PAC 1: Què son les dades i quin el seu cicle de vida?

Tipologia i cicle de vida de les dades

Nom: Marc Bracons Cucó

Aula 2

Enunciat

Exercici 1 [70%]

Després de llegir el recurs “Calvo, M., Pérez, D., Subirats, L. (2019). Introducció al cicle de vida de les dades.” respon les preguntes següents amb les teves pròpies paraules:

1. Què és la datificació i com es relaciona amb la societat de la informació? (Màxim 100 paraules.)

La datació és la transformació de les activitats i interaccions en dades que es puguin quantificar. La societat de la informació es basa en l'ús intensiu i la dependència de les tecnologies de la informació i la comunicació, facilitant la generació, emmagatzematge i anàlisi de grans volums de dades.

2. Quines són les principals diferències entre un científic de dades i un enginyer de dades? Adjunta un enllaç d'alguna oferta de Linkedin de científic de dades i una altra oferta per a enginyer de dades. (Màxim 100 paraules.)

Mentre que un científic de dades se centra en analitzar i interpretar grans volums de dades, un enginyer de dades se especialitza en la creació i manteniment de sistemes que recullen, guarden i processen les dades eficaçment.

Oferta de científic de dades:
https://www.linkedin.com/jobs/view/4040662688

Acerca del empleo

FIATC Seguros busca un/a **Data Scientist** altamente motivado/a y capacitado/a para unirse a su equipo en el área de Estrategia y Desarrollo Corporativo. El candidato/a seleccionado/a desempeñará un papel clave en la recopilación, análisis e interpretación de grandes volúmenes de datos con el objetivo de optimizar la toma de decisiones estratégicas, mejorar la eficiencia operativa y desarrollar nuevas oportunidades de negocio.

Responsabilidades:

- Analizar y procesar grandes conjuntos de datos relacionados con el sector de seguros y financieros.
- Desarrollar modelos predictivos y prescriptivos para la evaluación de riesgos, comportamiento del cliente y optimización de productos.
- Colaborar con equipos multidisciplinarios (marketing, ventas, desarrollo de productos) para identificar oportunidades de mejora y nuevos enfoques basados en datos.
- Automatizar procesos de análisis de datos y mejorar la eficiencia operativa mediante la implementación de herramientas avanzadas de análisis.
- Comunicar los hallazgos y recomendaciones a los líderes de negocio y partes interesadas, a través de informes claros y visualizaciones comprensibles.
- Aplicar técnicas de machine learning, inteligencia artificial y análisis estadístico para resolver problemas complejos relacionados con la estrategia empresarial.
- Monitorear y mejorar continuamente los modelos y las herramientas implementadas, asegurando su precisión y relevancia en el tiempo.
- Mantenerse al día con las tendencias emergentes en ciencia de datos, tecnologías financieras y aseguradoras.

Oferta de data enginyer:
https://www.linkedin.com/jobs/view/4039784158

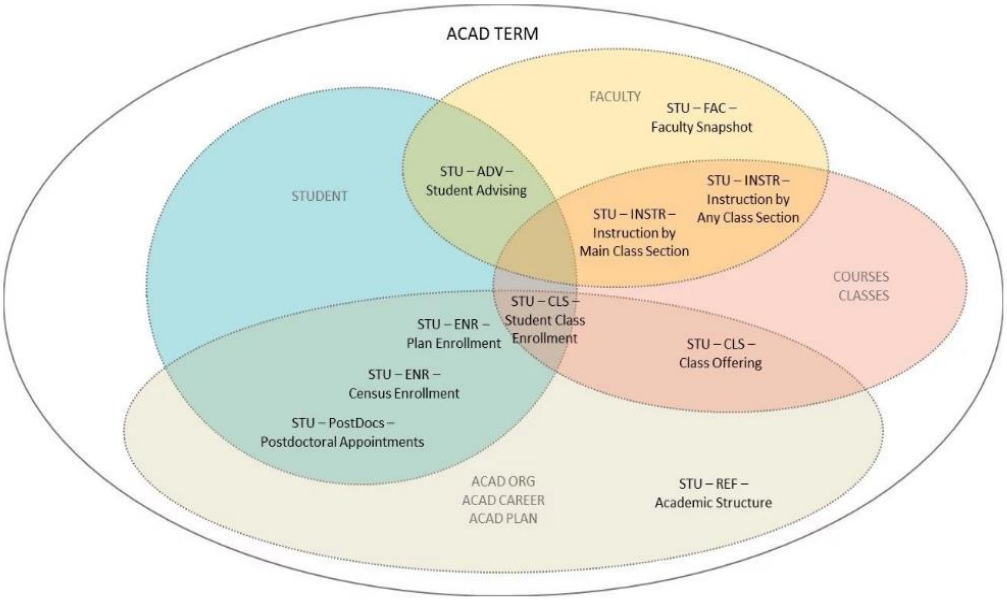
Puesto y Responsabilidades

Buscamos una persona con pasión por los datos y que le guste tener diferentes retos dentro del ámbito de los datos. Que disfrute interactuando con las diferentes áreas de la empresa y que sepa priorizar.
Nuestro/a DATA ENGINEER ideal es una persona que tenga formación técnica, con dominio amplio de SQL, experiencia en Python orientado a datos y que haya trabajado desarrollando ETL's leyendo de diferentes API's.
Nuestro stack de data se está montando en GCP, con Big Query como data warehouse, airbyte como sincronizador de datos, y dataform para definir y ejecutar flujos de trabajo, conjuntamente con ETL's desarrolladas en Python dentro de Google Cloud (cloud functions + cloud workflows). Por ello valoraremos el estar familiarizado/a con GCP.
El proyecto de data en Bobochoses es uno de los proyectos importantes para los siguientes años, en los que va a crecer de forma exponencial. Por ello buscamos una persona preparada, que no tenga miedo y que pueda aportar su experiencia.

En aquestes dues ofertes podem veure clarament com el científic de dades està més enfocat a la part annalística i el enginyer de dades a la part estructural, fent que l'arquitectura suporti bé les necessitats analítiques.

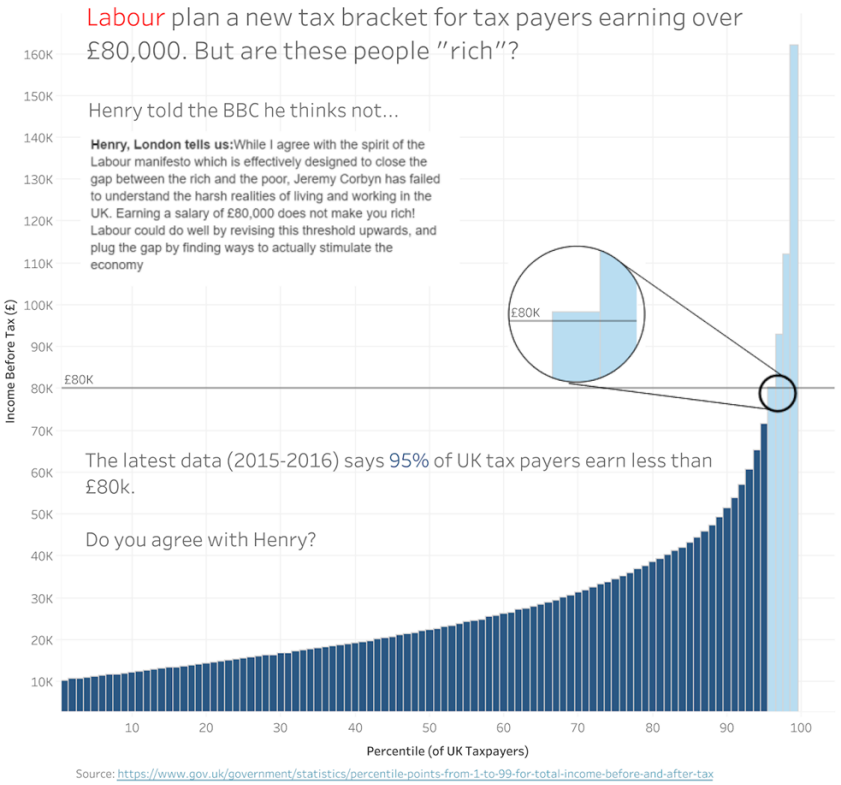
3. Esmenta les set tasques bàsiques que permeten un nivell més alt d'abstracció per a la visualització de dades i adjunta en imatges exemples on es mostrin almenys 4 d'aquestes tècniques (Màxim 100 paraules).

   - Panorama general: Tindre una vista completa dels conjunts de dades pot permetre obtenir una primera impressió de les dades disponibles.



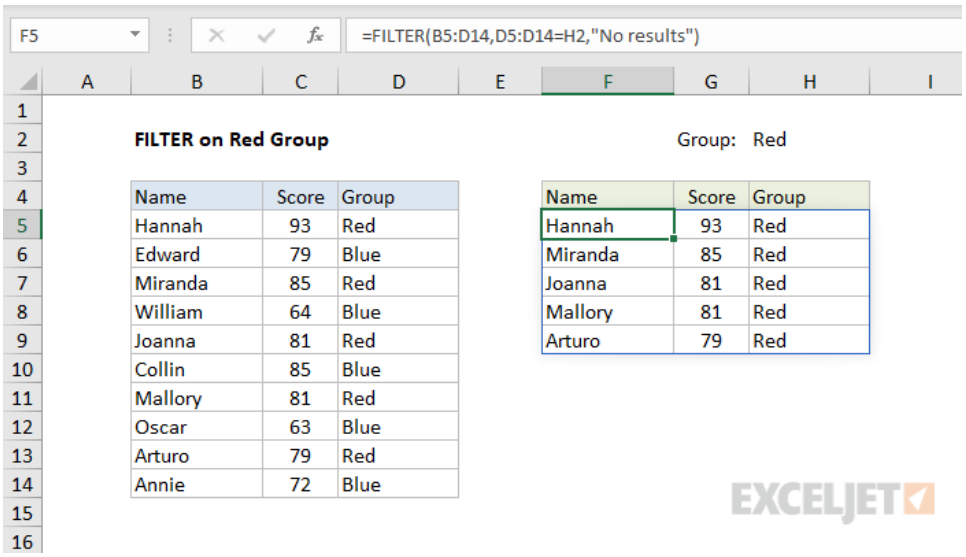
(https://siris.stanford.edu/SIRIS-Overview)

- Acostament: Ajuda a focalitzar en àrees específiques de les dades.



(https://www.flerlagetwins.com/2019/02/zoomable-charts.html)

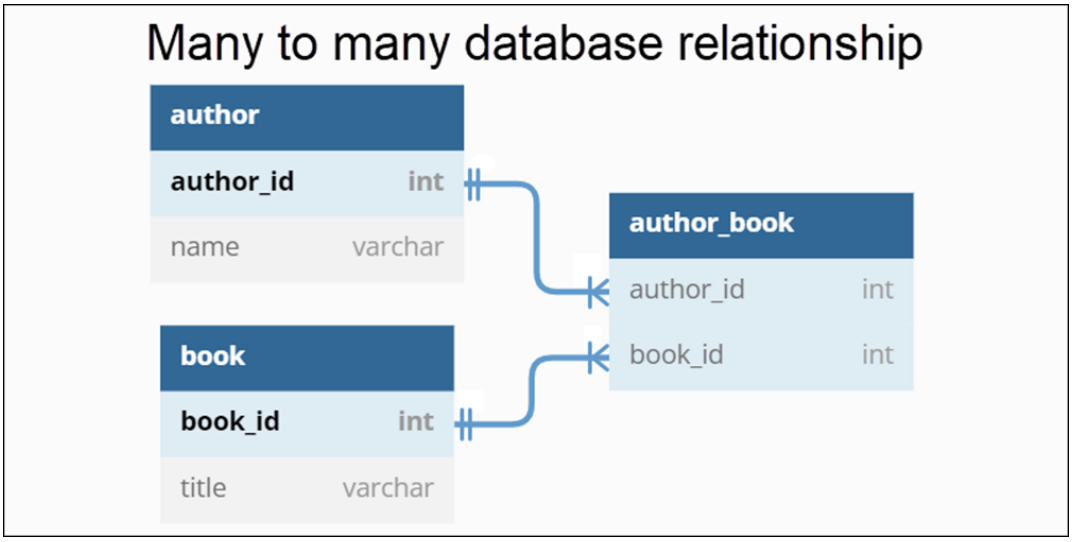
- Filtratge: Elimina informació menys rellevant o no desitjada.



(https://exceljet.net/functions/filter-function)

- Detalls a petició: Obtenció d'informació més detallada quan es requereix.

- Relacions: Identificar connexions i/o correlacions entre les dades.



(https://phoenixnap.com/kb/database-relationships)

- Historial: Permet revisar o desfer canvis anteriors i veure'n l'evolució.

- Extracció: Separar i guardar segments específics de dades.

4. Descriu les etapes del cicle de vida de les dades i proporciona un exemple per a cadascuna. (Màxim 200 paraules.)

   - Captura: Recollida de dades de diferents fonts. Exemple: Recopilar dades de clients des de formularis en línia.

   - Emmagatzematge: Guardar les dades recollides en sistemes que permetin un accés fàcil i la seva gestió. Exemple: Guardar les dades en una base de dades SQL.

   - Preprocessat: Netejar i transformacions preliminars per preparar les dades pel seu anàlisi. Exemple: Normalització de dades per a assegurar que estan en un format consistent.

   - Anàlisi: Examinar les dades processades per a l'obtenció de conclusions. Exemple: Fer servir anàlisi estadística per a identificar patrons de compra dels clients.

   - Visualització: Presentació de les dades analitzades en formats gràfics per facilitar la seva comprensió. Exemple: Crear dashboards interactius que mostren tendències de vendes.

   - Publicació: Distribució de les dades analitzades i visualitzades a stakeholders. Exemple: Enviar un informe analític mensuals als executius de l'empresa.

5. Explica la tècnica de reducció de dimensionalitat i esmenta dos mètodes utilitzats. (Màxim 100 paraules.)

Són el conjunts de mètodes que serveixen per reduir el nombre de variables aleatòries o atributs del joc de dades. Això ajuda a millorar l'eficiència en el processament i l'anàlisi.

En dos projectes que vaig treballar vam fer servir tècniques de reducció de la dimensionalitat. En el primer vam analitzar etapes del son amb senyals EEG, reduint de 10 a 3 canals via PCA. En el segon, per detectar miocardiopatia hipertròfica, vam entrenar una IA. Vam eliminar atributs que eren redundants com l'índex de massa corporal, ja que teníem l'alçada i el pes.

6. Realitza una breu recerca sobre dues eines (o llibreries) diferents per a la neteja de dades. Descriu les seves funcionalitats principals i avalua la seva conveniència per a un projecte on es requereixi eliminar duplicats, corregir errors tipogràfics i transformar formats. (Màxim 150 paraules.)

Pandas i Datacleaner són dues llibreries útils per a la neteja de dades en Python.

- Pandas: Permet manipular i analitzar dades de forma eficient. Té funcionalitats per eliminar registres duplicats, gestionar valors faltats i convertir tipus de dades. Molt útil per a la manipulació de grans conjunts de dades. (https://www.freecodecamp.org/news/data-cleaning-and-preprocessing-with-pandasbdvhj/)

- Dataclenaer: És més senzill que pandes, ja que automatitza processos com la substitució de valors faltants o la codificació de variables categòriques. Útil si es vol reduir el temps necessari per a netejar les dades (https://www.dataquest.io/blog/most-helpful-python-libraries-for-data-cleaning/)

Exercici 2 [30%]

Després de llegir el recurs “Subirats, L., Calvo, M. (2019). Web Scraping”, capítols 1 i 6. Contesta les següents preguntes amb les teves pròpies paraules:

1. Esmenta un exemple d'un lloc web real que ofereixi una API pública, però on l'ús de web scraping també podria ser beneficiós. Quines serien les raons per a considerar l'ús de totes dues tècniques i quina seria la viabilitat legal de realitzar web scraping en aquest lloc web en particular? (Màxim 100 paraules.)

X.com (l'antic Twitter) pot ser-ne un bon exemple. La seva API és publica però a la vegada fer web scarping també beneficiós. Amb la API es pot accedir a les dades de forma estructurada, però amb restriccions, com ara límits en el nombre de sol·licituds o el rang de dades accessibles (https://developer.x.com/en/docs/x-api/rate-limits). Amb web scraping podrem evitar aquests límits.

2. Posa un exemple real d'un lloc web diferent en el qual sigui interessant realitzar web scraping, però que presenti continguts dinàmics o un altre tipus d'elements que dificultin l'extracció de dades. Explica breument i amb les teves pròpies paraules quins passos utilitzaries per a avaluar la dificultat de realitzar web scraping en aquest lloc, i per què realitzaries cada pas. Ressalta en negreta una paraula per cada pas que funcioni com a títol per a aquest pas. A més, indica quines llibreries o tecnologies faries servir en el projecte. (Màxim 250 paraules.)

He triat https://www.reddit.com/, ja que aquest fòrum actualitza les dades de forma dinàmica.

1. Avaluació inicial: El primer pas es navegar per la web per comprendre l'estructura del lloc. És important fixar-se en com estan organitzades les publicacions, els comentaris i quines dades es mostren públicament.

2. Anàlisi de robots.txt: A continuació es miraria aquest fitxer per veure si hi ha restriccions específiques sobre el web scraping.
3. Monitorització de la xarxa: Analitzar les peticions que es fan quan carregues diferents parts de Reddit. Amb això es podria identificar les peticions HTPP que retornen dades JSON o fragments d'HTML amb informació útil.
4. Selecció d'eines: Tenint en compte la informació recollida als passos anteriors, es decidirà si Requests i BeatifulSoup són suficients per a les necessitats del projecte. També es podria fer servir wiresShark per a veure les peticions.

3. A l'hora de realitzar web scraping, quins serien tres bones pràctiques a tenir en compte per a evitar o reduir les possibilitats de ser bloquejat pel servidor del lloc objectiu. Expliqui-les breument. (Màxim 100 paraules)

- Revisar robot.txt: Aquest arxiu ens indica les pàgines o seccions del lloc web que estan restringides per al scraping.

- No saturar la web: Fer peticions a un ritme raonable i no de manera massiva en curts períodes de temps des de la mateixa IP.

- Ús de múltiples adreces IP: Pot ajudar a minimitzar el risc de ser identificat i bloquejat per realitzar múltiples peticions des de la mateixa adreça IP.