Tipología y ciclo de vida de los datos

Prueba de Evaluación Continua (PRACTICA2: Limpieza de Datos)

Alumno: Cèsar Comas Solé

7 de diciembre de 2018

Índice de contenidos

1 Enunciado

El objetivo de esta actividad será el tratamiento de un dataset, que puede ser el creado en la práctica 1 o bien cualquier dataset libre disponible en Kaggle (https://www.kaggle.com).

Algunos ejemplos de dataset con los que podéis trabajar son:

Red Wine Quality (https://www.kaggle.com/uciml/red-wine-quality-cortez-et-al-2009)

Titanic: Machine Learning from Disaster (https://www.kaggle.com/c/titanic)

El último ejemplo corresponde a una competición activa de Kaggle de manera que, opcionalmente, podéis aprovechar el trabajo realizado durante la práctica para entrar en esta competición.

Siguiendo las principales etapas de un proyecto analítico, las diferentes tareas a realizar (y justificar) son las siguientes:

1. Descripción del dataset. ¿Por qué es importante y qué pregunta/problema pretende responder?
2. Integración y selección de los datos de interés a analizar.
3. Limpieza de los datos.
    i. ¿Los datos contienen ceros o elementos vacíos? ¿Cómo gestionarías cada uno de estos casos?
    ii. Identificación y tratamiento de valores extremos.
4. Análisis de los datos.
    i. Selección de los grupos de datos que se quieren analizar/comparar (planificación de los análisis a aplicar).
    ii. Comprobación de la normalidad y homogeneidad de la varianza.
    iii. Aplicación de pruebas estadísticas para comparar los grupos de datos. En función de los datos y el objetivo del estudio, aplicar pruebas de contraste de hipótesis, correlaciones, regresiones, etc.
5. Representación de los resultados a partir de tablas y gráficas.
6. Resolución del problema. A partir de los resultados obtenidos, ¿cuáles son las conclusiones? ¿Los resultados permiten responder al problema?
7. Código: Hay que adjuntar el código, preferiblemente en R, con el que se ha realizado la limpieza, análisis y representación de los datos. Si lo preferís, también podéis trabajar en Python.

2 Respuesta

2.1 Proceso general de análisis

Para dar respuesta a esta práctica, seguiremos el siguiente flujo de trabajo:

1. Contexto
2. Definición del problema y el objetivo del análisis
3. Descripción del conjunto de datos
4. Selección tecnológica
5. Preparación de datos
    I. Análisis de conjunto
    II. Anonimización
    III. Limpieza (selección de atributos, vacíos, nulos, incompletos, duplicados, 'outliers', normalización de nombres, categorización, etc.)
    IV. Integración
    V. Reducción
    VI. Transformación
6. Análisis de datos
7. Presentación gráfica de resultados
8. Conclusiones

2.2 Contexto

El presente análisis se realiza en el marco de las prácticas de la asignatura de Tipología y ciclo de Vida de los Datos del Master de Data Science de la UOC.

Tiene como objetivo principal la ejecución de un proceso analítico de datos.

El repositorio de GitHub se encuentra en:

https://github.com/ccomassole/wine-review-data-analysis

2.3 Definición del problema y objetivo del análisis

Se pretende identificar si existe alguna correspondencia entre el precio de un vino y algunos de sus atributos (valoración y descripción dada por someliers)

Caso de descubrirse algún patrón en dichas descripciones, este podría utilizarse para desarrollar una herramienta la cual, a partir de una cata ciega previa, ayude a posicionar en precio de un nuevo vino en el mercado.

Los datos que permitirán realizar el análisis han sido recopilados por **Zack Thoutt** [1] mediante técnicas de 'scraping' de la web "**WineEnthusiast**" [2], y publicados en **Kaggle** a través del siguiente enlace [3].

2.4 Descripción del conjunto de datos.

2.4.1 Fuentes, origen, propiedad, formato y resumen de contenidos

Número de datasets: 1

Dataset 1:

- Origen de datos: Kaggle
- Título: Wine reviews
- Nombre: "winemag-data-130k-v2.csv"
- Volumen: 51.669 KB
- Autor: Zack Thoutt
- Fecha de la última actualización: Noviembre, 2017
- Enlace: https://www.kaggle.com/zynicide/wine-reviews
- Formato: csv
- Número de registros:
- Resumen de contenidos:
    - Fichero agregado de datos generado mediante técnicas de 'scraping' de la web 'WineEnthusiast' [2] que contiene descripciones de diferentes características de 130k vinos (país, región, variedad, bodega, precio, etc.) así como la descripción cualitativa de su cata realizada por maestros 'sommeliers' y la valoración que de los mismos se otorga desde la web por parte de sus usuarios.

2.4.2 Diccionario de características/atributos

El conjunto de datos del Dataset 1 se describe a través de **14** atributos:

- **id**: número identificador único de cada registro.
- **country**: país de origen del vino.
- **description**: algunas frases de un 'sommelier' que describe el sabor, olor, aspecto, sensación, y otros atributos del vino.
- **designation**: el viñedo particular dentro de la bodega de donde provienen las uvas que permitieron elaborar el vino.
- **points**: la cantidad de puntos de calificación del vino, en una escala de 1 a 100, otorgada en 'WineEnthusiast'. En la web indican que solo publican reseñas de vinos con puntajes superiores o iguales a 80.
- **price**: el precio de una botella de vino.
- **province**: la provincia o estado de origen del vino.
- **region_1**: el área de cultivo de vino en una provincia o estado (por ejemplo: Napa)

- region_2: a veces hay regiones más específicas dentro de un área de cultivo de vino (por ejemplo: Rutherford dentro del Valle de Napa).
- taster_name: nombre del sommelier que realizó y revisó la cata del vino.
- taster_twitter_handle: hashtag de Twitter del 'sommelier'.
- title: el título de la reseña del vino, que a menudo contiene también información acerca de la cosecha.
- variety: el tipo de uvas utilizadas para hacer el vino (por ejemplo: Pinot Noir).
- winery: la bodega que elaboró el vino.

2.5 Selección tecnológica.

La siguiente es una síntesis de posibles tecnologías usables en el contexto de la preparación y análisis de datos:

- Tecnologías de Bases de Datos
  - BBDD Estructuradas
  - BBDD No Estructuradas
  - BBDD orientadas a grafos
  - BigData
- Aplicaciones propietarias para la preparación de datos
  - Pentaho
  - Tableau Prep
  - Knime
  - OpenRefine
  - Trifacta
- Programación a medida:
  - R
  - Python

La selección de la tecnología a emplear dependerá, en general, de los siguientes factores:

1. volumen y características del dataset
2. requisitos impuestos por el cliente
3. requisitos asociados a nuestras capacidades y conocimientos.

En el caso particular del presente análisis, el pequeño tamaño del dataset, su formato y los requisitos impuestos a la entrega, implican la selección de R o bien Python para su ejecución.

En particular, en nuestro caso optaremos por Python (v. 2.7.14) puesto que disponemos de mayor 'know-how' y capacidad de adaptación sobre dicha tecnología.

El desarrollo se realizará a través de Notebook (Jupyter) por la flexibilidad y capacidad de anotación paralela.

2.6 Preparación de los datos.

2.6.1 Análisis de conjunto

Entendemos por análisis de conjunto, la identificación del número de registros de cada dataset, la identificación de sus variables, el análisis estadístico general (que permite identificar rápidamente posibles errores en los datos o bien posibles acciones a realizar con ellos), así como su correcto tipado una vez realizada la exportación del fichero hacia las diferentes herramientas de trabajo.

- Número de registros, número de atributos y tipo

```
path = "C:/Users/Usuario/CloudStation/02 Emprender/03 HolaParlem/10 Biblioteca/06 Master Data Science UOC/97 Entregas/06 Tipologi
file_name="winemag-data-130k-v2.csv"
data = pd.read_csv(path + "/" + file_name)

# Mostramos el número de muestras del fichero
print ("Número de registros: " + str(len(data)) + "\n")

# Mostramos el número de muestras del fichero
print ("Número de atributos: " + str(len(data.columns)) + "\n")

# Mostramos los atributos juntpo con el tipo asignado en la extracción
print "Analisis del tipado de los datos:"
print(data.dtypes)
print ("\n")
```

```
Número de registros: 129971

Número de atributos: 14

Analisis del tipado de los datos:
Unnamed: 0                 int64
country                   object
description               object
designation               object
points                     int64
price                   float64
province                  object
region_1                  object
region_2                  object
taster_name               object
taster_twitter_handle     object
title                     object
variety                   object
winery                    object
dtype: object
```

- Estadística básica

```
# Analisis estadistico básico
print (data.describe())
print ("\n")

# Nombramos los atributos que no están identificados (ID)
data.rename(columns={'Unnamed: 0':'id'}, inplace=True)
print (data.columns)
print ("\n")
```

```
          Unnamed: 0         points          price
count  129971.000000  129971.000000  120975.000000
mean    64985.000000      88.447138      35.363389
std     37519.540256       3.039730      41.022218
min         0.000000      80.000000       4.000000
25%     32492.500000      86.000000      17.000000
50%     64985.000000      88.000000      25.000000
75%     97477.500000      91.000000      42.000000
max    129970.000000     100.000000    3300.000000


Index([u'id', u'country', u'description', u'designation', u'points', u'price',
       u'province', u'region_1', u'region_2', u'taster_name',
       u'taster_twitter_handle', u'title', u'variety', u'winery'],
      dtype='object')
```

Se dispone de **129.971 registros y 14 características.**

Se observa que el **tipado de los datos se corresponde con la descripción** ofrecida para el conjunto de datos, por lo que no es necesario realizar ninguna transformación.

Se observa igualmente una variable sin nombre que corresponde al identificador/contador de registros. **Se renombra dicha variable como 'id'**.

La estadística general, para los atributos numéricos (puntuación y precio), nos indica una **amplia dispersión del precio (entre 4€ y 3.300 €),** con un precio medio de 35 € y una desviación estándar relativamente pequeña, lo que **permite considerar la presencia de valores extremos en la muestra** que deberán ser analizados.

La puntuación no presenta una dispersión apreciable y es consistente con la información proporcionada por la web de valoraciones superiores a 80 puntos.

2.6.2 Anonimización

Los datos **no contienen registros con información personal sensible** que requieran anonimización.

2.6.3 Consolidación e Integración de datos

El conjunto de datos bajo análisis ya se encuentra unificado en un único dataset, por lo que **no es necesario integrar diferentes fuentes**.

Los comentarios y conclusiones al resto de apartados pueden verse en el Notebook entregado.

2.6.4 Limpieza y conclusiones

Los comentarios y conclusiones al resto de apartados pueden verse en el Notebook entregado.

3 Referencias

[1] Zack Thoutt profile (linkedin). https://www.linkedin.com/in/zack-thoutt-57275655/

[2] WineEnthsiasts web. https://www.winemag.com/?s=&drink_type=wine

[3] Kaggle. Wine Reviews. https://www.kaggle.com/zynicide/wine-reviews