

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

Jo, Xavi Rubio, declaro que per realitzar aquest lliurament m'he inspirat, he mirat, i m'he informat però no he copiat ni plagiat ningun contingut ni un document d'un company ja sigui actual o passat.

Introducció a la ciència de dades

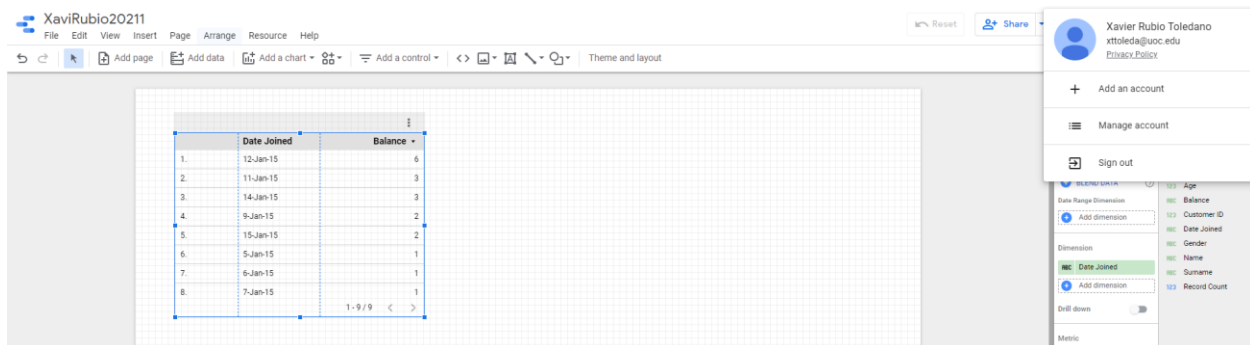
PAC1: Exemples pràctics d'anàlisi.

Pregunta 1 (35% puntuació)

1.1. Extreure les dades

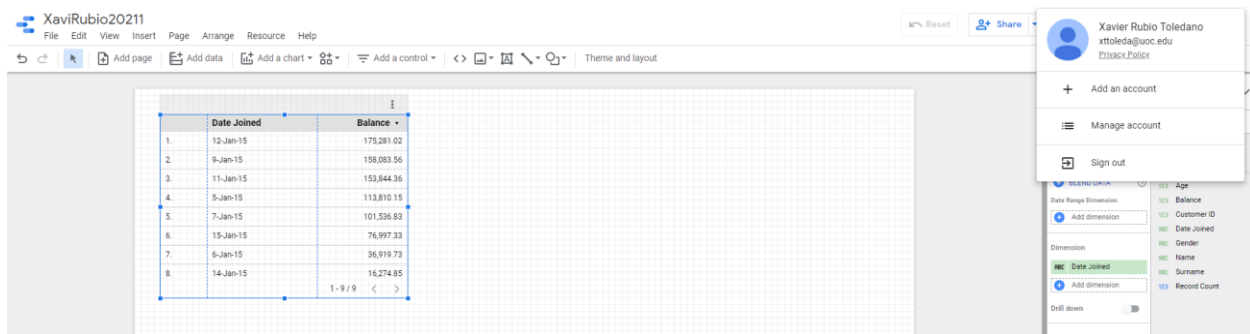
- Aquesta taula té 7 atributs que són: Customer ID, Name, Surname, Gender, Age, Date Joined, i, Balance. Dels quals cadascun dels atributs són dades simples, ja que veiem noms, dates, edat, números... Però si ens fixem en el document sencer, veiem que és una dada composta, ja que cada columna informa de les dades (que es pot deduir) dels treballadors d'una empresa, amb el seu nom i cognom, el dia de la seva entrada, el sexe i més informació.

1.2. Carregar les dades i analitzar-les



En aquesta captura tenim el report de les dades de l'arxiu P12-Bank-Customers-Demo.csv raw, és a dir, sense modificar, on li hem aplicat que sumi els sous de les mateixes dates. Però com podem veure ordena per data, però no fa ve el balanç, ja que hauria de ser nombres amb milers no només unitats.

1.3. Transformar les dades



	Date Joined	Balance
1.	12-Jan-15	175,281.02
2.	9-Jan-15	158,083.56
3.	11-Jan-15	153,844.36
4.	5-Jan-15	113,810.15
5.	7-Jan-15	101,536.83
6.	15-Jan-15	76,997.33
7.	6-Jan-15	36,919.73
8.	14-Jan-15	16,274.85

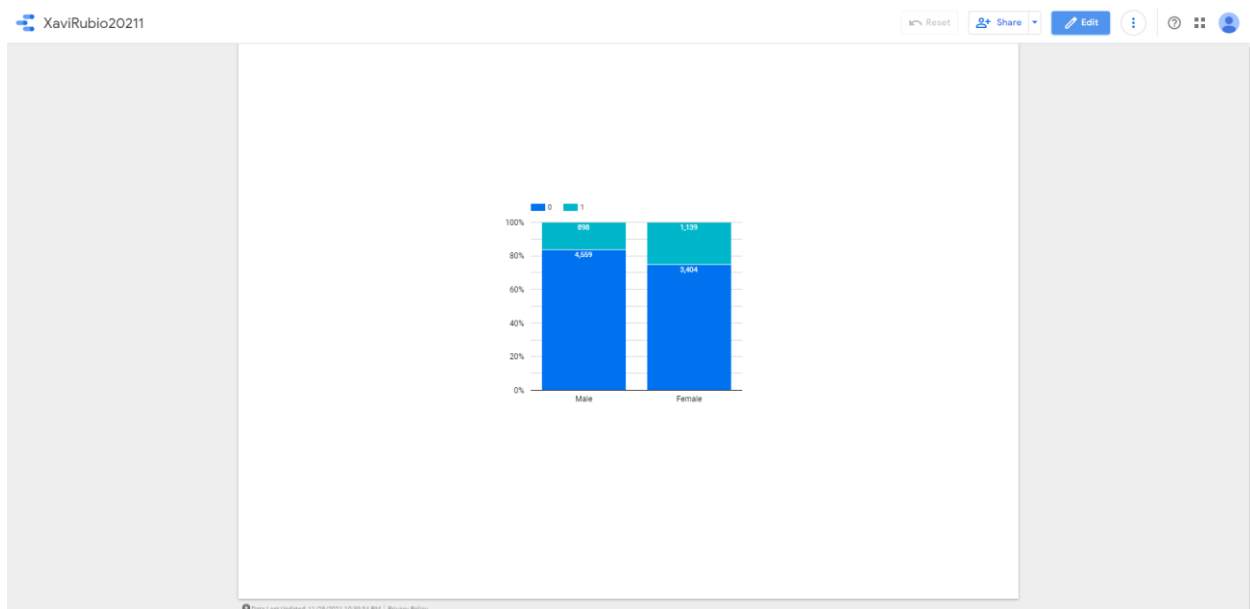
En aquesta captura tenim el report de les dades de l'arxiu P12-Bank-Customers-Demo.csv però modificada perquè el que abans eren els sous té tipus text, ara ho siguin de tipus número, on li hem aplicat que sumi els sous de les mateixes dates. I com podem veure ordena per data, però fa ve el balanç.

1.3.2. Transformar les dades

Respecte a veure les dades és una forma molt visual, ja que veus el dia, mes o l'any sense cap classe de problema, però a l'hora d'utilitzar-les o crear-ne algun tipus de model seria molt complicat i molt costós si és tinguessin que separar els tres nombres a més que l'any són 2 nombres i no 4 com haurien de ser, la data és text... El millor per a poder treballar amb ells és transformar la data en un nombre de la següent manera: dia 1 de gener any 0000 = 1, 1-gen-0002 = 365... I així successivament.

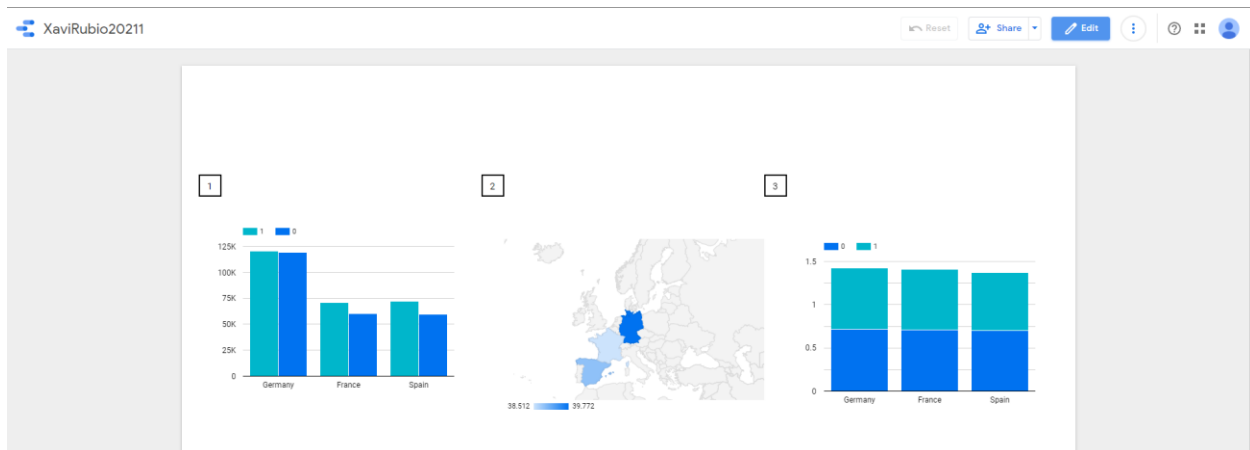
Pregunta 2 (35% puntuació)

2.1. Visualització de dades



En aquesta gràfica veiem segons el sexe (Masculí i Femení) si segueixen sent actius al banc o no, podem veure que en el 100% de la barra d'homes només són actius un 16.46% i la resta són inactius. En el cas de la dona passa més o menys el mateix, del 100% de dones només n'hi ha un 25.07% de dones que són actives. Tenint aquestes dades i haver-les analitzat podem assegurar que el banc no està en el seu millor moment, ja que la majoria dels seus clients tant homes com dones no són actius. Amb aquesta anàlisi podríem crear un model predictiu per a tenir una comparació de la resta de mesos i així poder veure quin canvi ha pogut tenir l'empresa per a tenir tants clients inactius. També seria una molt bona pràctica saber de quant són les dades on ajudaria molt a concretar-les més i a treure conclusions molt més exactes tant si es fa un model predictiu com perquè una persona entengui com està anant el seu negoci.

2.2. Presentació de resultats



Per aquesta anàlisi he escollit aquests tres gràfics, els quals estan relacionats entre ells per així poder fer una millor anàlisi de la situació del banc corresponent. Primer introduiré una mica de què està format cada gràfic per així poder explicar les meves conclusions de cada un i una conclusió conjunta.

El primer gràfic relaciona la localització de cada persona amb el valor de 'Exited' on es veu la mitjana del balanç de cada compte.

En el segon gràfic podem veure la mitjana d'edats per país dels clients.

I per últim tenim la mitjana de si els clients tenen o no targeta de crèdit relacionat amb 'Exited'.

1. El primer gràfic ens dona una idea del balanç en el compte de cada país, la meua idea era saber si els clients no actius del banc potser utilitzaven aquest compte per a estalvis o inversions, així podem veure que si el balanç dels comptes inactius és molt més gran que el de les actives cap a la possibilitat que estiguessin utilitzant aquest compte com a compte d'estalvis i és per això que no tenen el valor 'Exited', ja que només entren un parell de cops cada molt de temps i no per utilitzar el compte dia a dia.
2. Per al segon gràfic he analitzat la mitjana de l'edat per cada país per així poder tenir una resposta a la pregunta 'A quin públic enfoca l'empresa?'. Sabent aquest valor per cada

país com a banc podem adaptar els anuncis o la publicitat a pot ser un públic més jove o més gran. Un exemple d'utilització d'aquesta anàlisi seria per si en tenen un públic jove podrien canviar les tarifes o fer renting de cotxes més petits i així podrien maximitzar el públic jove, ja que en general no solen tenir un sou molt alt i no vendre cotxes més classe mitjana o alta. O també es podria donar més suport amb la hipoteca, ja que el públic jove en algun moment s'han d'anar a viure sols i es podrien enfocar els pisos a potser més petits i en la seva majoria de lloguer. Els valors d'aquest mapa estan en la mitja, no són un públic molt jove ni són un públic molt gran.

3. Com a últim gràfic he escollit veure segons l'activitat de 'Exited' per país, veure quins dels clients tenen targeta de crèdit i quins no. Amb aquest punt de vista podríem saber que si tenen targeta de crèdit són propens a fer pagaments dels dia a dia amb el compte d'aquell banc, però si no tinguessin targeta de crèdit significaria que són persones que no utilitzen molt aquest compte, o, que no l'utilitzen diàriament que pot ser només en ocasions especials, així que aquest gràfic el podríem comparar amb el primer. També podríem veure la utilitat de les targetes de crèdit avui en dia i que pot ser fan els pagaments amb el mòbil així que no cal que tinguin una targeta de crèdit i es podrien treure o reduir les targetes com també crear un model predictiu on indiqui mitjançant tècniques de regressió lineal com van amb el temps apareixent o desapareixent les targetes de crèdit i així poder fer una anàlisi més exhaustiu per a poder prendre decisions vàlides i reals per al banc en qüestió. Com es pot veure en el gràfic encara que molt sutilment els valors de la gent amb valor 1 en la columna 'Exited' fan servir un 70% la seva targeta de crèdit que és el mateix valor que en els clients com a valor 0. Una vegada explicats els tres gràfics m'agradaria treure una relació dels tres com a un conjunt o un mateix anàlisis.

El que volia saber era si en aquest banc s'utilitzaven els comptes com a estalvis o inversions així mirant el valor de 'Exited' podríem saber són comptes que utilitzen dia a dia o no. Com a conclusió d'aquests tres gràfics en conjunt podríem treure que a alemanya tenen una mitja molt alta de balanç al seu compte i a la resta de països no, això podria ser degut als sous de cada país on si són sous majors major balanç al compte de la persona d'un país en concret, però comparant el valor de la gent activa no n'hi ha molta variació així que també d'una altra forma amb el tercer gràfic podem veure si d'aquests països canviant el valor del balanç per si tenen targeta o no es podria confirmar que moltes persones la fan servir per guardar els diners o pot ser com en aquest cas el balanç està igualat, però la utilització de les targetes de crèdits bastant baix, ja que només utilitzen targeta de crèdit un 70% dels clients, però això també pot ser perquè utilitzen el telèfon mòbil i si ho comparem amb els clients actius o no ens dona un valor pràcticament idèntic en els dos casos, i utilitzant el segon gràfic es podria veure si el balanç és baix podria ser degut al fet que si els clients són molt joves com tenen un sou menor com a norma general podríem explicar-ho.

Com a conclusió de la meua anàlisi he pogut treure que els clients d'aquest banc no fan servir aquests comptes per a estalviar o fer inversions sinó que podem dir que n'hi ha tota classe de clients tant de rang d'edat com a rang de balanç al seu compte. Però hem deduït que només un

70% dels clients fan servir targeta de crèdit així que s'hauria de veure el perquè d'aquest fenomen.

Pregunta 3 (30% puntuació)

3.1. Tipus de tècniques d'aprenentatge automàtic

Tècniques supervisades: Són aquelles les quals mitjançant una sèrie de dades conegudes on també coneixem la variable resultat es poden deduir quines seran les noves dades. N'hi ha dues tècniques diferents dintre de les supervisades: les tècniques de classificació, i les de regressió. La diferència entre una i un altre és que les de classificació dedueixen respostes directes; com per exemple en les que volem que la resposta sigui una categoria (si el correu és spam o no, o, classificar un tipus de malaltia en unes ja conegudes), d'altra banda, les de regressió dedueixen respostes contínues o reals, com per exemple quan es crea un algoritme que es vol que dedueixi el que farà el mercat de valors en un futur, deduir la temperatura d'una ciutat...

Tècniques no supervisades: El que fa aquesta tècnica és mitjançant un conjunt de valor els quals no estan categoritzats ni tenen cap etiqueta, el mateix algoritme busca relacions o patrons ocults segons propietats estadístiques. N'hi ha de dos tipus, clusterització, el qual intenta agrupar les dades en grups de la mateixa família, com per exemple agrupa les paraules amb la mateixa arrel, i, reduint la seva dimensionalitat, aquesta tècnica analitza totes les variables i el que fa és reduir el número d'aquestes basant-se a deixar les més rellevants.

Tècniques de reforç: Aquesta tècnica el que fa és interactuar amb el seu entorn per així mitjançant la repetició d'accions poder trobar el camí correcte. Ex. Que aprengui a jugar als escacs, mitjançant les partides que juga i perdent cada vegada al final podrà aprendre dels seus errors i no tornar-los a fer.

Per acabar reduir la dimensionalitat de les dades pot ser molt beneficiós, però a la vegada contraproductiu. Un dels avantatges que jo li veig és que pots eliminar molt soroll de per exemple un gràfic de punts, ja que ensenyes les dades rellevants, però això també pot ser contraproductiu, ja que no treballes amb dades reals, sinó que són dades modificades les quals no s'hauria de treure una anàlisi molt profunda perquè podria donar porta oberta a sortir errors. Però jo ho veig com una bona tècnica quan es vol ensenyar les dades en un gràfic i exposar-les a persones per així donar-les una idea principal i clara dels resultats.

MathWorks [Online] // Por qué es importante el machine learning. -

<https://es.mathworks.com/discovery/machine-learning.html>.

Morris & Opazo [Online] // Técnicas de reducción de dimensionalidad. -

<https://www.morrisopazo.com/es/recursos/blog-espanol/tecnicas-de-reduccion-de-dimensionalidad/>.

Tsymbol Oleksii mobidev [Online] // 5 Essential Machine Learning Algorithms For Business Applications. - <https://mobidev.biz/blog/5-essential-machine-learning-techniques>.

3.2. Anàlisi d'un model d'aprenentatge automàtic

3.2.1.

Els resultats del meu model són que per les dues classes agafant dos exemples l'exactitud és d'1, és a dir que el model ha sigut 100% correcte en les 4 mostres.

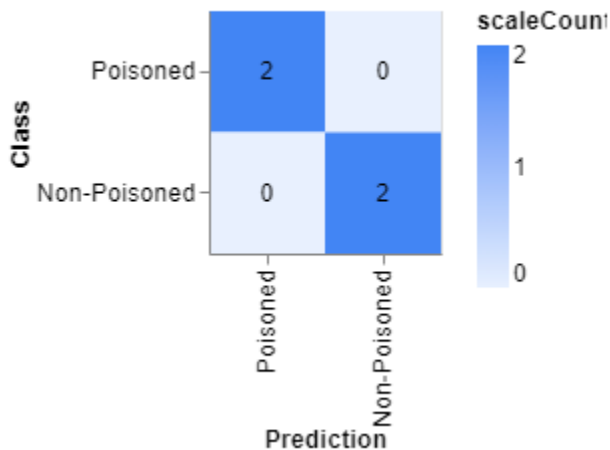
Accuracy per class



CLASS	ACCURACY	# SAMPLES
Poisoned	1.00	2
Non-Poisoned	1.00	2

I per al valor de confusion matrix passa una cosa similar, per a cada columna 'y' que és la classe va relacionada amb la fila 'x' que és la predicció al que el model ha arribat. Com podem veure a la imatge de 4 mostres ha deduït que dos són enverinats i els altres dos són sans, així que no hi ha hagut cap confusió, ja que 'x' i 'y' corresponen al mateix valor.

Confusion Matrix



3.2.2.

Per al primer cas el peix té un punt al ull així que és un peix enverinat, i el model prediu que el peix és un 83% enverinat.

Per al segon cas el peix té un guió al ull així que és un peix sa, i el model prediu que el peix és un 94% sa.

Per al tercer cas el peix té un punt al ull així que és un peix enverinat, i el model prediu que el peix és un 86% enverinat.

Per al quart cas el peix té un punt al ull així que es un peix enverinat, i el model prediu que el peix es un 82% sa.

El model ha fallat en un cas dels 4 que he provat.

3.2.3.

Aquesta eina emplea una tècnica d'aprenentatge supervisada de classificació, ja que segons les classes que en aquest cas és deduir si el peix està enverinat o no nosaltres li passem unes dades les quals el resultat que volem és que donar-li una nova dada o en aquest cas un nou peix, i pugui dir si el peix està enverinat o no, o fer una aproximació i deduir el cas de la imatge que li hem passat com a input i dir mitjançant el percentatge per a ensenyar-ho d'una forma més precisa o la possibilitat que té cada test de què classe s'apropa més.