

# Introducció a la ciència de dades.

## PAC2: Exemples pràctics d'anàlisi.

### Presentació

*Avaluació contínua (AC) dels continguts lectius corresponents:*

- *Bloc 2: La Ciència de Dades.*
  - o *Contingut: Els rols, àmbits i noms de la Ciència de Dades (autor: Marçal Mora).*
  - o *Aspectes vinculats a:*
    - *Processos.*
    - *Tècniques d'aprenentatge automàtic.*
- *Bloc 3: Exemples de projectes de Ciència de Dades.*
  - o *Contingut: Exemples de projectes en l'àmbit de la Ciència de Dades (autor: Marçal Mora).*
  - o *Aspectes vinculats a la visualització de les dades.*

### Objectius i competències

- *Entendre els conceptes utilitzats habitualment i relacionats amb el context de la Ciència de Dades.*
- *Entendre les fases del cicle de vida de la dada.*
- *Entendre la necessitat de posar en pràctica altres competències a l'hora de treballar posant-nos en el paper d'un científic de dades:*
  - o *Entendre les fases del procés ETL.*
  - o *Construir visualitzacions per a l'anàlisi de dades.*
  - o *Entendre els diferents tipus d'aprenentatge automàtic i metodologies de la Ciència de 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 en les respostes i en els raonaments.*
- *Capacitat de síntesi.*
- *Originalitat.*

### Criteris de valoració dels exercicis

- *Indicats a l'enunciat.*

### Format i data 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).*
- *Cal assegurar que **el nom de l'estudiant consti en el/s document/s** (per exemple: a la coberta i al peu de pàgina) i que els documents s'han enviat/carregat correctament.*
- *La **data màxima** per lliurar les respostes és el **28 de novembre, a les 23:59h**. No es corregiran les PACs que no compleixin aquest requisit, excepte en casos de força major i degudament justificats.*

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

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

## Pregunta 1: Extracció, Transformació i Càrrega de dades (ETL) (35%)

Com hem pogut veure en el subapartat 4.4. del segon mòdul teòric d'aquesta assignatura, el primer procés que es sol seguir en la Ciència de Dades és **Extreure, Transformar i Carregar** les dades, o el que es sol anomenar com el procés **ETL**.

Per poder analitzar les nostres dades, primer necessitem tenir-les emmagatzemades en algun lloc. Per exemple, podem tenir les nostres dades pujades en una base de dades, un full de càlcul d'Excel, una pàgina web, etc. No obstant això, independentment del sistema que utilitzem per emmagatzemar les nostres dades, no podem pujar-les directament sense eliminar primer possibles anomalies, assegurar-nos que el format sigui el correcte, etc. Per aquest motiu, necessitem definir el procés ETL que seguirem, de manera que sigui reproducible i escalable en qualsevol moment.

Per poder entendre la importància d'aquest procés, començarem experimentant els problemes que podem trobar sense un ETL correcte.

### 1.1. Extreure les dades

Per poder fer aquest exercici, el primer pas és obtenir les dades. En aquest sentit, ens descarregarem el fitxer "P12-Bank-Customers-Demo.csv" que podem trobar al següent enllaç: <https://www.superdatascience.com/pages/training>, a l'apartat *Part 3: Data Prep - Section 2. ETL Phase 1: Data Wrangling Before the Load*.

Una vegada que hagueu pogut descarregar el fitxer en una carpeta del vostre ordinador, el podeu obrir amb l'editor de text que preferiu. A continuació, responeu les següents preguntes:

1. Descriu els atributs que conté.
2. Sabries identificar el "tipus" de dada (simple, compost, ...) de cada atribut?

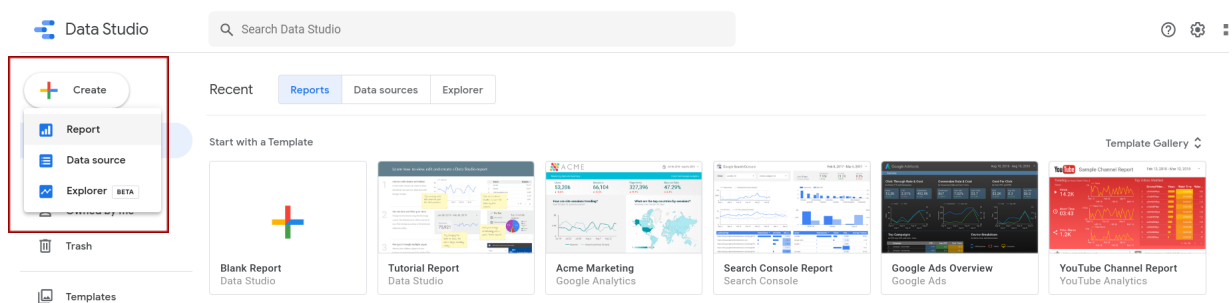
### 1.2. Carregar les dades i analitzar-les

Imagineu que volem obtenir el balanç total per cada dia de registre, és a dir, volem saber quants diners tenen en total els usuaris que es van registrar el dia 5 de gener, el 6 de gener, etc. El procediment correcte seria avaluar primer si cal transformar aquestes dades. Però, de moment, ometrem aquest pas (ho farem més endavant un cop comprovat què passa al carregar-les tal i com estan) i procedirem a carregar-les en una eina per analitzar-les.

Per això, farem servir l'eina gratuïta de Google anomenada “*Data Studio*”. Aquesta eina ens permet analitzar les nostres dades i crear visualitzacions de forma fàcil.

**1.2.1.** Per poder carregar les nostres dades, seguirem aquests passos:

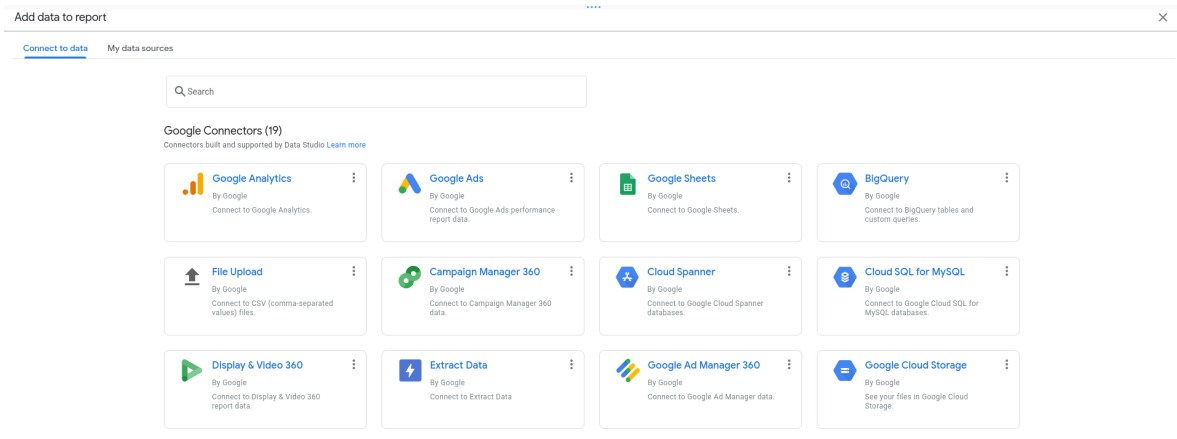
1. Entrarem amb el nostre usuari de la UOC a <https://datastudio.google.com>
2. Seleccionarem “**Create**”<sup>1</sup> i, a continuació, “**Report**” tal com es mostra en la imatge següent<sup>2</sup>:



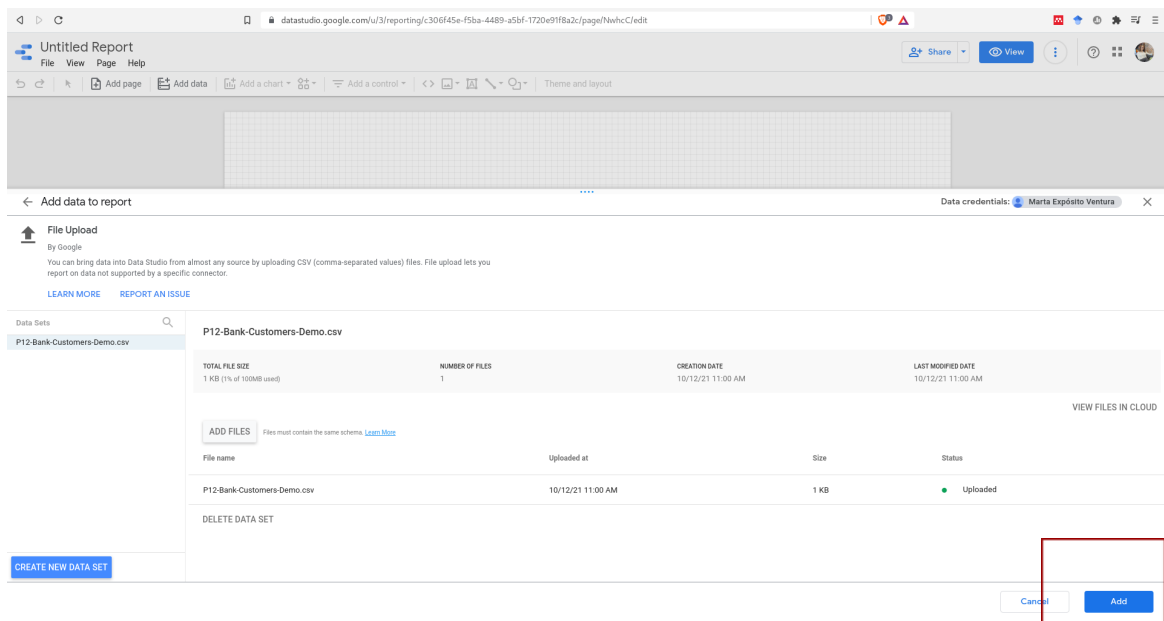
3. Ens apareixerà una pestanya per carregar les nostres dades, amb l'opció “**File Upload**” situada a l'esquerra i en la fila del mig.

<sup>1</sup> Pot ser que tingueu l'eina en un altre idioma. En aquest cas, pot ser que aquests noms us surtin diferents. De totes maneres, podeu seguir els passos mirant les imatges que es mostren en cadascun d'ells, ja que totes les opcions es trobaran en els mateixos llocs de l'eina.

<sup>2</sup> També pot ocórrer que se us sol·licitin dades de configuració (país, recepció d'emails amb novetats, etc.) si no heu utilitzat abans aquesta eina.



4. Seleccionarem l'opció **“File Upload”** i pujarem el fitxer que ens hem descarregat anteriorment.



5. Una vegada el **“Status”** aparegui com **“Uploaded”**, podem carregar les dades seleccionant el botó **“Add”**. Us apareixerà una pestanya per confirmar que voleu pujar aquestes dades. Simplement doneu-li a **“Add to Report”**.

## You are about to add data to this report

 P12-Bank-Customers-Demo.csv

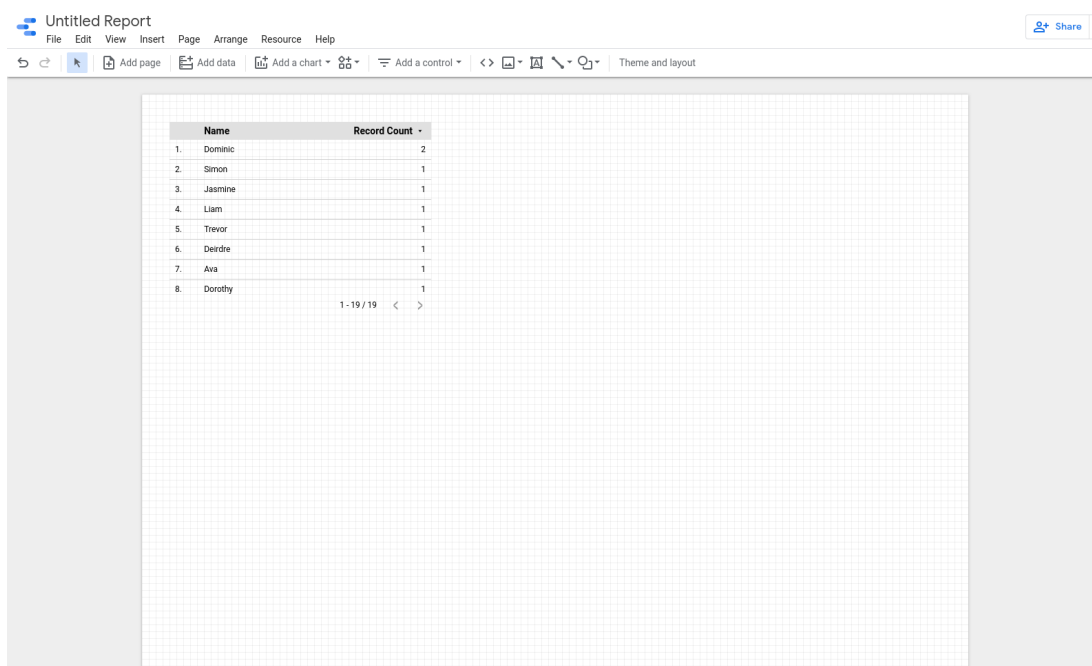
Note that **Report Editors** can create charts using the new data source(s), and can add dimensions and metrics not currently included in the report.

☐ Don't show me this again

CANCEL

ADD TO REPORT

### 6. Una vegada s'hagin pujat les dades, se us obrirà la següent finestra:



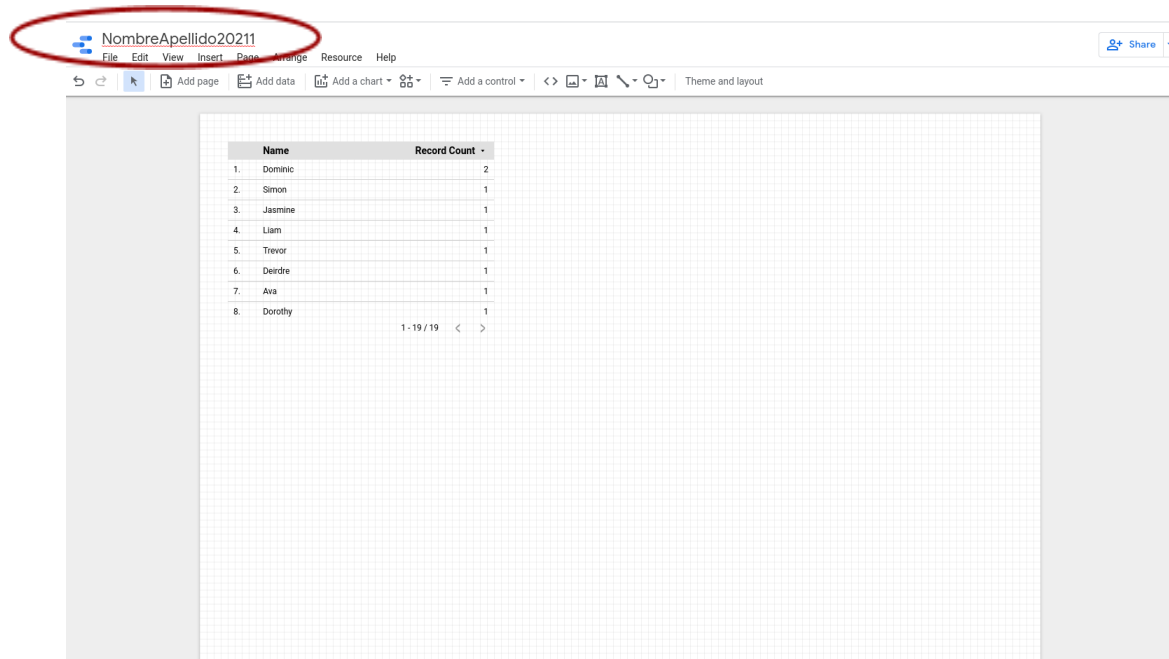
The screenshot shows a web application window titled 'Untitled Report'. The interface includes a menu bar (File, Edit, View, Insert, Page, Arrange, Resource, Help) and a toolbar with various icons for adding data, charts, and controls. The main content area displays a table with the following data:

	Name	Record Count
1.	Dominic	2
2.	Simon	1
3.	Jasmine	1
4.	Liam	1
5.	Trevor	1
6.	Deirdre	1
7.	Ava	1
8.	Dorothy	1

At the bottom of the table, there is a pagination indicator: '1 - 19 / 19' with navigation arrows.

Ja tenim les dades pujades!

A continuació, anomenarem el document amb el nostre nom seguint aquest format: **NombreApellido20211**.



**1.2.2.** Ara que ja tenim les nostres dades carregades, anem a crear una taula on es vegi el balanç total per data de registre. Per fer això, seguirem aquests passos:

1. Primer de tot, seleccionarem la taula que ens ha aparegut per defecte amb el Nom (*Name*) i el Nombre de Registres (*Record Count*). Una vegada seleccionat, en el panell de la dreta veurem aquestes dades:

Chart > Table

DATA

STYLE

Data source

P12-Bank-Custom...

+

BLEND DATA

Date Range Dimension

Date Joined

Dimension

ABC

Name

+

Add dimension

Drill down

Metric

AUT

Record Count

+

Add metric

Optional metrics

Metric sliders

Rows per Page

100

Summary row

Show summary row

Sort

AUT

Record Count

Descending

Ascending

Available Fields

Q

Type to search

123

Age

ABC

Balance

123

Customer ID

Date Joined

ABC

Gender

ABC

Name

ABC

Surname

123

Record Count

+

ADD A FIELD

+

ADD A PARAMETER

2. Eliminarem el Nom (*Name*) de les dimensions (*Dimension*) seleccionant la creu ("x") que apareix a l'esquerra:

Dimension

ABC

Name

+

Add dimension

3. A continuació, afegirem una nova dimensió, la Data de Registre (*Date Joined*):

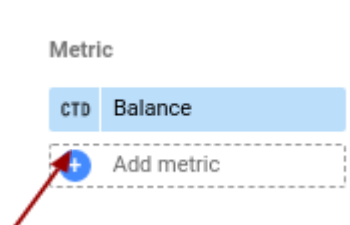
Dimension

Date Joined

+

Add dimension

4. Veurem que la taula que teníem abans en el document s'ha actualitzat automàticament, mostrant les dates. Ara ens falta afegir els Diners (*Balanç*) per a cada data, en lloc del nombre de registres. Per a això, anirem a l'apartat de les Mètriques (*Metric*) i treurem el nombre de registres (*Record Count*), tal com hem fet abans amb la dimensió dels noms (*Name*). A continuació afegirem *Balanç* com a mètrica, amb el botó “Add metric”, tal com es mostra en la imatge següent:

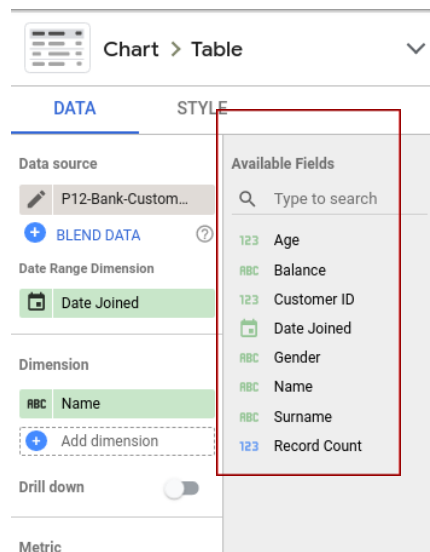


5. En aquest moment, se'ns ha actualitzat la taula amb uns nombres que no corresponen a la suma total dels diners que tenim per data. Això és a causa que en lloc de sumar el balanç de cada usuari per dia, Google Studio ha representat el nombre de registres amb diners per dia. Així per exemple, el dia 15 de gener ens apareix el valor “2” com a *Balanç*, ja que tenim dues files de dades en aquest dia:

A	B	C	D	E	F	G
Customer ID	Name	Surname	Gender	Age	Date Joined	Balance
100000001	Simon	Walsh	Male	21	January 5, 2015	\$113,810.15
400000002	Jasmine	Miller	Female	34	January 6, 2015	\$36,919.73
100000003	Liam	Brown	Male	46	January 7, 2015	\$101,536.83
300000004	Trevor	Parr	Male	32	January 8, 2015	\$1,421.52
100000005	Deirdre	Pullman	Female	38	January 9, 2015	\$35,639.79
300000006	Ava	Coleman	Female	30	January 9, 2015	\$122,443.77
100000007	Dorothy	Thomson	Female	34	January 11, 2015	\$42,879.84
200000008	Lisa	Knox	Female	48	January 11, 2015	\$36,680.17
300000009	Ruth	Campbell	Female	33	January 11, 2015	\$74,284.35
100000010	Dominic	Parr	Male	42	January 12, 2015	\$10,912.45
100000011	Dominic	Lewis	Male	40	January 12, 2015	\$39,667.83
100000012	Benjamin	Grant	Male	39	January 12, 2015	\$32,281.62
100000013	Ryan	MacDonald	Male	24	January 12, 2015	\$40,781.63
200000014	Thomas	Lawrence	Male	46	January 12, 2015	\$48,791.46
300000015	Madeleine	Marshall	Female	36	January 12, 2015	\$2,846.03
100000016	Nicholas	Newman	Male	42	January 14, 2015	\$2,116.85
200000017	Grace	Hill	Female	31	January 14, 2015	\$10,356.31
200000018	Samantha	Coleman	Female	42	January 14, 2015	\$3,801.69
100000019	William	Ince	Male	40	January 15, 2015	\$65,534.69
100000020	Audrey	Jones	Female	46	January 15, 2015	\$11,462.64

Per entendre per què passa això, podem mirar en la pestanya “Available Fields”. Google Studio ens mostra els atributs que conté la nostra base de dades i, a la seva dreta, el tipus de dades. Així, per exemple, veiem que l'atribut de l'edat (*Age*) l'interpreta com un nombre, ja que ens apareix el símbol “123”. No obstant això, l'atribut dels diners (*Balanç*) ens apareix el símbol “ABC”, per tant, l'interpreta com un text. Per aquest motiu, no pot realitzar la suma d'aquesta mètrica.





A continuació, **adjunteu una captura de pantalla** on es vegi el resultat del procés que has seguit, amb la data i els diners en la taula, i on es vegi el teu nom en el format que hem indicat anteriorment. En definitiva, has de facilitar-nos una captura de pantalla similar a la imatge que et mostrem a continuació.

NombreApellido20211

File Edit View Insert Page Arrange Resource Help

↶ ↷ ⚙ Add page 📄 Add data 📊 Add a chart ⚙ Add a control < > 🖨 Theme and layout

	Date Joi...	Balance
1.	Jan 15, 20...	2
2.	Jan 14, 20...	3
3.	Jan 12, 20...	6
4.	Jan 11, 20...	3
5.	Jan 9, 2015	2
6.	Jan 8, 2015	1
7.	Jan 7, 2015	1
8.	Jan 6, 2015	1

1 - 9 / 9 < >

## Exercici 1.3. Transformar les dades

Com hem vist, abans de posar-nos a analitzar les nostres dades, és important dedicar un temps a assegurar-nos que estan en el format correcte.

**1.3.1.** A continuació, anem a transformar les nostres dades per analitzar-les bé:

1. Anem a obrir de nou el nostre fitxer *P12-Bank-Customers-Demo.csv* amb un editor de text de la nostra elecció.
2. Una vegada obert, veureu que cada atribut de cada fila està separat per comes:

```

1 Customer ID,Name,Surname,Gender,Age,Date Joined,Balance
2 100000001,Simon,Walsh,Male,21,"January 5, 2015","$113,810.15"
3 400000002,Jasmine,Miller,Female,34,"January 6, 2015","$36,919.73"
4 100000003,Liam,Brown,Male,46,"January 7, 2015","$101,536.83"
5 300000004,Trevor,Parr,Male,32,"January 8, 2015","$1,421.52"
6 100000005,Deirdre,Pullman,Female,38,"January 9, 2015","$35,639.79"
7 300000006,Ava,Coleman,Female,30,"January 9, 2015","$122,443.77"
8 100000007,Dorothy,Thomson,Female,34,"January 11, 2015","$42,879.84"
9 200000008,Lisa,Knox,Female,48,"January 11, 2015","$36,680.17"
10 300000009,Ruth,Campbell,Female,33,"January 11, 2015","$74,284.35"
11 100000010,Dominic,Parr,Male,42,"January 12, 2015","$10,912.45"
12 100000011,Dominic,Lewis,Male,40,"January 12, 2015","$39,667.83"
13 100000012,Benjamin,Grant,Male,39,"January 12, 2015","$32,281.62"
14 100000013,Ryan,MacDonald,Male,24,"January 12, 2015","$40,781.63"
15 200000014,Thomas,Lawrence,Male,46,"January 12, 2015","$48,791.46"
16 300000015,Madeleine,Marshall,Female,36,"January 12, 2015","$2,846.03"
17 100000016,Nicholas,Newman,Male,42,"January 14, 2015","$2,116.85"
18 200000017,Grace,Hill,Female,31,"January 14, 2015","$10,356.31"
19 200000018,Samantha,Coleman,Female,42,"January 14, 2015","$3,801.69"
20 100000019,William,Ince,Male,40,"January 15, 2015","$65,534.69"
21 100000020,Audrey,Jones,Female,46,"January 15, 2015","$11,462.64"
22

```

3. D'altra banda, veureu que els valors de *Balanç* contenen cometes i el símbol del dòlar (per exemple: "\$65,534.69").
4. Per evitar problemes en pujar-ho a Google Studio (o a qualsevol altra base de dades o eina), anem a procedir a:
  - a. Eliminar les **cometes** del *Balanç*.
  - b. Eliminar el **símbol** del dòlar del *Balanç*.
  - c. Eliminar les **comes** que s'usen per indicar els milers en els nombres → Atès que el fitxer separa els atributs per comes, si no ho eliminem en *Balanç*, quan ho pugem a Google Studio creurà que hi ha un altre atribut més.

(*Depenent de l'editor de text que feu servir, aquest procés es pot fer bastant ràpid amb l'ajuda de les eines de "Buscar i reemplaçar"*)

5. Hauries de tenir al final d'aquest procés el fitxer així:

```

1 Customer ID,Name,Surname,Gender,Age,Date Joined,Balance
2 100000001,Simon,Walsh,Male,21,"January 5, 2015",113810.15
3 400000002,Jasmine,Miller,Female,34,"January 6, 2015",36919.73
4 100000003,Liam,Brown,Male,46,"January 7, 2015",101536.83
5 300000004,Trevor,Parr,Male,32,"January 8, 2015",1421.52
6 100000005,Deirdre,Pullman,Female,38,"January 9, 2015",35639.79
7 300000006,Ava,Coleman,Female,30,"January 9, 2015",122443.77
8 100000007,Dorothy,Thomson,Female,34,"January 11, 2015",42879.84
9 200000008,Lisa,Knox,Female,48,"January 11, 2015",36680.17
10 300000009,Ruth,Campbell,Female,33,"January 11, 2015",74284.35
11 100000010,Dominic,Parr,Male,42,"January 12, 2015",10912.45
12 100000011,Dominic,Lewis,Male,40,"January 12, 2015",39667.83
13 100000012,Benjamin,Grant,Male,39,"January 12, 2015",32281.62
14 100000013,Ryan,MacDonald,Male,24,"January 12, 2015",40781.63
15 200000014,Thomas,Lawrence,Male,46,"January 12, 2015",48791.46
16 300000015,Madeleine,Marshall,Female,36,"January 12, 2015",2846.03
17 100000016,Nicholas,Newman,Male,42,"January 14, 2015",2116.85
18 200000017,Grace,Hill,Female,31,"January 14, 2015",10356.31
19 200000018,Samantha,Coleman,Female,42,"January 14, 2015",3801.69
20 100000019,William,Ince,Male,40,"January 15, 2015",65534.69
21 100000020,Audrey,Jones,Female,46,"January 15, 2015",11462.64
22

```

6. Finalment, ho guardarem i ho pujarem de nou a Google Data Studio. A partir d'aquí heu de repetir tots els passos de l'exercici **1.2.2.** però utilitzant aquest nou fitxer.

A continuació, adjunteu una captura de pantalla on es vegi **el vostre nom i cognom en el títol** (amb el format d'abans) i la suma total dels diners per a cada dia de registre.

**1.3.2.** Què opines de l'atribut "Date Joined"? Creus que el format actual és el més indicat per poder aplicar qualsevol tipus d'anàlisi en el futur? Quina bona pràctica proposaries en relació a transformar el format d'aquest atribut?

## Criteris d'avaluació

- La pregunta 1.1 es valorarà amb **1 punt** com a màxim.
- La pregunta 1.2 es valorarà amb **1 punt** com a màxim.
- La pregunta 1.3.1 es valorarà amb **1 punt** com a màxim.
- La pregunta 1.3.2. es valorarà amb **0.5 punts** com a màxim.
- Les preguntes amb captures de pantalla que no mostrin el nom del document amb el format *NombreApellido20211* no es valoraran.
- Les captures de pantalla aportades han de contenir una explicació del seu contingut d'aproximadament unes 100 paraules.

## Pregunta 2: Exploració de dades (35 %)

Una vegada tenim les nostres dades carregades de forma correcta, podem procedir a analitzar-les. Per a això, la visualització de les dades pot ser una eina molt útil, ja que ens permet obtenir una representació gràfica de les nostres dades i treure unes primeres conclusions de forma ràpida i visual.

Per entendre millor els avantatges que la visualització de dades ens proporciona, utilitzarem el dataset *P12-Churn-Modelling.csv*. Aquest dataset es troba en <https://www.superdatascience.com/pages/training> en format Excel, però per poder-lo fer servir en aquest exercici necessitem convertir-ho a format “CSV”. En aquest sentit, **podeu trobar el document en format csv directament a l'aula**, al costat del document de l'enunciat d'aquesta PAC<sup>3</sup>.

Aquest document conté 10.000 files amb dades (inventades) de clients d'un banc. Aquestes dades contenen informació relativa al gènere del client, la seva nacionalitat, edat, entre d'altres. Farem servir aquest document per solucionar un cas d'ús de Ciència de Dades.

### Exercici 2.1. Visualització de dades

Imagineu que aquest banc us ha contractat perquè recentment ha detectat que molts clients han estat abandonant el seu banc. Amb base en la seva experiència, la taxa d'abandonament sol ser menor, i els agradaria detectar què està passant i com solucionar aquesta situació.

Per fer aquesta anàlisi, el banc va seleccionar 10.000 clients aleatòriament de la seva base de dades i després de 6 mesos, es va anotar per a cadascun d'ells si havia abandonat el banc. Aquesta dada, la tenim en el document amb el nom “**Exited**”, i pot prendre el valor 0 si la persona segueix al banc, o el valor 1 si ja no hi és.

Anem a procedir a fer una mica de visualització de dades per esbrinar per on podria venir el problema. Possiblement després podrem entrar a analitzar amb més detall, però la visualització de dades ens donarà un lloc per on començar.

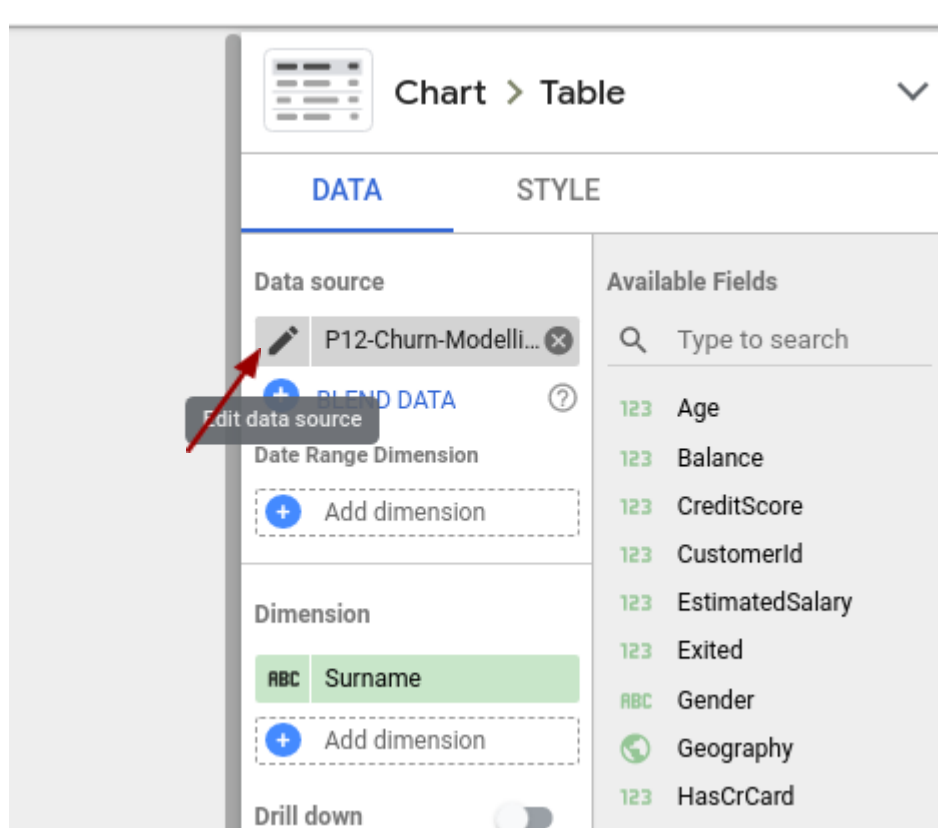
Així doncs, anem a seguir aquests passos:

1. Tornarem a Google Data Studio i obrirem un nou projecte.
2. Aquesta vegada, pujarem les dades del fitxer *P12-Churn-Modelling.csv*

<sup>3</sup> També pot ser descargable desde la pàgina web indicada, en: Part 2: Modelling - Section 5. Building a Robust Geodemographic Segmentation Model. Churn Modelling.

P12-Churn-Modelling.csv			
TOTAL FILE SIZE 659 KB (1% of 100MB used)	NUMBER OF FILES 1	CREATION DATE 10/12/21 6:45 PM	LAST MODIFIED DATE 10/12/21 6:45 PM
VIEW FILES IN CLOUD			
<div>ADD FILES</div> <div>Files must contain the same schema. <a href="#">Learn More</a></div>			
File name	Uploaded at	Size	Status
P12-Churn-Modelling.csv	10/12/21 6:45 PM	659 KB	● Uploaded
DELETE DATA SET			
<div>Cancel</div> <div>Add</div>			

3. A continuació, comprovarem que Google Data Studio ha interpretat bé les nostres dades per evitar futurs problemes. Per a això, anem a la pestanya dreta on posa “Data Source” i li donem a “*Edit Data Source*” (mireu la imatge següent per veure on se situa aquesta opció).



4. Se'ns obrirà una finestra amb totes les dimensions del nostre document i el seu tipus de dada. Al verificar si estan tots bé veiem que la dimensió “*Exited*” la interpreta com un valor numèric.

P12-Churn-Modelling.csv

← EDIT CONNECTION | FILTER BY EMAIL



Data source editors can now refresh fields, edit connections, and edit custom SQL.

Field ↓	Type ↓	Default Aggregation ↓	De
DIMENSIONS (14)			
Age	123 Number	Sum	
Balance	123 Number	Sum	
CreditScore	123 Number	Sum	
CustomerId	123 Number	Sum	
EstimatedSalary	123 Number	Sum	
Exited	123 Number	Sum	
Gender	ABC Text	None	

REFRESH FIELDS

5. Atès que en aquest cas és una variable categòrica, que ens indica si el client segueix estant actiu al banc o no, ho convertirem a tipus **Text**. A continuació, li donem a “**Done**” per continuar.

P12-Churn-Modelling.csv

← EDIT CONNECTION | FILTER BY EMAIL



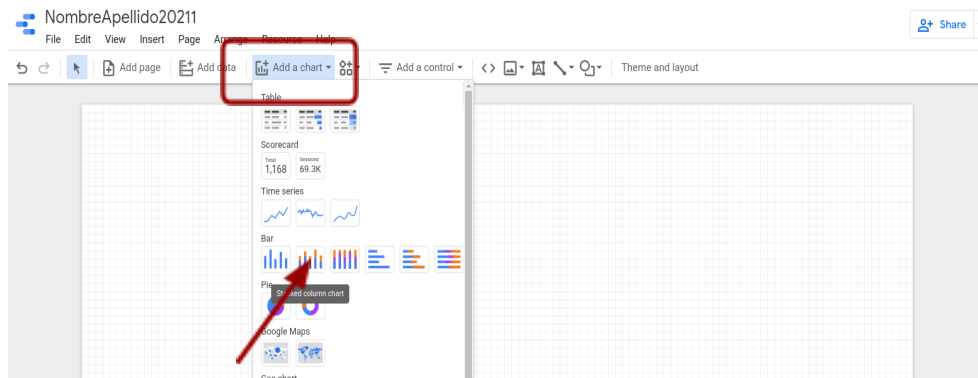
Data source editors can now refresh fields, edit connections, and edit custom SQL.

Field ↓	Type ↓	Default Aggregation ↓	
DIMENSIONS (14)			
Age	123 Number	Sum	
Balance	123 Number	Sum	
CreditScore	123 Number	Sum	
CustomerId	123 Number	Sum	
EstimatedSalary	123 Number	Sum	
Exited	ABC Text	None	
Gender	ABC Text	None	

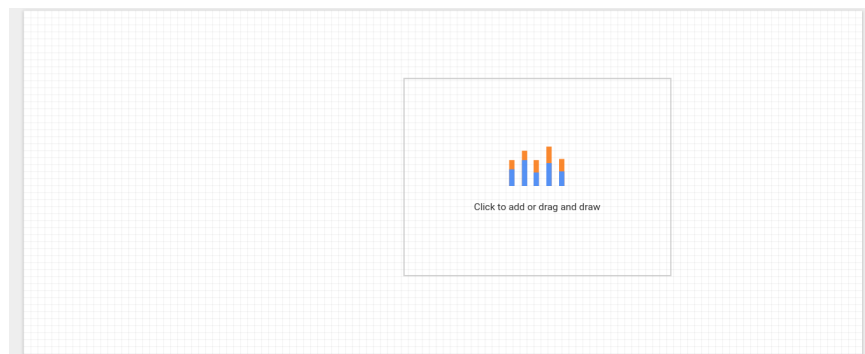
REFRESH FIELDS

6. Ara estem llestos per a realitzar la nostra primera visualització. Anem a fer un cop d'ull al comportament de la nostra variable categòrica (*Exited*) pel que fa al gènere dels nostres clients. Per a això, podem utilitzar una gràfica de barres. Seleccionarem

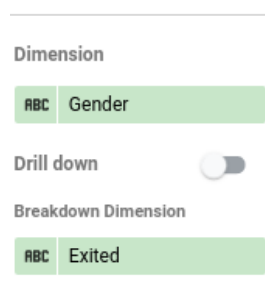
“**Add a chart**” en la nostra finestra principal, i a continuació la gràfica tipus “**Stacked column chart**”.



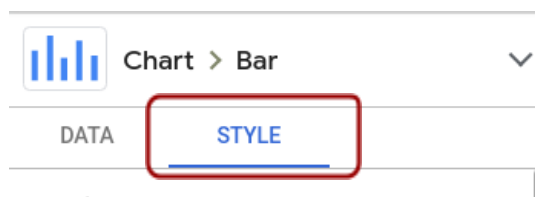
- Tot seguit, ens apareixerà un quadrat que podem arrossegar a on més ens agradi. Escollim un lloc i cliquem en ell.



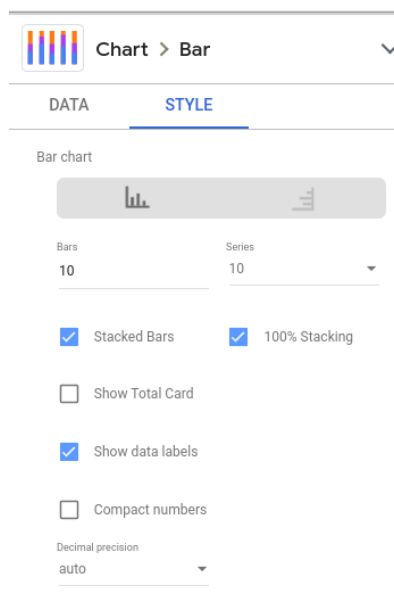
- Anem a procedir a modificar aquesta gràfica per a que ens mostri el gènere dels nostres clients i la seva relació amb la nostra variable *Exited*. Per a això, en el panell de la dreta seleccionem “**Gender**” com *Dimension* i “**Exited**” on posa “**Breakdown Dimension**”.



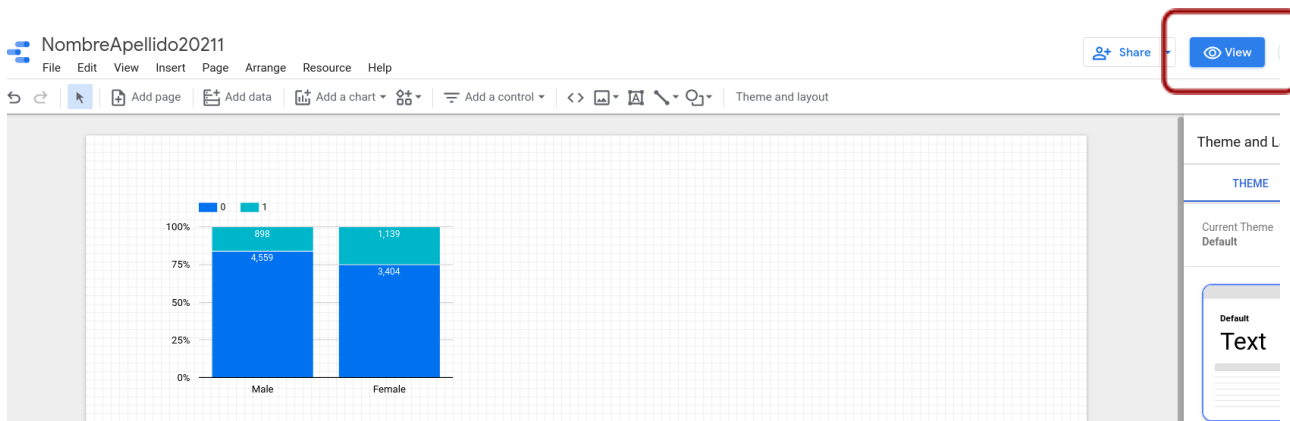
- Finalment, anem a editar l'estil de la nostra gràfica perquè sigui més informativa. Per a això, ens dirigim al panell de la dreta, i seleccionem “**STYLE**”.



10. A continuació, seleccionem les següents opcions: “**Stacked Bars**”, “**100% Stacking**” i “**Show data labels**”.



11. Veurem que la nostra gràfica de barres s'ha actualitzat automàticament. Ara que ja tenim la nostra gràfica al nostre gust, podem visualitzar el resultat final seleccionant el botó “**View**”.



12. Se'ns obrirà una pestanya nova on podem passar el cursor per sobre de la nostra gràfica de barres i ens sortirà informació complementària, com el percentatge d'usuaris masculins i femenins que s'han anat i que s'han quedat.

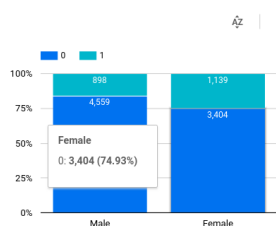


NombreApellido20211

Reset

Share

Edit



En aquest punt, **adjunteu una captura de pantalla** on es vegi aquesta visualització i en la qual **aparegui el vostre nom i cognoms** en el format *NombreApellido20211*.

**Quines conclusions podeu treure d'aquesta gràfica?** Redacteu les vostres conclusions amb una extensió entre 150 i 400 paraules.

## Exercici 2.2. Presentació de resultats

En l'exercici anterior hem vist un exemple de visualització de les nostres dades. No obstant això, tenim molts més tipus de gràfics i moltes altres dimensions que explorar. A continuació, **escolliu 3 tipus de gràfics diferents** que analitzin altres dimensions i **la seva relació amb la variable categòrica "Exited"**.

Una vegada els tingueu creats, feu una captura de pantalla del resultat final en la manera **"View"** (com en la pregunta 2.1.12).

Finalment, **redacteu les conclusions** que pugueu observar en funció d'aquestes gràfiques. Aquesta redacció vindrà acompanyada de la vostra captura de pantalla, així que tingueu present que el lector estarà veient les gràfiques conforme avanceu en la vostra explicació. Les vostres conclusions han d'estar desenvolupades, per la qual cosa cada gràfica ha de tenir una explicació de entre 150 i 400 paraules.

## Criteris d'avaluació

- La pregunta 2.1 es valorarà amb **1 punt** com a màxim.
- La pregunta 2.2. es valorarà amb **2.5 punts** com a màxim.
- Les respostes amb captures de pantalla que no mostrin el nom del document amb el format *NombreApellido20211* no es valoraran.
- Les respostes que no compleixin els requisits de nombre de paraules no es valoraran.
- En aquesta pregunta es valorarà la creativitat per crear visualitzacions capaces de donar respostes a l'enunciat.

## Pregunta 3: Minería de datos (30 %)

En les preguntes anteriors hem vist dos processos de la Ciència de Dades. D'una banda hem vist la importància de definir un bon procés de càrrega i transformació de les dades. Seguidament hem vist com podem realitzar una primera anàlisi de les nostres dades mitjançant la seva visualització.

No obstant això, en molts contextos, després d'entendre les nostres dades podem necessitar dissenyar aplicacions que les utilitzin amb finalitats més específiques i analítiques. Aquest procés es denomina *Data mining*, i pot desenvolupar-se de diverses maneres.

### Exercici 3.1. Tipus de tècniques d'aprenentatge automàtic

A continuació ens centrarem en les tècniques **d'aprenentatge automàtic**. Tal com es presenta a l'apartat 4.5 del mòdul teòric 2, en l'aprenentatge automàtic podem trobar **tècniques supervisades, tècniques no supervisades i tècniques de reforç**. En aquest sentit, desenvolueu els següents apartats:

**3.1.1.** Descriviu **amb les vostres paraules** les diferències entre els tres tipus d'aprenentatge automàtic.

**3.1.2.** Indiqueu tres exemples de cas d'ús on es podria aplicar cada tipus d'aprenentatge automàtic.

D'altra banda, independentment del tipus de tècnica d'aprenentatge que necessitem utilitzar, de vegades necessitem reduir la **dimensionalitat** de les nostres dades. En base a la teoria que hem vist a l'apartat 4.5.2 del mòdul teòric 2, i altres fonts d'informació que podeu buscar en internet, contesteu:

**3.1.3.** Quins beneficis i desavantatges presenta la reducció de dimensionalitat en les nostres dades?

Recordeu que totes les preguntes han d'estar **desenvolupades amb les vostres paraules**, és a dir, no copieu directament el contingut de cap font. D'altra banda, **recordeu citar la/les font/s** d'informació utilitzades per obtenir informació.

### Exercici 3.2. Anàlisi d'un model d'aprenentatge automàtic.

Finalment, com a Científics de Dades, necessitarem validar que el nostre model d'aprenentatge automàtic funcioni correctament. Per a això, ens pot ser útil tenir un conjunt de dades de prova amb els quals validar el model. És important que aquestes dades de

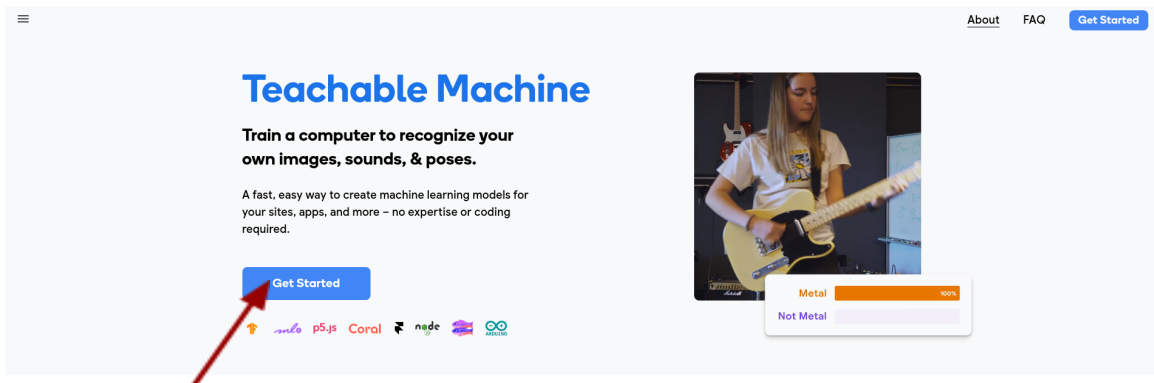
prova siguin noves, en el sentit que l'algorisme no les “hagi vist” al moment de l'entrenament, per evitar problemes de sobre entrenament (*overfitting*).

Per entendre millor com funciona aquesta validació, anem a veure un cas pràctic. Imagineu-vos que volem entrenar un model capaç d'identificar si un peix està enverinat. Per fer això, ens han facilitat un conjunt d'imatges amb exemples de peixos enverinats (peixos amb els ulls en forma de cercle) i de peixos sans (peixos amb els ulls amb un “-”). Podem trobar aquestes imatges en diferents carpetes en el següent enllaç: [Fish Images for Upload](#)<sup>4</sup>

Com podem veure, tenim els dos tipus de peixos separats en diferents carpetes. D'altra banda, també disposem d'una carpeta denominada “*Test Data*”, amb la qual podem realitzar la nostra validació.

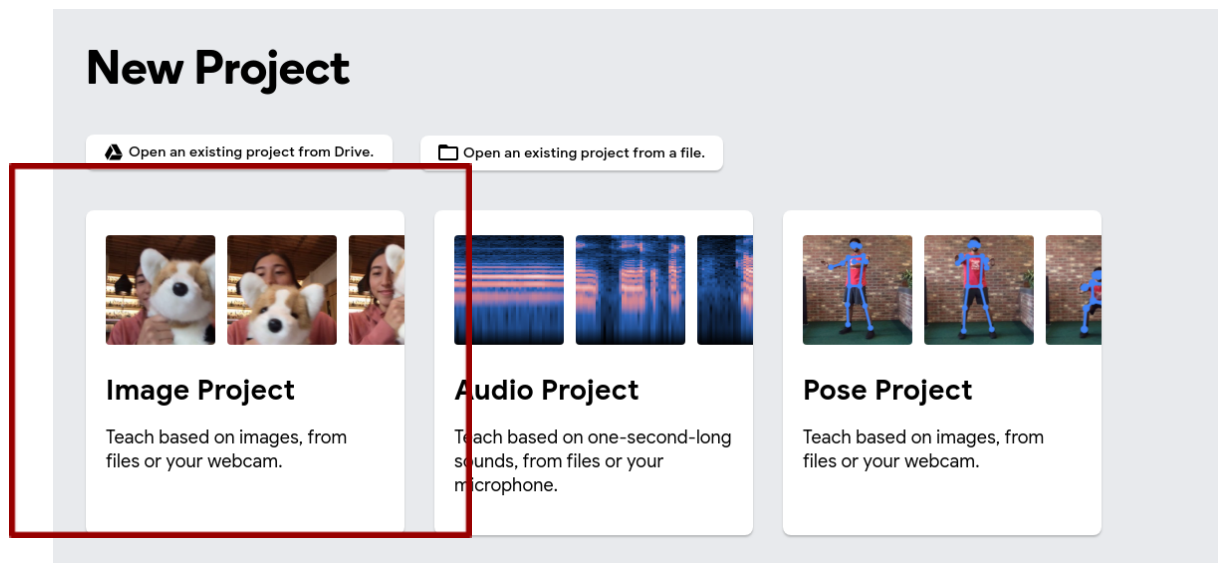
Primer que tot, anem a crear el model. Per a això, utilitzarem l'eina gratuïta de Google denominada “*Teachable Machine*”. En aquest sentit, seguiu els següents passos:

1. Accediu al següent enllaç: <https://teachablemachine.withgoogle.com/>
2. Una vegada obert, seleccioneu el botó “*Get Started*”.

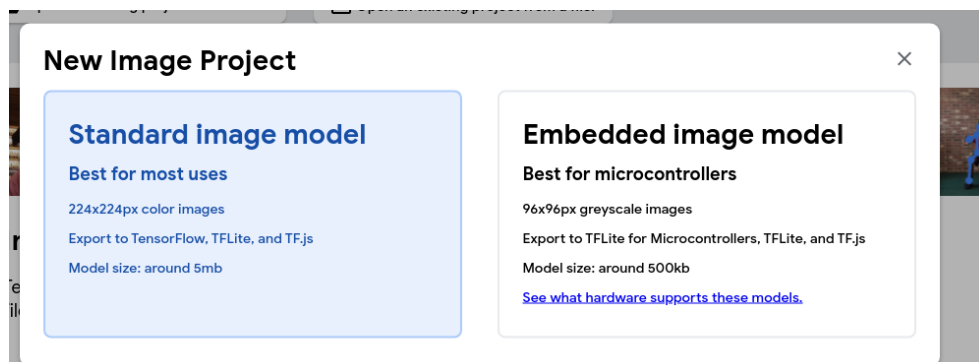


3. A continuació, crearem un nou “projecte d’imatge”.

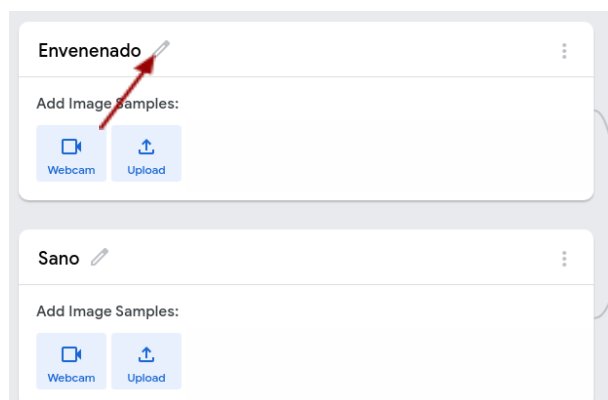
<sup>4</sup> Imatges obtingudes del curs gratuït <https://edu.readyai.org/courses/teachable-machine>



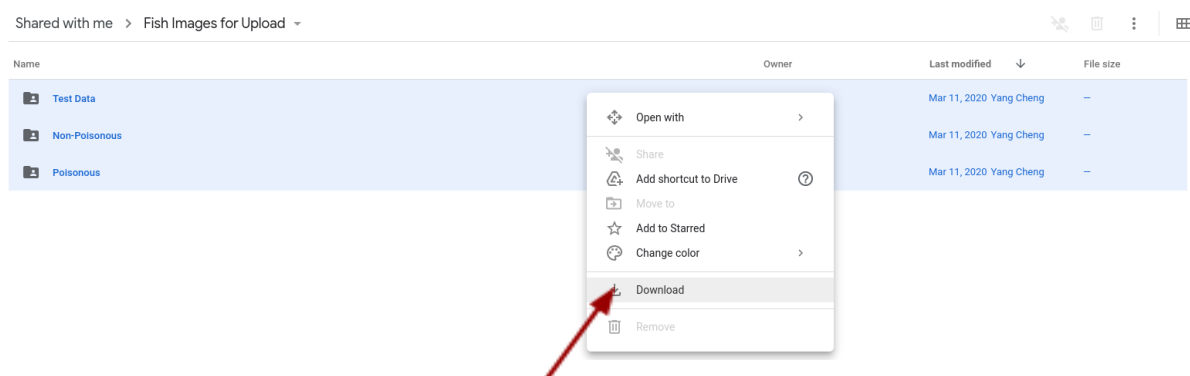
4. Ens apareixerà aquesta finestra i seleccionarem l'opció per defecte: “*Standard Image model*”



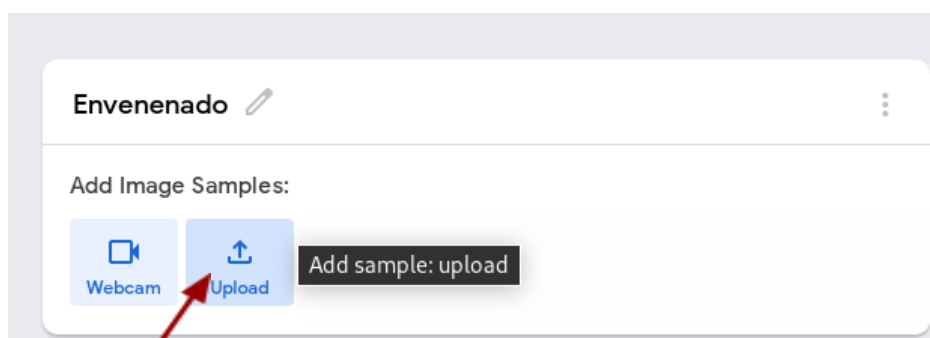
5. A continuació, anomenarem a la Classe 1 com “enverinat” i la Classe 2 com a “sa” seleccionant el botó del llapis:



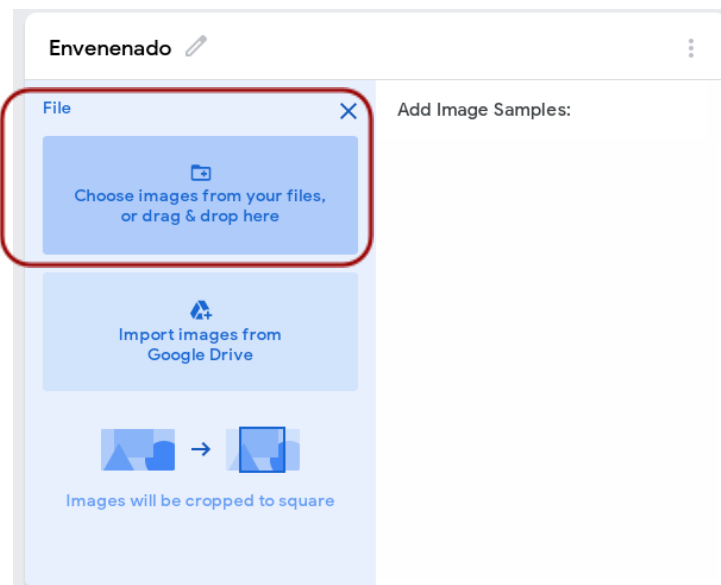
6. El següent pas és carregar les nostres imatges. Per a això, primer ens descarregarem les carpetes amb les imatges dels peixos al nostre ordinador:



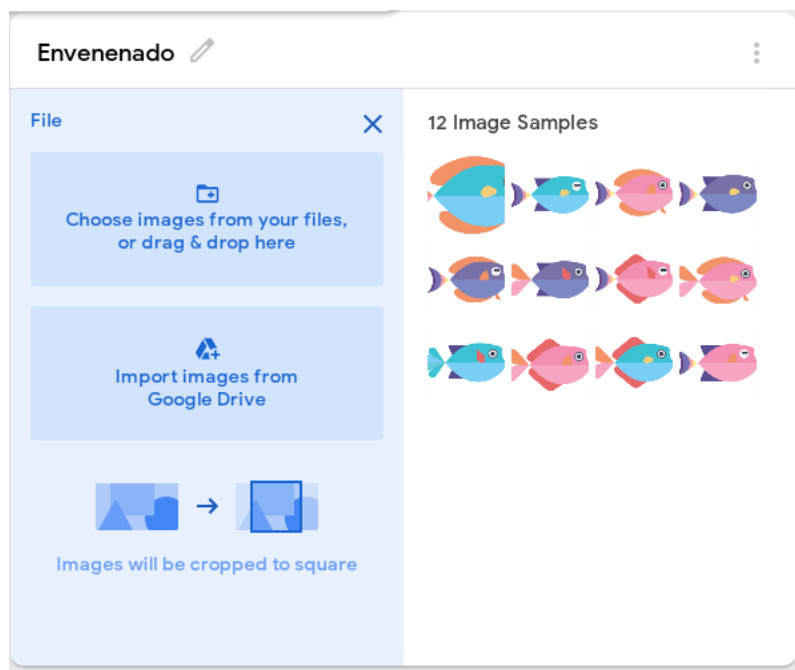
7. Una vegada descarregades les carpetes, les pujarem a l'eina mitjançant el botó "Upload":



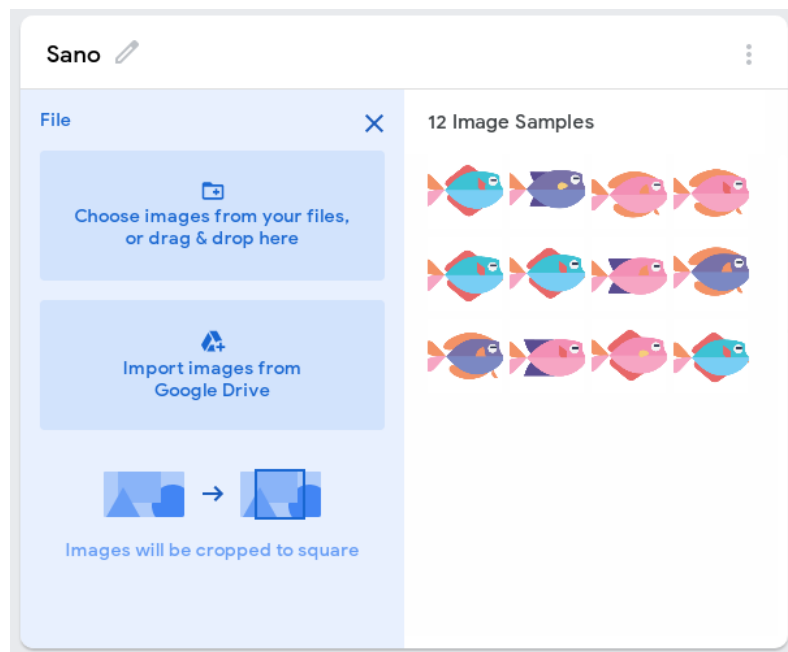
8. A continuació, seleccionarem l'opció "Choose images from your files, or drag & drop here"



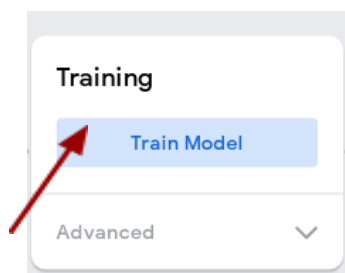
9. Seleccionarem les imatges de la carpeta de peixos enverinats (“Poisonous”) i les carregarem:



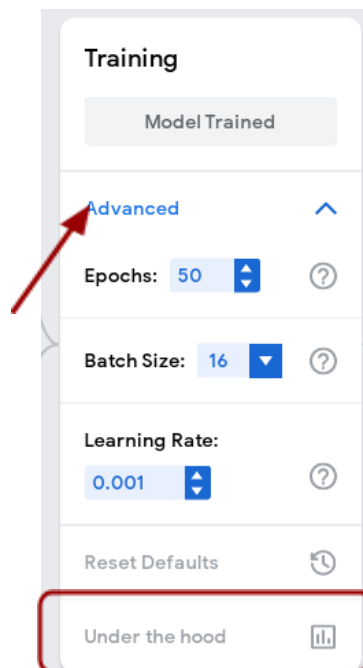
10. Fem el mateix per a la classe de peixos sans, carregant les imatges de peixos sans (“Non-Poisonous”).



11. Finalment, procedim a entrenar el model, seleccionat el botó “*Train Model*”.



L'eina entrenarà el model fent servir una tècnica denominada “*transfer learning*”, la qual està basada en l'aprenentatge per xarxes neuronals. Sense entrar molt en el seu funcionament, podem veure que conté una sèrie de paràmetres que es poden ajustar seleccionant el desplegable “*Advanced*”.



La modificació d'aquests paràmetres ens permet ajustar el nostre model fins a obtenir resultats òptims. Aquest procés, es coneix com “*fine-tuning*”. Ara com ara, ometem aquest pas, i passarem directament a l'avaluació del model. Per a això, seleccionarem el botó “*Under the hood*” (emmarcat en vermell en la imatge de dalt).

Se'ns obrirà un desplegable a la dreta amb algunes mètriques interessants per avaluar el nostre model. Entre elles, trobem:

- **Accuracy:** Es refereix a la “exactitud” que ha obtingut el nostre model, i es calcula sobre la base del percentatge de casos que el model ha encertat. És un valor molt utilitzat per avaluar models, encara que pot donar problemes quan les dades no estan balancejades correctament.
- **Confusion Matrix:** Ens indica quants valors s'han classificat correctament o incorrectament per a cada classe.

Si les seleccionem, podem observar els resultats que hem obtingut. Aquest algorisme realitza diverses iteracions per aprendre a classificar les nostres imatges, i en aquestes iteracions realitzarà diverses proves fins a aprendre a classificar correctament. Aquest procés pot generar resultats diferents cada vegada que s'executa, a causa del comportament propi de l'algorisme, per la qual cosa no us preocupeu si els vostres resultats són diferents als que mostrem en aquest enunciat.

En aquest cas, obtenim els següents resultats:

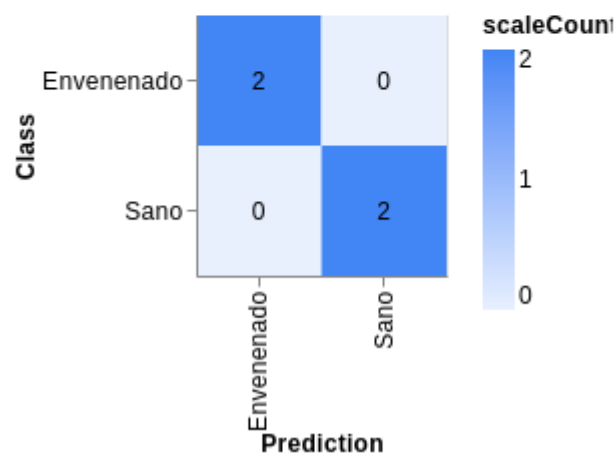


### Accuracy per class



CLASS	ACCURACY	# SAMPLES
Envenenado	1.00	2
Sano	1.00	2

### Confusion Matrix

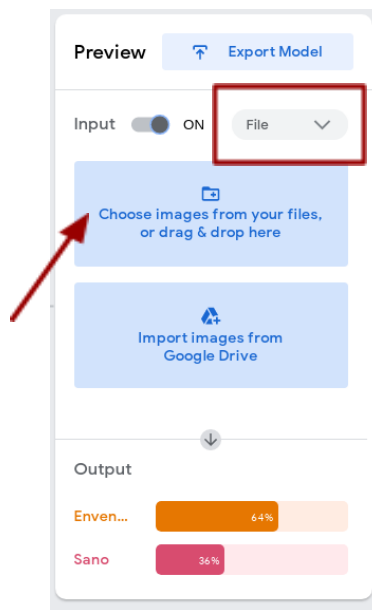


A continuació, responeu les següents preguntes:

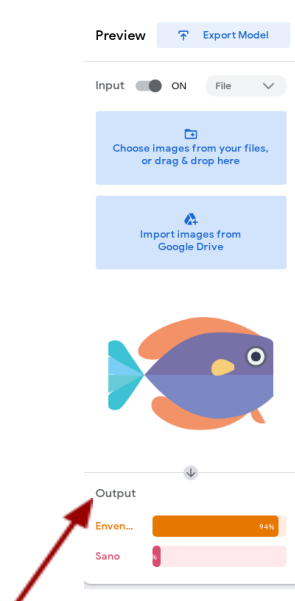
**3.2.1.** Quins resultats d'Accuracy i de Confusion matrix ha obtingut el vostre model? Quines conclusions podeu extreure?

Per finalitzar, anem a avaluar el model amb les imatges de "Test Data". Per a això, seguirem aquestes instruccions:

1. En la pestanya "Preview", seleccionarem en el desplegable l'opció "File". A continuació, pujarem la imatge a avaluar mitjançant el botó "Choose images from your files, or drag & drop here"



2. Anem a provar cadascuna de les 4 imatges que tenim de *Test* i anem a mirar el resultat del model (“*Output*”):



3. Veiem que ens retorna el percentatge de possibilitats de pertànyer a una classe o a una altra. Per a aquest exercici, suposarem que una probabilitat **major del 50%** significa que la imatge pertany a aquesta classe. En l'exemple de la imatge de dalt, tenim un 94% de probabilitats que la imatge sigui un peix enverinat, per tant, considerarem que l'algorisme ho classifica com aquesta classe.

A continuació, responeu:

**3.2.2.** Com classifica el vostre model cadascuna de les imatges? Comenta el percentatge obtingut per a cada imatge en relació a cada classe i si les classificacions són correctes.

**3.2.3.** Quin tipus d'aprenentatge automàtic creus que utilitza l'algorisme de *Teachable Machine*?

## Criteris d'avaluació

- La pregunta 3.1 es valorarà amb **1.5 punts** com a màxim.
- La pregunta 3.2 es valorarà amb **1.5 punts** com a màxim.
- Es valorarà la capacitat d'identificar i definir diferents tipus d'aprenentatges automàtics.
- Es valorarà l'aportació de **fonts externes** que sustentin els raonaments.
- Totes les respostes han d'estar **desenvolupades i argumentades**.
- Extensió **mínima** per pregunta: **100 paraules**.
- Les respostes que no compleixin els requisits de nombre de paraules es penalitzaran.