Universidades de Burgos, León y Valladolid

Máster universitario

Inteligencia de Negocio y Big Data en Entornos Seguros

TFM del Máster Inteligencia de Negocio y Big Data en Entornos Seguros

Sistema de recomendación de municipios españoles en línea basado en las preferencias del usuario y en la similitud entre municipios.

Presentado por Mario Varona Bueno
en Universidad de Burgos — 15 de enero de 2023
Tutor: Álvar Arnaiz González

Universidades de Burgos, León y Valladolid

Máster universitario en Inteligencia de Negocio y Big Data en Entornos Seguros

D. Álvar Arnaiz González, profesor del departamento de Ingeniería Informática, área de Lenguajes y Sistemas Informáticos.

Expone:

Que el alumno D. Mario Varona Bueno, con DNI 70918530P, ha realizado el Trabajo final de Máster en Inteligencia de Negocio y Big Data en Entornos Seguros titulado "Sistema de recomendación de municipios españoles en línea basado en las preferencias del usuario y en la similitud entre municipios".

Y que dicho trabajo ha sido realizado por el alumno bajo la dirección del que suscribe, en virtud de lo cual se autoriza su presentación y defensa.

En Burgos, 15 de enero de 2023

Vº. Bº. del Tutor:

Álvar Arnaiz González

Resumen

En España existen 8.131 municipios, compuestos por más de 60.000 entidades singulares de población. Sin embargo, la gran mayoría de ellos son totalmente desconocidos incluso para personas de la provincia.

La mejora de las telecomunicaciones, infraestructuras, desarrollos turísticos e incremento del trabajo remoto, entre otros muchos factores, permiten considerar nuevos destinos como lugares de ocio o asentamiento; y la minería de datos y las técnicas de procesamiento de grandes cantidades de información permiten ofrecer recomendaciones personalizadas para cada usuario en función de sus gustos.

En el presente trabajo de fin de máster se abordará el diseño y creación de un sistema de recomendación en línea para todos los municipios españoles, que ofrecerá al usuario aquellos que mejor se adapten a sus gustos y más se parezcan a otros cuya preferencia ha manifestado anteriormente, entre otras características.

Primeramente, se expondrá el problema a resolver y los diferentes enfoques existentes. Posteriormente, se presentará la solución elegida junto con los retos encontrados y superados. Finalmente, se realizará una evaluación del funcionamiento del sistema, con la intención de comprobar su utilidad real y de medir el grado de consecución de los objetivos propuestos.

Descriptores

Sistema de recomendación en línea, municipios, España, ciudades, pueblos, sistema de recomendación basado en contenido, sistema de recomendación con filtro colaborativo.

Abstract

There are 8,131 municipalities in Spain, which are compound by more than 64,000 singular entities of population. However, the vast majority of them are totally unknown even by local people.

The improvement of telecommunications, infrastructures, tourist developments and rise of remote working, among many other factors, allow considering new destinations as places of leisure or settlement; and data mining and processing techniques of large amounts of information allow offering personalized recommendations for each user according to their tastes.

This master's thesis will address the design and creation of an online recommendation system for all Spanish municipalities, which will offer to users those places that suit best their tastes and are most similar to others whose preference has been previously expressed, among other characteristics.

Firstly, the problem to be solved and the different existing approaches will be presented. Next, the chosen solution will be introduced together with the encountered and overcome challenges. Finally, an evaluation of the system will be carried out, in order to verify its actual utility and to measure the achievement of the proposed goals.

Keywords

Online recommendation system, municipalities, Spain, cities, towns, villages, content-based recommendation system, collaborative filter recommendation system.

Índice general

Índice de figuras

Memoria

Introducción

Los sistemas de recomendación se han convertido en una de las aplicaciones más usadas dentro del abanico de posibilidades que ofrecen las tecnologías que explotan *Big Data*; es decir, que se aprovechan de una gran cantidad de datos para funcionar. Encontramos estos sistemas de recomendación en redes sociales, portales audiovisuales o cualquier otra plataforma de consumo de contenido –como libros o noticias–, donde nos pueda interesar conocer más sobre alguna temática por la que hemos manifestado un interés previamente. Estos sistemas están, por tanto, especialmente indicados para casos de uso en los que la oferta posible supera con creces la facilidad natural con la que el contenido sería descubierto por los interesados en ausencia del mecanismo de sugerencias automáticas. Si se implementan correctamente, estos mecanismos se convierten en componentes del sistema en los que el usuario confía a la hora de descubrir nuevos grupos musicales, información relacionada o, por qué no, destinos vacacionales o permanentes, atribuyéndoles un argumento de autoridad que en el mundo analógico es habitualmente otorgado a personas de confianza para el usuario, o que considera expertos en la materia.

Partiendo de esa base, y aplicando todos los contenidos vistos en el Máster para el que esta memoria constituye el entregable principal del trabajo final, se ha deseado construir un sistema de recomendación con diversas características que se irán presentando a lo largo de la misma. Pese a que se hace un uso extenso de los conocimientos adquiridos en las asignaturas "Aprendizaje No Supervisado" y "Técnicas de Aprendizaje Automático Escalables", se ha querido aprovechar el mayor número posible de competencias adquiridas en el resto de materias, con especial mención a las procedentes de "Modelos de Programación para el Big Data", "Infraestructura para el Big Data", "Visualización de Datos", "Derecho en Seguridad de Datos" y

"Fundamentos de Ciberseguridad". En cualquier caso, las aportaciones de cada asignatura se señalarán convenientemente a lo largo de esta memoria.

Con la intención de aunar estos conocimientos en un trabajo teórico-práctico cuyo resultado pueda ser de utilidad, y dentro de la función de transferencia y aplicación de conocimiento que el alumno considera que debe exigirse al término de unos estudios universitarios, se propone como idea el diseño y creación de un sistema de recomendación de municipios de toda España, elegida por los motivos que a continuación se exponen:

- **Necesidad de tratar con una gran cantidad de datos.** Pese a que en un primer momento se barajó la idea de acotar el ámbito a municipios de Castilla y León (2.248 frente a los 8.131 nacionales), y aprovechando la gran cantidad de datos abiertos en formato reusable que esta comunidad ofrece y que facilitaría en gran medida las labores de extracción, transformación, limpieza e ingestión de datos en el sistema, se descartó por diversos motivos: En primer lugar, las labores de recogida y limpieza de datos son tan importantes en un proyecto de Big Data como cualquier otra fase, consumiendo habitualmente entre un 50 % y un 80 % de los recursos del proyecto [1]. Esta regla asomaba intuitivamente en asignaturas que requerían algún tipo de ingestión de datos, como "Arquitecturas Big Data" y "Almacenamiento Escalable", pero no ha sido hasta la elaboración de este trabajo –en el que los datos no se suministraban como parte de un enunciado– cuando el alumno la ha podido comprobar empíricamente. En su caso, y dado que ha procurado usar fuentes de datos abiertas en todo lo posible (de ámbito nacional siempre que han estado disponibles, con la intención de evitar manejar fuentes de 19 comunidades y ciudades autónomas distintas), alrededor de un tercio del tiempo de desarrollo total del trabajo han sido invertidos en esta fase, lo que está estrechamente relacionado con el siguiente punto.

- **Deseo de aplicar técnicas de minería y extracción de datos.** Si bien es cierto que en el máster cursado se ha profundizado en diferentes técnicas para mantener una gran cantidad de datos en infraestructuras adecuadas (que escalen horizontalmente y permitan lecturas y escrituras rápidas), en maneras de obtener conclusiones sobre ellos gracias a técnicas estadísticas y de inteligencia artificial, en formas de mantenerlos seguros a nivel técnico y jurídico, y de visualizarlos cómodamente; las limitaciones lógicas de temario y tiempo han impedido incidir en técnicas de extracción automática de datos.

Ha sido gracias a otros recursos aprendidos de forma autodidacta, como el alumno ha conocido técnicas de extracción automática de datos en Internet, cuya puesta en práctica se perfilaba casi obligada en el trabajo final de estos estudios.

- **Intención de crear un sistema de recomendación basado en contenido y en la retroalimentación de los usuarios.** El alumno deseaba poner en práctica sus conocimientos de los dos principales tipos de recomendadores: los basados en contenido y los que utilizan un filtro colaborativo empleando las preferencias de otros usuarios. A lo largo de esta memoria se abordarán ambos y su resultado final.

- **Voluntad de aportar una solución para descubrir ciudades y pueblos españoles.** En numerosas ocasiones se presenta el dilema de qué sitio elegir como destino vacacional o residencial. Hasta ahora, el lugar donde se encontrara trabajo condicionaba enormemente dónde se podría establecer una persona, y los destinos turísticos clásicos ocupaban la mayor parte de la oferta. En la actualidad, la masificación de los mismos ha traído consigo un deseo de explorar nuevas zonas, la mejora de las comunicaciones e infraestructuras ha provocado un resurgimiento en el interés por las zonas rurales, y el coste de vida asociado a las grandes urbes –destino de trabajo obligado para mucha gente– ha provocado la añoranza de una vida más cómoda, fácil de conseguir en una ciudad mediana o pequeña [2], [3]. Por su parte, el incremento del trabajo remoto desde la situación socio-sanitaria derivada de la crisis de la CoViD-19, ha materializado estos factores en un éxodo urbano experimentado en un porcentaje de población lo suficientemente significativo como para ser tenido en cuenta [4], [5]. En nuestro país, el caso paradigmático es el de la ciudad de Madrid, que hasta ahora ha aglutinado a personas emigrantes del interior peninsular en busca de oportunidades laborales, pero no es el único. En el sur y archipiélagos peninsulares se ha observado el caso de trabajadores remotos de todo el mundo que han acudido a ellos para realizar sus funciones desde allí, y la tendencia es claramente alcista [6]. Desde el ámbito mediático y político se puede apreciar como frecuentemente se iguala el concepto de "España vaciada" con territorios muy poco habitados o despoblados [7], olvidando que, para muchas personas, las limitaciones en municipios así son muy grandes como para plantearse el cambio desde una gran ciudad. Sin embargo, para el autor, esta España despoblada incluye inexorablemente a ciudades, capitales de provincia en muchos casos, que en las últimas

décadas han visto disminuir su población drásticamente hacia los sumideros de las grandes áreas urbanas nacionales. Es el caso de la comunidad de las tres universidades de este máster, y es un motivo más para que el alumno no quisiera excluir a ciudades tan válidas para visitar o vivir como Burgos, León, Valladolid o Zamora, por poner algunos ejemplos. Además, y como se detallará en secciones posteriores, el trabajo con las capitales de provincia ha permitido aportar datos muy relevantes, que complementan la información de los municipios cuando, por su tamaño, no es posible disponer de esos datos.

- **Aprovechamiento de experiencia previa con datos abiertos.** El alumno ha querido aprovechar su conocimiento y trabajo previo con diversos catálogos de datos abiertos, como los de la Junta de Castilla y León, los del Gobierno de España o los de la Unión Europea. A mayores, ha querido experimentar con catálogos de datos procedentes de iniciativas privadas o extracción de datos existentes en Internet.

- **Materialización de una idea premiada previamente.** La idea de este trabajo fue una de las seis premiadas el pasado curso en el marco del Programa de Prototipos TCUE (Transferencia de Conocimiento Universidad-Empresa) de la Fundación Universidad de Burgos, dirigido a potenciar la creación de prototipos a partir de trabajos de fin de grado, máster o tesis doctorales que puedan aportar algo novedoso al mercado [8]. Por otro lado, también fue premiada con el segundo premio en la categoría "Ideas" del V Concurso de Datos Abiertos de la Junta de Castilla y León [9], por lo que al recibir esta validación externa, se consideró que se debía intentar llevar a cabo.

- **Deseo de aportar una solución novedosa.** El alumno considera que, además de demostrar y aplicar los conocimientos adquiridos a lo largo de los estudios universitarios, los trabajos finales de grado y máster también deben poder ser publicables; en el sentido de poder devolver a la sociedad parte de lo adquirido gracias a una formación pública de calidad que ella misma ha sufragado. Por ello, se ha diseñado también una forma de distribución de los resultados de este trabajo adecuada al público objetivo, como se detallará posteriormente. Asimismo, si bien se ha encontrado un servicio similar para filtrar por diversos criterios los municipios de Castilla y León [10], o diversas iniciativas para descubrir pueblos de la España vaciada [11], [12], no se ha encontrado ninguna alternativa para descubrir ciudades y pueblos españoles, sin importar su densidad de población, y cuya recomendación esté personalizada para cada usuario.

Objetivos del proyecto

A continuación, se exponen los principales objetivos de este trabajo:

1. **Construir un sistema de recomendación de municipios basado en las preferencias del usuario.** Como parte de este objetivo, se acotará la definición de "preferencias del usuario" y la expectativa de cómo se han de manifestar y tratar. Además, queda establecido por el objetivo que habrá algún tipo de interacción que permitirá introducir datos y mostrar el resultado correspondiente.

2. **Recopilar los datos suficientes como para crear un perfil adecuado de cada municipio.** Este objetivo abarca toda la fase de extracción, transformación y limpieza de datos, con el objetivo de garantizar una adecuada cantidad y calidad en los datos con los que se alimentará al sistema.

3. **Crear un modelo que se comporte adecuadamente.** Para conseguir esta meta, el sistema deberá ser capaz de recomendar municipios similares y no necesariamente evidentes, cuya sugerencia realmente aporte una utilidad real, sea pertinente y coherente; es decir, se evaluará el rendimiento del sistema y se ajustará de manera que arroje las mejores métricas posibles.

4. **Explorar los dos tipos de sistemas de recomendación fundamentales.** Estos son los basados en contenido y los basados en filtros colaborativos. Se procurará explorar ambas ideas con la intención de buscar los casos de uso que mejor se adaptan a cada una, viendo cuál es la que mejores resultados ofrece y analizando los retos encontrados y superados en ambas.

5. **Ofrecer un sistema que pueda ser puesto en producción fácilmente.** El sentido de este objetivo es doble: Por una parte, garantizar la búsqueda de soluciones viables y eficientes, que permitan devolver el resultado de una recomendación fácilmente, sin una demora significativa que impida su uso de forma cómoda. Por otra parte, busca aplicar los conocimientos de las asignaturas de la rama de ciencia de datos, dado que será necesario almacenar grandes cantidades de información y operar con ellas en alguna de las soluciones de infraestructura escalable estudiadas.

6. **Crear el sistema velando por las mejores prácticas que permitan su escalado y mejora continua.** Con este fin se pretende que todas las fases del proyecto –diseño, extracción, transformación, limpieza, modelado, programación, implementación y despliegue– apliquen las mejores prácticas posibles, cada una en su ámbito, para garantizar el escalado del sistema en el futuro –en el hipotético caso de que aumente el número de municipios, por ejemplo–, o su mejora continua –en el supuesto de que se desee hacer actualizaciones periódicas de los datos para mantener el modelo fiable y actual. Este objetivo abarca realizar una adecuada modularización del proyecto software, desacoplar en la mayor medida posible los datos de sus fuentes, utilizar estructuras de datos agnósticas a la presentación final, y aplicar los principios de responsabilidad única, por ejemplo, junto con emplear las mejores prácticas posibles en la implementación técnica de cada fase, como se describirá posteriormente.

7. **Garantizar un adecuado nivel de seguridad jurídica y técnica en cuanto a la puesta en marcha se refiere.** Como parte integral del diseño y desarrollo del sistema, y aplicando los conocimientos de la parte del máster relacionada con las ramas de seguridad y derecho informáticos, se ha pretendido aplicar los principios de seguridad y privacidad por diseño y por defecto, incluidos en el Reglamento Europeo de Protección de Datos [13], que buscan considerar desde el primer momento de ideación la información personal con la que tratará el sistema, intentando minimizarla en la medida de lo posible y restringirla a la imprescindible para funcionar adecuadamente. Esto es especialmente deseado, tal como aboga el Reglamento, en aquellas actividades que conlleven la creación de perfiles individuales o que procesen cantidades de datos de forma masiva, características que podría tener el sistema resultante de este trabajo.

8. **Validar el sistema con usuarios potenciales reales.** Se buscará obtener retroalimentación de personas cercanas al autor que cumplan las condiciones para ser usuarios potenciales del resultado final del trabajo, a fin de validar cualitativamente diversos aspectos del mismo, como la utilidad, capacidad de interacción, tiempo de respuesta del sistema, facilidad de uso, satisfacción con la forma de organizar y presentar la información, coherencia de las respuestas ofrecidas frente a las esperadas, etc. Para ello, se hará uso de diversos conocimientos adquiridos en la asignatura "Visualización de Datos".

Conceptos teóricos

A continuación, se exponen los principales conceptos teóricos cuya aparición en el presente trabajo es relevante:

3.1. Sistema de recomendación

Un sistema de recomendación, también llamado a veces recomendador automático –para reforzar la ausencia de intervención humana–, es una herramienta informática que opera como un sistema de filtrado de información, de forma que, para un conjunto de datos, devuelve los candidatos más prometedores con respecto a la función para la que fue diseñado. Es el caso de los mecanismos que recomiendan qué vídeo, serie o película ver a continuación, qué noticias relacionadas existen o qué artista musical nos puede gustar descubrir. La principal característica de estos sistemas es su capacidad para encontrar cosas que el usuario no está buscando activamente, porque ignora que existen, pero que encajan con su perfil y, por lo tanto, resultan útiles y agradables.

Tal como se vio en la asignatura "Técnicas de Aprendizaje Automático Escalables", se debe tener en cuenta factores como la relevancia (fundamental para ofrecer sugerencias útiles), la novedad o aleatoriedad donde corresponda (puede ser positivo priorizar elementos recientes, como noticias o canciones; o sorprender con resultados con un toque algo más aleatorio), y la diversidad (a fin de ofrecer resultados menos evidentes que los más lógicamente esperables).

Existen dos aproximaciones fundamentales: los sistemas de recomendación basados en contenido y los que están basados en filtros colaborativos. Recientemente, han aparecido soluciones híbridas que aúnan ambos esquemas para ofrecer resultados más robustos, donde los artículos se parecen

individualmente pero las recomendaciones también cuentan con el respaldo de usuarios reales. En ambos casos se desea crear listas de los k elementos cuya recomendación es más favorable.

3.1.1. Sistema de recomendación basado en contenido

En estos sistemas se analiza el contenido de los artículos a comparar (libros, canciones, series...), y se extraen sus características para poder utilizar medidas de similitud entre las propiedades encontradas que aporten una idea de la similitud entre los ítems que forman.

En ocasiones se construyen creando primero perfiles de los usuarios para agrupar los elementos más parecidos, con técnicas similares al *clustering*, como variantes del algoritmo de k-vecinos más cercanos (*k-nearest neighbors*).

La principal limitación de usar este enfoque es el llamado problema de arranque en frío (o *cold start problem*, en inglés), que se caracteriza por no saber cómo actuar ante un caso nuevo, tanto de artículos como de usuarios. Por otra parte, si no queremos depender de los perfiles de los usuarios para encontrar elementos similares –lo que nos permite poder ofrecer recomendaciones incluso aunque no tengamos usuarios en el sistema o este aún no haya expresado ninguna preferencia histórica–, deberemos asegurarnos de que contamos con características suficientes (en número y en representatividad) como para poder encontrar parecidos razonables entre los artículos.

Aquí la definición de similitud no es agnóstica del contenido, sino totalmente dependiente; de hecho, las medidas de distancia empleadas variarán también en gran medida según el dominio del problema. Así, para vectores de números reales se usará habitualmente el coeficiente de correlación de Pearson –que mide la dependencia entre dos variables aleatorias continuas, sin necesidad de que los valores estén en la misma escala– o la similitud coseno –que mide la variación en el ángulo entre dos vectores–; mientras que para vectores binarios (asociados a características cualitativas) se usará la similitud de Jaccard –que mide el número de elementos comunes entre dos conjuntos–.

- Coeficiente de correlación de Pearson. Siendo X e Y un par de variables aleatorias continuas, es el cociente entre la covarianza entre ambas y el producto de su desviación estándar:

$$\rho_{X,Y} = \frac{\sigma_{XY}}{\sigma_X \sigma_Y} = \frac{\mathrm{Cov}(X,Y)}{\sqrt{\mathrm{Var}(X)\,\mathrm{Var}(Y)}}$$

- Similitud coseno. Siendo A y B dos vectores, la similitud coseno es su producto escalar dividido por el producto de sus normas:

$$\text{similitud} = \cos(\theta) = \frac{\mathbf{A} \cdot \mathbf{B}}{\|\mathbf{A}\|\|\mathbf{B}\|} = \frac{\sum_{i=1}^{n} A_i B_i}{\sqrt{\sum_{i=1}^{n} A_i^2}\sqrt{\sum_{i=1}^{n} B_i^2}},$$

- Índice de Jaccard. Siendo A y B dos conjuntos, este índice se calcula dividiendo el número de elementos que se encuentran en su intersección, entre el número de elementos que se encuentran en la unión:

$$\mathcal{J} = \frac{\mid A \cap B \mid}{\mid A \cup B \mid}$$

3.1.2. Sistema de recomendación basado en un filtro colaborativo

Otra manera de evitar las limitaciones de los sistemas de recomendación basados en contenido es utilizar la *sabiduría del grupo* (o *wisdom of the crowd* en inglés), que nos permite ofrecer sugerencias comparando los gustos implícitos (clicks, visitas...) o explícitos (valoraciones, opiniones...) de diferentes usuarios; de forma que, si dos usuarios se parecen en sus perfiles, podremos asumir que existirá una posibilidad significativa de que lo que le ha gustado anteriormente a uno, le guste al otro.

Además, este tipo de sistemas presenta una mayor independencia del dominio, dado que se basa en evaluaciones agnósticas de las características de los elementos que componen el conjunto; y una mayor capacidad de explicación, dado que las preferencias se asocian automáticamente, en lugar de depender de las características extraídas.

Entre los retos que presentan estos sistemas nos encontramos su escalabilidad y dispersión de la matriz de gustos, dado que un usuario valorará habitualmente muy pocos elementos del total de elementos existentes; la privacidad, ya que los usuarios no siempre querrán que se les asocie con los elementos con los que interactúen; y la dificultad para recomendar a usuarios poco comunes.

Los principales métodos de esta familia se basan en memoria y en factorización de matrices. Los primeros buscan similitudes entre usuarios o

artículos (filas o columnas, respectivamente de la matriz de gustos) para después aplicar algoritmos de los k-vecinos más próximos, pero tienen complejidades temporales y espaciales elevadas cuando crecen los datos. Por otra parte, los segundos se basan en modelos para implementar algún algoritmo de factorización de matrices, como ALS (siglas de *Alternating Least Squares*, Mínimos Cuadrados Alternos, en inglés), que permite reducir una matriz de rango alto a dos matrices de rango menor cuyo producto se aproxima a la original y donde el cálculo de similitud entre usuarios o artículos es inmediato y se resuelve con un producto escalar de vectores fila o columna, respectivamente. Para estimar el error global se puede minimizar la raíz del error cuadrático medio, o *Root Mean Squared Error* (RMSE, por sus siglas en inglés), que consiste en la raíz de la media del cuadrado de la diferencia entre los valores pronosticados y los valores observados. Siendo n el número de observaciones, R_{ij} el valor observado, y $U_i * I_j^T$ el valor pronosticado:

$$\text{RMSE} = \sqrt{\frac{1}{n}\Sigma_{i=1}^{n}(R_ij - U_i * I_j^T)^2}$$

3.2. *One Hot Encoding*

Es un tipo de codificación de datos ampliamente usado en tratamiento de características categóricas donde el conjunto de posibles valores se va a representar por un grupo de bits, de forma que cada valor posible se representará con el bit correspondiente en valor alto, y todos los demás en valor bajo.

Por ejemplo, si quisiéramos representar la presencia o ausencia de palabras en un texto de un corpus, crearíamos una matriz con tantas filas como documentos componen el corpus, y tantas columnas como valores posibles tiene nuestro diccionario, marcando con un valor binario la presencia o ausencia de cada palabra en cada texto.

3.3. Discretización

Es el proceso de convertir funciones continuas en homólogas discretas, habitualmente necesario al digitalizar señales analógicas o tratar con variables cualitativas. Puede conseguirse creando intervalos en el rango de valores que toma la función, y aproximando los valores al intervalo más cercano [14].

Siempre que se discretizan valores continuos, existe lo que se conoce como error de discretización, que deberá procurar reducirse a una cantidad aceptable para el modelado correspondiente.

3.4. Normalización

Este concepto tiene amplios significados en el campo de la estadística, pero por lo que a este trabajo respecta, nos referiremos por "normalización" a la conversión de escala de la distribución de una variable para poder realizar comparaciones adimensionales entre diversos elementos o con respecto a sus promedios u otras propiedades estadísticas. De esta manera, es posible comparar valores naturalmente distintos, procedentes de diferentes variables medidas con diversas unidades [15], [16].

Existen diversos tipos de normalización según su caso de aplicación:

- Puntuación estándar: $\frac{X-\mu}{\sigma}$

Es la diferencia entre el valor observado y la media, dividido entre la desviación típica. Se usa para normalizar errores cuando se conocen los parámetros de población, que debe seguir una distribución normal.

- T de Student: $\frac{X-\overline{X}}{s}$

Normaliza los residuos cuando se pueden estimar los parámetros de población pero no se conocen.

- Coeficiente de variación: $\frac{\sigma}{\mu}$

Se aplica en distribuciones positivas como la exponencial o la de Poisson, y utiliza la media como medida de escala para normalizar la dispersión.

- Característica *scaling*: $X' = \frac{X-X_{min}}{X_{max}-X_{min}}$

También se conoce como normalización basada en la unidad, dado que convierte los valores a valores del intervalo [0, 1].

3.5. Organización territorial española

En España existen actualmente 8.131 municipios [17]. Los municipios son, según la Ley reguladora de las Bases del Régimen Local, las entidades locales básicas de la organización territorial del Estado [18]. De acuerdo

con esta ley, los municipios poseen personalidad jurídica y capacidad para el cumplimiento de sus fines, contando con los elementos de su territorio, población y organización. Los municipios se organizan en provincias, y estas en Comunidades Autónomas, a excepción de las dos ciudades autónomas de Ceuta y Melilla, que no disponen de ningún municipio asociado ni pertenecen a otra comunidad autónoma.

Por su parte, existen 61.578 entidades singulares de población, que el Instituto Nacional de Estadística define como *«cualquier área habitable de un término municipal, habitada, o excepcionalmente deshabitada, claramente diferenciada dentro del mismo y que es conocida por una denominación específica que la identifica sin posibilidad de confusión»*. Las entidades singulares pueden estar formadas por varios núcleos de población, y/o tenerla diseminada, englobando edificaciones que no pueden considerarse núcleo. Además, cada entidad posee una calificación tradicionalmente reconocida, como villa, lugar, aldea, urbanización, poblado, caserío, monasterio...

Es importante señalar que la división del municipio en entidades singulares y colectivas (agrupación de varias entidades singulares, o de partes de ellas) de población no goza de carácter oficial, pero sí gran tradición. No es infrecuente debido al alto número de entidades existentes conocer a una o a varias, e incluso presuponer que se trata de un municipio en sí mismo. Esto es especialmente notable en la mitad norte peninsular, a causa de la dispersión geográfica y la orografía, que junto con razones históricas motiva la existencia de numerosos diseminados, muchos de ellos actualmente deshabitados [19].

Desde el año 1981, el Instituto Nacional de Estadística etiqueta cada núcleo de población con un código numérico formado por 11 cifras que, comenzando por la izquierda, tiene la siguiente forma:

- Las dos primeras cifras corresponden a la provincia.
- Las tres siguientes (tercera, cuarta y quinta) corresponden al municipio.
- La sexta y séptima corresponden a la entidad colectiva, si la hubiera. De no ser el caso, se usarán dos ceros.
- La octava y novena cifras codifican la entidad singular dentro de la entidad colectiva o municipal, según corresponda.
- La décima y undécima se refieren a los núcleos de población dentro de la entidad singular, usando "99"para diseminados.

Los códigos fueron actualizados en el año 1991 para mantener un orden alfabético, y corresponde a los ayuntamientos revisar, al menos anualmente, la relación de entidades y núcleos de población, para ser remitida al Instituto Nacional de Estadística, que las publica con la misma frecuencia. En caso de que se incorporen nuevos núcleos, se les asignará un código correlativo al último asignado, no reutilizando nunca los códigos de los núcleos que pudieran desaparecer. Conviene señalar que es posible que existan municipios o incluso entidades singulares de población con el mismo nombre –de hecho, ocurre en varios casos–, por lo que la identificación debe hacerse con el código del INE siempre que sea posible. Sin embargo, debido al alto número de instituciones que no emplean este código, la desambigüación debe hacerse con la provincia de pertenencia, donde no existen municipios duplicados.

3.6. Datos abiertos

La filosofía de datos abiertos (u *Open Data*, en inglés) busca liberar determinados tipos de datos, especialmente concernientes a instituciones públicas y privadas, sin que suponga un menoscabo de las libertades individuales, de forma libre; esto es, sin limitaciones de derechos de autor, patentes, regalías u otro tipo de mecanismos de control, amparándose en el derecho a saber de la ciudadanía, la transparencia como parte de un buen gobierno público y corporativo, y en la utilidad pública de herramientas que surgen de hacer la información accesible y reusable. [20]

Esto último se consigue con una limpieza previa de los datos, donde el organismo titular de los mismos se encarga de garantizar que los conjuntos de datos elegidos son aptos para ser liberados, tanto por su calidad (limpieza, corrección...), como por su capacidad jurídica para poder hacerlo (los datos están agregados, anonimizados o no se corresponden con personas físicas identificadas o identificables.

Además, es fundamental que las capacidades de adquisición y procesamiento técnicas de los datos sean grandes, con el objetivo de no obstaculizar el propósito mismo para el que los datos abiertos son diseñados: promover el desarrollo de contenidos periodísticos, soluciones técnicas, productos, servicios, etc. gratuita o comercialmente, que beneficien a la sociedad basándose en la idea de aprovechar datos, habitualmente de titularidad pública o considerados como de interés público (como datos de consumo, datos de calidad de aire y aguas, datos de tráfico, datos de contagios, etc...), para realizar un análisis y extraer conclusiones valiosas.

Esto se materializa en servicios que idealmente cuentan con APIs (interfaces de programación de aplicaciones, por sus siglas en inglés) y/o capacidad de exportación de los datos en múltiples formatos habitualmente reconocidos fácilmente por máquinas, como los formatos CSV (valores separados por comas, por sus siglas en inglés), JSON (*JavaScript Object Notation*, notación de Objetos JavaScript en inglés), XML (*Extensible Markup Language*, lenguaje de marcado extensible en inglés) o XLSX (formato de hojas de cálculo de Microsoft Excel, el menos reusable de los comentados por tratarse de archivos binarios de software propietario comercial; pero no por ello menos usado en administraciones públicas).

3.7. *Web scraping*

Traducido en español como *raspado Web*, abarca el conjunto de medios técnicos que posibilitan la extracción automática de información de sitios Web. Habitualmente, estas técnicas se basan en programas informáticos que simulan la navegación de un usuario legítimo en Internet, realizando las llamadas a través del protocolo de transferencia de hipertexto y su versión segura (HTTP y HTTPS, por sus siglas en inglés), o imitando los movimientos de un humano a través de un navegador Web con interfaz gráfica.

Dado que se basan en imitar el comportamiento de un humano, o la llamada a un servidor de un navegador Web, los resultados no son siempre los deseados, ya que existen diversas medidas técnicas que los autores de un sitio Web pueden adoptar, bien para evitar esta técnica en sus versiones más factibles, o bien como parte de las decisiones del diseño del sistema implementado, que pueden provocar que falle el rastreo Web. Un ejemplo de esto último son las páginas basadas en *scroll* (o deslizamiento vertical) infinito, o que completan en algún momento la carga de datos de manera dinámica tiempo después de haber terminado de enviar la página al cliente de red. Este enfoque complica la creación de un programa que extraiga la información, dado que es notablemente más factible comunicarse vía HTTP/S con un servidor, lo que normalmente provoca el cierre de la conexión cuando se recibe la respuesta, no siendo posible emular el desplazamiento vertical o interacción que realizaría un humano.

Por estos motivos, esta técnica se usa como último recurso para extraer datos, prevaleciendo siempre que sea posible el uso de APIs públicas de la institución, que están en sí mismas diseñadas para un uso mecanizado y

automático, y donde los cambios suelen estar indicados mediante mecanismos de versionado.

Es importante resaltar que el raspado Web, por definición, consiste en extraer datos que no necesariamente fueron concebidos para su extracción, por lo que puede conllevar implicaciones legales. Actualmente, la legislación no es clara en este sentido, y la jurisprudencia ha dictaminado veredictos cuyo resultado depende en gran medida de las características del *scrapper* y del uso que se fuera a dar a esos datos [21], [22].

Entre los usos más comunes del Web *scraping* encontramos la monitorización de precios, la inyección de datos de otros sitios Web de forma adaptada a las necesidades del que los inyecta, la lectura de contenido desde fuentes externas o la recopilación de datos para enriquecer búsquedas, creando fragmentos accesorios conocidos como *rich snippets*, que aportan información de diversas fuentes de confianza relacionadas con la búsqueda.

3.8. *Geocoding* y distancias

La geocodificación, referida habitualmente en inglés como *geocoding*, abarca el proceso de asignar coordenadas geográficas a puntos geográficos cualitativos (ciudades, pueblos, direcciones, edificios, lugares emblemáticos, puntos de interés, accidentes geográficos...). De esta manera, se transforma una dirección exacta o aproximada, en unas coordenadas que permiten su representación inequívoca en un plano de un Sistema de Información Geográfica [23].

Conociendo la posición exacta de dos puntos es posible calcular su distancia por diversos medios, siendo algunas de las más empleadas las siguientes:

- Distancia por carretera. Es una de las más populares por su utilidad, dado que tiene en cuenta la disponibilidad de infraestructura y las características orográficas del terreno para poder comunicar dos puntos cualesquiera por vía terrestre.
- Distancia geodésica. Una línea geodésica es aquella de menor longitud entre dos puntos en una superficie dada, estando contenida en la propia superficie. Habitualmente nos referimos a esta distancia como la distancia *en línea recta*, y no tiene en cuenta los accidentes geográficos, infraestructura disponible u otras características que limiten la comunicación ente los puntos.

Pese a que el cálculo de distancias entre dos puntos geolocalizados (identificados con coordenadas geográficas, habitualmente latitud y longitud) puede requerir de un tiempo de computación elevado –especialmente si se tienen en cuenta factores como la conveniencia de la ruta o posibles combinaciones con medios de transporte colectivo–, es un cálculo factible. Sin embargo, no ocurre lo mismo con determinadas distancias que los humanos consideramos aún más intuitivas, como la distancia de un punto a la costa o a la montaña.

Incluso aún siendo capaces de definir formalmente conceptos como *costa* y *montaña* –asumiendo, por ejemplo, lugares de altitud menor y mayor, respectivamente, a unos umbrales determinados sobre el nivel del mar– no es trivial elegir el punto costero o montañoso más cercano para poder calcular algún tipo de distancia geográfica.

Un posible enfoque podría ser calcular la distancia a todos los puntos que componen la frontera natural de un territorio con la costa, y elegir la menor, pero podría estar sujeto a problemas de solución arbitraria, como en el caso de playas interiores o costas incluidas en el territorio. En el escenario de sistemas montañosos, sería necesario contar con el perímetro de todos ellos, algo difícil de conseguir en el caso de las montañas de menor altitud o que se encuentran fuera de sistemas montañosos importantes. Además, en ambos supuestos, el coste lógico que supone realizar comparaciones para un gran número de pares de puntos es elevado.

Otro enfoque puede ser contar con datos geolocalizados de accidentes geográficos similares, como “playas” o “picos montañosos significativos”, y calcular la distancia hasta los mismos. De nuevo, se puede caer en el error de asimilar una “playa” con una “costa”, teniendo problemas similares en el caso de las montañas (dado que se forzaría a escoger solo los picos más significativos, que son los que cuentan con denominación y geolocalización conocidos), incurriendo en costes altos de igual manera y dependiendo notablemente de la cantidad y calidad de los datos auxiliares etiquetados.

Una de las formas de resolver problemas de distancias en tiempos razonables consiste en emplear heurísticas para aproximar la solución del problema. Por ejemplo, se podría tomar la heurística de considerar que los puntos pertenecientes a subconjuntos territoriales de ámbito superior que limitan con el hito geográfico de interés presentan la menor distancia posible, y que esta aumenta conforme aumenta la distancia de otras agrupaciones territoriales a las anteriores. De esta forma, los municipios pertenecientes a provincias costeras, serán los que menor distancia a la costa tendrán, los municipios de provincias colindantes a las costeras presentarán la segunda

menor distancia a la costa, los municipios de provincias colindantes a las anteriores, la tercera menor distancia, y así sucesivamente, teniendo en cuenta que ninguna agrupación puede tener dos asignaciones de distancia distintas. En el caso de la distancia a la montaña, se podría tomar una heurística similar con ligeras modificaciones según el dominio del problema.

3.9. *Paas*

Una plataforma como servicio, o *PaaS* por sus siglas en inglés, es un conjunto de servicios y herramientas proporcionadas por un proveedor de computación en la nube (o *cloud computing*) que permiten abstraer la infraestructura subyacente (controlada por el proveedor) y ofrecer a los desarrolladores de aplicaciones un entorno de despliegue de aplicaciones con diversas prestaciones que difieren de las capacidades de una instalación en sistemas propios (o que abaratan significativamente sus costes) [24].

Dado que los detalles de infraestructura son abstraídos para el programador, este puede despreocuparse de la escalabilidad, velocidad o instalación de actualizaciones de la plataforma, que será configurable, pero dependerá en última instancia del proveedor de la computación en la nube. Además, el uso de APIs le permiten comunicarse con el sistema como si realmente este fuera conocido para él, por lo que los resultados a nivel de usuario son equivalentes.

Ejemplos de plataformas como servicio son App Engine, de Google; Azure, de Microsoft; AWS Lambda, de Amazon; o Heroku, de Salesforce.

Técnicas y herramientas

A continuación, se presentan las principales técnicas y herramientas usadas en la construcción de este trabajo.

4.1. Metodología de trabajo

Como se ha estudiado en la asignatura "Técnicas de Aprendizaje Automático Escalables", existen diversas metodologías para proyectos de Big Data o ciencia de datos en grandes cantidades. En este punto, se procede a hacer un breve análisis de algunas de las más populares, y se indica también cuál ha sido la elegida para realizar este trabajo y por qué:

4.1.1. KDD (*Knowledge Discovery in Databases*)

KDD, del inglés "Extracción de conocimiento en bases de datos", se originó en la comunidad investigadora, y distingue entre etapas de identificación, preprocesado y modelado. Pese a que existe una retroalimentación entre etapas, pone el foco en las etapas de preprocesado, evaluación y presentación del conocimiento adquirido, en línea con el peso mencionado anteriormente que estos pasos tienen en cualquier proyecto de Big Data. Los procesos de esta metodología se pueden resumir en cinco etapas de la siguiente manera:

1. Selección de datos a explorar. En primer lugar, elegiremos los datos que queremos integrar en la base de datos.
2. Preprocesamiento de datos. Abarca las etapas de limpieza, donde se eliminará el ruido u otros datos incoherentes; la integración, que consiste en la unificación de datos procedentes de diversos orígenes; y

la selección final de datos, que toma de la base de datos existente en este punto los datos relevantes (es decir, un subconjunto de variables) para la exploración del fenómeno a estudiar.

3. Transformación de datos. Dentro de esta etapa, los datos adquieren formas apropiadas para su posterior procesado. Ejemplos de estas tareas son la creación de sumatorios, el uso de funciones de agregación, u otras similares para resumir los datos.

4. Minería de datos. Consiste en buscar patrones de interés o representativos en los datos por medio de diversas técnicas.

5. Evaluación de patrones e interpretación del conocimiento. Los últimos pasos de esta metodología, y que condicionan una posible regresión al comienzo, abarcan identificar los patrones que realmente aportan información novedosa o interesante, junto con su adecuada visualización para poder representar este nuevo conocimiento extraído al usuario.

4.1.2. SEMMA (*Sample, Explore, Modify, Model and Access*)

La metodología SEMMA (muestrear, explorar, modificar, modelar y acceder, por sus siglas en inglés) se basa en muestrear la base de datos principal, asumiendo que un procesamiento completo es complejo y lento de llevar a cabo. Pone el foco en la transformación y modelado de los datos, omitiendo aspectos de negocio. Sus principales etapas son las siguientes:

1. Muestrear. Consiste en tomar muestras de pequeño tamaño de la base de datos, con el objetivo de poder realizar manipulaciones de forma ágil.

2. Explorar. Esta fase busca aumentar el entendimiento de los datos, refinando el proceso mientras se buscan anomalías, tendencias o patrones.

3. Modificar. Abarca la creación, selección y transformación de variables para reducir su número, con la intención de contar con datos relevantes.

4. Modelar. Busca crear el modelo que mejor satisfaga los objetivos propuestos al inicio del proyecto.

5. Acceder. Evalúa la confianza y utilidad de los resultados extraídos, provocando una repetición del proceso en caso de que los objetivos no se hayan cumplido.

4.1.3. CRISP-DM (*Cross-Industry Standard Process for Data Mining*)

CRISP-DM (proceso estándar entre industrias para el minado de datos, por sus siglas en inglés) es la metodología más usada actualmente [25], y se diferencia de las anteriores por considerar el proceso de negocio dentro del propio ciclo de vida de los datos. Así, sus principales fases son:

1. Entendimiento de negocio. Desde una perspectiva de negocio, se busca entender los objetivos y requisitos del proyecto, con la intención de convertirlos en una definición de un problema de minado de datos, para diseñar posteriormente un plan inicial para alcanzar dichos objetivos. Esta fase es de gran importancia en esta metodología, ya que se determinan los antecedentes, metas estratégicas del proyecto y criterios de éxito. Además, se realiza un inventario de los recursos, un análisis de costes y beneficios estimados, se determinan los objetivos finales y se produce el plan preliminar mencionado.

2. Entendimiento de datos. Esta fase comienza con una colección de datos inicial y actividades para familiarizarse con los datos, explorarlos, identificar problemas de calidad, descubrir las primeras observaciones o detectar conjuntos interesantes para formar hipótesis a partir de información oculta.

3. Preparación de datos. En este punto, se realizan las transformaciones adecuadas para construir el conjunto de datos final que compondrá el modelo. Abarca transformaciones, limpieza, reducción de variables o cambios de formato, y su resultado es un entregable apto para ser introducido en la herramienta de modelado elegida.

4. Modelado. Se seleccionan y aplican varias técnicas de modelado, cuyos parámetros se ajustan hasta encontrar los valores óptimos. Habitualmente será necesario volver a la etapa de preprocesado anterior para adecuar los datos a una nueva técnica de modelado, ya que con frecuencia se requieren tratamientos o formatos especiales para ciertas técnicas.

5. Evaluación. Desde un punto de vista de negocio, se revisan los pasos ejecutados para construir el modelo junto con el grado de consecución de los objetivos inicialmente propuestos y la revisión de los criterios marcados para su evaluación. Se busca encontrar un objetivo clave de negocio que no haya sido suficientemente satisfecho, y al final de esta fase se decide si los resultados son aptos para continuar o no.

6. Despliegue. Al igual que en el resto de metodologías mencionadas, CRISP-DM entiende que el desarrollo del proyecto no termina normalmente con la creación del proyecto. Por el contrario, suele ser necesario interpretar y presentar los conocimientos obtenidos de una manera adecuada según los requisitos de negocio, pudiendo significar la creación de informes o cuadros de mando, la elaboración de procesos de minado repetibles, el acceso e interacción con sistemas que contengan los resultados obtenidos, etc.

Una de las principales diferencias entre esta metodología y las anteriores, es la posibilidad de volver hacia una etapa anterior en cualquier paso de la metodología CRISP-DM, promoviendo la mejora continua flexible y ágil, necesarias en muchos entornos empresariales.

Esta flexibilidad para mejorar los conjuntos de datos constantemente, junto con la gran popularidad de la que goza y el énfasis que realiza en los requisitos de negocio y en los criterios de evaluación y aceptación –cuya definición y consecución es fundamental en cualquier proyecto real, como el que se ha buscado realizar con este trabajo–, han propiciado su elección para el mismo.

Por último, para administrar y monitorizar el progreso del desarrollo de este trabajo, se ha utilizado el gestor de tareas Todoist [26], y las herramientas conectadas al sistema de control de cambios Git, como tickets o *issues*, de los repositorios en línea de GitHub [27].

4.2. Fuentes de datos

Debido a la importancia de los procesos de extracción, transformación y limpieza para el desarrollo de este trabajo, se considera relevante incluir las principales fuentes de datos consultadas o explotadas para el mismo:

- Altitud por municipio. Fuente: OpenElevation [28].
- Centros de Atención Primaria, Centros Hospitalarios y Centros de Atención Urgente Extrahospitalaria. Fuente: Ministerio de Sanidad [29].
- Cobertura de banda ancha por municipio. Fuente: Ministerio de Asuntos Económicos y Transformación Digital [30].
- Datos climáticos por municipio. Fuente: OpenWeather [31].

- Distancias a las capitales de provincias. Fuente: Cálculos por carretera de OpenTripPlanner [32], y en línea recta a través de la distancia geodésica de GeoPy [33] cuando la primera no se encontró.
- Extracto de información del municipio e imágenes del municipio. Fuente: Wikipedia [34].
- Geoposicionamiento por municipio. Fuente: PositionStack [35].
- Lugares más significativos por municipio. Fuente: GeoApify [36].
- Nomenclátor de Entidades Singulares del INE, con municipio de pertenencia asociado. Fuente: INE [37] y Francisco Ruiz [38].
- Nomenclátor Geográfico de Municipios y Entidades de Población. Fuente: Centro Nacional de Información Geográfica [39].
- Número de colegios o institutos por municipio. Fuente: Ministerio de Educación [40].
- Número de ofertas de empleo disponibles por provincia. Fuente: Infojobs [41].
- Número de universidades por municipio. Fuente: Ministerio de Educación [42].
- Precios medios de viviendas en venta, y precios medios de viviendas en alquiler por provincia. Fuente: Fotocasa [43].
- Precios y número de viviendas en venta, y precios y número de viviendas en alquiler por municipio. Fuente: Idealista [44], y Fotocasa [43], usada en los casos en los que la primera falló.
- Provincias de España y sus capitales. Fuente: Libretilla [45].
- Relieve de España. Fuente: El Orden Mundial [46].
- Rentas brutas medias por municipio de la Comunidad Foral de Navarra. Fuente: Hacienda Foral [47].
- Rentas brutas medias por municipio de País Vasco. Fuente: Eustat [48], [49].
- Rentas brutas medias por municipio. Fuente: Agencia Tributaria para todos los territorios excepto País Vasco [50], y la Comunidad Foral de Navarra [47].

- Rugosidad del terreno en España. Fuente: Fundación BBVA [51].
- Tasa de actividad, paro y empleo por provincia y sexo. Fuente: INE [52].

Como se detallará posteriormente, en diversos momentos de las fases de ETL y exploración de datos ha sido necesario utilizar diversas utilidades de terceros para procesar parte de los datos, verificar expresiones o comprobar la integridad de los mismos. Estas han sido las siguientes:

- Postman, de Postman [53]. Utilizada para realizar peticiones HTTP y HTTPS a los servicios Web ajenos (para conocer el formato de la respuesta devuelta), y propios (para ahorrar peticiones innecesarias durante la programación de los *scripts*).
- Comparador de listas, de Molbiotools [54]. Se ha utilizado para comparar listas y verificar la corrección de los cambios.
- Contador de líneas duplicadas en listas, de Shailesh N. Humbad [55]. Empleada para verificar la integridad de los datos.
- Conversor de CSV a JSON, de Flatfile [56]. Empleada para crear archivos JSON para la Web a partir de los archivos CSV manejados habitualmente.
- Coordinates Plotter, de Darrin J. Ward [57]. Usada para explorar los datos de geoposicionamiento.
- Detector de duplicados, de dCode [58]. Ha sido usada para buscar duplicados en listas, con información complementaria a la del resto de herramientas.
- Excel, de Microsoft [59], y Numbers, de Apple [60]. Utilizadas para abrir los archivos en formato CSV, XLS y XLSX, y verificar su integridad y corrección.
- Herramienta de explicación, construcción y pruebas de expresiones regulares, de gskinner [61]. Empleada para construir y comprobar la corrección de diversas expresiones regulares.

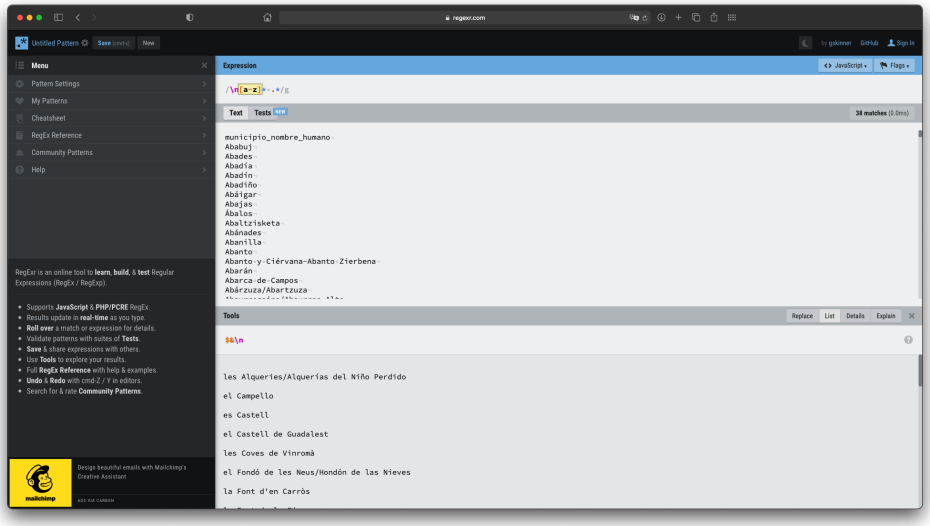

Figura 4.1: Manejo de expresiones regulares para detectar y corregir inconsistencias.

- Herramienta para ordenar una lista alfabéticamente, de Scott Smind [62]. Utilizada para ordenar puntos de la memoria y anexos alfabéticamente.
- Supresor de líneas duplicadas en listas, de autoría anónima [63]. Usada para eliminar duplicados en listas intermedias.
- Supresor de líneas vacías en listas, de CodeBeautify [64]. Usada para eliminar líneas en blanco tras supresiones manuales o automáticas.
- Validador de datos en formato JSON, de CircleCell [65]. Utilizado para comprobar la integridad y corrección de datos en formato JSON.
- Visualizador de datos en formato JSON, de CodeBeautify [66]. Utilizada para visualizar gráficamente el árbol jerárquico de un archivo en formato JSON.

4.3. Programación

Como se detallará posteriormente, el proyecto abarca *scripts* de programación escritos en Python y JavaScript que extraen, transforman, limpian y procesan información de diversas fuentes (tanto propias como ajenas), tomando los datos de archivos CSV (archivos separados por comas, por

sus siglas en inglés, o puntos y comas, en el caso de este trabajo), JSON (archivos de notación JavaScript), XML (lenguaje de marcado extensible) o páginas de Internet públicas; y devolviendo archivos en formato CSV o JSON.

Entre las bibliotecas de terceros usadas, destacan las siguientes, ordenadas alfabéticamente y respetando la grafía de los autores:

- beautifulsoup4. Desarrollada por Leonard Richardson bajo licencia MIT [67], permite extraer fácilmente información de páginas Web mediante mecanismos de interpretación y selección avanzada de elementos HTML y XML. Se ha utilizado en los archivos de extracción de datos de origen Web.

- Flask. Publicada por Armin Ronacher con licencia BSD [68], es un micro *framework* Web, o micro marco de desarrollo Web, para aplicaciones en línea basadas en Python. A diferencia de otros marcos, Flask es ligero y ofrece una gran flexibilidad en cuanto a arquitectura, herramientas, estructura de ficheros o dependencias, lo que lo convierte en una opción escalable y que se adapta bien a numerosos tipos de proyectos. Constituye la base sobre la que se ejecuta la aplicación Web final.

- Jinja2. Desarrollada por el mismo autor de Flask y bajo la misma licencia [69], Jinja es un motor de plantillas ligero y extensible, que permite crear páginas HTML dinámicas en tiempo real, facilitando enormemente la programación de sitios Web de gran tamaño. Se ha utilizado para generar las páginas HTML de la aplicación Web, con algunas particularidades que se detallarán posteriormente.

- pandas. Programada por el equipo de desarrollo de Pandas ("The Pandas Development Team"), y publicada bajo licencia BSD [70], proporciona estructuras de datos eficientes para manejar grandes cantidades de información, series temporales y estadísticas varias. Se ha utilizado para acceder y procesar la información de los archivos CSV y JSON.

- python-dotenv. Desarrollada por Saurabh Kumar con licencia BSD [71], permite utilizar variables de entorno en Python de fácilmente, evitando escribir secretos como claves de API o credenciales en el código en texto plano que queda registrado en el sistema de control de versiones. De esta manera, los secretos se almacenan codificados

mediante un par clave-valor en un archivo ".env", de variables de entorno. Este archivo se incluye en el fichero ".gitignore" para evitar su adición al repositorio. Tras incluir y llamar a la función principal de esta biblioteca, los secretos se cargan en el entorno del sistema para acceder a ellos de forma segura, quedando localmente en la máquina de destino y evitando su propagación en texto plano más allá del citado archivo. Se ha empleado para no exponer las credenciales de las APIs usadas en la fase de extracción Web y en algunos servicios del producto final. Esta es una de las medidas de seguridad estudiadas en la asignatura "Fundamentos de Ciberseguridad", y presente en diversos documentos de seguridad estándar, como el Plan Director de Seguridad del INCIBE [72] o el proyecto OWASP para la seguridad de aplicaciones Web [73].

- phpDotEnv. Desarrollada por Vance Lucas con licencia BSD (cláusula 3) [74], permite la misma medida de seguridad anterior en el servidor donde se ejecuta el código PHP.

- re. Este módulo, desarrollado por la fundación de software Python ("Python Software Foundation") forma parte de la distribución oficial de Python y utiliza su licencia propia (PSFL, por sus siglas en inglés) [75]. Permite operar con expresiones regulares, y se ha utilizado para buscar patrones, como por ejemplo precios o metros cuadrados en páginas Web, o paréntesis en los nombres de los municipios.

- requests. Publicada por Kenneth Reitz bajo la licencia Apache 2.0 [76], esta biblioteca ofrece implementar la capa de red para comunicaciones HTTP y HTTPS de forma sencilla. Se ha utilizado para realizar las llamadas en los archivos de extracción Web.

- scikit-learn. Desarrollada por Andreas Mueller con licencia BSD [77], es una serie de módulos para el aprendizaje automático y el minado de datos. Se ha utilizado para crear el modelo y los recomendadores.

- surprise. Incluida posteriormente como módulo de scikit-learn [78], es una biblioteca desarrollada por Nicolas Hug con licencia BSD para sistemas de recomendación [79]. Se ha utilizado para desarrollar el recomendador de filtro colaborativo.

- sys. Mencionado por su gran presencia en el código, el módulo del sistema de Python permite acceder a variables de entorno y funciones que interactúan con el intérprete [80]. Se ha utilizado para pasar argumentos a los *scripts* por línea de comandos, indicando valores como

los archivos de entrada o salida u otros parámetros de configuración de los mismos.

- wikipedia. Publicada por Jonathan Goldsmith con licencia MIT [81], ofrece la API pública de Wikipedia a través de Python. Se ha utilizado para extraer la información de Wikipedia de los distintos municipios.

Se ha aprovechado la existencia de un servidor personal para generar archivos de imitación (*mock*) de diversas fuentes a las que se quería acceder durante el proceso de extracción de datos (tanto vía API como vía HTTP), con la intención de probar los guiones de extracción escritos en Python y JavaScript contra ellos durante el desarrollo. De esta manera, se han utilizado estos archivos de imitación durante la fase de depuración de errores, verificando el correcto funcionamiento de los archivos antes de realizar las llamadas a los servicios de terceros, y evitando incurrir en gastos de cuotas innecesarios en los servicios accedidos mediante API, o en demasiadas llamadas a los servicios de terceros durante el proceso de *Web scraping*.

Además, se han creado archivos de HTML, CSS y JavaScript para la visualización e interacción con el resultado del proyecto, que constituye una aplicación Web desarrollada con el *micro-framework* Flask. Para su desarrollo, al igual que el resto de archivos de programación del proyecto, se ha empleado el editor de código Sublime Text [82], y los navegadores Web Safari [83], Chrome [84] y Firefox [85].

4.4. Despliegue

Para desplegar el resultado final del proyecto software se ha optado por la plataforma como servicio (PaaS, por sus siglas en inglés) App Engine, de Google [86]. Esta solución abstrae la infraestructura necesaria para desplegar una aplicación Web, dejando en manos del desarrollador la configuración del entorno a nivel de aplicación, y permitiendo que este pueda alojarse en los servicios de computación en la nube de Google a bajo coste. App Engine es una de las alternativas estudiadas en este máster dentro de la asignatura "Infraestructura para el Big Data", y ha sido elegida por la experiencia previa del autor con ella y su conveniencia frente a otras alternativas. Sin embargo, como se abordará posteriormente, no se descarta una posible migración en el futuro a otra alternativa, dada la facilidad para hacerlo entre las soluciones de computación en la nube existentes. Por otra parte, el dominio Web del proyecto ha sido adquirido con la empresa Soluciones Corporativas IP [87].

4.5. Otras herramientas

Como herramienta para la creación de la memoria y anexos en LaTeX se ha empleado el portal en línea Overleaf [88] y Texmaker [89]. Los iconos empleados a lo largo del trabajo, el producto software o cualquier otro artefacto relacionado con este trabajo, y que no sean de fuente propia, están extraídos de Icons8 [90].

Aspectos relevantes del desarrollo del proyecto

Trabajos relacionados

Conclusiones y Líneas de trabajo futuras

Apéndices

Apéndice A

Plan de Proyecto Software

Apéndice B

Especificación de Requisitos

Apéndice C

Especificación de diseño

Apéndice D

Documentación técnica de programación

Apéndice E

Documentación de usuario

Bibliografía

[1] A. Woodie. Data prep still dominates data scientists' time, survey finds, julio 2020. [Internet; accedido el 15 de enero de 2023].

[2] J. A. Fernández. Teletrabajar desde la españa vaciada, mayo 2021. [Internet; accedido el 15 de enero de 2023].

[3] Terrenos.es. Teletrabajo en entorno rural, ¿el remedio contra la 'españa vaciada'?, febrero 2022. [Internet; accedido el 15 de enero de 2023].

[4] M. Regidor. Conquistando la españa vaciada gracias al teletrabajo, junio 2021. [Internet; accedido el 15 de enero de 2023].

[5] 20 Minutos | Europa Press. ¿mudarse a la españa vaciada? 27 pueblos forman la red nacional de pueblos acogedores para el teletrabajo, julio 2021. [Internet; accedido el 15 de enero de 2023].

[6] M. Castillo. Esta es la españa vaciada donde los funcionarios podrán teletrabajar un 90[Internet; accedido el 15 de enero de 2023].

[7] S. Delgado. España vaciada y reconexión digital: ¿hemos podido teletrabajar desde el pueblo?, junio 2022. [Internet; accedido el 15 de enero de 2023].

[8] Fundación Universidad de Burgos. Convocatoria prototipos orientados al mercado - plan tcue 2021-2023 curso 2021/2022: Acta de resultados de valoración, marzo 2022. [Internet; accedido el 16 de enero de 2023].

[9] Junta de Castilla y león. La junta concede los premios datos abiertos de castilla y león 2021 y reconoce los mejores proyectos que han reutilizado información pública autonómica, octubre 2021. [Internet; accedido el 16 de enero de 2023].

[10] G. García J. García. Repuéblame: Pueblos con calidad de vida, 2022. [Internet; accedido el 10 de diciembre de 2022].

[11] VanWoow. Apadrina un pueblo, 2022. [Internet; accedido el 10 de diciembre de 2022].

[12] Pueblos Mágicos de España. Pueblos mágicos de españa, 2022. [Internet; accedido el 10 de diciembre de 2022].

[13] Agencia Española de Protección de Datos. Protección de datos por defecto, noviembre 2021. [Internet; accedido el 31 de diciembre de 2022].

[14] Colaboradores de Wikipedia. Discretización - wikipedia, la enciclopedia libre, marzo 2022. [Internet; accedido el 16 de enero de 2023].

[15] Colaboradores de Wikipedia. Normalización (estadística) - wikipedia, la enciclopedia libre, octubre 2022. [Internet; accedido el 16 de enero de 2023].

[16] P. Rodó. Normalización estadística, julio 2019. [Internet; accedido el 16 de enero de 2023].

[17] Colaboradores de Wikipedia. Municipio - wikipedia, la enciclopedia libre, diciembre 2022. [Internet; accedido el 16 de enero de 2023].

[18] Colaboradores de Wikipedia. Entidad singular de población - wikipedia, la enciclopedia libre, noviembre 2022. [Internet; accedido el 16 de enero de 2023].

[19] Instituto Nacional de Estadística. Glosario de conceptos: Entidad singular, n.d. [Internet; accedido el 16 de enero de 2023].

[20] Colaboradores de Wikipedia. Datos abiertos - wikipedia, la enciclopedia libre, diciembre 2022. [Internet; accedido el 16 de enero de 2023].

[21] Datstrats. ¿es legal el scraping en españa?, (n.d.). [Internet; accedido el 15 de enero de 2023].

[22] Gomez-Acebo Pombo Abogados. Screen scraping: condiciones generales: de la contratación, bases de datos y competencia desleal, enero 2013. [Internet; accedido el 15 de enero de 2023].

[23] Colaboradores de Wikipedia. Geocodificación - wikipedia, la enciclopedia libre, diciembre 2022. [Internet; accedido el 16 de enero de 2023].

[24] Salesforce. ¿qué es paas? - descripción de plataforma como servicio, (n.d.). [Internet; accedido el 16 de enero de 2023].

[25] KDNuggets. Crisp-dm, still the top methodology for analytics, data mining, or data science projects, (n.d.). [Internet; accedido el 2 de enero de 2023].

[26] Inc. Doist. Todoist | una to-do list app para organizar trabajo y vida, (n.d.). [Internet; accedido el 15 de enero de 2023].

[27] Inc. GitHub. Github: Let's build from here, (n.d.). [Internet; accedido el 15 de enero de 2023].

[28] Open-Elevation. Open-elevation api, 2022. [Internet; accedido el 18 de diciembre de 2022].

[29] Ministerio de Sanidad. Centros, servicios y establecimientos sanitarios del sistema nacional de salud, 2022. [Internet; accedido el 18 de diciembre de 2022].

[30] Ministerio de Asuntos Económicos y Transformación Digital. Cobertura banda ancha españa 2021 fija y móvil, junio 2021. [Internet; accedido el 21 de diciembre de 2022].

[31] OpenWeather. Weather api - openweather, 2022. [Internet; accedido el 19 de diciembre de 2022].

[32] OpenTripPlanner. Opentripplanner basic tutorial: An open source multi-modal trip planner, 2022. [Internet; accedido el 18 de diciembre de 2022].

[33] contribuyentes de GeoPy K. Esmukov. Geopy • pypi, 2018. [Internet; accedido el 18 de diciembre de 2022].

[34] Wikipedia. Api:portada - mediawiki, 2022. [Internet; accedido el 22 de diciembre de 2022].

[35] Positionstack. Accurate forward reverse batch geocoding rest api, 2022. [Internet; accedido el 18 de diciembre de 2022].

[36] Geoapify. Maps, apis and components | geoapify location platform, 2022. [Internet; accedido el 21 de diciembre de 2022].

[37] Instituto Nacional de Estadística. Relación de municipios y códigos por provincias y comunidades autónomas, enero 2021. [Internet; accedido el 17 de diciembre de 2022].

[38] F. Ruiz. Lista de entidades singulares según el nomenclátor oficial, enero 2011. [Internet; accedido el 17 de diciembre de 2022].

[39] Instituto Geográfico Nacional. Nomenclátor geográfico de municipios y entidades de población, 2022. [Internet; accedido el 17 de diciembre de 2022].

[40] Ministerio de Educación y Formación Profesional. Registro estatal de centros docentes no universitarios (rcd), 2022. [Internet; accedido el 9 de diciembre de 2022].

[41] Adevinta. Infojobs - bolsa de trabajo, ofertas de empleo, 2022. [Internet; accedido el 10 de diciembre de 2022].

[42] Ministerio de Universidades. Registro de universidades, centros y títulos (ruct), 2022. [Internet; accedido el 9 de diciembre de 2022].

[43] Fotocasa. Fotocasa.es: Alquiler de pisos, compra y venta, 2022. [Internet; accedido el 19 de diciembre de 2022].

[44] Idealista. idealista — casas y pisos, alquiler y venta. anuncios gratis, 2022. [Internet; accedido el 18 de diciembre de 2022].

[45] Libretilla. Las 50 provincias de españa y sus capitales, febrero 2020. [Internet; accedido el 18 de diciembre de 2022].

[46] El Orden Mundial. El mapa físico de españa, enero 2022. [Internet; accedido el 18 de diciembre de 2022].

[47] Hacienda Navarra. Estadísticas impuesto sobre la renta de las personas físicas, 2020. [Internet; accedido el 18 de diciembre de 2022].

[48] Euskal Estatistika Erakundea/Instituto Vasco de Estadística. Renta familiar media de la c. a. de euskadi por ámbitos territoriales, según tipo de renta (euros). 2020, 2020. [Internet; accedido el 18 de diciembre de 2022].

[49] Euskal Estatistika Erakundea/Instituto Vasco de Estadística. Producto interior bruto (pib) de la c.a. de euskadi por ámbitos territoriales. 1996 - 2020, 2020. [Internet; accedido el 26 de diciembre de 2022].

[50] Agencia Tributaria. Posicionamiento de los municipios mayores de 1.000 habitantes por renta bruta media, 2019. [Internet; accedido el 18 de diciembre de 2022].

[51] I. Cantarino F. J. Gisbert. Rugosidad del terreno: Una característica del paisaje poco estudiada. enero 2010.

[52] Instituto Nacional de Estadística. Tasas de actividad, paro y empleo por provincia y sexo, 2022. [Internet; accedido el 20 de diciembre de 2022].

[53] Inc. Postman. Postman api platform, 2022. [Internet; accedido el 17 de diciembre de 2022].

[54] Molbiotools.com. Compare lists - multiple list comparator, 2022. [Internet; accedido el 20 de diciembre de 2022].

[55] S. N. Humbad. Count duplicates in a list online tool, octubre 2022. [Internet; accedido el 26 de diciembre de 2022].

[56] Flatfile team and contributors. Csv to json - csvjson, 2022. [Internet; accedido el 21 de diciembre de 2022].

[57] MapMaker. Mapmaker - plot coordinates, make advanced maps analyze geographic data, 2022. [Internet; accedido el 21 de diciembre de 2022].

[58] dCode. Duplicates in a list finder, 2022. [Internet; accedido el 10 de diciembre de 2022].

[59] Inc. Microsoft. Excel | microsoft 365, (n.d.). [Internet; accedido el 31 de diciembre de 2022].

[60] Inc. Apple. Numbers - apple (es), (n.d.). [Internet; accedido el 15 de enero de 2023].

[61] GSkinner. Regexr: Learn, build, test regex, 2022. [Internet; accedido el 19 de diciembre de 2022].

[62] Elite Cafemedia Tech. Alphabetical order, 2022. [Internet; accedido el 21 de diciembre de 2022].

[63] I. Bradley. Remove duplicates from list of lines, 2022. [Internet; accedido el 15 de diciembre de 2022].

[64] CodeBeautify. Remove empty lines, 2022. [Internet; accedido el 15 de diciembre de 2022].

[65] CircleCell. Jsonlint - the json validator, 2022. [Internet; accedido el 21 de diciembre de 2022].

[66] CodeBeautify. Json viewer, 2022. [Internet; accedido el 23 de de diciembre de 2022].

[67] L. Richardson. beautifulsoup4 • pypi, abril 2022. [Internet; accedido el 18 de diciembre de 2022].

[68] A. Ronacher. Flask • pypi, agosto 2022. [Internet; accedido el 26 de diciembre de 2022].

[69] A. Ronacher. Jinja2 • pypi, abril 2022. [Internet; accedido el 26 de diciembre de 2022].

[70] The Pandas Development Team. Pandas • pypi, noviembre 2022. [Internet; accedido el 15 de enero de 2023].

[71] S. Kumar. python-dotenv • pypi, septiembre 2022. [Internet; accedido el 18 de diciembre de 2022].

[72] Instituto Nacional de Ciberseguridad. Plan director de seguridad, 2022. [Internet; accedido el 12 de enero de 2023].

[73] The Open Web Application Security Project. The open web application security project - top 10 vulnerabilities 2021, 2022. [Internet; accedido el 12 de enero de 2023].

[74] V. Lucas. Phpdotenv, octubre 2022. [Internet; accedido el 9 de enero de 2023].

[75] Python Software Foundation. re — regular expression operations, (n.d.). [Internet; accedido el 22 de diciembre de 2022].

[76] K. Reitz. Requests • pypi, junio 2022. [Internet; accedido el 17 de diciembre de 2022].

[77] A. Mueller. Scikit-learn • pypi, diciembre 2022. [Internet; accedido el 27 de diciembre de 2022].

[78] N. Hug. Scikit-surprise • pypi, septiembre 2022. [Internet; accedido el 27 de diciembre de 2022].

[79] N. Hug. Surprise • pypi, enero 2017. [Internet; accedido el 27 de diciembre de 2022].

[80] Python Software Foundation. sys — parámetros y funciones específicos del sistema, (n.d.). [Internet; accedido el 17 de diciembre de 2022].

[81] J. Goldsmith. Wikipedia • pypi, noviembre 2014. [Internet; accedido el 9 de enero de 2023].

[82] Sublime HQ Pty Ltd. Sublime text - text editing, done right, (n.d.). [Internet; accedido el 17 de enero de 2023].

[83] Inc. Apple. Safari - apple (es), (n.d.). [Internet; accedido el 17 de enero de 2023].

[84] Inc. Google. Google chrome, (n.d.). [Internet; accedido el 17 de enero de 2023].

[85] Mozilla Foundation. Descarga navegador firefox - rápido, privado y gratis - de mozilla, (n.d.). [Internet; accedido el 17 de enero de 2023].

[86] Inc. Google. Plataforma de aplicaciones app engine | app engine | google cloud, (n.d.). [Internet; accedido el 17 de enero de 2023].

[87] Soluciones Corporativas IP. Dondominio | registro de dominios, hosting, correo y ssl, (n.d.). [Internet; accedido el 17 de enero de 2023].

[88] Overleaf. Online latex editor overleaf, (n.d.). [Internet; accedido el 17 de enero de 2023].

[89] P. Brachet. Texmaker (free cross-platform latex editor), (n.d.). [Internet; accedido el 17 de enero de 2023].

[90] Visualpharm. Free icons, clipart illustrations, photos and music, (n.d.). [Internet; accedido el 17 de enero de 2023].