TAREA FINAL SEMANA 4 (parte1)

Título: LA BATALLA DE VECINDARIOS 1 EN LIMA – PERÚ

Contenido

1. Introducción y descripción del caso.

En el Perú el consumo del pan per cápita se incrementa año tras año. En el 2018 según el Instituto Nacional de Estadística – INEI, cada peruano consumía en promedio 28 kilogramos de pan al año. En el 2021 este valor se incrementó, llegando a 30 kilogramos por año y se prevé que hacia finales del presente año podría inclusive incrementarse en un kilogramo más por persona. (América Economía, 2022).

No es de extrañar este crecimiento ya que el peruano consume pan no solo en el desayuno o la merienda final del día, sino que lo consume en diversos tipos de sándwich llamados en el Perú “sanguches” con diversos tipos de carnes y acompañamientos y también el pan forma parte de los ingredientes de muchos platos de la afamada cocina peruana.

El pan no solo es fabricado en el Perú con harina de trigo, que es el pan más común que existe, sino que se aprovecha la variedad de granos andinos y también se preparan utilizando harina de diferentes tubérculos (Aspan,2022).

Es de interés de un empresario nacional el abrir una cadena de panaderías en donde se prepare y hornee el pan de manera artesanal y de manera moderna, es por esta razón que se requiere conocer la mejor ubicación para instalar los locales los que incluirán además una cafetería.

Existen alrededor de 20,000 panaderías a lo largo del país y donde casi el 50% de ellas se ubica en Lima la capital del Perú y sus 43 distritos. Es por esta razón que la ubicación de los nuevos locales tendrá que estar en aquellos distritos que tengan la mayor población con nivel socioeconómico AB.

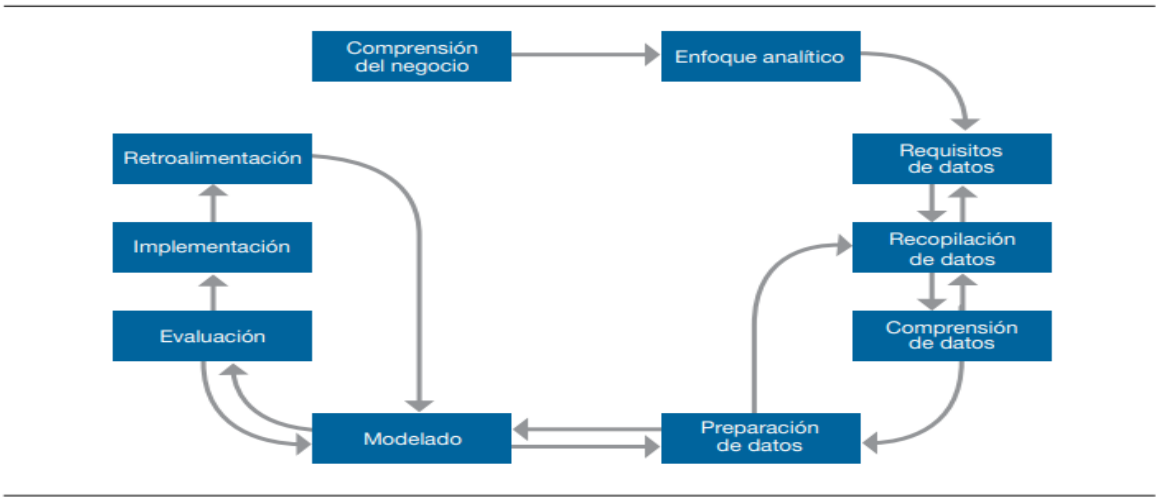
El objetivo del presente caso es el de determinar en qué distritos y zonas dentro de estos se ubican los lugares más convenientes para abrir las nuevas panaderías.

2. Aplicación de la metodología de 10 pasos de la ciencia de datos - Definición del enfoque analítico a emplear

Para la comprensión del problema y su resolución se aplicará la metodología de ciencia de datos de 10 pasos desarrollada por el Ing. Jhon Rollins y aplicada en IBM.

Figura Nro.1

Metodología fundamental de la ciencia de datos



Nota: Esquema que muestra las etapas de la metodología fundamental de la ciencia de datos. De "Metodología Fundamental para la Ciencia de Datos IBM", por IBM, 2022 (https://www.ibm.com/downloads/cas/6RZMKDN8)

2.1. Pregunta de ciencia de datos:

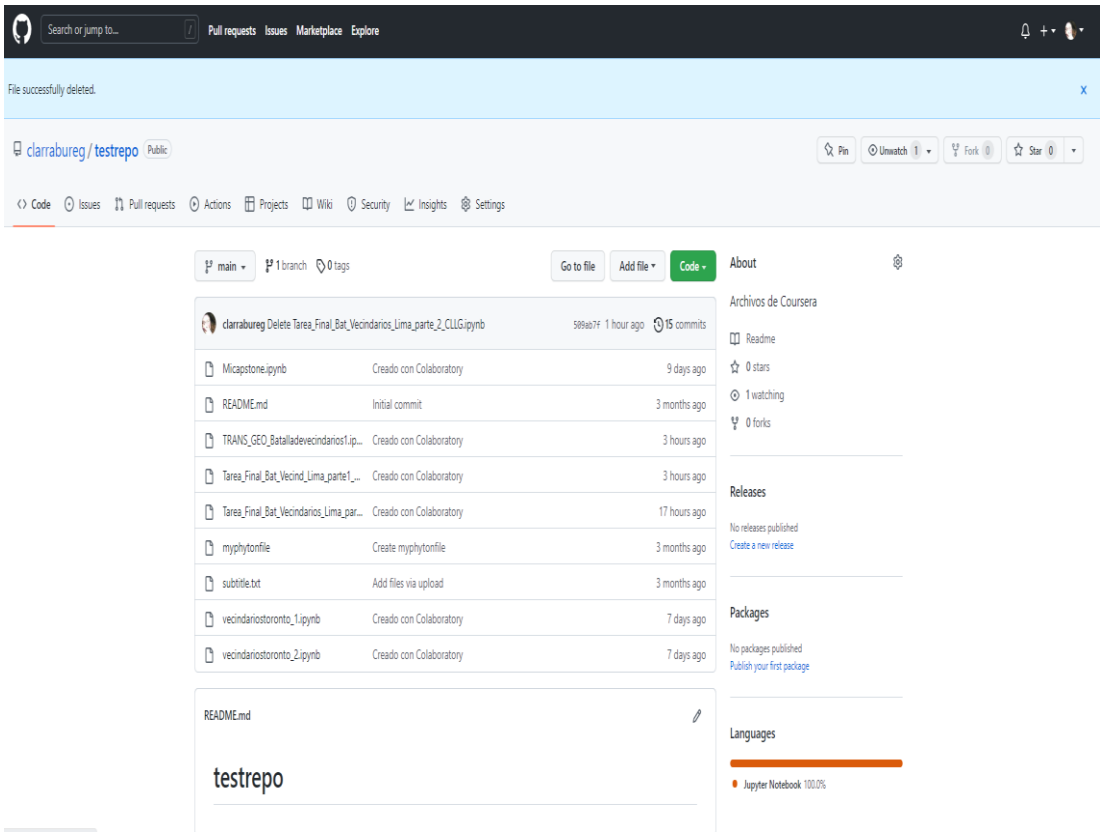
¿Dónde se recomendaría abrir las nuevas panaderías?

2.2. Requisitos de datos:

- Los datos que se requieren son la lista de distritos de Lima Metropolitana, seleccionar aquellos distritos con mayor población de personas de los niveles socio económicos AB. Incluir datos para posicionamiento (longitud y latitud).
- La información será tomada de fuentes de datos válidas y conjuntos de datos de las plataformas abiertas de las entidades gubernamentales del Perú.
- Para el trabajo se utilizará la herramientas Python, Google colab. Los datos se almacenarán en el repositorio github https://github.com/clarrabureg/testrepo/tree/main como se muestra en la figura 2.

Figura 2.

Repositorio de datos GITHUB



Nota: el esquema muestra el contenido parcial del repositorio github montado para el trabajo. https://github.com/clarrabureg/testrepo/tree/main

2.3. Recopilación de datos:

Los datos utilizados para este trabajo fueron recopilados de la plataforma de datos abiertos del gobierno del Perú y de organizaciones de prestigio como es el caso de la Asociación Peruana de Empresas de Inteligencia de Mercados APEIM, que es una organización sin fines de lucro que se dedica a promover y centralizar toda la información del sector de investigación de mercados peruano y es una fuente de datos reconocida en el Perú.

La relación de fuentes de datos utilizada para el presente trabajo se encuentra en la Tabla Nro.1 detallada a continuación.

Tabla Nro.1

Fuentes de datos plataforma de datos abiertos del Perú

Nombre del Archivo	Descripción del conjunto de datos	Ubicación en la plataforma de datos abiertos
TB_UBIGEOS.CSV	Es la tabla de códigos de ubigeos, donde se encuentran incluidos los datos de ubicación de todos los departamentos, provincias y distritos del Perú.	https://www.datosabiertos.gob.pe/dataset/codigos-equivalentes-de-ubigeo-del-peru/resource/4a035ef3-8c50-4a4c-a11b-45a0777aedb3 https://cloud.minsa.gob.pe/s/GkfcJD8xKHJeCqn/download
Dd_TB_UBIGEOS.XLSX	Diccionario de datos de la tabla TB_UBIGEOS.CSV	https://www.datosabiertos.gob.pe/dataset/codigos-equivalentes-de-ubigeo-del-peru/resource/3b48f627-9b02-45dc-96d0-15635de7b33c
2020.ZIP	Encuesta Nacional de Hogares (ENAHO) 2020 - [Instituto Nacional de Estadística e Informática - INEI]	https://www.datosabiertos.gob.pe/dataset/encuesta-nacional-de-hogares-enaho-2020-instituto-nacional-de-estad%C3%ADstica-e-inform%C3%A1tica-2
Geoperu-peru_dist.xlsx	GeoPeru plataforma de datos georeferenciados Información Por distritos	https://visor.geoperu.gob.pe/files/GeoPeru-peru_distritos.xlsx
Informe de Niveles socioeconómicos 2020-2021.pdf	Informe Niveles Socioeconómicos 2021 (APEIM)	https://apeim.com.pe/wp-content/uploads/2022/01/2021-APEIM-NSE-Presentacion_Comite-Vfinal2.pdf

Fuente: Elaboración propia

2.4. Comprensión de los datos

A partir de la información recopilada, se realiza el proceso de comprensión se los datos, sobre el que detalla lo siguiente:

Tabla TB_UBIGEOS:

"Ubigeo" es el termino oficial utilizado para denominar al código de una ubicación geográfica en el Perú. Éste es asignado a las diversas divisiones territoriales que tiene le país (Departamentos, provincias y distritos del Perú).

La tabla TB_UBIGEOS publicada en la plataforma de datos abiertos contiene el código de ubigeo equivalente que convalida los ubigeos otorgados por el Registro Nacional de Identificación y Estado Civil del Perú (RENIEC) y el Instituto Nacional de estadística (INEI). Además de esta información la tabla contiene datos de la superficie total de cada unidad geopolítica, s altitud, latitud y longitud.

Diccionario de datos

TB_UBIGEO

Campo	Descripción	Tipo de dato	Tamaño
id_ubigeo	id	Numérico	4
ubigeo_reniec	UBIGEO distrital de RENIEC	Alfanumérico	6
Ubigeo_inei	UBIGEO distrital de INEI	Alfanumérico	6
departamento_inei	UBIGEO departamental de INEI	Texto	
departamento	Nombre del Departamento	Texto	
provincia_inei	UBIGEO provincial de INEI	Alfanumérico	4
provincia	Nombre de la Provincia	Texto	
distrito	Nombre del Distrito	Texto	
region	Nombre de la Región	Texto	
macroregion_inei	Macroregión a la que pertenece la región, según INEI	Texto	
macroregion_minsa	Macroregión a la que pertene la región, según el MINSA	Texto	
iso_3166_2	Código ISO-3166-2 para la Región	Alfanumérico	6
fips	Código FIPS para la Región	Alfanumérico	4
superficie	Superficie en Km^2 del Distrito	Numérico	11,2
altitud	Altitud del Distrito en metros sobre el nivel del mar (msnm)	Numérico	12
latitud	Latitud del Distrito	Numérico	14,8
longitud	Longitud del Distrito	Numérico	14,8

La tabla fue tratada mediante el programa TRANS_UBIGEO_Batalladevecindarios1.jynb

Tabla Geoperu_peru_dist: (versión corta)

La tabla contiene los datos generales de cada departamento, provincia, distrito del Perú se incluye el total de personas mayores de edad, los códigos de cada departamento, provincia,

distrito asignado para su identificación de Ubigeo el número total de hogares y la población total

Diccionario de datos

Geoperu_peru_dist

Campo	Descripción	Tipo de dato	Tamaño
total_pers	Total de personas mayores de edad	Numérico	
cod_dpto	Código de departamento	Numérico	
departamento	nombre del departamento	Texto	
cod_prov	código de provincia	Numérico	
provincia	nombre de provincia	Texto	
cod_dist	código de distrito	Numérico	
distrito	nombre de distrito	Texto	
num_hog	número de hogares	Numérico	
pob_total	población total	Numérico	

La tabla fue tratada mediante el programa TRANS_GEO_Batalladevecindarios1.jynb

2.5. Preparación de datos

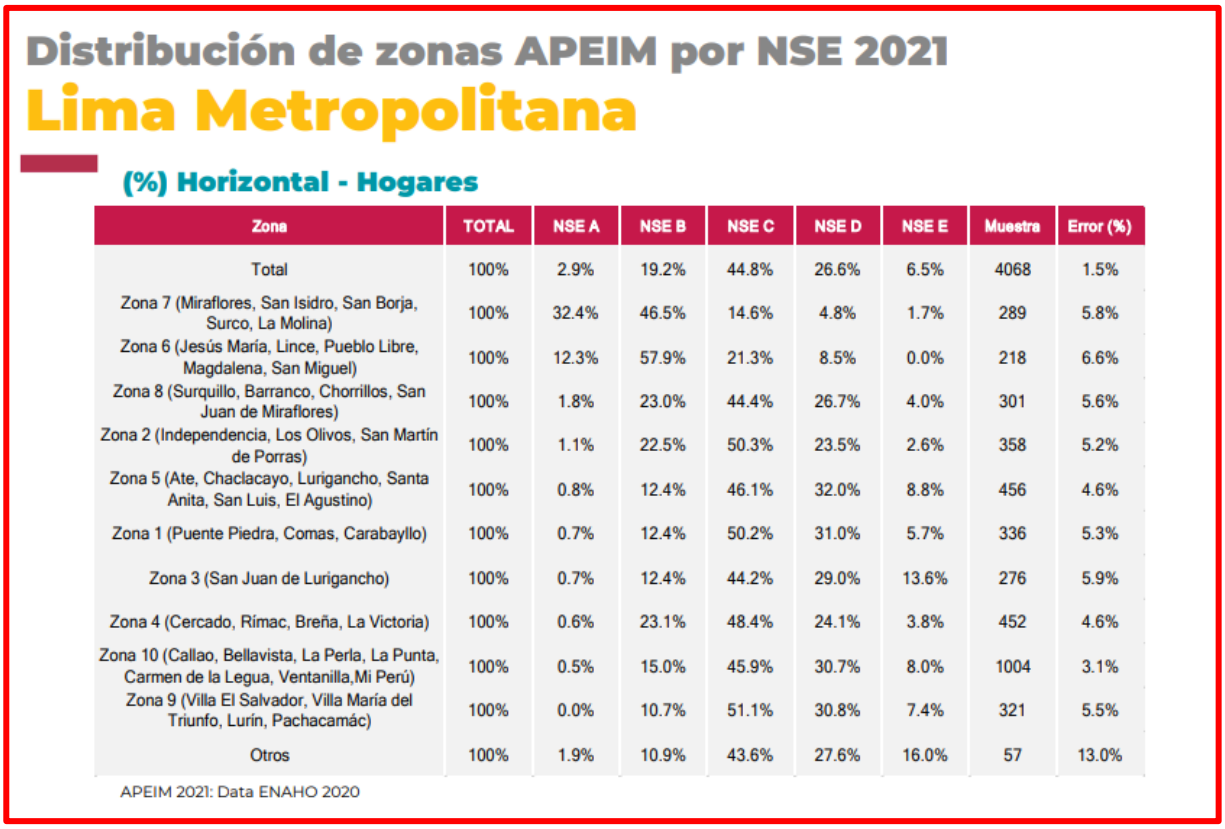
De las Tablas TB_ Ubigeo, Tabla Geoperu_peru_dist se eliminarán las columnas y filas que no son relevantes para el trabajo.

Se conservará solo aquellos distritos de Lima metropolitana que cumplan con las características de población y NSE indicadas para el caso.

Para la información de los niveles socioeconómicos se cuenta con la información de la Encuesta de Hogares ENAHO 2020- Esta información proviene de la plataforma de datos abiertos y del repositorio de datos del Instituto Nacional de Estadística del Perú y ha sido resumida en el informe niveles socioeconómicos 2021 del APEIM. A partir de la información que se muestra en la Figura 3 Se confecciona una tabla APEIM que permitirá trasladar hacia la tabla TB_UBIGEO en un nuevo campo llamado porcentaje_AB.

Figura Nº3

Distristribución de niveles socio económicos por zonas y distritos – Perú



Distribución de zonas APEIM por NSE 2021

Lima Metropolitana

(%) Horizontal - Hogares

Zona	TOTAL	NSE A	NSE B	NSE C	NSE D	NSE E	Muestra	Error (%)
Total	100%	2.9%	19.2%	44.8%	26.6%	6.5%	4068	1.5%
Zona 7 (Miraflores, San Isidro, San Borja, Surco, La Molina)	100%	32.4%	46.5%	14.6%	4.8%	1.7%	289	5.8%
Zona 6 (Jesús María, Lince, Pueblo Libre, Magdalena, San Miguel)	100%	12.3%	57.9%	21.3%	8.5%	0.0%	218	6.6%
Zona 8 (Surquillo, Barranco, Chorrillos, San Juan de Miraflores)	100%	1.8%	23.0%	44.4%	26.7%	4.0%	301	5.6%
Zona 2 (Independencia, Los Olivos, San Martín de Porras)	100%	1.1%	22.5%	50.3%	23.5%	2.6%	358	5.2%
Zona 5 (Ate, Chaclacayo, Lurigancho, Santa Anita, San Luis, El Agustino)	100%	0.8%	12.4%	46.1%	32.0%	8.8%	456	4.6%
Zona 1 (Puente Piedra, Comas, Carabayllo)	100%	0.7%	12.4%	50.2%	31.0%	5.7%	336	5.3%
Zona 3 (San Juan de Lurigancho)	100%	0.7%	12.4%	44.2%	29.0%	13.6%	276	5.9%
Zona 4 (Cercado, Rímac, Breña, La Victoria)	100%	0.6%	23.1%	48.4%	24.1%	3.8%	452	4.6%
Zona 10 (Callao, Bellavista, La Perla, La Punta, Carmen de la Legua, Ventanilla,Mi Perú)	100%	0.5%	15.0%	45.9%	30.7%	8.0%	1004	3.1%
Zona 9 (Villa El Salvador, Villa María del Triunfo, Lurín, Pachacamác)	100%	0.0%	10.7%	51.1%	30.8%	7.4%	321	5.5%
Otros	100%	1.9%	10.9%	43.6%	27.6%	16.0%	57	13.0%

APEIM 2021: Data ENAHO 2020

Nota: Tomado del informe niveles socioeconómicos 2021 – APEIM https://apeim.com.pe/wp-content/uploads/2022/01/2021-APEIM-NSE-Presentacion_Comite-Vfinal2.pdf

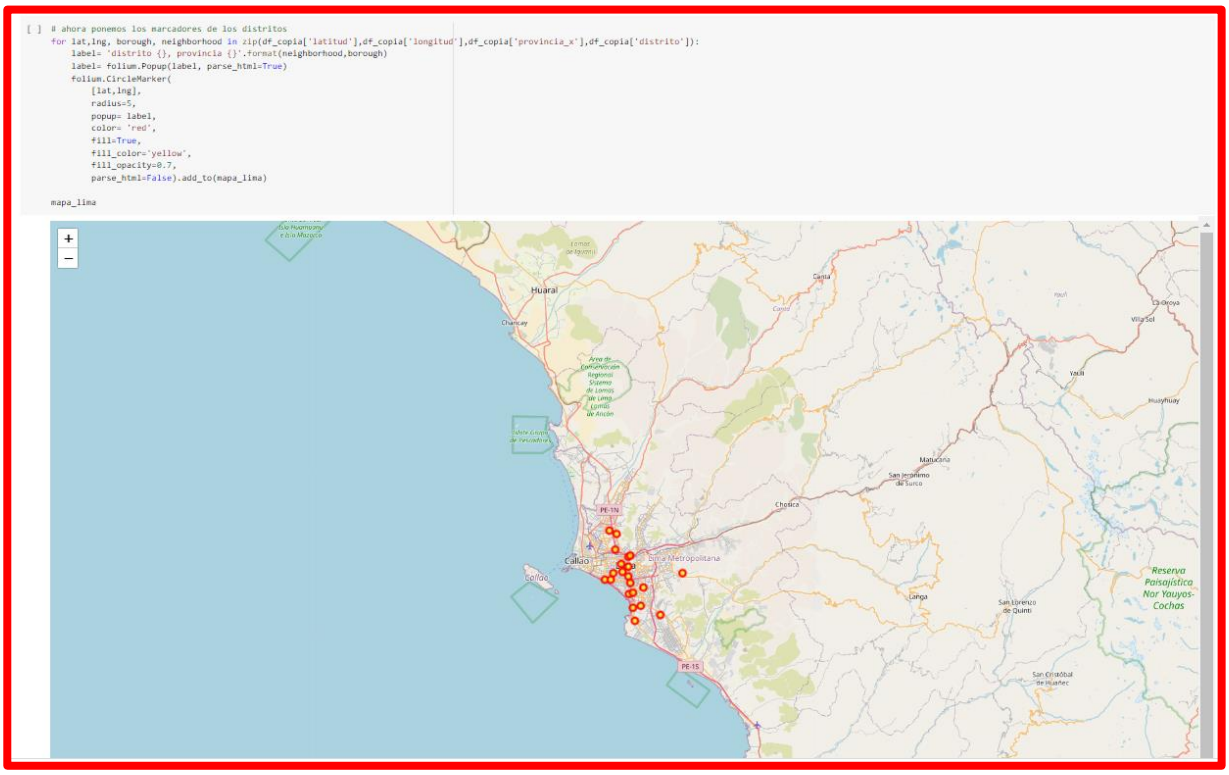
2.5.1. Información consolidada y esquema de procesos siguientes para modelo:

Los datos de las tablas TB_Ubigeos y Geoperu_peru_dist, se unifican mediante el programa COMBINA_Batalladevecindarios1.ipynb completando la nueva estructura de datos para los procesos siguientes.

Con los datos consolidados y a través de la librería Geopy en Python y con las funciones del módulo Nominatim de la misma herramienta permitirán mostrar el mapa de la ciudad de lima y marcar los distritos que cumplen con la condición de tener la mayor población del NSE AB.

Figura 4

Mapa de Lima Perú con los distritos con mayor población del NSE AB



Luego de haber logrado seleccionar los datos para el trabajo, se empleará el API de Foursquare para marcar aquellas ubicaciones favorables para la instalación de las panaderías en los distritos que tienen mayor población NSE AB.

Finalmente, para la segunda parte del informe y aplicando la metodología de ciencia de datos se utilizarán los algoritmos no supervisados desarrollados en los laboratorios de Coursera para la generación de clusters aplicando K-means.

3. Links de Fuentes de datos consultadas para la elaboración del trabajo

https://www.datosabiertos.gob.pe/dataset/mtc-codigo-postal-peru

https://cdn.www.gob.pe/uploads/document/file/1903877/Lima%20Metropolitana_Informaci%C3%B3n%20Territorial%20Completo.pdf

https://www.americaeconomia.com/negocios-industrias/consumo-capita-de-pan-en-el-peru-se-incrementara-en-3-durante-este-ano

https://aspanperu.com/

https://panaderos.info/?p=5779