**Declaració de treball original (no plagi) de l'estudiant**

*Jo, NomEstudiant,, declaro que per realitzar aquest lliurament ... (completeu la frase amb les vostres pròpies paraules)*

Introducció a la ciència de dades

**PAC1: Ciència a les dades?**

|  |
| --- |
| ***Criteris d'avaluació generals de la PAC.***   * *Aportació d'alguna referència externa que complementi o sustenti els raonaments que s'exposin.* * *No ometre cap de les preguntes de cada exercici.* * *Respectar l'extensió de paraules assenyalada a cada enunciat.* * *Claredat a les respostes i els raonaments.* * *Capacitat de síntesi.* * *Originalitat.*   ***Format de lliurament***   * *Les respostes es lliuraran en* ***format PDF****.* * *Es farà servir el* ***Registre d'Avaluació Contínua*** *per pujar les respostes.* * *El document* ***no ha d'incloure l'enunciat****, només les respostes.* * *El document ha d'estar* ***estructurat*** *i el text en un color que faciliti la lectura (negre o blau fosc).* * *La* ***data màxima*** *per lliurar les respostes és el* ***31 d'octubre, a les 23:59h****. No es corregiran les PACs que no compleixin aquest requisit, excepte en casos de força major i degudament justificats.* |

**Pregunta 1 (20% puntuació)**

**Enunciat**

Al primer mòdul de l'assignatura, hem vist la **importància de les dades** en l'actual societat de la informació i l'existència de diferents perfils professionals orientats a l'anàlisi de dades i la generació de coneixement. Concretament, s'ha parlat de perfils complementaris com els *data scientist*, *data engineer*, i *data miner*.

En aquest sentit, també podem trobar-nos amb **periodistes de dades**, que tracten d’ajudar-se de visualitzacions de dades per a la il·lustració de les notícies que estan exposant i explicant.

A la següent notícia publicada en *elDiario.es*, podeu observar diversos exemples de la utilització de les dades al món del clima:

<https://www.eldiario.es/datos/ola-calor-temperatura-mapas-graficos-maximas-provincia-calentamiento-septiembre-6_1_8157324.html>

Llegiu amb atenció la notícia i responeu, segons el treball realitzat pels periodistes:

1. Quin conjunt de dades s'ha fet servir? Quines característiques té? És una font de dades fiable?
2. Quines tipologies de dades hi trobes? Dóna exemples.
3. En quina categoria de dades classificaries els comentaris dels lectors, si els hi hagués?
4. A quines preguntes clau pretén donar respostes la narrativa de la notícia? S'aprofiten les possibilitats de les visualitzacions de dades com a eines del periodisme de dades?

**Criteris d’avaluació**

* Cada pregunta es valorarà amb **0.5 punts** com a màxim.
* Extensió màxima per pregunta: **100 paraules.**
* Es valorarà la capacitat d’identificar i definir diferents tipus de dades.
* Totes les respostes han d’estar **desenvolupades** i **argumentades**.

**Pregunta 2 (30% puntuació)**

**Enunciat**

En el primer mòdul també s'han exposat les diferents etapes del **cicle de vida de les dades**, tenint en compte les seves fases més conegudes i comunes: captura, emmagatzematge, preprocessat, anàlisi, visualització, i publicació. I en el segon mòdul s'han observat diferents aspectes vinculats a les tasques d'un **científic de dades, els seus àmbits d'aplicació, i fins i tot unes primeres tècniques analítiques**.

En aquest sentit, anem a aplicar aquests continguts a la situació que estem vivint durant aquests dies de setembre i octubre de 2021, en els quals hi ha hagut una erupció volcànica a la illa de La Palma que ha creat una situació d'emergència (<https://www3.gobiernodecanarias.org/noticias/category/consejeria-seguridad-y-emergencias/>).

D'aquesta manera, l'objectiu d'aquest exercici és assumir el rol d'un científic de dades que ha de vetllar pel bon compliment d'aquest cicle de vida de les dades. Per a això, per exemple, es pot tenir en compte la següent font de dades obertes existent i relacionada amb els moviments sísmics de la illa de La Palma ([https://www.opendatalapalma.es/datasets/lapalma::movimientos-s%C3%Adsmicos/explore?location=0.000000%2C0.000000%2C0.00](https://www.opendatalapalma.es/datasets/lapalma::movimientos-s%C3%ADsmicos/explore?location=0.000000%2C0.000000%2C0.00)).

Així doncs, desenvolupeu els següents apartats:

1. Sintetitza la situació existent, i proposa un problema o pregunta concreta que volguessis ajudar a resoldre o respondre (combinar aquestes dades amb les dades de la població resident a les diferents àrees geogràfiques, detecció de tendències de més moviments sísmics, etc.). Estableix clarament **un cas d'ús sobre el qual aplicar la pregunta o hipòtesi de treball**.
2. **Captura de les dades**. Quines fonts de dades faries servir? Com les buscaries? On podries obtenir-les? Com les classificaries com a vàlides per al teu treball?
3. **Emmagatzematge i/o Preprocessat**. Com emmagatzemaries les dades? Com podries portar a terme una primera validació de la seva qualitat? Com podries millorar la seva estructura i format?
4. **Anàlisi i/o Visualització**. Què indicadors, índexs o dades calculades podries oferir per contrastar i testear la hipòtesi de treball? Quines tècniques d'anàlisis de dades i visualització faries servir?
5. **Publicació**. Com seleccionaries els millors resultats obtinguts? Quin tipus de metodologia utilitzaries? Com plantejaries explicar-los? Qui creus que serien els destinataris preferents per als resultats obtinguts?

**Criteris d’avaluació**

* Cada pregunta es valorarà amb **0.6 punts** com a màxim.
* Extensió màxima per pregunta: **100 paraules**.
* Totes les respostes han d'estar **desenvolupades i argumentades**.
* Les respostes han de mostrar connexió amb el temari teòric, de manera que es pugui percebre el seu aprenentatge.
* **Sent un exercici d'assumpció d'un rol, els errors que puguin haver-hi, en temes que no siguin estrictament sobre el cicle de vida de les dades, no seran tinguts en compte**.

**Pregunta 3 (30% puntuació)**

**Enunciat**

Als nostres dies, existeixen nombrosos casos d'utilització de la ciència de dades en tots els àmbits socioeconòmics. Considerant un nivell de desenvolupament mundial afí a l'observat durant les últimes dècades, l'ús del *Big Data* i de l'*AI* és una tendència creixent, tot i que també és cert que existeixen certs riscos associats als algorismes (sempre en constant evolució), als biaixos introduïts pels propis científics de dades i programadors (persones amb les seves pròpies limitacions), i a la qualitat i quantitat de les pròpies dades (per exemple, excessives imatges d'un tipus concret).

En aquest sentit, assenyalem **dos casos d'iniciatives d'ús de la ciència de dades** en els nostres dies:

**Cas 1**: **La Ciència de dades com a eina aplicada per a la millora de diferents negocis**.

Per exemple, l'anàlisi de les persones usuàries d'una tecnologia, duta a terme per part d'*Apple*, mitjançant l'anàlisi de la seva conducta (tribalisme, efecte dotació, prova social, heurística, efecte halo, i preu):

* + <https://dataworld.blog/2021/02/20/el-big-data-de-apple-inc/>
  + <https://www.techopedia.com/behavioral-economics-how-apple-dominates-in-the-big-data-age/2/34344>

Un altre exemple, desenvolupat per part de *Zara*, amb l'objectiu de millorar la logística i gestió de l’*estoc* a les botigues físiques:

* + <https://www.elmundo.es/tecnologia/2018/06/16/5b23f09dca474194538b4576.html>
  + <https://digitalgyan.org/zara-the-use-of-big-data-to-create-business-value/>

**Cas 2**: **La Ciència de dades com a eina aplicada per a la millora de si mateixa**.

Per exemple, el conjunt d'eines del codi obert *IA Fairness 360* dissenyat per *IBM* i amb el qual evitar biaixos en conjunts de dades i models de *machine learning*:

* + <https://zephyrnet.com/es/m%C3%A1s-desarrolladores-de-inteligencia-artificial-se-centraron-en-dise%C3%B1ar-el-sesgo-de-la-inteligencia-artificial/>
  + <https://aif360.mybluemix.net/>

Un altre exemple, la iniciativa *KELM* (*Knowledge Enhanced Languaje Model*) de *Google* que redueix el biaix de fer servir fonts d'informació no confiables:

* + <https://conectasoftware.com/ecommerce/google-presenta-kelm-una-nueva-metodologia-para-reducir-el-sesgo/>
  + <https://github.com/google-research-datasets/KELM-corpus>

Considerant tot això, **escolliu un dels anteriors dos casos d'iniciatives presentats, i redacteu una anàlisi** de 500 i 1000 paraules, recolzant-vos en els dos exemples facilitats per a cada cas o en altres exemples vostres, i fent servir la **següent estructura**:

1. **Context** de l'aplicació de la Ciència de dades en aquest cas escollit.
2. **Objectius i expectatives**. Per a què van fer servir la Ciència de Dades?
3. **Impacte** per a l'empresa i/o la societat.

Els enllaços subministrats són introductoris, de manera que podeu consultar més informació *online* sobre el cas seleccionat, sempre que citeu les vostres fonts.

**Criteris d’avaluació**

* Cada pregunta es valorarà amb **1 punt** com a màxim.
* Entendre el valor de la dada com a actiu fonamental.
* Analitzar aplicacions de Ciència de dades.
* Contextualitzar l'avantatge competitiu que aporta la Ciència de dades.
* La redacció que contingui **menys de 500 paraules o més de 1000 paraules es penalitzarà**.

**Pregunta 4 (20% puntuació)**

**Enunciat**

Al segon mòdul heu tingut ocasió de conèixer breument les denominades *dades obertes*, o ***Open Data***, que són, cada dia més, una creixent i importantíssima font de dades per a empreses, investigadors, periodistes, ciutadans i administracions públiques. Aquestes dades són facilitades al món des de projectes i plataformes tecnològiques tant públiques com a privades.

Per exemple, a l'àmbit públic, trobem el portal de transparència de dades obertes del *Govern d'Espanya*: <https://datos.gob.es/es>

Així mateix, també trobem iniciatives privades com la desenvolupada per l'empresa *Esri Espanya*, orientada cap a dades vinculades amb Sistemes d'Informació Geogràfica (SIG): <https://opendata.esri.es/>

Entreu en aquest darrer portal d'iniciativa privada, i responeu:

1. Què són les dades obertes? Quines són les seves principals característiques?
2. Descriu les funcionalitats del portal, la manera de navegació, i la llicència d'ús de les dades.
3. Quines possibilitats d'accés i visualització de dades obertes s'ofereixen? Com és la descàrrega de dades?
4. Quins tipus de dades podem trobar al portal? Quines etapes del cicle de vida de les dades poden ser trobades al portal?

**Criteris d’avaluació**

* Cada pregunta es valorarà amb **0.5 punts** com a màxim.
* Comprendre el significat de *Open Data* i el seu ús.
* Explorar el portal de cerca de dades obertes i comprendre el seu funcionament i contingut.
* Extensió màxima per pregunta: **100 paraules**.