

A Data Science Project: Wine System Recommender

Laura Álvarez Mon

14 ed. Master in Data Science

Kschool Madrid

Junio 2019

Índice

1. Objetivo	2
2. Metodología	2
3. Business Understanding	2
4. Data Acquisition	2
5. Data Understanding	5
6. Data Preparation	5
7. Modeling	7
8. Visualization.....	8

1. OBJETIVO

El objetivo del presente proyecto es desarrollar un sistema de recomendación de vinos.

2. METODOLOGÍA

Para el desarrollo del proyecto se ha empleado la metodología CRISP-DM (Cross Industry Standard Process for Data Mining).

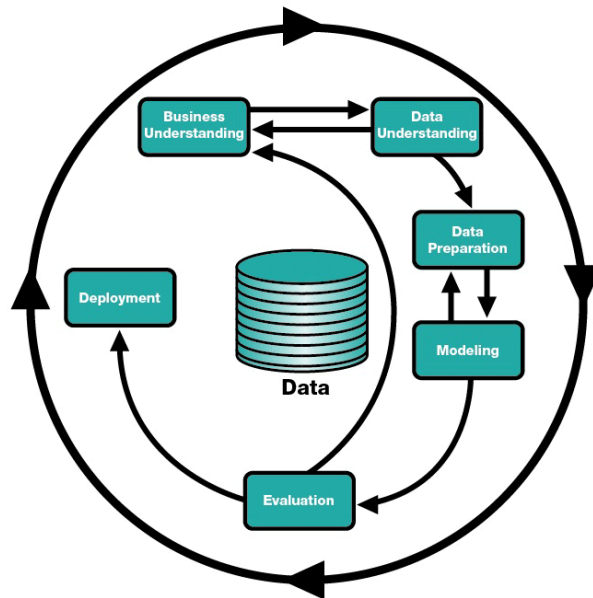


Imagen 1. Metodología.

En los siguientes epígrafes se describen las diferentes etapas del proceso.

3. BUSINESS UNDERSTANDING

Se plantea la necesidad de una herramienta para realizar una recomendación de vinos a los usuarios conociendo uno o varios vinos de su gusto.

Por ello, se define el objetivo de desarrollar un sistema de recomendación de vinos basado en el contenido, es decir, en las propiedades de los vinos.

4. DATA ACQUISITION

Dado que se trata de un proyecto académico, no se dispone de datos de negocio. De modo que se ha realizado una etapa intermedia (Data Acquisition) para la obtención de los datos mediante web scraping.

La técnica web scraping consiste en extraer información de sitios web mediante programas de software.



Imagen 2. Web scraping.

Para realizar el proceso de web scraping se ha utilizado BeautifulSoup, una biblioteca de Python para analizar documentos HTML. Esta biblioteca crea un árbol con todos los elementos del documento que puede ser utilizado para extraer información.

La página web de la que se han extraído los datos es: <https://labodega.consum.es/catalogo>

Se ha analizado el código fuente de la página para obtener para cada uno de los vinos:

- URL de la imagen.
- URL de la ficha.

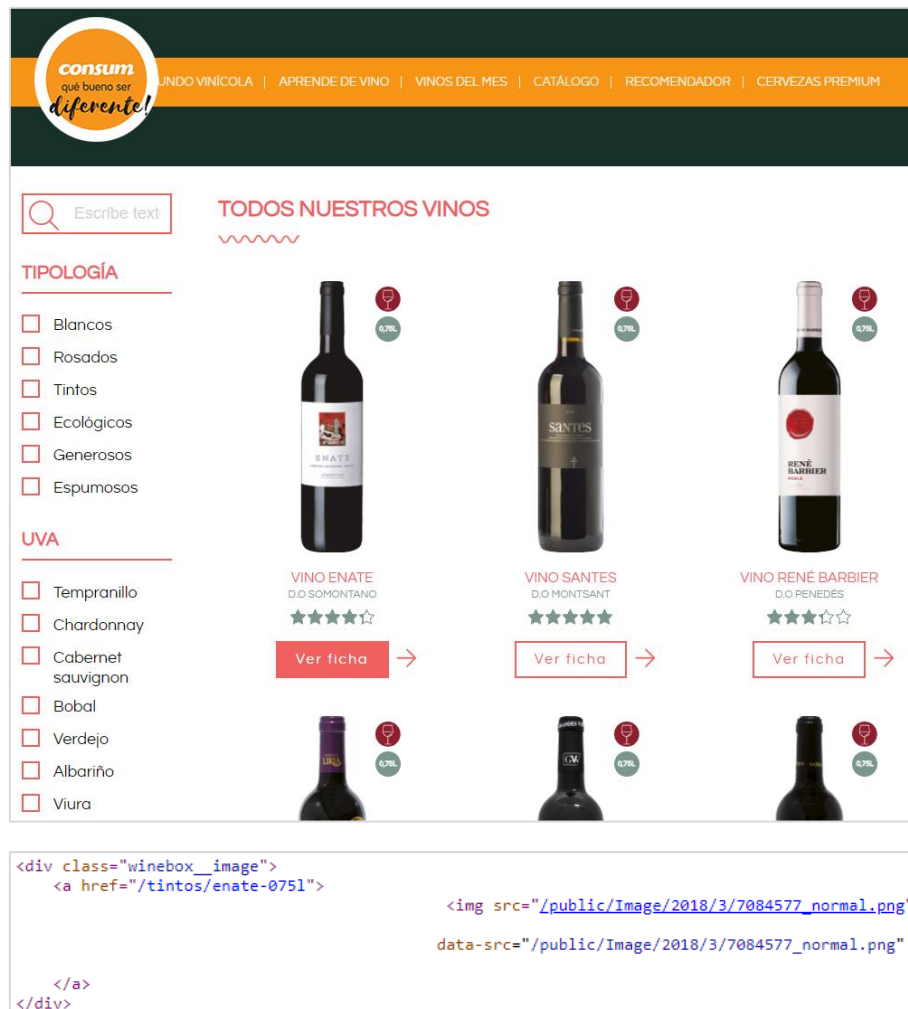


Imagen 3. Origen de los datos. Catálogo.

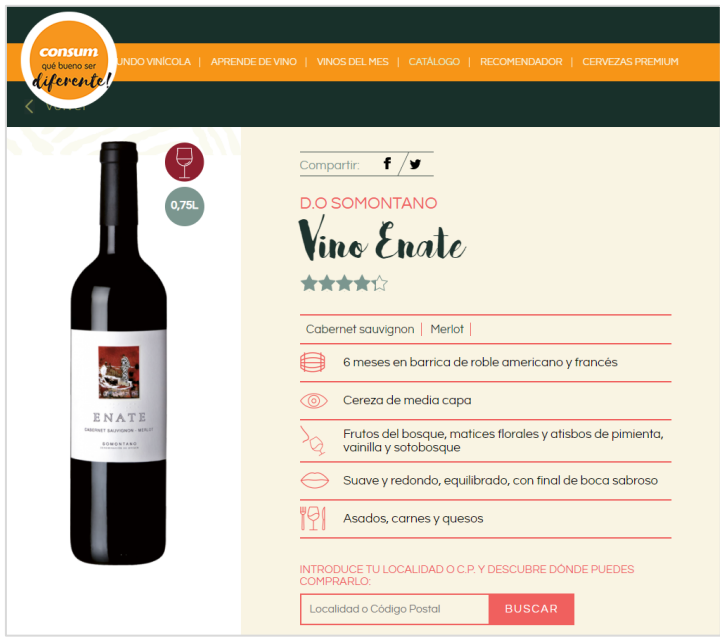
Esta información se ha almacenado en el fichero 'Links.csv' y en el DataFrame **links**.

	Wine_Id	product_link	image_link
0	0	https://labodega.consum.es/montesierra	https://labodega.consum.es/public/Image/2017/1...
1	1	https://labodega.consum.es/tintos/faustinovii-...	https://labodega.consum.es/public/Image/2018/3...
2	2	https://labodega.consum.es/tintos/claretdetard...	https://labodega.consum.es/public/Image/2018/3...

Imagen 4. Dataframe Links.

Asimismo, aunque para visualización se ha utilizado el image_link, se han descargado las imágenes por si fueran necesarias en un futuro.

A continuación, se ha accedido a la URL de cada vino, y se han obtenido sus propiedades:



tipo (type)

denominación de origen (DO)

Nombre (Name)

uva (grape)

crianza (ageing)

vista (sight)

olor (smell)

gusto (taste)

maridaje (pairing)

```

<dl class="wine-tags" itemprop="description">
  <div class="wine-tag-row">
    <dt><i class="ico ico-crianza">Crianza</i></dt> <dd>6 meses en barrica de roble americano y franc&eacute;s</dd>
  </div>
  <div class="wine-tag-row">
    <dt><i class="ico ico-visual">Visual</i></dt> <dd>Cereza de media capa</dd>
  </div>
  <div class="wine-tag-row">
    <dt><i class="ico ico-nariz">Nariz</i></dt> <dd>Frutos del bosque, matices florales y atisbos de pimienta, vainilla y sotobosque</dd>
  </div>
  <div class="wine-tag-row">
    <dt><i class="ico ico-boca">Boca</i></dt> <dd>Suave y redondo, equilibrado, con final de boca sabroso</dd>
  </div>
  <div class="wine-tag-row">
    <dt><i class="ico ico-maridaje">Maridaje</i></dt> <dd>Asados, carnes y quesos</dd>
  </div>
</dl>
  
```

Imagen 5. Origen de los datos. Producto.

Esta información se ha almacenado en el fichero 'Wines.csv' y en el DataFrame **wines**.

Wine_Id	Name	Type	DO	Size	Grape	Ageing	Sight	Smell	Taste	Pairing	
0	0	Montesierra	Tintos	D.O Somontano	Cabernet sauvignon, Merlot, Tempranillo	Sin crianza	Cereza picota	Aroma frescos a grosella negra, cerezas y menta	Carnoso, jugoso y con mucho sabor	Pastas, legumbres, tapas, quesos y embutidos	
1	1	Vino Faustino VII	Tintos	D.O Rioja	0,187L	Tempranillo	6 meses en barrica de roble americano	Rojo picota brillante con evolución a granate	Agradable, frutos rojos y toque dulce por el p...	Fresco y elegante, notas frutales y barrica	Carnes, patatas a la brasa, pescados y marisco...
2	2	Vino Claret de Tardor	Tintos	D.O Empordà	0,75L	Cabernet sauvignon, Garnacha	Sin crianza	Rojo guinda brillante con reflejos violáceos	Fruta roja y aromas de crianza	Ligero, agradable y equilibrado con acidez ref...	Arroces, entrantes, pastas, jamón, embutidos

Imagen 6. Dataframe Wines.

5. DATA UNDERSTANDING

Dataframe: links

- Wine_Id
Identificador del vino.
- producto_link
<https://labodega.consum.es/montesierra>
- image_link
https://labodega.consum.es/public/Image/2017/10/7298821.png_normal.jpg

Dataframe: wines

Almacena las propiedades de los 370 vinos y su Wine_Id identificador.

6. DATA PREPARATION

Se he llevado a cabo una limpieza previa de los datos, rellenando los nulos y eliminando los posibles duplicados (0).

No obstante, la tarea principal de preparación y tratamiento de los datos ha consistido en transformar las propiedades de los vinos en variables categóricas numéricas, hasta obtener un Dataframe del tipo:

		Wines			
Properties					

Las transformaciones realizadas se pueden clasificar en tres tipos:

- Transformación tipo 1

Esta transformación se ha aplicado a los grupos Type, DO y Grape, y ha consistido simplemente en quedarse con los valores únicos.

En el caso de Type y DO ha sido suficiente con agrupar respectivamente por cada columna.

Y para la columna Grape se ha hecho un Split y se han eliminado los duplicados.

- Transformación tipo 2

Esta transformación, aplicada sobre los grupos Sight, Smell y Taste, ha requerido un pequeño procesamiento del lenguaje.

El problema era que la descripción de estas columnas contenía palabras de una misma familia como, por ejemplo: fruta, fruto, frutos, afrutado, frutal, ...

Se ha tratado de realizar un embedding de palabras usando Word2Vec y similares, pero posteriormente se ha descartado pues surgían los siguientes inconvenientes:

- El texto de entrenamiento era insuficiente.
- Los modelos entrenados, generalmente, contenían vocabulario en inglés.

De modo que, dado que el vocabulario que se necesitaba procesar no era demasiado extenso, se ha procedido a construir el vocabulario de palabras mediante la función *CountVectorizer* (from *sklearn.feature_extraction.text* import *CountVectorizer*), se ha medido la similaridad entre todas las palabras mediante la función *fuzz.partial_ratio* (from *fuzzywuzzy* import *fuzz*, from *fuzzywuzzy* import *process*), y se ha considerado que las palabras pertenecen a una misma familia cuando su similaridad es > 80-90. Todo este proceso se ha almacenado en la función *FamilyOfWords*.

```
[[['cereza'], ['picota'], ['rojo'], ['brillante'], ['evolución'], ['granates', 'granate']]
[['frescos', 'fresco'], ['negras', 'negra']]
[['tostado'], ['redondos', 'redondo'], ['expresión'], ['tánica']]]
```

Imagen 7. Resultado de aplicar FamilyOfWords.

▪ Transformación tipo 3

Por último, el grupo Ageing se ha procesado en Microsoft Excel y se ha importado como csv.

Una vez realizadas estas transformaciones, se ha creado y cargado el Dataframe 'features'.

Éste está compuesto por las columnas: Group, Feature, Family_Id y Feature_Id.

Para entender mejor qué información contiene cada una de ellas, se muestra un ejemplo en la imagen inferior.

Dentro del Group = Smell, existen diferentes Features, cada uno con un diferente Feature_Id.

Sin embargo, tras haber empleado la función FamilyOfWords, se observa que algunos Features comparten el mismo Family_Id.

	Group	Feature	Family_Id	Feature_Id
220	Smell	maduras	220	220
221	Smell	madura	220	221
222	Smell	notas	222	222
223	Smell	balsámicas	223	223
224	Smell	balsámicos	223	224
225	Smell	ciruelas	225	225
226	Smell	ciruela	225	226
227	Smell	flor	227	227
228	Smell	floral	227	228
229	Smell	flores	227	229
230	Smell	florales	227	230

Imagen 8. Dataframe: features.

Se han creado una serie de diccionarios para agilizar el código.

A partir del campo Feature del Dataframe **features**, se ha hecho una búsqueda en el Dataframe inicial **wines**, para identificar coincidencias como 1 y 0 en caso contrario.

De este modo se ha cargado el DataFrame **wines_features**, que tiene por columnas todos los vinos, y por filas todos los features.

	0	1	2	3	4	5	6	7	8	9	...	360	361	362	363	364	365	366	367	368	369
0	0	0	0	0	0	0	0	0	0	0	...	1	0	0	1	0	0	0	1	0	0
1	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	1	0
2	0	0	0	0	0	0	0	0	0	0	...	0	1	1	0	0	1	0	0	0	0

Imagen 9. DataFrame: wines_features.

Finalmente, se han agrupado los features por familias, resultando el DataFrame **wines_families**.

	0	1	2	3	4	5	6	7	8	9	...	360	361	362	363	364	365	366	367	368	369
Family_Id																					
0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	1.0	0.0	0.0	1.0	0.0	0.0	0	1	0	0
1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	0.0	0	0	1	0
2	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	1.0	1.0	0.0	0.0	1.0	0	0	0	0

Imagen 10. DataFrame: wines_families.

El diagrama inferior resume el proceso llevado a cabo:

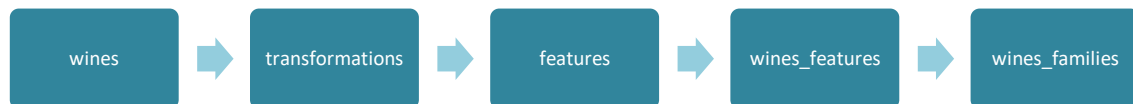


Imagen 11. Proceso Data Preparation

7. MODELING

Una vez finalizada la fase 'Data Preparation', se ha iniciado la fase 'Modeling'.

Basándose en el sistema de recomendación Collaborative filtering visto en clase, se ha realizado un sistema de recomendación basado en las propiedades del producto.

En este modelo, las propiedades de los vinos se usan para encontrar otros vinos similares, de manera análoga a como en el modelo visto en clase los rating's de usuarios se empleaban para encontrar películas similares.

El proceso seguido ha sido:

- Se ha calculado la matriz de co-ocurrencia aplicando álgebra lineal.

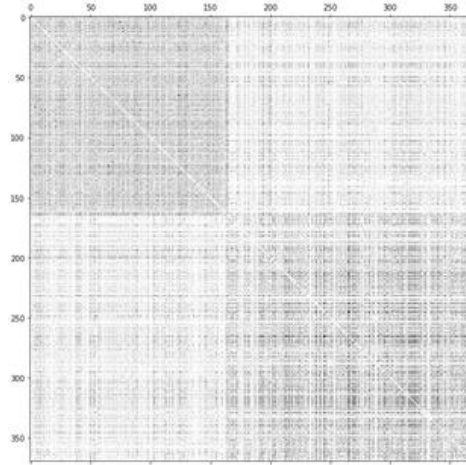


Imagen 12. Co-occurrence matrix.

- Se ha definido una función de similitud, que determina dado un vino, cuales son los más similares.

```
co_occurrence_similarity(150,coocc,5)

array([[ 5, 20],
       [98, 18],
       [38, 16],
       [72, 16],
       [92, 15]])
```

Imagen 13. Función de similitud.

- Se ha definido una función de recomendación, que determina dados como input uno o varios vinos, los mas similares de cada uno de ellos, eliminando el propio input.

```
co_occurrence_recommendation([150,5], coocc, 5)

array([ 98,  51, 131, 112,  88])
```

Imagen 14. Función de recomendación.

8. VISUALIZATION

Por último, para poder interactuar con este sistema de recomendación, se ha desarrollado un fichero '.py' en Jupyter Lab y se ha empleado la herramienta Dash.

Para ejecutar este fichero, se debe ejecutar el siguiente comando en la terminal:

```
Python -m WineRecSys_Visualization
```

El usuario deberá seleccionar uno o los vinos que desee de la lista desplegable, donde además podrá buscar por nombre introduciendo el texto de búsqueda en la barra.

Entonces, se mostrarán las imágenes de los vinos seleccionados en la parte izquierda de la ventana, y los vinos que se recomiendan en la parte derecha. Si además desea consultar propiedades específicas de alguno de los vinos, éstas se muestran tabuladas más abajo.

A continuación, se presentan algunas imágenes para mostrar este Dash de visualización del modelo.

Wine System Recommender

Please, select the wines you like:

... well, so we recommend this for you:

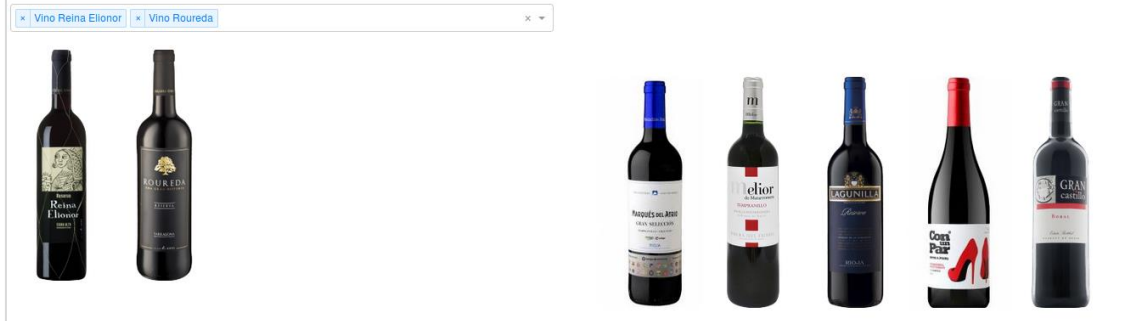


Imagen 15. Visualización.

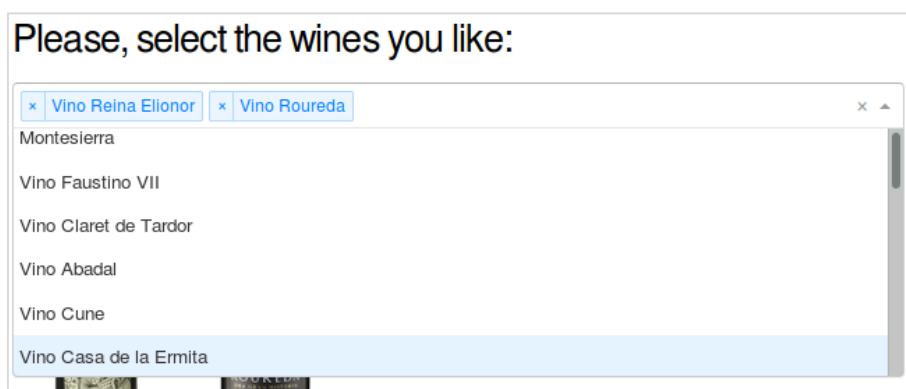


Imagen 16. Desplegable para selección de vinos.

See more details...

The wines you like:

NAME	TYPE	DO	GRAPE	AGEING	SIGHT	SMELL	TASTE	PAIRING
Vino Reina Elionor	Tintos	D.O Terra Alta	Mazuelo, Tempranillo, Garnacha	12 meses en barrica de roble	Rojo rubi intenso	Frutas rojas maduras, confituras y notas balsámicas	Armonioso, intenso, equilibrado con postgusto persistente y agradable	Carnes, quesos
Vino Roureda	Tintos	D.O Tarragona	Mazuelo, Tempranillo, Garnacha	12 meses en barrica de roble	Rojo rubi	Frutas maduras y notas especiadas	Suave, equilibrado y potente	Carnes rojas, caza, guisos, quesos

The wines we recommend for you:

NAME	TYPE	DO	GRAPE	AGEING	SIGHT	SMELL	TASTE	PAIRING
Vino Melior	Tintos	D.O Ribera del Duero	Tempranillo	6 meses en barrica	Cardenal oscuro con ribete morado	Frutas rojas maduras, regaliz, café y vainilla	Fresco, intenso y voluminoso	Carnes blancas, asados, legumbres y quesos de oveja
Vino Gran Castilla	Tintos	D.O Utiel-Requena	Bobal	Sin crianza	Rojo rubi intenso con reflejos violáceos	Frutas rojas maduras	Fresco, equilibrado y con buen paso por boca	Carnes a la plancha, pescados en salsa, arroces, pastas y tapas
Tinto selección Marqués del Tinto	Tintos	D.O Rioja	Tempranillo, Graciano	6 meses en barrica de roble americano y uno, dos meses en botella	Rojo rubi brillante de intensidad media	Aromas a frutas rojas maduras con elegantes toques de regaliz y	Equilibrado en boca con un largo y elegante postgusto final	Ideal para todas clases de carnes, pastas, tapas y

Imagen 17. Tabla con detalles de los vinos.