CONCEPTES BÀSICS DE LA CIÈNCIA DE DADES

Qualsevol estudi o ciència té un procés a seguir i la gestió de les dades no és una excepció. A continuació es presenten els passos que cal seguir per tal de realitzar una bona gestió de les dades.

- **Objectiu**. El primer de tot és establir un objectiu. És a dir, per què volem aquestes dades i quina finalitat tindrà el coneixement adquirit?
- **Obtenció de les dades.** Necessitem accedir i recopilar les dades amb les que volem treballar. Aquestes dades segurament arribaran de manera desordenada.
- **Preparar les dades.** Aquest pas es basa en ordenar les dades i depurar les interferències abans de començar a dissenyar models per trobar patrons.
- **Exploració de les dades.** Fase centrada en trobar patrons, correlacions i desviacions que puguin convertir les dades en un coneixement valuós i aplicable.
- **Construcció de models.** Aquesta etapa es basa en generar prediccions en funció del model escollit. En aquest cas, podrem aplicar el coneixement automàtic per anar millorant el model amb l'acumulació de més dades.
- **Presentació de resultats i anàlisis.** Aquest punt consistirà en fer públics els resultats a altres persones, per tal de dur a terme les noves accions o els canvis que els models predictius ens indiquin.

Cal dir que la majoria de vegades aquest procés no és lineal, sinó cíclic. La generació de dades és continua i cal revisar tots els passos contínuament per tal de millorar l'eficàcia i l'eficiència de les prediccions (que comportarà prendre decisions millors).

La ciència de dades ofereix, bàsicament, dos tipus de models predictius: el de regressió i el de classificació. Vegem-los:

El primer (model de regressió) té per objectiu la predicció d'un número, que depèn de la relació que pugui existir respecte a una o més variables. Aquest model predictiu es basa en els coeficients de correlació entre les variables independents i la dependent, de la qual volem saber-ne el valor. Les dades que pot necessitar aquest model compleixen els següents requisits:

- Les variables s'han de poder mesurar de manera contínua. El temps, les vendes o el pes corresponen a aquest tipus de mesures.
- Les observacions entre diferents variables no han de tenir interferències entre elles.
- Les dades no haurien de tenir valors atípics. Un valor atípic és aquell que queda molt allunyat de la resta i podria ser degut, probablement, a un error de mesura. En certs casos podria no haver cap error i el valor de la dada fos només una excepcionalitat.
- La variància no ha de canviar al llarg de la línia predictiva. Es pot veure la definició de variància a: https://ca.wikipedia.org/wiki/Variància.
- Els errors al llarg de la línia de predicció segueixen una distribució normal. En l'enllaç següent es pot aprofundir en el concepte de distribució normal: https://ca.wikipedia.org/wiki/Distribució_normal.

El model de classificació té per objectiu predir si una opció es durà a terme o no. Aquesta opció pot tenir només dues possibilitats diferents, el que s'anomena *classificació binària*, o més de dues, que, en aquest cas, s'anomena *classificació multinominal*. Un model binari, per exemple, por tenir les opcions de *veritat o fals*, si o no, etc. El model multinominal, en canvi, en té més: per exemple podríem incloure les opcions *Grup A, B o C,* o *Graus de satisfacció des de Molt satisfet, Poc Satisfet, Gens satisfet*, etc...

Gràcies al model de classificació podrem saber, per exemple, si un accident es produirà o no (dues possibilitats) o si un producte determinat serà comprat per persones joves, adultes o d'edat avançada (més de dues possibilitats).

Per tal de generar un model capaç de fer aquesta predicció, necessitarem trobar correlacions amb altres variables independents. Per això, haurem de tenir en compte aquests elements:

- Les variables independents han de ser vàlides.
- Hem d'evitar les dades contínues, com ara la temperatura, el temps, etc.
- Hem de procurar no fer servir variables que estiguin estretament relacionades entre elles.

Per tal d'entendre els propers capítols, és important que ens familiaritzem amb els termes següents:

- Base de dades: espai digital destinat a emmagatzemar informació al qual es pot accedir, sempre i quan es tinguin les credencials i els permisos necessaris per tal de llegir i importar les dades que s'hi troben.
- *Overfitting*: és l'efecte de proveir massa informació sobre un model del qual ja es coneix el resultat desitjat. Aquest efecte fa que el model quedi massa definit respecte a una entrada de dades i que no sigui capaç de generalitzar resultats en altres situacions.
- Correlació: mesura com estan relacionats directament els canvis d'una variable respecte a una altra.
- Mediana: en una col·lecció ordenada de dades és el valor que es troba just a la posició central.
- Normalització: consisteix en ajustar els valors mesurats en escales diferents sobre una mateixa escala.
- Valor atípic: és aquell que està lluny respecte a la resta del grup. El valor atípic es deu a causes excepcionals o, a vegades, senzillament a un error.
- Mineria de dades: procés que consisteix a extreure dades d'una font determinada per ser posteriorment examinades. Inclou des de la neteja, organització i depuració de dades fins a la cerca de patrons i relacions significatives.
- *Clustering*: consisteix a recopilar i agrupar un conjunt de punts suficientment similars o propers.
- *Web scraping*: procés d'extracció de dades des de pàgines web.

Descobreix tot el que Barcelona Activa pot fer per a tu

Acompanyament durant tot el procés de recerca de feina

barcelonactiva.cat/treball

Suport per posar en marxa la teva idea de negoci

barcelonactiva.cat/emprenedoria

Serveis a les empreses i iniciatives socioempresarials

barcelonactiva.cat/empreses

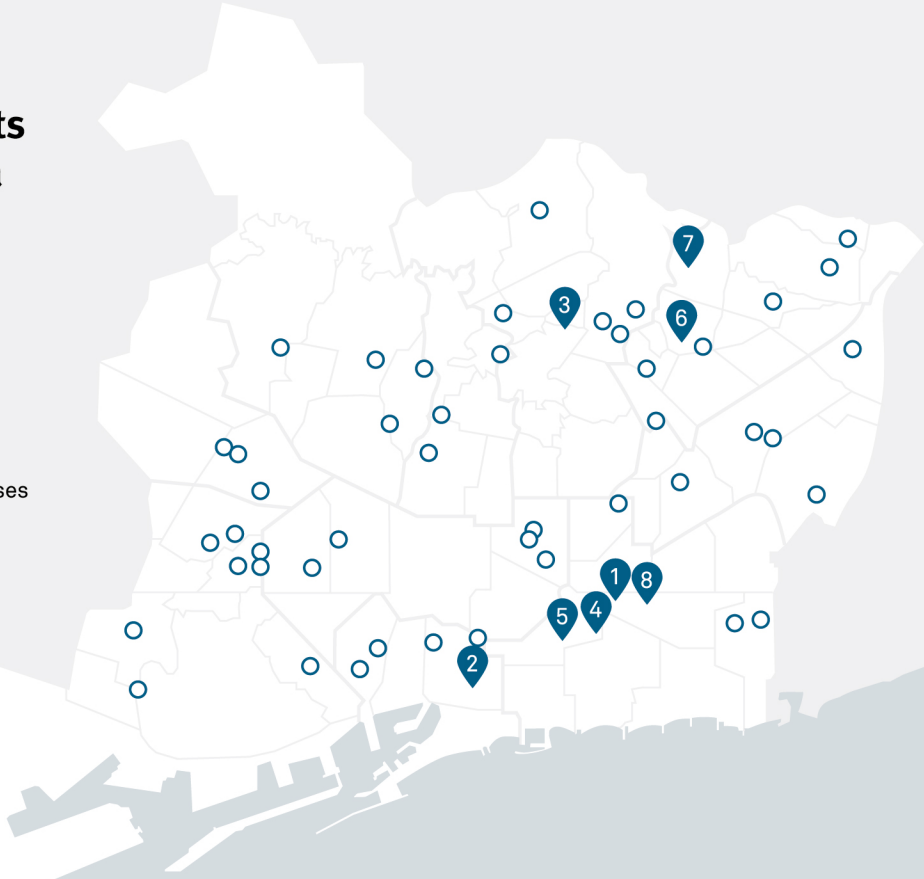
Formació tecnològica i gratuïta per a la ciutadania

barcelonactiva.cat/cibernarium

Xarxa d'equipaments de Barcelona Activa

1. Seu Central Barcelona Activa
   Porta 22
   Centre per a la Iniciativa Emprenedora Glòries
   Incubadora Glòries
2. Convent de Sant Agustí
3. Ca n'Andalet
4. Oficina d'Atenció a les Empreses
   Cibernàrium
   Incubadora MediaTIC
5. Incubadora Almogàvers
6. Parc Tecnològic
7. Nou Barris Activa
8. innoBA

○ Punts d'atenció a la ciutat



Barcelona Activa

Ajuntament de Barcelona

Darrera actualització 2019

Cofinançat per:

UNIÓ EUROPEA
Fons Europeu de Desenvolupament Regional

Segueix-nos a les xarxes socials:

barcelonactiva.cat/cibernarium
barcelonactiva
barcelonactiva
company/barcelona-activa