MACHINE LEARNING APLICADO A LA PREDICCIÓN DE PRECIOS DE ARTÍCULOS DE LUJO

El proyecto consiste en un asesor de reventa de artículos de moda. Existen muchas plataforma de reventa de ropa como Vestiaire que se centra en artículos de lujo o Vinted que en principio no apunta tanto hacia el lujo sino más por lo vintage o simplemente darle salida a ropa que no se usa, independientemente de la marca y eso hace que la media del precio de los productos (incluso los de lujo) sea más baja que la de Vestiaire, lo que es una oportunidad de negocio si sabes reconocer las prendas que pueden tener salida. Esto siempre ha sido un proceso mental mío durante todos estos años de dedicarme a la reventa y cuando empecé este curso me di cuenta de que con la inteligencia artificial se podría agilizar mucho este proceso.

LAS BASES DE DATOS DEL PROYECTO

Constamos de dos bases de datos:

- El dataset de **Vestiaire** consta de 900000 artículos que están subidos a la plataforma con sus correspondientes características (36 en total) que nos aportan información del artículo y del usuario. Lo hemos limitado a productos vendidos, lo cual es esencial para más adelante entrenar los modelos de regresión de precios, además hemos filtrado marcas y lo hemos limitado para agilizar procesos.

product_keywords	brand_name	product_color	product_condition	Tamaño	Suma de price_usd
Acne Studios Cotton - elasthane Jeans	Acne Studios	Anthracite	Very good condition	32 US	57.51
Acne Studios Cotton - elasthane Jeans	Acne Studios	Black	Never worn	28 US	166.75
Acne Studios Cotton - elasthane Jeans	Acne Studios	Blue	Very good condition	27 US	50.16
Acne Studios Cotton - elasthane Jeans	Acne Studios	Blue	Very good condition	28 US	25.56
Acne Studios Cotton - elasthane Jeans	Acne Studios	Blue	Very good condition	32 US	58.48
Acne Studios Cotton Dresses	Acne Studios	Blue	Very good condition	36 FR	51.12
Acne Studios Cotton Dresses	Acne Studios	Khaki	Very good condition		137.36
Acne Studios Cotton Dresses	Acne Studios	White	Very good condition		159.75
Acne Studios Cotton Jackets	Acne Studios	Beige	Never worn	34 FR	210.87
Acne Studios Cotton Jackets	Acne Studios	Black	Very good condition	42 IT	309.28
Acne Studios Cotton Jackets	Acne Studios	Blue	Never worn, with tag	50 FR	163.90
Acne Studios Cotton Jackets	Acne Studios	Multicolour	Very good condition	32 FR	131.63
Acne Studios Cotton Jeans	Acne Studios	Black	Good condition	28 US	219.63
Acne Studios Cotton Jeans	Acne Studios	Black	Never worn, with tag	31 US	115.97
Acne Studios Cotton Jeans	Acne Studios	Black	Very good condition	32 US	315.79
Acne Studios Cotton Jeans	Acne Studios	Black	Very good condition	38 FR	47.29
Acne Studios Cotton Jeans	Acne Studios	Blue	Never worn	38 US	206.93
Acne Studios Cotton Jeans	Acne Studios	Blue	Very good condition	25 US	166.14
Total					**772,018.56**

- Al scrapear el de **Vinted** aplicamos un filtro para scrapear como máximo productos de 150€ ya que a partir de ahí los productos tardan más en venderse y son inversiones más difíciles de rentabilizar. Además nos hemos encontrado algunos problemas porque tenía una estructura diferente y además la columna del precio estaba vacía. Como el precio sí que estaba incluido en la columna del título hemos tenido que hacer un código para extraer la cifra y agregarla a la columna. El código también traduce palabras del italiano y el francés y adapta las columnas al dataset de Vestiaire, extrayendo la información de la columna del título.

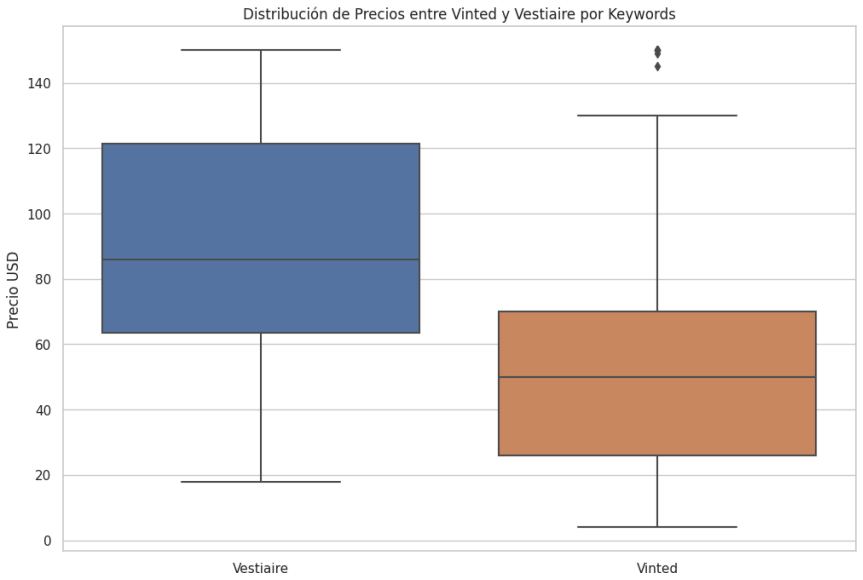
  El codigo consigue que pasemos de un dataset poco inteligible a otro mucho mas claro que incluye toda la informacion necesaria (https://github.com/alvarofn23/proyectofinal/blob/main/extraer-precios-vinted.ipynb)

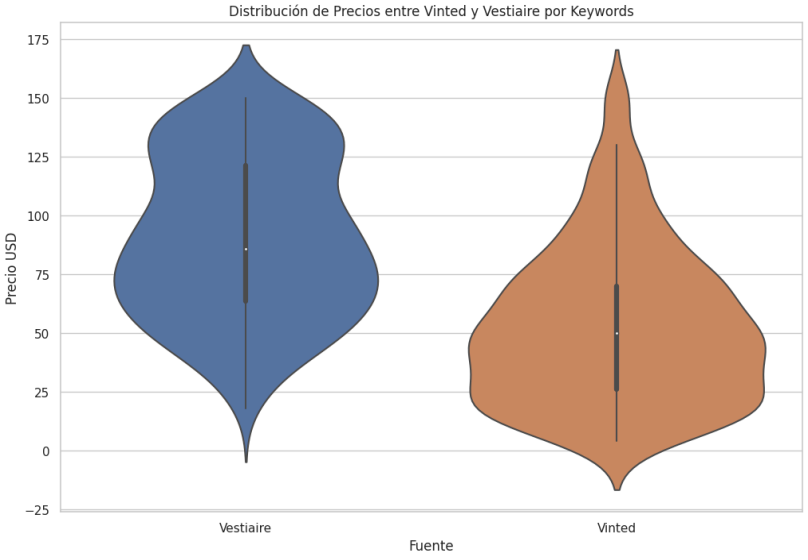
  En cuanto al tratamiento de valores nulos, habia pocos en los dos datasets pero se

hover_title	Precio_con_proteccion	URL	Marca	Tamaño	Estado	Color
2000's Dolce & Gabbana Hybrid Leather Jacket, marca: Dolce & Gabbana, estado: Muy bueno, tamaño: M, 120,00 €, 126,70 € Protección al comprador incluida		https://www.vinted.es/items/6111168953-2000s-dolce-gabbana-hybrid-leather-jacket	Dolce & Gabbana	M	Muy bueno	Gris
Acne Logo Stockholm 1996 Print Tshirt White, marca: Acne Studios, estado: Nuevo con etiquetas, tamaño: S, 130,00 €, 137,20 € Protección al comprador incluida		https://www.vinted.es/items/6083453909-acne-logo-stockholm-1996-print-tshirt-white?referrer=catalog	Acne Studios	S	Nuevo con etiquetas	Blanco, Crema
Acne Studios Black pleather jeans, marca: Acne Studios, estado: Muy bueno, tamaño: W31 | ES 41, 55,00 €, 58,45 € Protección al comprador incluida		https://www.vinted.es/items/6123028867-acne-studios-black-pleather-jeans?referrer=catalog	Acne Studios	W31 | ES 41	Muy bueno	Negro
Acne Studios Flannel Shirt Blue size 52, marca: Acne Studios, estado: Muy bueno, tamaño: L, 55,00 €, 58,45 € Protección al comprador incluida		https://www.vinted.es/items/6123059008-acne-studios-flannel-shirt-blue-size-52?referrer=catalog	Acne Studios	L	Muy bueno	Azul claro

product_keywords	brand_name	product_color	product_condition	Tamaño	Suma de price_usd
Acne Studios jacket	Acne Studios	Gris	Very good condition	S	145.00
Acne Studios jeans	Acne Studios	Marrón	Very good condition	W33 | ES 42	55.00
Acne Studios jeans	Acne Studios	Negro	Very good condition	W31 | ES 41	55.00
Acne Studios sweater	Acne Studios	Marrón	Very good condition	L	100.00
Acne Studios t-shirt	Acne Studios	Azul claro	Very good condition	L	55.00
Acne Studios t-shirt	Acne Studios	Beige	Very good condition	S	68.00
Acne Studios t-shirt	Acne Studios	Blanco	Very good condition	XL	100.00
Acne Studios t-shirt	Acne Studios	Blanco, Beige	Never worn	M	90.00
Acne Studios t-shirt	Acne Studios	Blanco, Crema	Never worn	S	130.00
Acne Studios t-shirt	Acne Studios	Gris	Very good condition	XL	65.00
Total					**19,604.76**

Para visualizar la diferencia de precio de los dos datasets y la **oportunidad de negocio** hemos limitado el de Vestiaire también a 150 y hemos hecho unos gráficos:





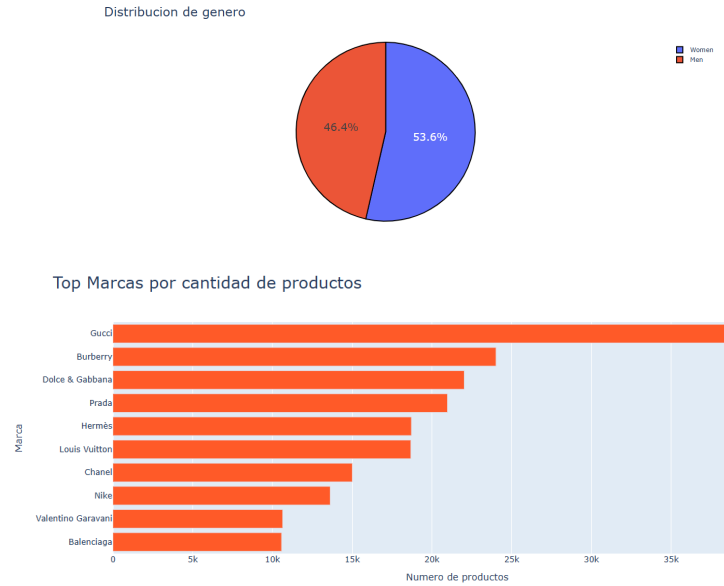
APARTADO LEGAL

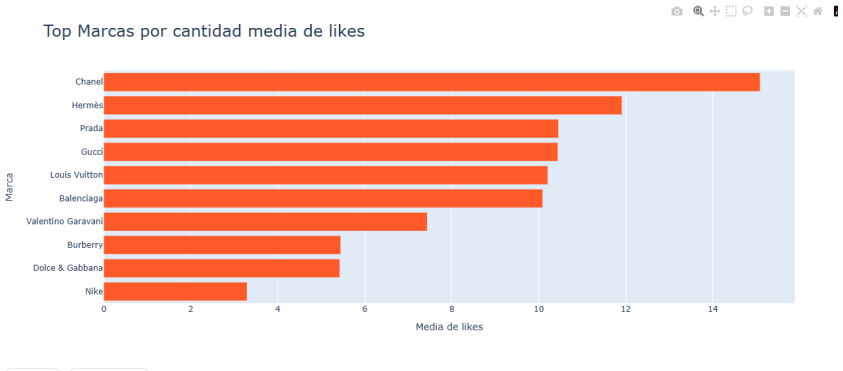
El scraping de datos de Vinted o Vestiaire para uso interno y analítico, centrándote en precios y atributos de producto, es legal siempre que:

- No incumplas gravemente sus ToS (limitando volumen y velocidad).
- No infrinjas el derecho sui generis (no copiar bases completas).
- No proceses datos personales sin justificarlo y anonimices todo lo posible.
- No redistribuyas imágenes o descripciones protegidas por copyright.

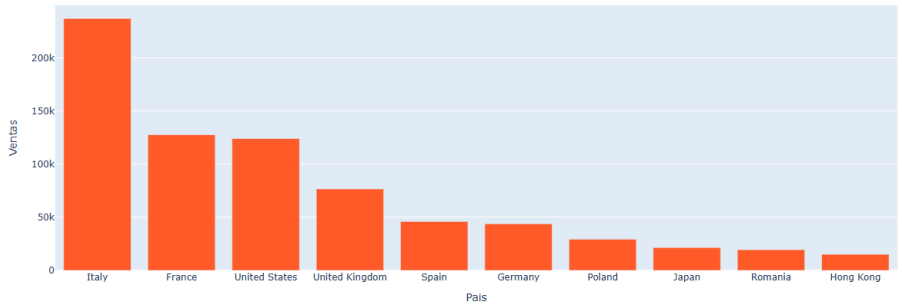
ANÁLISIS EXPLORATORIO DE LOS DATOS

Para la toma de decisiones en la inversión es esencial hacer un buen EDA que nos ayude a saber cuales son las características de los productos mas vendidos.

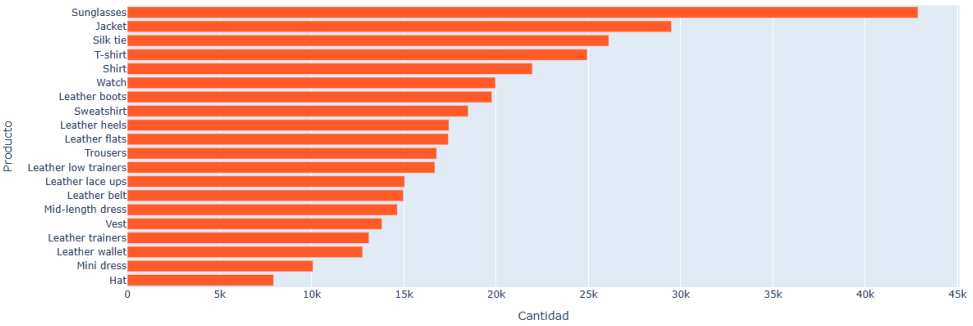




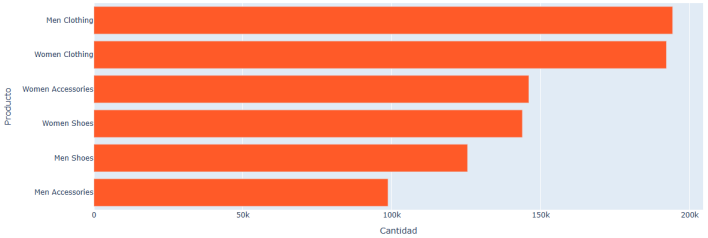
Top 10 Paises por cantidad de productos vendidos



Top 20 Productos

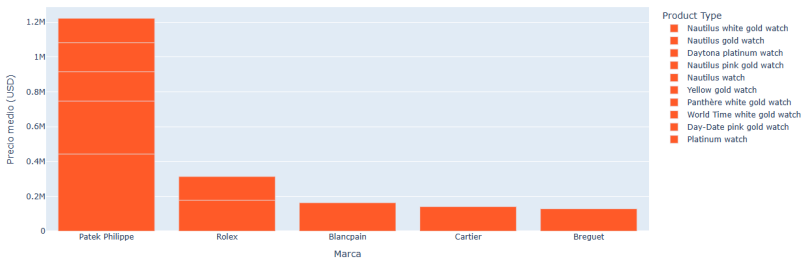


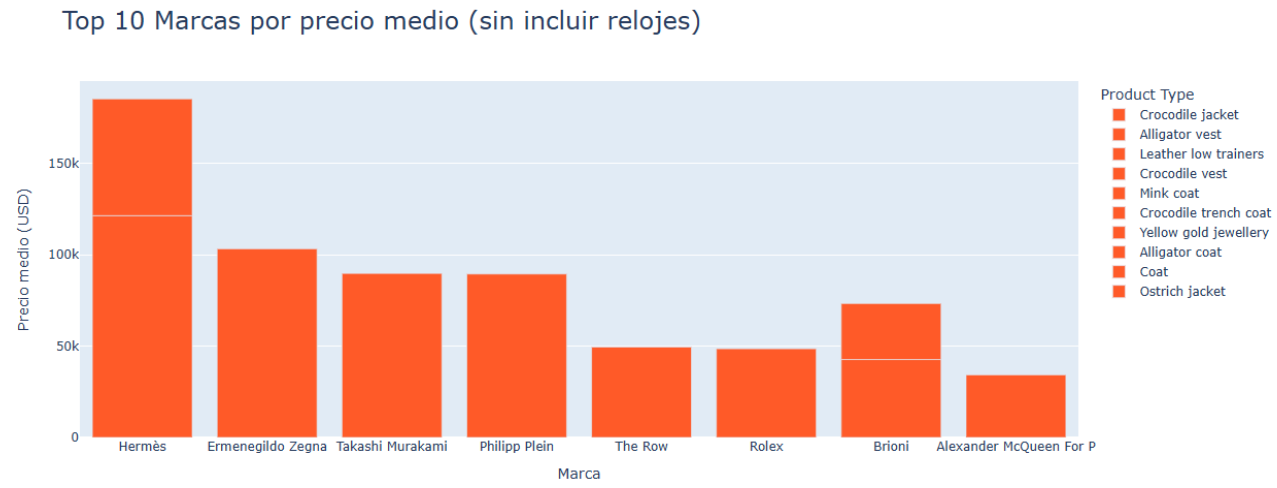
Distribucion de productos por categoria

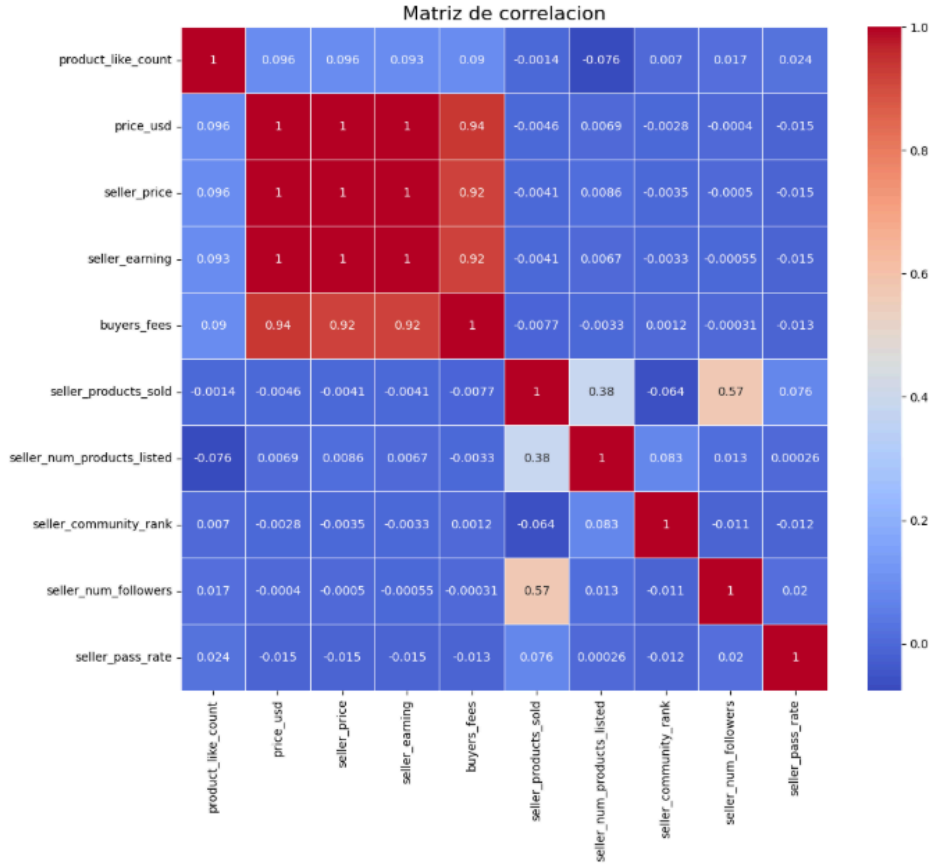


+ Code + Markdown

Top 10 Marcas por precio medio





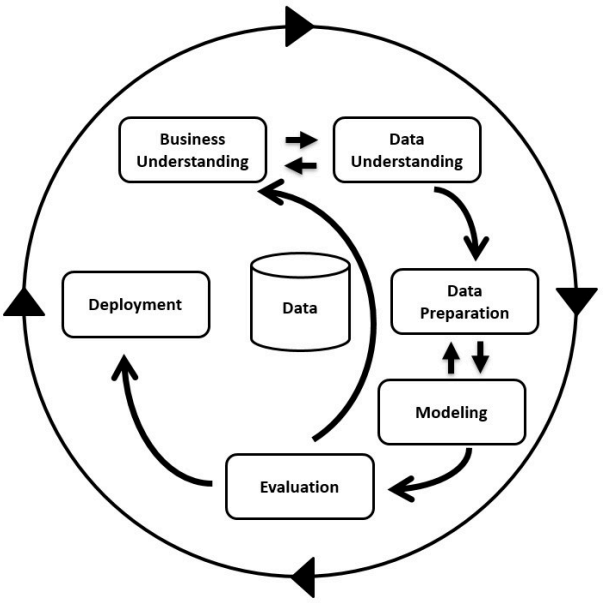


Con estas visualizaciones podemos saber en qué tipo de marcas, colores y productos conviene más invertir.

METODOLOGÍA Y MINERÍA DE DATOS

En este proyecto, estamos aplicando una metodología de minería de datos predictiva y seguimos una serie de pasos que forman parte de una metodología comúnmente conocida como CRISP-DM (Cross-Industry Standard Process for Data Mining).

El CRISP-DM (*Cross-Industry Standard Process for Data Mining*) es una metodología estándar y ampliamente aceptada para llevar a cabo proyectos de minería de datos. Fue desarrollada en los años 90 por un consorcio de empresas (incluyendo IBM) con el objetivo de crear un enfoque estructurado, aplicable a múltiples industrias.



¿Qué es exactamente?

- Fases del CRISP-DM

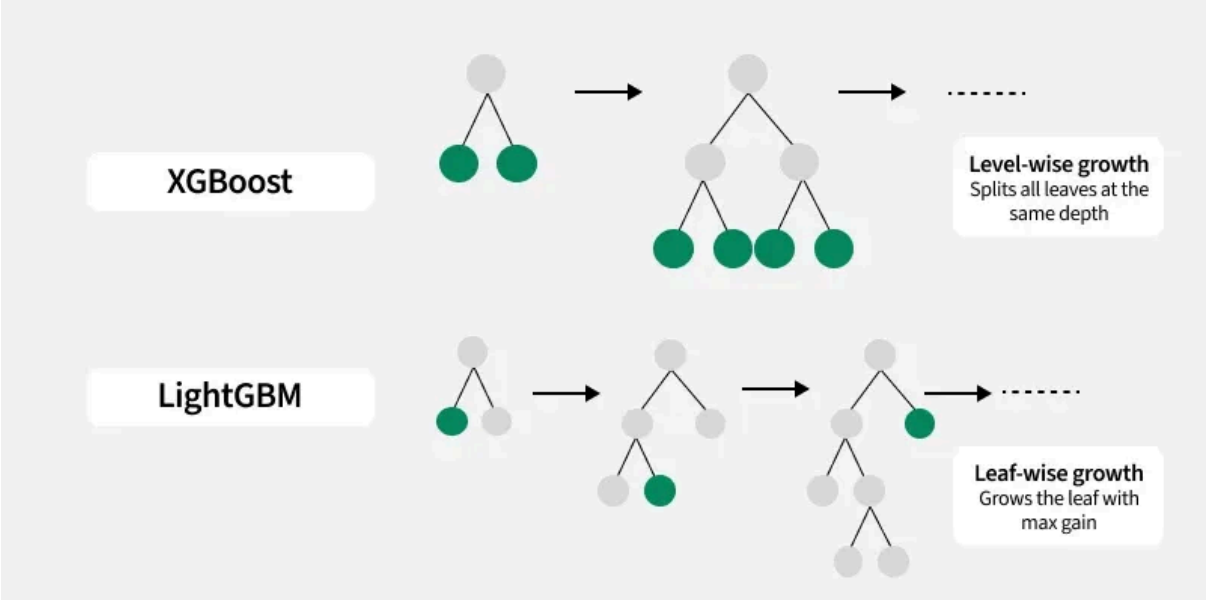
Fase	Descripción
1. Comprensión del negocio	Entender los objetivos del negocio y cómo la minería de datos puede ayudar a alcanzarlos.
2. Comprensión de los datos	Recopilar, explorar y describir los datos disponibles para entender su calidad y relevancia.
3. Preparación de los datos	Limpiar, transformar, seleccionar y estructurar los datos necesarios para el modelado.
4. Modelado	Aplicar técnicas de modelado (como regresión, árboles de decisión, etc.) y ajustar parámetros.
5. Evaluación	Verificar si los modelos cumplen los objetivos de negocio y funcionan correctamente.
6. Implementación	Integrar los resultados en el entorno empresarial (puede ser un informe, una app, un dashboard, etc.).

Predicción de precios

El primer paso para la predicción de los precios que alcanzarán los artículos de vinted en vestiaire es entrenar el modelo de regresión con los propios datos de vestiaire. Hemos filtrado el dataset por productos vendidos y con unas marcas especificas que serán las mismas que hay en el dataset de vinted. Para ello primero vamos a probar con LightGBM y luego con otros para comprobar.

LightGBM es un algoritmo de **machine learning** basado en árboles de decisión, optimizado para **alta velocidad y eficiencia**. Utiliza técnicas como **histogramas** y **crecimiento por hojas** (en lugar de por niveles) para construir árboles más profundos y precisos. Soporta grandes volúmenes de datos y alta dimensionalidad sin pérdida de rendimiento. Funciona muy bien en tareas de clasificación, regresión y ranking. Está diseñado para ser **más rápido y consumir menos memoria** que otros métodos de boosting como XGBoost.

LightGBM funciona construyendo múltiples árboles de decisión de forma secuencial para **minimizar el error de predicción**. En cada iteración, ajusta un nuevo árbol para corregir los errores del modelo anterior, usando el **gradiente de la función de pérdida** (por eso es "gradient boosting"). A diferencia de otros algoritmos, crece los árboles **por hojas** (leaf-wise), eligiendo la hoja con mayor ganancia para dividir, lo que mejora la precisión pero puede sobreajustar si no se controla. Además, usa técnicas como **binning de histogramas** para acelerar el entrenamiento y reducir memoria.



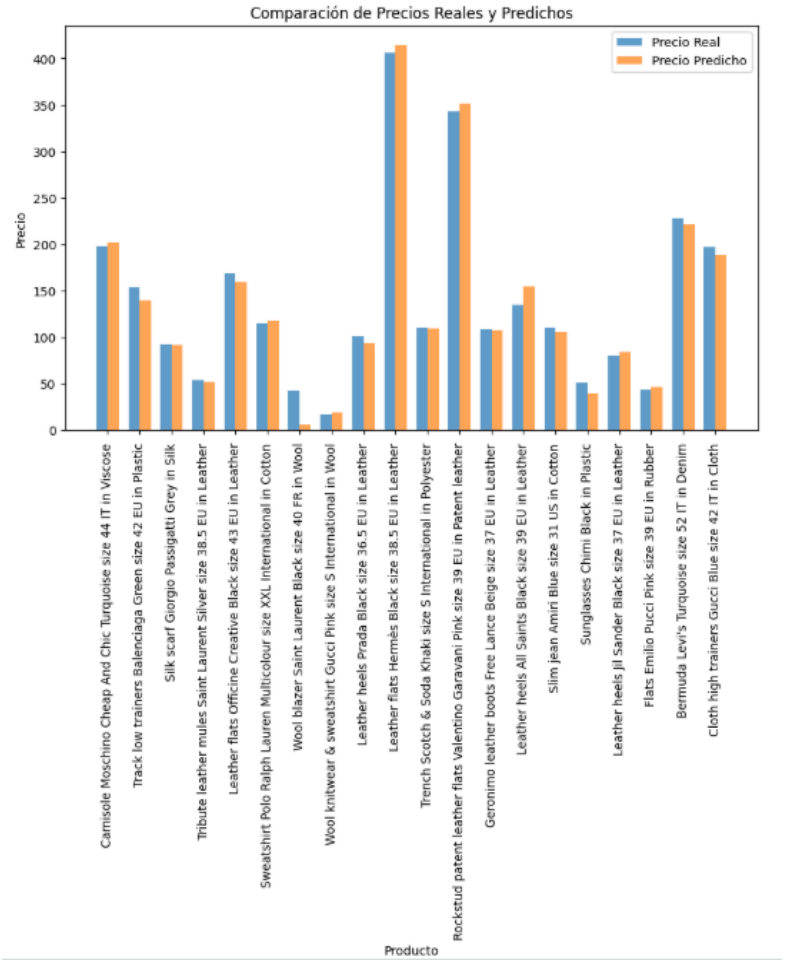
Parámetros clave que he usado

Parámetro	Función principal
`objective='rmse'`	Minimiza el error cuadrático medio.
`learning_rate=0.1`	Paso de actualización al añadir cada nuevo árbol.
`n_estimators=10000`	Número máximo de árboles (se detiene antes si converge).
`max_depth=5`	Profundidad máxima de cada árbol.
`num_leaves=62`	Número de hojas por árbol (complejidad).
`subsample=0.9`	Uso de una fracción de datos para cada árbol (bagging).
`colsample_bytree=0.5`	Fracción de features usadas por árbol (feature bagging).
`reg_alpha=0.1`	Penalización L1 en pesos de los árboles.
`reg_lambda=1.0`	Penalización L2 en pesos de los árboles.
`min_child_samples=10`	Mínimo número de observaciones por hoja.

Estos hyperparámetros buscan un equilibrio entre **sesgo** (bias) y **varianza** (overfitting).

Una vez explicado el modelo veamos los resultados:

	Nombre	Marca	Precio_real	Prediccion
6378	Camisole Moschino Cheap And Chic Turquoise siz...	Moschino Cheap And Chic	198.09	201.642451
658237	Track low trainers Balenciaga Green size 42 EU...	Balenciaga	153.93	139.868570
588467	Silk scarf Giorgio Passigatti Grey in Silk	Giorgio Passigatti	92.40	91.543355
777931	Tribute leather mules Saint Laurent Silver siz...	Saint Laurent	53.77	51.635607
709694	Leather flats Officine Creative Black size 43 ...	Officine Creative	168.82	159.432734
255210	Sweatshirt Polo Ralph Lauren Multicolour size ...	Polo Ralph Lauren	114.66	117.423695
104136	Wool blazer Saint Laurent Black size 40 FR in ...	Saint Laurent	42.17	5.221632
193258	Wool knitwear & sweatshirt Gucci Pink size S I...	Gucci	16.61	18.848472
849334	Leather heels Prada Black size 36.5 EU in Leat...	Prada	100.42	93.372535
861441	Leather flats Hermès Black size 38.5 EU in Lea...	Hermès	406.09	414.614543
312708	Trench Scotch & Soda Khaki size S Internationa...	Scotch & Soda	110.02	109.158049
842641	Rockstud patent leather flats Valentino Garava...	Valentino Garavani	342.70	351.506623
873847	Geronimo leather boots Free Lance Beige size 3...	Free Lance	108.63	107.713029
782085	Leather heels All Saints Black size 39 EU in L...	All Saints	135.16	154.833929
327391	Slim jean Amiri Blue size 31 US in Cotton	Amiri	110.47	105.929583
403877	Sunglasses Chimi Black in Plastic	Chimi	51.12	39.515579
898509	Leather heels Jil Sander Black size 37 EU in L...	Jil Sander	80.25	84.305423
791683	Flats Emilio Pucci Pink size 39 EU in Rubber	Emilio Pucci	43.44	45.699694
195571	Bermuda Levi's Turquoise size 52 IT in Denim	Levi's	228.34	221.199365
698175	Cloth high trainers Gucci Blue size 42 IT in C...	Gucci	197.07	188.096062

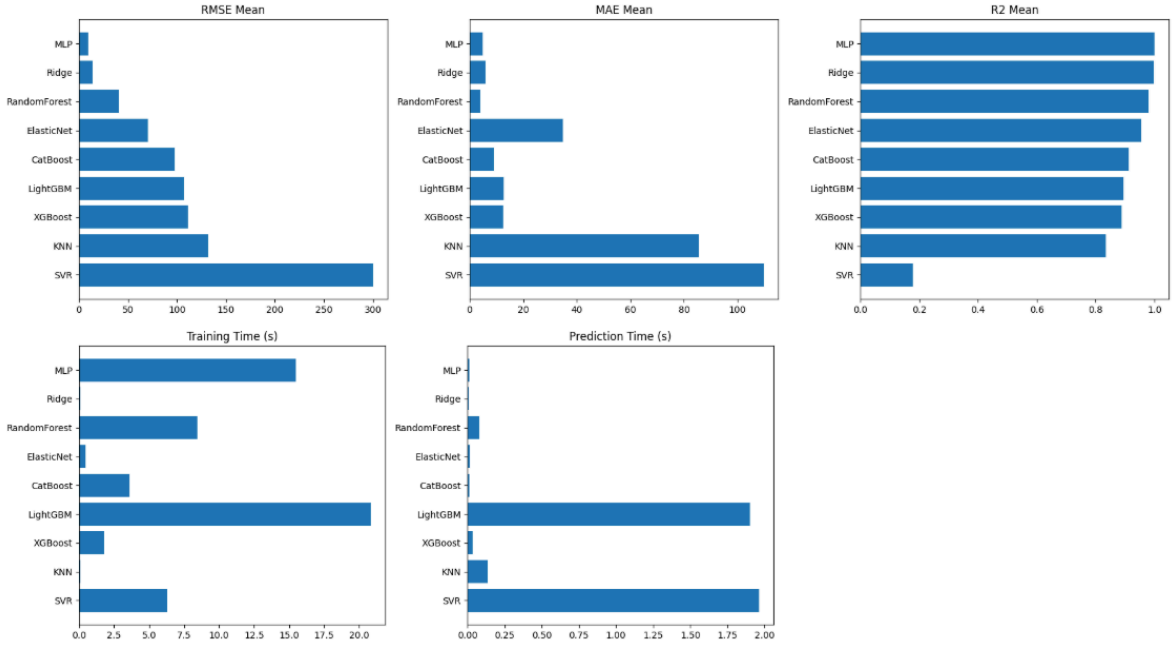


Como podemos apreciar, el modelo predice muy bien los precios de los artículos, pasemos ahora a la comparación de modelos para quedarnos con el mejor.

Otros modelos a considerar para comparar

Para validar que LightGBM es la mejor opción, conviene compararlo frente a otros enfoques:

Modelo	Ventajas	Cuándo usarlo
XGBoost	Otro GBDT muy optimizado, con regularización robusta.	Si quieres comparar otras implementaciones GBDT.
CatBoost	Maneja categorías nativamente, menos necesidad de codificar.	Cuando hay muchas variables categóricas.
Random Forest Regressor	Ensamble de árboles independientes, menos tuning fino.	Para tener un baseline rápido y estable.
Ridge / Lasso / ElasticNet	Modelos lineales penalizados, interpretables.	Si sospechas de relación lineal fuerte o quieres interpretabilidad.
Support Vector Regressor (SVR)	Robusto en datasets medianos, kernels no lineales.	Con pocas muestras y features, kernel RBF.
k-Nearest Neighbors	Modelo muy simple, sin entrenamiento pesado.	Para detectar patrones locales en el espacio.
Redes Neuronales (MLPRegressor)	Modela relaciones complejas no lineales.	Si dispones de mucha data y potencia de cómputo.
Gaussian Process Regressor	Predicción probabilística y cuantifica incertidumbre.	Cuando quieras intervalos de confianza en predicción.



	RMSE Mean	RMSE Std	MAE Mean	MAE Std	R2 Mean	R2 Std	Training Time (s)	Prediction Time (s)	Model
0	9.611277	0.569033	4.862961	0.324185	0.999079	0.000305	15.491604	0.013476	MLP
1	13.762767	0.999710	5.918284	0.105623	0.998145	0.000546	0.093246	0.010739	Ridge
2	41.133282	36.058924	4.048566	1.233146	0.980884	0.026495	8.445175	0.077299	RandomForest
3	70.474110	12.147419	34.772556	1.372108	0.954554	0.003888	0.469356	0.015769	ElasticNet
4	98.029694	28.091802	8.969338	1.417208	0.912716	0.025360	3.600682	0.015117	CatBoost
5	107.099238	33.922080	12.750134	1.443249	0.895284	0.037964	20.827288	1.901079	LightGBM
6	111.583307	34.105141	12.535145	1.771565	0.887538	0.036791	1.780115	0.034398	XGBoost
7	131.813844	14.800854	85.648622	1.884074	0.835616	0.032781	0.081542	0.134902	KNN
8	300.030455	53.015119	109.701136	3.980721	0.179625	0.045735	6.309944	1.962479	SVR

Para determinar cuál de los modelos es mejor, es importante tener en cuenta varias métricas de evaluación que se muestran en las gráficas y la tabla. Las métricas clave para un modelo de regresión suelen ser:

1. **RMSE (Root Mean Squared Error)**: Indica la magnitud promedio de los errores en las predicciones. Cuanto más bajo sea el valor de RMSE, mejor será el modelo en cuanto a precisión.

2. **MAE (Mean Absolute Error)**: Mide el error absoluto promedio de las predicciones. Al igual que el RMSE, un valor más bajo es mejor.

3. **R^2 (Coeficiente de determinación)**: Mide la calidad del ajuste del modelo, es decir, cuánto de la varianza en los datos se explica por el modelo. Un valor cercano a 1 indica un buen ajuste.

4. **Tiempo de entrenamiento**: Indica cuánto tiempo tarda el modelo en entrenarse. Aunque no afecta la precisión, un tiempo de entrenamiento más bajo es generalmente preferido si no compromete el rendimiento.

5. **Tiempo de predicción**: Similar al tiempo de entrenamiento, este valor indica cuánto tarda el modelo en realizar las predicciones, siendo importante en situaciones de producción en tiempo real.

Análisis de los resultados:

1. **RMSE y MAE**:
    - **SVR** (Support Vector Regression) tiene el mayor RMSE y MAE, lo que indica que tiene un rendimiento inferior en comparación con otros modelos.
    - **KNN** (K-Nearest Neighbors) también muestra un rendimiento deficiente en estas métricas, especialmente en RMSE y MAE altos.
    - **MLP** (Multi-layer Perceptron) es el mejor modelo en términos de RMSE y MAE, con los valores más bajos de ambas métricas, lo que significa que es el más preciso en la predicción.
2. **R^2**:
    - **MLP** es el modelo con el mayor R^2, lo que indica que es el que mejor explica la varianza en los datos y, por lo tanto, tiene el mejor ajuste entre los modelos evaluados.
    - Modelos como **RandomForest**, **ElasticNet**, y **Ridge** también tienen un buen desempeño, con valores altos de R^2, aunque no superan al MLP.
3. **Tiempo de entrenamiento**:
    - El **MLP** tiene el mayor tiempo de entrenamiento, lo que podría ser un inconveniente si el tiempo de cómputo es una preocupación.
    - Los modelos como **KNN** y **SVR** tienen los tiempos de entrenamiento más bajos, lo que los hace más adecuados para escenarios donde la rapidez de entrenamiento es crucial.
4. **Tiempo de predicción**:
    - **LightGBM** y **XGBoost** tienen los tiempos de predicción más rápidos, lo que puede ser beneficioso en aplicaciones donde las predicciones deben hacerse en tiempo real.
    - **MLP** y **SVR** tienen tiempos de predicción más largos, lo que podría ser un problema en aplicaciones sensibles al tiempo.

Conclusión:

- **Mejor modelo**: **MLP (Multi-layer Perceptron)** es el mejor modelo en términos de precisión (bajos RMSE y MAE) y ajuste (alto R^2). Sin embargo, tiene un tiempo de entrenamiento largo.

- **Modelos alternativos**:

    - Si el tiempo de entrenamiento es crucial, **KNN** y **SVR** ofrecen tiempos de entrenamiento más rápidos, aunque con un rendimiento inferior en términos de precisión.

    - **LightGBM** y **XGBoost** ofrecen un buen equilibrio entre precisión y tiempos de predicción rápidos, siendo buenos para aplicaciones en tiempo real.

En resumen, **MLP** es el modelo más preciso, pero si el tiempo de entrenamiento o de predicción es una preocupación, **LightGBM** o **XGBoost** podrían ser mejores opciones.

El **MLP (Multilayer Perceptron)** es un tipo de red neuronal compuesta por una **capa de entrada**, **una o más capas ocultas** y una **capa de salida**. Cada neurona combina entradas con pesos, aplica una función no lineal (como ReLU) y transmite la señal.El modelo aprende ajustando los pesos usando **retropropagación** y **descenso del gradiente**.Sirve para tareas de **clasificación y regresión**, capturando relaciones complejas entre variables.Es un modelo supervisado que requiere datos etiquetados para entrenar.

Una vez hemos dado con el mejor modelo, ahora solo nos queda usarlo para predecir los precios del dataset de Vinted y tendremos una aproximación del beneficio que le podemos sacar a cada uno. Como extra para la toma de decisión, hemos agregado otra predicción basada en la comparación de características usando la técnica de **similitud del coseno** entre representaciones vectoriales TF-IDF de texto.

- Se calcula una **matriz de similitud** donde cada fila representa un producto de Vinted y cada columna un producto de Vestiaire, usa **TF-IDF** sobre el `product_name` y `product_keywords` para representar los textos como vectores. luego mide la **similitud del coseno** entre estos vectores para saber qué productos son “parecidos” en términos de descripción textual.

```
similarities = cosine_similarity(X_vinted, X_train)
```

- Para cada producto de Vinted (`idx`), selecciona los **top 5 productos más similares** en Vestiaire, luego **calcula el precio medio** de esos productos y lo usa como estimación del precio para el producto de Vinted.

```
def estimate_price_from_similarity(idx, sims, vest_data, top_k=5):
    top_idxs = sims[idx].argsort()[::-1][:top_k]  # top K productos más similares
    return vest_data.iloc[top_idxs]['price_usd'].mean()  # promedio de sus precios
```

Es útil porque no requiere entrenamiento y se basa solo en texto, siendo ideal cuando los productos están bien descritos. Sirve como referencia simple e interpretable frente a modelos complejos como LightGBM o XGBoost.

product_keywords	product_type	product_condition	Suma de Precio Vinted	Suma de predicted_price_lightgbm	Suma de predicted_price_mlp	Suma de predicted_price_similarity
Acne Studios jacket	jacket	Very good condition	145.00	68.45	89.99	340.80
Acne Studios jeans	jeans	Very good condition	110.00	352.91	256.70	364.76
Acne Studios sweater	sweater	Very good condition	100.00	105.82	63.44	294.83
Acne Studios t-shirt	t-shirt	Never worn	310.00	341.24	326.85	704.13
Acne Studios t-shirt	t-shirt	Very good condition	288.00	593.45	657.05	626.98
Alexander McQueen t-shirt	t-shirt	Very good condition	35.00	204.23	270.24	301.98
Christian Dior t-shirt	t-shirt	Very good condition	40.00	277.63	596.51	334.17
Dior jeans	jeans	Very good condition	50.00	268.43	551.53	470.67
Dior sweater	sweater	Fair condition	35.00	241.26	473.92	426.48
Dior sweater	sweater	Never worn	85.00	247.51	500.88	502.41
Total			**19,604.76**	**52,158.67**	**64,640.15**	**60,670.26**

En esta tabla de PowerBi podemos ver el precio del producto en Vinted, y el precio estimado que podría alcanzar en Vestiaire. Aunque vimos que el MLP era el que mejores resultados dio, da unos precios demasiado optimistas para dejarnos llevar solo por su predicción.

CATEGORIZACIÓN

En el código se realizan **dos formas distintas de categorización** de inversión para los productos de Vinted, basadas en la relación entre el precio real y el precio estimado:

1. Categorización manual por regla lógica (condicional):

Se calcula el promedio de todas las predicciones de precio (`predicted_price_avg`) y luego se aplica una **función de decisión** que compara ese promedio con el precio real (`price_usd`):

```
def categorize_investment(price_usd, predicted_price_avg):
    if predicted_price_avg > 1.5 * price_usd:
        return 'High Investment'
    elif 1 <= predicted_price_avg <= 1.5 * price_usd:
        return 'Moderate Investment'
    else:
        return 'Low Investment'
```

Esto evalúa cuánto más (o menos) vale el producto según los modelos comparado con su precio real:

- **Alta inversión (High Investment)**: el producto parece estar **muy infravalorado** (es una buena oportunidad).
- **Inversión moderada (Moderate Investment)**: el producto está **ligeramente infravalorado**.
- **Baja inversión (Low Investment)**: el producto está **sobrevalorado**.

2. Categorización automática con KMeans (clustering no supervisado):

- Aquí se usan **todas las predicciones y el precio real** como características para aplicar `KMeans`, se estandariza con `StandardScaler()` para que todas las variables pesen igual, y luego agrupa los productos en **3 clústeres** automáticamente, según patrones en sus precios reales y estimados.

```
X = vinted_price_estimates[['price_usd', 'predicted_price_lightgbm', 'predicted_price_xgboost',
                            'predicted_price_catboost', 'predicted_price_mlp', 'predicted_price_si
```

```
kmeans = KMeans(n_clusters=3)
vinted_price_estimates['investment_cluster'] = kmeans.fit_predict(X_scaled)
```

Finalmente, se analizan los clústeres para asignarles etiquetas:

```
cluster_order = vinted_price_estimates.groupby('investment_cluster')['predicted_price_avg'].mean().sort_values(ascending=False)
label_map = {cluster: label for cluster, label in zip(cluster_order.index, ['High Investment', 'Moderate Investment', 'Low Investment'])}
```

Resultado:

- `investment_category`: categorización explícita por reglas.
- `investment_category_from_cluster`: categorización descubierta por agrupamiento automático.

product_keywords	Suma de price_usd	Suma de predicted_price_mlp	Suma de predicted_price_similarity	investment_category_from_cluster	investment_category
Acne Studios jacket	145.00	89.99	340.80	Low Investment	Moderate Investment
Acne Studios jeans	110.00	256.70	364.76	High Investment	High Investment
Acne Studios sweater	100.00	63.44	294.83	Low Investment	Moderate Investment
Acne Studios t-shirt	55.00	206.27	175.05	High Investment	High Investment
Acne Studios t-shirt	65.00	113.92	138.44	Low Investment	High Investment
Acne Studios t-shirt	410.00	480.23	842.57	Low Investment	Moderate Investment
Acne Studios t-shirt	68.00	183.48	175.05	Moderate Investment	High Investment
Alexander McQueen t-shirt	35.00	270.24	301.98	Moderate Investment	High Investment
Christian Dior t-shirt	40.00	596.51	334.17	Moderate Investment	High Investment
Dior jeans	50.00	551.53	470.67	Moderate Investment	High Investment
Dior sweater	120.00	974.80	928.89	Moderate Investment	High Investment
Dior t-shirt	450.00	2,289.89	1,324.72	Low Investment	High Investment
Dior t-shirt	657.90	5,319.73	3,545.74	Moderate Investment	High Investment
Dolce & Gabbana jacket	95.00	374.33	754.32	High Investment	High Investment
Dolce & Gabbana jacket	230.00	1,340.23	1,031.11	Low Investment	High Investment
Dolce & Gabbana jacket	240.00	385.11	492.73	Low Investment	Moderate Investment
Dolce & Gabbana jacket	320.00	1,514.74	1,568.02	Moderate Investment	High Investment
Dolce & Gabbana jeans	327.00	1,462.05	2,100.26	High Investment	High Investment
Dolce & Gabbana jeans	70.00	188.85	269.28	Low Investment	High Investment
Dolce & Gabbana jeans	100.00	252.02	265.44	Low Investment	Moderate Investment
Dolce & Gabbana jeans	249.24	912.44	742.66	Moderate Investment	High Investment
Dolce & Gabbana sweater	50.00	58.84	413.23	High Investment	High Investment
Dolce & Gabbana t-shirt	622.00	4,134.34	3,369.41	High Investment	High Investment
Dolce & Gabbana t-shirt	345.00	1,004.23	937.33	Low Investment	High Investment
Dolce & Gabbana t-shirt	220.00	533.76	427.47	Low Investment	Moderate Investment
Dolce & Gabbana t-shirt	301.00	1,926.70	1,809.33	Moderate Investment	High Investment
Dsquared2 jacket	99.00	28.26	145.07	Low Investment	Moderate Investment
Dsquared2 jeans	1,400.79	3,656.47	3,812.77	High Investment	High Investment
Dsquared2 jeans	220.00	198.60	384.12	High Investment	Moderate Investment
Dsquared2 jeans	210.00	119.34	262.29	Low Investment	High Investment
Dsquared2 jeans	1,968.60	2,031.82	1,954.90	Low Investment	Moderate Investment
Total	**19,604.76**	**64,640.15**	**60,670.26**		

CONCLUSIÓN

Este proyecto demuestra cómo el machine learning puede aplicarse eficazmente al mercado de reventa de artículos de lujo, automatizando la estimación del valor de productos mediante modelos entrenados con datos reales de Vestiaire. A partir del scraping y procesamiento del contenido de Vinted —incluyendo limpieza, extracción de precios desde el título y normalización multilingüe— se construyó un dataset comparable, lo que permitió entrenar y aplicar modelos como LightGBM, XGBoost y MLP.

El modelo MLP fue el más preciso en términos de error y ajuste, mientras que LightGBM y XGBoost ofrecieron un rendimiento más equilibrado y tiempos de predicción rápidos. Además, se implementó un sistema de estimación por similitud semántica (TF-IDF + coseno) entre productos de ambas plataformas, útil como referencia adicional sin necesidad de entrenamiento.

Finalmente, se categorizaron los productos según su potencial de inversión utilizando dos enfoques: uno lógico basado en reglas y otro automático mediante clustering (KMeans), ofreciendo así una herramienta de apoyo a la decisión robusta, práctica y aplicable al análisis real de oportunidades de reventa.