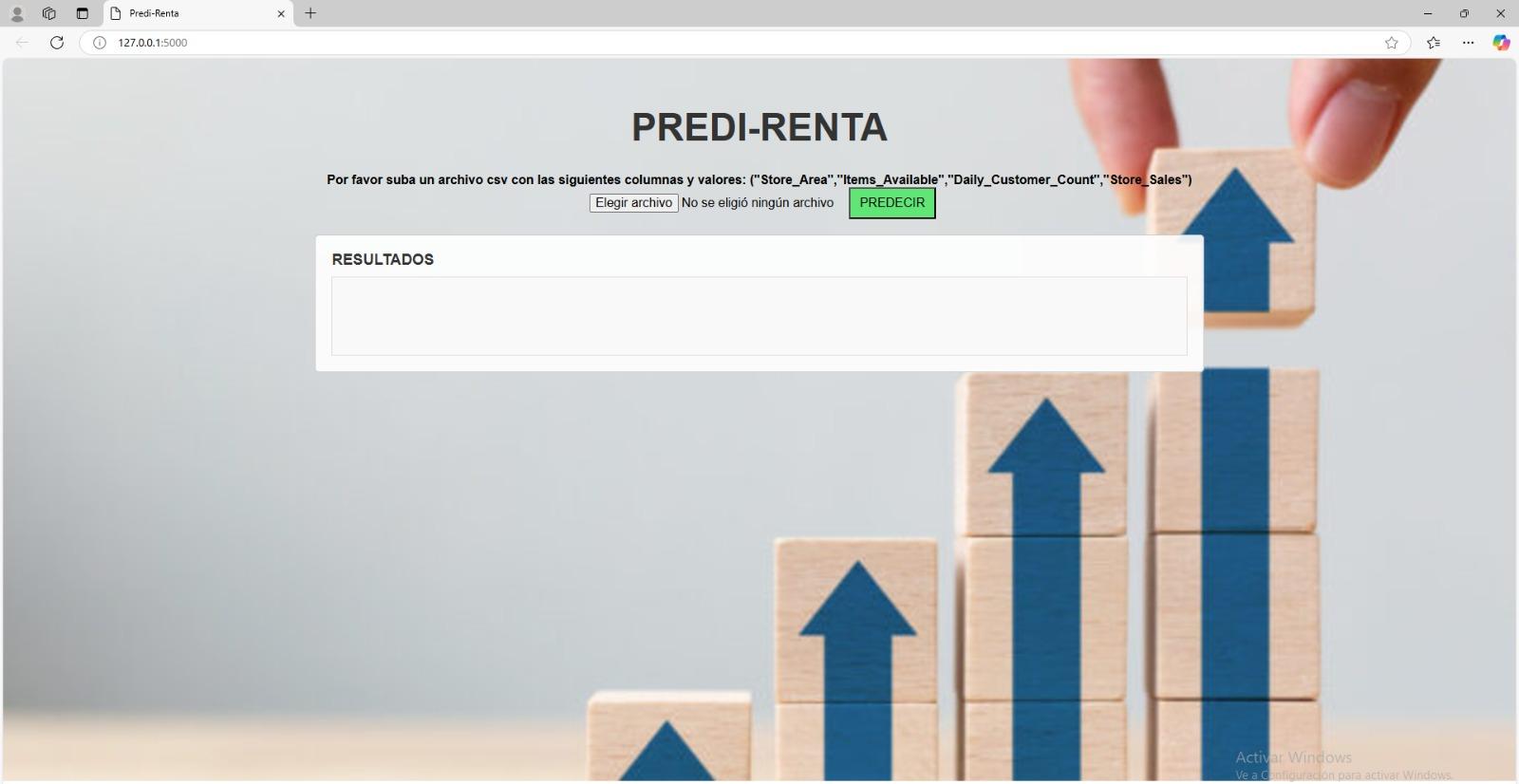
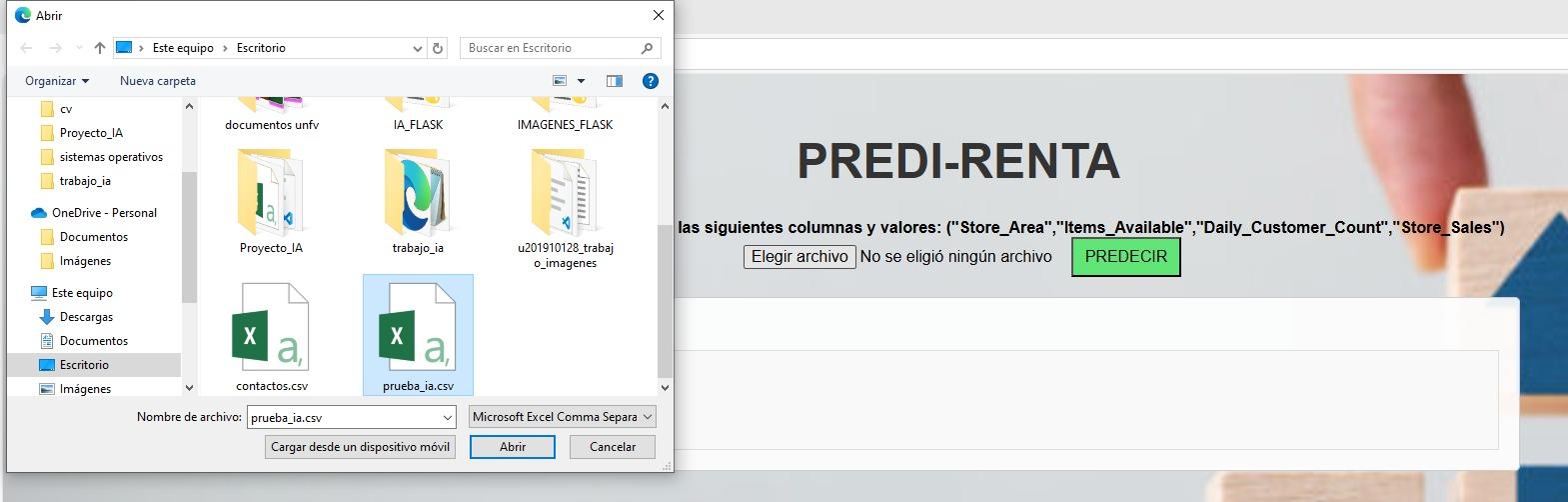
-Login:

-Aplicación principal.

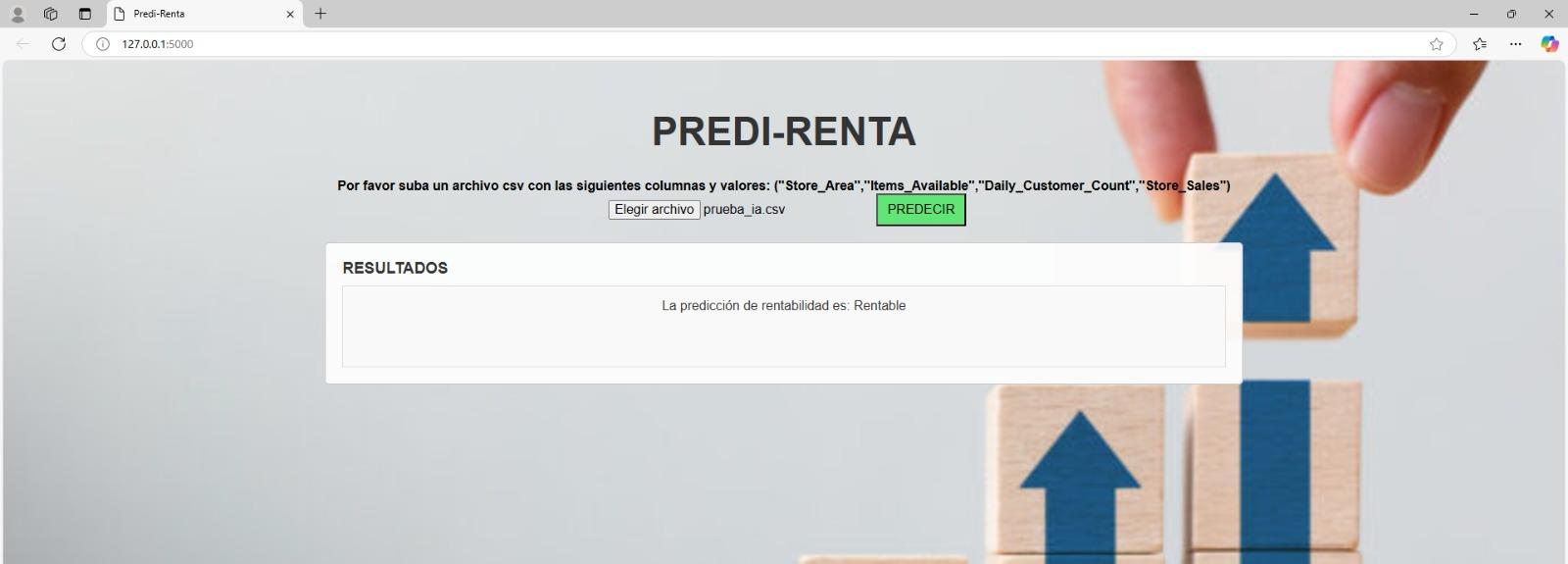
# **Validación de resultados y pruebas**

**Entradas:** El sistema requiere un archivo CSV que contenga datos clave relacionados con el desempeño de las tiendas comerciales. Las columnas aceptadas incluyen:

* **Store\_Area**: Área de la tienda (m²).
* **Items\_Available**: Número total de artículos disponibles en la tienda.
* **Daily\_Customer\_Count**: Promedio de clientes diarios.
* **Items\_Sales**: Ventas totales mensuales de la tienda.

El formato del archivo debe ser consistente con las columnas mencionadas para asegurar la compatibilidad con el modelo predictivo.

**Salidas:** El sistema genera una predicción basada en los datos ingresados, que refleja la estimación de ventas o rentabilidad esperada. Esta salida es presentada en un cuadro de texto en la interfaz, permitiendo al usuario analizar los resultados de manera sencilla.

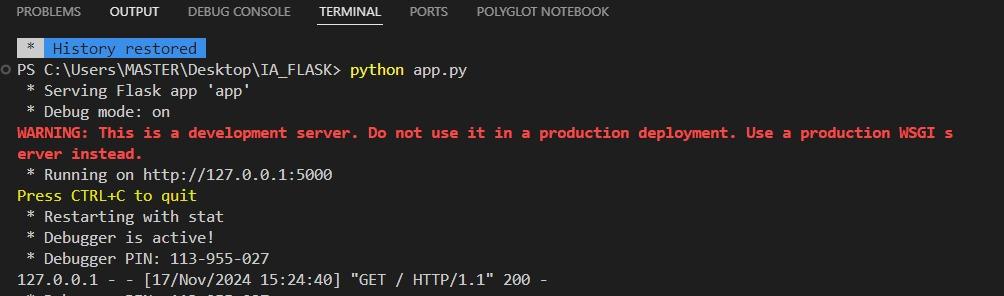


**Interpretación de Resultados:**

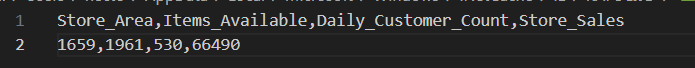
Los resultados predichos representan la proyección de la rentabilidad de una tienda en función de las características provistas. Por ejemplo:

* Una predicción alta indica que la tienda tiene una combinación favorable de área, inventario y flujo de clientes, lo que sugiere un buen potencial de ventas.
* Predicciones más bajas pueden reflejar deficiencias en alguno de los factores clave, lo que podría requerir ajustes operativos.

El objetivo es ofrecer a los gerentes de tiendas una herramienta confiable para evaluar la viabilidad de sus negocios o identificar oportunidades de mejora.



**Data de ejemplo ( Correcta estructura)**



# **Conclusiones**

El análisis llevado a cabo en este trabajo demuestra cómo las técnicas de inteligencia artificial pueden ser efectivamente aplicadas para predecir la rentabilidad y viabilidad de tiendas comerciales mediante el uso de modelos avanzados como Random Forest Regressor. El cambio de un enfoque inicial basado en regresión lineal a un modelo no lineal evidenció mejoras significativas en la precisión de las predicciones, resaltando la importancia de seleccionar algoritmos adecuados al contexto y la naturaleza de los datos. Las variables derivadas introducidas en el análisis jugaron un papel fundamental para enriquecer la representación de los datos y capturar interacciones complejas que contribuyen al desempeño financiero de las tiendas.

Futuras investigaciones podrían explorar la integración de otros modelos de aprendizaje automático, como redes neuronales o gradient boosting, para evaluar su desempeño en este tipo de problemas.

**Referencias bibliográficas**

* JS. (2021 - 2022). *Supermarket store branches sales analysis* [Dataset]. Kaggle.<https://www.kaggle.com/datasets/surajjha101/stores-area-and-sales-data>
* McKinney, W. (2017). *Python for Data Analysis: Data Wrangling with Pandas, NumPy, and IPython* (2da ed.). O'Reilly Media. <https://www.oreilly.com/library/view/python-for-data/9781491957653/>
* Geron, A. (2019). *Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow* (2da ed.). O'Reilly Media. <https://www.oreilly.com/library/view/hands-on-machine-learning/9781492032632/>