

Dataset: Proximitat de productes venuts a Bonpreu-Esclat

Joan Morral Ventura i Nicolás González Soler

November 5, 2021

Amb la voluntat de reduir la petjada de carboni i potenciar el producte de proximitat, s'ha realitzat un dataset que aconsegueix les característiques dels articles del supermercat BonPreu-Esclat i en consulta la seva localització geogràfica. Tenint les dades de longitud i latitud dels productes, resulta senzill calcular-ne el quilometratge respecte la nostra població. A partir d'aquest dataset, podem prendre consciència de quins són els productes més llunyans de la nostra llista de la compra o que el sistema ens ofereixi el més pròxim i de menor preu.

La idea que volem transmetre és la necessitat de posar el quilometratge en els tiquets de compra, especialment, els de format virtual, per tal de conscienciar la població en quan a sostenibilitat. El dataset té per objectiu contribuir en els objectius de desenvolupament sostenible de l'agenda 2030 de les Nacions Unides.

1 Context

El dataset conté el conjunt de productes del supermercat i tots els seus camps d'informació per tal que els propis usuaris puguin realitzar compres orientades a les dades. El posicionament que cerquen els autors de la pràctica, no és la de beneficiar exclusivament el client, sinó de fomentar el *win-win* amb el supermercat, per tal de beneficiar el territori.

S'ha cregut convenient incloure tots els camps per poder respondre a un major nombre de preguntes a través de aprenentatge automàtic. No obstant, degut a la seva extensió, la pràctica s'ha focalitzat en un dels seus prismes: la ubicació del producte en la seva zona geogràfica.

S'ha escollit la pàgina web de BonPreu-Esclat per tres motius.

- **Voluntat d'informar el client.** BonPreu-Esclat és un supermercat que ofereix una plataforma online amb molta informació sobre els seus productes. Per tant,

a través de tècniques de *webscraping* podem capturar del preu a la composició de tots els articles.

- **Nivell tècnic accessible.** La pàgina web té una estructura de navegació simple degut a l'ús reiterat de les mateixes plantilles, no s'han detectat *antibots* i no disposa de botons que requerissin l'ús de la llibreria Selenium, tal i com passa a Veritas, entre d'altres. La complexitat rau en l'alt volum de pàgines que cal visitar per a obtenir la informació de tots els seus articles, així com l'alteració dels camps en funció de la seva categoria.
- **Interès personal.** Com a clients del supermercat, em pensat en com podríem fer conscienciar a la població sobre la petjada de carboni dels productes adquirits.

2 Títol: Proximitat de productes venuts a Bonpreu-Esclat

El títol del dataset inclou la data de captura de les dades, el nom del supermercat i la terminació productes. Atès que els preus dels productes d'un supermercat van variant, cal documentar-ne l'extracció. El nom del supermercat també s'ha extret per a reduir el nombre de camps, atès que la seva iteració no aporta nova informació. Emprant l'extensió *.CSV ja es dona a entendre que la solució és tabular.

3 Descripció del dataset

El dataset presenta la informació dels productes disponible a la plataforma online de BonPreu-Esclat i n'adjunta la geoposició del nom de l'operador, de forma que permet posicionar-lo en un mapa. A més, amb llibreries python com Geopy, es permet calcular la distància entre el productor i el comprador.

Finalment, també és un bon dataset per poder treballar amb dades d'un supermercat real, ja que s'inclouen els camps de categoria, nom de producte, format de l'envàs, preu, preu per volum, informació del producte, marca, direcció, latitud, longitud, ingredients, dades nutricionals i instruccions.

4 Representació gràfica

La pàgina web de BonPreu-Esclat té un apartat de tenda online. El funcionament del nostre agent es basa en seguir la estructura de la pàgina web a través de categories i múltiples subcategories, capturant-ne les URL. Aquestes URL són executades de forma continuada a posteriori amb la voluntat de capturar tota la informació disponible sobre el producte. En la següent imatge es pot veure el funcionament de forma gràfica.

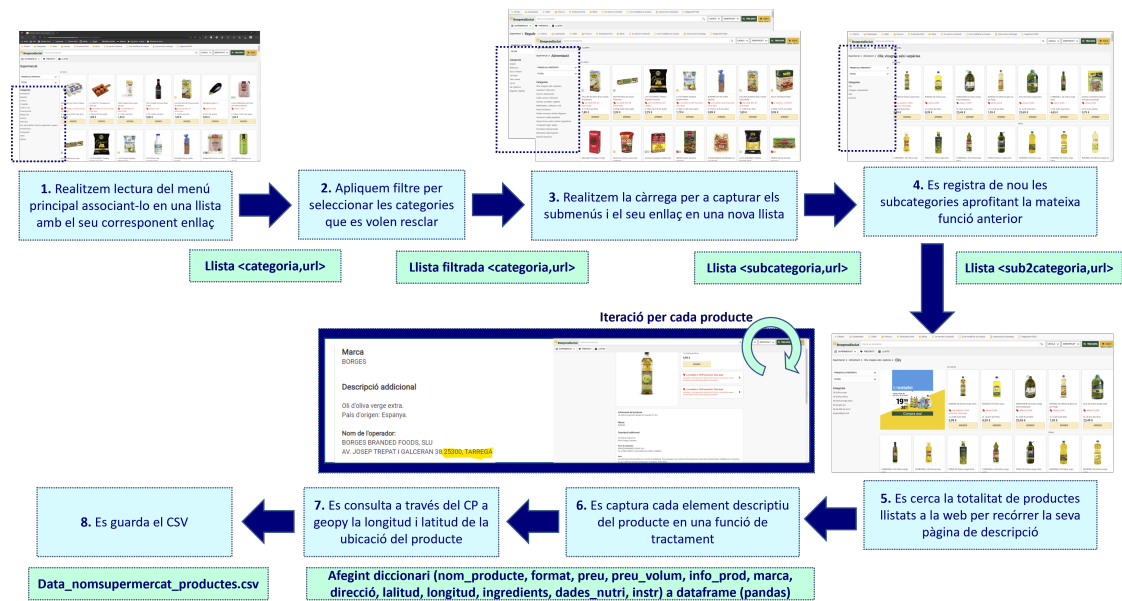


Figure 1: Procediment del funcionament del webscraper

5 Contingut

La taula següent presenta els camps del dataset.

Camp	Descripció
data	Camp que presenta la data en format YYYYMMDD, permet comparar preus entre dies.
subcategoria	Llista la categoria a la que pertany el producte, permeten relacionar productes similars.
nom_producte	Nom descriptiu intern de BonPreu que identifica el producte.
format	Format de l'envàs amb el que es presenta el producte
preu	Preu actual del producte
preu_volum	Preu per unitat de volum amb el que es presenta el producte.
info_prod	Informació general del producte
marca	Empresa responsable del producte.
direccio	Direcció del producte.
latitud	Distància angular mesurada sobre un meridià, entre una localització terrestre i l'Equador.
longitud	Distància angular sobre un meridià, entre una localització terrestre i Greenwich.
ingredients	Components del producte separats per una coma (només aliments).
dades_nutri	Informació nutricional del producte (només aliments).
instr	Instruccions dels productes

Table 1: Definició dels camps dels que es compona el dataset.

Els nom del supermercat s'ha considerat que era millor incorporar-los al nom del fitxer per reduir el pes del fitxer amb informació repetida. La data no s'ha considerat treure-la

per a la harmonia amb el concepte de versions de Zenodo.

6 Agraïments

La font d'informació d'on s'han recollit les dades és la pàgina web de <https://www.compraonline.bonpreuesclat.cat/products?source=navigation> i la seva plataforma en línia. Per aconseguir les dades s'han emprat tècniques de *Web Scraping* emprant el llenguatge Python i les llibreries Request i BeautifulSoup4. Per aconseguir els posicionament del producte (longituds i latituds), així com el càlcul en quilòmetres s'ha emprat la llibreria GeoPy.

7 Inspiració

Cada vegada que anem a comprar, no únicament estem escollint productes, sinó que els votem. Com a consumidors tenim la força de que els productes i les seves empreses productores persisteixin en el temps. Per tant, l'acció de comprar un producte suposa un impacte en la sostenibilitat socio-econòmica i en el medi ambient. Informar al consumidor de la distància del centre de producció pot despertar la consciència de reduir les cadenes de subministrament i beneficiar la comunitat propera.

Aquesta pràctica intenta treballar el punt 12 dels ODS, els 17 objectius de desenvolupament sostenible marcats per l'agenda 2030 de les Nacions Unides [1].

Aquest punt parla de **garantir patrons de consum i producció sostenibles**, i afecta especialment en dos dels punts:

- 12.1 Implementar el marc de consum i producció sostenibles en 10 anys
- 12.6 Animar les empreses a adoptar pràctiques sostenibles i informes de sostenibilitat

Si en la llista de la compra s'indiqués la distància a la que es troba el centre productor de cada producte, es permetria entendre quins d'aquests recorren major distància. És a dir, ens permetria donar-nos compte de quins articles podríem canviar per alternatius més pròxims. També ens permetria veure si hi ha algun producte sense productor pròxim en la nostra zona. Aquest podria ser un bon sistema per identificar possibilitats de negoci.

Tot i aquest enfocament, la base de dades de producte s'ha estructurat per poder donar resposta a moltes més preguntes:

- Seguiment de preus. Llançant el programa de forma cíclica es podria avaluar quina evolució de preu tenen els diversos productes. Pregunta: Quin és el millor dia del mes per anar a comprar? Per exemple, a través d'una rutina Cron en un sistema Linux corrent en una SBC com la Raspberry.



Figure 2: Captura dels 17 objectius de l'agenda 2030

- Validar quins productes són els que ofereixen major oferta o nombre de productes
- Detectar quins productes no tenen marca blanca a través de validar subcategories i marques.
- Validar quin increment suposen els productes sense gluten davant de productes amb gluten. Quin percentatge de la compra suposen aquests articles especials?
- Comprovar quins aliments són els que disposen de major nombre de grasses saturades a través del camp de valors nutricionals. Quins són els aliments amb menors grasses saturades?

8 Llicència

L'elecció de la llicència pel dataset s'ha basat en la consideració per part de l'equip dels següents punts:

- Manifestar que es tracta del resultat d'un treball acadèmic, del qual cal preservar-ne l'esperit
- Relatar l'autoria, en el marc institucional de la UOC
- Posar el dataset a disposició de la comunitat per a usos no comercials
- Promoure el respecte i seguiment dels principis esmentats en totes aquelles obres derivades

En conseqüència, s'ha optat per acollir-se a la llicència **CC BY-NC-SA 4.0** de Creative Commons [2]

Aquesta opció compleix amb els principis anteriors: atribució (BY), no comercial (NC) i compartició en idèntiques condicions (SA).

Cal remarcar que l'alternativa *Database released under Open Database License* [3], *individual contents under Database Contents License* [4] s'ha descartat perquè els drets contemplats inclouen l'ús comercial, aspecte que nosaltres volem limitar.

9 Codi

El codi font d'extracció i el dataset creat a tall d'exemple són accessibles a través de GitHub. Per a poder accedir al contingut pot visitar el següent enllaç: <https://github.com/joanmorrall/producteProximitatScraper>

El programa emprat per a la fase d'investigació a sigut jupyter notebook, ja que permet iterar amb dades descarregades. Per a realitzar la codificació del programa final en format *.py s'ha emprat Pycharm, ja que es requeria un fitxer format *.py i

10 Dataset

El dataset generat s'ha emmagatzemat al repositori públic Zenodo i es troba accessible de manera pública sense restriccions.

A efectes informatius d'aquest treball confirmem que el nom i la descripció del dataset informats al repositori coincideixen amb els especificats anteriorment en aquest document. Així mateix, també confirmem que el fitxer CSV associat al dataset es denomina `productesProximitat_bonpreu.csv`.

Seguint els requisits del present treball acadèmic, el DOI del dataset publicat a Zenodo es troba indicat al fitxer `Readme.md` del repositori Github.

11 Contribucions

La taula de contribucions és la següent:

Contribucions	Signatura
Investigació prèvia	JMV, NGS
Redacció de les respostes	JMV, NGS
Desenvolupament del codi	JMV, NGS

Table 2: Contribucions en cada apartat.

References

- [1] UN. <https://www.un.org/sustainabledevelopment/es/sustainable-development-goals/>.
- [2] Wikipedia. https://es.wikipedia.org/wiki/Licencias_Creative_Commons.
- [3] Open Data Commons. <https://opendatacommons.org/licenses/odbl/1.0/>.
- [4] Open Data Commons. <https://opendatacommons.org/licenses/dbcl/1.0/>.