Pràctica 2 (35% nota final)

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 (https://github.com) 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. Malgrat que no es tracta del mateix enunciat, els següents exemples d'edicions anteriors us poden servir com a guia:

- Exemple: https://github.com/Bengis/nba-gap-cleaning
- Exemple complex (fitxer adjunt).

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.

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.
- Desenvolupar la capacitat de cerca, gestió i ús d'informació i recursos en l'àmbit de la ciència de dades.

Descripció de la Pràctica a realitzar

L'objectiu d'aquesta activitat serà el tractament d'un dataset, que pot ser el creat a la pràctica 1 o bé qualsevol dataset lliure disponible a Kaggle (https://www.kaggle.com). Alguns exemples de dataset amb els que podeu treballar són:

- 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).

L'últim exemple correspon a una competició activa a Kaggle de manera que, opcionalment, podeu aprofitar el treball realitzat durant la pràctica per entrar en aquesta competició.

Seguint les principals etapes d'un projecte analític, les diferents tasques a realitzar (i **justificar**) són les següents:

1. Descripció del dataset. Perquè és important i quina pregunta/problema pretén respondre?
2. Integració i selecció de les dades d'interès a analitzar.
3. Neteja de les dades.
   - 3.1. Les dades contenen zeros o elements buits? Com gestionaries aquests casos?
   - 3.2. Identificació i tractament de valors extrems.
4. Anàlisi de les dades.
   - 4.1. Selecció dels grups de dades que es volen analitzar/comparar (planificació dels anàlisis a aplicar).
   - 4.2. Comprovació de la normalitat i homogeneïtat de la variància.
   - 4.3. Aplicació de proves estadístiques per comparar els grups de dades. En funció de les dades i de l'objectiu de l'estudi, aplicar proves de contrast d'hipòtesis, correlacions, regressions, etc. Aplicar almenys tres mètodes d'anàlisi diferents.
5. Representació dels resultats a partir de taules i gràfiques.
6. Resolució del problema. A partir dels resultats obtinguts, quines són les conclusions? Els resultats permeten respondre al problema?
7. Codi: Cal adjuntar el codi, preferiblement en R, amb el que s'ha realitzat la neteja, anàlisi i representació de les dades. Si ho preferiu, també podeu treballar en Python.

Recursos

Els següents recursos són d'utilitat per la realització de la pràctica:

- Calvo M., Subirats L., Pérez D. (2019). Introducción a la limpieza y análisis de los datos. Editorial UOC.

- Megan Squire (2015). *Clean Data*. Packt Publishing Ltd.
- Jiawei Han, Micheine Kamber, Jian Pei (2012). *Data mining: concepts and techniques*. Morgan Kaufmann.
- 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.
- Wes McKinney (2012). *Python for Data Analysis*. O'Reilley Media, Inc.
- Tutorial de Github https://guides.github.com/activities/hello-world.

Criteris de valoració

Tots els apartat són obligatoris. La ponderació dels exercicis és la següent:

- Els apartats 1, 2 i 6 valen 0,5 punts.
- Els apartats 3,5 i 7 valen 2 punts.
- L'apartat 4 val 2,5 punts.

Es valorarà la idoneïtat de les respostes, que han de ser clares i completes. Les diferents etapes han d'estar ben justificades i acompanyades del codi corresponent. També es valorarà la síntesi i claredat, a través de l'ús de comentaris, del codi resultant, així com la qualitat de les dades finals analitzades.

Format i data de lliurament

Durant la setmana del 23 de desembre el grup podrà lliurar al professor una entrega parcial opcional. Aquesta entrega parcial és molt recomanable per tal de rebre assessorament sobre la pràctica i verificar que la direcció presa és la correcta. Es lliuraran comentaris als estudiants que hagin efectuat l'entrega parcial però no comptarà per la nota de la pràctica. En l'entrega parcial els estudiants hauran de lliurar per correu electrònic (mcalvogonza@uoc.edu) l'enllaç al repositori Github amb el que hagin avançat.

Pel que fa a l'entrega final, cal lliurar un únic fitxer que contingui l'enllaç a Github on hi hagi:

1. Una Wiki on hi hagi els noms dels components del grup i una descripció dels fitxers.
2. Un document Word, Open Office o PDF amb les respostes a les preguntes i els noms dels components del grup. A més, al final de document, haurà d'aparèixer la següent taula de contribucions al treball, la qual ha de signar cada integrant del grup amb les seves inicials. Les inicials representen la confirmació de que l'integrant ha participat en aquell apartat. Tots els integrants han de participar en cadascun dels apartats, de manera que, idealment, els apartats hauran d'estar signats per tots els integrants.

Contribuciones	Firma
Investigació prèvia	Integrant 1, Integrant 2, ...
Redacció de les respostes	Integrant 1, Integrant 2, ...
Desenvolupament codi	Integrant 1, Integrant 2, ...

3. Una carpeta amb el codi generat per analitzar les dades.
4. El fitxer CSV amb les dades originals.
5. El fitxer CSV amb les dades finals analitzades.

Aquest document de l'entrega final de la Pràctica 2 s'ha de lliurar a l'espai de Lliurament i Registre d'AC de l'aula abans de les **23:59 del dia 7 de gener**. No s'acceptaran lliuraments fora de termini.