Mineria de Dades aplicada a indicadors d'exclusió social per barris de Barcelona

Gustavo Talavera Patón
Grau d'Enginyeria Informàtica
Àrea d'Intel·ligència Artificial

David Isern Alarcón
Carles Ventura Royo

Data Lliurament Juny 2019

Aquesta obra està subjecta a una llicència de
Reconeixement-NoComercial 3.0 Espanya de
CreativeCommons

FITXA DEL TREBALL FINAL

Títol del treball:	*Mineria de Dades aplicada a indicadors d'exclusió social per barris de Barcelona*
Nom de l'autor:	*Gustavo Talavera Patón*
Nom del consultor/a:	*David Isern Alarcón*
Nom del PRA:	*Carles Ventura Royo*
Data de lliurament (mm/aaaa):	*06/2019*
Titulació o programa:	*Grau d'Enginyeria Informàtica*
Àrea del Treball Final:	*Intel·ligència Artificial*
Idioma del treball:	*Català*
Paraules clau	*Mineria de dades, Open Data, Weka*

Resum del Treball (màxim 250 paraules): *Amb la finalitat, context d'aplicació, metodologia, resultats i conclusions del treball*

L'actual rellevància en la gestió de dades en combinació amb el compromís, cada cop més habitual, de transparència d'institucions públiques està generant multitud d'iniciatives de dades obertes (*Open Data*). Les motivacions comuns a totes aquestes iniciatives són la transparència en la gestió pública i l'accés i reutilització de dades per maximitzar els recursos públics.

En aquest treball de fi de grau, s'utilitza el servei de dades obertes de l'Ajuntament de Barcelona, *Open Data BCN*, amb la finalitat d'abordar una problemàtica creixent a les grans ciutats, que preocupa des de les autoritats locals fins a les autoritats europees, l'exclusió social.

Aquest TFG proposa l'aplicació de tècniques de Machine Learning, utilitzant algoritmes de *clustering* i de *classificació* amb els objectius d'obtenir un model que permeti classificar els barris de Barcelona, segons indicadors d'exclusió social, i conèixer quins indicadors mostren més influència en l'agrupació dels barris. Per a això, s'inicia un procés de descobriment de coneixement en bases de dades (*KDD*, *knowledge discovery in databases*). S'apliquen les fases corresponents a aquests tipus de processos, focalitzant a la fase de mineria de dades (*Data Mining*) i fent servir algorismes de Machine Learning amb l'eina *Weka*.

Els resultats obtinguts determinen que els indicadors d'immigració, atur,

educació i estat dels habitatges, són els indicadors amb més influència en l'agrupació dels barris, a més, l'algorisme de veïns més propers és qui obté els millors resultats de classificació.

Abstract (in English, 250 words or less):

The current relevance in the management of data, in combination with the increasingly commitment of transparency for public institutions, is generating a multitude of open data initiatives. The common motivations for all these initiatives, are transparency in public management and access and reuse of data to maximize public resources.

In this end-of-degree project, the open data service of the Barcelona City Council, Open Data BCN, is used to address a growing problem in large cities, which is concerned with local authorities to European authorities, social exclusion.

This work proposes the application of Machine Learning techniques, using clustering and classification algorithms with the aims of obtaining a model that allows classifying the neighborhoods of Barcelona, according to indicators of social exclusion, and know which indicators show more influence in the clustering of neighborhoods. For this, a database discovery process begins (KDD, knowledge discovery in databases). The phases corresponding to these types of processes are applied, focusing on the Mining Data phase (Data Mining) and using Machine Learning algorithms with the Weka tool. Finally, the model obtained is analyzed, interpreted and evaluated to discover knowledge of the obtained patterns.

The results obtained indicate that the indicators of immigration, unemployment, education and housing status are the indicators with the greatest influence on the clustering of neighborhoods, and the nearest neighbors algorithm is who obtains the best classification results .

Índex

Llista de figures

1. Introducció

Aquest primer capítol comença definint el context i motivació del treball, s'enumeren tant els objectius generals com específics. Es continua, presentant l'enfocament i la metodologia escollida a desenvolupar per assolir els objectius.

A continuació, s'identifiquen les tasques a realitzar i la consegüent temporalització, es finalitza descrivint els productes obtinguts i la resta de capítols del treball.

1.1 Context i justificació del Treball

L'***exclusió social*** és un fenomen que està generant una creixent preocupació d'entitats governamentals i no governamentals d'àmbit mundial. Principalment institucions i organismes europeus han interioritzat el discurs sobre polítiques socials i aprofundit en el debat de l'exclusió social. Aquest debat no és nou, però amb altres denominacions i concepcions, com pobresa o desigualtat, sempre ha estat controvertit en la seva concepció i definició, i ara no ho és menys.

La creació del terme *exclusió social* s'atribueix al Secretari d'Estat d'Acció Social francès, ***René Lenoir***, el qual va exposar públicament una problemàtica creixent a França al seu primer llibre *Les exclus: un Français sur Dix* (1974), assenyalant una fractura del 10% de la població que restava al marge de la xarxa de seguretat social pública. Des de llavors el terme *exclusió social* ha estat estudiat, analitzat i debatut, però igual que les societats evolucionen el concepte d'*exclusió social* també ho fa. En conseqüència, no resulta fàcil donar una definició consensuada i tancada, i encara més si considerem els tres aspectes principals que el caracteritzen: origen estructural, caràcter multidimensional i perspectiva dinàmica. No obstant això, la definició que ofereix l'***European Foundation*** **[1]** acompleix per simplicitat:

> *"L'exclusió social és el procés a través del qual persones o grups queden totalment o parcialment exclosos de la plena participació en la societat en què viuen"*

Però si existeix controvèrsia respecte la concepció *d'exclusió social*, aquesta s'amplifica quan investigadors socials i institucions intenten posar-se d'acord en com mesurar-la, quins indicadors s'han de considerar i amb quina ponderació. La lluita contra l'exclusió social és al nucli de l'***Estratègia Europa 2020*** de la Unió Europea per un creixement intel·ligent, sostenible i integrador **[2]**. Intenta ser un marc de referència per les iniciatives en els àmbits de la Unió Europea, nacionals i regionals. Fixant com a objectiu (en matèria d'*exclusió social*) la reducció de almenys de 20 milions de persones en situació o risc d'*exclusió social* al període 2010-2020.

En l'àmbit nacional, el Govern d'Espanya mitjançant el Ministeri de Foment ha emprès diverses iniciatives a llarg termini, per detectar zones urbanes vulnerables al risc d'*exclusió social*, Com ***l'Atlas de vulnerabilidad urbana en España*** **[3]**, una aplicació de cartografia dinàmica en línia que permet analitzar la vulnerabilitat urbana a nivell de secció censal a tots els municipis. El principal inconvenient és que utilitza el cens de població i vivendes, i aquest es realitza amb periodicitat decennal (l'últim és del 2011). Cal destacar entre el gran nombre d'estudis analitzant l'*exclusió social* a Espanya, el informe Foessa 2008 **[4]** i Subirats (2004) **[5]**.

Per la seva banda, el ***Institut d'estadística de Catalunya*** (*Idescat*) té la missió de proveir informació estadística rellevant i d'alta qualitat, amb l'objectiu de contribuir a la presa de decisions, la recerca i la millora de les polítiques públiques. Entre aquesta informació es troben els *indicadors territorials de risc de pobresa i exclusió social* (**INTPOBR**) **[6]**, utilitzant 19 indicadors en 4 àmbits i presentant els resultats per àrees bàsiques de serveis socials, per comarques i pel conjunt de Catalunya, la freqüència és anual i l'últim registre és de 2017.

Amb totes aquestes iniciatives i recollida de dades en continu creixement, no és suficient amb analitzar les dades per mesurar el nivell de vulnerabilitat de les zones urbanes i detectar quines poden patir *exclusió social*. És fa necessari

extreure'n el màxim profit a aquestes dades per millorar la resposta a aquest fenomen i poder impulsar accions i polítiques socials de prevenció i regeneració urbana.

Una branca de la *Intel·ligència Artificial* que permet aprofundir en l'anàlisi de dades per obtenir nou coneixement no obvi, és el ***Machine Learning***, disciplina d'anàlisi de dades que automatitza la construcció de models per reconèixer patrons. No és una disciplina nova, però amb l'augment exponencial de les capacitats computacionals ha agafat un gran impuls, com tot el relacionat amb l'anàlisi de dades: *Big Data*, *Mineria de Dades*, Magatzems de dades,...

No obstant aquest impuls, encara queda molt camí a recórrer en la relació entre *Machine Learning* i *exclusió social*. Potser pel caràcter multidimensional i dinàmic del concepte, per la falta de consens en els indicadors a utilitzar o en la dificultat en la recol·lecció de dades, però la realitat és que no existeix una bibliografia gaire extensa en l'aplicació del *Machine Learning* a la problemàtica de l'*exclusió social*. Destacar entre les estudis realitzats, *Lafuente i Faura* **[7]** utilitzen el *clustering* i la *regressió logística* aplicats a 31 indicadors obtinguts de l'enquesta de condicions de vida 2009, *Ramos i Varela* **[8]** utilitzen la *regressió logística* per detectar l'*exclusió social* crònica amb les dades d'una ONG d'ajuda als sense sostre, *Serrano et al*. **[9]** també utilitzen la *regressió logística* per detectar l'*exclusió social crònica*, afegint-hi valor proporcionant una aplicació als treballadors dels serveis socials. Per últim, . *Hile i Cova* **[10]** utilitzen *Xarxes Artificials Neuronals (ANN*) per mesurar la vulnerabilitat social per blocs censals, amb les dades d'una enquesta (*American Community Survey*).

Aquests estudis tenen diverses coses en comú, però cal posar el focus en com els tres primers estudis analitzen i/o prediuen casos individuals, la qual cosa dificulta com s'ha comentat anteriorment la recollida de dades, avui dia no es poden obtenir les dades necessàries (indicadors) per cada individu, per això s'utilitzen enquestes i bases de dades referents a un conjunt de població concret, i consegüentment, per una banda no tenim la certesa que l'enquesta reflecteixi la realitat, i per l'altra banda, el grup d'individus que són objecte i final de l'estudi és molt reduït respecte al total de la població.

Aquest treball proposa l'estudi i definició d'indicadors d'exclusió social, per l'aplicació de tècniques de *Machine Learning* realitzant dos processos de Mineria de Dades sobre dades del servei de dades obertes de l'Ajuntament de Barcelona, ***Open Data BCN***. En primer lloc, s'apliquen tècniques de *clustering* intentant extreure coneixement no obvi per comprendre les desigualtats entre barris de Barcelona (veure els barris a l'**Annex 1**) i quins indicadors són els que més influència tenen. A continuació s'apliquen tècniques de *classificació* per obtenir el millor model possible per classificar i detectar al més aviat possible les vulnerabilitats i riscos d'*exclusió social* per barris de Barcelona.

1.2 Objectius del Treball

L'objectiu del treball és posar en pràctica els coneixements i competències adquirides en el Grau d'Enginyeria Informàtica, concretant en les assignatures de Mineria de Dades i Aprenentatge Computacional.

1.2.1 Objectius Generals

- ✓ Analitzar el fenomen de l'exclusió social i els estudis realitzats en relació al Machine Learning
- ✓ Estendre els coneixements en Mineria de dades i Machine Learning
- ✓ Desenvolupar un projecte de KDD (procés de descobriment de coneixement en bases de dades) focalitzant a la fase de Mineria de Dades
- ✓ Obtenir coneixement vàlid, útil i comprensible dels resultats obtinguts

1.2.2 Objectius específics

- ✓ Analitzar el servei de dades obertes *Open Data BCN*
- ✓ Definir els indicadors d'*exclusió social* considerant les dades disponibles
- ✓ Obtenir, Adequar i transformar les dades segons els indicadors
- ✓ Preprocessar les dades considerant l'algorisme a aplicar

- ✓ Crear models d'anàlisi de dades aplicant algorismes de clustering i classificació
- ✓ Analitzar i interpretar els resultats obtinguts

1.3 Enfocament i mètode seguit

Amb els objectius definits, el procés metodològic escollit per assolir-los consisteix en el procés de descobriment de coneixement en bases de dades (*KDD, Knowledge Discovery in Databases*). Segons *Fayyad,* Piatetsky-Shapiro i Smith **[11]**: "*és el procés no trivial d'identificar patrons vàlids, nous, potencialment útils i, finalment, comprensibles en les dades*". A més, *Maimon i Roka*ch **[12]** el descriuen com iteratiu i interactiu, sent iteratiu en cada fase pot ser necessari ajustar les fases anteriors. També destaquen la necessitat d'entendre profundament les necessitats i possibilitats de cada fase, i de tot el procés complet.

Tant *Fayyad,* Piatetsky-Shapiro i Smith com *Maimon i Roka*ch coincideixen en dividir el procés de *KDD* en nou fases o passos, els quals són la guia d'aquest treball i es poden veure a la **figura 1**.

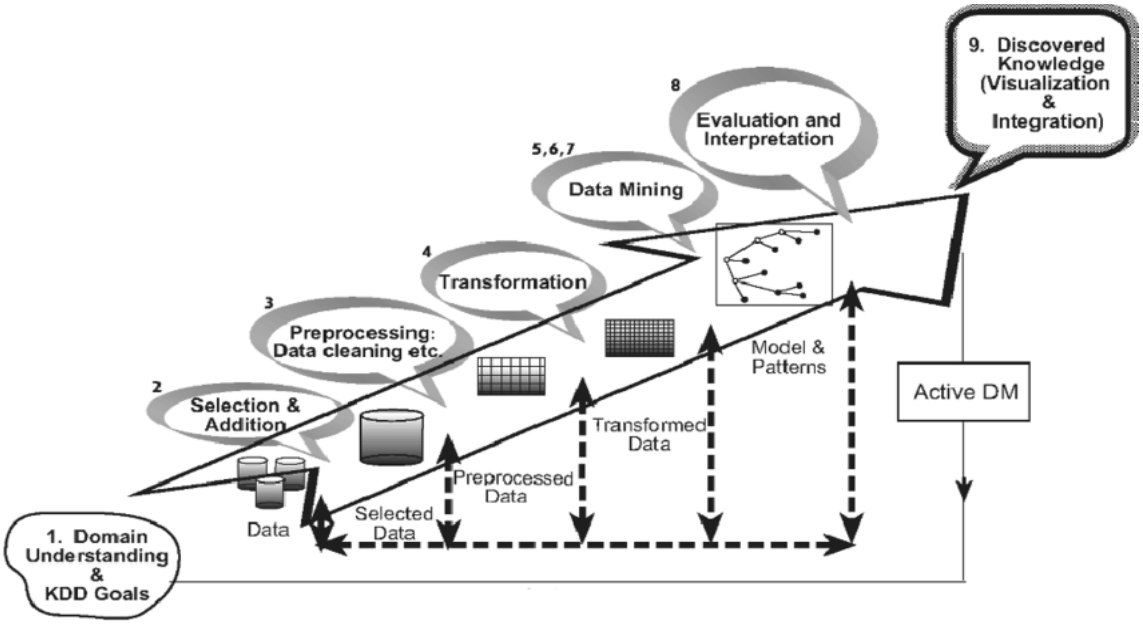


Figura 1. Knowledge Discovery in Databases (KDD).

Les nou fases o etapes es detallaren en el següent capítol:

1. Comprensió del domini de l'aplicació
2. Seleccionar i crear el conjunt de dades
3. Preprocessament i neteja
4. Transformació de dades
5. Escollir la tasca de mineria de dades
6. Escollir l'algorisme
7. Aplicar l'algorisme
8. Avaluar els resultats obtinguts
9. Utilitzar el coneixement descobert

El treball es focalitza en les fases de Mineria de Dades, fases 5, 6 i 7, com es pot veure a la **figura 1**. Entre les diferents àrees o disciplines de *Mineria de Dades* s'utilitza el *Machine Learning* o *Aprenentatge automàtic*, es realitza una visió general dels algorismes a aplicar focalitzant en algorismes de *clustering* o *agrupació* i de *classificació*.

Per la recollida de dades s'utilitza el servei de dades obertes de *l'Ajuntament de Barcelona,* ***Open Data BCN*** **[13]**, el qual té el principal objectiu d'aprofitar al màxim els recursos públics disponibles, exposant la informació generada per organismes públics, permetent el seu accés i reutilització per al bé comú i per al benefici de persones i entitats interessades.

Actualment ofereix 432 datasets organitzats en 5 temes, Administració, Ciutat i Serveis, Economia i Empresa, Població i Territori, en aquest treball s'utilitzaran datasets de Població. Dóna accés als desenvolupadors a l'API de catàleg i acompleix amb 3 estrelles el grau d'obertura de les dades, permetent manipular les dades a qualsevol format que vulguis, sense limitació d'ús d'algun tipus de programari en particular (OF, Open Format,arxius csv), **Figura 2**.

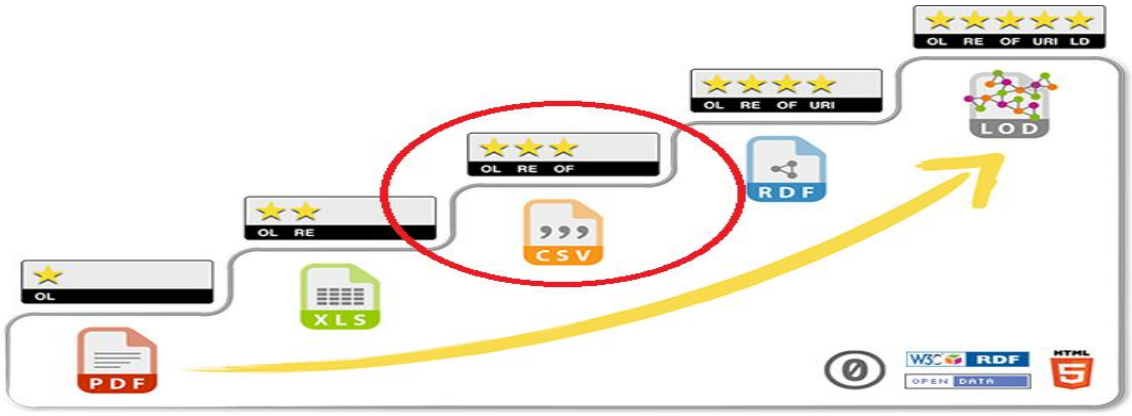


Figura 2. 5-star deployment scheme of Tim Berners-Lee

Per desenvolupar les fases de *Transformació de dades* i les 3 fases de Mineria de Dades s'utilitza el programari lliure ***WEKA*** (*Waikato Environment for Knowledge Analysis*), distribuït sota llicència GNU-GPL, basat en un conjunt de llibreries Java i desenvolupat a la *Universitat de Waikato*. És una col·lecció d'algorismes d'aprenentatge automàtic per a tasques de mineria de dades. Conté eines per a la preparació de dades, classificació, regressió, agrupació, regles d'associació i visualització.

1.4 Planificació del Treball

En aquest apartat es realitza la descripció dels recursos utilitzats per realitzar el projecte, la definició de tasques a realitzar i la planificació temporal considerant les tasques i utilitzant un diagrama de Gantt. La planificació temporal ve donada per les dates de lliurament de les PACs.

1.4.1 Recursos emprats

Recursos de maquinari, s'ha utilitzat un ordinador portàtil amb les següents característiques:

- Processador: Intel Core i5 M450 2.40GHz
- Memòria RAM: 4 GB
- Sistema Operatiu: Windows 10

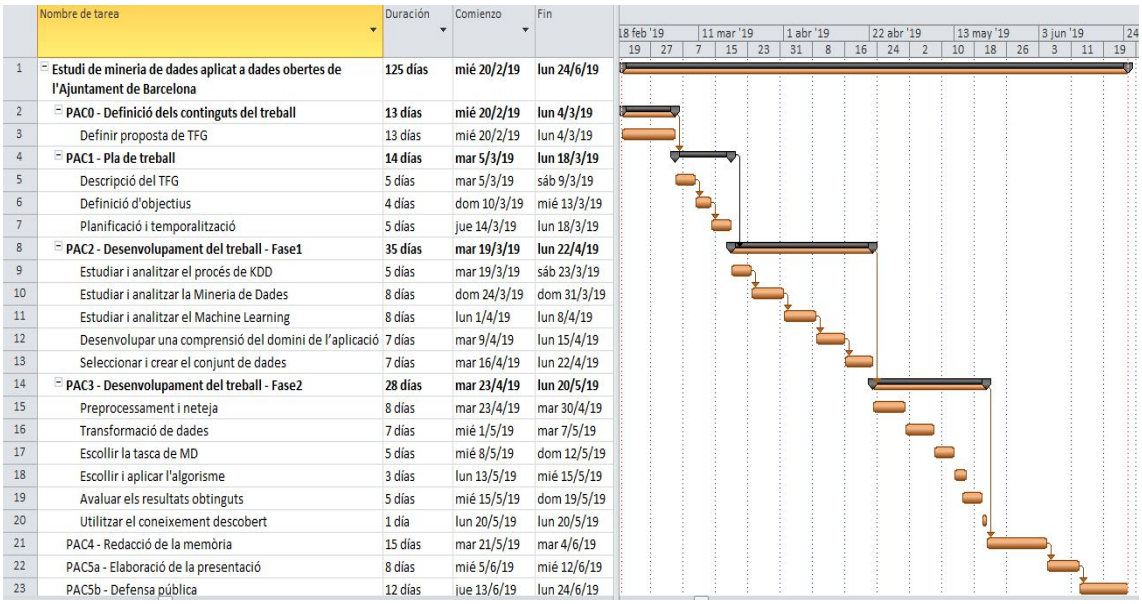
Recursos de programari:

- WEKA 3.8.3, utilitzat per la transformació i aplicació d'algorismes de Machine Learning a les dades
- Sublime Text 3, editor de text utilitzar per netejat les dades i la conversió al format d'arxiu propi de Weka (arff)
- Microsoft Excel 2010, utilitzat per la creació del dataset a utilitzar
- Microsoft Project 2010, utilitzat per realitzar el diagrama de Gantt
- Microsoft Project 2010, utilitzat per realitzar la presentació

1.4.2 Definició de tasques

- Definir proposta de TFG
- Descripció de TFG
- Definició d'objectius
- Planificació i temporalització
- Estudiar i analitzar el procés de KDD
- Estudiar i analitzar la Mineria de Dades
- Estudiar i analitzar el Machine Learning
- Comprensió del domini de l'aplicació
- Seleccionar i crear conjunt de dades
- Preprocessament i neteja
- Transformació de dades
- Escollir la tasca de mineria de dades
- Escollir l'algorisme
- Aplicar l'algorisme
- Avaluar els resultats obtinguts
- Utilitzar el coneixement descobert

1.4.3 Planificació temporal



	Nombre de tarea	Duración	Comienzo	Fin
1	**Estudi de mineria de dades aplicat a dades obertes de l'Ajuntament de Barcelona**	**125 días**	**mié 20/2/19**	**lun 24/6/19**
2	**PAC0 - Definició dels continguts del treball**	**13 días**	**mié 20/2/19**	**lun 4/3/19**
3	Definir proposta de TFG	13 días	mié 20/2/19	lun 4/3/19
4	**PAC1 - Pla de treball**	**14 días**	**mar 5/3/19**	**lun 18/3/19**
5	Descripció del TFG	5 días	mar 5/3/19	sáb 9/3/19
6	Definició d'objectius	4 días	dom 10/3/19	mié 13/3/19
7	Planificació i temporalització	5 días	jue 14/3/19	lun 18/3/19
8	**PAC2 - Desenvolupament del treball - Fase1**	**35 días**	**mar 19/3/19**	**lun 22/4/19**
9	Estudiar i analitzar el procés de KDD	5 días	mar 19/3/19	sáb 23/3/19
10	Estudiar i analitzar la Mineria de Dades	8 días	dom 24/3/19	dom 31/3/19
11	Estudiar i analitzar el Machine Learning	8 días	lun 1/4/19	lun 8/4/19
12	Desenvolupar una comprensió del domini de l'aplicació	7 días	mar 9/4/19	lun 15/4/19
13	Seleccionar i crear el conjunt de dades	7 días	mar 16/4/19	lun 22/4/19
14	**PAC3 - Desenvolupament del treball - Fase2**	**28 días**	**mar 23/4/19**	**lun 20/5/19**
15	Preprocessament i neteja	8 días	mar 23/4/19	mar 30/4/19
16	Transformació de dades	7 días	mié 1/5/19	mar 7/5/19
17	Escollir la tasca de MD	5 días	mié 8/5/19	dom 12/5/19
18	Escollir i aplicar l'algorisme	3 días	lun 13/5/19	mié 15/5/19
19	Avaluar els resultats obtinguts	5 días	mié 15/5/19	dom 19/5/19
20	Utilitzar el coneixement descobert	1 día	lun 20/5/19	lun 20/5/19
21	PAC4 - Redacció de la memòria	15 días	mar 21/5/19	mar 4/6/19
22	PAC5a - Elaboració de la presentació	8 días	mié 5/6/19	mié 12/6/19
23	PAC5b - Defensa pública	12 días	jue 13/6/19	lun 24/6/19

Figura 3. Diagrama de Gantt de la planificació temporal

1.5 Breu sumari de productes obtinguts

Podem considerar com productes resultants:

- ✓ Els models obtinguts en l'aplicació del procés de KDD
- ✓ Les conclusions i valoracions del projecte
- ✓ Memòria del projecte
- ✓ L'arxiu de vídeo amb l'exposició oral
- ✓ Presentació amb la perspectiva general del projecte

1.6 Breu descripció dels altres capítols de la memòria

La memòria s'estructura fonamentalment en dos blocs diferenciats, per una banda, els **capítols 2** i **3** es dediquen a fer una descripció de les metodologies a utilitzar, analitzant les característiques i situant-les dins del procés del projecte. Es fa una revisió del procés de *KDD* i la importància de les fases de *Mineria de Dade*s, es descriuen les àrees i tecnologies implicades en la *MD*,

focalitzant en el *Machine Learning*. Per finalitzar, es classifiquen les tècniques de *Machine Learning* considerant la supervisió de les dades, i destacant les tècniques principals, *aprenentatge no supervisat i aprenentatge supervisat*.

Per altra banda, al **capítol 4** es desenvolupa el cas pràctic del procés de KDD definit a l'**apartat 1.3**, detallant en cada fase les decisions presses, els passos a seguir i les dificultats trobades. Aquest capítol és el més extens i més temps ha necessitat.

Per finalitzar, Al **capítol 5** s'exposen les conclusions i valoracions obtingudes com a resultat del projecte.

2. Mineria de Dades

Aquest capítol tracta de contextualitzar i interrelacionar les tres principals tecnologies a utilitzar en el treball, *Mineria de Dades* (***MD***), procés de *Descobriment de Coneixement en Bases de dades* (***KDD***) i *Machine Learning* (***ML***) o *Aprenentatge Automàtic*. Comença introduint la *Mineria de Dades*, continua relacionant-la amb el procés de ***KDD***, i analitzant totes les fases d'aquest últim, a continuació s'enumeren les diferents àrees o tecnologies que utilitza la ***MD*** i finalitza relacionant-la amb el ***ML***.

2.1 Introducció a la Mineria de Dades

Com defineixen *Witten i Frank* **[14]** :

> *"La mineria de dades és l'extracció d'informació implícita, prèviament desconeguda i potencialment útil a partir de dades."*

La Mineria de Dades és multidisciplinar, combina diferents disciplines o tecnologies per extreure patrons de les dades i és nodreix de la investigació i avanços d'aquestes. L'Estadística, el Machine Learning, les Bases de Dades, són alguns exemples.

El coneixement extret pot representar-se en relacions, patrons o regles inferides de les dades i desconegudes, o descripcions. Aquestes representacions constitueixen el model de les dades.

La *Mineria de Dades* (*Data Mining*) no és una camp nou, comença a escoltar-se en la dècada dels seixanta, juntament amb altres termes com *data fishing* o *data archaeology*. La continua evolució de la tecnologia ha influenciat en la pròpia evolució de la *Mineria de Dades*, fent-la créixer i madurar, des del inici amb la capacitat de recol·lectar dades, passant per les bases de dades relacionals, el *Data Warehouse* i suport a la presa de decisions, i els algorismes avançats.

Un altre aspecte rellevant ha sigut la possibilitat d'afegir noves disciplines facilitant l'extracció de coneixement, com *la Computació paral·lela i Distribuida* o el *Machine Learning* . Però el impuls més important està produint-se en el Segle XXI amb el increment quasi exponencial de les capacitats del maquinari d'emmagatzemar dades i computar-les, la proliferació de dispositius mòbils interconnectats que no cessen de registrar dades, les xarxes socials, etc. Es podria dir que s'ha generat la moda per la *Mineria de Dades*, com a sortida a l'exorbitant recol·lecció de dades que es produeix diàriament, no té sentit acumular dades si no s'aprofiten.

2.2 Mineria de Dades com a nucli del procés de KDD

Per continuar comprenent la *Mineria de Dades* és necessari contextualitzar-la dins del procés de *KDD*, analitzant les definicions donades poden semblar conceptes sinònims però no ho són. La *Mineria de Dad*es és la part del procés de *KDD* que obté patrons i models utilitzant disciplines o àrees relacionades, com l'*Estadística* o el *Machine Learning* entre d'altres, i *KDD* és el procés complet de descobrir coneixement útil des de les bases de dades. Es podria considerar un procés (*MD*) dintre d'un altre procés (*KDD*).

El procés de *KDD* com s'ha definit al primer capítol d'aquest treball, defineix les propietats del coneixement extret, patrons vàlids, nous, potencialment útils i, finalment, comprensibles:

- ✓ ***vàlid***: fa referència a que els patrons han de continuar sent precisos per noves dades
- ✓ ***nou***: l'aportació ha de ser desconeguda
- ✓ ***potencialment útil***: la informació ha de comportar algun benefici
- ✓ ***comprensible***: la informació ha de ser comprensible per l'usuari, per tant, no només depèn dels patrons. En cas contrari dificulta la interpretació i qüestiona també les altres propietats

2.2.1 Fases del procés de KDD

El procés de KDD consta de 9 fases *(Fayyad, 1996),* és iteratiu ja que depenen de la sortida d'alguna fase pot ser necessari tornar a fases anteriors, i pot requerir de diverses iteracions per extreure coneixement. I és interactiu per que necessita de intervenció humana per la preparació i transformació de les dades, avaluació dels patrons, etc.

1. **Comprensió del domini de l'aplicació**. Entendre i definir els objectius del procés i l'entorn en que es durà a terme. Es pot tornar a aquesta fase si requereix una revisió.
2. **Seleccionar i crear el conjunt de dades**. Decidir les dades que seran objectiu del procés. Analitzar les dades disponibles, obtenir dades addicionals si és necessari i integrar-les en un conjunt. Fase molt important per definir la base per construir els models. Especial consideració als atributs, en aquesta fase no és moment de descartar cap.
3. **Preprocessament i neteja**. Es tracten les dades incompletes i incorrectes, la redundància de dades, els sorolls o els valors atípics. Pot requerir implementat mètodes estadístics complexos o algorismes específics pel context. L'objectiu principal és millorar la qualitat de les dades.
4. **Transformació de les dades.** En aquesta fase s'acaben de preparar les dades per les fases de *Mineria de Dades*, depenen dels objectius i algorismes a aplicar. Aquest pas inclou la *reducció de la dimensionalitat* (millorant l'eficiència) i la transformació d'atributs, com la *Discretització*, la *Normalització* i la conversió de tipus.
5. **Escollir la tasca de mineria de dades**. És el moment d'escollir la tasca, depenent principalment dels objectius a assolir i les fases anteriors. La primera decisió és si es persegueix un objectiu predictiu (*aprenentatge supervisat*) o descriptiu (*aprenentatge no supervisat*), si es predictiu les principals tècniques són la *classificació* i la *regressió*, i si es descriptiu la principal tècnica és el *clustering*.

6. **Escollir l'algorisme**. Es decideix el mètode específic que s'utilitzarà per la cerca de patrons i que concordi completament amb els objectius del procés, per exemple, *arbres de decisió*, *K-means*, etc. Les fases anteriors tenen gran rellevància en l'elecció, com també els paràmetres i les tàctiques d'aprenentatge dels algorismes.
7. **Aplicar l'algorisme**. Finalment, s'aplica l'algorisme per obtenir els models, és possible tornar aplicar l'algorisme ajustant els paràmetres, o els conjunts d'entrenament i test.
8. **Avaluació dels resultats obtinguts**. En aquesta fase s'avaluen i interpreten els models resultants respecte, els objectius fixats al inici del procés. S'analitza la influència de les decisions preses en cada fase anterior, possibilitant tornar a fases anteriors.
9. **Utilitzar el coneixement descobert**. Finalment, s'incorpora el coneixement a un altre sistema per obtenir el benefici final buscat o es fa difusió dels resultats.

2.3 Àrees i tecnologies de Mineria de Dades

La Mineria de Dades utilitza i es recolza en diferents àrees i tecnologies, que tenen una gran influència en el desenvolupament de mètodes de *Mineria de Dades*(**Figura 4**):

Estadística

És una disciplina que estudia la recopilació, anàlisi, interpretació i presentació de dades. Construeix un model estadístic amb un conjunt de funcions matemàtiques, les quals descriuen el comportament dels objectes en una classe objectiu.

L'Estadística proporciona moltes tècniques, conceptes i algorismes que s'apliquen en Mineria de Dades, com la mitja, la variància, la validació creuada, la regressió, les tècniques bayesianes, etc.

Machine Learning (Aprenentagte automàtic)

És l'àrea de la *Intel·ligència Artificial* (**IA**) que s'encarrega de desenvolupar algorismes que puguin aprendre a reconèixer patrons complexes i prendre decisions intel·ligents basades en les dades.

Existeixen moltes semblances en els principis a seguir, tant de *Mineria de Dades* com de *Machine Learning*, es construeix un model a partir de dades i s'aplica per resoldre el problema. Però mentre que el *Machine Learning* es centra principalment en la precisió del model, la *Mineria de Dades* també dóna molta rellevància a l'eficiència i escalabilitat en grans conjunts de dades.

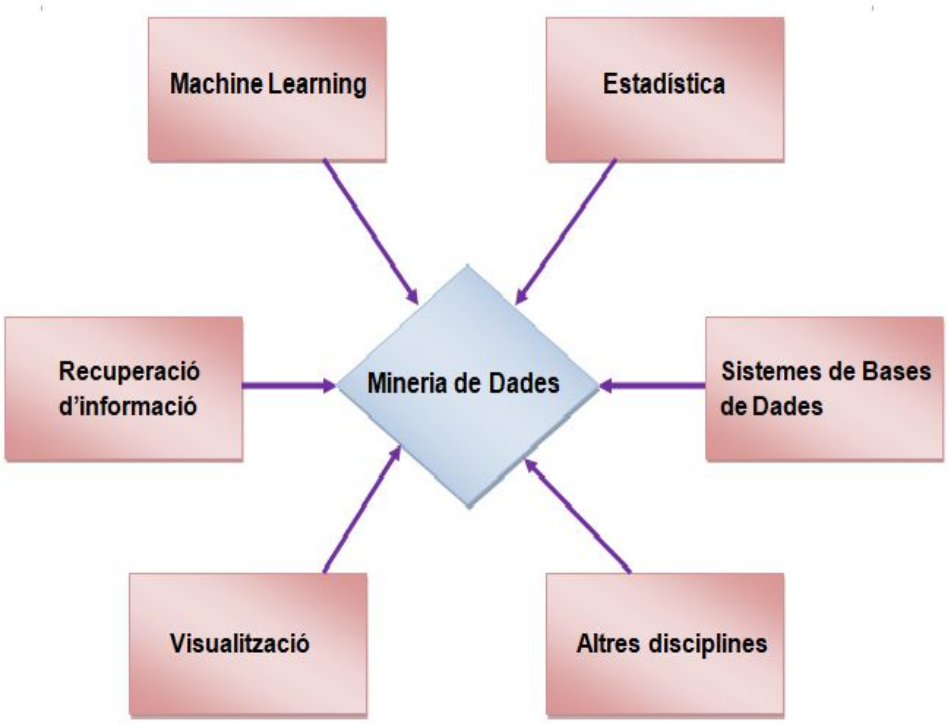


Figura 4. Disciplines que contribueixen a la Mineria de Dades

Recuperació de la informació (Information Retrieval, IR)

Disciplina que s'encarrega de la recerca de documents o d'informació en documents, ja siguin de text o multimèdia. Els enfocaments típics en la recuperació d'informació es basen models probabilístics, com a exemple, considerant un document de text com un conjunt de paraules, el model de llenguatge del document és la funció de densitat de probabilitat que genera el conjunt de paraules al document.

El desmesurat augment de dades de text i multimèdia disponibles, a causa del creixement d'*Internet*, ha provocat que la *Mineria de Text* (*text Mining*) i la *Mineria de Dades* multimèdia, integrats amb mètodes de Recuperació d'Informació esdevinguin cada vegada més rellevants.

Les principals desavantatges respecte les bases de dades són, la impossibilitat de realitzar consultes complexes, principalment utilitza paraules clau, i que les dades buscades no estan estructurades.

Sistemes de Bases de Dades

Tecnologia que s'encarrega de la creació, manteniment i ús de bases de dades. S'han establert uns principis en llenguatges de consulta, mètodes de processament i optimització de consultes, model de dades, emmagatzematge de dades i mètodes d'indexació i accés.

Una característica reconeguda és l'alta escalabilitat en el processament de conjunts de dades molt grans i relativament estructurats, la Mineria de Dades té la possibilitat d'utilitzar bases de dades amb aquesta característica per aconseguir una alta eficiència.

Els últims *Sistemes de Bases de Dades* integren anàlisi de dades sistemàtiques utilitzant eines de *Mineria de Dades* i *Data Warehousing*. Un *Data warehouse* (magatzem de dades) agrupa dades de diferents fonts i històriques, i és capaç d'integrar sistemes de presa de decisions.

Visualització de dades

L'anàlisi visual és una forma potent i intuïtiva d'extreure coneixement a partir de les dades. La utilització de tècniques o eines de visualització permet descobrir, intuir o entendre patrons difícilment descoberts a partir de resultats textuals o matemàtics.

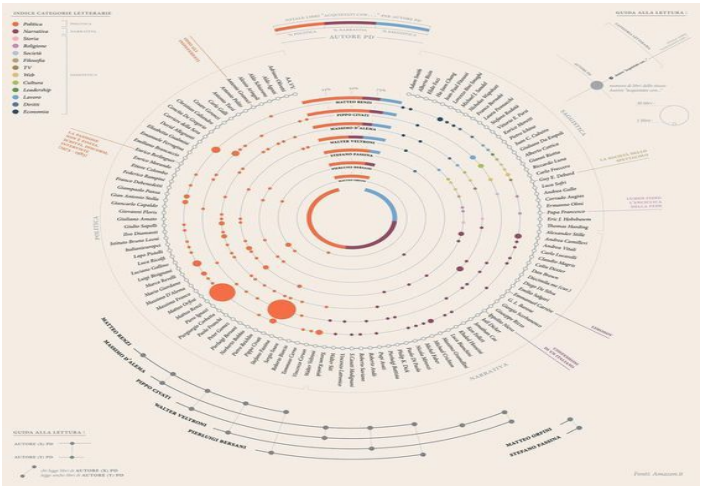


Figura 5. Exemple de gràfic circular

Les eines de visualització aporten llum a la interpretació dels models resultants, i combinades amb eines d'estadística descriptiva són molt útils en la preparació de les dades. Entre moltes d'altres es troben els histogrames, els diagrames de dispersió, els diagrames de caixes, tècniques de projecció en 2D i 3D, i els gràfics circulars.

Altres disciplines

La Mineria de Dades utilitza altres disciplines com la computació paral·lela i distribuïda, sistemes per la pressa de decisions, reconeixement de patrons, tècniques de llenguatge natural, anàlisi d'imatges, processament de senyals, etc.

2.4 *Machine Learning* com a impulsor de la *MD*

La *Mineria de Dades* és nodreix de la investigació i avanços de les disciplines i tecnologies relaciones amb la *MD* que s'acaben d'apuntar en l'apartat anterior, però segurament és el Machine Learning el principal culpable de l'actual rellevància de la *MD*. La millora en els algorismes i la generació de nous ha propiciat la reducció de dades d'entrenament, millora en el temps d'execució i nivell de confiança, a més, ha sorgit nous camps com el ***Deep Learning***, sub-àrea del *Machine Learning* amb vida pròpia i moltes possibilitats de present i futur.

En el **capítol 3** es tractarà més en detall aquesta àrea de la *IA*, descrivint els tipus de *Machine Learning* segons la supervisió de les dades.

3. Tècniques de *Machine Learning*

Les tècniques representen l'enfocament conceptual per extreure coneixement de les dades, i acostumen a ser implementades per diferents algorismes. Per tant, cada algorisme aplica en la pràctica el desenvolupament d'una tècnica concreta pas a pas, requerint una alta comprensió dels algorismes per poder discernir que tècnica és la més adequada pel problema en qüestió.

Existeixen diferents raonaments o metodologies per classificar les tècniques de *Machine Learning*, s'ha escollit possiblement la visió més freqüent i compartida amb la *Mineria de Dades* (com s'ha comentat existeixen moltes similituds en els principis de *ML* i *MD*), classificar en base a la supervisió de les dades. Existeixen quatre categories: *Aprenentatge supervisat, Aprenentatge no supervisat, Aprenentatge semi-supervisat i Aprenentatge per reforç*, però l'estudi es centra en les dues primeres categories que són les principals amb molta diferència. No és la intenció d'aquest treball aprofundir en l'anàlisi de les tècniques i algorismes de *Machine Learning*, s'analitzen les principals i que tinguin relació directa amb el cas d'estudi del treball (capítol 4). A **la figura 6** es pot veure un gràfic simplificat de l'estudi d'aquest capítol.

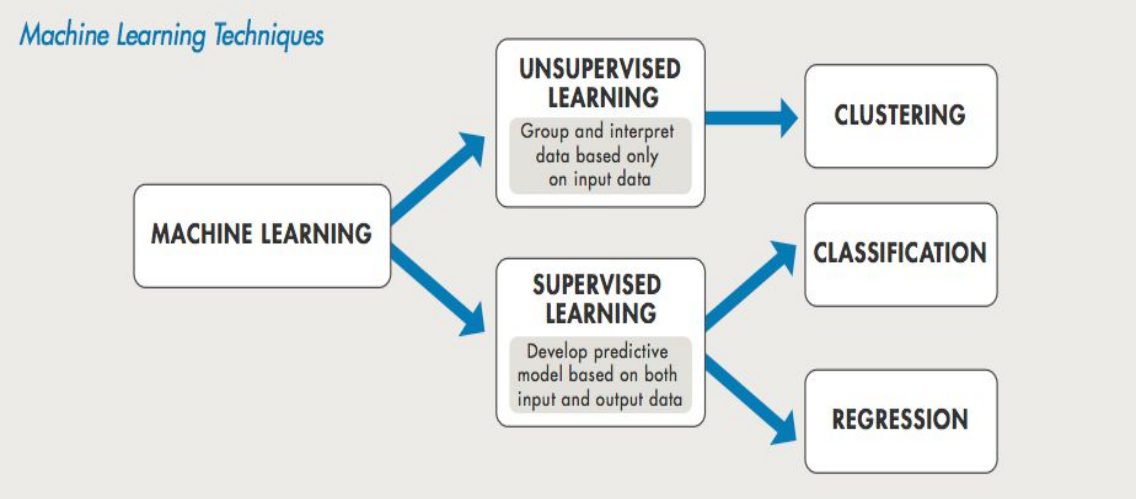


Figura 6, Tècniques de *Machine Learning*

Una altra classificació de les tècniques de Machine Learning molt utilitzada, és classificar per objectiu que es vol obtenir de les dades, *predicció*, *i descripció*, però en la pràctica és igual que la classificació aquí utilitzada, ja que *predicció* i *Aprenentatge supervisat* s'utilitzen indistintament, i passa el mateix amb *descripció* i *Aprenentatge no supervisat*.

En l'*Aprenentatge supervisat* es parteix d'una situació de cert coneixement, ja que es coneix el valor de sortida que es vol predir. Utilitza tècniques que intenten descobrir la relació entre els atributs d'entrada i l'atribut de classe (sortida), el qual descriu i explica informació oculta de les dades i, s'utilitza per predir el valor de l'atribut de classe només coneixent els valors d'entrada.

En l'*Aprenentatge no supervisat* es parteix d'una situació amb menys coneixement, no existeix cap variable de sortida, les tècniques que s'utilitzen identifiquen relacions o patrons ocults descobrint l'estructura subjacent de les dades.

Existeixen moltes altres tècniques i algorismes, però en aquest capítol els algorismes escollits són els que s'han considerat més adequats, per les condicions de les dades, pels objectius del cas d'estudi i per la disponibilitat dels algorismes implementats en *Weka*.

Les tècniques analitzades en aquest capítol no s'apliquen exclusivament als tipus de problema on es classifiquen, algunes poden aplicar-se a d'altres objectius.

3.1. Aprenentatge supervisat

L*'Aprenentatge supervisat* el poden dividir en dos models principals: models de ***classificació*** i models de ***regressió***, es distingeixen pel tipus de variable objectiu, mentre que un model de *classificació* prediu una categoria com a valor de sortida, un model de *regressió* prediu un nombre real.

Juntament amb l'estimació de la probabilitat i la regressió, la classificació és un dels models més estudiats, possiblement el que té major rellevància pràctica. El potencial del progrés en la classificació és molt gran, al tenir un gran impacte en altres àrees, tant en *Data Mining* com en les seves aplicacions.

Les tècniques representen l'enfocament conceptual per extreure coneixement de les dades, i acostumen a ser implementades per diferents algorismes. Per tant, cada algorisme aplica en la pràctica el desenvolupament d'una tècnica concreta pas a pas, requerint una alta comprensió dels algorismes per poder discernir que tècnica és la més adequada pel problema en qüestió.

Analitzant les tècniques de *regressió,* els objectius i dades del cas d'estudi i la disponibilitat dels algorismes implementats en *Weka*, no s'aplicarà cap tècnica de regressió a les dades del cas. Indiferentment, s'analitzaran superficialment algunes tècniques.

Regressió

En els models de *regressió lineal* tant les variables dependents com la independent (atribut de classe) han de ser continues. En la ***regressió lineal simple*** només es disposa d'una variable dependent, mentre que a la ***regressió lineal múltiple*** es disposa de més d'una variable dependent.

Altres tècniques de *regressió* són les *xarxes neurals artificials, arbres de decisió (CART), màquines de vector de suport, algorismes genètics*, etc, algunes es veuran a continuació per *classificació*.

3.1.1 Classificació

Les tècniques de *classificació* prediuen una categoria com a valor objectiu, per tant, el conjunt de resultats possibles és finit. Quan aquest conjunt només té dos categories possibles com a objectiu, s'anomena *classificació binaria*, i quan en té més *classificació multicategoria*. A continuació les tècniques de *classificació* a destacar:

Arbres de decisió (arbres de classificació). És de les tècniques més fàcils d'entendre, la *classificació* és la que millor s'adequa però també s'utilitza per *regressió* i *clustering*. Consisteix en un conjunt de condicions en estructura jeràrquica de dalt a baix, que finalitza en la decisió final de classe, on als nodes hi són els atributs i als arcs els possibles valors dels atributs.

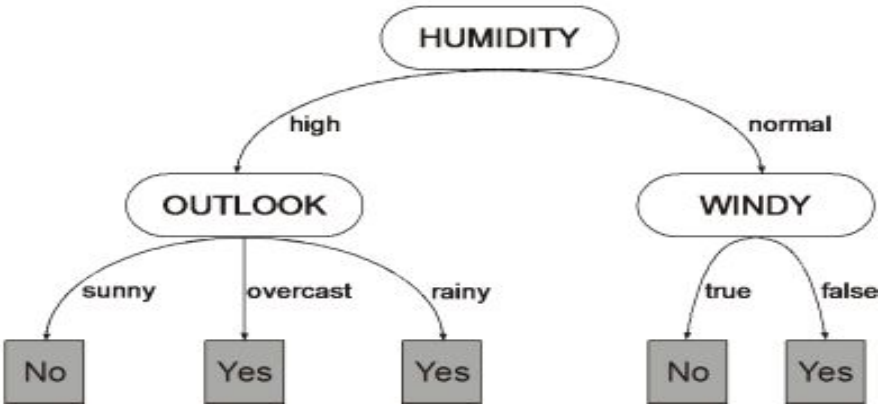


Figura7. Exemple de representació d'un arbre de decisió

La característica principal consisteix que tant les classes (classificació final de la instància) com les condicions són disjuntes, és a dir, una instància només pot pertànyer a una classe, i només pot acomplir una condició a l'arribar a un node. Per tant, quan una instància va descendent per l'arbre només pot seguir un camí.

Els primers algorismes d'arbres de decisió no permetien les particions numèriques (atributs numèrics), ja que només utilitzaven particions discretes, posteriors modificacions han permès aquesta possibiltat ampliant el camp d'aplicament.

L'algorisme ***C4.5*** desenvolupat per Ross Quinlan (millora de l'algorisme ID3), permet les particions numèriques separant-les en intervals. Per definir els intervals, ordena els valors treien els repetits i obté el valor intermedi entre els dos valors. Aquest algorisme s'aplicarà al cas d'estudi del treball amb el programari Weka.

Un altre algorisme que s'aplicarà al cas d'estudi del treball, és l'algorisme ***Random Forest***, el qual es basa en el desenvolupament de molts *arbres de classificació*, la instància o exemple recorre tots els arbres i la millor sortida és l'escollida.

L'últim algorisme d'*arbres de classificació* que s'aplicarà, és l'algorisme ***LMT***, peculiar ja que combina models de *regressió logística* amb *arbres de decisió*. Consisteix en una estructura d'*arbre de decisió* amb funcions de *regressió logística* a les fulles. Per a atributs numèrics, el node té dos nodes fills i la condició consisteix a comparar el valor de l'atribut amb un llindar.

Existeixen molts altres algorismes de *classificació* relacionats amb arbres, però s'ha decidit que els tres algorismes escollits són els més adequats, per les condicions de les dades, pels objectius del cas d'estudi i per la disponibilitat en *Weka*.

Classificació Bayesiana. Les tècniques de classificació Bayesiana utilitzen models probabilístics i es basen el teorema de Bayes, el qual s'utilitza per calcular la probabilitat de un succés tenint informació avançada sobre el succés.

El representant més conegut i simple dels classificadors Bayesians és l'algorisme ***Naïve Bayes***, el qual classifica nous exemples assignant-li la classe que maximitza la probabilitat condicional de la classe, donada la seqüència observada d'atributs en l'entrenament. Es basa en la suposició que tots els atributs són independents i que segueixen una distribució normal, la qual cosa pot resultar un inconvenient.

Veí més proper. És una tècnica i algorisme senzill de *classificació* basat en l'emmagatzematge d'un conjunt d'exemples, on a un nou objecte se li assigna la classe de l'exemple més proper, aquest algorisme es conegut com **1-NN**.

Una variant és l'algorisme k veïns més pròxim (*k-nearest neighbors*, ***kNN***), on a un nou objecte se li assigna la classe majoritària entre els *k* veïns més propers. Aquest algorisme s'aplicarà al cas d'estudi del treball amb el programari *Weka*.

Xarxes Neurals Artificials (XNA). És un mètode inspirat en el comportament observat en les xarxes neurals biològiques. Es basa en la presumpció de que a causa de la naturalesa biològica del nostre cervell té la capacitat de processar informació i intentar imitar-ho, el perceptró és la representació artificial de una neurona.

Un dels primers algorismes és el ***Perceptró simple***, amb una estructura de varis nodes o neurones d'entrada i un o més de sortida, sense capa oculta. L'algorisme que interessa pel cas d'estudi és l'algorisme ***Perceptró Multicapa***, evolució de l'algorisme anterior que estableix una xarxa neuronal en forma de cascada, amb una o més capes ocultes. Millora substancialment en casos que el conjunt de dades no és linealment separable.

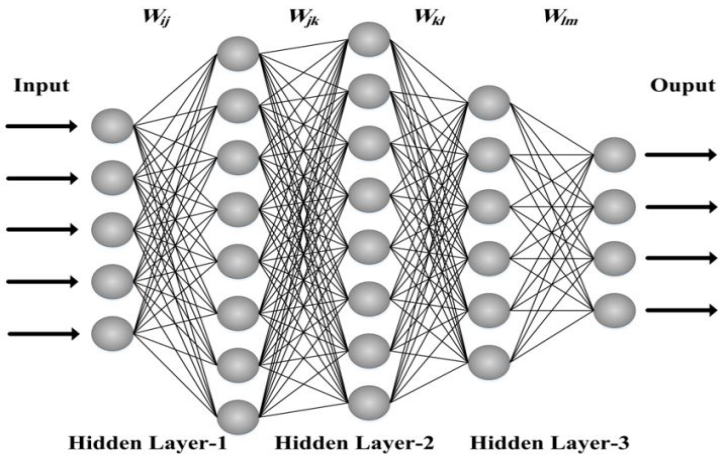


Figura 8. Exemple Perceptró Multicapa

3.2 Aprenentatge no supervisat

L'*Aprenentatge no supervisat* el poden dividir en dos models principals: models de ***clustering o agrupació*** i models de ***regles d'associació***. En un model de *clustering* s'agrupen els objectes en grups amb tècniques iteratives, de manera que els objectes dins d'un grup, siguin molts semblants entre ells (homogenis), i aquests molt diferents a qualsevol d'altre grup (heterogenis). Mentre que un model de *regles d'associació* permet descobrir patrons en comú entre els objectes que pertanyen a un conjunt de dades. Per a això, es consideren totes les possibles combinacions d'atribut-valor de les dades del conjunt.

Analitzant les tècniques de *regles d'associació*, els objectius i dades del cas d'estudi i la disponibilitat dels algorismes implementats en Weka, no s'aplicarà cap tècnica a les dades del cas. Indiferentment, s'analitzaran superficialment algunes tècniques.

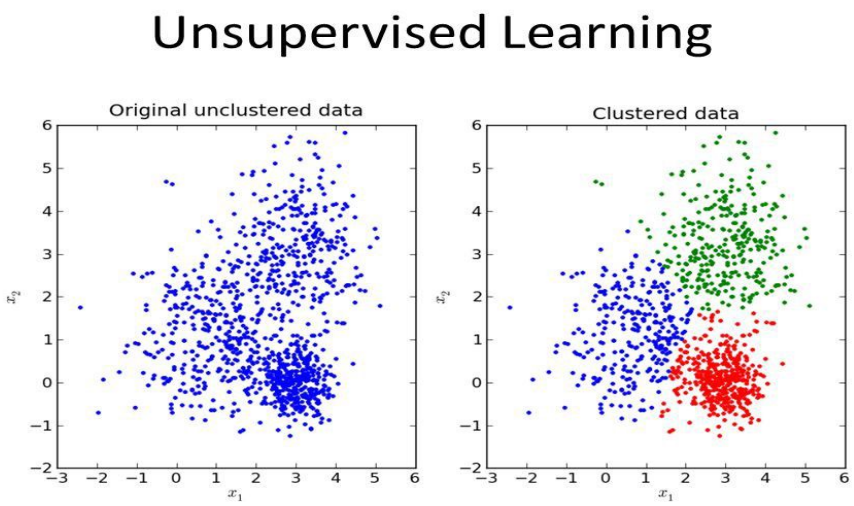


Figura 9. Exemple d'Aprenentatge no supervisat

Regles d'associació

Són tècniques que es basen en la confiança i cobertura (suport), s'utilitzen amb l'objectiu de realitzar anàlisis exploratoris, buscant regles que acompleixin un

nivell mínim d'aquestes mesures. Les possibles relacions o correlacions es poden utilitzar per predir comportaments.

No existeixen gaires algorismes de *regles d'associació*, el més conegut i utilitzat és l'algorisme ***Apriori***, el qual es basa en la cerca dels conjunts d'ítems amb una determinada cobertura o suport, per després cercar les regles que tinguin un nivell mínim de confiança.

3.2.1 Clustering o agrupació

Les tècniques de *clustering* es poden classificar en dos tipus, les jeràrquiques, que van formant grups progressivament, i els *particionals*, que només calculen una partició de les dades.

És freqüent que el clustering precedeixi a la classificació de nous objectes utilitzant els grups obtinguts amb el clustering, per això, són dos tècniques estretament relacionades.

La principal característica és l'ús de mesures de similaritat, basada en els atributs que descriuen els objectes, definint-se habitualment per proximitat en un pla multidimensional. Per atributs amb dades numèriques és necessari estandarditzar-los.

Agrupament numèric. L'algorisme d'agrupament numèric més utilitzat és el ***k-means***, utilitza *k* elements aleatòriament i prèviament definits que representen els centres (centroides) dels clústers inicials, a continuació utilitza la distància Euclidea per assignar els nous exemples al clúster (k) més proper. Finalitzada la primera iteració, tots els exemples estan assignats, es torna a calcular els centroides amb els exemples classificats a cada clúster, amb aquests nous centroides es tornen a classificar tots els exemples, el procés continua iterant fins que en dos iteracions consecutives es repeteixen els mateixos centroides.

Agrupament probabilístic. El principal algorisme d'agrupament probabilístic és el **EM** (*Expectation Maximization*), calcula el nombre de clústers (k) més adequat automàticament, per tant, no cal definir-lo prèviament com al *k-means*. Més complex i menys eficient que altres algorismes, es basa en les probabilitats de pertànyer a un clúster. Més adient per atributs numèrics.

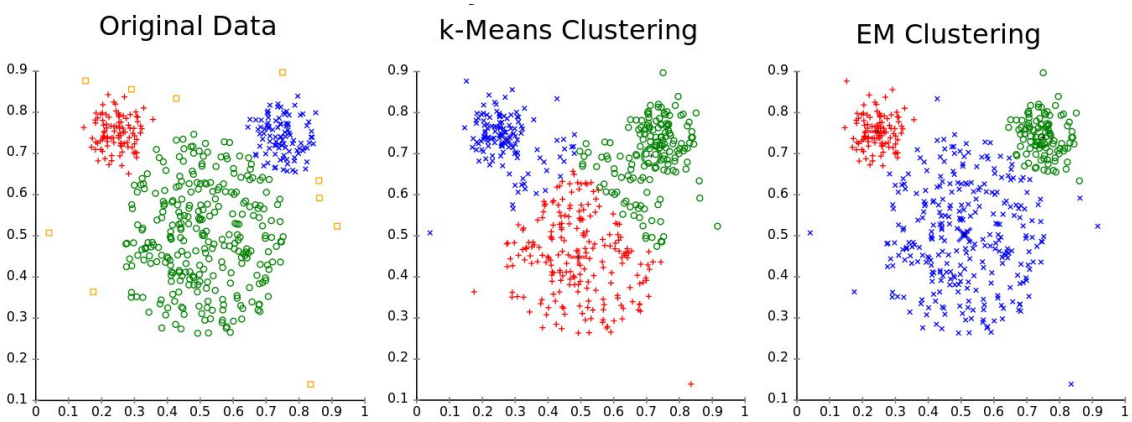


Figura 10. Diferència entre K-means i EM sobre les mateixes dades

En aquest capítol, només s'han tractat les tècniques i algorismes que s'aplicaran al cas d'estudi del treball, considerant principalment el tipus de dades que s'utilitza i els algorismes i les implementacions d'aquests que utilitza *Weka*.

4. Procés de KDD aplicat a dades d'exclusió social

Un cop realitzada la síntesi de les tècniques a aplicar en el cas d'estudi, es desenvoluparà la metodologia definida al primer capítol, per començat introduint els criteris i l'elecció dels indicadors d'*exclusió social*, qüestió complexa i controvertida.

En un principi es va plantejar realitzar el cas d'estudi amb tècniques de *clustering*, amb l'objectiu d'agrupar les dades per comprendre-les i trobar patrons interessants. Després d'un període d'anàlisi i proves embrionàries es va decidir aplicar tècniques de classificació per completar l'anàlisi de les dades preparades. Consegüentment, el capítol es divideix en dos parts, en la primera part s'apliquen tècniques de *clustering*, es presenten els models resultants i es valoren. A la segona part, es realitza el mateix procés amb tècniques de classificació.

Entrant en matèria, la definició conceptual i elecció del indicadors d'exclusió social són força controvertits entre els investigadors socials i experts. Deixant a banda qüestions conceptuals i terminològics, ens centrarem en quin tipus d'indicadors són els adequats pel context del cas d'estudi.

En primer lloc, és conegut en l'àmbit que tracta aquesta problemàtica, les diferències que es produeixen a l'analitzar els indicadors depenent de la regió, país o continent, fins i tot en delimitacions geogràfiques més reduïdes (fins i tot es podria afegir la falta de consensos existents dintre de les pròpies delimitacions).

Una altra dificultat per escollir els indicadors es troba en quines dades s'utilitzen i com es recullen, les principals iniciatives en aquest sentit per part d'institucions públiques es basen en enquestes, per tant, es podria considerar que generen certa informació i coneixement de la situació, però no representen anàlisi exhaustius i reals aplicables a polítiques socials. El principal problema és intentar fer una anàlisi individual, quan avui dia no és possible tenir les dades per analitzar la situació de tots els ciutadans, en un altre sentit, les accions i iniciatives de serveis socials i altres institucions enfocades a sectors

de la població en perill d'exclusió, es poden considerar mesures atenuadores però no arriben a mitigar tots els factors que influeixen en l'exclusió social (bàsicament són mesures de supervivència, menjar, sostre...), a més, molta població vulnerable es queda fora d'aquestes mesures, ja sigui per voluntat pròpia o per falta de recursos en les iniciatives.

El caràcter multidimensional de l'exclusió social implica analitzar diferents àmbits, com el deteriorament de les àrees urbanes, el nivell d'educació de la població, el nivell d'atur, taxa d'immigració, composició de les unitats familiars, nivell de delinqüència, etc. És fonamental afrontar el problema de manera global, amb objectius de govern i no com a polítiques marginals de departaments (socials) amb pocs recursos, aquest tipus de polítiques han de ser complementàries. En aquest sentit, iniciatives que ja s'han comentat com *l'Atlas de vulnerabilidad Urbana en España* i els *indicadors territorials de risc de probresa i d'exclusió social* (***INTPOBR***) són un referent per aquest treball, encara que en el primer cas es realitza amb freqüència decennal (1991-2001-2011) i en el segon la delimitació geogràfica sembla massa àmplia per extreure resultats aplicables. L'última iniciativa per destacar és el *Análisis Urbanístico de Barrios Vunerables*, també emprès pel *Ministeri de Foment* del *Govern d'Espanya* i realitzat els anys 1991, 2001, 2006 i 2011. El qual servirà com a marc general per la delimitació geogràfica, els barris, i per l'elecció dels indicadors, es poden veure a **[15]** i més endavant..

4.1 Procés de KDD amb tècniques de clustering

En aquest apartat es desenvoluparà el cas d'estudi aplicant-li tècniques d'agrupació, les fases inicials no difereixen gaire del cas d'estudi desenvolupat al següent apartat amb tècniques de classificació.

Les primeres fases, comprensió del domini, recollir les dades segons indicadors i crear el conjunt de dades a analitzar són les fases que més temps han comportat.

4.1.1 Comprensió del domini de l'aplicació

Aquesta fase com s'ha pogut veure està detallada al principi d'aquest capítol (4), la raó és per que considerant la seva rellevància i donat que es realitzen dos casos d'estudi, es considera que és el lloc més adient, evitant repetir-ho en els dos casos d'estudi. Per tant, després d'estudiar i analitzar la problemàtica de l'exclusió social, la controvèrsia amb els indicadors i les dificultats de recollida de dades, i com ja s'ha comentat repetidament, l'objectiu és agrupar les dades per comprendre-les i trobar patrons interessants, s'intenta extreure coneixement no obvi per comprendre les desigualtats entre barris de Barcelona i quins indicadors són els que més influència tenen.

Com també s'ha comentat abans, al primer capítol, els recursos necessaris per realitzar el cas d'estudi són mínims i de fàcil abast, el recurs fonamental és el programari lliure ***WEKA*** (*Waikato Environment for Knowledge Analysis*), distribuït sota llicència GNU-GPL, basat en un conjunt de llibreries Java i desenvolupat a la *Universitat de Waikato*. És una col·lecció d'algorismes d'aprenentatge automàtic per a tasques de mineria de dades. Conté eines per a la preparació de dades, classificació, regressió, agrupació, regles d'associació i visualització.

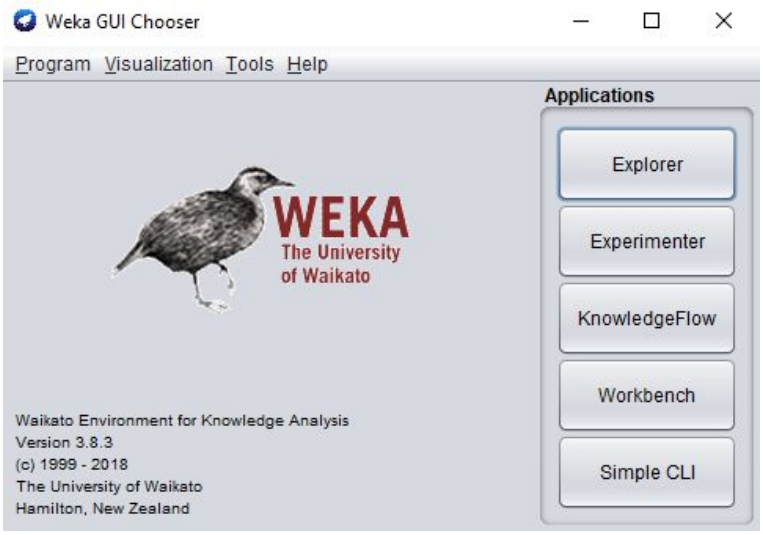


Figura 11. Interfície inicial de *Weka*

Weka ofereix la possibilitat d'ampliar memòria, si és necessari, iniciant el programa des de la consola o modificant l'arxiu *RunWeka.ini*, però pels dos casos d'estudi no serà necessari ampliar la memòria, els datasets a utilitzar són molt petits.

4.1.2 Seleccionar i crear el conjunt de dades

És el moment de definir els indicadors abans de començar a seleccionar les dades necessàries, el marc escollit per definir els indicadors és el *Análisis Urbanístico de Barrios vulnerables*, i en concret, els indicadors de vulnerabilitat urbana, 20 indicadors dividits en 4 temàtiques, Vulnerabilitat Sociodemogràfica, Vulnerabilitat Socioeconòmica, Vulnerabilitat Residencial i Vulnerabilitat Subjectiva.

Nº	Denominació	Descripció	Tipus
Vulnerabilitat Sociodemogràfica			
1	Percentatge de llars unipersonals majors de 64 anys	Llars constituïdes per una sola persona major de 64 anys respecte al conjunt de les llars	Numèric amb el rang 1-100
2	Índex de sobre-envelliment	Persones majors de 74 anys respecte al total	Numèric amb el rang 1-100
3	Índex de població estrangera en edat infantil	Nens menors de 15 anys de nacionalitat estrangera, respecte al total de nens menors de 15 anys	Numèric amb el rang 1-100
4	Índex de població estrangera	Percentatge d'immigrants estrangers respecte al total de la població	Numèric amb el rang 1-100
5	Percentatge de llars monoparentals	Llars amb un adult i un o més menors, respecte al conjunt de llars	Numèric amb el rang 1-100
Vulnerabilitat Socioeconòmica			
6	Taxa d'atur	Percentatge d'aturats respecte al total de la població activa	Numèric amb el rang 1-100
7	Taxa d'atur juvenil	Percentatge de la població de 16 a 29 anys en situació d'atur respecte al total de població activa de 16 a 29 anys	Numèric amb el rang 1-100
8	Taxa d'ocupats eventuals	Percentatge d'ocupats que són treballadors per compte d'altri amb caràcter eventual, temporal... sobre el total d'ocupats	Numèric amb el rang 1-100

9	Taxa de treballadors no qualificats	Percentatge de treballadors no qualificats respecte el total d'ocupats	Numèric amb el rang 1-100
10	Taxa de població sense estudis	Percentatge de població major de 16 anys que no disposa de cap titulació	Numèric amb el rang 1-100
Vulnerabilitat Residencial			
11	Percentatge d'habitatges amb una superfície útil menor a 31 m^2	Habitatges principals que tenen una superfície útil menor de 31 m^2 respecte el total d'habitatges principals	Numèric amb el rang 1-100
12	Superfície mitja de l'habitatge per ocupant	Metres quadrats per ocupant en els habitatges principals	Numèric continu
13	Percentatge de persones residents en habitatges sense servei i bany auxiliar	Percentatge de persones residents en habitatges principals que no tenen servei o bany auxiliar a dins de l'habitatge respecte al total de persones residents en habitatges principals convencionals	Numèric amb el rang 1-100
14	Percentatge d'habitatges situades en edificis en un estat de conservació dolent	Habitatges principals convencionals situades en edificis en situació ruïnosa o deficient respecte al total d'habitatges	Numèric amb el rang 1-100
15	Percentatge d'habitatges situades en edificis construïts abans de 1951	Habitatges principals convencionals situades en edificis construïts abans de 1951	Numèric amb el rang 1-100
Vulnerabilitat Subjectiva			
16	Percentatge d'habitatges la persona dels quals de referència considera que el seu habitatge aquesta afectada per sorolls exteriors		Numèric amb el rang 1-100
17	Percentatge d'habitatges la persona dels quals de referència considera que el seu habitatge està afectat per contaminació o males olors provocades per la indústria, el trànsit....		Numèric amb el rang 1-100
18	Percentatge d'habitatges la persona dels quals de referència considera que el seu lloc de residència té males comunicacions		Numèric amb el rang 1-100
19	Percentatge d'habitatges la persona dels quals de referència considera que el seu lloc de residència té poques zones verdes en la seva proximitat (parcs, jardins...)		Numèric amb el rang 1-100
20	Percentatge d'habitatges la persona dels quals de referència considera que el seu lloc de residència està afectat per un mitjà social on la delinqüència i el vandalisme són un problema		Numèric amb el rang 1-100

Figura 12. Indicadors d'exclusió social

Respecte els valors dels atributs, es considera que a valors més alts més risc d'exclusió social, excepte el indicador 12 que es considera el contrari. També és important valorar la dependència entre indicadors, en aquest cas es consideren independents per dues raons, la primera és que la tria d'indicadors es realitza amb l'objectiu d'analitzar diferents dimensions, i la segona és a causa del tipus i rang dels valors, ja que tots els indicadors (excepte el 12)

tenen la mateixa mida absoluta (per cent). Encara que és fàcil veure una correlació entre els dos indicadors d'estrangers (indicadors 3 i 4).
Amb els indicadors definits, és a dir, les dades que seran objectiu del procés, el següent pas és analitzar les dades disponibles al servei de dades obertes de Barcelona, *Open Data BCN*, per intentar aconseguir els màxims indicadors possibles.

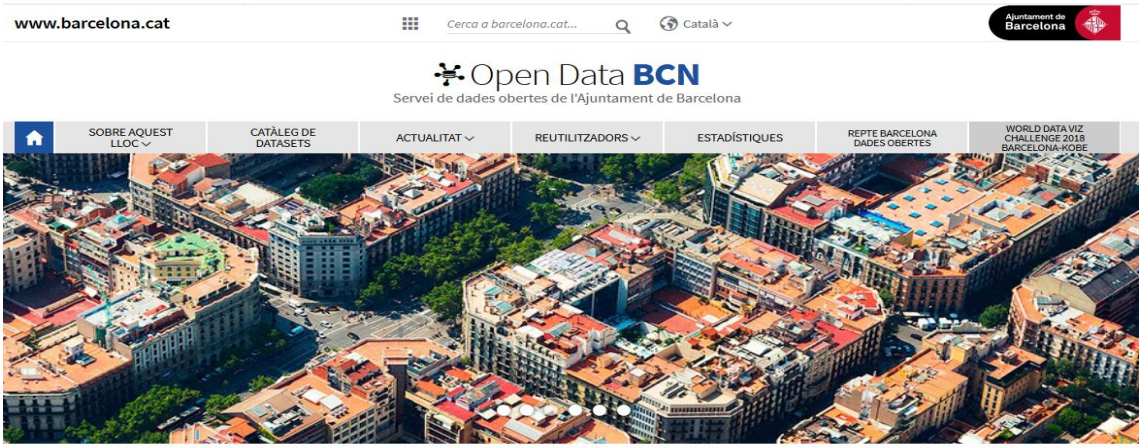


Figura 13. Portal de l'Open Data BCN

Aquesta fase és la més extensa en dedicació amb molta diferència, principalment per la dificultat en trobar dades i adaptar-les als indicadors considerats. La situació òptima seria fer la recollida de dades considerant prèviament els indicadors i el seu tipus, però donada les evidents limitacions d'aquest treball, es realitza un important esforç per la cerca i preparació de les dades. En el moment d'inici del treball es realitza un primer anàlisi dels datasets disponibles al servei de dades obertes, i es decideix desenvolupar l'estudi a dades de l'any 2017, al mancar algun dataset de l'any 2018. Es decideix utilitzar els datasets en format csv per unificar, ja que és l'únic format disponible en tots els datasets (no tots els datasets estan en format excel), a pesar de que el necessari tractament de les dades es realitza amb excel.

ÉS important comentar que al llarg de la realització del treball, l'Open Data BCN ha realitzat canvis en els datasets, modificant, creant i esborrant, i no sempre per millorar la comprensió i el tractament de les dades que es necessita per aquest treball. Com a exemple, el dataset utilitzat per obtenir les dades pels

indicadors 1 i 5 (indicadors de composició de les llars) anomenat *Estructura dels domicilis de la ciutat de Barcelona* (**figura 14**).

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P
1	Dte.,"Barris","TOTAL","Una dona sola de 18 a 64 anys","Un home sol de 18 a 64 anys","Una dona sola de 65 anys i més","Un home sol de 65 anys i més","Dues persones de 18 a 64 anys","Dues persones															
2	de 18 a 64 anys","Dues persones															
3	de 65 anys i més","Dues persones,															
4	una de 18-64 anys i															
5	l'altra de 65 i més","Dues persones															
6	o més, totes															
7	majors de 18 anys","Dues persones															
8	o més: una dona de															
9	18 anys i més amb															
10	altres menors de 18 anys","Dues persones															
11	o més: un home de															
12	18 anys i més amb															
13	altres menors de 18 anys","Tres persones															
14	o més: dues de															
15	18 anys i més i la															
16	resta menor de 18 anys","Altres domicilis															
17	amb una o més															
18	persones menors de															
19	18 anys"															
20	1,"1. el Raval",16.851,1.727,2.343,1.278,614,2.475,644,721,3.297,303,56,1.376,2.118															
21	1,"2. el Barri Gòtic",6.450,873,1.014,432,227,1.169,249,293,1.277,110,32,418,370															
22	1,"3. la Barceloneta",6.850,882,922,700,223,1.193,380,345,1.206,158,27,503,373															
23	1,"4. Sant Pere, Santa Caterina i la Ribera",9.954,1.401,1.504,816,329,1.838,424,453,1.637,241,38,750,577															
24	2,"5. el Fort Pienc",12.805,1.152,1.123,1.332,404,1.615,1.110,730,2.557,268,83,1.635,831															
25	2,"6. la Sagrada Família",21.844,2.169,1.730,2.540,678,2.994,1.875,1.306,4.224,463,105,2.391,1.366															
26	2,"7. la Dreta de l'Eixample",18.391,1.960,1.784,1.900,681,2.474,1.305,976,3.321,469,111,2.299,1.078															
27	2,"8. l'Antiga Esquerra de l'Eixample",18.104,1.988,1.832,2.058,579,2.407,1.379,1.063,3.308,428,106,1.965,1.013															
28	2,"9. la Nova Esquerra de l'Eixample",24.515,2.257,2.162,2.790,807,3.155,2.171,1.510,4.850,516,114,2.657,1.495															
29	2,"10. Sant Antoni",16.383,1.501,1.592,1.996,553,2.209,1.446,961,2.962,319,70,1.699,1.080															
30	3,"11. el Poble Sec - Parc Montjuïc",15.983,1.648,1.632,1.514,519,2.435,1.055,800,2.914,336,64,1.646,1.462															

	A	B	C	D	E	F	G	H	I
1	Any,"Codi_Districte","Nom_Districte","Codi_Barri","Nom_Barri","Estructura_domicili","Nombre"								
2	2017,1,"Ciutat Vella",1,"el Raval","Una dona sola de 18 a 64 anys",1727								
3	2017,1,"Ciutat Vella",2,"el Barri GÃ²tic","Una dona sola de 18 a 64 anys",873								
4	2017,1,"Ciutat Vella",3,"la Barceloneta","Una dona sola de 18 a 64 anys",882								
5	2017,1,"Ciutat Vella",4,"Sant Pere, Santa Caterina i la Ribera","Una dona sola de 18 a 64 anys",1401								
6	2017,2,"Eixample",5,"el Fort Pienc","Una dona sola de 18 a 64 anys",1152								
7	2017,2,"Eixample",6,"la Sagrada FamÃ-lia","Una dona sola de 18 a 64 anys",2169								
8	2017,2,"Eixample",7,"la Dreta de l'Eixample","Una dona sola de 18 a 64 anys",1960								
9	2017,2,"Eixample",8,"l'Antiga Esquerra de l'Eixample","Una dona sola de 18 a 64 anys",1988								
10	2017,2,"Eixample",9,"la Nova Esquerra de l'Eixample","Una dona sola de 18 a 64 anys",2257								
11	2017,2,"Eixample",10,"Sant Antoni","Una dona sola de 18 a 64 anys",1501								
12	2017,3,"Sants-MontjuÃ¯c",11,"el Poble Sec - Parc MontjuÃ¯c","Una dona sola de 18 a 64 anys",1648								
13	2017,3,"Sants-MontjuÃ¯c",12,"la Marina del Prat Vermell - Zona Franca","Una dona sola de 18 a 64 anys",32								
14	2017,3,"Sants-MontjuÃ¯c",13,"la Marina de Port","Una dona sola de 18 a 64 anys",781								
15	2017,3,"Sants-MontjuÃ¯c",14,"la Font de la Guatlla","Una dona sola de 18 a 64 anys",397								
16	2017,3,"Sants-MontjuÃ¯c",15,"Hostafrancs","Una dona sola de 18 a 64 anys",693								
17	2017,3,"Sants-MontjuÃ¯c",16,"la Bordeta","Una dona sola de 18 a 64 anys",585								
18	2017,3,"Sants-MontjuÃ¯c",17,"Sants - Badal","Una dona sola de 18 a 64 anys",901								
19	2017,3,"Sants-MontjuÃ¯c",18,"Sants","Una dona sola de 18 a 64 anys",1808								
20	2017,4,"Les Corts",19,"les Corts","Una dona sola de 18 a 64 anys",1652								
21	2017,4,"Les Corts",20,"la Maternitat i Sant Ramon","Una dona sola de 18 a 64 anys",697								

Figura 14. Canvis en els datasets d'Open Data BCN

Com es pot veure en la **figura 14**, en el dataset de dalt (el primer en descarregar i el utilitzat per definir els indicadors) la primera fila amb els noms dels atributs està mal constituïda, en canvi, en la modificació realitzada pel servei de dades (dataset de sota) està millor estructurada, però no conté el camp del total de llars per barri (imprescindible per poder calcular el percentatge, per tant, es necessita aquesta informació d'un altre dataset per després realitzar els càlculs). S'han trobat datasets amb dades mal estructurades, diferents denominacions dels barris per dataset (un exemple entre d'altres, el barri *el Poble Sec* en alguns datasets es denomina *el Poble Sec – Montjuïc)*, o també tenir unificat els camps de número de barri i nom,

dificultant la realització dels càlculs al no permetre ordenar pel número de barri (per exemple, tenir el valor “1. El Raval”, “2. El Barri Gòtic,etc.).

Es decideix utilitzar el software propietari Microsoft Excel per transformar les dades, en un principi es volien utilitzar els macros d’Excel per automatitzar els càlculs, però la poca experiència amb aquest programari ha fet descartar aquesta opció i realitzar els càlculs d’una manera més manual. Com es veu a la **figura 12** tots els indicadors són percentatges i, malauradament només es troba un indicador que no sigui necessari calcular, la taxa d’atur.

Després d’analitzar detingudament el catàleg de datasets d’Open Data BCN i realitzar una exhaustiva cerca per trobar les dades en altres fonts, es decideix descartar els 5 últims indicadors (Vulnerabilitat Subjectiva) per dues raons, la primera i principal, per que no es troben dades i la segona, que com són valors subjectius (percepcions) es considerant poc fiables. Dels 15 indicadors restants, s’aconsegueixen recollir i transformar 11 indicadors, 10 obtinguts del servei de dades Open Data BCN i un extret manualment de l’Observatori de Barris del Departament d’Estadística de l’Ajuntament de Barcelona **[16]** (l’índex d’atur juvenil, indicador 7). A continuació una relació dels 11 indicadors (es manté el número d’indicador de la **figura 12**) a aplicar i els datasets o fonts utilitzats:

1. **Percentatge de llars unipersonals majors de 64 anys**.
   Dataset: Estructura dels domicilis de la ciutat de Barcelona 2017. Conté 876 registres i 7 atributs, dades completes.
2. **Índex de sobre-envelliment**
   Dataset: Població per barris, edat quinquennal per nacionalitat i sexe de la ciutat de Barcelona 2017. Conté 6132 registres i 10 atributs, dades completes.
3. **Índex de població estrangera en edat infantil**
   Dataset: Població per barris, edat quinquennal per nacionalitat i sexe de la ciutat de Barcelona 2017. Conté 6132 registres i 10 atributs, dades completes.
4. **Índex de població estrangera**
   Dataset: Població per barris, edat quinquennal per nacionalitat i sexe de la ciutat de Barcelona 2017. Conté 6132 registres i 10 atributs, dades completes.
5. **Percentatge de llars monoparentals**

Dataset: Estructura dels domicilis de la ciutat de Barcelona 2017. Conté 876 registres i 7 atributs, dades completes.

6. **Taxa d'atur**

   Dataset: Pes de l'atur registrat sobre la població de 16 a 64 anys per barris de la ciutat de Barcelona 2017. Conté 876 registres i 9 atributs, dades completes.

7. **Taxa d'atur juvenil**

   Dades extretes manualment de l'Observatori de Barris de l'Ajuntament de Barcelona

10. **Taxa de població sense estudis**

    Dataset: Població per barris, nivell acadèmic i sexe de la ciutat de Barcelona 2017. Conté 876 registres i 9 atributs, dades completes.

11. **Percentatge d'habitatges amb una superfície útil menor a 31 m^2**

    Habitatges principals de la ciutat de Barcelona segons superfície útil 2011. Conté 73 registres i 13 atributs, dades completes.

14. **Percentatge d'habitatges situades en edificis en un estat de conservació dolent**

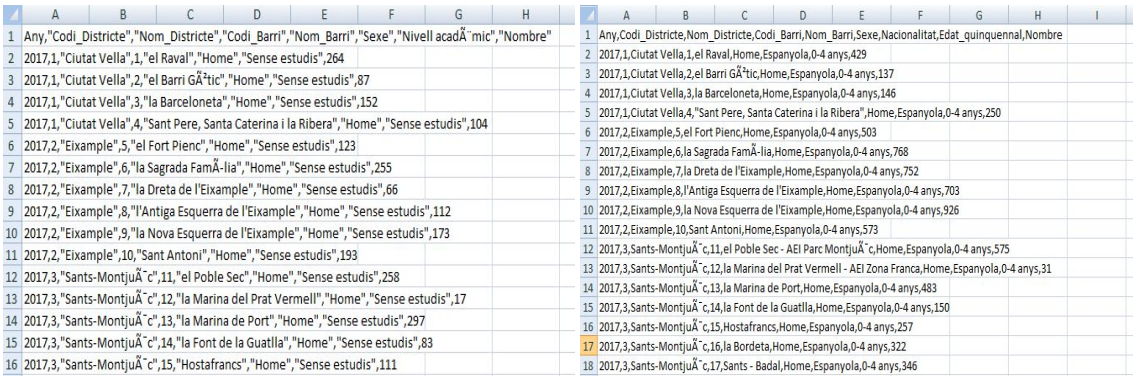
    Dataset: Habitatges principals de la ciutat de Barcelona segons estat de conservació de l'edifici 2011. Conté 74 registres i 9 atributs, dades incompletes.

15. **Percentatge d'habitatges situades en edificis construïts abans de 1951**

    Dataset: Habitatges de la ciutat de Barcelona segons any de construcció 2011. Conté 74 registres i 23 atributs, dades incompletes.

La dades dels cinc indicadors que pertanyen a la Vulnerabilitat Residencial només estan disponibles les del 2011, pertanyen al *Censo de Población y Viviendas del 2011* (***INE**, Instituto Nacional de Estadística*), l'últim realitzat, aquest cens es realitza decennalment. Dels dos indicadors (12 i 13) s'han trobat dades però són insuficients, en un cas només existeixen valors per tres barris (Habitatges principals de la ciutat de Barcelona segons instal·lacions II: aigua corrent, lavabo i bany).

Respecte els cinc indicadors descartats de Vulnerabilitat Subjectiva, s'ha trobat dades en format de fitxes estadístiques amb els 20 indicadors aquí referenciats, pertanyen a l'anàlisi de barris vulnerables **[15]**, però les dades trobades són dels anys 1991, 2001 i 2006, a més, l'àrea delimitada no és exactament per barris si no per àrees estadístiques vulnerables (AEV) i a més no hi són completes en cap any.

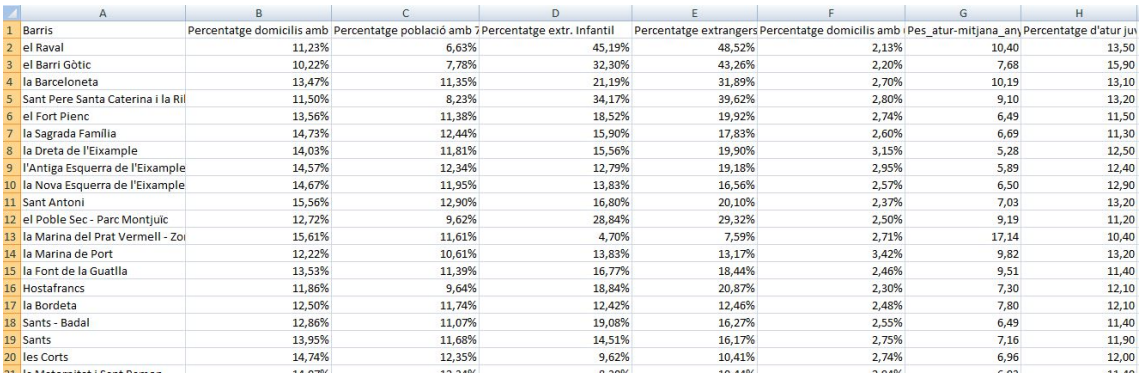


	A	B	C	D	E	F	G	H
1	Any,"Codi_Districte","Nom_Districte","Codi_Barri","Nom_Barri","Sexe","Nivell acadÃ¨mic","Nombre"							
2	2017,1,"Ciutat Vella",1,"el Raval","Home","Sense estudis",264							
3	2017,1,"Ciutat Vella",2,"el Barri GÃ²tic","Home","Sense estudis",87							
4	2017,1,"Ciutat Vella",3,"la Barceloneta","Home","Sense estudis",152							
5	2017,1,"Ciutat Vella",4,"Sant Pere, Santa Caterina i la Ribera","Home","Sense estudis",104							
6	2017,2,"Eixample",5,"el Fort Pienc","Home","Sense estudis",123							
7	2017,2,"Eixample",6,"la Sagrada FamÃlia","Home","Sense estudis",255							
8	2017,2,"Eixample",7,"la Dreta de l'Eixample","Home","Sense estudis",66							
9	2017,2,"Eixample",8,"l'Antiga Esquerra de l'Eixample","Home","Sense estudis",112							
10	2017,2,"Eixample",9,"la Nova Esquerra de l'Eixample","Home","Sense estudis",173							
11	2017,2,"Eixample",10,"Sant Antoni","Home","Sense estudis",193							
12	2017,3,"Sants-MontjuÃ¨c",11,"el Poble Sec","Home","Sense estudis",258							
13	2017,3,"Sants-MontjuÃ¨c",12,"la Marina del Prat Vermell","Home","Sense estudis",17							
14	2017,3,"Sants-MontjuÃ¨c",13,"la Marina de Port","Home","Sense estudis",297							
15	2017,3,"Sants-MontjuÃ¨c",14,"la Font de la Guatlla","Home","Sense estudis",83							
16	2017,3,"Sants-MontjuÃ¨c",15,"Hostafrancs","Home","Sense estudis",111							

	A	B	C	D	E	F	G	H	I
1	Any,Codi_Districte,Nom_Districte,Codi_Barri,Nom_Barri,Sexe,Nacionalitat,Edat_quinquennal,Nombre								
2	2017,1,Ciutat Vella,1,el Raval,Home,Espanyola,0-4 anys,429								
3	2017,1,Ciutat Vella,2,el Barri GÃ²tic,Home,Espanyola,0-4 anys,137								
4	2017,1,Ciutat Vella,3,la Barceloneta,Home,Espanyola,0-4 anys,146								
5	2017,1,Ciutat Vella,4,"Sant Pere, Santa Caterina i la Ribera",Home,Espanyola,0-4 anys,250								
6	2017,2,Eixample,5,el Fort Pienc,Home,Espanyola,0-4 anys,503								
7	2017,2,Eixample,6,la Sagrada FamÃlia,Home,Espanyola,0-4 anys,768								
8	2017,2,Eixample,7,la Dreta de l'Eixample,Home,Espanyola,0-4 anys,752								
9	2017,2,Eixample,8,l'Antiga Esquerra de l'Eixample,Home,Espanyola,0-4 anys,703								
10	2017,2,Eixample,9,la Nova Esquerra de l'Eixample,Home,Espanyola,0-4 anys,926								
11	2017,2,Eixample,10,Sant Antoni,Home,Espanyola,0-4 anys,573								
12	2017,3,Sants-MontjuÃ¨c,11,el Poble Sec - AEI Parc MontjuÃ¨c,Home,Espanyola,0-4 anys,575								
13	2017,3,Sants-MontjuÃ¨c,12,la Marina del Prat Vermell - AEI Zona Franca,Home,Espanyola,0-4 anys,31								
14	2017,3,Sants-MontjuÃ¨c,13,la Marina de Port,Home,Espanyola,0-4 anys,483								
15	2017,3,Sants-MontjuÃ¨c,14,la Font de la Guatlla,Home,Espanyola,0-4 anys,150								
16	2017,3,Sants-MontjuÃ¨c,15,Hostafrancs,Home,Espanyola,0-4 anys,257								
17	2017,3,Sants-MontjuÃ¨c,16,la Bordeta,Home,Espanyola,0-4 anys,322								
18	2017,3,Sants-MontjuÃ¨c,17,Sants - Badal,Home,Espanyola,0-4 anys,346								

Figura 15. Exemples de datasets utilitzats d'Open Data BCN

Definides les dades a utilitzar i transformades, es recopilen en un arxiu d'Excel.



	A	B	C	D	E	F	G	H
1	Barris	Percentatge domicilis amb	Percentatge població amb 7	Percentatge extr. Infantil	Percentatge extrangers	Percentatge domicilis amb	Pes_atur-mitjana_any	Percentatge d'atur juv
2	el Raval	11,23%	6,63%	45,19%	48,52%	2,13%	10,40	13,50
3	el Barri Gòtic	10,22%	7,78%	32,30%	43,26%	2,20%	7,68	15,90
4	la Barceloneta	13,47%	11,35%	21,19%	31,89%	2,70%	10,19	13,10
5	Sant Pere Santa Caterina i la Rib	11,50%	8,23%	34,17%	39,62%	2,80%	9,10	13,20
6	el Fort Pienc	13,56%	11,38%	18,52%	19,92%	2,74%	6,49	11,50
7	la Sagrada Família	14,73%	12,44%	15,90%	17,83%	2,60%	6,69	11,30
8	la Dreta de l'Eixample	14,03%	11,81%	15,56%	19,90%	3,15%	5,28	12,50
9	l'Antiga Esquerra de l'Eixample	14,57%	12,34%	12,79%	19,18%	2,95%	5,89	12,40
10	la Nova Esquerra de l'Eixample	14,67%	11,95%	13,83%	16,56%	2,57%	6,50	12,90
11	Sant Antoni	15,56%	12,90%	16,80%	20,10%	2,37%	7,03	13,20
12	el Poble Sec - Parc Montjuïc	12,72%	9,62%	28,84%	29,32%	2,50%	9,19	11,20
13	la Marina del Prat Vermell - Zo	15,61%	11,61%	4,70%	7,59%	2,71%	17,14	10,40
14	la Marina de Port	12,22%	10,61%	13,83%	13,17%	3,42%	9,82	13,20
15	la Font de la Guatlla	13,53%	11,39%	16,77%	18,44%	2,46%	9,51	11,40
16	Hostafrancs	11,86%	9,64%	18,84%	20,87%	2,30%	7,30	12,10
17	la Bordeta	12,50%	11,74%	12,42%	12,46%	2,48%	7,80	12,10
18	Sants - Badal	12,86%	11,07%	19,08%	16,27%	2,55%	6,49	11,40
19	Sants	13,95%	11,68%	14,51%	16,17%	2,75%	7,16	11,90
20	les Corts	14,74%	12,35%	9,62%	10,41%	2,74%	6,96	12,00

Figura 16. Arxiu final amb les dades seleccionades (no supervisat)

4.1.3 Preprocessament i neteja de dades

A causa del procediment per obtenir el conjunt de dades, no és necessari implementar aquesta fase, enlloc d'obtenir les dades en brut i netejar-les (tractament de les dades incompletes i incorrectes, la redundància de dades, els sorolls o els valors atípics), en aquest cas les dades s'han construït al inrevés, cercant atribut a atribut i afegint-les al conjunt ja netes.

Però abans d'aplicar la transformació de les dades amb Weka a la següent fase, és necessari convertir l'arxiu resultant de la fase anterior (aquest arxiu s'ha anomenat *dataset_no_supervisat_exclusio_social.xlsx*) al format nadiu de Weka, *arff* (*Attribute-Relation File Format*). Els arxius *arff* tenen el següent format i estructura: tenen format de text pla en ASCII, poden ser visualitzats i modificats amb qualsevol editor de text, han de seguir les següents pautes:

- Les línies de comentari han de començar amb el símbol %
- Les línies de declaració de relació, atributs i dades han de començar amb el símbol @
- El text que inclogui espais ha d'anar entre cometes

S'estructuren en dues parts: capçalera i dades, la capçalera inclou el nom de la relació i la declaració dels atributs:

Capçalera

Inclou el nom de la relació de tipus String:

@relation <nom_de_la_relació>

I la declaració d'atributs:

@attribute <nom_de_l'atribut> <tipus_de_dades>

Dades

La secció de dades comença amb:

@data

Seguit de les dades, cada línea és una instància separant els atributs per comes, a més, els atributs han d'anar en el mateix ordre que es declaren a la capçalera.

A continuació un exemple:

```
@relation weather
@attribute outlook { sunny, overcast, rainy }
@attribute temperature numeric
@attribute humidity numeric
@attribute windy { true, false }
@attribute play { yes, no }
@data
%
% 14 instances
%
sunny, 85, 85, false, no
sunny, 80, 90, true, no
overcast, 83, 86, false, yes
rainy, 70, 96, false, yes
rainy, 68, 80, false, yes
rainy, 65, 70, true, no
```

```
overcast, 64, 65, true, yes
sunny, 72, 95, false, no
sunny, 69, 70, false, yes
```

Per construir l'arxiu arff en primer lloc es converteix l'arxiu Excel a csv, a continuació, s'obre l'arxiu csv amb un editor de text, en aquest cas s'utilitza Sublime Text 3, es modifica l'estructura amb les pautes abans definides, s'eliminen els símbols de percentatge, les comes per punts pels nombres reals, el punt i coma per comes pels separadors, etc.

Figura 17. Arxiu final arff generat amb Sublime Text 3

Finalment, l'arxiu generat i preparat per utilitzar a Weka s'anomena *dataset_no_supervisat_exclusio_social.arff*

4.1.4 Transformació de les dades

En aquesta fase es continua millorant la qualitat de les dades, transformant les dades per adaptar-les a les tècniques que més tard s'aplicaran. La transformació de dades es pot dividir en dos passos, en el primer pas es tracten els atributs, ja sigui per reduir atributs (reducció de la dimensionalitat), augmentar-los (augment de la dimensionalitat) o crear-ne de nous (descobriment de característiques).

En el segon pas, es tracten les dades transformant els seus valors, ja sigui canviant el tipus de dada o la distribució que segueix. Entre aquestes

operacions es poden trobar la conversió de tipus numèric a nominal, de nominal a binari, la discretització de valors, la normalització, etc.

Per la realització d'aquest fase i les tres següents (fases de *Data Mining*) s'utilitza el programari lliure ***Weka***. En concret, es selecciona l'aplicació *Explorer* de la interfície d'inici (En les últimes versions de Weka està disponible l'opció *Workbench*, que permet utilitzar totes les aplicacions de la interfície d'inici en una sola, però per aquest treball només s'utilitza l'*Explorer*). A la primera pestanya (Preprocess) es selecciona des de que font es vol obrir el dataset a utilitzar i es trobant els filtres per transformar les dades.

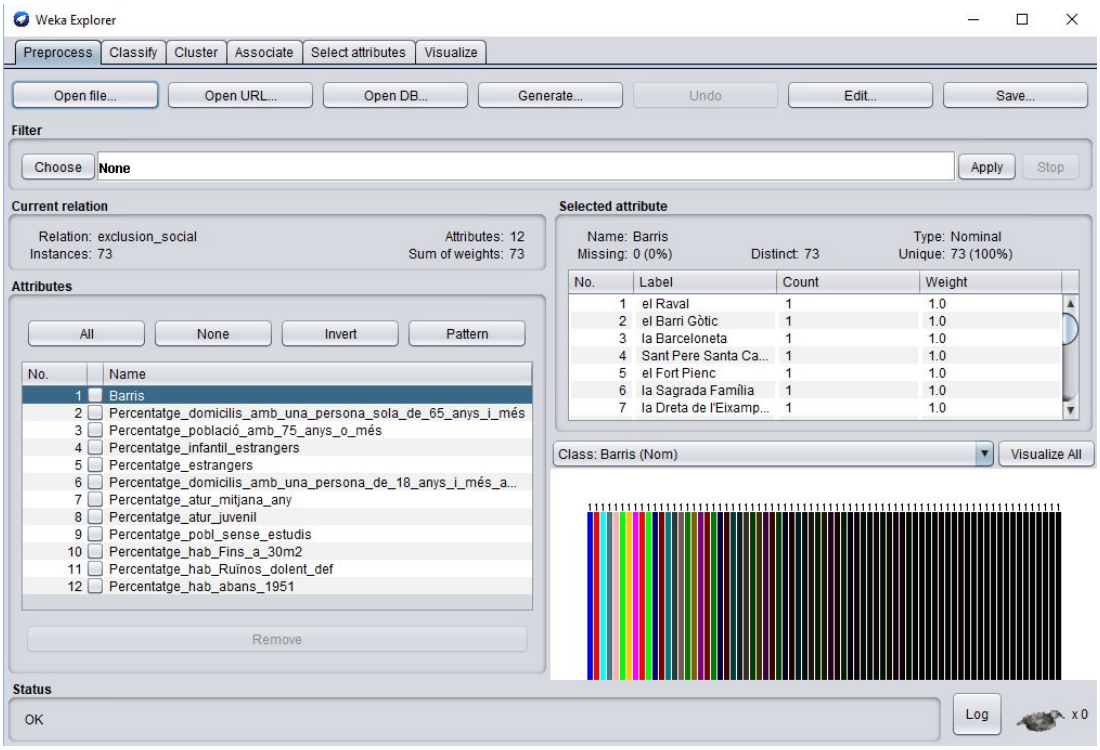


Figura 18. Pestanya Preprocess de l'Explorer de Weka

Es realitza un primer anàlisi visual de tots els atributs (*Visualize All*), on es pot veure la distribució de cadascun i els rangs en què es mouen.

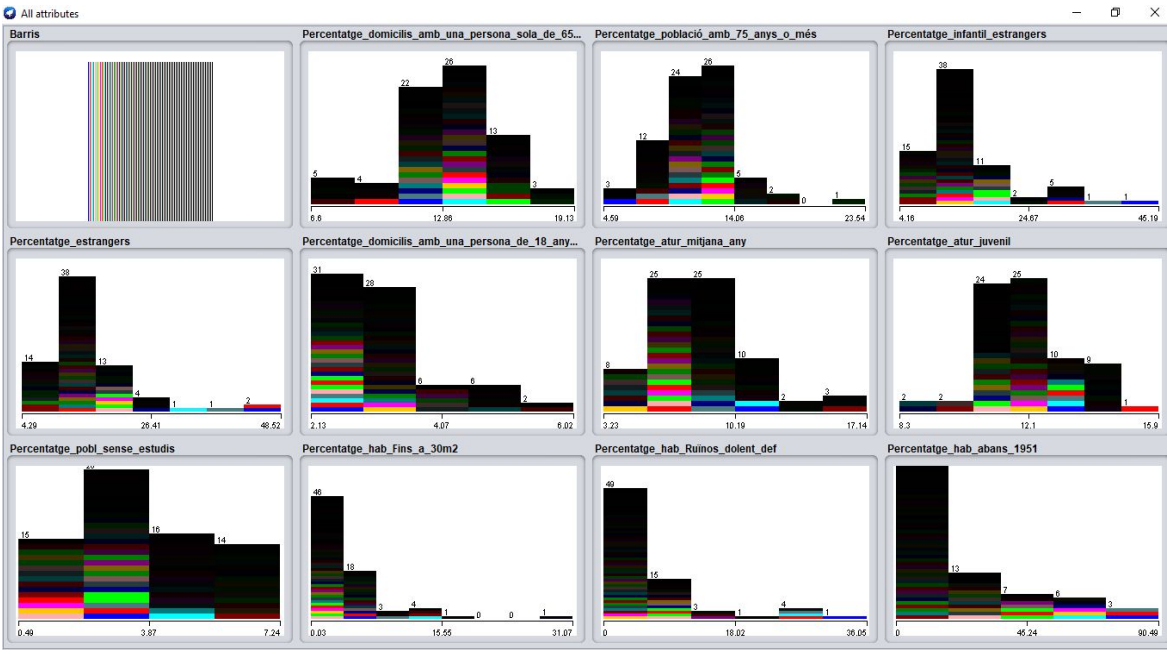


Figura 19. Resum de tots els atributs amb Weka

Depenent del tipus de dades i de les tècniques i algorismes a aplicar, s'aplica una operació, cap operació o més d'una operació (una en cada pas dels dos passos comentats). En el cas del dataset a tractar (dataset_no_supervisat_exclusio_social.arff) es suposa que tots els atributs són independents i necessaris, per tant, no es realitza cap operació referent a la dimensionalitat. Respecte la transformació de valors i tipus dels atributs, donat que tots els atributs tenen la mateixa escala (percentatges) i són numèrics, no és necessari tractar les dades (la transformació de valors pot dificultar la interpretació dels grups), en tot cas es realitzen proves normalitzant tots els atributs, per comprovar si la diferència de rang en alguns atributs influeix.

4.1.5 Escollir la tasca de Mineria de Dades

En aquesta fase s'escull el tipus de tasca que es vol desenvolupar, ja sigui tasca de classificació, regressió, clustering o associació. Veritablement, aquesta decisió es pren al inici del procés on es defineixen els objectius i es comprèn el domini de l'aplicació, de fet, totes les fases anteriors a aquesta van encaminades depenent de la tasca i dels algorismes a aplicar. No tindria gaire sentit transformar les dades en la fase anterior, per exemple, canviar el tipus d'un atribut de numèric a nominal, si després s'aplica un algorisme que només

accepta dades numèriques, o tenir dades no supervisades i voler aplicar la predicció.

En aquest cas la tasca a desenvolupar és el clustering, per tant, al programari Weka es selecciona la pestanya *Cluster.*

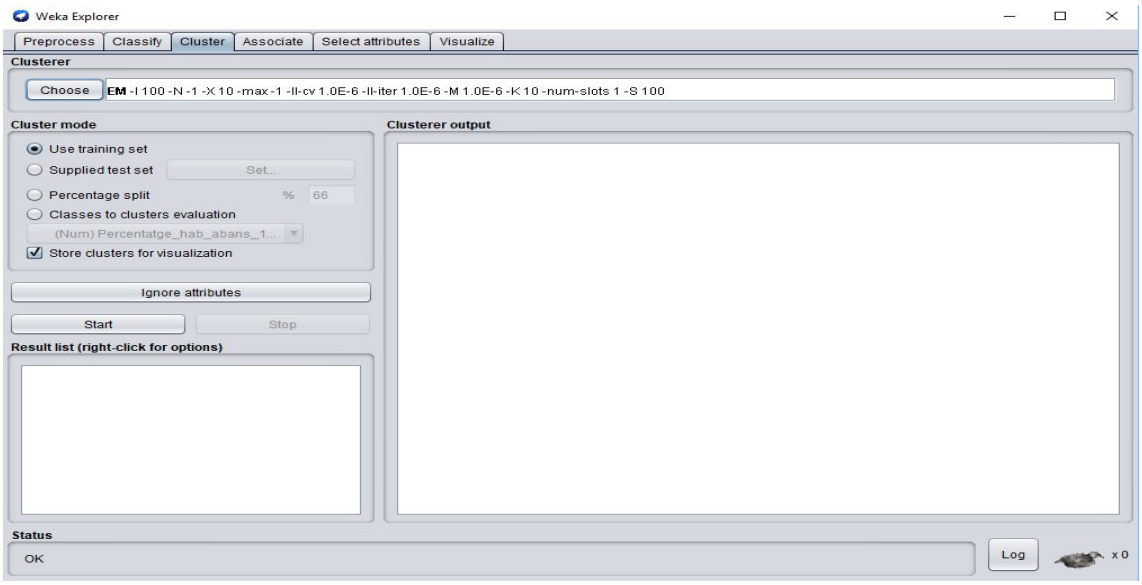


Figura 20. Interfície de clustering de Weka

4.1.6 Escollir l'algorisme

En aquesta fase s'ha d'escollir algorisme a aplicar a les dades, pel cas d'estudi del treball s'escullen dos algorismes, detallats al **capítol 3**:

- ✓ K-means
- ✓ EM

L'elecció d'aquests algorismes és principalment a causa del tipus de dades a tractar (tipus numèric) i el reduït nombre d'instàncies, descartant algorismes com el Cobweb, gens idoni per dades numèriques, o el Make Density Based Clusterer, més adient per dades geoespacials. Un punt apart mereix l'algorisme Hierarchical Clusterer, algorisme jeràrquic adequat per petits datasets del qual es realitzen proves en el següent apartat, donats els resultats i per raons d'extensió del treball és decideix focalitzar en els dos algorismes escollits (K-means i EM).

4.1.7 Aplicar l'algorisme

Es decideix aplicar en primer lloc l'algorisme ***EM*** al dataset (dataset_no_supervisat_exclusio_social.arff), degut al seu funcionament. L'algorisme *EM* comença utilitzant la validació creuada amb 10 conjunts per determinar quin és el nombre òptim de clústers (dada que s'utilitzarà per fixar el nombre de clústers amb *K-means*), per després executar l'algorisme sobre el conjunt d'entrenament iterant fins que convergeix al màxim o s'arriba al nombre d'iteracions fixat als paràmetres. Aquest màxim pot ser que no sigui el màxim global, és necessari provar amb diferents valors del paràmetre que determina els conjunts inicials. Per tant, cal modificar el paràmetre seed (llavor) per trobar el valor màxim de la probabilitat de registre (*log likelihood*).

A la pestanya *Cluster*, es selecciona el tipus de clúster i com s'avalua aquest (*Cluster mode*), per les característiques del dataset, pocs registres però tots importants (cada registre és un barri), es selecciona l'opció Use training set per utilitzar el dataset complet per entrenar el model. S'escull l'algorisme EM mantenint tots els paràmetres per defecte (el nombre de clústers -1 per trobar el nombre òptim de clústers) excepte la llavor (seed) , amb el qual es realitzaren proves amb diferents valors per trobar el millor model. Es decideix que el millor model obtingut té com a valor de llavor 100 (per defecte) i s'obtenen 3 clústers. A continuació una taula amb diferents valors de *seed* i els valors obtinguts de *log likelihood*:

seed	**1**	**20**	**50**	**100**
Log likelihood	**-34,26296**	**-34,26296**	**-31,22027**	**-31,22027**

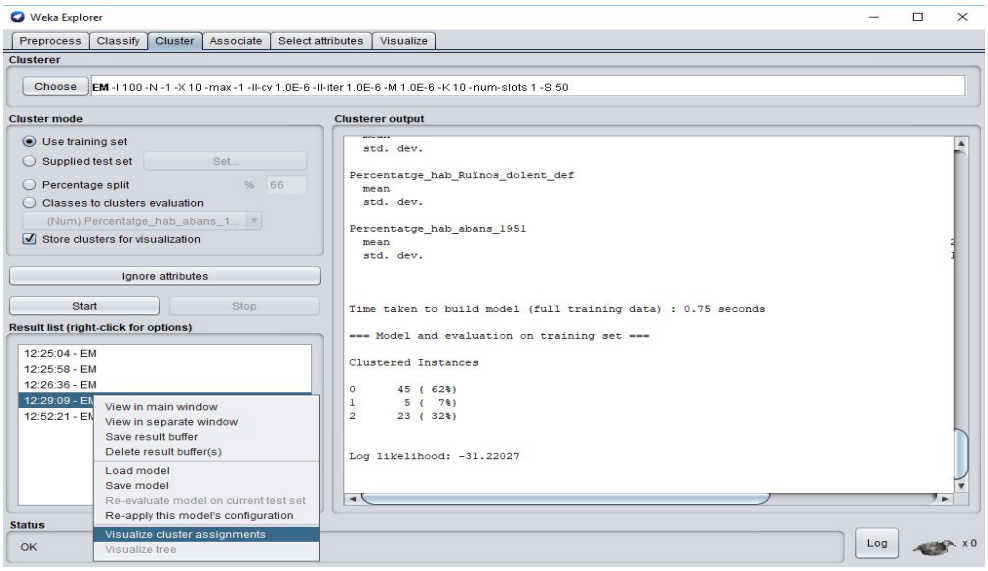


Figura 21. Resultats de l'algorisme EM

Ara es comprova l'agrupació per clúster visualitzant els resultats, com es pot veure a la **figura 21** (*Visualize cluster assignments*) amb el botó dret del ratolí sobre l'aplicació de l'algorisme.

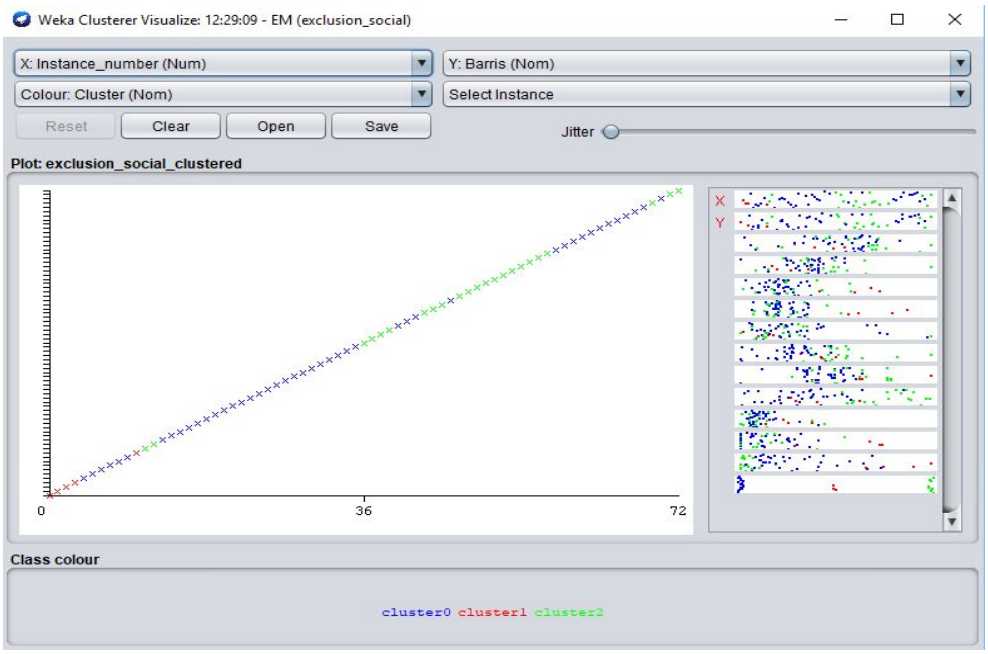


Figura 22. Visualització dels clústers (EM)

A continuació, s'aplica l'algorisme ***K-means*** amb el nombre de clústers obtinguts amb l'algorisme *EM* (3 clústers), realitzant proves amb el valor de la llavor (*seed*), per minimitzar l'error quadrat (*Within cluster sum of squared errors*), el qual és la suma de les distàncies euclidianes (o qualsevol altra distància a utilitzar) entre cada instància i el seu centre de clúster, i deixant la

resta de paràmetres per defecte. El model resultant amb 3 clústers i el mínim valor de "*Within cluster sum of squared errors"*, s'obté amb valor 20 per la llavor (seed). A continuació una taula amb diferents valors de *seed* i els valors obtinguts de *Within cluster sum of squared errors*:

seed	**1**	**20**	**50**	**100**
Within cluster sum of squared errors	**91,51025**	**89,12712**	**92,55789**	**89,91480**

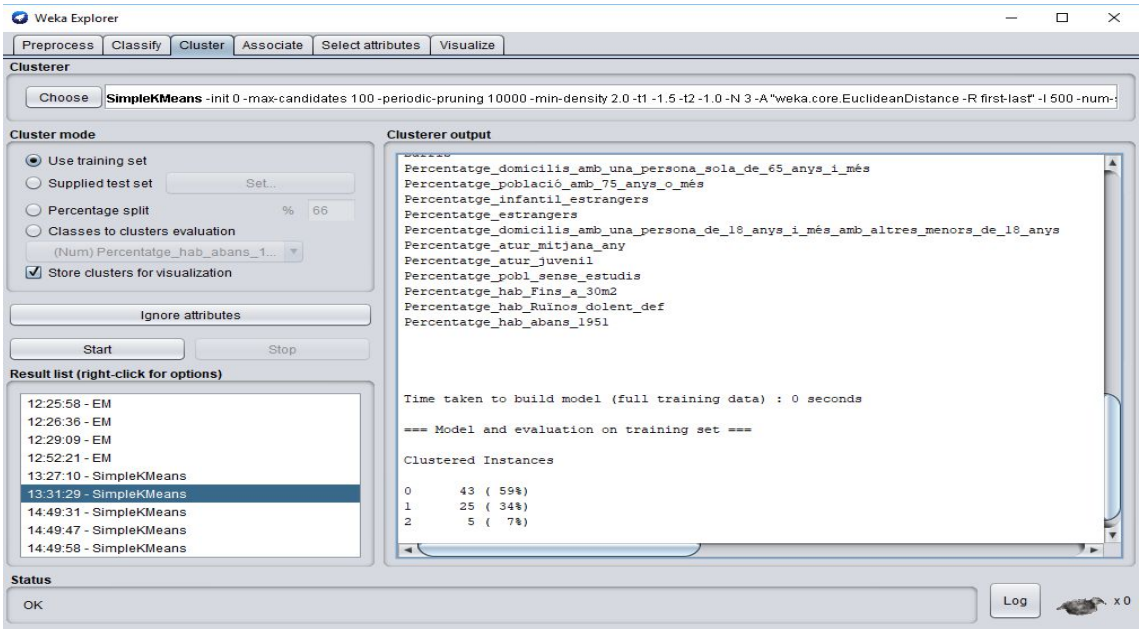


Figura 23. Resultats de l'algorisme K-means

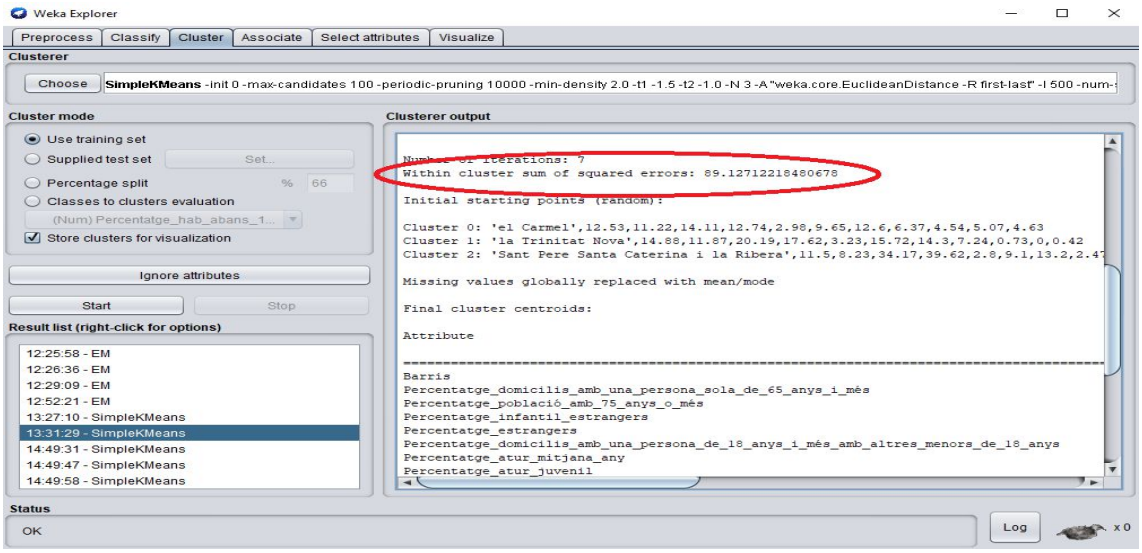


Figura 24. Valor mínim trobat de Within cluster sum of squared errors

Ara es comprova l'agrupació per clúster visualitzant els resultats, com es pot veure a la **figura 25**. S'han definit els mateixos colors per clúster que amb l'algorisme EM, per facilitar la comparació.

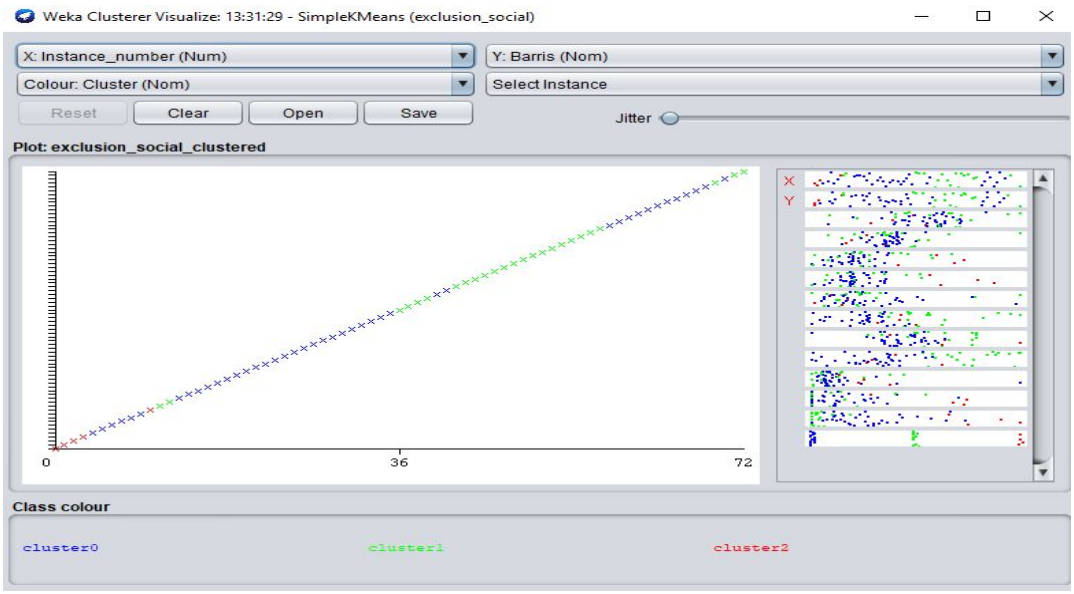


Figura 25. Visualització dels clústers (K-means)

4.1.8 Avaluar els resultats obtinguts

En aquesta fase s'analitzen i interpreten els models obtinguts, en primer lloc es comproven els centroides per atribut de cada clúster i de totes les dades de l'aplicació del K-means.

Final cluster centroids:

Attribute	Full Data	Cluster# 0	1	2
	(73.0)	(43.0)	(25.0)	(5.0)
Barris	el Raval	el Fort Pienc	la Marina del Prat Vermell	el Raval
Percentatge_domicilis_amb_una_persona_sola_de_65_anys_i_més	13.1607	13	13.7036	11.828
Percentatge_població_amb_75_anys_o_més	11.5018	11.1823	12.6072	8.722
Percentatge_infantil_estrangers	14.7697	12.5798	15.0228	32.338
Percentatge_estrangers	15.7314	14.1149	13.9536	38.522
Percentatge_domicilis_amb_una_persona_de_18_anys_i_més_amb_altres_menors_de_18_anys	3.1918	3.2749	3.194	2.466
Percentatge_atur_mitjana_any	8.4505	6.7898	11.1348	9.312
Percentatge_atur_juvenil	12.1096	11.514	12.88	13.38
Percentatge_pobl_sense_estudis	3.643	2.4856	5.7548	3.038
Percentatge_hab_fins_a_30m2	4.3026	3.9849	4.1768	7.664
Percentatge_hab_Ruïnos_dolent_def	5.6979	5.0116	3.0412	24.884
Percentatge_hab_abans_1951	21.1053	23.3919	6.1016	76.46

Figura 26. Resultats dels clústers del K-means

S'ha decidit analitzar els resultats del K-means per que pràcticament són idèntics als resultats del EM, només varien un parell de registres. Com es pot comprovar amb els centroides de cada clúster, el clúster 0 i més nombrós (43) està molt a prop dels centroides (mitjana) de les dades completes, per altra banda, el clúster 1 (25 registres) concentra els barris amb més atur i població sense estudis, i molts menys habitatges construïts abans de 1951. En l'últim clúster (2) només agrupa 5 barris però concentra els barris amb més immigració (també la juvenil), més habitatges en males condicions i construïts abans del 1951, aquest 4 indicadors estan molt per sobre de la mitjana, a més, els 5 barris d'aquest clúster corresponen als 4 del districte de Ciutat Vella i al barri de Poble Sec.

A la següent figura es pot veure com al clúster 1 es concentren els barris amb més atur.

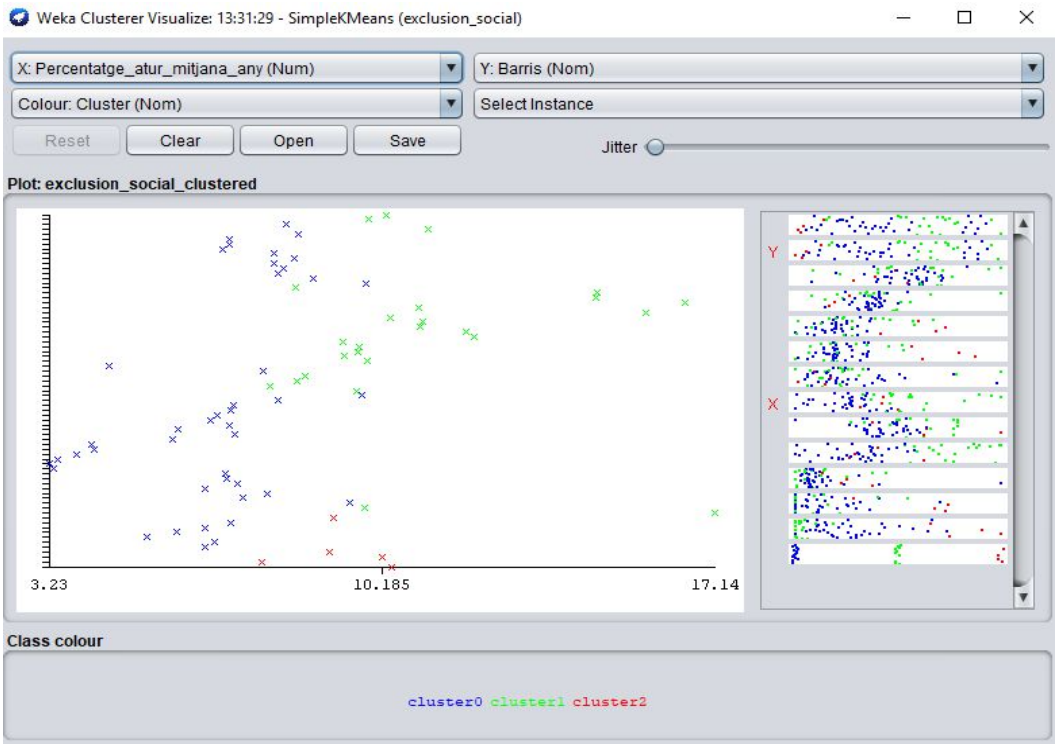


Figura 27. Visualització de l'atur per barris

A la següent figura es pot veure com al clúster 1 es concentren els barris amb més població sense estudis.

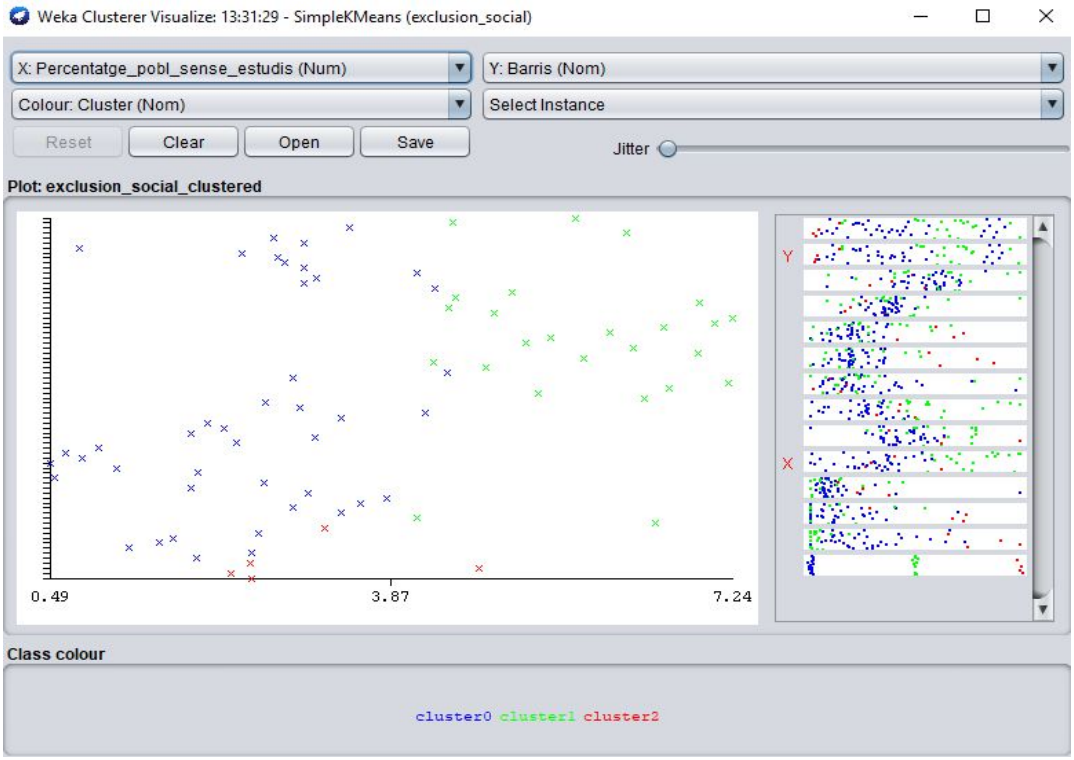


Figura 28. Visualització de la població sense estudis per barris

A la següent figura es pot veure com al clúster 1 es concentren els barris amb menys habitatges construïts abans del 1951, i en canvi, al clúster 2 els barris amb més.

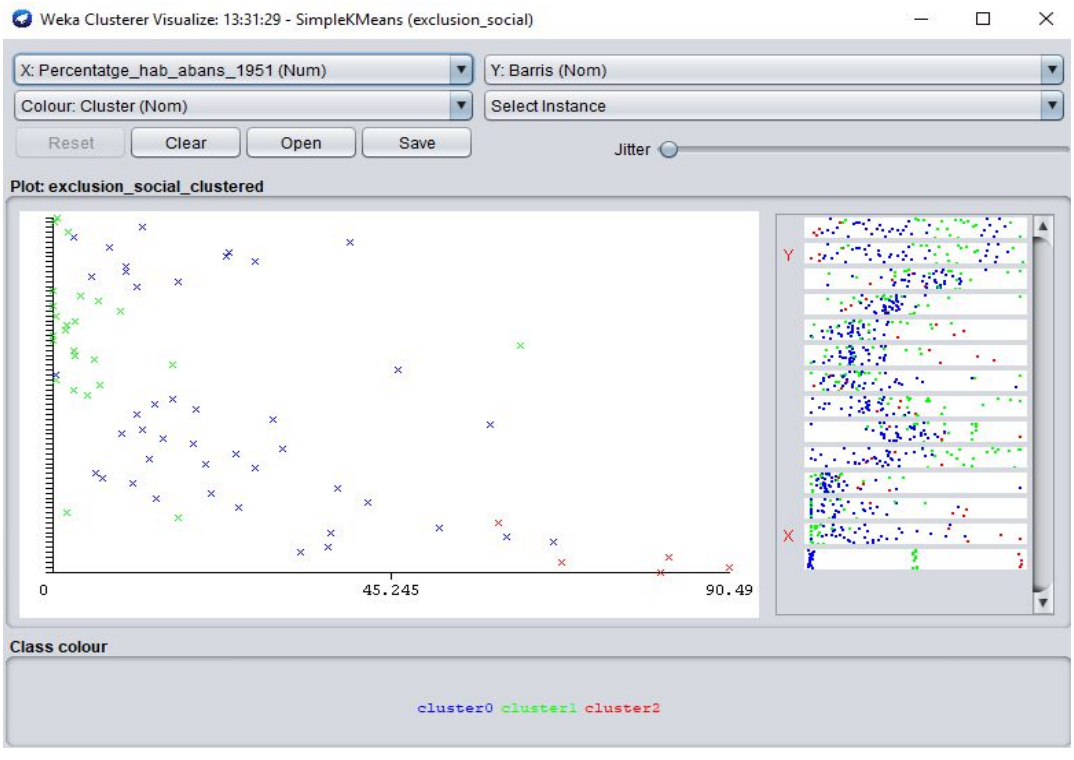


Figura 29. Visualització dels habitatges construïts abans del 1951

A la següent figura es pot veure com al clúster 2 es concentren els barris amb més immigració i habitatges en edificis en mal estat de conservació.

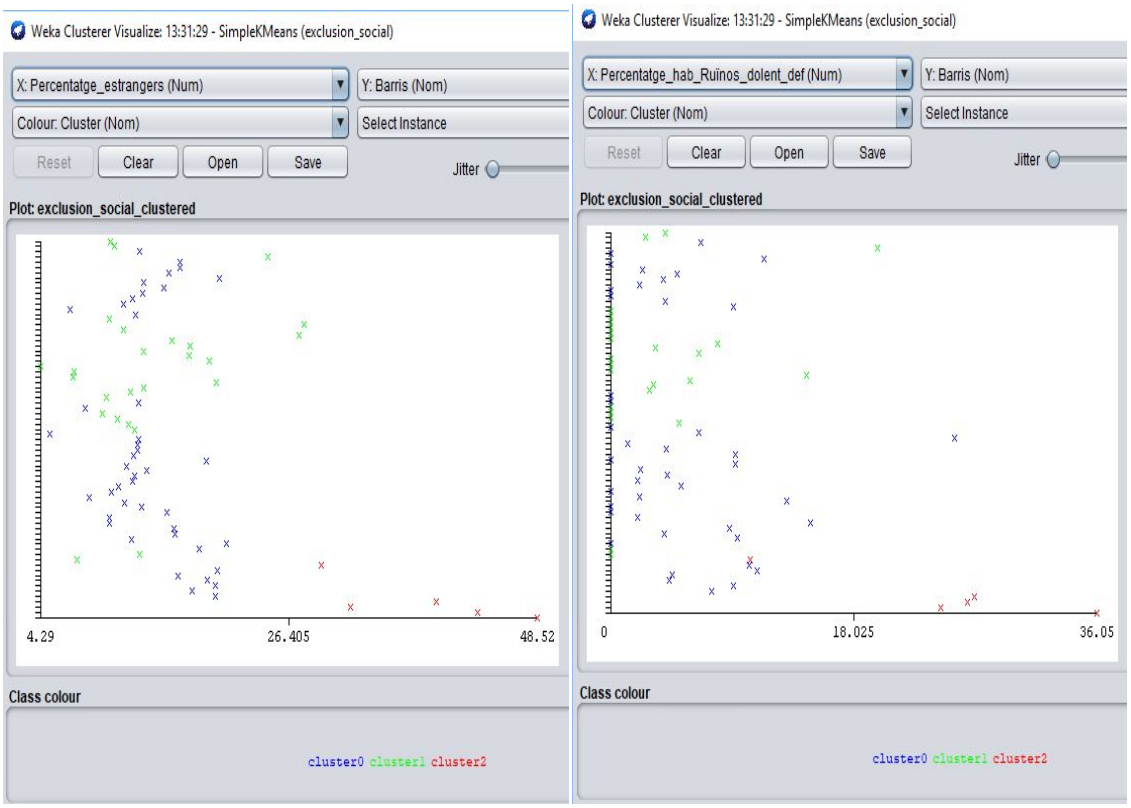


Figura 30. Visualització d'estrangers i edificis en mal estat de conservació

Es pot deduir quins atributs (indicadors) són els que tenen més incidència alhora de formar els clústers:

- Percentatge de població estrangera en edat juvenil
- Percentatge de població estrangera
- Percentatge d'atur
- Percentatge de població sense estudis
- Percentatge d'habitatges en edificis en mal estat de conservació
- Percentatge d'habitatges construïts abans de 1951

No obstant això, analitzant barri per barri les dades dels clústers, es troben barris als clústers amb valors alts d'aquests atributs i que no formen part del clúster abans definit, per exemple, a l'atribut *Percentatge d'habitatges construïts abans de 1951*, la qual cosa faria pensar que hauria de pertànyer al

clúster 2, però en canvi no pertanyen a aquest clúster ja que no coincideixen amb valors alts d'immigració i habitatges en edificis en mal estat de conservació.

Donat que per valors alts d'atributs o indicadors es considera que són més vulnerables a l'exclusió social, es possible definir els clústers formats de menys a més risc, és a dir, el Clúster 0 és el conjunt de barris amb menys risc, el Clúster 1 amb un risc entremig i el Clúster 2 el conjunt de barris amb més risc.

Al Clúster 0 es troben els barris dels districtes de l'Eixample, Les Corts, Sarrià-Sant Gervasi, Gràcia, i part dels districtes de Sants-Montjuïc, Horta-Guinardó, Sant Andreu i Sant Martí. Per la seva part, al Clúster 1 es troben la resta de barris dels districtes de Sants-Montjuïc, Horta-Guinardó, Sant Andreu i Sant Martí i el districte complet de Nou Barris. I com ja s'ha comentat, al Clúster 2 es troben els barris del districte de Ciutat Vella i el Poble Sec, del barri de Sants_Montjuïc.

Atès que l'objectiu és obtenir coneixement per lluitar contra les desigualtats entre barris, és necessari focalitzar les polítiques i accions en reduir aquests sis indicadors acabats de declarar o com a mínim mitigar els seus efectes, si no és possible la seva reducció.

4.1.9 Utilitzar el coneixement descobert

Aquesta fase queda fora de l'àmbit del projecte, en tot cas, es fa difusió dels resultats del treball en el repositori públic de la UOC.

4.2 Procés de KDD amb tècniques de classificació

En aquest apartat es desenvoluparà el cas d'estudi aplicant-li tècniques de classificació, les fases inicials no difereixen gaire del cas d'estudi desenvolupat a l'apartat anterior.

4.2.1 Comprensió del domini de l'aplicació

Aquesta fase com s'ha pogut veure està detallada al principi d'aquest capítol (4), la raó és per que considerant la seva rellevància i donat que es realitzen dos casos d'estudi, es considera que és el lloc més adient, evitant repetir-ho en els dos casos d'estudi. Per tant, després d'estudiar i analitzar la problemàtica de l'*exclusió social*, la controvèrsia amb els indicadors i les dificultats de recollida de dades, l'objectiu és trobar un model que sigui un bon classificador per detectar al més aviat possible les vulnerabilitats i riscos d'*exclusió social* per barris de Barcelona.

4.2.2 Seleccionar i crear el conjunt de dades

La implementació d'aquesta fase és molt semblant a la implementada a l'**apartat 4.1.2**, amb la diferència que el procés que ara s'enceta utilitza tècniques de *classificació* i l'objectiu és ben diferent.

Es parteix del conjunt seleccionat i creat al procés de KDD amb tècniques de *clustering*, però és necessari tenir un atribut de classe que classifiqui els barris per risc d'*exclusió social*. No es disposa d'aquesta dada al servei de dades obertes de l'Ajuntament de Barcelona, *Open Data BCN*, per tant, es cerca entre les estudis i iniciatives relacionades amb aquest àmbit d'actuació.

Per obtenir aquesta informació, s'han definit els barris de que són objecte en *El pla de barris de Barcelona* **[17],** com a barris en risc d'exclusió social, a més dels barris resultants de *l'Estudi i detecció a la ciutat de Barcelona d'àmbits de vulnerabilitat residencial* **[18]**.

Amb aquesta informació, es genera un nou atribut de classe binari (Risc d'exclusió social) que classifica els barris entre els que tenen risc d'exclusió social i els que no.

Dels 73 barris de Barcelona es classifiquen 22 amb risc d'exclusió social:

- ✓ El Raval
- ✓ El Barri Gòtic
- ✓ El poble Sec
- ✓ La Marina del Prat Vermell
- ✓ La Marina de Port
- ✓ Hostafrancs
- ✓ Vallvidrera, el Tibidabo i les Planes
- ✓ El Carmel
- ✓ La Teixonera
- ✓ Sant Genís dels Agudells
- ✓ La Clota
- ✓ Can Peguera
- ✓ Les Roquetes
- ✓ La Trinitat Nova
- ✓ Torre Baró
- ✓ Ciutat Meridiana
- ✓ Vallbona
- ✓ La Trinitat Vella
- ✓ Baró de Viver
- ✓ El Bon Pastor
- ✓ El Besòs i el Maresme
- ✓ La Verneda i la Pau

Per tant, es crea un conjunt de dades amb 73 registres i 13 atributs, 11 indicadors, 1 atribut amb els noms dels barris i 1 atribut de classe.

4.2.3 Preprocessament i neteja de dades

La implementació d'aquesta fase és molt semblant a la implementada a l'**apartat 4.1.3**, amb la diferència que es parteix de l'arxiu *arff* generat en l'apartat esmentat (*dataset_no_supervisat_exclusio_social.arff*), afegint-hi l'atribut de classe *Risc d'exclusió social*.

Per tant, s'afegeix aquest atribut amb el programari Sublime Text 3.

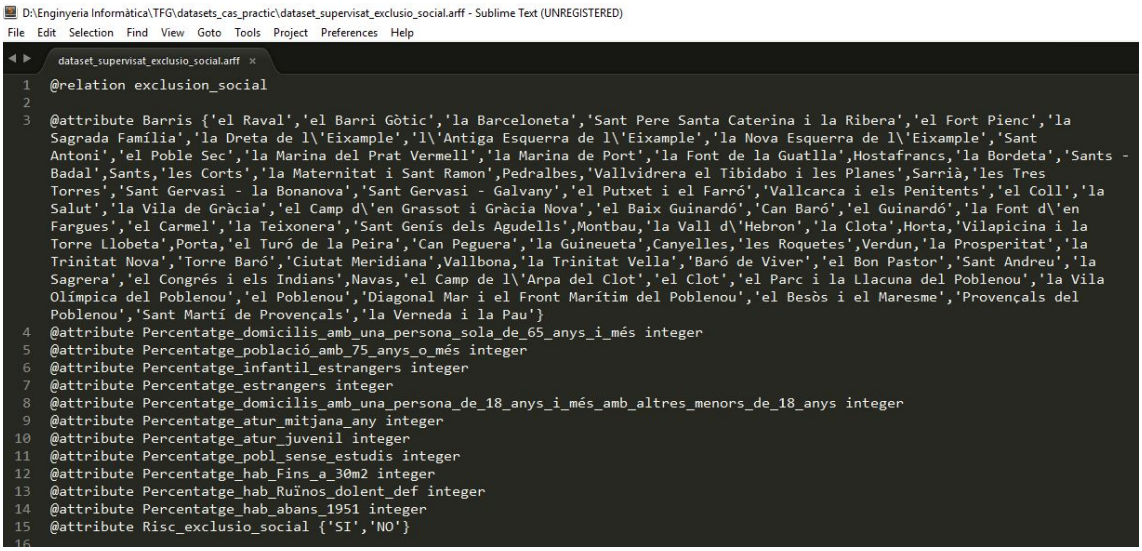


Figura 31. Arxiu final arff generat amb Sublime Text 3

Finalment, l'arxiu generat i preparat per utilitzar a Weka s'anomena *dataset_supervisat_exclusio_social.arff*

4.2.4 Transformació de les dades

En aquesta fase es tenen les mateixes consideracions i decisions que a la mateixa fase del procés de KDD amb tècniques de *clustering* (4.1.4). Per tant, no es realitza cap transformació de les dades respecte els atributs numèrics. En canvi, l'atribut de classe no està balancejat (22-51), la qual cosa pot produir que classifiqui millor la classe majoritària, per tant, es decideix utilitzar el filtre SMOTE per balancejar l'atribut de classe. Entre els paràmetres del filtre, és realitzen proves i es decideix canviar el paràmetre per defecte de veïns més propers (*nearest neighbors*), fixant el valor a 3, duplicant la classe minoritària.

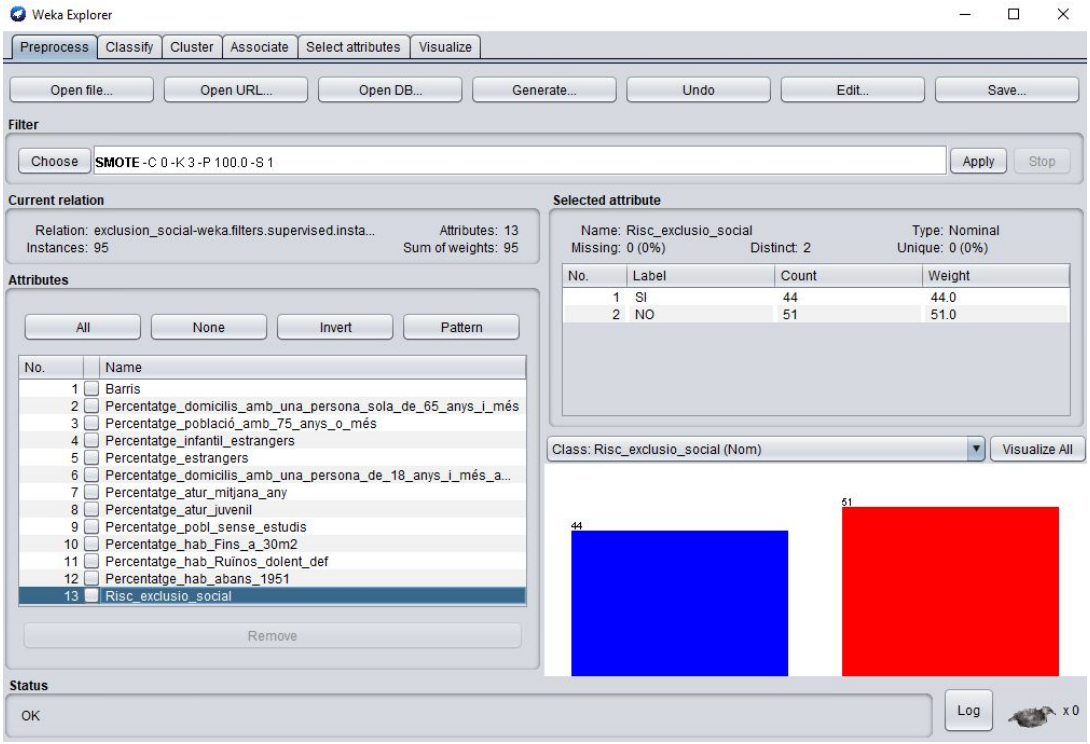


Figura 32. Filtre SMOTE balancejant l'atribut de classe

Es realitza un primer anàlisi visual de tots els atributs (*Visualize All*), on es pot veure la distribució de cadascun per l'atribut de classe.

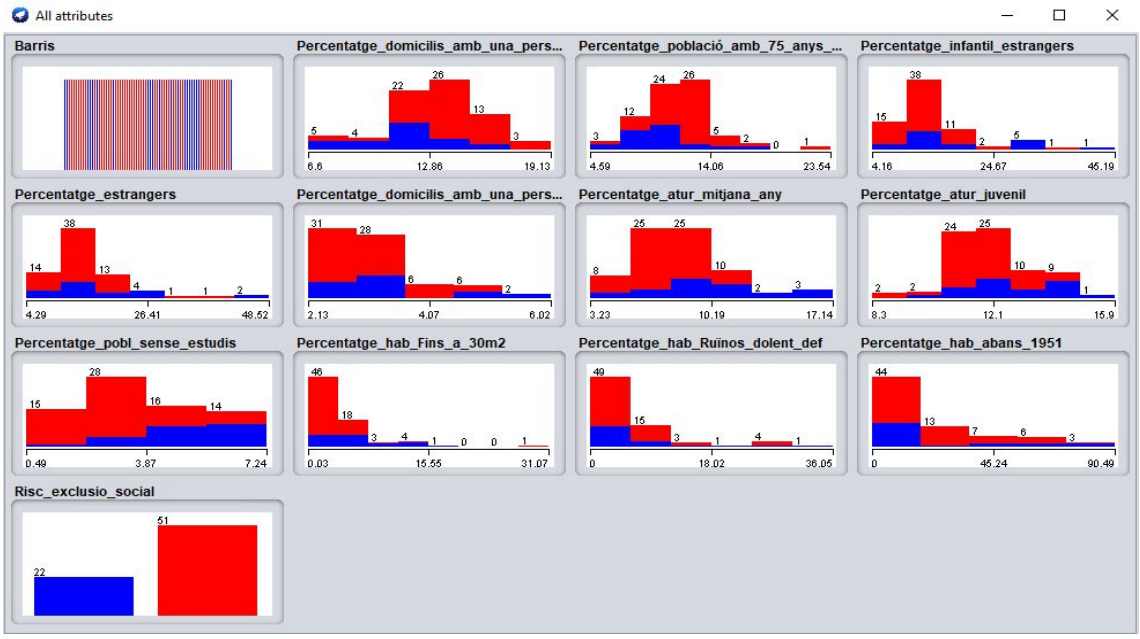


Figura 33. Resum de tots els atributs amb Weka

Com es pot veure, existeixen atributs (indicadors) que tenen a la part alta del rang, barris amb la classe positiva per risc d'exclusió social (zona blava).

4.2.5 Escollir la tasca de Mineria de Dades

Com s'ha comentat a **l'apartat 4.1.5**, en aquesta fase s'escull el tipus de tasca que es vol desenvolupar, ja sigui tasca de classificació, regressió, clustering o associació. En aquest cas la tasca a desenvolupar és la *classificació*, per tant, al programari Weka es selecciona la pestanya *Classify.*

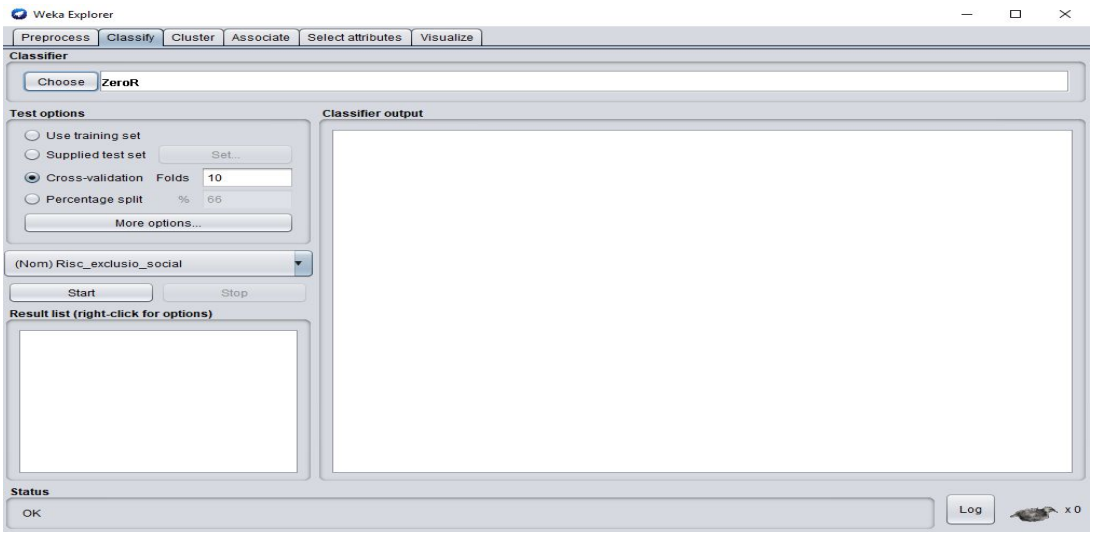


Figura 34. Interfície de classificació en Weka

4.2.6 Escollir l'algorisme

En aquesta fase s'ha d'escollir algorisme a aplicar a les dades, pel cas d'estudi del treball s'escullen sis algorismes, detallats al **capítol 3**:

- ✓ C4.5 (J48)
- ✓ Random Forest
- ✓ LMT
- ✓ Veí més proper (ibK)
- ✓ Perceptrón Multicapa (Multilayer Perceptron)
- ✓ Naïve Bayes

Els algorismes escollits són dels més utilitzats i coneguts, admeten atributs numèrics i són els més adients per aplicar a les dades obtingudes, per aconseguir els objectius fixats.

4.2.7 Utilitzar l'algorisme

Es decideix utilitzar el mode Cross-Validation 10-folds per realitzar l'avaluació, a causa de les característiques ja comentades sobre les dades. Després d'exposar els resultats de cada algorisme, es presenta una taula resum de tots els algorismes amb els següents coeficients:

- **Precisió**, instàncies correctament classificades
- **Kappa statistic**, mesura la precisió considerant la influència de l'atzar
- **F-measure,** mesura la bondat per cada classe
- **Matriu de confusió**, es veuen les instàncies classificades per classe, tant les correctes com les incorrectes

A l'algorisme J48 es modifica el paràmetre confidence factor, el qual controla la mida de l'arbre, a més alt (rang 0-1) construeix un arbre més complex. Realitzant proves es decideix pujar el valor per defecte (0.25) a 0.5. Respecte l'algorisme del veí més proper, es decideix utilitzar k=3 després de diverses proves. Aquests són els únic paràmetres a modificar, la resta de paràmetres de tots els algorismes es deixen per defecte.

J48

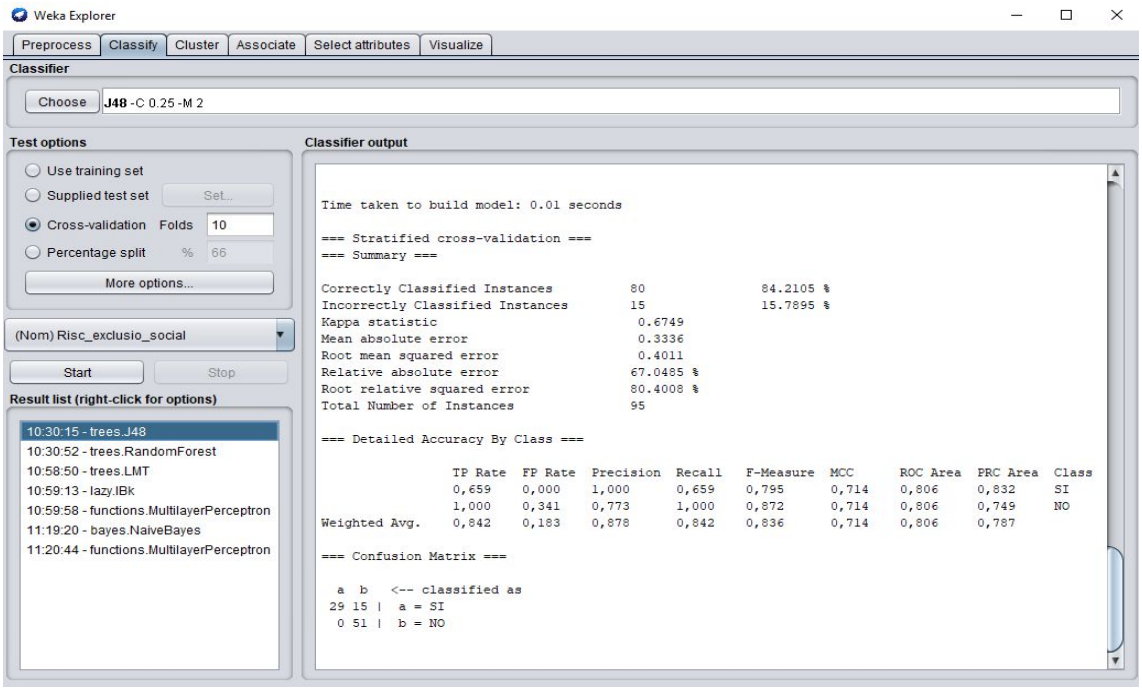


Figura 35. Resultats de l'algorisme J48

Random Forest

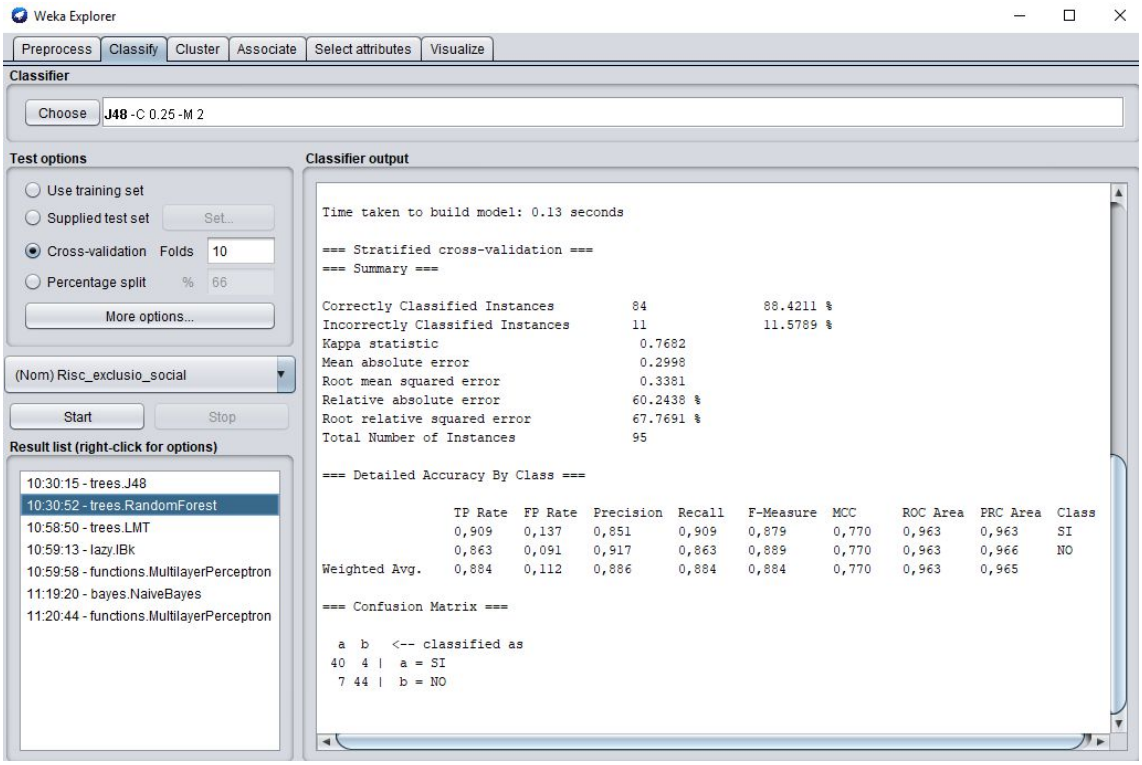


Figura 36. Resultats de l'algorisme Random Forest

LMT

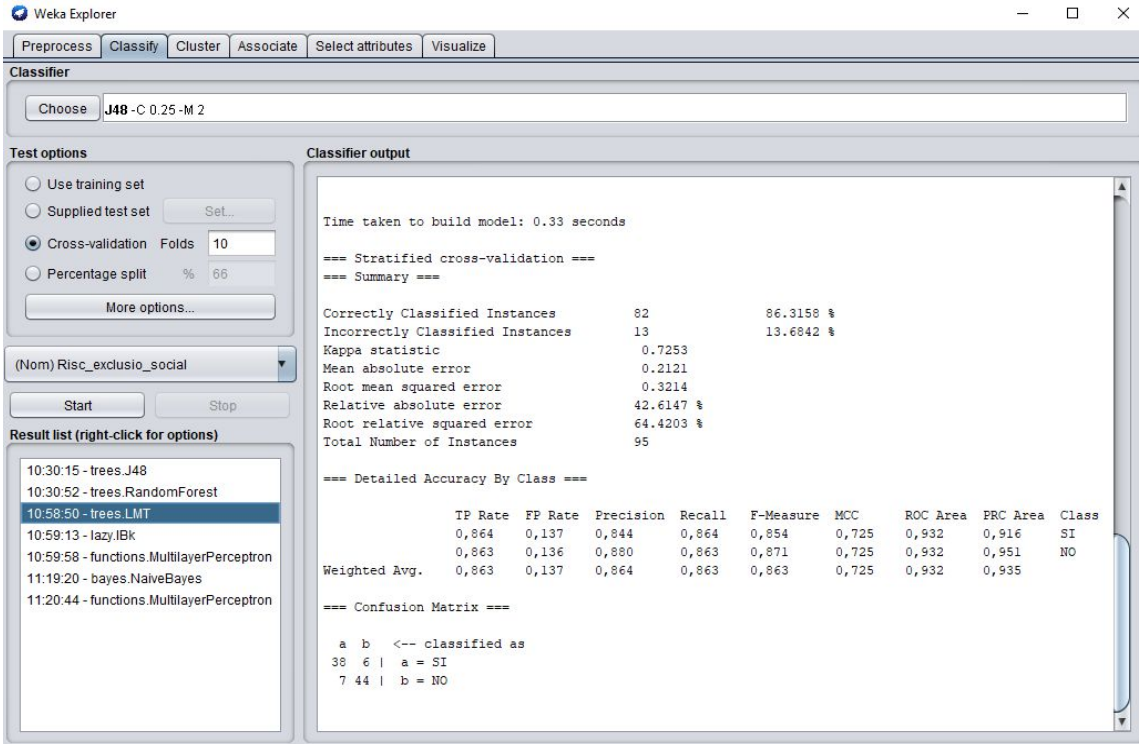


Figura 37. Resultats de l'algorisme LMT

Veïns més propers (ibK)

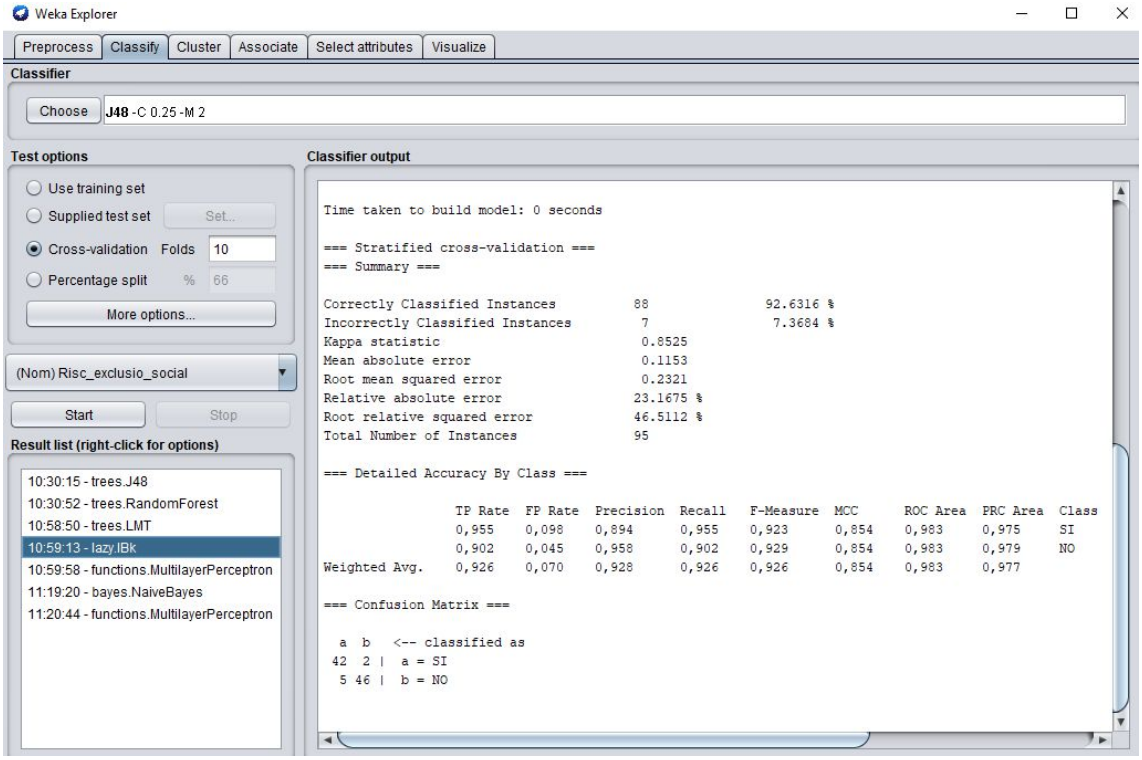


Figura 38. Resultats de l'algorisme Veïns més propers

Perceptrón Multicapa (Multilayer Perceptron)

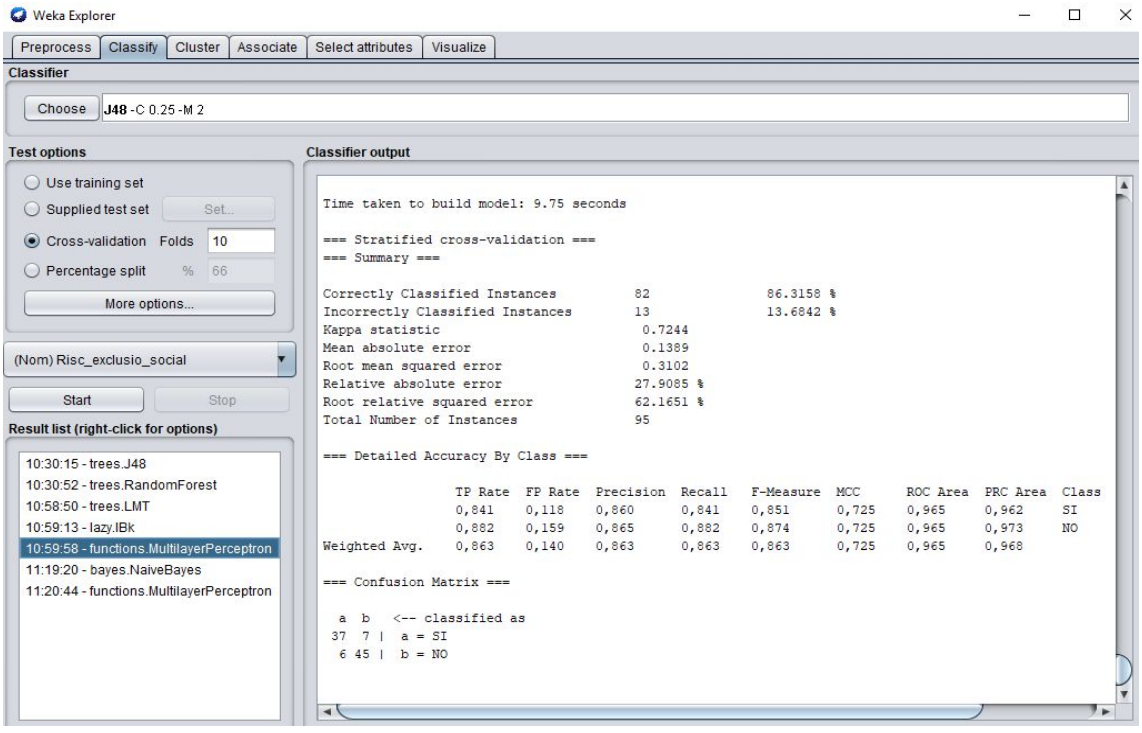


Figura 39. Resultats de l'algorisme Perceptrón Multicapa

Naïve Bayes

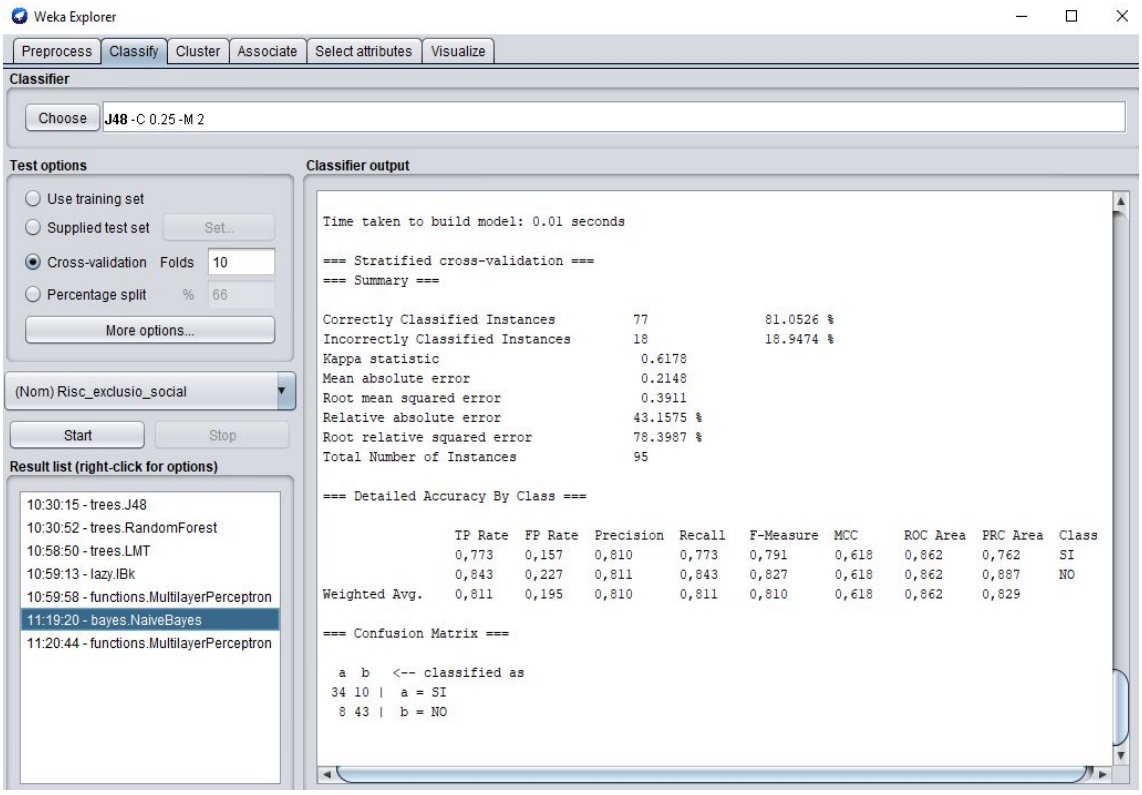


Figura 40. Resultats de l'algorisme Naïve Bayes

Taula resum

	J48	Random Forest	LMT	ibK	Multilayer Perceptron	Naïve Bayes
Precisió	84.2%	88.4%	86.3%	92.6%	86.3%	81%
Kappa statistic	0.6749	0.7682	0.7253	0.8525	0.7244	0.6178
F-measure(SI)	0.795	0.879	0.854	0.923	0.851	0.791
F-measure(NO)	0.872	0.889	0.871	0.929	0.874	0.827
Matriu de confusió	29 15 0 51	40 4 7 44	38 6 7 44	42 2 5 46	37 7 6 45	31 10 8 43

Figura 41. Taula resum dels models de classificació

4.2.8 Avaluar els resultats

Comprovant els resultats de la figura 41, es veu com tots els algorismes classifiquen prou bé, a més, aconsegueixen classificar més o menys igual tant les instàncies positives com les negatives, és a dir, el filtre *SMOTE* per balancejar l'atribut de classe millora els models. L'únic algorisme que classifica bastant pitjor la classe positiva és l'algorisme *J48*.

L'algorisme que presenta el millor model de classificació és l'algorisme de veïns més propers (*ibK*), millora en tots els coeficients o mesures però especialment és destacable el valor del coeficient *Kappa statistic* (0.8525) i que classifica millor la classe positiva. Una suposició del bon resultat d'aquest algorisme respecte els altres, podria ser la influència de les instàncies més properes en la classificació d'una instància, la qual cosa podria ser extrapolable a les relacions i proximitats entre barris. Per alguna cosa es parla de veïns i veïnats (barris);-))

4.2.9 Utilitzar el coneixement descobert

El model obtingut amb l'algorisme de veïns més propers, es podria utilitzar per detectar barris en risc d'*exclusió social* amb noves dades, cosa que queda fora de l'àmbit de treball per qüestions de temps. De la mateixa manera que amb el coneixement descobert amb el procés de *clustering*, es fa difusió dels resultats del treball en el repositori públic de la *UOC.*

5. Conclusions

La realització d'aquest TFG m'ha permès aprofundir una mica més en un àmbit tant d'actualitat com complex, la Mineria de Dades. És cert que ja havia utilitzat aquestes eines en assignatures del itinerari de Computació, però he de reconèixer el repte que ha sigut el desenvolupament del treball, elaborant un procés de KDD des de zero, cercant una temàtica que em motivés, documentant-me en la situació i complexitat de l'exclusió social, prenent decisions de com abordar aquesta problemàtica, cercant i generant un conjunt de dades que permeti un mínim de credibilitat per l'anàlisi, etc.

He après a valorar la constància, l'esforç i sobretot el no rendir-se encara que et sentis sol i no sàpigues cap a on tirar, no ha sigut gens fàcil mantenir un mínim de confiança i començar a caminar quan has perdut algunes eines i veus una muntanya que no has pujat mai.

Estic prou satisfet del resultat final del TFG i en la millora de les meves competències. En un principi els objectius plantejats es focalitzaven en tècniques de *clustering*, però al comprovar la dificultat en la cerca i creació de un conjunt de dades, vaig creure convenient ampliar els objectius amb tècniques de classificació.

Respecte la planificació del treball, he sofert diverses contingències, no trivials, que han endarrerit les entregues parcials, arribant al punt de posar en perill la finalització del treball, la qual cosa m'ha obligat a realitzar un esforç final molt important que espero no perjudiqui el resultat final del TFG. En canvi, considero que la metodologia i el fil conductor del treball han sigut correctes.

Aquest endarreriment no m'ha permès explorar noves línies de treball, com per exemple, analitzar l'evolució dels indicadors d'exclusió social amb dades històriques pels barris de Barcelona, ampliar l'anàlisi a altres delimitacions geogràfiques o modificar algun algorisme per adequar-lo a les dades.

Aquest projecte no pretén ser una referència però, sí posar el focus en una problemàtica creixent a les grans ciutats, evidenciant la falta de consens i dificultats en l'anàlisi de l'exclusió social. A causa de les múltiples dimensions

implicades, treball, immigració, habitatge, educació, etc., és imprescindible abordar l'exclusió social amb polítiques globals.

M'agradaria tornar a comentar la dificultat en la recollida de dades dels indicadors d'exclusió social, considero que queda molta feina a fer per consensuar els indicadors, recollir les dades en el format més adequat i sobretot la periodicitat d'aquestes. La majoria d'estudis sobre l'exclusió social fan servir bases de dades reduïdes o enquestes, i els que utilitzen indicadors com els utilitzats en aquest treball, es basen en el cens de població i vivendes amb periodicitat decennal, amb la consegüent inconsistències entre indicadors, ja que mentre alguns indicadors utilitzen dades actualitzades (per exemple, 2018), altres indicadors utilitzen dades del 2011.

Per últim, a pesar de no tenir un apartat d'agraïments, vull agrair a poques persones però molt importants per mi. En primer lloc, vull agrair per la seva comprensió i suport durant tant de temps al meu tutor del Grau d'Enginyeria Informàtica, *Javier Martí Pintanel*, també vull agrair per haver-me deixat espai en moments difícils per mi i per la seva paciència al consultor d'aquest TFG, *David Isern*, tampoc m'oblido (bé...dels noms si que m'he oblidat;-)) dels professors que em van guiar en el curs d'accés a la Universitat per majors de 25 anys, els quals em van ensenyar a escriure amb més de trenta anys i motivar per emprendre aquest grau. I per finalitzar, vull agrair per sobre de tothom a la meva dona **IRENE** i al meu fill **POL**, els quals són els que més han patit les meves absències i el meu mal caràcter en aquest llarg camí.

6. Bibliografia

[1] European Foundation. Disponible en: https://europeanfoundation.org/. Data consulta: 02/03/19.

[2] Comisión Europea, Estrategia Europa 2020. Disponible en: https://ec.europa.eu/info/business-economy-euro/economic-and-fiscal-policy-coordination/eu-economic-governance-monitoring-prevention-correction/european-semester/framework/europe-2020-strategy_es. Data consulta: 02/03/19.

[3] Atlas de vulnerabilidad urbana en España. Disponible en: https://www.fomento.gob.es/areas-de-actividad/arquitectura-vivienda-y-suelo/urbanismo-y-politica-de-suelo/observatorio-de-la-vulnerabilidad-urbana/atlas-de-la-vulnerabilidad-urbana/atlas-de-las-vulnerabilidad-urbana-en-espan%CC%83a Data consulta: 04/03/19

[4] Laparra, M. y Pérez, B. *Exclusión social en España: un espacio diverso y disperso en intensa transformación*. Madrid, Fundación FOESSA, 2008.

[5] Subirats, J. [dir.]; Riba C.; Giménez L., et al. *Pobreza y exclusión social. Un análisis de la realidad española y europea*. Barcelona: Fundación «la Caixa», 2004. (Colección Estudios Sociales,núm. 16).

[6] Institut d'Estadística de Catalunya, Idescat. Indicadors territorials de risc de pobresa i exclusió social (INTPOBR). Disponible en: http://www.idescat.cat/pub/?id=intpobr Data consulta: 06/03/19

[7] Lafuente Lechuga, M.; Faura Martínez, U. *Análisis de los individuos vulnerables a la exclusión social en España en 2009*. Anales de ASEPUMA, nº 21, 2013.

[8] Ramos, J.; Varela, A. *Beyond the margins: Analyzing social exclusion with a homeless client dataset.* Social Work & Society, 104p-120p, Volume 8, Issue 1, 2010.

[9] Serrano, E. et al. *Predicting the risk of suffering chronic social exclusion with machine learning*. Distributed Computing and Artificial Intelligence, 14th International Conference. Advances in Intelligent Systems and Computing, vol 620. Springer, Cham. 2017.

[10] Hile, R.; Cova, T.J. *Exploratory Testing of an Artificial Neural Network*

Classification for Enhancement of the Social Vulnerability Index, ISPRS International Journal of Geo-Information, 1774p-1790p, 4, 2015.

[11] Fayyad, U. M.; Piatetsky-Shapiro, G.; and Smyth, P. *Advances in Knowledge Discovery and Data Mining,* AAAI Press, Menlo Park, CA,1996.

[12] Maimon, O.; Rokach, L. *Data Mining and Knowledge Discovery Handbook*, 2nd ed., Springer, New York, 2010.

[13] Open Data BCN, Servei de dades obertes de l'Ajuntament de Barcelona. Disponible en: https://opendata-ajuntament.barcelona.cat/ca
Data consulta: 10/03/19

[14] Witten, I.H.; Frank, E. *Data Mining. Practical Machine Learning tools with Java implementations*, Morgan Kaufmann Publishers, San Francisco, USA, 2000

[15] Análisis Urbanístico de Barrios Vulnerables. Disponible en: https://www.fomento.es/recursos_mfom/pdf/C88DB66D-8669-497C-BEE4-442AE027E2FB/111287/SOBRE_vulnerabilidad.pdf
Data consulta: 21/03/19

[16] Observatori de Barris, Ajuntament de Barcelona. Disponible en: http://www.bcn.cat/estadistica/catala/documents/barris/
Data consulta: 01/04/19

[17] Pla de barris de Barcelona, Ajuntament de Barcelona. Disponible en: https://pladebarris.barcelona/ca/el-pla-dels-barris-de-barcelona
Data consulta: 15/04/19

[18] Garcia-Almirall, P., Vila, G., Moix, M., Ferrer, M., Vima-Grau, S. "*Estudi i detecció a la ciutat de Barcelona d'àmbits de vulnerabilitat residencial*", UPC, 2017.

Adriaans, P.; Zantinge, D. *Data Mining*, Addison-Wesley Longman Limited, Boston, USA, 1997

Han, J.; Kamber, M.; Pei, J; *Data Mining: Concepts and techniques*, 3rd edition, Morgan Kaufmann, 2011

Williams, G.; Simoff, S., *Data Mining: Teory, Methodology, Techniques, and Applications*, Springer, 2006.

Hernández, J.; Ramírez, M.J.; Ferri, C., Introducción a la Minería de Datos, Pearson Prentice Hall, Madrid, 2004

7. Annexos

7.1 Annex 1

Barris i districtes de Barcelona

Barcelona està dividida en 10 districtes i 73 barris:

1. **Districte de Ciutat Vella**
    1. El Raval
    2. El Barri Gòtic
    3. La Barceloneta
    4. Sant Pere, Santa Caterina i la Ribera
2. **Districte de l'Eixample**
    5. El Fort Pienc
    6. La Sagrada Família
    7. La Dreta de l'Eixample
    8. L'Antiga Esquerre de l'Eixample
    9. La Nova Esquerre de l'Eixample
    10. Sant Antoni
3. **Districte de Sants – Montjuïc**
    11. El Poble Sec
    12. La Marina del Prat Vermell
    13. La Marina de Port
    14. La Font de la Guatlla
    15. Hostafrancs
    16. La Bordeta
    17. Sants – Badal
    18. Sants
4. **Districte de Les Corts**
    19. Les corts
    20. La Maternitat i Sant Ramon
    21. Pedralbes

5. **Districte de Sarrià – Sant Gervasi**
    22. Vallvidrera, el Tibidabo i Les Planes
    23. Sarrià
    24. Les Tres Torres
    25. Sant Gervasi – la Bonanova
    26. Sant Gervasi – Galvany
    27. El Putxet i el Farró
6. **Districte de Gràcia**
    28. Vallcarca i els Penitents
    29. El Coll
    30. La Salut
    31. La Vila de Gràcia
    32. El camp d'en Grassot i Gràcia Nova
7. **Districte d'Horta – Guinardó**
    33. El Baix Guinardó
    34. Can Baró
    35. El Guinardó
    36. La Font d'en Fargues
    37. El Carmel
    38. La Teixonera
    39. Sant Genís dels Agudells
    40. Montbau
    41. La Vall d'Hebron
    42. La Clota
    43. Horta
8. **Districte de Nous Barris**
    44. Vilapicina i la Torre Llobeta
    45. Porta
    46. El Turó de la Peira
    47. Can Peguera
    48. La Guineueta
    49. Canyelles
    50. Les Roquetes
    51. Verdun

52. La Prosperitat
53. La Trinitat Nova
54. Torre Baró
55. Ciutat Meridiana
56. Vallbona

9. Districte de Sant Andreu

57. La Trinitat Vella
58. Baró de Viver
59. El Bon Pastor
60. Sant Andreu
61. La Sagrera
62. El congrés i els Indians
63. Navas

10. Districte de Sant Martí

64. El Camp de l'Arpa del Clot
65. El Clot
66. El Parc i la Llacuna del Poblenou
67. La Vila Olímpica del Poblenou
68. El Poblenou
69. Diagonal Mar i el Front Marítim del Poblenou
70. El Besòs i el Maresme
71. Provençals del Poblenou
72. Sant Martí de Provençals
73. La Verneda i la Pau

7.2 Annex 2

Datasets del cas d'estudi

Dataset: *dataset_supervisat_exclusio_social.arff*

@relation supervisat_exclusion_social

@attribute Barris {'el Raval','el Barri Gòtic','la Barceloneta','Sant Pere Santa Caterina i la Ribera','el Fort Pienc','la Sagrada Família','la Dreta de l\'Eixample','l\'Antiga Esquerra de l\'Eixample','la Nova Esquerra de l\'Eixample','Sant Antoni','el Poble Sec','la Marina del Prat Vermell','la Marina de Port','la Font de la Guatlla',Hostafrancs,'la Bordeta','Sants - Badal',Sants,'les Corts','la Maternitat i Sant Ramon',Pedralbes,'Vallvidrera el Tibidabo i les Planes',Sarrià,'les Tres Torres','Sant Gervasi - la Bonanova','Sant Gervasi - Galvany','el Putxet i el Farró','Vallcarca i els Penitents','el Coll','la Salut','la Vila de Gràcia','el Camp d\'en Grassot i Gràcia Nova','el Baix Guinardó','Can Baró','el Guinardó','la Font d\'en Fargues','el Carmel','la Teixonera','Sant Genís dels Agudells',Montbau,'la Vall d\'Hebron','la Clota',Horta,'Vilapicina i la Torre Llobeta',Porta,'el Turó de la Peira','Can Peguera','la Guineueta',Canyelles,'les Roquetes',Verdun,'la Prosperitat','la Trinitat Nova','Torre Baró','Ciutat Meridiana',Vallbona,'la Trinitat Vella','Baró de Viver','el Bon Pastor','Sant Andreu','la Sagrera','el Congrés i els Indians',Navas,'el Camp de l\'Arpa del Clot','el Clot','el Parc i la Llacuna del Poblenou','la Vila Olímpica del Poblenou','el Poblenou','Diagonal Mar i el Front Marítim del Poblenou','el Besòs i el Maresme','Provençals del Poblenou','Sant Martí de Provençals','la Verneda i la Pau'}
@attribute Percentatge_domicilis_amb_una_persona_sola_de_65_anys_i_més integer
@attribute Percentatge_població_amb_75_anys_o_més integer
@attribute Percentatge_infantil_estrangers integer
@attribute Percentatge_estrangers integer
@attribute Percentatge_domicilis_amb_una_persona_de_18_anys_i_més_amb_altres_menors_de_18_anys integer
@attribute Percentatge_atur_mitjana_any integer
@attribute Percentatge_atur_juvenil integer
@attribute Percentatge_pobl_sense_estudis integer
@attribute Percentatge_hab_Fins_a_30m2 integer
@attribute Percentatge_hab_Ruïnos_dolent_def integer
@attribute Percentatge_hab_abans_1951 integer
@attribute Risc_exclusio_social {'SI','NO'}

@data
'el Raval',11.23,6.63,45.19,48.52,2.13,10.40,13.50,2.49,7.47,36.05,81.40,SI
'el Barri Gòtic',10.22,7.78,32.30,43.26,2.20,7.68,15.90,2.28,7.05,24.51,90.49,SI
'la Barceloneta',13.47,11.35,21.19,31.89,2.70,10.19,13.10,4.74,11.84,26.5,68.21,NO
'Sant Pere Santa Caterina i la Ribera',11.50,8.23,34.17,39.62,2.80,9.10,13.20,2.47,7.84,26.99,82.46,NO
'el Fort Pienc',13.56,11.38,18.52,19.92,2.74,6.49,11.50,1.94,2.15,7.51,33.17,NO
'la Sagrada Família',14.73,12.44,15.90,17.83,2.60,6.69,11.30,2.49,2.45,9.09,36.91,NO
'la Dreta de l\'Eixample',14.03,11.81,15.56,19.90,3.15,5.28,12.50,1.28,2.18,4.35,67.01,NO
'l\'Antiga Esquerra de l\'Eixample',14.57,12.34,12.79,19.18,2.95,5.89,12.40,1.57,1.73,4.60,60.72,NO

'la Nova Esquerra de
l\'Eixample',14.67,11.95,13.83,16.56,2.57,6.50,12.90,1.71,1.57,10.90,37.20,NO
'Sant Antoni',15.56,12.90,16.80,20.10,2.37,7.03,13.20,2.55,2.04,10.33,51.85,NO
'el Poble Sec',12.72,9.62,28.84,29.32,2.50,9.19,11.20,3.21,4.12,10.37,59.74,SI
'la Marina del Prat
Vermell',15.61,11.61,4.70,7.59,2.71,17.14,10.40,6.48,11.78,0.00,16.79,SI
'la Marina de
Port',12.22,10.61,13.83,13.17,3.42,9.82,13.20,4.12,1.55,0.00,2.00,SI
'la Font de la
Guatlla',13.53,11.39,16.77,18.44,2.46,9.51,11.40,3.37,2.08,0.00,24.90,NO
'Hostafrancs',11.86,9.64,18.84,20.87,2.30,7.30,12.10,2.89,4.18,9.40,42.17,SI
'la Bordeta',12.50,11.74,12.42,12.46,2.48,7.80,12.10,3.57,1.24,3.99,13.88,NO
'Sants - Badal',12.86,11.07,19.08,16.27,2.55,6.49,11.40,3.83,1.70,8.83,21.29,NO
'Sants',13.95,11.68,14.51,16.17,2.75,7.16,11.90,3.05,3.21,14.83,38.22,NO
'les Corts',14.74,12.35,9.62,10.41,2.74,6.96,12.00,1.89,2.08,2.00,10.69,NO
'la Maternitat i Sant
Ramon',14.07,12.34,8.30,10.44,2.94,6.93,11.40,2.61,2.08,0.00,6.71,NO
'Pedralbes',14.13,13.10,19.37,15.60,3.53,3.34,8.30,0.54,2.74,0.00,5.86,NO
'Vallvidrera el Tibidabo i les
Planes',7.61,7.41,12.13,13.29,5.95,3.23,11.00,1.95,12.73,13.06,27.07,SI
'Sarrià',12.52,11.79,11.88,11.84,4.73,3.40,9.20,1.15,4.64,2.16,20.56,NO
'les Tres Torres',12.28,11.22,6.55,8.69,4.22,3.82,10.90,0.49,1.89,0.00,12.92,NO
'Sant Gervasi - la
Bonanova',13.44,12.06,7.94,10.62,4.12,4.18,10.30,0.81,3.81,5.23,24.53,NO
'Sant Gervasi -
Galvany',15.60,12.51,8.45,11.23,3.77,4.12,10.90,0.64,2.74,2.03,30.76,NO
'el Putxet i el
Farró',14.41,10.54,8.06,12.51,3.45,5.83,11.30,0.98,3.39,4.18,18.89,NO
'Vallcarca i els
Penitents',14.12,11.97,9.64,12.72,3.83,7.11,12.30,2.34,4.15,2.23,14.78,NO
'el Coll',12.08,9.93,12.47,13.77,3.70,5.93,12.30,3.11,4.31,9.24,9.22,NO
'la Salut',15.02,12.37,9.46,12.02,2.63,7.00,12.40,1.89,5.85,0.00,12.09,NO
'la Vila de
Gràcia',13.29,11.21,13.41,19.12,3.30,6.62,12.80,2.22,5.37,9.26,58.55,NO
'el Camp d\'en Grassot i Gràcia
Nova',14.35,12.36,10.35,12.64,2.91,6.74,11.60,2.05,2.35,4.14,29.44,NO
'el Baix
Guinardó',14.66,14.08,12.36,12.98,2.32,7.02,11.50,3.37,1.88,1.31,11.27,NO
'Can Baró',13.73,12.18,12.32,12.97,2.80,7.09,12.40,4.21,5.07,25.56,19.28,NO
'el Guinardó',14.44,12.33,11.94,13.10,3.07,8.03,12.10,2.97,3.86,6.59,13.67,NO
'la Font d\'en
Fargues',11.57,11.19,4.16,5.15,3.61,9.78,12.60,2.63,7.40,0.00,16.14,NO
'el Carmel',12.53,11.22,14.11,12.74,2.98,9.65,12.60,6.37,4.54,5.07,4.63,SI
'la Teixonera',12.30,11.18,12.44,12.14,3.04,7.85,12.20,5.32,4.41,0.00,2.86,SI
'Sant Genís dels
Agudells',15.75,17.02,11.36,11.13,3.50,8.43,12.10,6.62,2.56,0.00,6.40,SI
'Montbau',19.13,23.54,11.05,9.82,2.13,8.59,12.10,7.21,4.39,0.00,0.54,NO
'la Vall
d\'Hebron',15.57,12.57,10.31,8.25,3.03,7.71,12.10,2.89,31.07,0.00,0.37,NO
'la Clota',9.09,10.17,11.49,13.05,5.14,4.50,12.10,4.42,9.72,0.00,46.21,SI
'Horta',14.56,13.78,10.38,10.20,3.20,9.88,12.40,4.80,5.59,2.88,16.18,NO
'Vilapicina i la Torre
Llobeta',15.53,13.57,14.25,12.37,2.65,9.40,12.40,4.29,2.68,3.22,5.55,NO
'Porta',15.41,14.39,15.51,13.53,2.66,9.68,13.30,5.77,1.52,5.92,2.98,NO
'el Turó de la
Peira',18.40,17.18,23.13,19.94,2.42,9.73,12.10,6.91,1.64,14.55,2.78,NO
'Can Peguera',13.58,14.98,5.50,7.22,2.96,9.39,12.70,6.27,3.71,0.00,62.61,SI

'la Guineueta',16.20,15.29,7.48,7.28,2.60,12.13,14.20,5.20,0.10,0.00,0.01,NO
'Canyelles',14.24,11.65,4.48,4.29,3.00,11.94,14.20,5.44,0.24,0.00,0.10,NO
'les Roquetes',11.59,8.85,20.87,19.38,3.31,10.99,14.20,6.03,4.68,6.57,1.73,SI
'Verdun',15.01,12.83,21.80,17.59,3.13,11.05,14.20,6.56,2.67,3.34,1.85,NO
'la Prosperitat',15.32,13.87,15.75,13.48,3.01,10.38,13.10,7.07,2.85,7.98,3.04,NO
'la Trinitat
Nova',14.88,11.87,20.19,17.62,3.23,15.72,14.30,7.24,0.73,0.00,0.42,SI
'Torre Baró',7.73,6.15,15.53,15.99,6.02,10.95,14.30,4.89,13.90,0.00,9.04,SI
'Ciutat Meridiana',11.71,8.53,31.95,27.36,2.77,16.53,14.30,4.44,0.03,0.00,0.00,SI
'Vallbona',6.60,11.37,10.42,11.67,4.68,14.66,14.30,6.92,18.59,0.00,6.24,SI
'la Trinitat Vella',9.33,7.39,28.07,27.81,3.34,14.70,14.30,4.50,7.97,0.00,3.85,SI
'Baró de Viver',11.63,9.00,9.53,10.43,5.19,8.39,11.80,5.06,4.05,0.00,0.00,SI
'el Bon Pastor',10.78,9.27,10.20,12.76,3.50,9.86,11.80,4.30,3.86,9.13,11.27,SI
'Sant Andreu',12.39,10.36,6.72,6.93,3.12,8.75,11.90,3.00,4.93,4.04,16.89,NO
'la Sagrera',12.51,11.35,12.07,11.73,2.78,8.01,11.00,3.13,0.96,0.00,5.22,NO
'el Congrés i els
Indians',17.07,14.97,12.63,12.53,2.86,8.14,11.00,4.13,2.28,0.00,9.92,NO
'Navas',15.17,12.49,14.49,13.41,2.37,7.94,11.10,3.01,0.96,2.16,9.81,NO
'el Camp de l\'Arpa del
Clot',14.20,11.60,14.23,15.33,2.46,8.35,11.20,2.82,2.50,3.95,27.16,NO
'el Clot',11.71,9.61,11.48,13.52,2.92,7.93,11.00,2.74,2.97,4.92,23.26,NO
'el Parc i la Llacuna del
Poblenou',11.40,9.26,17.32,20.21,3.10,6.87,10.80,2.39,2.87,2.39,23.66,NO
'la Vila Olímpica del
Poblenou',6.90,4.59,13.87,15.78,4.91,7.00,10.60,0.78,2.34,0.00,7.69,NO
'el Poblenou',10.87,9.09,13.40,16.73,3.73,7.00,10.60,3.01,3.18,11.41,39.87,NO
'Diagonal Mar i el Front Marítim del
Poblenou',7.26,6.99,15.21,16.77,4.99,8.44,10.90,2.71,1.55,0.00,2.78,NO
'el Besòs i el
Maresme',13.74,11.38,27.75,24.59,3.04,11.15,10.90,6.19,3.21,19.83,2.04,SI
'Provençals del
Poblenou',10.20,9.24,14.08,13.14,3.37,8.19,11.00,3.45,1.29,6.68,11.99,NO
'Sant Martí de
Provençals',15.29,13.97,11.54,10.94,2.48,9.93,11.20,4.48,0.61,2.57,0.29,NO
'la Verneda i la
Pau',14.30,13.95,13.95,10.56,2.38,10.29,11.20,5.69,0.42,4.10,0.61,SI

Dataset: *dataset_no_supervisat_exclusio_social.arff*

@relation no_supervisat_exclusion_social

@attribute Barris {'el Raval','el Barri Gòtic','la Barceloneta','Sant Pere Santa Caterina i la Ribera','el Fort Pienc','la Sagrada Família','la Dreta de l\'Eixample','l\'Antiga Esquerra de l\'Eixample','la Nova Esquerra de l\'Eixample','Sant Antoni','el Poble Sec','la Marina del Prat Vermell','la Marina de Port','la Font de la Guatlla',Hostafrancs,'la Bordeta','Sants - Badal',Sants,'les Corts','la Maternitat i Sant Ramon',Pedralbes,'Vallvidrera el Tibidabo i les Planes',Sarrià,'les Tres Torres','Sant Gervasi - la Bonanova','Sant Gervasi - Galvany','el Putxet i el Farró','Vallcarca i els Penitents','el Coll','la Salut','la Vila de Gràcia','el Camp d\'en Grassot i Gràcia Nova','el Baix Guinardó','Can Baró','el Guinardó','la Font d\'en Fargues','el Carmel','la Teixonera','Sant Genís dels Agudells',Montbau,'la Vall d\'Hebron','la Clota',Horta,'Vilapicina i la Torre Llobeta',Porta,'el Turó de la Peira','Can Peguera','la Guineueta',Canyelles,'les Roquetes',Verdun,'la Prosperitat','la Trinitat Nova','Torre Baró','Ciutat Meridiana',Vallbona,'la Trinitat Vella','Baró de Viver','el Bon Pastor','Sant Andreu','la Sagrera','el Congrés i els Indians',Navas,'el Camp de l\'Arpa del Clot','el Clot','el

Parc i la Llacuna del Poblenou','la Vila Olímpica del Poblenou','el Poblenou','Diagonal Mar i el Front Marítim del Poblenou','el Besòs i el Maresme','Provençals del Poblenou','Sant Martí de Provençals','la Verneda i la Pau'}
@attribute Percentatge_domicilis_amb_una_persona_sola_de_65_anys_i_més integer
@attribute Percentatge_població_amb_75_anys_o_més integer
@attribute Percentatge_infantil_estrangers integer
@attribute Percentatge_estrangers integer
@attribute Percentatge_domicilis_amb_una_persona_de_18_anys_i_més_amb_altres_menors_de_18_anys integer
@attribute Percentatge_atur_mitjana_any integer
@attribute Percentatge_atur_juvenil integer
@attribute Percentatge_pobl_sense_estudis integer
@attribute Percentatge_hab_Fins_a_30m2 integer
@attribute Percentatge_hab_Ruïnos_dolent_def integer
@attribute Percentatge_hab_abans_1951 integer

@data
'el Raval',11.23,6.63,45.19,48.52,2.13,10.40,13.50,2.49,7.47,36.05,81.40
'el Barri Gòtic',10.22,7.78,32.30,43.26,2.20,7.68,15.90,2.28,7.05,24.51,90.49
'la Barceloneta',13.47,11.35,21.19,31.89,2.70,10.19,13.10,4.74,11.84,26.5,68.21
'Sant Pere Santa Caterina i la Ribera',11.50,8.23,34.17,39.62,2.80,9.10,13.20,2.47,7.84,26.99,82.46
'el Fort Pienc',13.56,11.38,18.52,19.92,2.74,6.49,11.50,1.94,2.15,7.51,33.17
'la Sagrada Família',14.73,12.44,15.90,17.83,2.60,6.69,11.30,2.49,2.45,9.09,36.91
'la Dreta de l\'Eixample',14.03,11.81,15.56,19.90,3.15,5.28,12.50,1.28,2.18,4.35,67.01
'l\'Antiga Esquerra de l\'Eixample',14.57,12.34,12.79,19.18,2.95,5.89,12.40,1.57,1.73,4.60,60.72
'la Nova Esquerra de l\'Eixample',14.67,11.95,13.83,16.56,2.57,6.50,12.90,1.71,1.57,10.90,37.20
'Sant Antoni',15.56,12.90,16.80,20.10,2.37,7.03,13.20,2.55,2.04,10.33,51.85
'el Poble Sec',12.72,9.62,28.84,29.32,2.50,9.19,11.20,3.21,4.12,10.37,59.74
'la Marina del Prat Vermell',15.61,11.61,4.70,7.59,2.71,17.14,10.40,6.48,11.78,0.00,16.79
'la Marina de Port',12.22,10.61,13.83,13.17,3.42,9.82,13.20,4.12,1.55,0.00,2.00
'la Font de la Guatlla',13.53,11.39,16.77,18.44,2.46,9.51,11.40,3.37,2.08,0.00,24.90
'Hostafrancs',11.86,9.64,18.84,20.87,2.30,7.30,12.10,2.89,4.18,9.40,42.17
'la Bordeta',12.50,11.74,12.42,12.46,2.48,7.80,12.10,3.57,1.24,3.99,13.88
'Sants - Badal',12.86,11.07,19.08,16.27,2.55,6.49,11.40,3.83,1.70,8.83,21.29
'Sants',13.95,11.68,14.51,16.17,2.75,7.16,11.90,3.05,3.21,14.83,38.22
'les Corts',14.74,12.35,9.62,10.41,2.74,6.96,12.00,1.89,2.08,2.00,10.69
'la Maternitat i Sant Ramon',14.07,12.34,8.30,10.44,2.94,6.93,11.40,2.61,2.08,0.00,6.71
'Pedralbes',14.13,13.10,19.37,15.60,3.53,3.34,8.30,0.54,2.74,0.00,5.86
'Vallvidrera el Tibidabo i les Planes',7.61,7.41,12.13,13.29,5.95,3.23,11.00,1.95,12.73,13.06,27.07
'Sarrià',12.52,11.79,11.88,11.84,4.73,3.40,9.20,1.15,4.64,2.16,20.56
'les Tres Torres',12.28,11.22,6.55,8.69,4.22,3.82,10.90,0.49,1.89,0.00,12.92
'Sant Gervasi - la Bonanova',13.44,12.06,7.94,10.62,4.12,4.18,10.30,0.81,3.81,5.23,24.53
'Sant Gervasi - Galvany',15.60,12.51,8.45,11.23,3.77,4.12,10.90,0.64,2.74,2.03,30.76
'el Putxet i el Farró',14.41,10.54,8.06,12.51,3.45,5.83,11.30,0.98,3.39,4.18,18.89

'Vallcarca i els
Penitents',14.12,11.97,9.64,12.72,3.83,7.11,12.30,2.34,4.15,2.23,14.78
'el Coll',12.08,9.93,12.47,13.77,3.70,5.93,12.30,3.11,4.31,9.24,9.22
'la Salut',15.02,12.37,9.46,12.02,2.63,7.00,12.40,1.89,5.85,0.00,12.09
'la Vila de Gràcia',13.29,11.21,13.41,19.12,3.30,6.62,12.80,2.22,5.37,9.26,58.55
'el Camp d\'en Grassot i Gràcia
Nova',14.35,12.36,10.35,12.64,2.91,6.74,11.60,2.05,2.35,4.14,29.44
'el Baix Guinardó',14.66,14.08,12.36,12.98,2.32,7.02,11.50,3.37,1.88,1.31,11.27
'Can Baró',13.73,12.18,12.32,12.97,2.80,7.09,12.40,4.21,5.07,25.56,19.28
'el Guinardó',14.44,12.33,11.94,13.10,3.07,8.03,12.10,2.97,3.86,6.59,13.67
'la Font d\'en
Fargues',11.57,11.19,4.16,5.15,3.61,9.78,12.60,2.63,7.40,0.00,16.14
'el Carmel',12.53,11.22,14.11,12.74,2.98,9.65,12.60,6.37,4.54,5.07,4.63
'la Teixonera',12.30,11.18,12.44,12.14,3.04,7.85,12.20,5.32,4.41,0.00,2.86
'Sant Genís dels
Agudells',15.75,17.02,11.36,11.13,3.50,8.43,12.10,6.62,2.56,0.00,6.40
'Montbau',19.13,23.54,11.05,9.82,2.13,8.59,12.10,7.21,4.39,0.00,0.54
'la Vall d\'Hebron',15.57,12.57,10.31,8.25,3.03,7.71,12.10,2.89,31.07,0.00,0.37
'la Clota',9.09,10.17,11.49,13.05,5.14,4.50,12.10,4.42,9.72,0.00,46.21
'Horta',14.56,13.78,10.38,10.20,3.20,9.88,12.40,4.80,5.59,2.88,16.18
'Vilapicina i la Torre
Llobeta',15.53,13.57,14.25,12.37,2.65,9.40,12.40,4.29,2.68,3.22,5.55
'Porta',15.41,14.39,15.51,13.53,2.66,9.68,13.30,5.77,1.52,5.92,2.98
'el Turó de la Peira',18.40,17.18,23.13,19.94,2.42,9.73,12.10,6.91,1.64,14.55,2.78
'Can Peguera',13.58,14.98,5.50,7.22,2.96,9.39,12.70,6.27,3.71,0.00,62.61
'la Guineueta',16.20,15.29,7.48,7.28,2.60,12.13,14.20,5.20,0.10,0.00,0.01
'Canyelles',14.24,11.65,4.48,4.29,3.00,11.94,14.20,5.44,0.24,0.00,0.10
'les Roquetes',11.59,8.85,20.87,19.38,3.31,10.99,14.20,6.03,4.68,6.57,1.73
'Verdun',15.01,12.83,21.80,17.59,3.13,11.05,14.20,6.56,2.67,3.34,1.85
'la Prosperitat',15.32,13.87,15.75,13.48,3.01,10.38,13.10,7.07,2.85,7.98,3.04
'la Trinitat Nova',14.88,11.87,20.19,17.62,3.23,15.72,14.30,7.24,0.73,0.00,0.42
'Torre Baró',7.73,6.15,15.53,15.99,6.02,10.95,14.30,4.89,13.90,0.00,9.04
'Ciutat Meridiana',11.71,8.53,31.95,27.36,2.77,16.53,14.30,4.44,0.03,0.00,0.00
'Vallbona',6.60,11.37,10.42,11.67,4.68,14.66,14.30,6.92,18.59,0.00,6.24
'la Trinitat Vella',9.33,7.39,28.07,27.81,3.34,14.70,14.30,4.50,7.97,0.00,3.85
'Baró de Viver',11.63,9.00,9.53,10.43,5.19,8.39,11.80,5.06,4.05,0.00,0.00
'el Bon Pastor',10.78,9.27,10.20,12.76,3.50,9.86,11.80,4.30,3.86,9.13,11.27
'Sant Andreu',12.39,10.36,6.72,6.93,3.12,8.75,11.90,3.00,4.93,4.04,16.89
'la Sagrera',12.51,11.35,12.07,11.73,2.78,8.01,11.00,3.13,0.96,0.00,5.22
'el Congrés i els
Indians',17.07,14.97,12.63,12.53,2.86,8.14,11.00,4.13,2.28,0.00,9.92
'Navas',15.17,12.49,14.49,13.41,2.37,7.94,11.10,3.01,0.96,2.16,9.81
'el Camp de l\'Arpa del
Clot',14.20,11.60,14.23,15.33,2.46,8.35,11.20,2.82,2.50,3.95,27.16
'el Clot',11.71,9.61,11.48,13.52,2.92,7.93,11.00,2.74,2.97,4.92,23.26
'el Parc i la Llacuna del
Poblenou',11.40,9.26,17.32,20.21,3.10,6.87,10.80,2.39,2.87,2.39,23.66
'la Vila Olímpica del
Poblenou',6.90,4.59,13.87,15.78,4.91,7.00,10.60,0.78,2.34,0.00,7.69
'el Poblenou',10.87,9.09,13.40,16.73,3.73,7.00,10.60,3.01,3.18,11.41,39.87
'Diagonal Mar i el Front Marítim del
Poblenou',7.26,6.99,15.21,16.77,4.99,8.44,10.90,2.71,1.55,0.00,2.78
'el Besòs i el
Maresme',13.74,11.38,27.75,24.59,3.04,11.15,10.90,6.19,3.21,19.83,2.04
'Provençals del
Poblenou',10.20,9.24,14.08,13.14,3.37,8.19,11.00,3.45,1.29,6.68,11.99

'Sant Martí de
Provençals',15.29,13.97,11.54,10.94,2.48,9.93,11.20,4.48,0.61,2.57,0.29
'la Verneda i la
Pau',14.30,13.95,13.95,10.56,2.38,10.29,11.20,5.69,0.42,4.10,0.61