Práctica 2: Limpieza y análisis de datos

Rubén Moya Vázquez

Tipología y ciclo de vida de los datos

Máster Universitario en Ciencia de Datos

Universitat Oberta de Catalunya

Nota del autor

En la redacción de este documento se tiene en cuenta la normativa APA. En caso de necesitar incluir imágenes o tablas para ilustrar nuestras respuestas, se añadirán junto a la respuesta que desean ilustrar, en lugar de en el anexo correspondiente, para facilitar la lectura y la corrección del documento por parte del personal docente.

Esta práctica ha sido realizada única y exclusivamente por el alumno Rubén Moya Vázquez, tal y como firma a continuación.

Resumen

En esta práctica aplicaremos los conocimientos adquiridos a lo largo del estudio de la asignatura que enmarca la misma para que realizar un caso práctico de limpieza y análisis de un conjunto de datos. En este caso, se ha seleccionado el conjunto de datos relativo a la clasificación de estrellas en base a sus características de tamaño, luminosidad, temperatura, etc. Este documento sirve de lectura principal de la práctica, pero a su vez, necesita de la lectura del documento “star_cleaner.rmd” o el reporte generado por el mismo para su comprensión total.

Palabras clave: Limpieza de Datos, Análisis de Datos, Data Science, Estrellas, R

Práctica 2: Limpieza y análisis de datos

En esta práctica llevaremos a cabo el tratamiento de un conjunto de datos seleccionado de la web Kaggle (https://www.kaggle.com) con el objetivo de demostrar los conocimientos adquiridos durante el estudio de esta asignatura.

Descripción del dataset

El dataset seleccionado para la ejecución de esta práctica ha sido "*Star dataset to predict star types*" (Baidya, 2019). Este dataset contiene 240 elementos con los siguientes 7 atributos.

1. **Temperature (K)**. Temperatura absoluta en grados kelvin. Dato medido.
2. **Luminosity(L/Lo)**. Luminosidad relativa respecto al Sol. Dato calculado tomando como referencia la luminosidad del Sol.
3. **Radius(R/Ro)**. Radio relativo respecto al sol. Dato calculado tomando como referencia el radio del Sol.
4. **Absolute magnitude(Mv)**. Magnitud visual absoluta.
5. **Star type**. Tipo de estrella. Variable categórica con 6 valores posibles:
    0. Brown Dwarf.
    1. Red Dwarf.
    2. White Dwarf.
    3. Main Sequence.
    4. Supergiant.
    5. Hypergiant.
6. **Star color**. Valor medido del color en base al análisis espectral.

7. **Spectral Class**. Clase dentro de la clasificación espectral de las estrellas. Variable categórica con 7 valores posibles (O,B,A,F,G,K,M).

¿Por qué es importante? ¿Qué pregunta pretende responder?

Este dataset contiene las mediciones de 240 estrellas del espacio observable, siendo uno de sus parámetros la clasificación de la estrella. La relevancia de este conjunto de datos radica en que nos permite, en base al diagrama "*Hertzsprung-Russell*" (Wikipedia, 2021), clasificar las estrellas observadas en una de las 6 clases descritas, en función del resto de variables definidas. Esto nos permite crear un modelo de clasificación no supervisado, que pueda, en función de las mediciones obtenidas, clasificar los cuerpos celestes observados por los telescopios de manera automatizada. Es decir, sienta un primer precedente para la cartografía automatizada del universo observable.

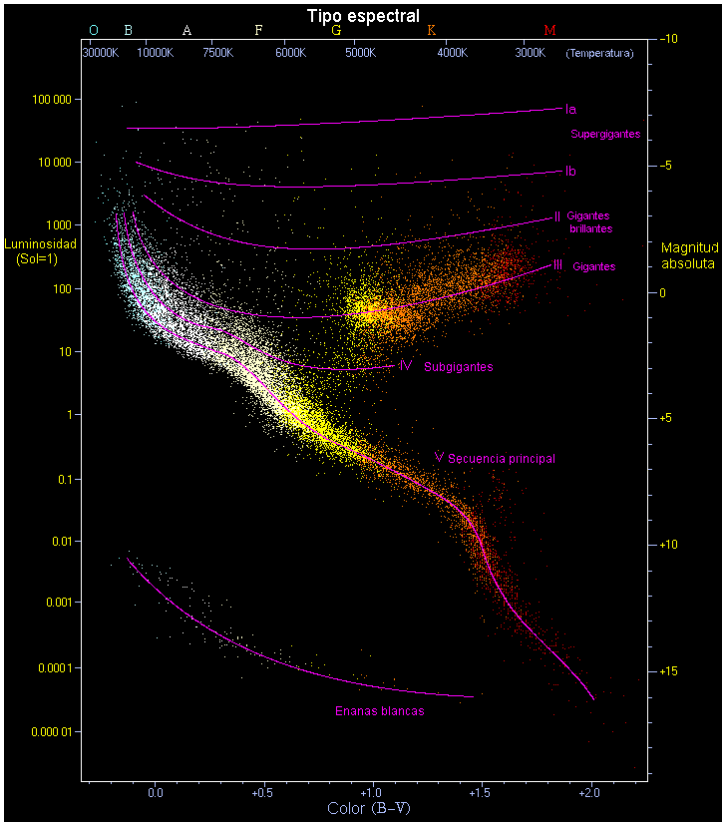

Ilustración 1. Diagrama Hertzsprung-Russell.

Integración y selección de los datos de interés a analizar

En este caso, los datos no requieren de ninguna integración ya que se nos presentan en un único fichero csv ("6_class_csv.csv"). El único cambio realizado ha sido la substitución de los espacios por '_' en el título del fichero, para evitar problemas al leerlo desde RStudio.

Por otra parte, consideramos que, partiendo del diagrama Hertzsprung-Russell podríamos considerar inicialmente que el radio es una variable que no aporta información relevante a la hora de realizar modelo de clasificación, pero tratándose de un conjunto de datos pequeño, consideramos que podemos mantener todas las variables en nuestro conjunto de datos ya que podríamos encontrar relaciones ocultas que fuesen interesantes.

Limpieza de datos

El proceso de limpieza de datos se visualiza mejor en el fichero rmd (star_cleaner.rmd) adjunto. Por ello, esta pregunta la responderemos en extensión, directamente sobre el fichero en cuestión.

Para no dar un salto disruptivo en la lectura y corrección de este documento, extraeremos las conclusiones del proceso de tratamiento de valores nulo y outliers.

Conclusiones del proceso de limpieza de datos

En cuanto a los valores nulos podemos decir que no se ha encontrado ninguno en ninguna de las variables de nuestro dataset.

Respecto a los outliers podemos decir que, pese a que sí que hemos encontrado cierta cantidad en las variables numéricas, finalmente hemos decidido contemplarlos tal y como

aparecen, sin realizar modificaciones, al considerar que son valores validos extraídos de mediciones correctas.

Análisis de los datos

Selección de los grupos de datos que se quieren analizar/comparar (planificación de los análisis a aplicar)

En el fichero rmd adjunto se puede ver como hemos realizado las agrupaciones pertinentes en base a distintos criterios como el color, el tipo o la clase espectral.

Comprobación de la normalidad y homogeneidad de la varianza

Al igual que en el caso anterior, podemos ver los resultados del estudio de la normalidad y homogeneidad de la varianza en el fichero adjunto.

Para el estudio de la normalidad se han realizado las pruebas de Anderson-Darling sobre nuestras variables numéricas. A su vez, se ha estudiado si alguna de dichas variables, que no siguen distribuciones normales, podrían ser candidatas a la normalización. Finalmente, dicha normalización ha sido descartada por criterios de rigurosidad sobre los datos.

Respecto a la homogeneidad de la varianza, la prueba utilizada para comprobarlo ha sido la de Levene, aplicado sobre los grupos formados por los diferentes tipos de estrella. En este caso hemos podido percatarnos de diferencias en la homogeneidad de dichas varianzas según el grupo.

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. Aplicar al menos tres métodos de análisis diferentes.

La resolución de este apartado se puede ver en el rmd adjunto en el proyecto. De todas formas, pasamos a realizar una descripción de los análisis realizados a continuación:

1. En primer lugar, hemos realizado un pequeño contraste de hipótesis para ver si era cierto que las enanas blancas son menos luminosas que las estrellas de la secuencia principal. Este contraste no se ha podido realizar correctamente por falta de datos.
2. En segundo lugar, hemos creado un modelo de predicción para clasificar estrellas basado en Random Forest con K-fold cross validation para garantizar la calidad del modelo y minimizar las posibilidades de sobre-entrenamiento.
3. Finalmente hemos realizado una pequeña regresión lineal con el objetivo de ver la posible correlación entre el radio (la variable que estuvimos a punto de dejar fuera del dataset) y nuestra clase. En dicha regresión lineal vimos que la variable radio si que tiene cierta correlación con nuestra clase y, por tanto, es relevante.

Resolución del problema

¿Cuáles son las conclusiones? ¿Los resultados permiten responder al problema?

Como conclusiones podemos extraer que, si bien los datos si son validos para poder realizar un modelo que nos permita la catalogación automatizada de cuerpos celestes, el volumen de los datos de nuestro dataset es demasiado pequeño para poder realizar estudios exhaustivos o en mayor detalle. Si contásemos con un mayor numero de estrellas en nuestro

dataset (varios miles) podríamos desarrollar un modelo más efectivo y usar el conjunto de datos para realizar contrastes de hipótesis más en detalle.

Repositorio Git.

El enlace al repositorio git en el que se encontrará todos los entregables deseados es el siguiente: https://github.com/rmoyav/TIPOLOGIA_PRA_2

Bibliografía

Baidya, D. (21 de 10 de 2019). *Star dataset to predict star types*. Obtenido de kaggle: https://www.kaggle.com/deepu1109/star-dataset