

Proyecto de la asignatura

La Accesibilidad en Alicante

Adquisición y Preparación de Datos

Ingeniería en Inteligencia Artificial - UA

Integrantes

- Ignacio Mendoza Díaz
- Carlos Vidal Rodríguez
- Santiago Álvarez Geanta

Índice

1. Definición del proyecto centrado en los datos	3
1. Selección de la temática	3
2. Problema a resolver	3
3. Lienzo del problema	3
4. Objetivos del proyecto	4
5. Casos de uso	4
6. Métricas clave y evaluación	4
2. Analizar y evaluar necesidades de datos	5
2.1. Requisitos y naturaleza de los datos	5
2.2. Análisis de las fuentes disponibles	6
2.3. Reajuste de objetivos y filtrado	6
2.4. Selección final de conjuntos de datos	6
3. Realizar el diseño conceptual, lógico y físico del almacén de datos	7
3.1. Diseño conceptual	7
3.2. Diseño lógico	7
3.3. Diseño físico	8
4. Limpiar, transformar y normalizar los datos	10
4.1. Estandarización y Filtrado Mediante Pentaho Data Integration	10
4.1.1. Primera transformación	10
4.1.2. Segunda transformación	10
4.2. Feature Engineering, Enriquecimiento y Consolidación Mediante Python	11
4.2.1. Reducción de Redundancia y Feature Engineering Avanzado	11
4.2.2. Tratamiento de Outliers y Enriquecimiento de Datos	11
4.2.3. Fusión Final de los Datos	11
5. Transformar los datos	12
5.1. Análisis Semántico y Diseño del Modelo	12
5.2. Implementación y Enriquecimiento Programático	13
5.3. Validación del Grafo de Conocimiento	14
6. Visualización	14
6.1. Mapa por niveles de accesibilidad	14
6.2. Mapa de calor de accesibilidad	15
6.3. Mapa por agrupación de clusters	16

1. Definición del proyecto centrado en los datos

1. Selección de la temática

Para este proyecto hemos elegido la temática de Movilidad y Discapacidad, centrada específicamente en la accesibilidad urbana. Nos interesa analizar cómo las personas con movilidad reducida se desplazan por la ciudad y qué obstáculos encuentran en su día a día. La idea es aprovechar los datos disponibles para crear un sistema que permita visualizar, analizar y planificar recorridos y visitas a lugares accesibles, contribuyendo a una ciudad más inclusiva y sostenible.

2. Problema a resolver

El problema que nos planteamos es el siguiente:

¿Cómo podemos crear un almacén de datos que ayude a las personas de movilidad reducida a encontrar lugares accesibles en la ciudad?

La elección partió de la situación actual, donde la información sobre la accesibilidad de los edificios y espacios públicos suele estar dispersa y no siempre es precisa. Esto dificulta que las personas con movilidad reducida puedan planificar sus rutas de manera autónoma y segura. Con nuestro proyecto, pretendemos centralizar y estructurar estos datos para ofrecer una herramienta útil, práctica y confiable.

3. Lienzo del problema

Nuestro proyecto va dirigido principalmente a **personas con movilidad reducida**¹, como usuarios de sillas de ruedas o con problemas de movilidad. Organizaciones y asociaciones que promueven la inclusión y accesibilidad urbana. Autoridades locales y urbanistas interesados en mejorar la accesibilidad.

El reto principal es moverse de manera segura y eficiente por la ciudad, encontrando lugares que sean accesibles. Actualmente, la información sobre accesibilidad no está centralizada ni categorizada de forma clara.

Para ello deberemos tratar el origen del problema centrándonos en lugares públicos de la ciudad que forman parte de la vida cotidiana, como supermercados, estaciones, restaurantes, oficinas públicas, y espacios culturales. En el contexto del día a día, cuando las personas necesitan desplazarse sin asistencia externa.

Por tanto, nuestro objetivo es aumentar la autonomía y seguridad de las personas con movilidad reducida, permitiéndoles planificar sus rutas y actividades sin depender

¹ En este proyecto nos centramos en personas con movilidad reducida, no considerando casos como usuarios temporales con muletas o problemas transitorios.

constantemente de terceros. Además, pensamos que contribuir a la mejora de la accesibilidad urbana tiene un impacto positivo en la sociedad, fomentando la inclusión y la igualdad de oportunidades.

4. Objetivos del proyecto

- O1: Identificar los lugares accesibles. Recopilar y centralizar información sobre accesibilidad en distintos tipos de espacios públicos.
- O2: Clasificar las zonas de la ciudad. Analizar las áreas urbanas en función de su accesibilidad, identificando zonas totalmente accesibles, parcialmente accesibles o inaccesibles.
- O3: Ubicar los lugares accesibles. Permitir la visualización geográfica de los lugares y zonas accesibles para facilitar la planificación de rutas.

5. Casos de uso

En nuestro proyecto, imaginamos diversas situaciones en las que el almacén de datos puede resultar especialmente útil para las personas con movilidad reducida. Por ejemplo, cuando planificamos visitas a lugares públicos, como restaurantes, bibliotecas o estaciones de transporte, nuestra herramienta permitirá a los usuarios verificar previamente si dichos lugares son accesibles, evitando desplazamientos innecesarios o incómodos. De esta manera, fomentamos que la autonomía del usuario no dependa de la disponibilidad de asistencia externa.

Asimismo, consideramos el uso cotidiano en el día a día. Las personas con movilidad reducida necesitan poder organizar sus rutas para ir al trabajo, a estudiar o realizar gestiones, y nuestro proyecto les proporciona la información necesaria para hacerlo de forma segura y eficiente. En este sentido, la posibilidad de planificar recorridos sin depender de terceros se convierte en un factor clave para garantizar su independencia y confianza en el entorno urbano.

Otro caso relevante es aquel en el que los usuarios se encuentran en situaciones donde no disponen de ayuda externa. En estas circunstancias, nuestro sistema permite anticipar posibles obstáculos y seleccionar rutas que aseguren la accesibilidad, contribuyendo a que la persona pueda desenvolverse de manera autónoma. Finalmente, nuestro almacén de datos facilita la organización en el tiempo, optimizando desplazamientos y considerando tanto la ubicación de los lugares como su grado de accesibilidad, lo que ayuda a planificar las actividades diarias de manera más eficiente y segura.

6. Métricas clave y evaluación

Para poder evaluar de manera precisa el cumplimiento de los objetivos de nuestro proyecto, hemos definido un conjunto de métricas de accesibilidad que nos permiten clasificar los lugares de acuerdo con su nivel de adecuación para personas con movilidad reducida.

En primer lugar, definimos la **Accesibilidad Alta (4)**, que se aplica a aquellos lugares donde cualquier persona con movilidad reducida puede acceder de manera completamente autónoma, sin necesidad de asistencia externa ni adaptaciones especiales. A continuación, la **Accesibilidad Media (3)** corresponde a los lugares que requieren algún tipo de ayuda externa o mecánica, como rampas móviles, asistencia de otra persona o dispositivos específicos para superar barreras físicas.

La categoría de **Accesibilidad Baja (2)** engloba aquellos lugares donde el acceso es limitado y presenta obstáculos significativos, de modo que solo algunas personas podrían utilizarlos con dificultades o con ayuda considerable. Por su parte, la **Accesibilidad Muy Baja (1)** indica que el acceso es prácticamente imposible, incluso con asistencia, lo que representa una barrera total para los usuarios con movilidad reducida. Finalmente, incluimos la categoría **Sin Datos (0)** para aquellos casos en los que no se dispone de información suficiente para evaluar la accesibilidad, asegurando así que nuestro análisis reconozca las lagunas en la información recopilada.

Estas métricas no solo nos permiten evaluar de manera sistemática la efectividad de nuestro almacén de datos, sino que también nos proporcionan criterios claros para identificar zonas o lugares que requieren mejoras en accesibilidad. Además, facilitan la toma de decisiones informadas para garantizar que la información proporcionada cumpla con su objetivo principal: aumentar la autonomía y seguridad de las personas con movilidad reducida en la ciudad.

2. Analizar y evaluar necesidades de datos

Para dar respuesta al problema de la autonomía de las personas con movilidad reducida en Alicante, el proyecto requiere una base de datos que combine precisión geográfica con información detallada sobre barreras arquitectónicas. Tras una evaluación inicial, definimos que los datos necesarios debían permitir no solo ubicar un punto en el mapa, sino también categorizar de forma granular el grado de accesibilidad de este.

2.1. Requisitos y naturaleza de los datos

Para que el almacén de datos fuera funcional y permitiera los casos de uso definidos, como la planificación de rutas y la verificación de locales, establecimos que los datos fundamentales debían ser de tipo espacial y categórico. En cuanto a la tipología, necesitábamos coordenadas precisas y atributos que describen los niveles de accesibilidad de forma estandarizada. Se priorizaron formatos estructurados como **CSV** y salidas de **API** procesables para facilitar el flujo posterior de **ETL**. Un aspecto técnico relevante detectado en esta fase fue que la fuente oficial de la Generalitat utilizaba el sistema de referencia **EPSG:25830 (UTM)**, lo que supuso una necesidad técnica de transformación al sistema estándar **WGS84 (EPSG:4326)** utilizado por la mayoría de servicios de mapas modernos. Finalmente, se determinó que el contenido debía centrarse exclusivamente en la accesibilidad para usuarios de sillas de ruedas.

2.2. Análisis de las fuentes disponibles

Realizamos un estudio exhaustivo de las fuentes de datos reales para evitar el uso de datos ficticios, garantizando que la herramienta tuviera una utilidad práctica real. La fuente primaria seleccionada fue el **Portal de Datos Abiertos de la Generalitat Valenciana**, donde localizamos el dataset "**Estudio de accesibilidad universal en edificios públicos en la Comunidad Valenciana de titularidad municipal (2020)**". Esta fuente es crítica porque ofrece una evaluación técnica experta dividida en cinco ejes: entrada principal, itinerarios, atención al público, aseos y elementos de comunicación. Sin embargo, presentaba el reto de estar limitado a edificios de gestión municipal y contener mucha información en valenciano o con caracteres diacríticos que requerían normalización. Para cubrir lugares de la vida cotidiana que no son de titularidad pública, como comercios o restaurantes, recurrimos a **OpenStreetMap (OSM)** mediante la **Overpass API**. Implementamos una consulta técnica en el lenguaje **Overpass QL** para extraer nodos y áreas en Alicante con las etiquetas **wheelchair** y **toilets:wheelchair**. Como complemento, incorporamos la **API de Nominatim** como fuente de enriquecimiento secundaria para obtener información contextual como el barrio o el código postal.

2.3. Reajuste de objetivos y filtrado

Durante la fase de análisis, decidimos realizar dos ajustes estratégicos en el alcance del proyecto basados en la disponibilidad y utilidad de los datos. En primer lugar, optamos por una especialización en discapacidad física y gente en silla de ruedas, inicialmente consideramos incluir otros problemas de movilidad como la gente en muletas, pero los datos disponibles eran demasiado heterogéneos y habrían dispersado el foco del análisis. Decidimos especializarnos en accesibilidad para sillas de ruedas para ofrecer un sistema de métricas mucho más robusto. En segundo lugar, se procedió a la exclusión del transporte público. Aunque el objetivo inicial planteaba analizar transportes, el análisis de los datos del **TRAM** y los autobuses de Alicante mostró que la flota es casi 100% accesible por normativa. Al no aportar una variabilidad relevante al sistema de búsqueda de lugares, que es donde reside el verdadero obstáculo para el usuario, decidimos centrar nuestros esfuerzos en la accesibilidad de los edificios y puntos de interés.

2.4. Selección final de conjuntos de datos

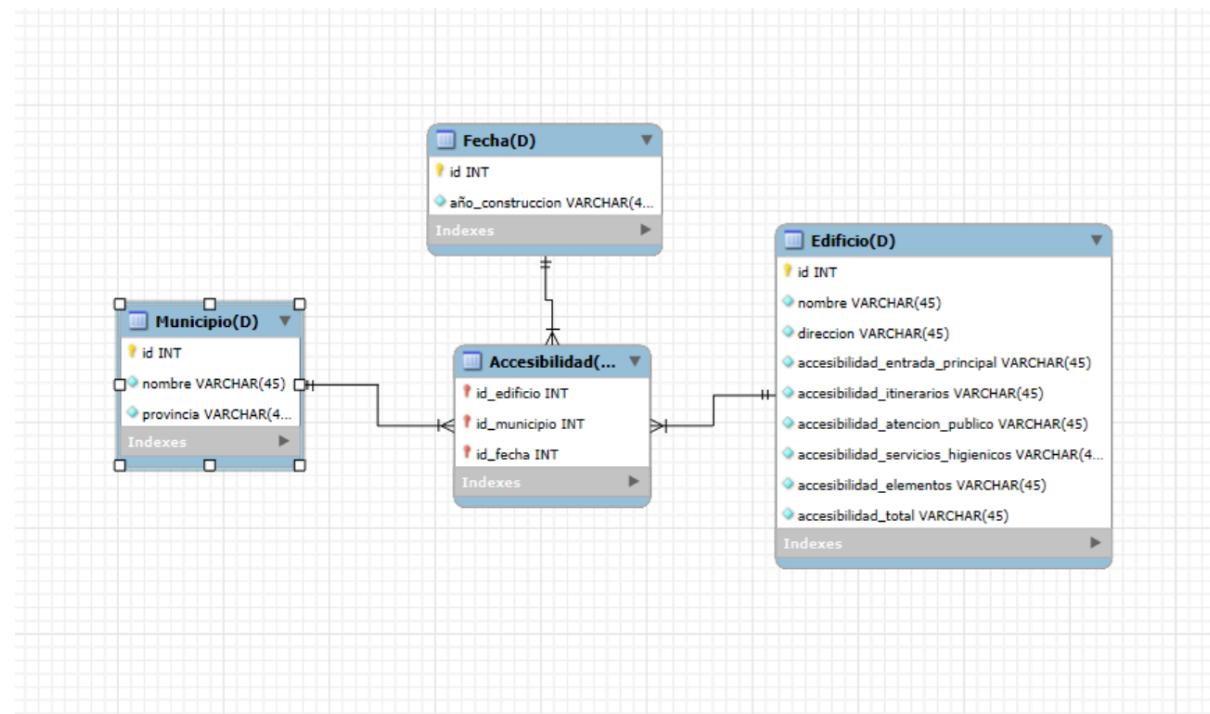
Finalmente, seleccionamos los conjuntos de datos de la **GVA** por su precisión técnica en infraestructuras públicas y los datos de **OpenStreetMap** por su diversidad de puntos de interés de uso cotidiano. Esta combinación se vio reforzada por el uso de la geocodificación inversa para dotar a cada registro de un contexto geográfico claro dentro de los barrios de Alicante. Todos los datos utilizados en este proyecto son 100% reales y han sido extraídos de fuentes oficiales y colaborativas. Este enfoque asegura que el almacén de datos resultante sea una base sólida para la creación de los grafos de conocimiento y las visualizaciones geográficas que permitan responder a las preguntas de investigación planteadas al inicio de este trabajo.

3. Realizar el diseño conceptual, lógico y físico del almacén de datos

Para la construcción de una solución robusta que permita analizar la accesibilidad urbana, hemos optado por un enfoque de modelado multidimensional. El objetivo principal es estructurar la información de manera que las consultas sobre niveles de accesibilidad, ubicaciones geográficas y tipos de edificios sean eficientes y claras. A continuación, se detalla la evolución de este diseño desde su concepción abstracta hasta su implementación en una base de datos relacional.

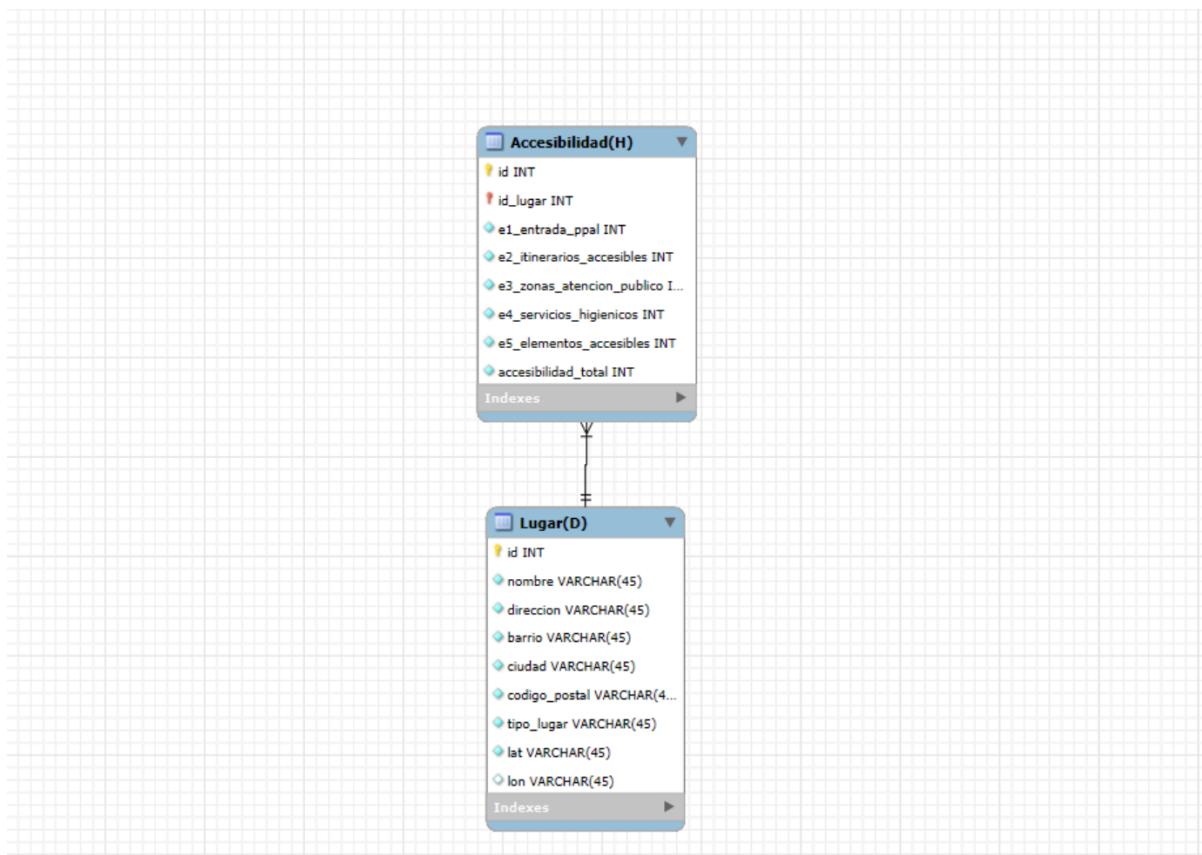
3.1. Diseño conceptual

El diseño conceptual constituye la primera aproximación al sistema, donde identificamos las entidades clave y los procesos de medida sin entrar en detalles técnicos de implementación. Para abordar la problemática de la accesibilidad en Alicante, hemos optado por un **modelo multidimensional basado en un esquema en estrella (Star Schema)**. El núcleo de este diseño es el **Hecho (Fact)** de la evaluación, que representa la medición técnica de un edificio. Este hecho se contextualiza mediante tres dimensiones principales: **Edificio/Lugar**, que identifica el sujeto de la evaluación; **Municipio**, que permite la segmentación geográfica y administrativa; y **Fecha**, centrada en el año de construcción. Esta separación es estratégica: nos permite analizar, por ejemplo, si existe una correlación entre la antigüedad de los edificios de un barrio concreto y su nivel de accesibilidad actual. El "grano" de nuestro almacén se define como una evaluación individual por cada edificio físico identificado, permitiendo una trazabilidad total de la información.



3.2. Diseño lógico

En la fase de diseño lógico, traducimos las entidades conceptuales a una estructura de tablas relacionales, definiendo sus atributos, tipos de datos lógicos y relaciones de integridad. Hemos consolidado el modelo en dos tablas principales interconectadas. La tabla de hechos, denominada **Accesibilidad(H)**, actúa como el repositorio de métricas cuantitativas; en ella, cada columna (desde **e1_entrada_ppal** hasta **e5_elementos_accesibles**) almacena un valor numérico que representa el nivel de adecuación técnica. Al utilizar tipos de datos **INT** para estas medidas, optimizamos el almacén para realizar operaciones de agregación y cálculo de medias. Por otro lado, la tabla **Lugar(D)** funciona como una dimensión descriptiva de movimiento lento que almacena atributos contextuales: el **nombre** del sitio, su **dirección** exacta, el **barrio** al que pertenece y su **tipo_lugar** (supermercado, oficina, etc.). Una decisión de diseño lógica fundamental fue la inclusión de los campos **lat** y **lon** en esta tabla, vinculando cada evaluación a una coordenada geográfica única. La relación se establece mediante una clave primaria en Lugar(D) que se propaga como clave foránea (**id_lugar**) en la tabla de hechos, garantizando que ninguna evaluación exista sin un lugar físico asociado.



3.3. Diseño físico

El diseño físico representa la implementación final del almacén en un sistema gestor de bases de datos real, en nuestro caso **MySQL**. Utilizando la herramienta **MySQL Workbench**, realizamos un proceso de **Forward Engineering** para generar el esquema físico a partir del modelo lógico. En esta etapa, afinamos la eficiencia del almacenamiento mediante la definición precisa de longitudes de campo, como el uso de **VARCHAR(45)** para nombres y direcciones, y el establecimiento de restricciones de **NOT NULL** para asegurar la calidad de los datos.

obligatorios. Un elemento crítico del diseño físico es la creación de **índices**, como el índice **fk_lugar_idx**, diseñado específicamente para acelerar las operaciones de *JOIN* entre la tabla de hechos y la dimensión de lugares durante las consultas de visualización. El motor de almacenamiento elegido ha sido **InnoDB**, lo que nos permite gestionar de forma segura las claves foráneas y asegurar la integridad referencial. El resultado de este proceso es un **script SQL** robusto que automatiza la creación de la infraestructura necesaria para albergar los datos ya transformados y normalizados, listos para ser explotados por los algoritmos y las herramientas de mapas.

```
-- MySQL Script generated by MySQL Workbench
-- Model: New Model    Version: 1.0
-- MySQL Workbench Forward Engineering

SET @OLD_UNIQUE_CHECKS=@UNIQUE_CHECKS, UNIQUE_CHECKS=0;
SET @OLD_FOREIGN_KEY_CHECKS=@FOREIGN_KEY_CHECKS, FOREIGN_KEY_CHECKS=0;
SET @OLD_SQL_MODE=@SQL_MODE,
SQL_MODE='ONLY_FULL_GROUP_BY,STRICT_TRANS_TABLES,NO_ZERO_IN_DATE,NO_ZERO_DATE,ERROR_FOR_DIVISION_BY_ZERO,
NO_ENGINE_SUBSTITUTION';

-- -----
-- Schema mydb
-----

-- -----
-- Schema mydb
-----

CREATE SCHEMA IF NOT EXISTS `mydb` DEFAULT CHARACTER SET utf8 ;
USE `mydb` ;

-- -----
-- Table `mydb`.`Lugar(D)`

CREATE TABLE IF NOT EXISTS `mydb`.`Lugar(D)` (
  `id` INT NOT NULL,
  `nombre` VARCHAR(45) NOT NULL,
  `direccion` VARCHAR(45) NOT NULL,
  `barrio` VARCHAR(45) NOT NULL,
  `ciudad` VARCHAR(45) NOT NULL,
  `codigo_postal` VARCHAR(45) NOT NULL,
  `tipo_lugar` VARCHAR(45) NOT NULL,
  `lat` VARCHAR(45) NOT NULL,
  `lon` VARCHAR(45) NULL,
  PRIMARY KEY (`id`)
)
ENGINE = InnoDB;

-- -----
-- Table `mydb`.`Accesibilidad(H)`

CREATE TABLE IF NOT EXISTS `mydb`.`Accesibilidad(H)` (
  `id` INT NOT NULL,
  `id_lugar` INT NOT NULL,
  `e1_entrada_ppal` INT NOT NULL,
  `e2_itinerarios_accesibles` INT NOT NULL,
  `e3_zonas_atencion_publico` INT NOT NULL,
  `e4_servicios_higienicos` INT NOT NULL,
  `e5_elementos_accesibles` INT NOT NULL,
  `accesibilidad_total` INT NOT NULL,
  PRIMARY KEY (`id`, `id_lugar`),
  INDEX `fk_lugar_idx` (`id_lugar` ASC) VISIBLE,
  CONSTRAINT `fk_lugar`
    FOREIGN KEY (`id_lugar`)
    REFERENCES `mydb`.`Lugar(D)` (`id`)
    ON DELETE NO ACTION
    ON UPDATE NO ACTION
)
ENGINE = InnoDB;

SET SQL_MODE=@OLD_SQL_MODE;
SET FOREIGN_KEY_CHECKS=@OLD_FOREIGN_KEY_CHECKS;
SET UNIQUE_CHECKS=@OLD_UNIQUE_CHECKS;
```

4. Limpiar, transformar y normalizar los datos

Nuestra estrategia adoptada para el tratamiento de datos se centra en la creación de un flujo de trabajo (ETL) robusto que integra dos herramientas: **Pentaho Data Integration (PDI)** para la manipulación y estandarización de grandes volúmenes de datos, y Python para la aplicación de lógicas complejas de Feature Engineering, enriquecimiento y validación de calidad. Este enfoque dual nos ha permitido maximizar la eficiencia en la unificación y preparación de los dos conjuntos de datos de accesibilidad de edificios.

4.1. Estandarización y Filtrado Mediante Pentaho Data Integration

Iniciamos nuestro proceso con la ingesta y el tratamiento preliminar de las dos fuentes de datos CSV mediante PDI, lo que nos garantiza una base estandarizada antes de la fusión.

4.1.1. Primera transformación

En la primera transformación, **transformation1.ktr**, nos concentraremos en la fuente de datos de la Comunidad Valenciana. El primer paso crítico consiste en la **conexión con la fuente de datos** y el **Filtrado de filas**, seleccionando exclusivamente aquellos registros donde la columna provincia coincidiera con "Alacant/Alicante". Tras este filtro geográfico, procedemos a la Selección de campos, **eliminando columnas** redundantes o con información marginal (p. ej., campos de coste y diversas variantes lingüísticas, mayormente valencianas, del nombre del municipio), lo que optimiza la estructura de nuestro dataset.

A continuación, implementamos una rigurosa **Corrección de errores y estandarización textual**. Mediante el paso String operations, convertimos las columnas clave (nombre, y todas las relacionadas con accesibilidad) a minúsculas para evitar inconsistencias en el matching. Esta normalización se complementa con una cascada de pasos Replace String, debido a que no encontramos otra alternativa más sencilla, para neutralizar los caracteres diacríticos, un factor común de ruido en datos multilingües. Sustituimos todas las vocales acentuadas (agudas y graves), así como caracteres especiales como la ‘ñ’ por ‘ny’ y los indicadores ordinales ‘º’ y ‘ª’, garantizando la máxima uniformidad textual entre los dos conjuntos de datos que se fusionaran.

4.1.2. Segunda transformación

La segunda transformación, **transformation2.ktr**, se destina a la fuente obtenida de *OpenStreetMap Overpass API*. Aquí, la tarea inicial ha sido el **Renombrado de campos** para alinear la nomenclatura técnica con la convención del primer conjunto de datos (p. ej., name se renombra a nombre, y wheelchair a accesibilidad_total). Acto seguido, aplicamos el mismo proceso de limpieza de caracteres que en la primera transformación. Un elemento diferenciador en esta etapa ha sido la **Codificación de variables categóricas iniciales**. Utilizando Replace String, mapeamos las etiquetas booleanas y limitadas de la fuente OSM (yes, no, limited) a categorías descriptivas más detalladas en español ("accesibilidad alta", "accesibilidad baja", "accesibilidad media"), facilitando la interpretación y la unificación de

los esquemas de accesibilidad. Finalmente, el paso **IfNull** nos ha permitido la Gestión de valores que faltan en campos cruciales, como nombre y las coordenadas, rellenando con cadenas vacías o valores predeterminados.

4.2. Feature Engineering, Enriquecimiento y Consolidación Mediante Python

Gracias a nuestro pipeline de PDI hemos generado dos archivos limpios que nos sirven de entrada para los scripts en Python, donde ejecutamos las tareas que requieren lógicas de programación más elaboradas, como la *geotransformación* y el *enriquecimiento de datos* (en este caso concreto nos referimos a la acción de añadir columnas).

4.2.1. Reducción de Redundancia y Feature Engineering Avanzado

En el script **transformation.py**, abordamos primeramente la **Eliminación o minimización de la redundancia de datos** mediante la identificación y eliminación de registros duplicados basados en el campo nombre. Esto es vital para asegurar la unicidad de las entidades.

Posteriormente, implementamos el Feature Engineering mediante la **Creación de nuevas características**. El aspecto más complejo encontrado fue la transformación de coordenadas. La columna *WKT* utiliza el sistema de referencia *EPSG:25830*, el cual debe ser transformado. Empleamos la librería **pyproj** para realizar una *geotransformación* precisa de las coordenadas a **latitud (lat)** y **longitud (lon)** en el sistema *WGS84 (EPSG:4326)*.

Adicionalmente, refinamos la Codificación de variables categóricas iniciada en PDI. Por una gestión de diseño decidimos convertir las categorías descriptivas de accesibilidad del primer dataset en una escala numérica de enteros del 1 a 4, realizando un **Label Encoding** (ej. accesibilidad muy baja, pasó a tener un valor de 1; accesibilidad alta, pasó a tener un valor de 4, etc.). Esto prepara nuestros datos para su uso en modelos de Machine Learning, análisis cuantitativos y mejor gestión para un almacén de datos.

4.2.2. Tratamiento de Outliers y Enriquecimiento de Datos

En cuanto a la calidad de los datos nos aseguramos de ella con el Tratamiento de los Outliers. Principalmente implementamos una validación rigurosa de las coordenadas, eliminando cualquier registro cuyas coordenadas estuvieran fuera de los límites geográficos esperados para la provincia de Alicante (Latitud entre 37.8° y 38.9°; Longitud entre -0.9° y 0.1°). Este paso mitiga el riesgo de introducir ruido espacial en nuestro análisis.

El punto culminante del proceso de transformación ha sido nuestro *Enriquecimiento de datos*. Hemos aprovechado las coordenadas geográficas (*lat*, *lon*) recién creadas para realizar una **Geocodificación Inversa** consultando la *API de Nominatim (OpenStreetMap)*. Gracias a esta consulta hemos podido crear cinco nuevas características: *dirección*, *barrio*, *ciudad*, *código postal* y *tipo de lugar*.

4.2.3. Fusión Final de los Datos

Finalmente, el script **combination.py** ejecuta la **Fusión de nuestras fuentes de datos** ya transformadas. Utilizamos la función *merge* de pandas con un tipo de unión *outer* en la columna clave nombre, lo que asegura la unión de todos los registros de ambas fuentes.

Además, implementamos la fusión utilizando el método `combine_first`, lo que nos permite priorizar los valores de la fuente de datos primaria sobre los de la fuente secundaria si es que todavía existen duplicados. Nuestro resultado final es un único dataset, totalmente integrado y normalizado, junto con una nueva columna `id`, para dotar a cada entidad de un identificador incremental único.

El conjunto de estas transformaciones, desde el filtrado inicial en PDI hasta el enriquecimiento geográfico en Python, y su orquestación mediante un flujo de trabajo (Job) de Pentaho, asegura que el dataset resultante esté en condiciones óptimas para su explotación y análisis subsiguiente.

5. Transformar los datos

El objetivo central de la transformación de nuestros datos, es transformar nuestro conjunto de datos combinado, en un grafo de conocimiento estructurado siguiendo los principios de los Datos Enlazados (Linked Data). Esto nos permite describir los datos no solo como valores aislados, sino como entidades relacionadas entre sí.

Para ello hemos adoptado el formato *RDF* (*Resource Description Framework*) estudiado en clase, para expresar los datos como **tripletas** (sujeto–predicado–objeto). Donde los sujetos y objetos son los nodos del grafo y los predicados definen las relaciones que los conectan.

Como herramienta hemos utilizado el lenguaje de programación Python, apoyándonos en la librería *rdflib*, y hemos empleado el vocabulario **Schema.org** como marco de modelado. Dado que proporciona clases y propiedades ampliamente aceptadas para describir ubicaciones físicas y características de forma estandarizada.

5.1. Análisis Semántico y Diseño del Modelo

Antes de llevar a cabo la transformación, realizamos un análisis del CSV final con el objetivo de definir un mapeo conceptual coherente. Dado que el dataset describe ubicaciones físicas con información geográfica, categórica y de accesibilidad, decidimos seleccionar la clase *schema:Place* como la entidad principal. Para construir el grafo, definimos tres espacios de nombres. El espacio **schema** se utiliza para el vocabulario estándar de Schema.org. El espacio **ex** se usa para las direcciones internas del proyecto. Y el espacio **wd** se utiliza para conectar nuestros datos con entidades de Wikidata.

En cuanto al diseño del mapeo, decidimos asociar cada columna del CSV a la propiedad semántica más adecuada de Schema.org. Los atributos de identificación y descripción básica, como nombre, dirección, lat, lon y código_postal, se mapearon directamente a propiedades canónicas como `schema:name`, `schema:address`, `schema:latitude`, `schema:longitude` y `schema:postalCode`.

Para la información de localización y accesibilidad, aplicamos mapeos más específicos: la ciudad se representó mediante `schema:addressLocality`, el barrio mediante `schema:addressRegion` y el tipo de lugar mediante `schema:category`. Este diseño permite mantener una correspondencia clara entre la estructura original del CSV y su representación semántica en el grafo.

5.2. Implementación y Enriquecimiento Programático

Nuestra transformación se ejecutó mediante la iteración sobre cada fila del CSV, generando un sujeto (recurso RDF) único por cada lugar a partir de su identificador (`ex:lugar_i`) y asignándole el tipo principal `schema:Place`. Este proceso garantiza que cada registro del dataset se convierta en una entidad identifiable de forma inequívoca dentro del grafo.

Para asegurar la integridad de los datos, se tiparon los literales generados de forma explícita, empleando tipos de datos adecuados como `xsd:float` para las coordenadas geográficas y `xsd:int` para el valor de accesibilidad global. Esta tipificación mejora la precisión semántica y facilita el posterior procesamiento de la información.

Un aspecto importante de esta fase fue agregar más información utilizando fuentes externas, como **Wikidata**. Creamos dos formas de conectar la información. Primero, mejoramos la ubicación geográfica enlazando cada lugar con su ciudad correspondiente en

```
- □ ×  
  
ex:lugar_42 a schema:Place ; # sujeto  
    schema:name "Biblioteca Pública" ; # propiedad  
    schema:containedInPlace wd:Q11975 . # enlace Wikidata
```

Wikidata, siempre que hubiera una conexión clara. Esto se hizo usando la propiedad `schema:containedInPlace`, lo que muestra claramente a qué lugar pertenece cada recurso.

En segundo lugar, aplicamos un enriquecimiento adicional al tipo de lugar. Para ello, utilizamos un mapeo similar que permitió enlazar la descripción categórica (tipo_lugar) con su entidad equivalente en Wikidata, utilizando la propiedad `schema:additionalType`. Esta doble estrategia de enlace proporciona contexto adicional y sitúa los recursos del proyecto dentro de un grafo de conocimiento global, sin alterar los valores originales del dataset.

Finalmente, para las métricas de accesibilidad granular, consolidamos los valores de las cinco columnas específicas (e1_entrada_ppal a e5_elementos_accesibles) en un único literal descriptivo. Este literal se asoció al recurso mediante la propiedad

schema:accessibilityFeature, lo que permitió preservar la información detallada de accesibilidad de forma agrupada y semánticamente coherente.

5.3. Validación del Grafo de Conocimiento

Al concluir el proceso de transformación de datos, serializamos el grafo en formato Turtle (.ttl) para su distribución y posterior consumo. La validación del resultado confirmó que la implementación fue exitosa: la transformación del dataset completo generó un total de 2,467 tripletas, cubriendo todas las propiedades de Schema.org diseñadas en el modelo e incluyendo los enlaces a Wikidata. Este volumen de registros y la verificación del esquema confirman que la información ha sido descrita correctamente y está lista para ser consultada como Linked Data.

6. Visualización

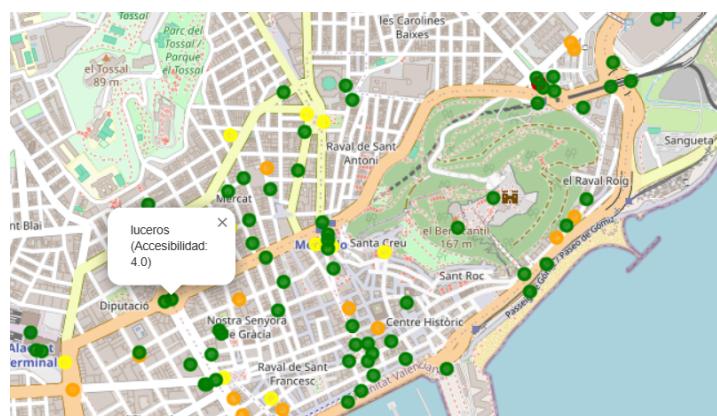
El objetivo de las visualizaciones en nuestro proyecto es facilitar la interpretación de los datos generados. Permitiendo de un solo vistazo que cualquier persona responder a las preguntas iniciales. Hemos implementado varias visualizaciones geográficas usando Python y la librería Folium, que nos ayuda a representar información espacial a partir de las coordenadas geográficas (latitud y longitud) de cada lugar.

Estas representaciones visuales permiten analizar la accesibilidad tanto a nivel individual de cada lugar como de forma global, identificando patrones y zonas con distintos niveles de adecuación para personas con movilidad reducida.

6.1. Mapa por niveles de accesibilidad

Esta primera visualización consiste en un mapa interactivo que muestra todos los lugares con un punto de color cada zona según su nivel de accesibilidad total. Donde:

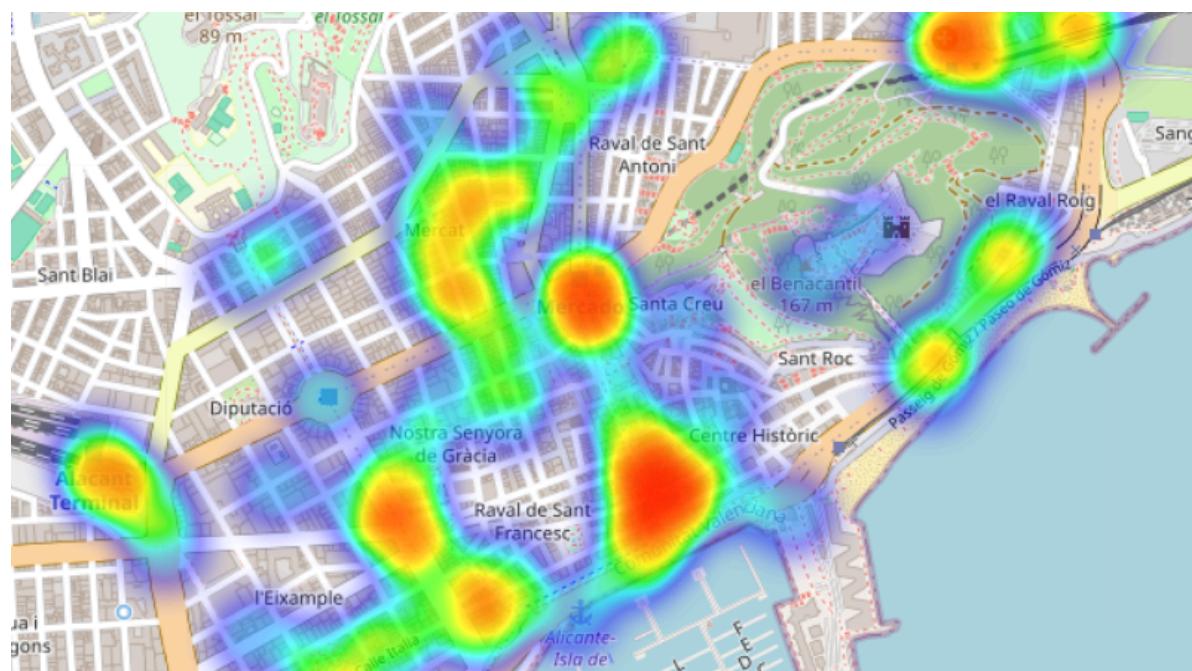
- Azul: 0 (nada accesible)
- Rojo: 1
- Naranja: 2
- Amarillo: 3
- Verde: 4 (máximo)



Adicionalmente en la leyenda tenemos la opción de activar o desactivar cada nivel de accesibilidad. Consiguiendo menor ruido visual y navegar más rápidamente. Esta funcionalidad resulta especialmente útil para centrarse en los niveles de accesibilidad más relevantes y comparar visualmente la distribución de los lugares accesibles dentro de la ciudad.

6.2. Mapa de calor de accesibilidad

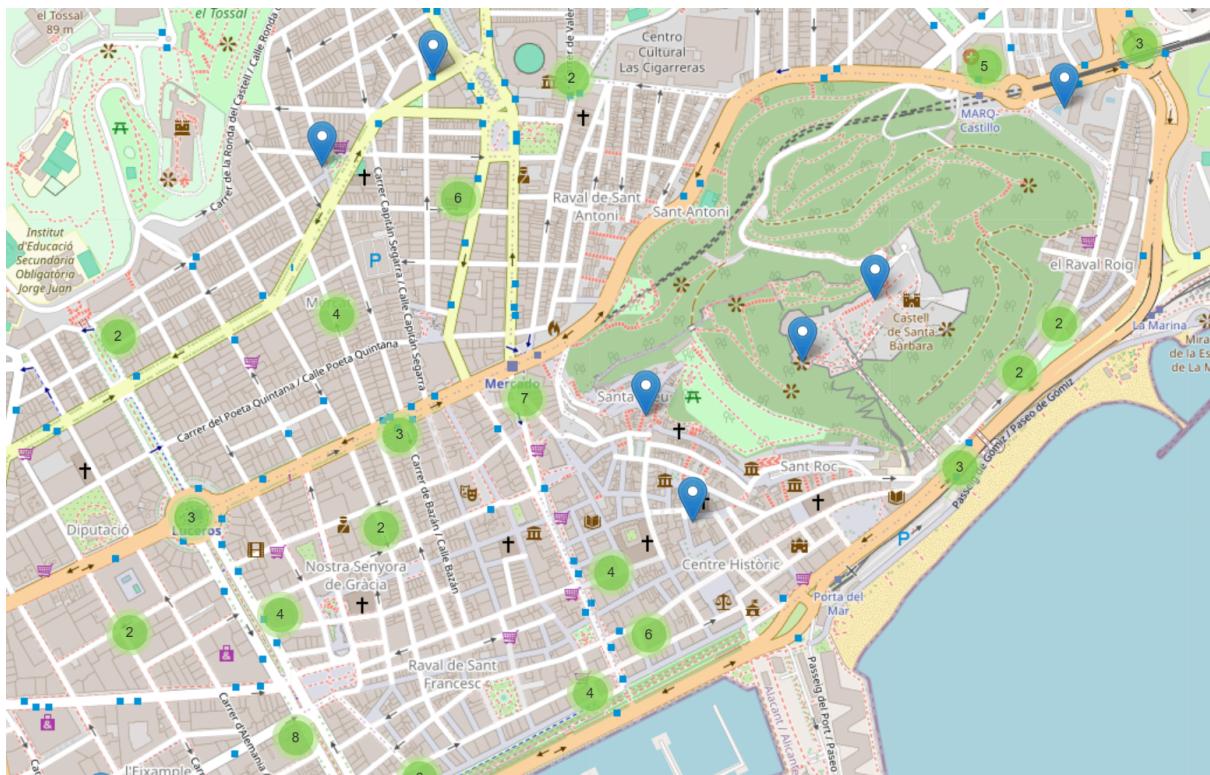
La segunda visualización que hemos implementado es un mapa de calor (heatmap) para representar la densidad de la accesibilidad en la provincia. Con el número y posición de los lugares accesibles generamos una superficie continua que resalta las áreas con mayor concentración de estos lugares. De esta forma, no solo se visualizan puntos individuales, sino que se obtiene una visión agregada que permite identificar patrones espaciales de accesibilidad en el entorno urbano



Este mapa es especialmente útil para analizar la accesibilidad desde una perspectiva global y detectar posibles áreas prioritarias de mejora. Facilitando la comparación entre distintas zonas de la ciudad, permitiendo observar de manera intuitiva qué áreas concentran una mayor oferta de espacios accesibles y cuáles presentan una menor cobertura.

6.3. Mapa por agrupación de clusters

La última visualización consiste en un mapa de clusters. En este tipo de visualización, los puntos cercanos entre sí se agrupan automáticamente en un único marcador cuando el nivel de zoom es bajo. A medida que el usuario aumenta el nivel de zoom, los clústeres se van desagregando progresivamente, permitiendo visualizar los lugares individuales de forma más detallada. Este comportamiento facilita la exploración interactiva del mapa y evita la saturación visual que se produciría al mostrar todos los marcadores simultáneamente.



Cada marcador individual incluye información básica del lugar, como su nombre y su nivel de accesibilidad, lo que permite acceder rápidamente a los datos relevantes. ofreciendo una forma más cómoda de explorar el conjunto completo de lugares accesibles distribuidos por la ciudad.