Practica Final

Alberto Jiménez Serrano

January 2024

Índice

1. Introducción

Durante esta práctica vamos a realizar un análisis de un dataset de transacciones a lo largo de un año, luego crear una serie de modelo de machine learning y por último tratar con Label Studio.
Label Studio es una herramienta de etiquetado de datos open-source. Nos permite etiquetar muchos tipos de datos como son audio, texto, imágenes, vídeos y series temporales para luego exportar a varios formatos de modelos.

2. Análisis de datos

Al importar los datos observamos que tiene las siguientes columnas:
InvoiceNo: Es un identificador único para cada transacción.
StockCode: Un código que representa un producto en especifico en el inventario.
Description: La descripción del producto.
Quantity: La cantidad de cada producto en cada transacción.
InvoiceDate: La fecha y hora de la transacción.
UnitPrice: El precio por unidad del producto.
CustomerID: Identificador único para cada cliente.
Country: El país donde se ha hecho la transacción o donde el consumidor se encuentra.

2.1. Información de datos

	Quantity	UnitPrice	CustomerID
count	541909.000000	541909.000000	406829.000000
mean	9.552250	4.611114	15287.690570
std	218.081158	96.759853	1713.600303
min	-80995.000000	-11062.060000	12346.000000
25%	1.000000	1.250000	13953.000000
50%	3.000000	2.080000	15152.000000
75%	10.000000	4.130000	16791.000000
max	80995.000000	38970.000000	18287.000000

Hay elementos que tienen cantidad negativa y el precio por unidad negativos por lo que eliminaremos los que tengan valor negativo o 0.
También tenemos menos datos en CustomerID por lo que hay valores nulos en esta columna.

2.2. Tipo de datos

```
InvoiceNo       object
StockCode       object
Description     object
Quantity         int64
InvoiceDate     object
UnitPrice      float64
CustomerID     float64
Country         object
dtype: object
```

Tenemos la fecha en tipo object y es preferible tratarlos como datos del tipo datetime por lo que vamos a cambiarle el tipo.

2.3. Valores nulos

```
CustomerID     24.926694
Description     0.268311
InvoiceNo       0.000000
StockCode       0.000000
Quantity        0.000000
InvoiceDate     0.000000
UnitPrice       0.000000
Country         0.000000
dtype: float64
```

En CustomerID tenemos un 25 % de valores nulos y algunas descripciones. Por lo tanto eliminaremos la columna ya que no nos va a dar suficientes datos buenos.
Para las descripciones nulas, vemos a ver que datos son aquellos con descripciones nulas y dependiendo de que datos sean tomaremos una decisión.

	InvoiceNo	StockCode	Description	Quantity	InvoiceDate	UnitPrice	Country
622	536414	22139	NaN	56	2010-12-01 11:52:00	0.0	United Kingdom
1970	536545	21134	NaN	1	2010-12-01 14:32:00	0.0	United Kingdom
1971	536546	22145	NaN	1	2010-12-01 14:33:00	0.0	United Kingdom
1972	536547	37509	NaN	1	2010-12-01 14:33:00	0.0	United Kingdom
1987	536549	85226A	NaN	1	2010-12-01 14:34:00	0.0	United Kingdom

Podemos observar que todos los que no tienen descripción tienen valor 0 en el precio por unidad por lo tanto no pasará nada si eliminamos esas transacciones ya que no tiene sentido hacer una transacción de 0 euros.

2.4. Valores erróneos

La siguiente imagen nos muestra los valores que hay en la columna de cantidad y las veces que aparece ese valor.

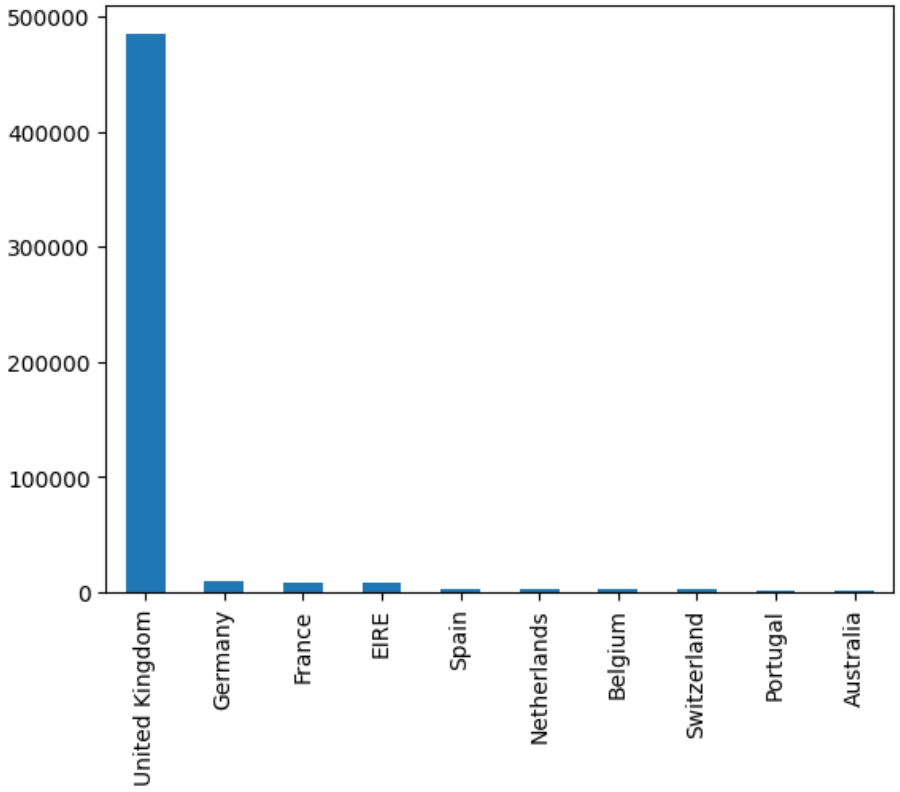
```
1        148101
2         81763
12        61049
6         40846
4         38461
          ...
3186          1
291           1
172           1
3100          1
-80995        1
Name: Quantity, Length: 671, dtype: int64
```

Hay elementos que tienen cantidad negativa por lo que eliminamos que tengan valor negativo o 0 ya que no tiene sentido tener una cantidad negativa de algo.
En el precio también hay valores negativos y ceros por lo que los eliminaremos.

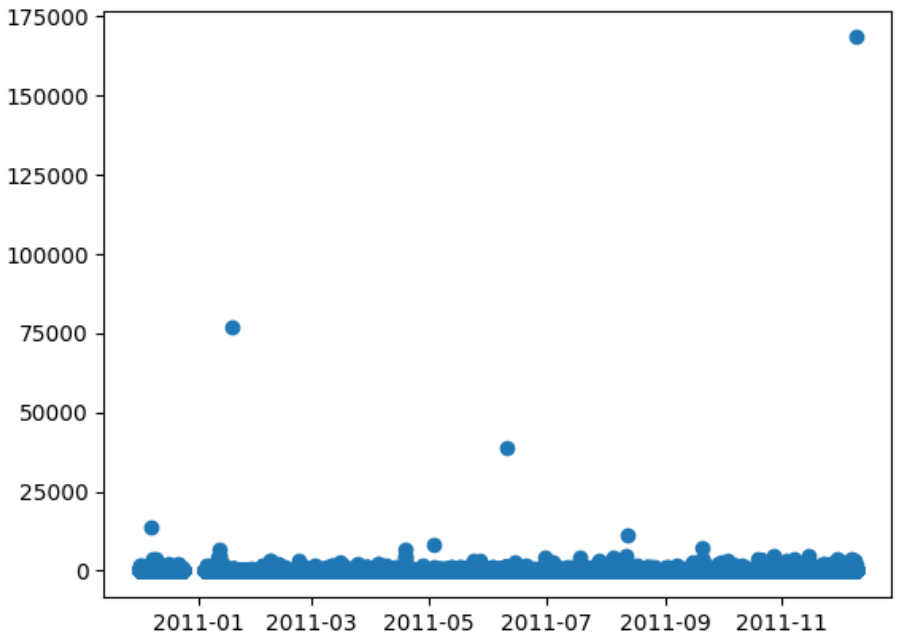
	InvoiceNo	StockCode	Description	Quantity	InvoiceDate	UnitPrice	Country
299984	A563187	B	Adjust bad debt	1	2011-08-12 14:52:00	-11062.06	United Kingdom
299983	A563186	B	Adjust bad debt	1	2011-08-12 14:51:00	-11062.06	United Kingdom
41478	539856	22624	IVORY KITCHEN SCALES	3	2010-12-22 14:41:00	0.00	United Kingdom
41504	539856	21888	BINGO SET	2	2010-12-22 14:41:00	0.00	United Kingdom
41505	539856	21539	RED RETROSPOT BUTTER DISH	3	2010-12-22 14:41:00	0.00	United Kingdom
...	...	...	...	...	...	...	...
268028	560373	M	Manual	1	2011-07-18 12:30:00	4287.63	United Kingdom
297723	562955	DOT	DOTCOM POSTAGE	1	2011-08-11 10:14:00	4505.17	United Kingdom
173382	551697	POST	POSTAGE	1	2011-05-03 13:46:00	8142.75	United Kingdom
299982	A563185	B	Adjust bad debt	1	2011-08-12 14:50:00	11062.06	United Kingdom
15017	537632	AMAZONFEE	AMAZON FEE	1	2010-12-07 15:08:00	13541.33	United Kingdom

3. Visualización de datos

Al mirar el número de transacciones por país, podemos observar que la mayoría de los datos pertenecen a Reino Unido por lo que podremos separar los datos entre aquellos que pertenecen a Reino Unido y aquellos que no.

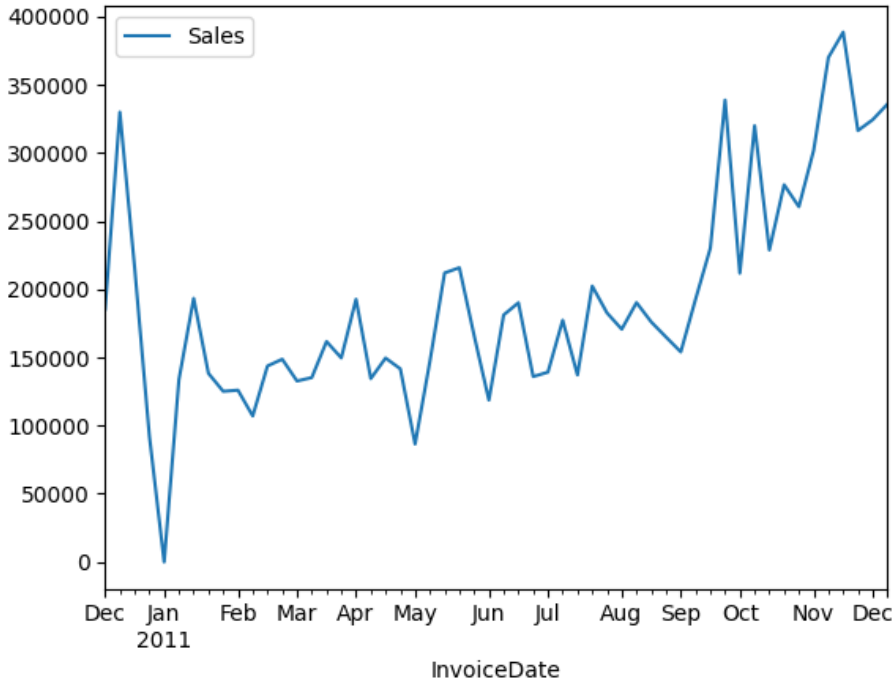


En el siguiente gráfico podemos ver las ventas y observamos valores extraños que no siguen la normal por lo que al ser pocos casos podemos prescindir de ellos.



	Quantity	UnitPrice	Sales
0.05	1.0	0.42	1.25
0.95	30.0	9.95	59.70
0.98	72.0	14.95	121.68
0.99	100.0	16.98	183.60

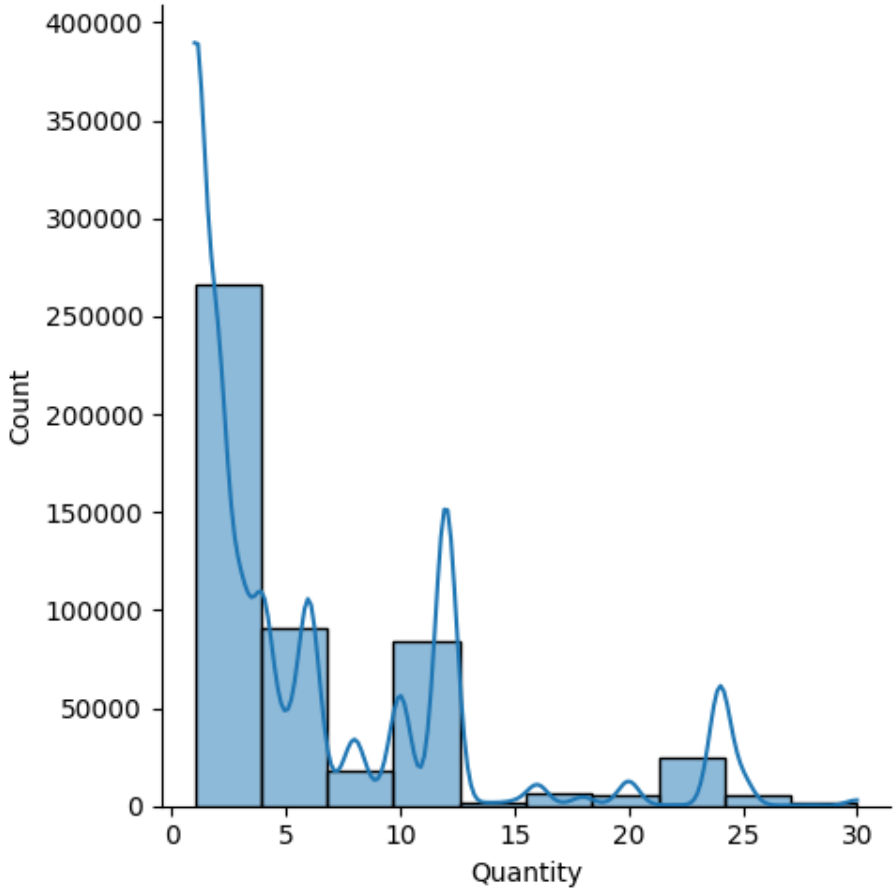
Podemos observar que tomando los datos del 98 % inferior nos quedamos con datos más balanceados y mejores para la predicción. El precio por unidad en ese cuantil es de 14.95 que lo tomaremos luego para filtrar.
Ahora miraremos cuantas ventas se han hecho en cada semana del año:



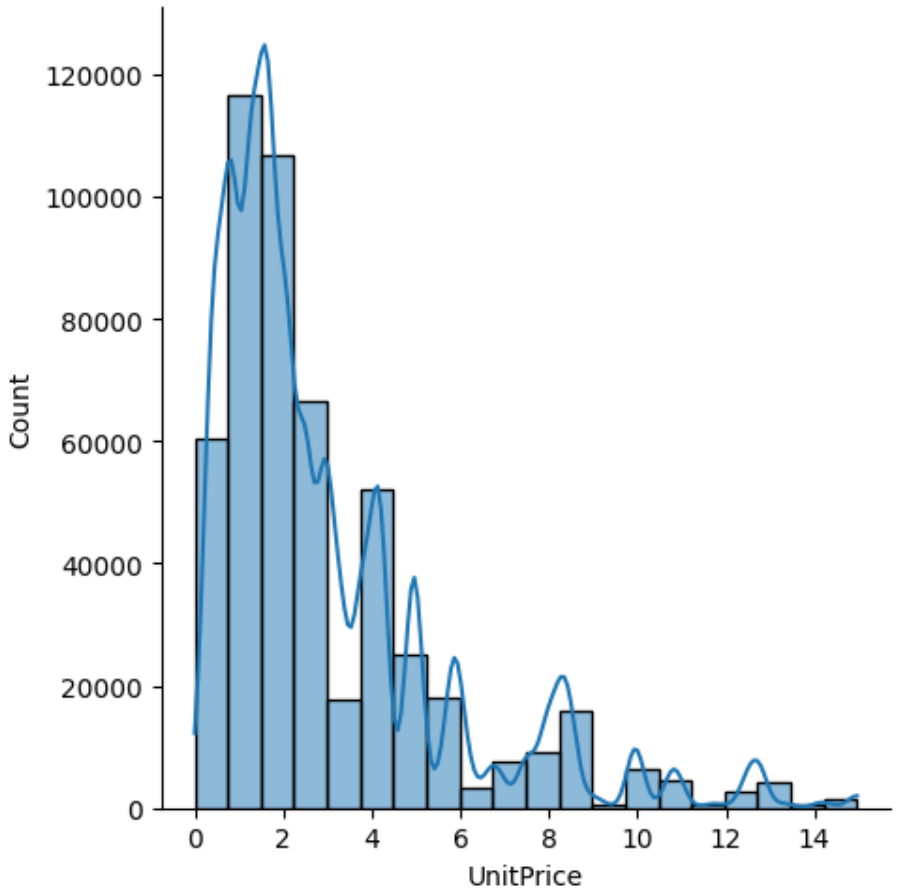
Hay un valor raro en una semana de enero que tenemos 0 ventas.
Al observar, vemos que es la semana de año nuevo por lo que en esa semana habrán cerrado por vacaciones y por eso no habrá ventas.

4. Limpieza de datos

Ahora realizaremos una selección de los datos con respecto a lo observado anteriormente.
Primero seleccionaremos los elementos cuyo precio por unidad sea inferior a 15.
Después podemos crear una columna para ver la cantidad de productos hay en cada transacción ya que en cada transacción hay más de un producto.
Para los datos de cantidad y de precio vamos a tomar rangos para que no sean solo números separados y predecir mejor.



Tomaremos los datos separándolos del 0 al 4 el primer grupo, el segundo desde el 4 al 8, el tercero hasta el 9, el cuarto grupo del 9 al 13, y el último del 13 en adelante.



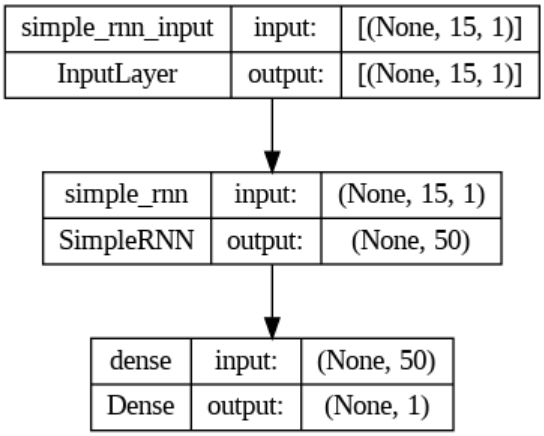
En el precio por unidad, tomaremos el primer grupo desde el 0 hasta el 1, luego el siguiente grupo 1 euro más hasta el 2, el tercer grupo hasta el 3, el cuarto entre el 3 y el 5 y el último hasta el final que hemos indicado antes que sea 15.
En la fecha también unificaremos y tomaremos valores por trimestres.
Ahora separamos los datos entre datos de Reino Unido y datos del resto del mundo y utilizaremos los datos de Reino Unido. Luego simplificaremos los datos para coger los más representativos como se ve en la siguiente tabla.

	Sales	QuantityInv	QuantityRange	PriceRange	DateRange
0	15.30	40	(4, 8]	(2, 3]	(9, 12]
1	20.34	40	(4, 8]	(3, 5]	(9, 12]
2	22.00	40	(4, 8]	(2, 3]	(9, 12]
3	20.34	40	(4, 8]	(3, 5]	(9, 12]
4	20.34	40	(4, 8]	(3, 5]	(9, 12]

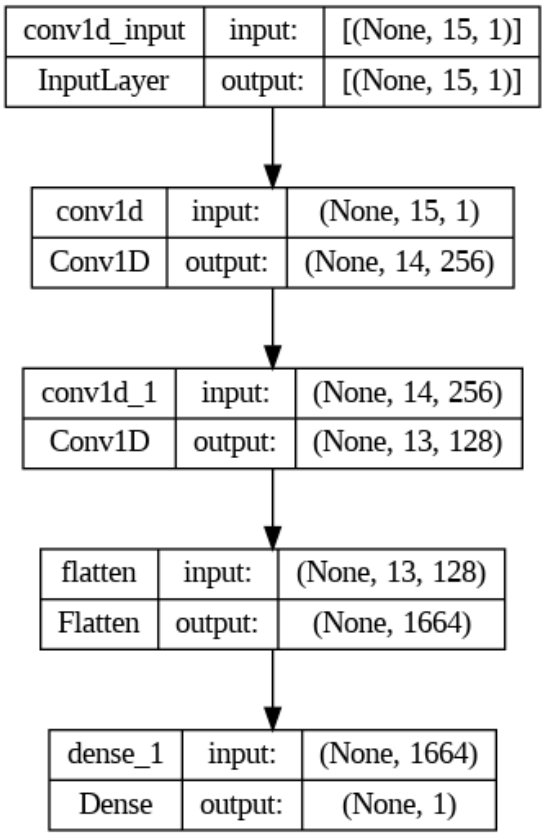
Por último hacemos dummies de todas las variables menos de Sales y QuantityInv (Cantidad de artículos por transacción).

5. Modelo

Primero escalaremos la cantidad de elemento por transacción.
Después separaremos los datos en train y test y ya estarán preparados para el modelo. Dado que se está trabajando con datos de series temporales y tenemos una variable objetivo, cogeremos modelos como las redes neuronales recurrentes (RNN) y las redes neuronales convolucionales 1D.
La primera red será RNN que tendrá la siguiente estructura y los resultados se pueden ver en el Jupyter Notebook.

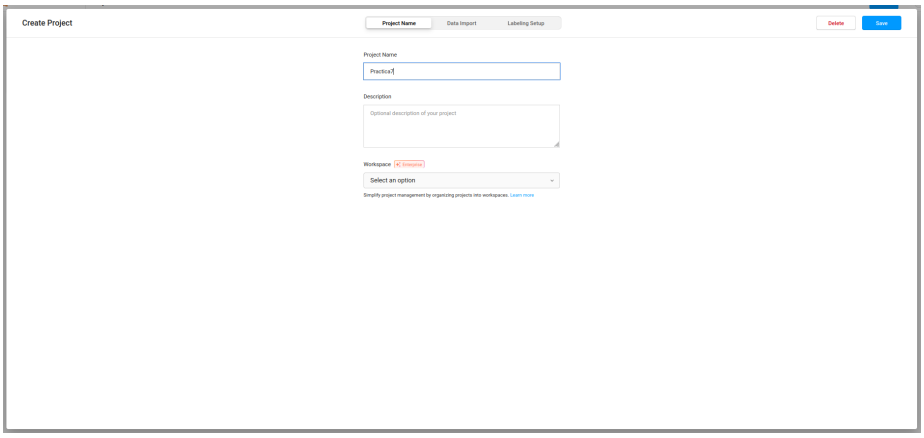


La segunda red será una red neuronal convolucional 1D que tendrá la siguiente estructura y los resultados se pueden ver en el Jupyter Notebook.

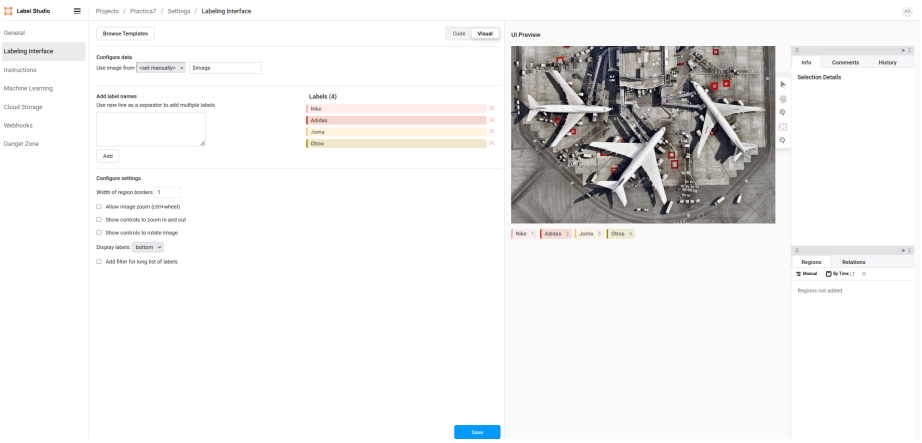


6. Label Studio

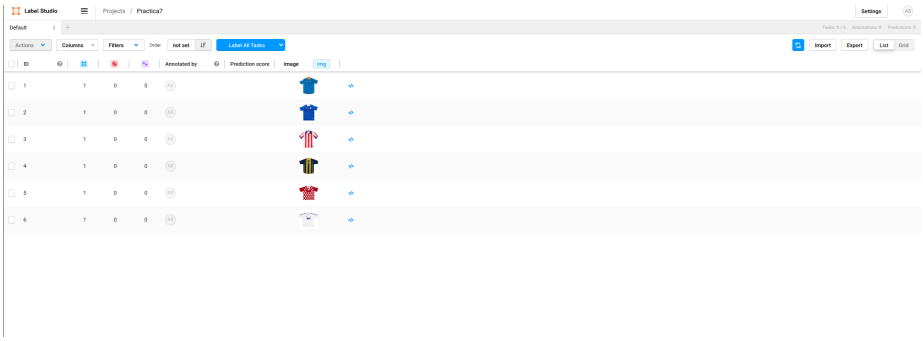
Label Studio es una herramienta de etiquetado de datos open-source. Nos permite etiquetar muchos tipos de datos como son audio, texto, imágenes, vídeos y series temporales para luego exportar a varios formatos de modelos.
Se puede utilizar para preparar los datos en bruto o mejorar los datos de entrenamiento existentes para obtener modelos de machine learning más precisos.
Para los datos de label Studio cogeremos un dataset con las urls de distintas camisetas y utilizando la librería PIL podremos ver y descargar esas imágenes.
Al iniciar Label Studio podemos crear un proyecto en el que le indicamos distintas características:



Podemos añadir las etiquetas que queramos y podemos elegir el método de selección de los elementos, en este caso hemos elegido uno en el que seleccionamos un rectángulo en el que le indicamos la etiqueta.



Una vez configurado todo e insertado las imágenes podemos ver la siguiente pestaña.



En cada imagen podemos poner distintas etiquetas, aunque en nuestro ejemplo no nos interesa por lo que vamos a etiquetar van a ser las marcas que crean las camisetas ya que en función de que empresa haga las camisetas tendrá un valor u otro.
En la siguiente imagen podemos observar como hay un rectángulo de color naranja alrededor del logo de Joma y corresponde con la tercera etiqueta "Joma".

7. Conclusión

Durante este proyecto hemos hecho un análisis exhaustivo de los datos obteniendo distintos parámetros sobre los que predecir los precios mejor además de usar Label Studio para hacer un preprocesamiento de datos de camisetas de fútbol filtrando por las marcas de las camisetas.