PAC 2 (10% nota final)

Presentació

En aquesta Prova d'Avaluació Continuada es treballen els conceptes generals d'integració, validació i anàlisi dels diferents tipus de dades.

Competències

En aquesta PAC 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 Prova d'Avaluació Continuada són:

- Conèixer els efectes de l'utilització de dades de qualitat en els processos analítics.
- Conèixer les principals eines de neteja i anàlisi dels diferents tipus de dades.
- 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.
- 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 PAC a realitzar

Exercici 1 [30%]

Després de llegir el recurs “Clean Data”, contesta les següents preguntes amb les teves pròpies paraules.

1. Quines són les principals etapes d’un projecte analític en ciència de dades? Posa un exemple real indicant les principals tasques realitzades en cadascuna d’aquestes etapes.

2. Una part important de la neteja de dades consisteix en convertir diferents tipus de dades per tal d’integrar subsets homogenis. No obstant això, aquesta transformació té un risc. Quin és aquest risc i quins són els principals factors que poden donar lloc a aquesta situació? Posa un exemple per cada cas.

3. Quina és la diferència entre els zeros i les dades buides? Posa un exemple on l’etapa de data cleaning consistirà en emplenar dades perdudes amb zeros, així com un altre exemple on es deixaran com dades buides.

Exercici 2 [30%]

Després de llegir el recurs “The power of outliers”, contesta les següents preguntes:

1. En l’anàlisi estadístic d’hipòtesis, quins tipus d’errors existeixen? Posa un exemple pràctic de cadascun.

2. Què són els *extreme scores*? Enumera els diferents tipus de valors extrems i explica com aquests afecten a l’anàlisi estadístic.

3. Els *extreme scores* sempre provenen d’errors a les dades? Explica amb les teves pròpies paraules les diferents causes que poden donar lloc a l’aparició d’*outliers*. Posa un exemple pràctic de cadascuna.

Exercici 3 [20%]

Després de llegir el recurs “The art of data analysis” contesta les següents preguntes:

1. Quina és la diferència entre les dades quantitatives i les qualitatives? Són analitzades de la mateixa manera? Posa un exemple de cada cas.

2. A què fa referència el teorema central del límit i quines són les 3 propietats més importants? Segons aquest teorema, l'utilització de més dades és sempre un aspecte positiu?

Exercici 4 [20%]

Imagina que realitzes un estudi sobre una malaltia on obtens resultats molt interessants relacionats amb la predicció de risc d'una sèrie de pacients donats uns indicadors. Posteriorment a la publicació d'aquest estudi, diferents hospitals i organitzacions et contacten per afegir dades al teu estudi i, per tant, arribar a resultats més concluents. Anomena i justifica almenys 5 processos que aplicaries per l'integració, neteja i validació de totes les dades disponibles.

Recursos

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

Bàsics

- Megan Squire (2015). *Clean Data*. Packt Publishing Ltd. Capítols 1 i 2.
- Jason W. Osborne and Amy Overbay (2004). *The power of outliers (and why researchers should ALWAYS check for them)*. Practical Assessment, Research & Evaluation, 9 (6).
- Kristin H. Jarman (2013). *The art of data analysis: how to answer almost any question using basic statistics*. John Wiley & Sons, Inc. Capítols 3 i 5.

Criteris de valoració

La ponderació dels exercicis és la següent:

- Exercici 1: 30%
- Exercici 2: 30%
- Exercici 3: 20%
- Exercici 4: 20%

Es valorarà la idoneïtat de les respostes, que han de ser clares i completes. Quan sigui necessari, hauran d'estar acompanyades d'exemples representatius i ben justificats.

Format i data de lliurament

Cal lliurar un únic document Word, Open Office o **PDF** (preferiblement aquest últim) amb les respostes a les preguntes.

Aquest document s'ha de lliurar a l'espai de Lliurament i Registre d'AC de l'aula abans de les **23:59** del dia **4 de desembre**. No s'acceptaran lliuraments fora de termini.