2023 - 2024

ACTIVIDAD GUIADA 1

Máster en Big Data y Data Science

05MBID - Minería de datos

Nombre: **Antonio Manuel Palma Bautista**

Fecha: **11 de junio de 2023**

Curso 2023 - Ed. Abril

viu Universidad Internacional de Valencia

1. Introducción, motivación y objetivo

El contexto del que se partirá es el siguiente:

> *Una empresa de gestión de viviendas de alquiler turístico, sin presencia en Andalucía, quiere introducirse en la región invirtiendo en algunos municipios.*

Y el problema a resolver:

> *Quiere mejorar su conocimiento del mercado turístico en Andalucía.*

El objetivo específico que se busca es obtener es:

> *Elaborar un perfil del sector de las viviendas de alquiler turístico de cada municipio de la región, para comparar unos perfiles con otros y así detectar oportunidades de negocio.*
>
> *Se pretende obtener como resultado final cuadros y gráficos que sirvan para que los directivos de la empresa, ajenos al mundo del análisis de datos, puedan entender la situación del mercado y comprender los patrones y tendencias que se siguen en él, identificando también aquellos municipios que se desvían de esos patrones (lo cual puede suponer la existencia de un nicho de mercado en esos municipios).*

2. Fuente de datos y selección

Los datos que se van a utilizar provienen de OpenRTA, el Registro de Turismo de Andalucía.

Es un registro de acceso público disponible en el Portal de Datos Abiertos de la Junta de Andalucía en https://www.juntadeandalucia.es/datosabiertos/portal/dataset/openrta, en el que se encuentra disponible la información más completa y actualizada sobre la oferta de todos los servicios turísticos en Andalucía, entre otros las viviendas con fines turísticos, establecimientos y servicios turísticos (por ejemplo, los hoteles o las empresas de turismo activo).

La inscripción, modificación y cancelación de los datos del registro se realiza en línea por los propios usuarios, debiendo identificarse mediante un certificado digital para poder hacer la gestión.

La tabla de datos que se va a utilizar se corresponde con el dataset completo en formato CSV de la OpenRTA, descargado comprimido en ZIP el día 11 de junio de 2023 y con datos actualizados a 25 de mayo de 2023 (según la fecha del fichero all-openRTA_csv.csv contenido en el interior del fichero comprimido). Está publicada bajo la licencia Reconocimiento 4.0 Internacional (CC BY 4.0) de Creative Commons, aprobada para "obras culturales libres".

En la tabla, cada registro representa un establecimiento o servicio turístico en Andalucía.

Cuenta con un total de 140.009 registros, cada uno de ellos con 76 atributos algunos numéricos (enteros, fechas o teléfonos, entre otros) y otros categóricos (por ejemplo elementos de una lista, textos libres, códigos postales o DNI).

Los atributos pueden ser agrupados en varias categorías:

- Atributos identificativos del establecimiento
- Atributos de contacto del establecimiento (con sus datos de contacto)
- Atributos exclusivos de algún (o algunos) tipo concreto de establecimiento
- Atributos de localización geográfica del establecimiento (coordenadas)
- Atributos sobre el seguro obligatorio del establecimiento
- Atributos identificativos del titular del establecimiento

Por el alto número de atributos de la tabla (y porque no se van a utilizar), se obvía el detalle de los atributos identificativos del titular del establecimiento y de los exclusivos de campamentos de turismo, oficinas de turismo u otros tipos de establecimiento que tampoco serán estudiados aquí.

A continuación se detallan las características del resto de atributos (los problemas que presentan se analizan al final del listado):

Atributos identificativos del establecimiento

- **RN** (Número de registro). Tipo: Numérico, con valores enteros
- **ID** (Identificador de registro). Tipo: Numérico, con valores enteros

- **COD_REGISTRO** (Código de registro). Tipo: Categórico, con valores códigos alfanuméricos de la forma [Id.Tipo]/[Id.Provincia]/[NºRegistro], por ejemplo CR/CO/00238 para una casa rural en Córdoba o VFT/MA/22372 para una vivienda con fines turísticos en Málaga
- **NOMBRE** (Nombre del establecimiento). Tipo: Categórico, con valores de texto libre
- **DESCRIPCION** (Descripción del establecimiento). Tipo: Categórico, con valores de texto libre
- **FEC_INSCRIPCION** (Fecha de inscripción del establecimiento en el registro de turismo). Tipo: Numérico, con valores enteros
- **FEC_INICIO_ACTIVIDAD** (Fecha de inicio de actividad establecimiento). Tipo: Numérico, con valores enteros
- **TIPO_OBJETO** (Tipo del establecimiento). Tipo: Categórico, con valores de una lista: Vivienda con fines turísticos, establecimiento hotelero, apartamento turístico, campamento turístico (camping), oficina de turismo,...
- **TIPO_OBJETO_ID** (Identificador del tipo de establecimiento). Tipo: Numérico, con valores enteros
- **GRUPO** (Se especifica si el servicio pertenece a una unidad mayor). Tipo: Categórico, con valores de una lista: Conjunto, Edificio, Hotel,…
- **CATEGORIA** (Categoría del establecimiento). Tipo: Categórico, con valores de una lista: 1 Estrella, 2 Estrellas, 1 Llave,…
- **MODALIDAD** (Modalidad del establecimiento). Tipo: Categórico, con valores de una lista: Playa, Ciudad, Rural, Carretera
- **ESPECIALIDADES** (Especialidades del servicio). Tipo: Categórico, con valores de una lista: Agroturismo, Familiar, Albergue,...

Atributos de contacto del establecimiento

- **DOMICILIO_ESTAB** (Domicilio del establecimiento, calle y número). Tipo: Categórico, con valores de texto libre
- **CODIGO_POSTAL** (Código postal del establecimiento). Tipo: Numérico, con valores enteros
- **LOCALIDAD** (Localidad del establecimiento). Tipo: Numérico, con valores de texto libre
- **ID_MUNICIPIO** (Identificador del municipio). Tipo: Numérico, con valores de texto libre
- **MUNICIPIO** (Municipio del establecimiento). Tipo: Categórico, con valores de texto libre
- **ID_PROVINCIA** (Identificador de la provincia). Tipo: Numérico, con valores
- **PROVINCIA** (Provincia del establecimiento). Tipo: Categórico, con valores de texto libre
- **TELEFONO** (Número de teléfono del establecimiento). Tipo: Numérico, con valores enteros

- **MOVIL** (Número de móvil del establecimiento). Tipo: Numérico, con valores enteros
- **FAX** (Número de fax del establecimiento). Tipo: Numérico, con valores enteros
- **CORREO_ELECTRONICO** (Dirección de correo electrónico del establecimiento). Tipo: Categórico, con valores de texto libre

Atributos exclusivos de los establecimientos con alojamientos

- **TOT_GEN_UA** (Número total de habitaciones, apartamentos, viviendas o alojamientos del establecimiento). Tipo: Numérico, con valores enteros
- **TOT_GEN_PLAZAS** (Número total de plazas del establecimiento). Tipo: Numérico, con valores enteros
- **IND_TIPO_ALQUILER** (Establecimiento que se alquila por completo, C, o por habitaciones, H). Tipo: Categórico, con valores de una lista: C, H

Atributos de localización geográfica del establecimiento

- **COORD_X** (Coordenada X de la localización del establecimiento). Tipo: Numérico, con valores en coma flotante
- **COORD_Y** (Coordenada Y de la localización del establecimiento). Tipo: Numérico, con valores en coma flotante
- **SRID** (Sistema de coordenadas en que están expresadas las coordenadas del establecimiento). Tipo: Numérico, con valores enteros

Los atributos que no se detallan son:

- **exclusivos de algún (o algunos) tipo concreto de establecimiento:**

  IND_COMPARTIDA
  ACTIVIDADES_TURISMO_ACTIVO
  IND_USO_PRIVADO
  NUM_PARCELAS_ACAMPADA
  NP_PARCELAS_ACAMPADA
  SUP_ZONA_ACAMPADA
  NUM_INSTALACIONES_FIJAS
  NP_INSTALACIONES_FIJAS
  SUPERFICIE_INSTALACIONES_FIJAS
  TOTAL_PLAZAS_CAMPAMENTO
  CAPACIDAD_MAXIMA_CAMPAMENTO
  IND_OFICINA_TUR_INTEGRADA_RED
  TITULARIDAD_OFICINA_TURISMO
  PADRE_ID
  IND_ESPECIFICO_ZONAL
  IND_FIJO_MOVIL
  IND_ON_LINE
  URL
  IDIOMAS

- **sobre el seguro obligatorio del establecimiento:**

  FEC_PRESENT_RTADSEG
  ESTADO_PRESENT_RTADSEG
  FEC_VERIF_RTADSEG
  ESTADO_VERIF_RTADSEG
  RESULT_VERIF_RTADSEG

- **identificativos del titular del establecimiento:**

  ID_TIPOVIA
  COD_VIA
  NOMBRE_VIA
  KM
  CALIF_NUM
  BLOQUE
  PORTAL
  ESCALERA
  PISO
  PUERTA
  REF_CATASTRAL
  IND_PUB_OPEN_RTA
  COMPLEMENTODOM
  ID_NUCLEO
  KM_NUM
  TIPO_NUMERACION
  NUM_DOC_IDENTIFICATIVO
  TITULAR
  GRUPO_ID
  CATEGORIA_ID
  MODALIDAD_ID
  LISTA_ESPEC

Problemas que presentan los atributos

El origen de los datos es un registro público en el que los campos del formulario de entrada no están limitados a unos valores concretos (o al menos no estaban limitados cuando se introdujeron algunos de los datos del registro), lo que conlleva que muchos de los atributos presenten problemas que hubieran sido fácilmente evitables en el origen limitando la entrada a una lista de elementos o a unos valores con un formato específico, o haciendo una mínima validación de los datos antes de cargarlos en el Registro.

En concreto, algunos de los problemas que se presentan son:

- Datos erróneos en atributos (por ejemplo, texto en atributos numéricos, correos electrónicos sin el carácter “@” o códigos postales con solo una o dos cifras en lugar de los cuatro o cinco dígitos habituales)
- Valores intercambiados entre atributos (por ejemplo, correos electrónicos en el campo TELEFONO o en el campo TOT_GEN_PLAZAS, o un municipio en el campo PROVINCIA)
- Datos erróneos en la mayoría de los atributos con valores de una lista (algunos registros tienen datos fuera de los elementos de la lista de valores posibles)

- Fechas en FEC_INSCRIPCION y FEC_INICIO_ACTIVIDAD sin formato fecha, escritas como entero (por ejemplo, la fecha 30/08/2016 aparece almacenada como “20160830”)
- Números de teléfono y fax con seis y siete dígitos (números introducidos sin el prefijo telefónico provincial, hasta hace unos años no obligatorio pero ya incluido al inicio del propio número)
- Números de móvil con menos de nueve dígitos
- Valores mal formateados en COD_REGISTRO
- Coordenadas dentro del atributo SRID (identificador del sistema de coordenadas)

3. Preparación, limpieza y transformación

Como ya se adelantó en el apartado anterior, se decide quitar los atributos identificativos del titular del establecimiento, porque son atributos que no se incluirán en el modelo.

Ocurre de igual manera con los atributos exclusivos de campamentos de turismo u oficinas de turismo, por lo que tampoco se incluyen. Además, en este caso y por la propia naturaleza de los atributos, aparecen sin datos en la mayoría de registros.

También se eliminan los atributos sobre el seguro obligatorio del establecimiento. Tampoco se incluyen en el modelo.

Para mejorar los datos, se convertirá a formato fecha los valores de los atributos FEC_INSCRIPCION y FEC_INICIO_ACTIVIDAD.

También se podrían corregir los números de teléfono y de fax que aparecen almacenados con seis y siete dígitos añadiéndoles al principio su prefijo provincial, para que pasen a tener los nueve dígitos que tienen en la actualidad todos los números de teléfono o fax en España.

Además, en los atributos críticos para el análisis, se realizarán las siguientes acciones:

- Se eliminarán todos los registros con atributo PROVINCIA diferente a ALMERÍA, CÁDIZ, CÓRDOBA, GRANADA, HUELVA, JAÉN, MÁLAGA o SEVILLA (se ha verificado que en este atributo no existen las variantes de estos valores sin acento)
- Se eliminarán todos los registros con valores no enteros en los atributos TOT_GEN_UA (total de habitaciones o alojamientos) y TOT_GEN_PLAZAS (total de plazas).
- Se eliminarán todos los registros con outliers en los atributos TOT_GEN_UA (total de habitaciones o alojamientos) y TOT_GEN_PLAZAS (total de plazas). Analizando los datos, se considerarán outliers los datos mayores a 4.000 habitaciones u 8.000 plazas.
- Se eliminarán todos los registros con atributo MUNICIPIO con valor entero (es decir, que sean un número en lugar de texto)

Respecto a los atributos con datos faltantes:

- Si falta el dato en el atributo TOT_GEN_UA (total de habitaciones o alojamientos) pero sí existe el dato en TOT_GEN_PLAZAS (total de plazas) para ese mismo registro, se insertará como valor “número de habitaciones” un valor igual a la mitad del “número de plazas”, considerándose que de media las habitaciones son de dos plazas.
- De forma análoga, si falta el dato en el atributo TOT_GEN_PLAZAS (total de plazas) pero sí existe el dato en TOT_GEN_UA (total de habitaciones o alojamientos) para ese mismo registro, se insertará como valor “número de plazas” un valor igual al doble del “número de habitaciones”, volviendo a considerarse que de media las habitaciones son de dos plazas.

- En los registros en los que falta el dato PROVINCIA también faltan los datos que nos podrían informar sobre la provincia (también falta, por ejemplo, el municipio, el identificador de la provincia y el código postal), por lo que no se puede recuperar ese dato.
- Se eliminarán todos los registros con atributo MUNICIPIO sin valor.

En los demás atributos no afecta que falte algún dato, como por ejemplo ocurre con el número de fax o el correo electrónico.

Dado el alto número de registros que tenemos, y que estas eliminaciones afectarán a un porcentaje pequeño del total de registros, la eliminación de los registros propuestos no afectará en gran medida a la calidad final del proceso.

4. Data mining y utilización del conocimiento

El modelo que se utilizará para realizar este proceso KDD descriptivo será un agrupamiento o clustering, pretendiéndose obtener grupos a partir de los datos para encontrar patrones en la distribución de las viviendas de alquiler turístico en los diferentes municipios en función de diferentes atributos.

En un primer momento se utilizaría un algoritmo de agrupamiento K-medias, al ser el algoritmo de aprendizaje no supervisado más simple.

En caso de que los datos no siguieran un formato circular y por lo tanto no se agruparan correctamente con un K-medias, se utilizaría un algoritmo de mezcla gaussiana (ya que es un algoritmo que no necesita datos con forma circular para que funcione bien).

El algoritmo de mezcla gaussiana calcula la probabilidad de que un punto de datos pertenezca a una distribución gaussiana específica y ése es el grupo en el que se ubicará.

Como resultado final, para utilizar el conocimiento se pretende obtener cuadros y gráficos que describan los escenarios estudiados según las tipologías y condicionantes planteados.