Informe Ejecutivo

Proyecto Júpiter

Eszter Patkai
Joel Pascual Moreno
Ingrid López Ulric

Índice

Acceso a Drive.

1. Equipo y objetivos

El proyecto ha sido desarrollado por **Eszter Patkai, Joel Pascual Moreno e Ingrid López Ulric**, combinando conocimientos en análisis de datos, *Machine Learning* (ML) y visualización.

Las responsabilidades se distribuyeron de la siguiente forma:

- **Eszter Patkai** se encargo del desarrollo de modelos de ML*, Deep Learning* (DL) y de la creación de *dashboards* interactivos en Power BI.
- **Joel Pascual Moreno** participó en el análisis exploratorio de datos (EDA), pipeline, las *queries* en MySQL, la definición del caso de uso de IA Generativa y la redacción del informe ejecutivo.
- **Ingrid López Ulric** participó con el EDA, la implementación de bases de datos en MySQL, el desarrollo de modelos de ML y visualización en Power BI.

El objetivo principal del proyecto es impulsar la transformación digital de una empresa del sector de distribución de alimentos, mediante la automatización de sus procesos mecánicos y operativos. Con objetivos específicos como automatizar el etiquetado mediante visión artificial, tomar decisiones basada en datos, facilitando la detección de errores, incidencias o anomalías en los sistemas.

2. Modelo relacional

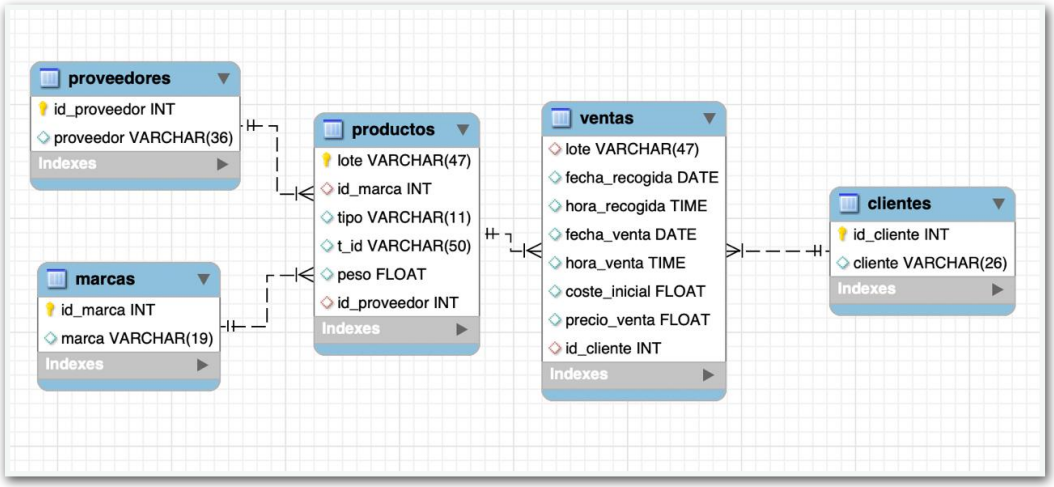
Se diseñó e implementó un modelo relacional en SQL con el objetivo de estructurar la información, atendiendo a los valores reales presentes en los datos y a las necesidades posteriores de análisis.

La elección de los tipos de datos, en particular de los campos VARCHAR, se basó en la longitud máxima observada en los datos originales, ajustando cada campo para evitar sobredimensionar. Identificadores alfanuméricos como el lote requirieron longitudes superiores, mientras que campos categóricos presentaron una longitud claramente acotada, lo que permitió limitar su tamaño sin pérdida de información.

Posteriormente, se realizaron uniones entre tablas mediante operaciones JOIN y se respondieron diversas preguntas de negocio, extrayendo información relevante mediante el uso de filtros, agrupaciones GROUP BY. Este proceso permitió identificar patrones, realizar comparaciones entre distintos grupos y obtener métricas de interés de forma estructurada y eficiente.

En conjunto, el trabajo realizado en SQL permitió organizar los datos de manera coherente y responder de forma precisa a las preguntas planteadas, aportando valor al análisis global del proyecto y sentando las bases para futuros análisis más avanzados.

Modelo relacional: tablas en MySQL Workbench



3. Limpieza de datos

Como parte del análisis exploratorio de datos, se llevó a cabo una revisión exhaustiva del conjunto de datos con el objetivo de identificar posibles problemas de calidad, tales como la presencia de valores atípicos (*outliers*), pesos inconsistentes, registros duplicados y valores nulos.

Durante esta fase inicial se detectó un elevado número de inconsistencias en los archivos originales (*.json*). Dado que este tipo de errores puede afectar negativamente a la fiabilidad de los análisis posteriores, se realizaron diversas correcciones orientadas a garantizar la coherencia, integridad y calidad de los datos utilizados en el proyecto. Se realizaron transformaciones en el formato de algunas variables. En particular, las variables de tipo fecha, que inicialmente se encontraban representadas como valores enteros, fueron convertidas a un formato de fecha y hora.

Asimismo, se procedió a la eliminación de registros duplicados y a la detección de pesos inconsistentes (negativos). Debido a la naturaleza de los datos, se decidió dividir el conjunto de datos en distintos archivos CSV según el tipo de fruta, ya que no resulta adecuado comparar directamente productos con características tan diferentes, como melones y peras.

A pesar de esta segmentación, se identificaron valores anómalos y registros claramente inconsistentes, como ventas de productos antes de su recogida, *missing values* o pesos completamente desorbitados, como una manzana de 2kgs cuando la media de un peso razonable está entre 150-200g.

Estos hallazgos llevaron a la conclusión de que el proceso de toma de peso o la forma en la que los datos fueron almacenados no es completamente fiable, y tras realizar las comprobaciones necesarias, se decidió mantenerlos ya que eliminarlos nos dejaría en algunos casos con pocos datos. En otros *bugs* cómo la duplicidad de los lotes o la falta de información, se implementaros técnicas de depuración y corrección necesarias para garantizar la coherencia, integridad y calidad de los datos utilizados en el proyecto.

4. Metodología de Machine Learning

En esta fase del proyecto se definió la metodología seguida para el desarrollo del sistema de etiquetado automático de frutas mediante técnicas de *Machine Learning* y *Deep Learning*.

El **objetivo principal** fue implementar y entrenar distintos modelos, evaluarlos bajo las mismas condiciones y analizar su comportamiento sobre el conjunto de datos.

En primer lugar, el conjunto de datos fue dividido en subconjuntos de entrenamiento y prueba, garantizando una evaluación objetiva del rendimiento de los modelos. Para todos los casos se utilizaron las mismas métricas de evaluación: *accuracy, classification report* y matriz de confusión.

Como primer modelo de ML, se implementó el algoritmo *K-Nearest Neighbors* (KNN), con el fin de obtener una referencia inicial debido a su simplicidad e interpretabilidad. Durante su entrenamiento, se analizó el impacto del número de vecinos (k) en el rendimiento del modelo, seleccionando el valor óptimo en función de los resultados obtenidos.

Posteriormente, se desarrolló un modelo *RandomForestClassifier,* basado en un conjunto de árboles de decisión. Este modelo fue elegido por su capacidad para manejar datos complejos y reducir el riesgo de sobreajuste, además de ofrecer buenos resultados en problemas de clasificación multiclase. Su entrenamiento se realizó bajo las mismas condiciones que el modelo KNN para garantizar una comparación justa.

En cuanto a los enfoques de *Deep Learning*, se implementó una Red Neuronal Convolucional (CNN) utilizando *TensorFlow*, diseñada para extraer automáticamente características relevantes de las imágenes y capturar relaciones no lineales entre las variables. Asimismo, se entrenó el modelo *MobileNetV2*, una arquitectura optimizada para eficiencia computacional, que permite obtener buenos resultados manteniendo un menor coste en términos de recursos.

Todos los modelos fueron entrenados y evaluados sobre el mismo conjunto de datos, lo que permitió establecer una base común para la posterior comparación de resultados. La aplicación de esta metodología proporcionó una visión integral del comportamiento de los distintos algoritmos y sentó las bases para la selección del modelo más adecuado para el sistema de etiquetado automático.

5. Comparación de modelos

En cuanto al rendimiento obtenido, se observaron diferencias significativas entre los modelos evaluados. El algoritmo de *K-Nearest Neighbors* (KNN) presentó el desempeño más bajo, con una *accuracy* aproximada del 72%. El análisis de su matriz de confusión evidenció un comportamiento desigual entre las distintas clases, mientras algunas frutas fueron clasificadas correctamente con una precisión aceptable, otras presentaron un alto nivel de error, lo que afectó negativamente al rendimiento global del modelo. Por este motivo, el modelo KNN fue descartado.

Por otro lado, el modelo de *Deep Learning* desarrollado con *TensorFlo*w alcanzó una *accuracy* cercana al 96%, lo que demuestra una mayor capacidad para capturar patrones complejos y generalizar correctamente sobre el conjunto de validación. Este resultado sugiere que las redes neuronales son especialmente eficaces en problemas con múltiples clases y relaciones no lineales entre las variables. No obstante, tras examinar las matrices de confusión, se concluyó que los resultados obtenidos por la CNN (Red Neuronal Convolucional) no eran lo suficientemente consistentes como para ser utilizada como base del sistema de etiquetado automático.

Finalmente, al comparar los modelos *RandomForestClassifier* y *MobileNetV2*, se observó que las diferencias de rendimiento eran mínimas. Sin embargo, todas las métricas evaluadas respaldaron la elección del *RandomForestClassifier*, al que se suma una ventaja clave, su tiempo de entrenamiento es considerablemente menor en comparación con los modelos de *Deep Learning*. Por estas razones, este modelo se posiciona como la opción más adecuada para la implementación del sistema automático de etiquetado de frutas.

	ML		DL	
	RandomForrestClassifier	KNN	MobilNetV2	CNN con TensorFlow
Accuracy	0.994472005669738	0.864209780297661	0.988163590431213	0.956410825252533
%	99.45%	86.42%	98.82%	95.64%