Depuración y calidad de datos en R (I)

SISTEMAS INTELIGENTES PARA LA GESTIÓN DE LA EMPRESA
CURSO 2016-2017

Documentación adicional

R

- J. Adler, R in a nutshell. O'Reilly Media, 2012. http://proquest.safaribooksonline.com/book/programming/r/9781449358204
- P. Teetor, R Cookbook. O'Reilly Media, 2011. http://proquest.safaribooksonline.com/book/programming/r/9780596809287
- H. Wickham, G. Grolemund. R for Data Science. O'Reilly Media, 2016. http://proquest.safaribooksonline.com/book/programming/r/9781491910382

ggplot2

- H. Wickham, Elegant Graphics for Data Analysis. Springer, 2016. https://github.com/hadley/ggplot2-book
- W. Chang, R Graphics Cookbook. O'Reilly Media, 2012. http://proquest.safaribooksonline.com/book/programming/r/9781449363086

Dataset - adult

Información del censo de EE.UU: de 1994

Objetivo: Predecir si una persona ganará más 50.000$ a partir de datos demográficos

Variables (14):

- Age = [17.0, 90.0]
- Workclass = {Private, Self-emp-not-inc, Self-emp-inc, Federal-gov, Local-gov, State-gov, Without-pay, Never-worked}
- Fnlwgt = [12285.0, 1490400.0]
- Education = {Bachelors, Some-college, 11th, HS-grad, Prof-school, Assoc-acdm, Assoc-voc, 9th, 7th-8th, 12th, Masters, 1st-4th, 10th, Doctorate, 5th-6th, Preschool}
- Education-num = [1.0, 16.0]
- Marital-status = {Married-civ-spouse, Divorced, Never-married, Separated, Widowed, Married-spouse-absent, Married-AF-spouse}
- Occupation = {Tech-support, Craft-repair, Other-service, Sales, Exec-managerial, Prof-specialty, Handlers-cleaners, Machine-op-inspct, Adm-clerical, Farming-fishing, Transport-moving, Priv-house-serv, Protective-serv, Armed-Forces}
- Relationship = {Wife, Own-child, Husband, Not-in-family, Other-relative, Unmarried}
- Race = {White, Asian-Pac-Islander, Amer-Indian-Eskimo, Other, Black}
- Sex = {Female, Male}
- Capital-gain = [0.0, 99999.0]
- Capital-loss = [0.0, 4356.0]
- Hours-per-week =[1.0, 99.0]
- Native-country {United-States, Cambodia, England, Puerto-Rico, Canada, Germany, Outlying-US(Guam-USVI-etc), India, Japan, Greece, South, China, Cuba, Iran, Honduras, Philippines, Italy, Poland, Jamaica, Vietnam, Mexico, Portugal, Ireland, France, Dominican-Republic, Laos, Ecuador, Taiwan, Haiti, Columbia, Hungary, Guatemala, Nicaragua, Scotland, Thailand, Yugoslavia, El-Salvador, Trinadad&Tobago, Peru, Hong, Holand-Netherlands}
- **Class** = {>50K, <=50K}

Dataset - adult

Original: https://archive.ics.uci.edu/ml/datasets/Adult

Modificados:

- Entrenamiento (80%): *adult.training.csv*
  - Valores perdidos (3.620)
  - Ruido de clase (10%)
- Validación (20%): *adult.test.csv*

Dataset - titanic

Objetivo: Predecir si un pasajero sobrevive o no en función de una serie de variables relativas a la edad, género, etc.

Variables (9):

- **survival =** {0, 1}
- pclass = {1st, 2nd, 3rd}
- sex: sexo
- age: edad
- sibsp: número de parientes (hermano/a, hermanastro/a) / cónyuge (esposo/a) a bordo
- parch: número de padres (madre/padre) / hijos a bordo (hijo/a, hijastro/a)
- ticket: número de ticket
- fare: precio del ticket
- cabin: número del camarote
- embarked: puerto de embarque

Dataset - titanic

Datasets:
- https://www.kaggle.com/c/titanic/data

Métrica:
- % de pasajeros correctamente clasificados (*accuracy*)

Formato:
- Entrenamiento *(train.csv*)
- Validación (*test.csv*):
  - No incluye el valor objetivo
  - Enviar un fichero .csv con cabecera + 418 entradas
  - Cada entrada incluye dos columnas: PassengerId, Survided
  - Ejemplo: https://www.kaggle.com/c/titanic/data

Tutoriales:
- Kaggle R tutorial on Machine Learning (Datacamp) https://www.datacamp.com/community/open-courses/kaggle-tutorial-on-machine-learing-the-sinking-of-the-titanic#gs.null
- Getting started with R (Trevor Stephens) http://trevorstephens.com/kaggle-titanic-tutorial/getting-started-with-r/
- Exploring the Titanic Dataset (Megan L. Risdal) https://www.kaggle.com/mrisdal/titanic/exploring-survival-on-the-titanic/notebook