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

Gustavo Talavera Patón

Àrea d'Intel·ligència Artificial

Consultor: David Isern Alarcón

TFG – Juny 2019

UOC
Universitat Oberta de Catalunya
www.uoc.edu

Agenda

- Motivacions del projecte
- Anàlisi de l'exclusió social
- Revisió de la Mineria de Dades
- Metodologia a desenvolupar
- Aplicació de la metodologia al cas d'estudi
- Resultats obtinguts
- Conclusions

Motivacions

- Comprendre i revisar la problemàtica de l'exclusió social focalitzant a l'anàlisi de dades
- Aprofitar el creixement d'iniciatives de dades obertes
- Aplicar i ampliar els coneixements adquirits en l'anàlisi de dades

Exclusió social

- "*Procés a través del qual persones o grups queden totalment o parcialment exclosos de la plena participació en la societat en què viuen*" [European Foundation]

- Fenomen d'origen estructural, caràcter multidimensional i perspectiva dinàmica

Exclusió social

Iniciatives

- La Unió Europea impulsa l'Estratègia Europa 2020 per un creixement intel·ligent, sostenible i integrador
- Atlas de Vulnerabilitat Urbana en Espanya
- Institut d'Estadística de Catalunya, indicadors territorials de risc de pobresa i exclusió social

Mineria de Dades

- *"És l'extracció d'informació implícita, prèviament desconeguda i potencialment útil a partir de dades"*[Witten i Krank]
- Tecnologia multidisciplinar que es nodreix de diferents disciplines y àrees
- És el nucli del procés de descobriment de coneixement en bases de dades (**KDD**)

Mineria de Dades

Nucli del procés de KDD

- La Mineria de Dades és una part del procés de KDD, especialitzada en l'extracció d'informació

- KDD (*Knowledge Discovery in Databases*) "és el procés no trivial d'identificar patrons **vàlids**, **nous**, potencialment **útils** i, finalment, **comprensibles** en les dades" [Fayyad]

Mineria de Dades

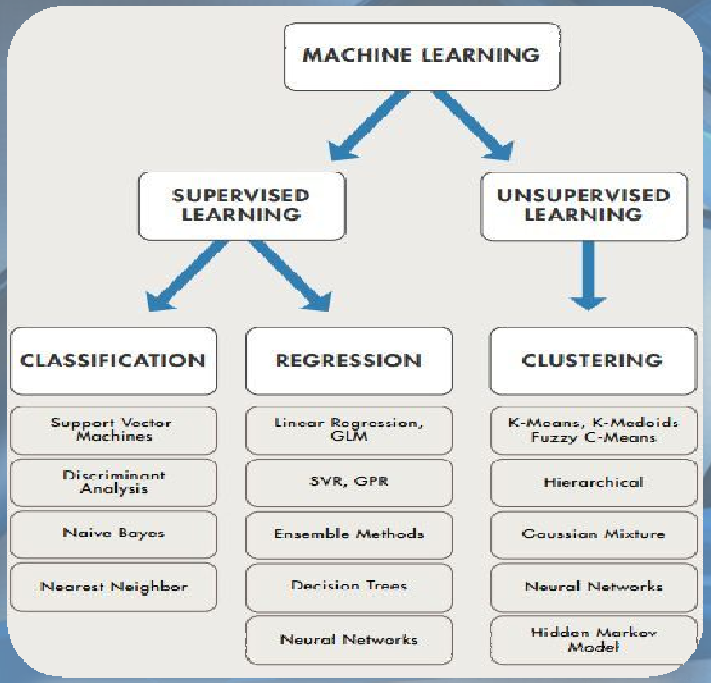
Disciplines i àrees

- Estadística
- Recuperació de la informació
- Sistemes de bases de dades
- Visualització
- **Machine Learning**
- Altres

Mineria de Dades

Machine Learning (ML)

Tècniques de ML



Algorismes de ML

- **Classificació**

Arbres de decisió

Veïns més propers

Naïve Bayes

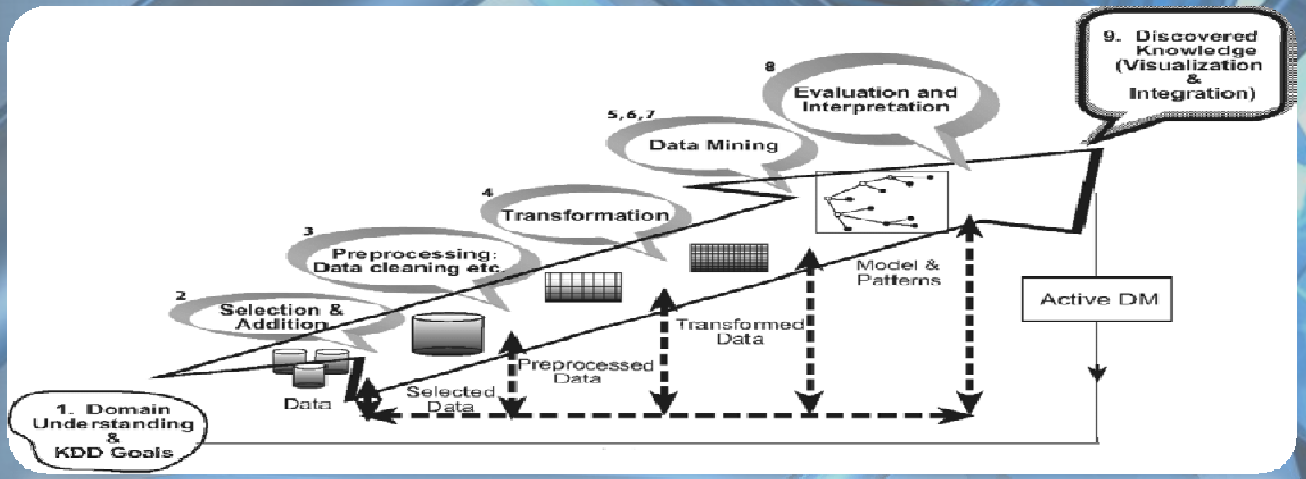
- **Clustering**

K-means, EM

Metodologia

Knowledge Discovery in Databases

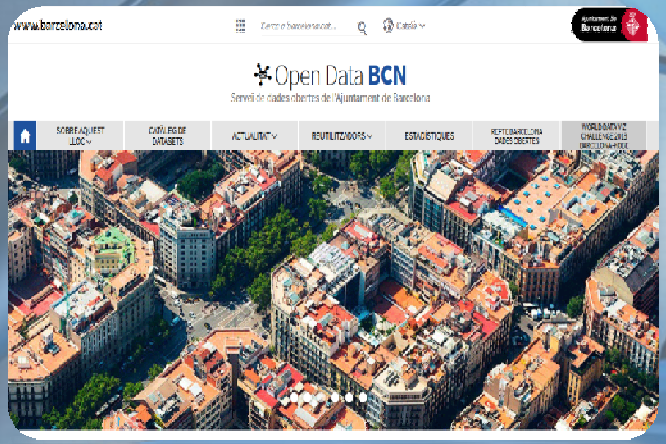
- S'implementa el procés de KDD analitzant i desenvolupant cada fase (9 fases)



Metodologia

Bases de dades

Servei de dades obertes de l'Ajuntament de Barcelona, Open Data BCN



Programari ML

Programari lliure Weka (*Waikato Environment for Knowledge Analysis)*

Cas d'estudi

- Es desenvolupen dos processos independents de *KDD utilitzant les dues tècniques principals,* ***classificació*** *i* ***clustering***
- Depèn del cas d'estudi concret, no és necessari desenvolupar totes les fases
- Després de comprendre el domini es decideix utilitzar com referència els indicadors d'exclusió social de l'Anàlisi Urbanístic de Barris Vulnerables(20 indicadors)

Cas d'estudi

- Analitzat el domini es defineixen els objectius de cada procés(**fase 1**):

  - **Clustering**. Extreure coneixement no obvi per comprendre les desigualtats entre barris de Barcelona i quins indicadors són els que més influència tenen.
  - **Classificació**. Trobar un bon model classificador per detectar els barris de Barcelona amb risc d'*exclusió social*.

Cas d'estudi

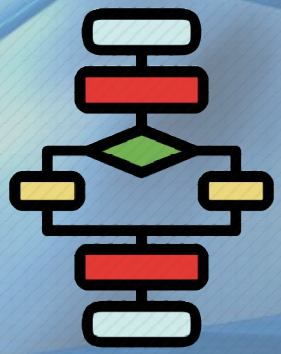
- Es seleccionen les dades del servei Open Data BCN, recollint, transformant i generant el conjunt de dades segons els indicadors (obtenint 11 de 20 indicadors)
- En el procés de classificació es complementa aquest conjunt amb un atribut de classe binari{SI,NO}, que determina si està en situació d'alt risc (**fase 2**)
- Es converteix el conjunt de dades en format d'Excel al format nadiu de *Weka*, **arff**, per continuar amb el procés (**fase 3**)

Cas d'estudi

Clustering

S'utilitzen els següents algorismes de Weka a la pestanya Cluster (**fase 6**):

- K-means
- EM



Classificació

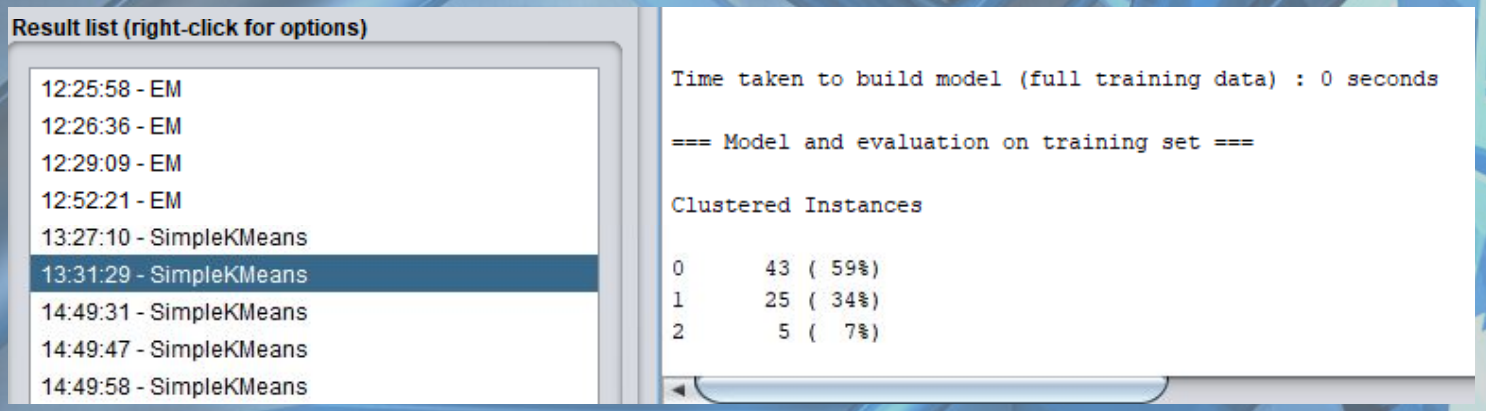
S'aplica el filtre **SMOTE** per balancejar les dades(**fase4**). S'utilitzen els següents algorismes de Weka a la pestanya Classify (**fase 6**):

- J48, Random Forest
- Multilayer Perceptron, ibK
- Naïve Bayes, LMT

Cas d'estudi

Clustering

- (**Fase 7**) S'apliquen els algorismes K-means i EM obtenint resultats pràcticament idèntics, agrupant els barris en 3 clústers



Resultat clustering

- ✓ (**fase 8**)Els resultats mostren 43 barris al Clúster 0 amb valors mitjans o baixos de tots els indicadors, són els barris en situació de menys risc
- ✓ El Clúster 1 agrupa 25 barris amb alts valors d'atur i població sense estudis, classificant aquest clúster com a situació en risc
- ✓ Per últim, el Clúster 2 mostra els 5 barris en pitjor situació, amb una alta immigració i problemes d'habitatge

Resultat clustering

- Indicadors (atributs) amb més incidència en la classificació d'exclusió social:
  - ✓ 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

Cas d'estudi

Classificació

- S'apliquen els algorismes J48, Random Forest, LMT, ibK, Multilayer Perceptron i Naïve Bayes obtenint:

	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

Resultat classificació

- ✓ Tots els models presenten un bon nivell de precisió, destacar en negatiu l'algorisme J48 al classificar molt pitjor la classe minoritària (positiva)
- ✓ Un model destaca entre els altres, el de veïns més propers (ibK, amb K=3), tant per la precisió (92,6%) com per classificar millor la classe minoritària. Però sobretot per valor de Kappa statistic (0.8525), el qual estima la confiabilitat de les dades

Conclusions

- ✓ 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