titanicDataClean

Analiza los datos del dataset de supervivientes del Titanic de la web www.kaggle.com y los analiza para encontrar una pauta de comportamiento general y verificable.

Práctica 2: Limpieza y validación de los datos

Descripción

Esta práctica se ha realizado bajo el contexto de la asignatura *Tipología y ciclo de vida de los datos*, perteneciente al Máster en Ciencia de Datos de la Universitat Oberta de Catalunya.

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) * Predict Future Sales (https://www.kaggle.com/c/competitive-data-sciencepredict-future-sales/).

Los últimos dos ejemplos corresponden a competiciones activas de Kaggle de manera que, opcionalmente, podríais aprovechar el trabajo realizado durante la práctica para entrar en alguna de estas competiciones.

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:
   3.1. ¿Los datos contienen ceros o elementos vacíos? ¿Cómo gestionarías cada uno de estos casos? 3.2. Identificación y tratamiento de valores extremos.
4. Análisis de los datos.
   4.1. Selección de los grupos de datos que se quieren analizar/comparar (planificación de los análisis a aplicar).
   4.2. Comprobación de la normalidad y homogeneidad de la varianza.
   4.3. 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.

El conjunto de datos de análisis escogido ha sido finalmente el Titanic de Kaggle [https://www.kaggle.com/c/titanic/data].

Miembros del equipo

La actividad ha sido realizada de manera individual por **Ricardo García Ruiz**.

Licencia

La licencia utilizada finalmente ha sido la *CC BY-NC-SA 4.0 International*. La licencia CC BY-NC-SA 4.0 International es una licencia de software libre muy utilizada y constituye un documento fundamental para el movimiento de software libre. CC BY-NC-SA 4.0 International es una licencia acorde al marco internacional

de derechos de autor y al nacional en España, siendo flexible y compatible con otras licencias de software libre.

Se permite con nuestro trabajo y la base de datos extraída de la web:

- *Compartir* — copiar y redistribuir el material en cualquier medio o formato
- *Adaptar* — remezclar, transformar y crear a partir del material

Por otro lado, la licencia activa las siguientes restricciones:

- **Reconocimiento**: Debe reconocer adecuadamente la autoría, proporcionar un enlace a la licencia e indicar si se han realizado cambios. Puede hacerlo de cualquier manera razonable, pero no de una manera que sugiera que tiene el apoyo del licenciador o lo recibe por el uso que hace.
- **NoComercial**: No puede utilizar el material para una finalidad comercial.
- **CompartirIgual**: Si remezcla, transforma o crea a partir del material, deberá difundir sus contribuciones bajo la misma licencia que el original.
- **No hay restricciones adicionales**: No puede aplicar términos legales o medidas tecnológicas que legalmente restrinjan realizar aquello que la licencia permite.

Ficheros del código fuente

- **src/titanicDataClean**: Es el código completo de gestión de los datos de la web www.kaggle.com que se ha utilizado para el análisis, limpieza y gestión del modelo final de trabajo.

Esta actividad se encuentra en la dirección de GitHub siguiente: https://github.com/rgarciarui/titanicDataClean

Recursos

1. Squire, Megan (2015). Clean Data. Packt Publishing Ltd.
2. Jiawei Han, Micheine Kamber, Jian Pei (2012). Data mining: concepts and techniques. Morgan Kaufmann.
3. Jason W. Osborne (2010). Data Cleaning Basics: Best Practices in Dealing with Extreme Scores. Newborn and Infant Nursing Reviews; 10 (1): pp. 1527-3369. Peter Dalgaard (2008). Introductory statistics with R. Springer Science & Business Media.
4. Wes McKinney (2012). Python for Data Analysis. O'Reilley Media, Inc. Tutorial de Github (https://guides.github.com/activities/hello-world/)