Pràctica2: Neteja i Validació de les Dades

Joaquim Dalmases i Juanjo Díez

4 de junio, 2019

Contents

1 Introducció

1.1 Presentació

En aquesta pràctica s'elabora un cas pràctic orientat a aprendre a identificar les dades rellevants per un projecte analític i usar les eines d'integració, neteja, validació i anàlisi de les mateixes. Per fer aquesta pràctica haureu de treballar en grups de 2 persones. Haureu de lliurar un sol fitxer amb l'enllaç Github on es trobin les solucions incloent els noms dels components de l'equip. Podeu utilitzar la Wiki de Github per descriure el vostre equip i els diferents arxius que corresponen a la vostra entrega. Cada membre de l'equip haurà de contribuir amb el seu usuari Github.

1.2 Competències

En aquesta pràctica es desenvolupen les següents competències del Màster de Data Science:

- Capacitat d'analitzar un problema en el nivell d'abstracció adequat a cada situació i aplicar les habilitats i coneixements adquirits per abordar-lo i resoldre'l.
- Capacitat per aplicar les tècniques específiques de tractament de dades (integració, transformació, neteja i validació) per al seu posterior anàlisi. Desenvolupar la capacitat de cerca, gestió i ús d'informació i recursos en l'àmbit de la ciència de dades.

1.3 Objectius

Els objectius concrets d'aquesta pràctica són:

- Aprendre a aplicar els coneixements adquirits i la seva capacitat de resolució de problemes en entorns nous o poc coneguts dintre de contextos més amplis o multidisciplinaris.
- Saber identificar les dades rellevants i els tractaments necessaris (integració, neteja i validació) per dur a terme un projecte analític.
- Aprendre a analitzar les dades adequadament per abordar la informació continguda en les dades.

- Identificar la millor representació dels resultats per tal d'aportar conclusions sobre el problema plantejat en el procés analític.
- Actuar amb els principis ètics i legals relacionats amb la manipulació de dades en funció de l'àmbit d'aplicació.
- Desenvolupar les habilitats d'aprenentatge que els permetin continuar estudiant d'una manera que haurà de ser en gran manera autodirigida o autònoma.

1.1 Repositori triat

Per l'elaboració de la pràctica s'ha triat:

- el repositori de *Kaggle* **Red Wine Quality**
- que correspon amb el repositori de *UCI* **Wine Quality Data Set** i
- l'accés a les dades completes es pot trobar a *aquest enllaç*.

Aquesta pràctica s'ha desenvolupat seguintla bibliografia recomanada: (Calvo M 2019; Squire 2015; Jiawei Han 2012; Dalgaard 2008)

Referències

Calvo M, Pérez D, Subirats L. 2019. *Introducción a La Limpieza Y Análisis de Los Datos.* Editorial UOC.

Dalgaard, Peter. 2008. *Introductory Statistics with R.* Springer Science & Business Media.

Jiawei Han, Jian Pei, Micheine Kamber. 2012. *Data Mining: Concepts and Techniques.* Morgan Kaufmann.

Squire, Megan. 2015. *Clean Data.* Packt Publishing Ltd.