Cas d'estudi: pous de petroli

Luis Carlos Molina Félix
Ramon Sangüesa i Solé

P03/05054/01041

UOC
www.uoc.edu

Índex

Introducció

Aquest mòdul presenta un cas d'estudi que pretén que l'estudiant s'atansi a la realitat de la mineria de dades a partir d'una situació real.

El cas presentat en aquest mòdul es basa en una situació real. Les dades que es donen s'han extret d'un cas concret i s'han adaptat perquè es puguin presentar com a exemple vàlid de les tècniques que empra la mineria de dades.

Objectius

Els materials didàctics associats a aquest mòdul permetran a l'estudiant d'assolir els objectius següents:

1. Posar el participant en contacte amb una situació real en què la mineria de dades és una metodologia potent que permet de fer una anàlisi acurada d'un problema i exteure'n conclusions.

2. Permetre a l'estudiant de conèixer les dificultats que es plantegen amb les tècniques de mineria de dades i els avantatges que presenten aquestes tècniques.

1. El problema dels pous de petroli

Per a decidir la conveniència o inconveniència i la dificultat de perforar un pou petrolífer i tenir una idea dels possibles costos de manteniment, la permeabilitat d'un pou és un factor important. El fet de conèixer la permeabilitat d'un pou és, doncs, molt important. Es tracta d'un problema de predicció. Un dels atributs importants que cal considerar és la porositat del terreny, que dóna una idea de la permeabilitat i la profunditat dels pous veïns.

Presentarem la manera com es va afrontar el problema. En concret, ens fixarem en els aspectes següents:

a) Com es van escollir les dades que cal considerar com a factors predictius.
b) Les tècniques de preparació de dades emprades.
c) Els diferents mètodes aplicats i els models resultants.
d) L'avaluació i comparació dels diversos resultats.
e) La comparació amb el coneixement dels experts sobre el mateix tema.

El cas té com a característiques més interessants la influència de les tècniques de preparació de dades en el resultat final i la importància relativa del coneixement aportat pels experts del domini.

1.1. Quin és el problema exactament?

El treball desenvolupat es basa en mètodes desenvolupats i dades recollides per S.J. Rogers, H.C. Chen, D.C. Kopaska-Merkeli i J.H. Fang publicades a *Predicting Permeability from Porosity Using Artificial Networks*. En aquest treball es van fer servir xarxes neuronals per a predir la permeabilitat d'un pou petrolífer segons la porositat i profunditat. El conjunt de dades procedeix de mesures preses en sis pous petrolífers a Big Escambia Creek, Alabama, als Estats Units. A partir d'aquestes dades i de consideracions geològiques aportades per experts en la matèria es van preparar tres escenaris de prova diferents:

Lectura complementària

Trobareu les dades del problema considerat en aquest apartat a l'obra següent:

S.J. Rogers; H.C. Chen; D.C. Kopaska-Merkel; J.H. Fang (1995, desembre). "Predicting Permeability from Porositat Using Artificial Neural Networks". *American Association of Petroleum Geologists Bulletin* (vol. 79, pàg. 1786-1797).

1) En el primer escenari es van fer servir les dades procedents dels pous 1.877 i 1.928, com a conjunt d'entrenament; les del pou 1.802, com a conjunt de validació, i les del pou 1.930, com a pou que cal predir.

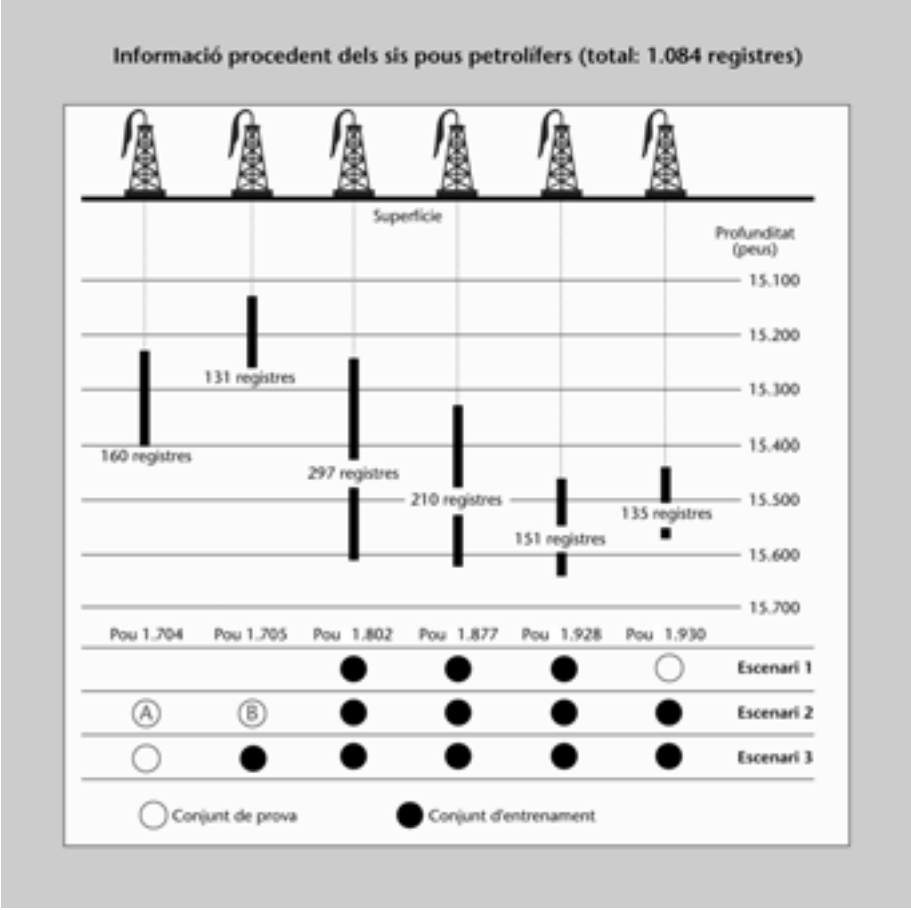
2) En el segon escenari, el conjunt d'entrenament estava format pels pous 1.802, 1.877 i 1.930; el conjunt de validació eren les dades procedentes del pou 1.928 i els pous que cal predir eren el 1.704 i el 1.705.

3) Finalment, en el tercer escenari els pous 1.928, 1.877 i 1.930 eren els que aportaven les dades del conjunt d'entrenament, el pou 1.705 contribuïa amb

les seves dades al conjunt de validació i el pou 1.704, com a conjunt de prova, era el pou sobre el qual es volia fer la predicció.

Una de les característiques que fan aquest problema prou difícil és la absència de dades, tant pel que fa a la porositat i la permeabilitat com pel que fa a la manca de seqüència entre les profundidats.

En la figura següent es poden veure les profunditats relatives dels sis pous (en peus), el nombre de casos que corresponen a cadascuna de les profunditats, i també els tres escenaris considerats:



Compararem dos tipus de model:

- Els obtinguts mitjançant la traducció a regles del mètode C4.5, que construeix inicialment arbres de decisió.

- Els obtinguts amb l'algorisme d'inducció de regles de classificació CN2.

2. Metodologia

Precisarem a continuació el tipus d'eines emprades en aquest cas i els diversos passos que s'han portat a terme.

2.1. Eines i mètodes

Per a les tasques d'anàlisi prèvia es va recórrer a les eines vSTATISTICA™, versió 5.0, de StatSoft Inc. i Statistics i Visualizer™ de Silicon Graphics. L'Scatter Visualizer™ de Silicon Graphics es va utilitzar per a fer l'anàlisi multidimensional de la conducta de les dades procedents dels sis pous.

La discretització es va fer amb el programari MineSet™, versió 2.01, una eina de l'empresa Silicon Graphics (Silicon Graphics, 1998) que conjuga mineria de dades amb eines de visualització multidimensional i ofereix, alhora, algunes eines de discretització i d'anàlisi de dades que cal trobar, per exemple, classificacions i associacions entre elements de la base de dades.

Com s'ha dit més amunt, es van utilitzar dos sistemes d'inducció de regles: CN2 (Clark i Niblett, 1989) i C4.5. Amb el seu ajut es van poder determinar i comparar les taxes d'error que corresponien a diversos tipus de discretització per a l'atribut de classe *Permeabilitat*.

Lectures complementàries

Amb relació a l'algorisme C4.5 consulteu les obres següents:

J.R. Quinlan (1987). "Generating Production Rules from Decision Trees". A: *Proceedings of 4th International Machine Learning Workshop* (pàg. 304-307). San Mateo (Califòrnia, EUA): Morgan Kaufmann Publishers.

J.R. Quinlan (1993). *C4.5 Programs for Machine Learning*. San Mateo (Califòrnia, EUA): Morgan Kaufmann Publishers.

1) El **mètode CN2** és un algorisme no incremental que pren un conjunt d'exemples i genera regles del tipus "si... aleshores..." per classificar els exemples.

2) El **mètode C4.5** és un generador d'arbres de decisió i la C4.5 crea regles del tipus "si... aleshores..." a partir de l'arbre de decisió resultant.

Els dos algorismes es van implementar utilitzant la llibreria MLC++, *Machine Learning Library in C++* (Kohavi i altres, 1994), un paquet de programari desenvolupat a la Universitat de Stanford que conté diversos algorismes d'aprenentatge automàtic i que permet no haver de canviar el format de les dades d'entrada quan es canvia d'algorisme. A més, ofereix mètodes normalitzats per portar a terme experimentació amb els diversos models obtinguts.

2.2. Fases del projecte

En aquest subapartat presentarem amb detall les fases del projecte, tal com les hem anat explicant al llarg de tot el curs.

2.2.1. Anàlisi preliminar i preparació de les dades

El domini es descriu mitjançant els atributs següents:

- **Permeabilitat**: es defineix com la facilitat amb la qual un gas o un líquid travessa un material a través dels seus porus quan està sotmès a pressió (mesurada en darcis).

- **Porositat**: és la propietat que té un material de contenir porus o intersticis (clivelles que separen les molècules d'un sòlid). Es defineix com la relació entre el volumen d'intersticis i el volum de la massa del material, i depèn del nombre, forma i distribució dels espais buits. S'expressa en forma percentual.

- **Profunditat**: expressada en peus, del nivell de perforació assolit en un pou.

Les dades originals sobre profunditat, porositat i permeabilitat es descriuen en la taula, que veiem en el subapartat següent, en forma d'atributs continus.

2.2.2. Selecció i neteja de dades

Les dades originals es presenten en la taula que veiem a continuació:

Dades utilitzables del conjunt inicial			
	Profunditat	**Porositat**	**Permeabilitat**
Total de registres	1.084	1.084	1.084
Valors absents	0	96	96
Valors perduts	0	6	6
Valors coneguts	1.084	982	982

D'aquest conjunt de dades, primer es van eliminar noranta-sis casos en què mancaven valors, tant per a la permeabilitat com per a la porositat. També es van eliminar sis casos considerats com a "perduts", que corresponien a casos en què l'instrument de mesura havia perdut els valors tant de porositat com de permeabilitat per a una profunditat determinada. En total, van quedar nou-cents vuitanta-dos casos per efectuar la primera anàlisi.

2.2.3. Anàlisi prèvia de dades

Per a veure què diuen les dades i tenir una millor comprensió del domini, es va desenvolupar una primera fase d'estadística descriptiva i es van trobar els

paràmetres següents: la mitjana, la desviació estàndard, la moda, i els valors mínim i màxim.

En la taula següent es presenten els resultats obtinguts aplicant STATISTICA™ sobre el conjunt dels mil vuitanta-quatre registres de dades:

Distribució dels valors nuls i inferiors a 0,01 per a la permeabilitat				
Pou	**Casos**	**Valor 0**	**Valors < 0,01**	**%**
1.704	159	66	0	41,5
1.705	129	0	78	60,4
1.802	249	121	0	48,6
1.877	182	0	58	31,8
1.928	130	0	38	29,2
1.930	133	0	15	11,3
Total	982	187	189	38,3

Es va observar que les dades originals mostraven força valors de permeabilitat més petits de 0,01, i que molts altres valors eren igual a zero.

Els valors nuls i inferiors a 0,01 es van substituir pel valor 0,009. Per tant, cal interpretar els resultats que s'obtenen tant amb CN2 com amb C4.5 en què apareix el valor *Permeabilitat* = 0,009 com a *Permeabilitat* < 0,01.

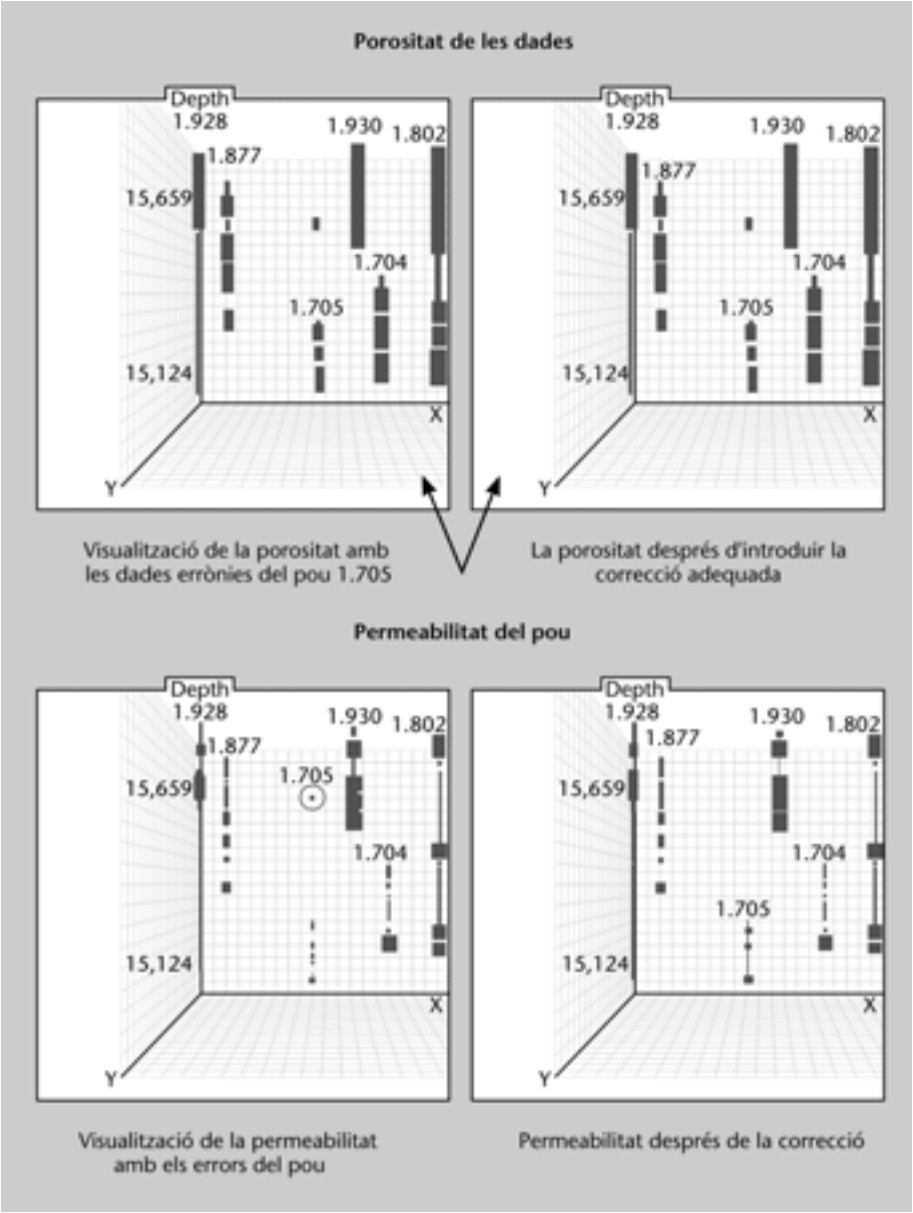
2.2.4. Visualització de dades

La possibilitat de visualitzar les dades prèviament permet una millor comprensió del domini. Potser és tant o més important disposar d'un coneixement previ que ens permeti de saber quin tipus d'eina de visualització cal aplicar per a obtenir el tipus de gràfic que retorni millor informació.

Vam emprar Scatter Visualizer® per a obtenir una primera visualització de les dades. Els quadrats dels gràfics representen els valors dels atributs. Un quadrat petit indica que l'atribut corresponent té un valor petit i un quadrat gran, el contrari. En la figura que presentem més avall es pot veure les gràfiques que corresponen a l'anàlisi de la permeabilitat i la porositat en relació amb la profunditat. Es pot apreciar la clara absència de dades que mostra el pou 1.705.

Vam tornar a analitzar les dades originals i vam poder concloure que segurament s'hi havia introduït un error a causa de la seqüència de valors que mostrava i dels valors que tenien altres atributs. Observeu que després d'una

profunditat de 1.520 peus, el valor següent que hi apareix és de 1.521 peus, i després hi ha un salt de 1.541 i 1.542:



Si s'analitzen els gràfics que apareixen en les figures anteriors es pot observar una gran entropia entre els valors de la permeabilitat i els de la porositat que tenen un comportament més homogeni.

També es pot observar que el pou 1.930 té una porositat i una permeabilitat més altes que les dels altres pous. És important remarcar que els pous 1.705 i 1.802 tenen valors de permeabilitat més baixos, però molta més porositat respecte als altres pous.

Un cop efectuada la neteja i la correcció de dades es va procedir a una anàlisi estadística de les dades (mitjana, mediana, desviació estàndard, histo-

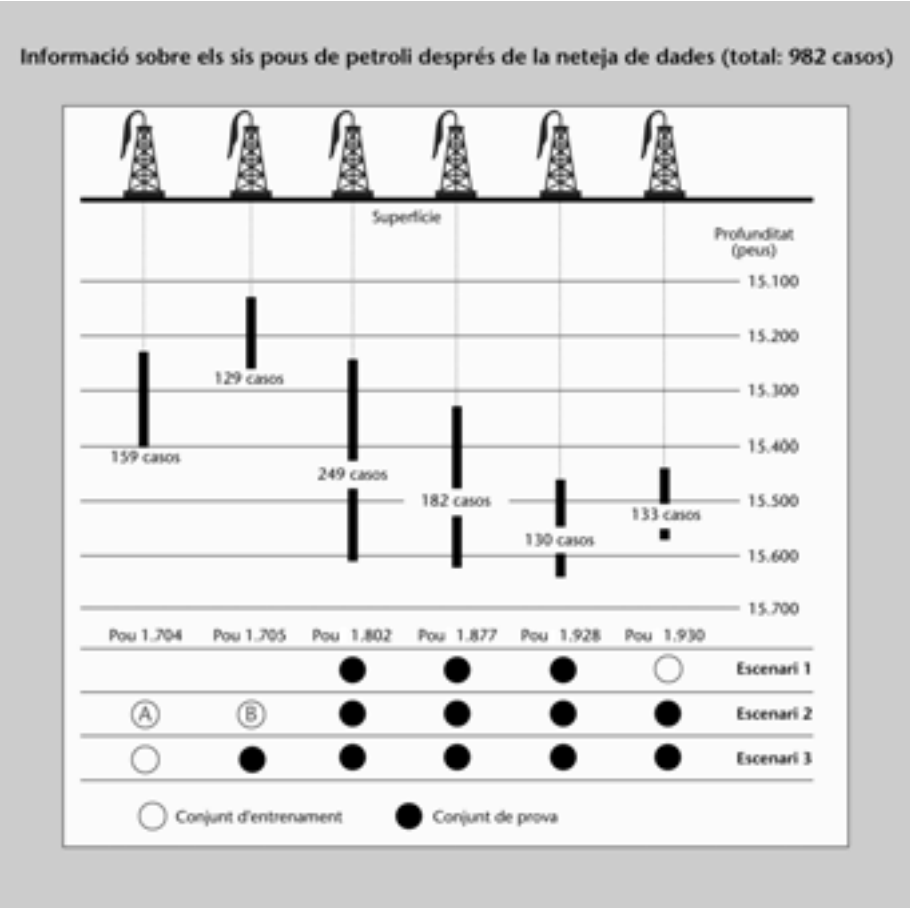
grames i quartils) que es mostra en forma gràfica en la taula que presentem a continuació:

Detecció i correcció d'errors en les dades del pou 1.705					
Dades originals			Dades corregides		
Profunditat	Porositat	Permeabilitat	Profunditat	Porositat	Permeabilitat
15.217	2	< 0,01	15.217	2	0,009
15.218	0,7	< 0,01	15.218	0,7	0,009
15.219	0,7	< 0,01	15.219	0,7	0,009
15.220	0,6	< 0,01	15.220	0,6	0,009
15.521	0,7	< 0,01	15.221	0,7	0,009
15.522	0,7	< 0,01	15.222	0,7	0,009
15.523	1,3	< 0,01	15.223	1,3	0,009
15.524	1,1	< 0,01	15.224	1,1	0,009
...	...	...	...	...	...
...	...	...	...	...	...
15.538	9,4	0,9	15.238	9,4	0,9
15.539	9,8	0,8	15.239	9,8	0,8
15.540	7,8	0,05	15.240	7,8	0,05
15.541	0,8	< 0,01	15.241	0,8	0,009
15.242	7,1	0,57	15.242	7,1	0,57
15.243	8,3	0,05	15.243	8,3	0,05
15.244	15,1	0,34	15.244	15,1	0,34
15.245	11,7	0,29	15.245	11,7	0,29

Mostrem els resultats d'aplicar estadística descriptiva amb Statistics Visualizer™ a nou-cents vuitanta-dos casos en la figura següent:

Descripció estadística dels sis pous (total: 982 casos)			
	Profunditat	Porositat	Permeabilitat
Valors diferents	500	213	142
Valors [Max.,Min.]	[15.568, 15.124]	[34,1,0,6]	[55,0,009]
Mitjana	15.415 ± 127,9	8,2 ± 6,5	2,5 ± 6,7
Mediana	15.439	8,2	0,05

Només es presenten els resultats que corresponen als nou-cents vuitanta-dos casos resultants de la selecció i neteja de dades.



2.2.5. Preparació de dades: procés de discretització

Un cop es van fer l'anàlisi inicial de dades i les correccions de dades que hem esmentat anteriorment, es va procedir a discretitzar l'atribut de classe *Permeabilitat*. D'aquesta manera, ja estàvem en condicions d'aplicar els algorismes CN2 i C4.5.

Vam utilitzar tres dels mètodes de discretització que ofereix MineSet™:

1) Discretització *stand-alone* de l'atribut de classe *Permeabilitat* basada en la freqüència de distribució.

2) Un mètode de discretització empírica suggerit per un petrolier expert que discretitza l'atribut de classe *Permeabilitat* segons els valors de la porositat.

3) Un mètode híbrid que intenta millorar la discretització suggerida per l'expert respecte a la precisió dels algorismes CN2 i C4.5.

Primer mètode de discretització

El nombre d'intervals que es van generar per a l'atribut de classe *Permeabilitat* es va especificar d'entrada sense considerar els altres atributs (porositat i profunditat) i es va mirar d'obtenir una bona distribució de les dades entre tots els intervals. Per a obtenir la discretització es van emprar dades dels sis pous de petroli (nou-cents vuitanta-dos casos) per a poder fer les comparacions correctes per als tres escenaris de prova. Es van obtenir, doncs, tres discretitzacions amb dos, tres i cinc grups.

Discretització suggerida per l'expert

Segons l'expert, cada terreny on s'efectua una perforació és diferent, amb la qual cosa resulta molt difícil determinar quins són els millors intervals de discretització per a cadascun dels tres escenaris. L'expert que aporta la millor discretització és el que coneix molt bé una classe particular de terreny, atès que els valors de permeabilitat són diferents per a cada cas.

Segons l'expert, també és important saber que en la majoria de casos una porositat alta en les roques és indicativa de l'existència de petroli. Per tant, segons l'expert, per a establir els criteris de discretització és convenient considerar la porositat del pou. Tot i que l'expert no va determinar els valors de discretització, sí que va suggerir discretitzar la permeabilitat utilitzant el valor corresponent de la porositat com a factor de ponderació. MineSet™ ofereix la possibilitat de discretitzar un atribut segons el pes d'un altre atribut. Els resultats de les diferents discretitzacions obtingudes amb la utilització d'un mètode *stand-alone* amb pesos uniformes i el que suggeria l'expert es mostren en la taula següent:

Intervals de discretització obtinguts per dos mètodes diferents

	Discretització *stand-alone*	**Discretització suggerida per l'expert**
2 intervals	< 0,055 > 0,055	< 0,445 > 0,445
3 intervals	< 0,0095 [0,0095, 0,235] > 0,235	< 0,155 [0,155, 2,4] > 2,4
	Discretització *stand-alone*	**Discretització suggerida per l'expert**
5 intervals	< 0,0095 [0,0095, 0,075] [0,075, 0,245] [0,245, 2,4] > 0,235	< 0,055 [0,055, 0,215] [0,215, 1,05] [1,05, 7,85] > 7,85

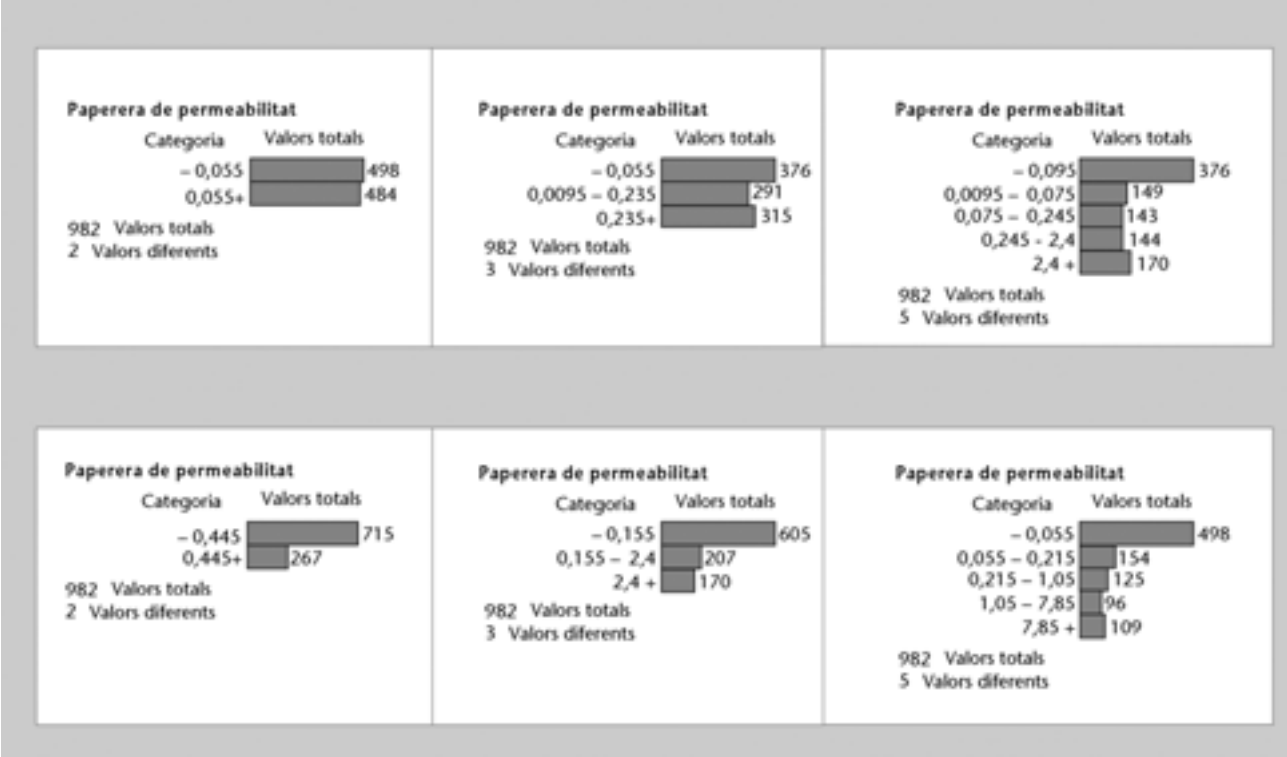
Lectura complementària

Consulteu altres experiments semblants en l'obra següent:

J. Dougherty; R. Kohavi; M. Sahami (1995). "Supervised and Unsupervised Discretizations of Continuous Features". A: *Proceedings of the 12th International Conference on Machine Learning* (pàg. 194-202). Morgan Kaufmann Publishers.

En la figura següent es presenten les diferents distribucions dels valors de discretització segons els dos mètodes comentats. L'algorisme de discretització uti-

litzat va ser el que es basa en l'entropia, perquè havia presentat els millors resultats en altres experiments semblants:



Farem una discussió més profunda dels resultats obtinguts amb aquestes discretitzacions més endavant.

També es van aplicar altres criteris de discretització. En concret es va aplicar un algorisme automàtic que intenta trobar el nombre d'intervals. El resultat va ser la proposta de dotze intervals discrets que generaven una gran quantitat d'errors en aplicar els algorismes CN2 i C4.5.

Lectures complementàries

Podeu consultar els mètodes CN2 i C4.5, respectivament, en les obres següents:

P. Clark; T. Niblett; (1989). "The CN2 Induction Algorithm". *Machine Learning* (vol. 3, pàg. 261-283).

J. Dougherty; R. Kohavi; M. Sahami (1995). "Supervised and Unsupervised Discretizations of Continuous Features". A: *Proceedings of the 12th International Conference on Machine Learning* (pàg. 194-202). Morgan Kaufmann Publishers.

2.2.6. Mineria de dades: models de classificació

Es van emprar els mètodes CN2 i C4.5 per a determinar quines taxes d'errors s'obtenien amb les diferents discretitzacions proposades per a l'atribut de classe *Permeabilitat*. Tots dos algorismes es van implementar amb MLC++, *Machine Learning Library in C++* (Kohavi i altres, 1994). En tots dos casos els algorismes es van executar amb els paràmeteres per defecte de la llibreria MLC++. Les taxes d'error per a cada algorisme i escenari, i també l'error de classificació del conjunt d'entrenament es mostren en les tres taules que veiem a continuació.

Dels resultats es pot inferir que només en un conjunt petit de casos la taxa d'error és relativament baixa per a ambdós mètodes (CN2 i C4.5). Es pot veure tam-

bé que hi ha una gran variabilitat en les taxes d'error. Tant en l'escenari 1, amb tres i cinc discretitzacions, com en l'escenari 2, amb les discretitzacions indicades per l'expert, la taxa d'error és més alta que la taxa d'error de classificació majoritària. Això és del tot inacceptable.

Taxes d'error per a l'escenari 1						
	Discretització *stand-alone*			**Discretització suggerida per l'expert**		
Nombre de discretitzacions	**Taxa d'error amb CN2**	**Taxa d'error C4.5**	**Taxa d'error de la classe majoritària**	**Taxa d'error amb CN2**	**Taxa d'error amb C4.5**	**Taxa d'error de la classe majoritària**
2	12,8%	22,6%	47,2%	23,3%	13,5%	26,7%
3	36,1%	27,1%	61,3%	42,1%	28,6%	36,5%
5	46,6%	41,4%	61,3%	67,7%	68,4%	47,2%

Taxes d'error per a l'escenari 2						
	Discretització *stand-alone*			**Discretització suggerida per l'expert**		
Nombre de discretitzacions	**Taxa d'error amb CN2**	**Taxa d'error C4.5**	**Taxa d'error de la classe majoritària**	**Taxa d'error amb CN2**	**Taxa d'error amb C4.5**	**Taxa d'error de la classe majoritària**
2 c. prova *A*	13,8%	6,9%	47,3%	8,8%	10,1%	30,8%
2 c. prova *B*	13,2%	9,3%	47,3%	6,2%	5,4%	30,8%
3 c. prova *A*	17,0%	16,4%	64,7%	**30,2%**	**17,0%**	**42,2%**
3 c. prova *B*	0,8%	7,8%	64,7%	5,4%	7,8%	42,2%
5 c. prova *A*	28,9%	27,7%	66,6%	34,0%	32,7%	52,7%
5 c. prova *B*	22,5%	22,5%	66,6%	25,6%	19,4%	52,7%

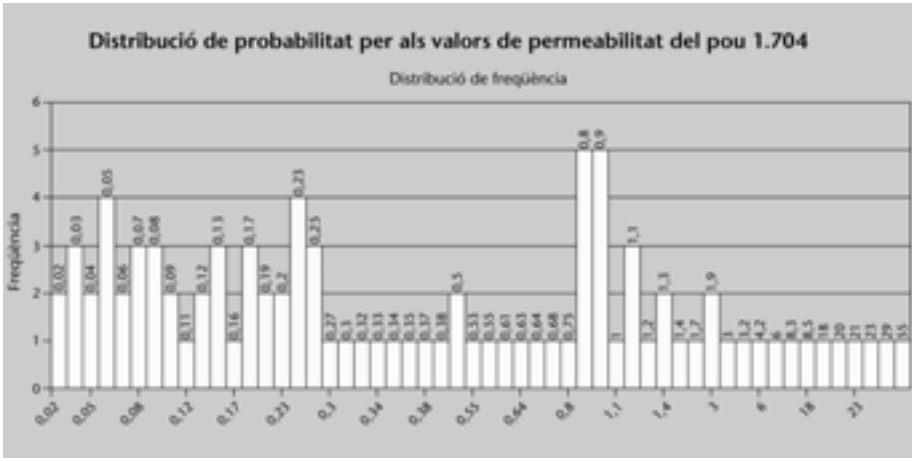
Taxes d'error per a l'escenari 3						
	Discretització *stand-alone*			**Discretització suggerida per l'expert**		
Nombre de discretitzacions	**Taxa d'error amb CN2**	**Taxa d'error C4.5**	**Taxa d'error de la classe majoritària**	**Taxa d'error amb CN2**	**Taxa d'error amb C4.5**	**Taxa d'error de la classe majoritària**
2	14,5%	6,9%	48,8%	8,8%	10,1%	27,2%
3	21,4%	15,7%	62,3%	29,6%	18,2%	37,8%
5	27,7%	25,2%	62,3%	28,9%	32,1%	48,8%

Ens concentrem ara a analitzar un escenari concret, amb una discretització fixa, per tal de millorar la classificació i obtenir regles de classificació d'una qualitat més alta.

Basant-nos en les consideracions prèvies, vam escollir l'escenari 2 (amb tres discretitzacions donades per l'expert amb el conjunt de prova *A*). La matriu de confusió per a aquests supòsits es presenta en la taula següent.

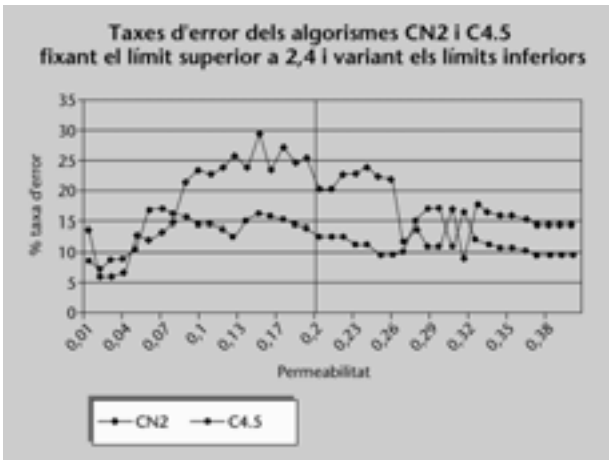
Matriu de confusió per a l'escenari 2 amb tres discretitzacions amb ajuda de l'expert utilitzant el conjunt de prova A										
Discretització	**CN2**					**C4.5**				
		(a)	(b)	(c)	Error		(a)	(b)	(c)	Error
< 0,155 [0,155, 2,4] > 2,4	(a)	90	3	0	3,2%	(a)	92	1	0	1,0%
	(b)	41	9	14	83,3%	(b)	25	28	1	48,1%
	(c)	0	0	12	0%	(c)	0	0	12	0%

La distribució dels valors de permeabilitat dels valors del pou 1.704, que era el que es va utilitzar com a conjunt de prova d'aquest escenari, es mostren en la figura següent. Al valor 0,009 li corresponen seixanta-sis elements que no es mostren en la gràfica.

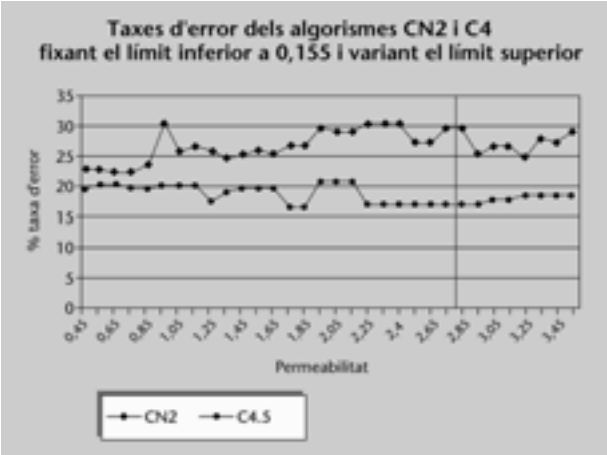


Per a millorar la precisió dels resultats tant del CN2 com de C4.5 vam fixar un dels límits de l'interval a un valor prefixat i vam fer variar l'altre límit amb valors més alts i més baixos. Per a cada variació es van determinar les taxes d'error de cadascun dels algorismes.

Vam començar per fixar un límit superior de 2,4 i per fer variar el límit inferior (0,15). En la figura següent es poden veure els resultats obtinguts. Les taxes d'error més baixes es van obtenir amb l'algorisme CN2, fixant el límit inferior a 0,015, 0,025 i 0,035. Per al cas de l'algorisme C4.5, les taxes d'error més baixes es van obtenir amb límits inferiors de 0,015, 0,025 i 0,035.



Després vam fixar el límit inferior a 0,155 i vam fer variar el límit superior fent servir valors més alts i més baixos. En la figura següent mostrem els resultats obtinguts. Amb CN2 els valors 0,65 i 0,75 van donar les taxes d'error més baixes. Per al cas del C4.5 els valors que corresponien a la taxa d'error més baixa van ser 1,75 i 1,85. És important remarcar en aquest moment que només es va fer servir el valor 1,75 per a les altres proves, i no pas el valor 1,85, atès que els dos valors donaven resultats semblants.



Utilitzant com a referència les taxes d'error mínimes es van escollir els millors límits inferiors i superiors propers (però no iguals) a 0,015 i 2,4. Un cop fet això, es van generar les matrius de confusió per a cada cas per a intentar determinar quin era l'interval amb la taxa d'error mínima i considerant la distribució dels conjunts de dades. En aquest escenari es van fer servir sis-cents noranta-quatre casos d'entrenament i cent cinquanta-nou de prova.

Les matrius de confusió pel mètode CN2, que empren les taxes d'error mínimes d'aquest algorisme, es mostren en l'annex 1. Les que corresponen al mètode C4.5 es poden veure en l'annex 2.

Vegeu les matrius de confusió pel mètode CN2 en l'annex I i les que corresponen al mètode C4.5, a l'annex 2, al final d'aquest mòdul.

Basant-nos en aquests resultats vam poder establir que les discretitzacions amb taxa d'error mínima i distribució relativa a les seves dades eren 0,015 i 0,75 i 0,015 i 1,75. Atès que vam considerar que aquests punts de tall donaven un bon resultat en aquest escenari, es van utilitzar en d'altres per a determinar-ne la conducta. Les matrius de confusió per a aquests dos nous intervals de discretització es mostren en l'annex 3.

Vegeu les matrius de confusió per als nous intervals de discretització en l'annex 3 que hi ha al final d'aquest mòdul.

3. Resultats

La taula següent mostra la comparació entre els resultats de la discretització suggerida per l'expert i l'obtinguda amb els punts de 0,015 i 1,75, que és amb la que es van obtenir els millors resultats per a tots els escenaris.

Comparació de la discretització suggerida per l'expert amb la discretització amb punts de tall de 0,015 i 1,75						
	Discretització suggerida per l'expert amb [0,155, 2,4]			Discretització usant els punts de tall [0,015, 1,75]		
Escenari	Taxa d'error amb CN2	Taxa d'error amb C4.5	Taxa d'error de la classe majoritària	Taxa d'error CN2	Taxa d'error amb C4.5	Taxa d'error de la classe majoritària
1	42,1%	28,6%	36,5%	38,3%	39,8%	64,4%
2 c. prova *A*	30,2%	17,0%	42,2%	13,8%	6,9%	59,9%
2 c. prova *B*	5,4%	7,8%	42,2%	3,9%	3,1%	59,9%
3	29,6%	18,2%	37,8%	8,2%	5,7%	56,7%

Es pot observar que, a excepció de l'escenari 1, la nova discretització millora la precisió en la classificació.

L'algorisme C4.5 va generar deu regles a partir del conjunt d'entrenament de l'escenari 2 (sis-cents noranta-quatre casos). Aquestes regles no cobreixen els noranta-nou casos dels sis-cents noranta-quatre casos totals (14,3%).

Vegeu les regles generades a partir de l'algorisme C4.5 en l'annex 4 que trobareu al final d'aquest mòdul.

Presentem una mica més avall el resultat d'aplicar tres regles al conjunt de prova.

a) Regla 20

Porositat > 15.7 ⇒
⇒ classe > 1,75 [11,1%] [112 14] [14,3%] [12 2]

b) Regla 1

Porositat ≤ 2,9 ⇒
⇒ classe < 0,015 [0,0%] [203 0] [1,6%] [60 1]

c) Regla 19

Classe > 3,9 ⇒
⇒ classe ≤ 15,7
⇒ classe [0,015:1,75] [17,5%] [188 40] [7,3%][76 6]

La conclusió de cada regla mostra la classe, a més d'informació en el format següent:

- [AA%]: error de classificació en aplicar la regla al conjunt d'entrenament.
- [BB i CC]: nombre d'exemples classificats correctament [BB] i classificats incorrectament [CC] dins el conjunt d'entrenament.
- [DD%]: error de classificació en aplicar la regla al conjunt de prova.
- [EE i FF]: nombre d'exemples classificats correctament [EE] i classificats incorrectament [FF] dins el conjunt de prova.

Amb el mateix escenari i conjunt d'entrenament CN2, va generar cent vint regles que deixen sense cobrir divuit casos (2,6%). Ara bé, més del 50% d'aquestes regles són especialitzacions per a poder cobrir un, dos o tres exemples.

3.1. Valoració

Els resultats indiquen que la tria d'un mètode de discretització o un altre afecta directament els mètodes CN2 i C4.5.

Per al cas de la discretització *stand-alone*, l'algorisme C4.5 ha donat millors resultas en el sentit que la precisió és millor que la del CN2 en un 83,3% dels casos, i la mateixa en un 8,3%.

Si s'empra una discretització que recull les indicacions de l'expert C4.5 s'obté un resultat millor que el del CN2 en un 41,6% dels casos, i en un 25% presenta els mateixos resultats.

A més, cal tenir en compte la matriu de confusió per a poder analitzar la distribució dels elements. Per exemple, es mostra una precisió del 100% amb CN2 per a dotze elements del conjunt *C*, però això no té cap rellevància si ho comparem amb el conjunt *A* amb CN2, que presenta una precisió del 96,8% per a norantatres elements.

La discretització híbrida augmenta la precisió en tots els escenaris menys en l'escenari 1.

Tot i la gran millora, el paper de l'expert és fonamental en la definició de la discretització. La discretització amb mètodes no supervisats ofereix a l'expert una bona ajuda per determinar els millors intervals. Finalment, el nombre d'intervals també pot influir en la precisió dels mètodes de mineria de dades aplicats. El mètode híbrid tendeix a incrementar la complexitat en incloure més intervals.

Bibliografia

Breiman, L.; Friedman, J.; Olshen, R.; Stone, C. (1993). *Classification and Regression Trees*. Nova York: Chapman and Hall.

Catlett, J. (1991). *Megainduction: Machine Learning on Very Large Databases*. Tesi doctoral. Universitat de Sydney.

Chmielewski, M.; Grzymala-Busse, J. (1995). "Global Discretization of Continuous Attributes as Preprocessing for Machine Learning". A: T.Y. Lin; A.M. Wildberger (ed.). *Soft Computing, Society for Computer Simulation* (pàg. 294-301). San Diego.

Clark, P.; Niblett, T.; (1989). "The CN2 Induction Algorithm". *Machine Learning* (vol. 3, pàg. 261-283).

Dougherty, J.; Kohavi, R.; Sahami, M. (1995). "Supervised and Unsupervised Discretizations of Continuous Features". A: *Proceedings of the 12th International Conference on Machine Learning* (pàg. 194-202). Morgan Kaufmann Publishers.

Fayyad, U.M.; Irani, K.B. (1993). "Multi-interval discretization of continuous-valued attributes for classification learning". A: *Proceedings of the 13th International Conference on Machine Learning* (pàg. 1.022-1.027). Morgan Kaufmann Publishers.

Fayyad, U.M.; Piatesky-Shapiro, G.; Smyth, P.; Uthurusamy, R. (1996). *Advanced in Knowledge and Data Mining*. Menlo Park (Califòrnia): AAAI / MIT Press.

Grzymala-Busse, W. (1992). "Lers - A System for Learning from Examples Based on Rough Sets". A: R. Slowinski (ed.). *Intelligence Decision Support. Handbook of Applications and Advances of the Rough Set Theory* (pàg. 3-18). Dordrecht: Kluwer Academic Publishers.

Kerber, R. (1992). "ChiMerge: Discretization of Numeric Attributes". *Proceedings of the 10th National Conference on Artificial Intelligence* (pàg. 123-127).

Kohavi, R.; John, G.; Long, R.; Manley, D.; Pfleger, K. (1994). "MLC++: A Machine Learning Library in C++". A: *Tool with Artificial Intelligence* (pàg. 740-743). IEEE Computer Society Press.

Lenarcik, A.; Piasta, Z. (1992). "Discretization of Condition Attribute Space". A: R. Slowinski (ed.). *Intelligence Decision Support. Handbook of Applications and Advances of the Rough Set Theory* (pàg. 373-389). Dordrecht: Kluwer Academic Publishers.

Lenarcik, A.; Piasta, Z. (1993). "Probabilistic Approach to Decision Algorithm Generation in the Case of Continuous Condition Attributes". *Foundations of Computing and Decision Sciences* (vol. 18, núm. 3-4, pàg. 213-224). Poznan.

Lenarcik, A.; Piasta, Z. (1995). "Minimizing the Number of Rules in Deterministic Rough Classifiers". A: T.Y. Lin; A.M. Wildberger (ed.). *Soft Computing* (pàg. 32-35). San Diego: Society for Computer Simulation.

Molina, F.L.C.; Oliveira, R.S.; Doi, C.Y.; Paula, M.F. de; Romanato, M.J. (1998). "MLC++: Biblioteca de Aprendizado de Máquina em C++". *Technical Report 72*. São Paulo: ICMSC, USP.

Nguyen, S.H.; Skowron, A. (1995). "Quantization of Real Value Attributes: Rough Set and Boolean Reasoning Approach". *Proceedings of the Second Joint Annual Conference on Information Sciences* (pàg. 34-37). Wrightsville Beach.

Pfahringer, B. (1995)."Compression-based discretization of continuous attributes". A: A. Prieditis; S. Russel (ed.). *Proceedings of the 12th International Conference on Machine Learning*. Morgan Kaufmann Publishers.

Quinlan, J.R. (1987). "Generating Production Rules from Decision Trees". A: *Proceedings of 4th International Machine Learning Workshop* (pàg. 304-307). San Mateo. Morgan Kaufmann.

Quinlan, J.R. (1990). "Induction of Decision Trees". A: J.W. Shavlik; T.G. Dietterich (ed.). *Readings in Machine Learning* (pàg. 57-69). San Mateo: Morgan Kaufmann Publishers.

Quinlan, J.R. (1993). *C4.5 Programs for Machine Learning*. San Mateo: Morgan Kaufmann Publishers.

Rogers, S.J.; Chen, H.C.; Kopaska-Merkel, D.C.; Fang, J.H. (1995, desembre). "Predicting Permeability from Porositat Using Artificial Neural Networks". *American Association of Petroleum Geologists Bulletin* (vol. 79, pàg. 1786-1797).

Silicon Graphics. *MineSet* (1997). http://www.sgi.com/Products/software/MineSet/.

Ventura, D.; Martinez, T.R. (1994). "BRACE: A Paradigm For the Discretization of Continuously Valued Data". A: *Proceedings of the 7th Florida Artificial Intelligence Research Symposium* (pàg. 117-121).

Ventura, D.; Martinez, T.R. (1995). "An Empirical Comparison of Discretization Methods". A: *Proceedings of the 10th International Symposium on Computer and Information Sciences* (pàg. 443-450).

Water Resource Research Center (1998). *Glossary of Organizations and Acronyms*. College of Agriculture. The University of Arizona. En línia. http://Ag.Arizona.Edu/Azwater/ Gloss.Html.

Annexos

Annex 1

Taxes d'error mínimes per a l'algorisme CN2

Interval	CN2					C4.5					
		(a)	(b)	(c)	Error (%)		(a)	(b)	(c)	Error (%)	Taxa d'error de la classe majoritària (%)
< 0,015	(a)	61	5	0	7,5	(a)	60	6	0	9,1	
[0,015,0,65]	(b)	11	41	5	28	(b)	1	50	6	12,2	
> 0,65	(c)	4	4	28	22,2	(c)	0	7	29	19,4	
Total					18,2					12,6	59,9

Interval	CN2					C4.5					
		(a)	(b)	(c)	Error (%)		(a)	(b)	(c)	Error (%)	Taxa d'error de la classe majoritària (%)
< 0,015	(a)	63	3	0	4,5	(a)	60	6	0	9,1	
[0,015,0,75]	(b)	2	58	0	3,3	(b)	1	54	5	10	
> 0,75	(c)	0	18	15	54,5	(c)	0	7	26	21,2	
Total					14,5					11,9	59,9

Interval	CN2					C4.5					
		(a)	(b)	(c)	Error (%)		(a)	(b)	(c)	Error (%)	Taxa d'error de la classe majoritària (%)
< 0,025	(a)	64	4	0	5,8	(a)	61	7	0	10,2	
[0,025,0,65]	(b)	15	35	5	36,3	(b)	0	49	6	10,9	
> 0,65	(c)	2	6	28	22,2	(c)	0	7	29	19,4	
Total					20,1					12,6	57,8

Interval	CN2					C4.5					
		(a)	(b)	(c)	Error (%)		(a)	(b)	(c)	Error (%)	Taxa d'error de la classe majoritària (%)
< 0,025	**(a)**	62	6	0	8,8	**(a)**	61	7	0	10,2	
[0,025,0,75]	**(b)**	8	44	6	24,1	**(b)**	0	53	5	8,6	
> 0,75	**(c)**	0	7	26	21,2	**(c)**	0	7	26	21,2	
Total					17					11,9	57,8

Interval	CN2					C4.5					
		(a)	(b)	(c)	Error (%)		(a)	(b)	(c)	Error (%)	Taxa d'error de la classe majoritària (%)
< 0,035	**(a)**	65	6	0	8,4	**(a)**	61	10	0	14	
[0,035,0,65]	**(b)**	7	40	5	23	**(b)**	0	46	6	11,5	
> 0,65	**(c)**	1	7	28	22,2	**(c)**	0	7	29	19,4	
Total					16,4					14,5	56,5

Interval	CN2					C4.5					
		(a)	(b)	(c)	Error (%)		(a)	(b)	(c)	Error (%)	Taxa d'error de la classe majoritària (%)
< 0,035	**(a)**	65	6	0	8,4	**(a)**	61	10	0	14	
[0,035,0,75]	**(b)**	5	45	5	18,1	**(b)**	0	54	1	1,8	
> 0,75	**(c)**	0	7	26	21,2	**(c)**	0	12	21	36,3	
Total					14,5					14,5	56,5

Annex 2

Taxes d'error mínimes per a l'algoritme C4.5-*rules*.

Interval	CN2					C4.5					
		(a)	(b)	(c)	Error (%)		(a)	(b)	(c)	Error (%)	Taxa d'error de la classe majoritària (%)
< 0,015	**(a)**	62	4	0	6	**(a)**	60	6	0	9	
[0,015,1,75]	**(b)**	11	63	5	20,2	**(b)**	1	76	2	3,7	
> 1,75	**(c)**	1	1	12	14,2	**(c)**	0	2	12	14,2	
Total					13,8					6,9	59,9

Interval	CN2					C4.5					
		(a)	(b)	(c)	Error (%)		(a)	(b)	(c)	Error (%)	Taxa d'error de la classe majoritària (%)
< 0,025	**(a)**	62	6	0	8,8	**(a)**	61	7	0	10,2	
[0,025,1,75]	**(b)**	17	54	6	29,8	**(b)**	0	75	2	2,5	
> 1,75	**(c)**	1	1	12	14,2	**(c)**	0	2	12	14,2	
Total					19,5					6,9	57,8

Interval	CN2					C4.5					
		(a)	(b)	(c)	Error (%)		(a)	(b)	(c)	Error (%)	Taxa d'error de la classe majoritària (%)
< 0,035	**(a)**	63	8	0	11,2	**(a)**	61	10	0	14	
[0,035,1,75]	**(b)**	16	52	6	29,7	**(b)**	0	72	2	2,7	
> 1,75	**(c)**	0	2	12	14,2	**(c)**	0	2	12	14,2	
Total					20,1					8,8	56,5

Annex 3

Matrius de confusió corresponents als valors 0,015 i 0,75 i 0,015 i 1,75.

Escenari	CN2					C4.5					
		(a)	(b)	(c)	Error (%)		(a)	(b)	(c)	Error (%)	Taxa d'error de la classe majoritària (%)
1	**(a)**	15	5	2	31,8	**(a)**	14	7	1	36,3	
	(b)	10	23	18	54,9	**(b)**	2	32	17	37,2	
	(c)	4	10	46	23,3	**(c)**	0	14	46	23,3	
Total					36,8					30,8	54,4

Matrius de confusió amb els punts de tall 0,015 i 0,75 en tots els escenaris.

Escenari	CN2					C4.5					
		(a)	(b)	(c)	Error (%)		(a)	(b)	(c)	Error (%)	Taxa d'error de la classe majoritària (%)
2 B	**(a)**	78	0	0	0	**(a)**	77	1	0	1,2	
	(b)	5	35	4	20,4	**(b)**	1	39	4	11,3	
	(c)	2	2	3	57,1	**(c)**	0	5	2	71,4	
Total					10,1					8,5	59,9

Escenari	CN2					C4.5					
		(a)	(b)	(c)	Error (%)		(a)	(b)	(c)	Error (%)	Taxa d'error de la classe majoritària (%)
2 B	**(a)**	62	4	0	6	**(a)**	62	4	0	6	
	(b)	7	46	7	23,3	**(b)**	1	58	1	3,3	
	(c)	0	7	26	21,2	**(c)**	0	12	21	36,3	
Total					15,7					11,3	56,7

Matrius de confusió amb els punts de tall 0,015 i 1,75 en tots els escenaris.

Escenari	CN2					C4.5					
		(a)	(b)	(c)	Error (%)		(a)	(b)	(c)	Error (%)	Taxa d'error de la classe majoritària (%)
1	**(a)**	17	4	1	22,7	**(a)**	16	2	4	27,2	
	(b)	14	30	16	50	**(b)**	8	21	31	65	
	(c)	1	15	35	31,3	**(c)**	0	8	43	15,6	
Total					38,3					39,8	54,4

Escenari	CN2					C4.5					
		(a)	(b)	(c)	Error (%)		(a)	(b)	(c)	Error (%)	Taxa d'error de la classe majoritària (%)
2 B	**(a)**	77	1	0	1,2	**(a)**	77	1	0	1,2	
	(b)	1	47	0	2	**(b)**	1	47	0	2	
	(c)	1	2	0	0	**(c)**	0	2	1	66,6	
Total					3,9					3,1	59,9

Escenari	CN2					C4.5					
		(a)	(b)	(c)	Error (%)		(a)	(b)	(c)	Error (%)	Taxa d'error de la classe majoritària (%)
3	**(a)**	62	4	0	6	**(a)**	62	4	0	6	
	(b)	6	72	1	8,8	**(b)**	1	76	2	3,7	
	(c)	0	2	12	14,2	**(c)**	0	2	12	14,2	
Total					8,2					5,7	56,7

Annex 4

Regles generades per l'algorisme C4.5-*rules*

```
C4.5 [release 8] rule generatorThu Jul  9 03:04:38 1998
----------------------------------------
   Options:
      Rulesets evaluated on unseen cases
      File stem </var/tmp/AAAa000ox>

Read 694 cases (2 attributes) from /var/tmp/AAAa000ox
----------------------------------------
Processing tree 0
Final rules from tree 0:

Rule 20:
      porositat > 15.7
      ->  class 1.75+  [86.4%]

Rule 18:
      profunditat > 15599
      porositat > 3.9
      ->  class 1.75+  [79.4%]

Rule 12:
      profunditat > 15426
      profunditat <= 15496
      porositat > 10.8
      ->  class 1.75+  [78.7%]

Rule 17:
      profunditat > 15582
      porositat > 11.1
      ->  class 1.75+  [64.5%]

Rule 1:
      porositat <= 2.9
      ->  class - 0.015  [99.3%]

Rule 4:
      profunditat > 15412
      porositat <= 3.9
      ->  class - 0.015  [96.3%]
```

```
Rule 16:
      profunditat > 15582
      profunditat <= 15599
      porositat <= 11.1
      ->  class - 0.015  [86.7%]

Rule 14:
      profunditat > 15515
      porositat <= 10.3
      ->  class - 0.015  [77.2%]

Rule 19:
      porositat > 3.9
      porositat <= 15.7
      ->  class 0.015-1.75  [65.3%]

Rule 3:
      porositat > 2.9
      porositat <= 3.2
      ->  class 0.015-1.75  [54.6%]

Default class: 0.015-1.75

Evaluation on training data (694 items):
```

Rule	Size	Error	Used	Wrong	Advantage	Class
20	1	13,6%	126	14 (11,1%)	53 (64|11)	> 1,75+
18	2	20,6%	4	0 (0,0%)	4 (4|0)	> 1,75+
12	3	21,3%	33	11 (33,3%)	11 (22|11)	> 1,75+
17	2	35,5%	6	2 (33,3%)	3 (4|1)	> 1,75+
1	1	0,7%	203	0 (0,0%)	56 (56|0)	< 0,015
4	2	3,7%	22	4 (18,2%)	1 (4|3)	< 0,015
16	3	13,3%	8	1 (12,5%)	3 (3|0)	< 0,015
14	2	22,8%	60	27 (45,0%)	6 (33|27)	< 0,015
19	2	34,7%	228	40 (17,5%)	0 (0|0)	[0,015, 1,75]
3	2	45,4%	2	0 (0,0%)	0 (0|0)	[0,015, 1,75]

```
Tested 694, errors 99 (14.3%)

       (a)    (b)    (c) classified as
- - - - - - - - - - - - - - - - - - - - - - - - -
      261     16      1   (a): class - 0,015
       32    192     26   (b): class 0,015-1.75
        0     24    142   (c): class 1,75+
```

Evaluation on test data (159 items):

Rule	Size	Error	Used	Wrong	Advantage	Class
1	1	0,7%	61	1 (1,6%)	59 (60|1)	< 0,015
19	2	34,7%	82	6 (7,3%)	0 (0|0)	[0,015, 1,75]
20	1	13,6%	14	2 (14,3%)	10 (12|2)	> 0,75

```
Tested 159, errors 11 (6.9%)

        (a) (b)   (c)   classified as
- - - - - - - - - - - - - - - - - - - - - - - - -
       60    6     0   (a): class < 0,015
        1   76     2   (b): class [0,015,1,75]
        0    2    12   (c): class > 1,75
```