TIPOLOGIA I CICLE DE VIDA DE LES DADES

PAC1: Què son les dades i quin el seu cicle de vida?

Marc Cervera Rosell

Semestre: febrer 2025 - juny 2025

Màster en ciència de dades

Índex

Índex de figures

Exercici 1

Pregunta 1

Per aplicar la tècnica de normalització caldria, en primer lloc, seleccionar aquelles variables numèriques a les quals volem aplicar la tècnica i després ajustar-les a una escala més reduïda, per tal de, fer-les més comprensibles. Típicament, les escales són entre -1.0 i 1.0 o entre 0.0 i 1.0.
Per aplicar la tècnica de discretització cal, com en el cas anterior, seleccionar aquelles variables numèriques a les quals volem aplicar la tècnica. Seguidament, cal substituir els valors numèrics per una sèrie d'etiquetes les quals poden ser conceptuals o intervals. Aquesta tècnica permet consolidar un possible criteri d'agrupació de dades. Per exemple, un *dataset* amb els resultats dels balanços d'empreses. Es podria substituir els resultats dels balanços per "positiu", "negatiu" i "neutre" i, posteriorment, agrupar pel resultat del balanç.
Aquestes tècniques són molt importants per als models de predicció, ja que permeten que les dades siguin més senzilles d'interpretar i això comporta que el rendiment del model sigui millor i que aquest tingui la capacitat de trobar patrons més fàcilment, és a dir, el model es fa més senzill.

Pregunta 2

Mentre un enginyer de dades treballa en el desenvolupament, construcció, prova i manteniment d'arquitectures, un científic de dades és una persona que es dedica a obtenir informació rellevant sobre les dades, a partir de preguntes que s'ha plantejat, i a transmetre aquesta informació de manera senzilla. A més, tot i que ambdós perfils tenen coneixements de programació, l'enginyer de dades no aprofundeix tant en coneixements d'estadística i matemàtiques com el científic de dades.
Oferta de *data scientist*
Oferta de *data engineer*

Pregunta 3

Les set tasques que permeten un major nivell d'abstracció en la visualització de dades són:

- Panorama general → Permet una visió de totes de les dades.
- Acostament → Permet focalitzar un punt concret de les dades.
- Filtratge → Permet filtrar els elements que no són d'interès.
- Detalls a petició → Permet obtenir detalls de les dades.

- Relacions → Permet veure com es relacionen les dades.
- Historial → Permet revisar/repetir versions anteriors.
- Extracció → Permet l'extracció de parts de les dades i els seus detalls.

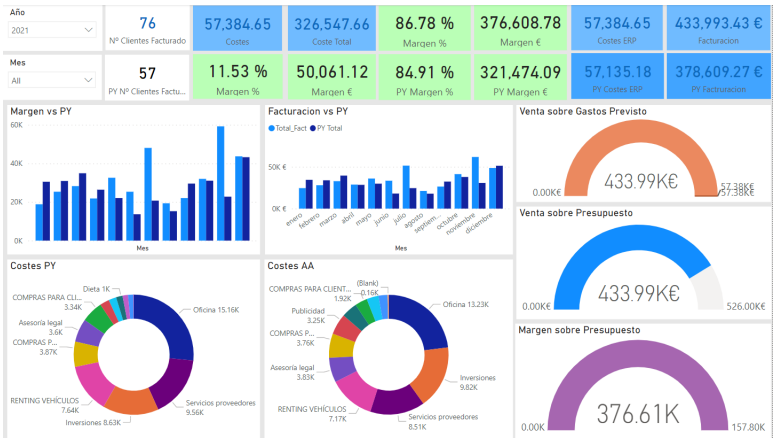

Figura 1: Exemple de la tasca "Panorama general" [1]

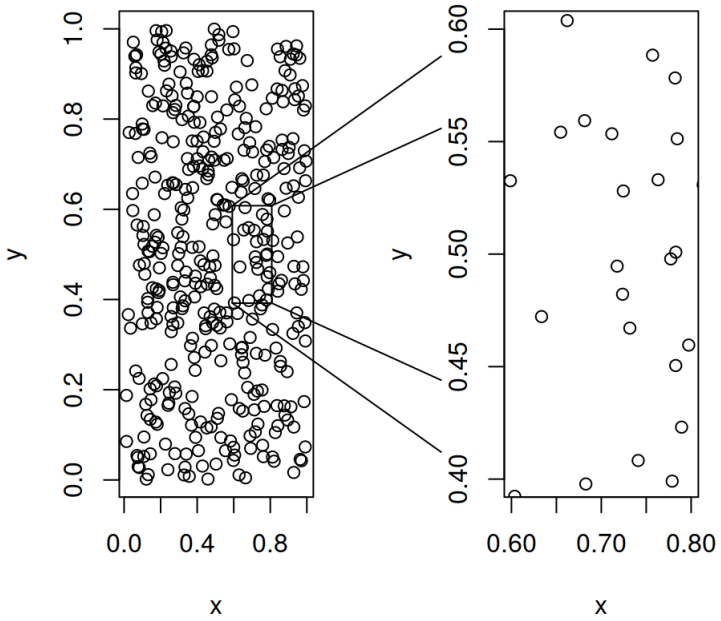

Figura 2: Exemple de la tasca "Zoom" [2]

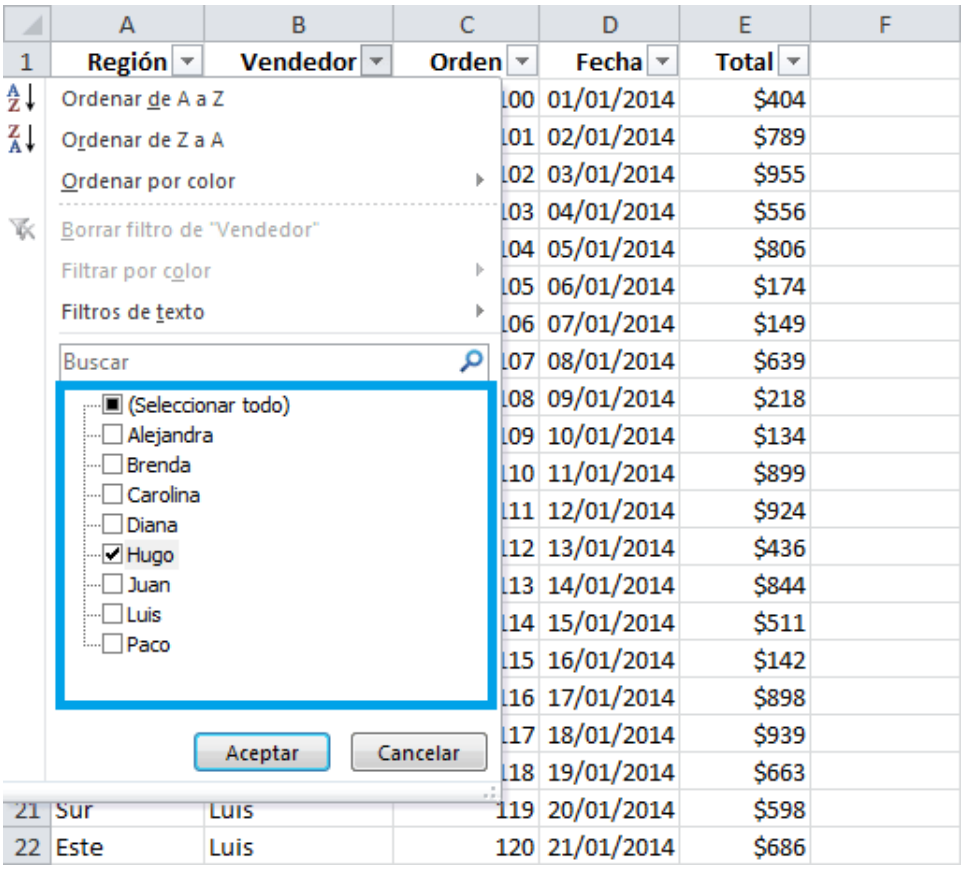

Figura 3: Exemple de la tasca “Filtratge” [3]

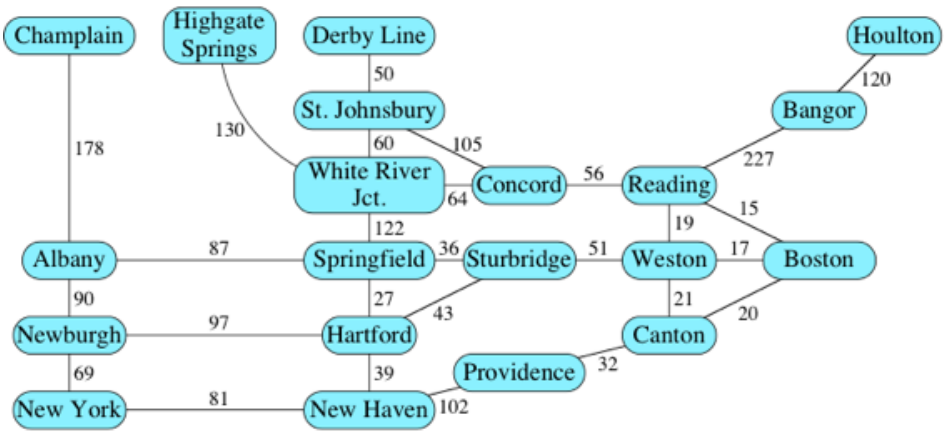

Figura 4: Exemple de la tasca “Relcions” [4]

Pregunta 4

La primera etapa és la captura, que té com a objectiu la recol·lecció de les dades que es generen en un procés. Un exemple d'aquesta etapa podria ser, les dades d'un sensor IoT que recol·lecta un broker de Kafka.
La segona etapa és l'emmagatzematge i té com a objectiu guardar les dades capturades en bases de dades o fitxers en un format adequat per a la posterior explotació. L'exemple en aquest cas, seria l'emmagatzemament de les lectures del sensor IoT en una base de dades.
La tercera etapa és el preprocessament i té com a objectiu preparar les dades per la seva posterior anàlisi i explotació. L'exemple podria ser la neteja de les dades.
La quarta etapa és l'anàlisi, que té com a objectiu crear models per explicar les dades. L'exemple seria construir un model de predicció de falles del sensor IoT.
La penúltima etapa és la visualització que té com a objectiu presentar les dades. L'exemple seria un panell de PowerBI.
Finalment, la publicació, que documenta els resultats. L'exemple seria la redacció d'un *report* amb els resultats obtinguts.

Pregunta 5

La millor opció per a un sistema de monitoratge ambiental basat en sensors seria una base de dades no relacional a causa del fet que els sensors poden generar dades molt diverses. És a dir els sensors poden generar dades de tipus text, numèriques, imatges, so, etc. Per tant, considerant aquesta gran varietat que deriva en diferents esquemes per cada sensor es considera una millor opció una base de dades no relacional gràcies a la seva flexibilitat amb els esquemes de les dades i els seus tipus. A més, cal destacar que aquest tipus de bases de dades tenen un rendiment alt de lectura i escriptura. Per tant, en el sistema que es planteja, que captura volums de dades molt grans, és important poder consultar i inserir dades de la manera més ràpida possible. Finalment, un altre motiu pel qual una base de dades NoSQL és una millor opció és la capacitat d'escalar horitzontalment. És a dir, la capacitat d'afegir més recursos per a poder treballar amb volums de dades majors que en les bases de dades relacionals.[5]

Pregunta 6

La primera llibreria és Pandas. Les seves funcionalitats clau són la lectura i escriptura de dades, la manipulació de dades, el tractament de *nulls*, l'agrupació, fusió i combinació i el tractament i anàlisi de *time series*. Aquesta llibreria conté una funció per eliminar duplicats (*drop_duplicates()*), una funció que es pot usar per a corregir errates tipogràfiques, tot i que no està dissenyada específicament per això(*replace()*) i una funció per a la con-

versió de formats (*astype()*). [6] [7] La segona eina és PyJanitor. És una llibreria basada en Pandas, per tant, les seves funcionalitats se centren simplificar algunes operacions de Pandas. Per eliminar duplicats es pot utilitzar la funció *remove_na()*. Per corregir errors tipogràfics i transformar formats, aquesta llibreria, no implementa cap funcionalitat especifia. En conseqüència, aquestes dues operacions s'haurien de realitzar amb Pandas o amb terceres llibreries. [8]

Exercici 2

Pregunta 1

Abans de procedir amb el *web scraping*, caldrà tenir en compte que la violació de les normes de *robots.txt*, pot comportar l'infringiment dels *termes d'ús*. També, cal considerar la legislació vigent. En el cas del nostre país, el Regne d'Espanya, cal revisar dues lleis; la de propietat intel·lectual (*Real Decreto Legislativo 1/1996 de 12 de abril*) i la de protecció de dades personals i garantia de drets digitals (*Ley orgáica 3/2018 de 5 de diciembre*). Èticament, cal respectar la "sagrada" voluntat del propietari de la *web*, per això la millor manera de no infringir les normes del lloc *web* i evitar incórrer en problemes legals és sol·licitar un permís per escrit del propietari on es demani consentiment per accedir a les dades.

Pregunta 2

Com a escenari es proposa un portal web que no permet l'accés a usuaris de països considerats dictatorials. Quan un usuari amb una adreça IP d'un d'aquests països vulgui accedir al lloc web, el servidor denegarà l'accés amb un codi d'error 403. [9] Aquest seria un escenari que dificulta el *web scraping* pel mètode de càrrega de dades, ja que, s'està impedint a usuaris de localitzacions concretes accedir a les dades.
Els possibles passos per superar aquesta barrera serien:

- **IDENTIFICAR** que es tracta d'un bloqueig per localització intentant accedir al portal web des de diferents xarxes, atès que així es probaria d'accedir al portal amb diferents IP.
- **CONFIGURAR** un servei VPN. Creant així la il·lusió d'estar en un altre país.
- **PROVAR** d'accedir al portal web des de diferents ubicacions fins a trobar un servidor d'un país que permeti l'accés.
- Un cop s'ha pogut accedir al portal web, **ESTUDIAR** els termes i condicions d'ús del lloc.

- **CONFIGURACIÓ** d'excepcions per a *timeouts* i *broken conncetions*.
- Configurar la profunditat de la pàgina per **EVITAR** els paranys d'aranya.

Pregunta 3

Modificació del *user agent* i les capçaleres HTTP $\longrightarrow$ Aquestes modificacions són una poderosa tècnica per simular que les peticions es fan des d'un navegador real i no des d'un d'un *script*.
Espaiat de peticions HTTP $\longrightarrow$ Fer que hi hagi un temps aleatori entre peticions "humanitza"el procés i fa que l'automatització de peticions sigui més difícil de detectar. Encara i així és la tècnica menys poderosa de les tres.
Ús de múltiples IPs $\longrightarrow$ És la tècnica més poderosa de les 3, ja que, el fet de tenir múltiples IPs evita el "baneig" per fer massa peticions.

Bibliografia

- [1] Imatge de l'exemple de "Panorama general"
- [2] Imatge de l'exemple de "Zoom"
- [3] Imatge de l'exemple de "Filtratge"
- [4] Imatge de l'exemple de "Relacions"
- [5] Bases de datos NoSQL: qué son, tipos y ventajas.
- [6] Pandas: La Herramienta Esencial para Data Science en Python
- [7] Pandas: Cambiar los tipos de datos en los DataFrames
- [8] *Most Helpful Data Cleaning Tools in Python for 2025*
- [9] Codigos de estado de respuesta HTTP