Sistema Expert
en
Anàlisi de Patrons Socials

Iván de Benito Bordoy
Enginyeria en Informàtica
Projecte Final de Carrera

Índex

1. INTRODUCCIÓ 3
2. INVESTIGACIÓ 4
2.1 Introducció 4
2.2 Elecció d'atributs 4
2.3 Anàlisi de relacions 5
2.4 Selecció del conjunt entrenament 17
3. CLASSIFICADORS 18
a) Estudi amb classificador arbre 19
b) Estudi amb classificador mapa de Kohonen 28
c) Estudi amb perceptró multicapa 37
c) Comparacions, conclusions i elecció del millor 43
4. CONCLUSIONS FINALS 45
5. BIBLIOGRAFIA 46

1. INTRODUCCIÓ

El projecte a desenvolupar consisteix en l'elaboració mitjançant tècniques d'Intel·ligència Artificial d'una eina capaç de classificar usuaris d'una xarxa social infantil segons el seu comportament.

Aquest sistema proporciona innovació donat que els sistemes actuals tenen un control reactiu, és a dir, només podem saber de quin tipus es un usuari quan una acció s'ha completat. El nostre sistema innovador detectarà patrons de comportament i classificarà als usuaris en un grup o altre. Per aconseguir una classificació predictiva necessitarem una bona base de patrons de comportament i prediccions basades en la conducta actual.

La classificació d'usuaris és un dels punts més importants a l'hora de definir estratègies de creixement, orientar diferents campanyes per a cada diferent tipus d'usuari que interactua a la nostra xarxa social o fins i tot comprovar la nostra evolució mes a mes amb respecte als diferents tipus d'usuari (potser ens interessa més un tipus que un altre).

Es proposa doncs treballar un sistema intel·ligent que es responsabilitzi de gran part de les labors d'un moderador en la monitorització i seguiment de tot tipus de conducta i fer un plantejament teòric del seguiment de conductes violentes o de risc a xarxes socials, fòrums i mons virtuals.

Aquest sistema, registrarà totes les dades dels usuaris per a la generació d'informes interns per als gestors de la pròpia xarxa (usuaris per zona, temps mig de connexió, tipus d'activitat, líders,...) i monitoritzarà les labors realitzades pels moderadors, convertint-se així en una eina de monitorització i gestió d'aquest tipus de xarxa molt completa.

En concret permetrà:

- Obtindre informació dels usuaris de la xarxa social, fòrum o mon virtual, que pot ésser valuosa tant per les aplicacions de moderació com per tenir un coneixement intern dels usuaris de la pròpia xarxa a partir del qual es poden definir estratègies de creixement o consolidació i observar la seva evolució (exemple: si els usuaris categoritzats com a líders redueixen el seu temps mitjà de connexió ens interessa incentivar-los per que no es produeixin efectes negatius sobre la seva xarxa d'influència).

- Realitzar un seguiment i gestió correcta per part dels moderadors de la xarxa. Ser moderador d'una comunitat es una tasca que du força feina així que qualsevol facilitat que els hi podem donar serà benvinguda.

Aleshores, els objectius generals d'aquest projecte fi de carrera seran:

- Realitzar una investigació prèvia sobre les dades subministrades per "Minics" (la nostra xarxa social) per extreure informació sobre el sistema i els usuaris del mateix.

- Elaborar mitjançant tècniques d'Intel·ligència Artificial un classificador capaç d'ubicar usuaris presents i futurs dins els diferents grups segons el seu comportament social.

- Presentar la informació obtinguda per a poder veure els canvis del sistema i que sigui fàcilment observable amb un CRM[1] (Customer Relationship Manager)

[1] http://es.wikipedia.org/wiki/Customer_relationship_management

2. INVESTIGACIÓ

2.1 Introducció

Aquesta primera fase d'investigació la dedicarem a fer l'anàlisi previ de la informació de la que disposem. Com hem comentat anteriorment, es de vital conèixer en primer lloc la informació de la que disposem dels usuaris de cara a poder classificar correctament als usuaris segons el grup al que pertanyen. Aquest precisament es l'objectiu d'aquest punt. En primer lloc revisarem els atributs dels que disposem i explicarem el seu rang de valors permesos, en segon lloc cercarem relacions existents entre els atributs que haguem seleccionat i en tercer lloc explicarem els criteris que hem seguit per a triar el conjunt entrenament. El conjunt entrenament es un conjunt de dades emprat en l'àmbit de la Intel·ligència Artificial per tal de descobrir relacions predictives.

2.2 Elecció d'atributs

En aquest primer punt ens centrarem en la elecció dels atributs que emprarem per a poder estudiar la seva correlació posterior. Disposem d'una base de dades de uns 77.000 exemples obtinguts de la xarxa social Minics abans de fer el filtrat.

En quant als atributs rellevants tenim:

- Identificador d'usuari. És un valor enter positiu que pot acceptar fins a 10 dígits. És a dir, va del valor 0 fins al 9999999999. Es incremental, és a dir, cada cop que s'insereix a base de dades un registre se li associa aquest identificador d'usuari a partir del darrer valor sumant-li un.
- Nombre de monedes. Diners al joc, que pot ser aconseguit mitjançant objectius o diners reals. És un valor enter positiu que pot acceptar fins a 10 dígits. És a dir, va del valor 0 fins al 9999999999.
- Felicitat del personatge. Dada que millora a base de millorar en el joc. És un valor enter positiu que pot acceptar fins a 10 dígits. És a dir, va del valor 0 fins al 9999999999.
- Nombre d'elements al inventari. Objectes totals que ha comprat el jugador. És un valor enter positiu que pot acceptar fins a 5 dígits. És a dir, va del valor 0 fins al 99999.
- Missatges al seu tauler. Missatges deixats per l'usuari a un tauló públic que tots els usuaris poden llegir. És un valor enter positiu que pot acceptar fins a 5 dígits. És a dir, va del valor 0 fins al 99999.
- Sancions rebudes. Valor que indica les sancions rebudes per l'usuari per comportament inadequat. És un valor enter positiu que pot acceptar fins a 2 dígits. És a dir, va del valor 0 fins al 99.
- Nombre de dies que l'usuari ha estat membre de pagament. És un valor enter positiu que pot acceptar fins a 10 dígits. És a dir, va del valor 0 fins al 9999999999.
- Chats privats que ha rebut l'usuari. És un valor enter positiu que pot acceptar fins a 5 dígits. És a dir, va del valor 0 fins al 99999.
- Chats privats que ha enviat l'usuari. És un valor enter positiu que pot acceptar fins a 5 dígits. És a dir, va del valor 0 fins al 99999.
- Temps total. Temps que l'usuari ha estat dintre el sistema. És un valor enter positiu que pot acceptar fins a 10 dígits. És a dir, va del valor 0 fins al 9999999999.
- Nombre d'amics de l'usuari. Es una acció que es fa quan dos perfils tenen relació i s'ofereix que siguin amics. És un valor enter positiu que pot acceptar fins a 10 dígits. És a dir, va del valor 0 fins al 9999999999.

A partir d'aquestes dades, les seves relacions i emprant tècniques i algoritmes d'Intel·ligència Artificial establirem el perfil social de cada jugador potencial.

Potser podria sorprendre que emprem un atribut com a felicitat, el motiu és que és una de les bases clau del nostre joc i a partir de la qual veiem cóm de 'feliç' és el jugador en concret. Un cop iniciat el joc, segons la interacció del propi jugador al joc, va millorant al llarg de tot el termini del joc al mon virtual. Quan més feliç és, més puntuació té a aquest camp.

Referent a l'atribut sancions rebudes, en la majoria dels casos és 0 donat que el comportament dels usuaris de la nostra xarxa social és exemplar. No obstant, és interessant disposar del nombre exacte de sancions que ha rebut un individu per tal de tenir-ho en compte a l'hora d'aplicar penalitzacions, poder estudiar la reincidència (no és el mateix 5 sancions seguides que 5 sancions al llarg de 2 anys), etc

El domini a aprendre són les diferents conductes i patrons socials que es poden observar a una xarxa social com Minics, per això es disposa d'una base de dades de 77.000 exemples abans de ser filtrats dels quals es posarà especial atenció als 12 atributs explicats anteriorment.

Finalment comentar que podria ésser una bona opció considerar un altre atribut no numèric o categòric per tal de veure més semblances entre atributs no numèrics però només tenim 2 atributs no numèrics i no són rellevants per l'anàlisi: nom (el qual és irrellevant de cara a estudiar una possible correlació) i classe, que és precisament el que volem predir amb la qual cosa tampoc ens serveix.

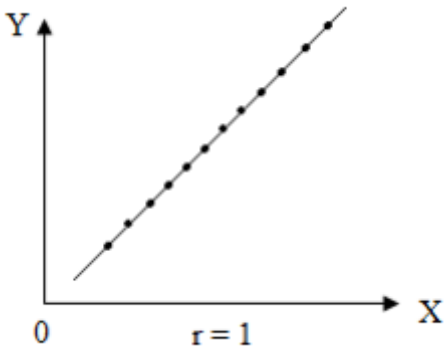
2.3 Anàlisi de relacions

A continuació procedirem a realitzar un anàlisi exhaustiu a partir dels atributs triats anteriorment per tal de poder trobar relacions entre ells. Hem pres la decisió de no normalitzar cap atribut de cara a fer la comparació entre dades.

Per a començar s'elimina d'aquest anàlisi l'identificador de l'usuari donat que, com hem comentat, és un nombre auto incremental que es genera cada cop que hi ha una alta al sistema i no té sentit de cara a trobar relacions entre atributs.

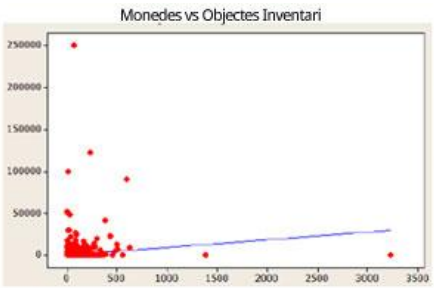
A continuació mostrarem unes gràfiques en forma de diagrames de dispersió en les quals l'eix Y correspondrà al primer atribut i l'eix X correspondrà al segon atribut de la comparació. El diagrama de dispersió ens pot suggerir varis tipus de correlacions entre variables (positiva, negativa o nul·la). Podem fer servir una línia d'ajust amb la finalitat d'estudiar la correlació entre variables.

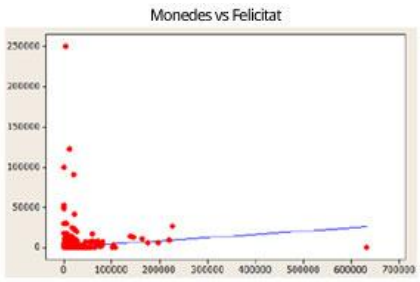
Aleshores, per una banda representarem com a punts vermells els valors individuals i per una altra banda tindrem una línea blava que serà la nostra línia d'ajust o línia de tendència[2] segons els valors indicats a la matriu base. Aleshores, a més de la concentració dels valors de forma gràfica amb respecte a la línia d'ajust podrem deduir més correlació. Juntament amb això tindrem en compte la segregació a nivell visual per a establir si una relació es clara ja que la relació ideal, si fos 1 seria similar al següent

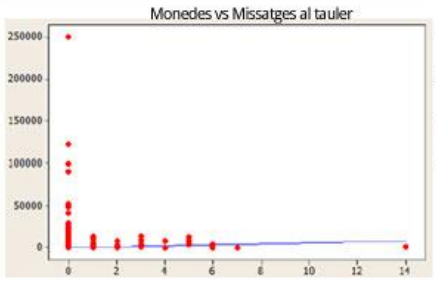


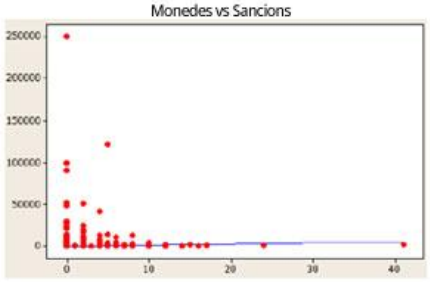
[2] http://es.wikipedia.org/wiki/Diagrama_de_dispersi%C3%B3n

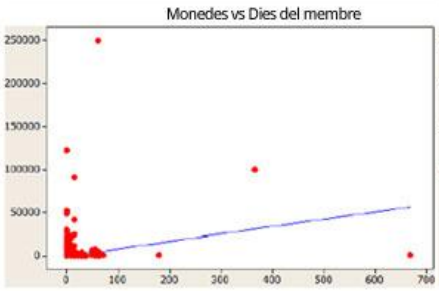
Tindrem en compte aquestes apreciacions comentades a l'hora de comparar les diferents gràfiques en que relacionem atributs.

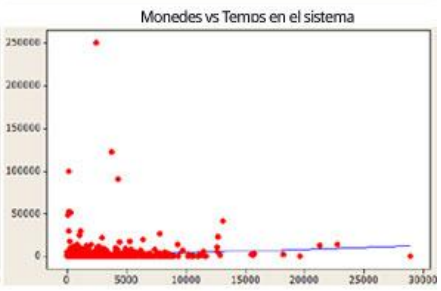


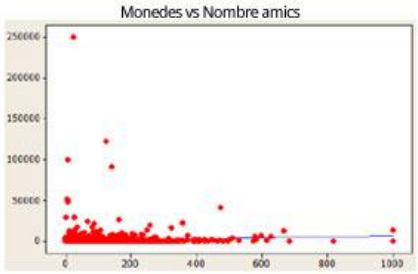


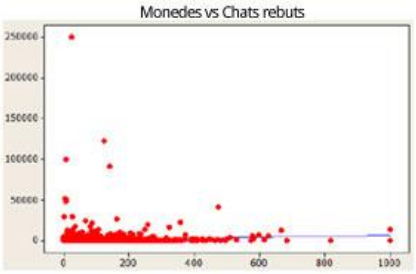


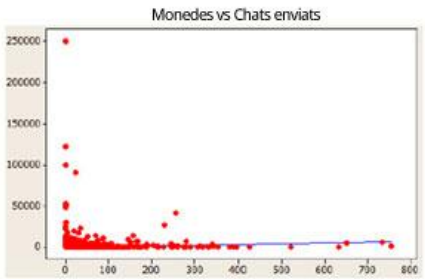




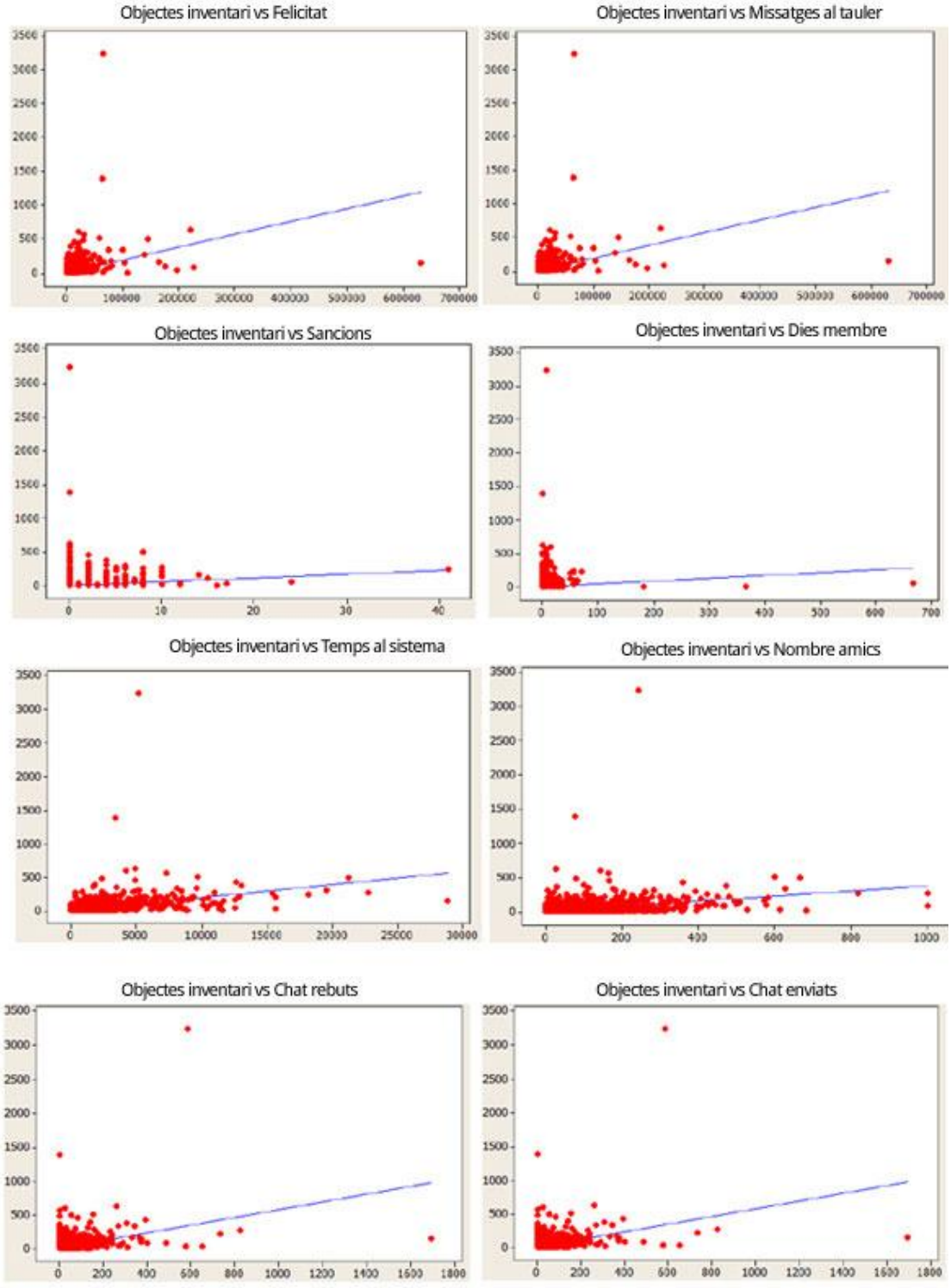






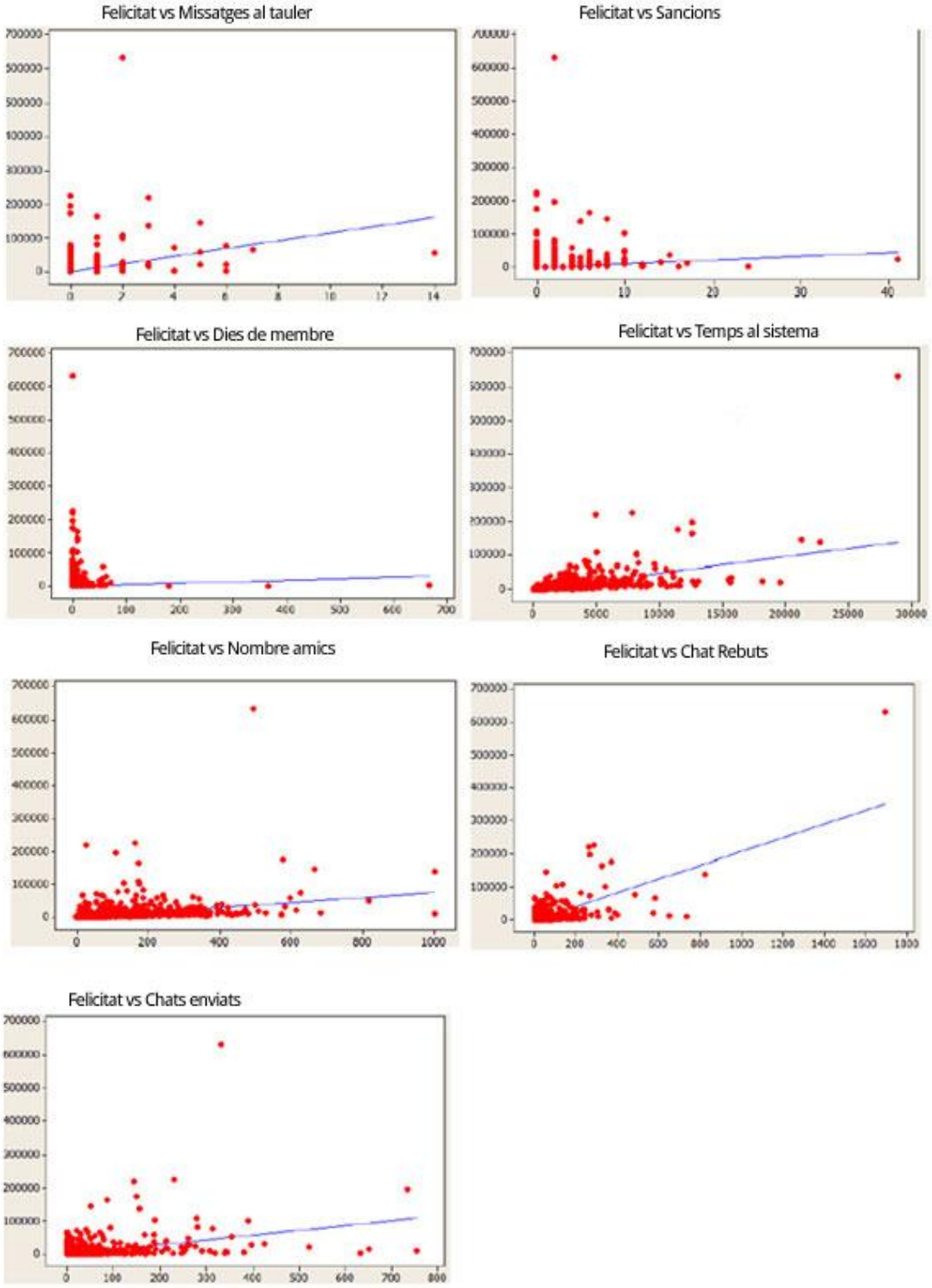


Podem veure a les gràfiques representades comparant l'atribut monedes que existeix una relació nul·la. Aquest fet és podia preveure donat que als valors de l'atribut monedes són els que hi havia en el moment d'agafar la base de dades. Per exemple, és evident que les persones amb més objectes han tingut més monedes però es possible que al moment d'obtenir les dades se les han pogut gastar. A totes les gràfiques podem observar que major increment de X no implica major increment de Y i també podem veure que la concentració de punts en la major part dels casos queda força allunyada de la línia d'ajust.

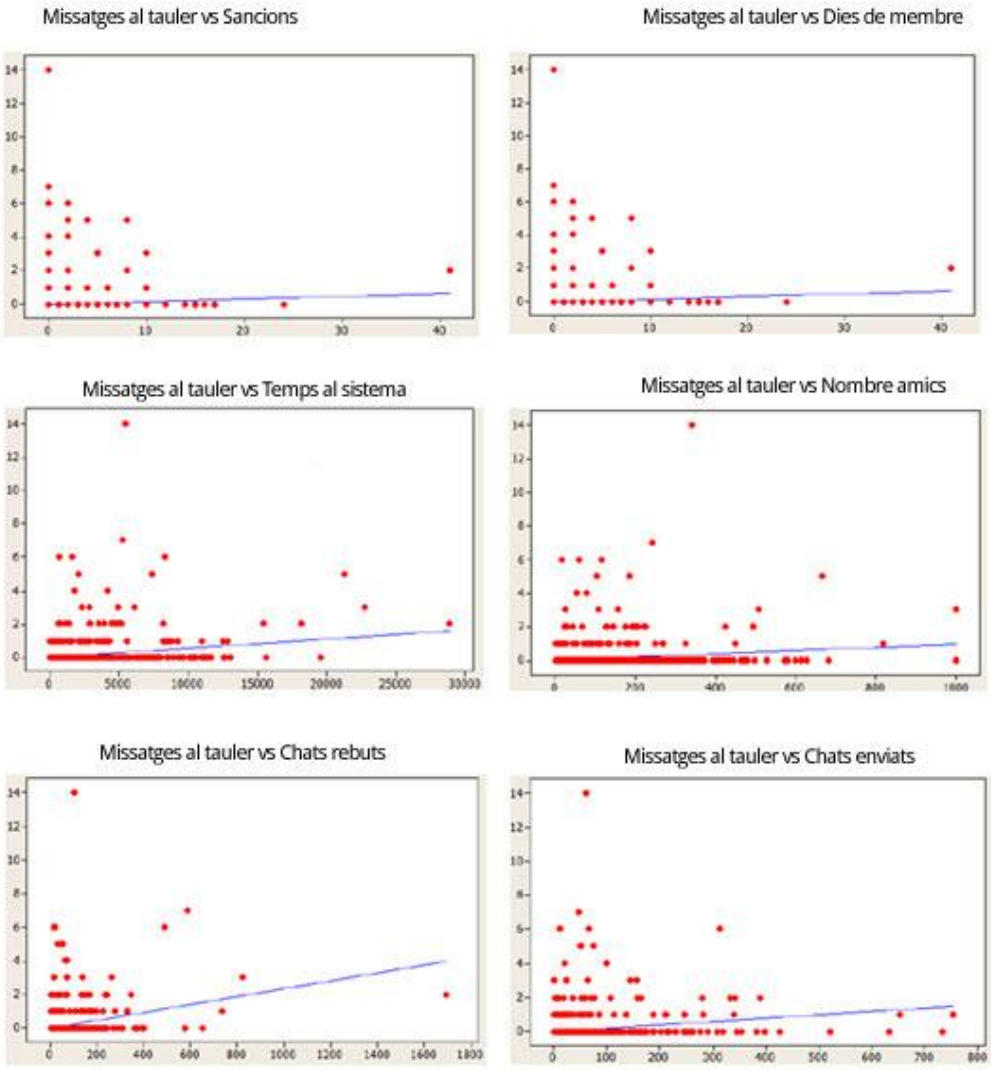


Les gràfiques en que es compara l'atribut objectes inventari amb la resta presenten certa semblança. La més significativa la veiem al cas de chats rebuts, chats enviats, felicitat i missatges al tauler. En aquests casos senyalats veiem que X i Y van més correlatives que als altres casos i que la dispersió de punts queda més propera a la línia d'ajust.

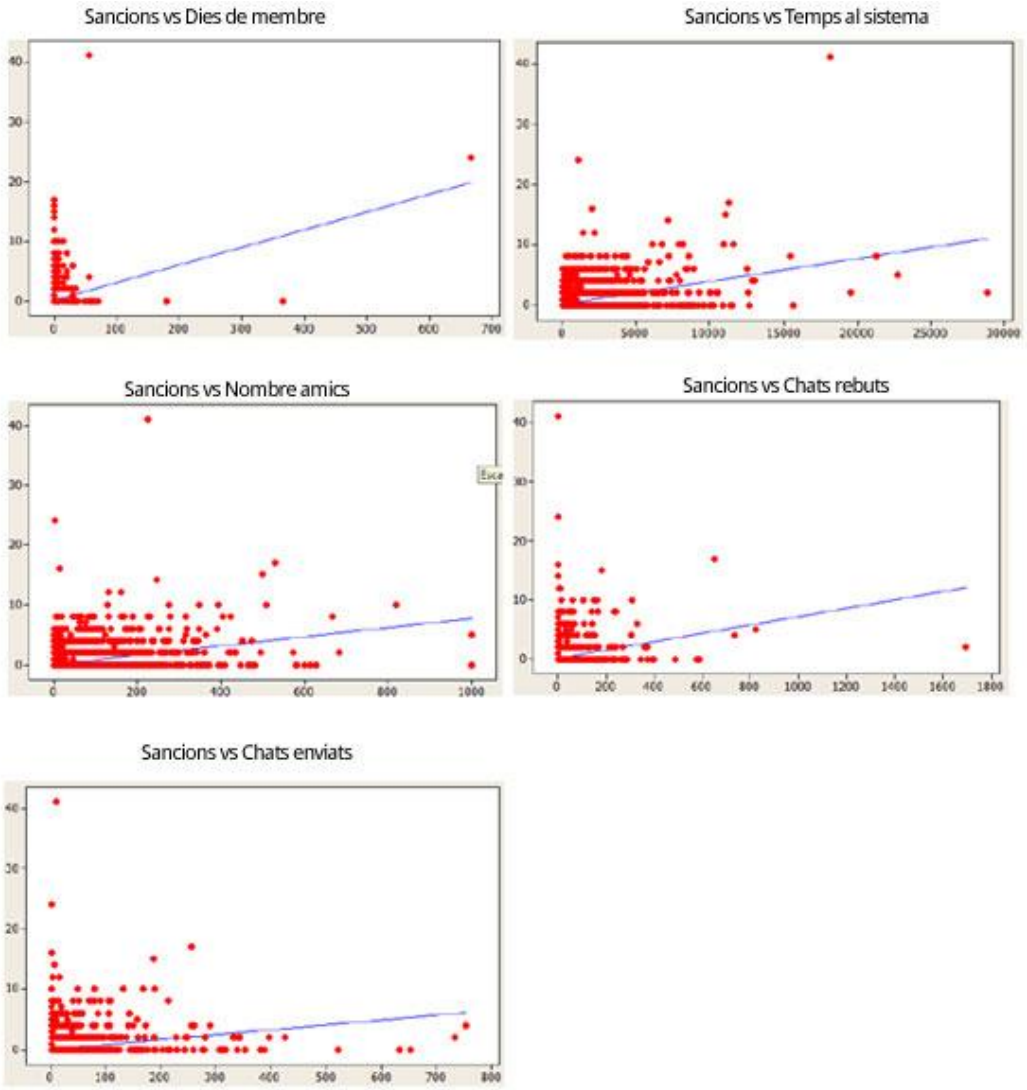
Tot i això no podem observar una relació suficientment representativa com per a concloure que una variable es totalment correlativa a l'altre.



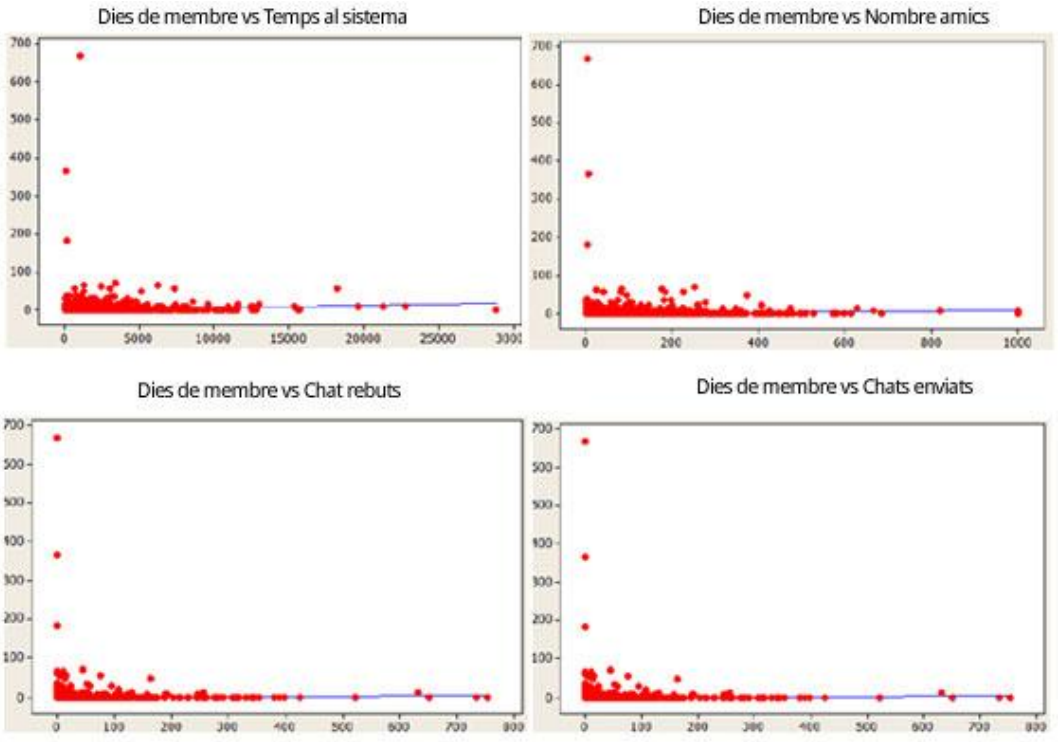
Podem observar la existència de relacions gairebé nul·les al cas de felicitat vs sancions o felicitat vs dies de membre ja que els valors es troben dispersos i força allunyats de la línia d'ajust. Altres gràfiques com a felicitat vs temps al sistema i felicitat vs chats rebuts tenen una correlació bastant alta que analitzarem a un altre apartat.



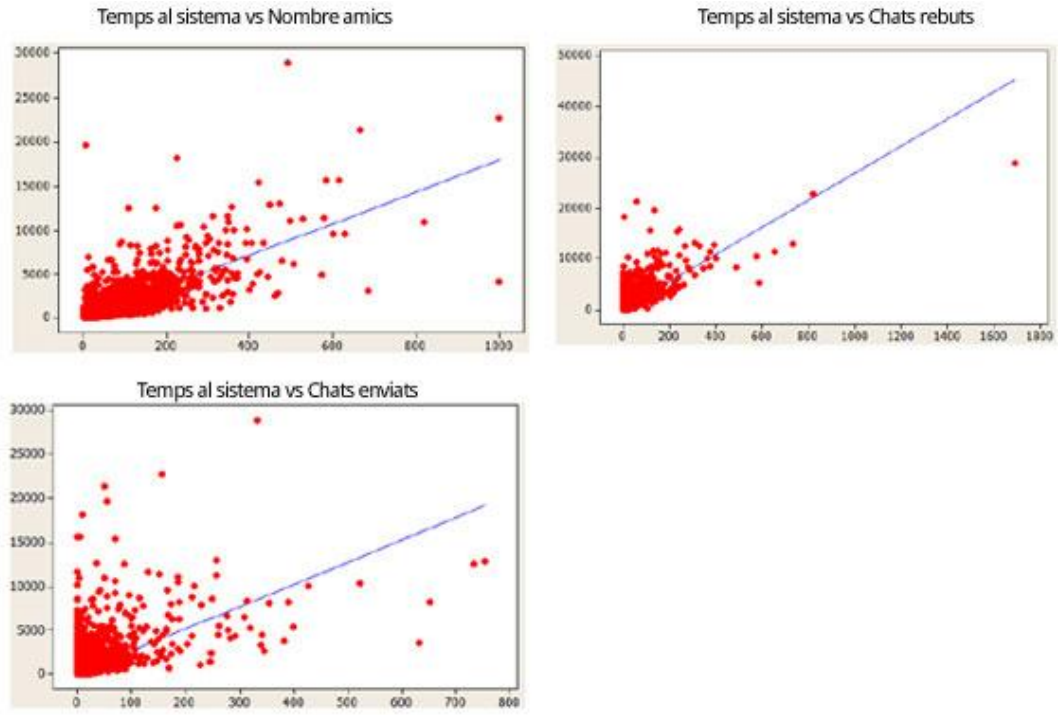
Observant les gràfiques veiem que tots els punts es troben lluny de la línea d'ajust i molt dispersos. Junt amb això si ens fixem a la gràfica en la pendent de la línea d'ajust veurem que està força lluny de semblar una pendent de 45º que indiqui una certa dependència i relació entre un atribut i l'altre motiu pel qual no podríem concloure que els atributs tenen qualque relació. Potser la gràfica que més s'adapta sutilment es la de chats rebuts però com veurem als valors posteriors no es ni molt manco significativa.



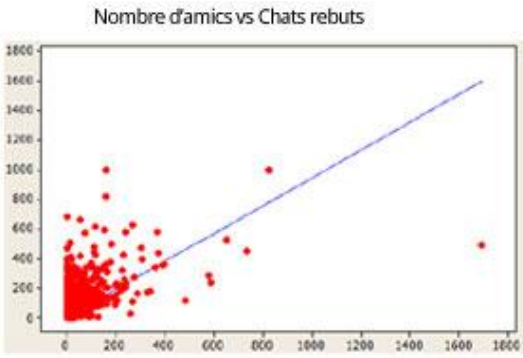
Comparant totes les gràfiques veiem que tenen certa semblança en la distribució dels seus punts al llarg de les diverses gràfiques. Potser la més diferent es la comparació de sancions vs dies de membre. En qualsevol dels casos els punts es troben força dispersos i allunyats de la línea d'ajust. Estudiarem la correlació més endavant per arribar a més conclusions.

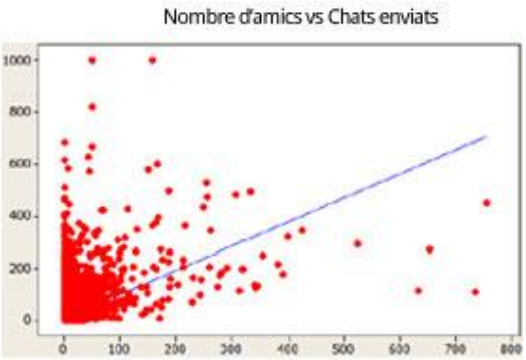


A totes les gràfiques veiem distribucions de punts semblants tot i que la línea d'ajust i la seva pendent ens fa veure que la correlació es nul·la. No hi ha cap atribut dels comparats que depengui directament de dies de membre.



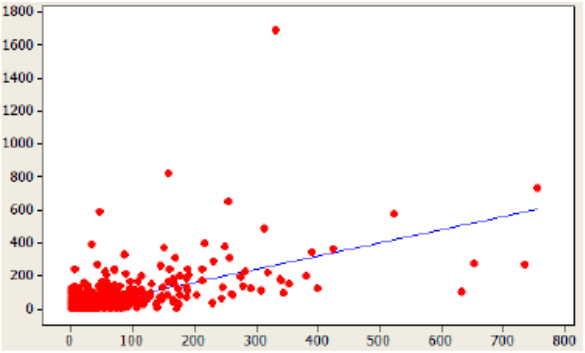
Gran semblança en aquestes 3 variables, casi totes les gràfiques s'apropen a una pendent de 45º, molt relacionades, com és lògic pels atributs que representen. Es veurà posteriorment que les seves correlacions són pròximes a 1 i haurà que veure si s'eliminen.





En aquest cas tot ens fa indicar una gran relació entre els atributs comparats amb nombre d'amics. Una pendent propera als 45º essent quasi perfecta i una distribució de punts molt propera a la línia d'ajust. De totes les estudiades fins ara es la que major dependència té i haurem d'estudiar si s'elimina o no detingudament.

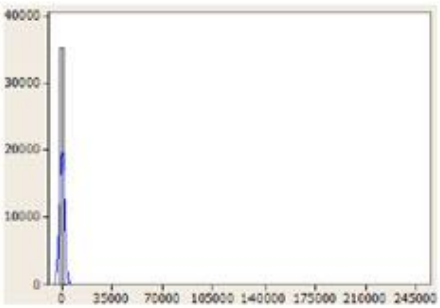
Chats rebuts vs Chats enviats



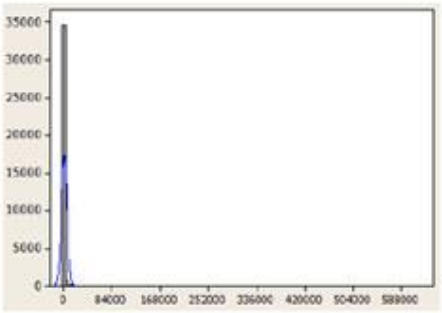
Pel que podem apreciar en la gràfica la relació es apreciable però manco significativa que al cas anterior. Té una lleugera pendent i uns punts pròxims a la línia d'ajust. En aquest cas puntual són dues variables que ens interessen molt ja que donen la interacció a nivell privat de l'usuari a la nostra xarxa social, motiu pel qual es descarta a priori la seva eliminació.

A continuació mostrarem amb histogrames la distribució de cada variable. En totes les gràfiques la barra que és més gran és la del 0, això no significa que sigui 0, si no que està entre 0-X essent X un valor que ve determinat per la longitud del interval. Això passa perquè l'eix reflexa el valor més alt, pot haver un usuari com en el cas del 660 chats enviats però el valor normal és que els usuaris estiguin entre 0-40

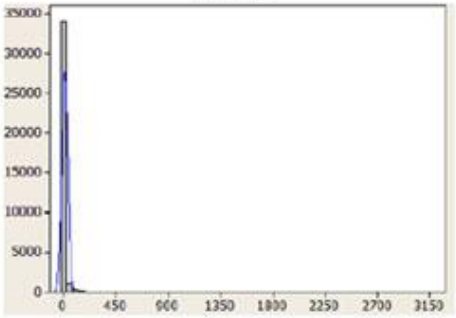
Monedes



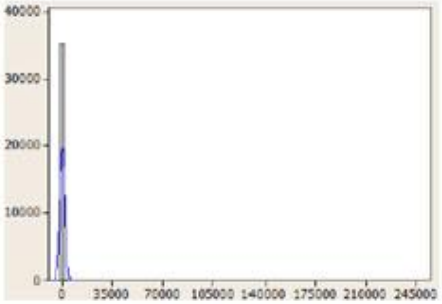
Felicitat



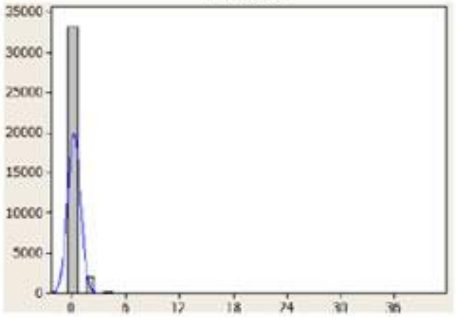
Inventari



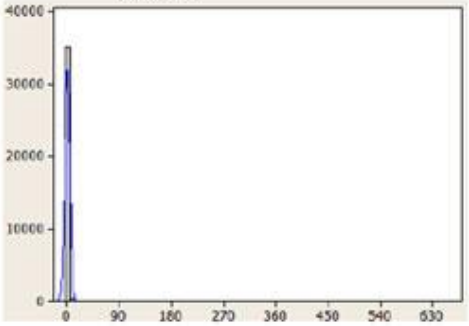
Tauler



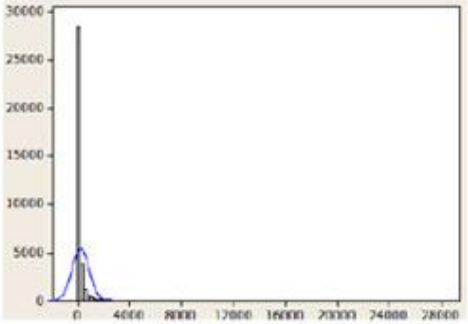
Sancions



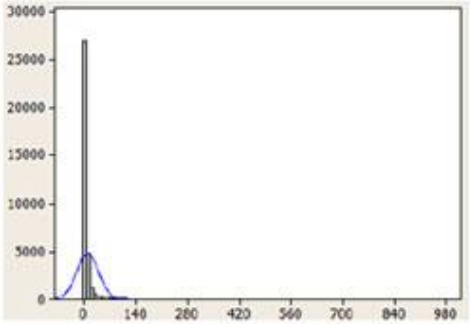
Membres



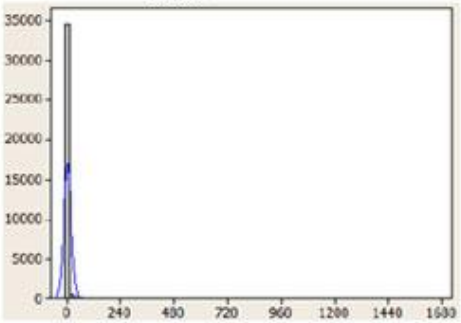
Temps



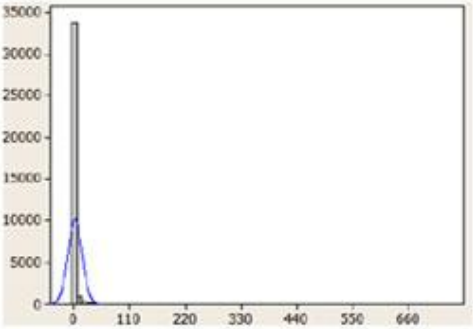
Amics



Chat-in



Chat-out



Monedes. Mitja 77,94 Variància 3237487 Asimetria 95,10
Felicitat. Mitja 505,14 Variància 24423314 Asimetria 71,68
Inventari. Mitja 8,75 Variància 657,39 Asimetria 64,02
Tauler. Mitja 0,045 Variància 0,017 Asimetria 55,19
Sancions. Mitja 0,16 Variància 0,5 Asimetria 11,83
Membre. Mitja 0,16 Variància 19,5 Asimetria 114,87
Temps. Mitja 175,6 Variància 426719 Asimetria 13,63
Amics. Mitja 7,95 Variància 868 Asimetria 10,71
Chatin. Mitja 1,58 Variància 279 Asimetria 114,87
Chatout. Mitja 1,57 Variància 196 Asimetria 25,58

Monedes: Les 'monedes' al joc Minics són escasses, per aquest motiu la gran majoria dels usuaris tenen entre 0-500.

Felicitat: La felicitat és un factor que disminueix amb el temps, i l'usuari ha d'esforçar-se per pujar, amb la qual cosa el valor quasi en tots els casos és bastant baix.

Inventari: Els objectes que es compren al inventari es compren mitjançant monedes per tant són també un bé escàs.

Tauler: Rara vegada algú escriu alguna cosa al tauler, però quan ho fa és important donat que és un comportament públic que tothom pot veure.

Sancions: La major part dels usuaris són modèlics motiu pel qual el nombre de sancions és 0 en quasi tots els casos i en uns pocs tenen valors petits.

Membres: La qualitat membre és una qualitat que s'obté mitjançant diners reals amb el que són pocs usuaris els que opten per aquesta opció.

Temps: Com en tot el joc el temps que passa l'usuari al sistema es pareixen molts de casos. Alguns juguen un poc més i uns pocs juguen molt.

Amics: L'usuari comença amb 0 amics i ha d'anar pujant, la qual cosa li aportarà felicitat, és per aquest motiu que tots es troben entre 0 i 140, però hi ha usuaris que poden tenir 980 amics.

Chat-in: Els chats privats rebuts reflecteixen la popularitat de l'usuari i són escassos com es pot veure.

A continuació mostrarem un estudi estadístic de les variables des d'un punt de vista numèric, per a corroborar i decidir respecte al vist en les gràfiques.

El que farem serà estudiar el coeficient de correlació per examinar la relació entre els dos atributs comparats.

$$Correl(X,Y) = \frac{\sum (x-\bar{x})(y-\bar{y})}{\sqrt{\sum (x-\bar{x})^2 \sum (y-\bar{y})^2}}$$

Al llarg del càlcul de les correlacions següents tindrem en compte que quan més s'aproximin a 1 més relacionades estan els dos atributs que es comparen, és a dir, que quasi podríem dir que un depèn de l'altre. Quan més pròxims a 0 més independents són, és a dir, que quasi no tenen relació i poden ser estudiades de forma independent.

Monedes
vs Felicitat: 0,112
vs Inventari: 0,131
vs Tauler: 0,042
vs Sancions: 0,043
vs Membre: 0,205
vs Temps: 0,142
vs Amics: 0,102
vs ChatIn: 0,079
vs ChatOut: 0,057

Com hem comentat en altres ocasions l'atribut monedes és un bé que es dona a l'inici del joc i sol esser entre 0 i 500 en la gran majoria de casos. Per tant, pel seu propi caràcter consumible no es veuen relacions aparents entre el nivell de monedes i el nivell de felicitat.

Felicitat
vs Inventari: 0,361
vs Tauler: 0,31
vs Sancions: 0,156
vs Membre: 0,038
vs Temps: 0,628
vs Amics: 0,439
vs ChatIn: 0,701
vs ChatOut: 0,407

En aquest cas podem observar que existeixen relacions mitges (amb el temps) i fortes (amb el cas del chatin). Com hem comentat anteriorment la felicitat és un valor que incrementa al llarg del joc segons la interacció que té el jugador amb el mateix. Per tant, és lògic veure una relació temps-felicitat i el mateix amb chats entrants i enviats. Al cas del nombre d'amics veiem també que encara que no té la mateixa relació alguna cosa té a veure amb els chats i el temps que els usuaris passen al joc.

Inventari
vs Tauler: 0,313
vs Sancions: 0,155
vs Membre: 0,072
vs Temps: 0,492
vs Amics: 0,428
vs ChatIn: 0,373
vs ChatOut: 0,22

Al cas de l'atribut inventari comparat amb altres atributs no veiem cap relació força significativa. L'atribut inventari guarda els objectes que es compren amb les monedes per tant amb el que té més relació si la volem veure es amb el temps. A més temps, més capacitat per recol·lectar més objectes del inventari.

Tauler
vs Sancions: 0,084
vs Membre: 0,032
vs Temps: 0,271
vs Amics: 0,209
vs ChatIn: 0,296
vs ChatOut: 0,208

En aquest cas no tenim cap relació que necessiti ser destacada. Totes tenen una baixa correlació ja que no arriben ni de lluny a superar 0,2. Té sentit per la pròpia característica de l'atribut ja que és només un missatge al tauler i és un comportament que no es fa molt sovint.

Sancions vs Membre: 0,184
vs Temps: 0,346
vs Amics: 0,316
vs ChatIn: 0,166
vs ChatOut: 0,156

Hem comentat en altres ocasions que el nombre de sancions sol esser molt baix degut al comportament exemplar del nostres jugadors. És per tant bastant lògic que no tinguem relacions significatives d'aquest atribut amb la resta i que les més destacables siguin amb temps i amics. A més temps al joc més possibilitats de caure en qualque sanció.

Membre vs Temps: 0,082
vs Amics: 0,060
vs ChatIn: 0,027
vs ChatOut: 0,025

L'atribut membre és una qualitat que s'obté amb diners reals pagant amb la qual cosa, degut a la baixa conversió al joc, tenim pocs usuaris que optin a aquesta opció. Per tant, és força raonable que gairebé existeixi relació.

Temps vs Amics: 0,808
vs ChatIn: 0,683
vs ChatOut: 0,542

Aquest sens dubte és l'atribut amb relacions més rellevants de tot el conjunt analitzat. Podem veure relacions properes a 1. Recordem que el temps que passa un jugador al nostre joc és un atribut força relatiu que ens fa veure el nivell de engagement (implicació) del usuari amb el nostre joc, amb el seu mon virtual, amics, etc.

Amics vs ChatIn: 0,533
vs ChatOut: 0,440

És obvi que l'atribut amics té una relació mitja amb els chats enviats i chats rebuts. Això es produeix ja que hi ha usuaris que gairebé fan ús dels chats, altres que en fan un ús bastant abundant diàriament i d'altres que l'empren una mica manco.

ChatOut vs ChatIn: 0,672

Els chats enviats veiem que tenen una relació significativa amb els chats rebuts. Això és degut al propi caràcter de la comunicació. Habitualment quan enviem missatges ens els solen respondre.

2.4 Selecció del conjunt entrenament

Els exemples proposats representen perfectament el domini a aprendre donat que es tracta de la base de dades completa del joc real Minics amb el que es desenvolupa el projecte, són dades totalment reals.

Filtrarem en primer lloc el nombre d'exemples per a aconseguir una representativitat, eliminant aquells amb les següents característiques:

- Usuaris amb valors nuls.
- Usuaris no representatius. Potser s'han registrat però no han interactuat amb el joc i no tenen activitat aparent, per exemple els que tenen temps inferior a 5 minuts jugats.
- Usuaris molt estranys. Pels valors de les dades que tenen i la relació entre elles es veu clarament que són fruit de proves.

Amb això ens queden 36.000 casos per estudiar.

Després d'un estudi a consciència dels valors alts de correlació i donada la importància en el resultat final s'ha decidit no eliminar cap de les variables, la seva forta correlació no és suficient per causar problemes d'estudi i es pot afegir qualque detall que resulti important pels perfils socials.

Les condicions que reuneix aquest conjunt d'entrenament són la representativitat i diversitat i un tamany adequat pel tamany de les dades totals. Representen perfectament els usuaris del joc i tenen tota la diversitat possible començant en avatars que no fan pràcticament res fins altres que són molt actius. El tamany quan es tenen 77.000 usuaris queda reduït a 36.000 individus representatius.

Els exemples estan ben caracteritzats amb atributs representatius de la seva activitat al sistema.

3. CLASSIFICADORS

Arribats a aquest punt procedirem a emprar classificadors per tal d'organitzar i categoritzar les dades en classes diferents.

Per una banda emprarem arbres donat que ens permetran generalitzar a partir de casos particulars els diferents conceptes que identifiquen, per una altra, les xarxes neuronals (Kohonen i multicapa) s'han revelant com un útil instrument per obtenir informació a partir de grans quantitats de dades.

Per aquest objectiu emprarem diferents mètodes de classificació com

- Classificador en arbre. Emprarem un arbre amb WEKA[3] donat que la col·lecció d'algorismes per l'anàlisi de dades i el modelatge predictiu que ens proporciona és extensa i de gran fiabilitat. Amb això aconseguirem generalitzar a partir de casos particulars els diferents conceptes que identifiquem. Un cop tinguem l'arbre creat a partir del conjunt inicial farem una segona prova amb dades reals no emprades per a la construcció de l'arbre que ens permetin verificar la fiabilitat del conjunt de regles creat.
- Mapa de Kohonen. És un tipus de xarxa neuronal artificial que s'entrena emprant aprenentatge no supervisat per produir una representació discreta del espai d'entrada de les mostres entrants (mapes). Formen un mapa bidimensional de característiques a partir de les dades d'entrada per quedar agrupat en classes de major semblança. Al nostre exemple serà capaç de classificar 10 classes i anirem modificant les proves per tal d'aconseguir la millor classificació possible. Aquestes xarxes neuronals aprenen amb un mecanisme estímul - resposta igual que les del nostre cervell i reconeixen característiques similars i les associen a respostes apreses.
- Perceptró multicapa. És una xarxa neuronal artificial formada per múltiples capes que li permet resoldre problemes que no són linearment separables. Es representa de forma successiva i de forma reiterada per parells de capes d'entrada i sortida i crea un model que ajusta els seus pesos en funció dels vectors d'entrenament per arribar a produir un vector sortida similar al esperat.

L'objectiu és veure cóm de bé es classifiquen emprant cada un dels classificadors les dades inicials que tenim i fer una comparativa final a fi de triar i emprar la millor opció.

Al llarg d'aquest capítol emprarem les següents inicials o tipologies de jugadors indistintament per a fer referència a diferents perfils de jugadors que reuneixen unes característiques determinades.

Líder (L): Usuari amb molta repercussió en els demés, és a dir, el que fa es té molt en compte per part de tots els participants del món virtual.

Vip (V): Usuari amb pagaments freqüents o grans pagaments.

Sociable (S): Usuari amb molts amics i ganes de participar en les converses tant públiques com privades, li agrada socialitzar.

Popular (P): Usuari amb el que tothom vol posar-se en contacte. També té molts amics.

Marginal (M): Oposat al popular. No té amics i ningú vol posar-se en contacte amb ell.

Triomfador (T): Usuari involucrat i motivat pels objectius del joc.

Agressiu (A): Usuari sancionat freqüentment per comportament agressiu.

[3] http://www.cs.waikato.ac.nz/ml/weka/downloading.html versió Windows x64

Col·leccionista (C): Usuari preocupat pels bens materials del joc.

Normal (N): Usuari habitual.

a) Estudi amb classificador arbre

L'objectiu de la classificació és construir un arbre que sigui capaç de separar diferents patrons socials.

És una tècnica d'aprenentatge automàtic per inducció que permet identificar conceptes (classes d'objectes) a partir de les característiques d'un conjunt d'exemples que els representen. La informació obtinguda dels mateixos queda organitzada jeràrquicament en forma d'arbre. És un procés de generalització a partir de casos particulars.

Es representen per un graf dirigit que consta de nodes i arcs. Els nodes es corresponen a una pregunta o a un test que es fa als exemples.

L'arbre de decisió es construeix a base d'anar fent preguntes sobre les característiques determinades als exemples i classificant-los segons la resposta. Per tant un arbre de decisió treballa com un classificador. Les diferents opcions de classificació (resposta a les preguntes) són excloents entre si, el que fa que a partir de casos desconeguts i seguint l'arbre adequadament, s'arriba a una única conclusió o decisió a prendre.

Els possibles atributs separadors de l'arbre són:

- Nombre de monedes (diners al joc, que pot ésser aconseguit mitjançant objectius o diners reals).
- Felicitat del personatge, dada que es millora a base de millorar al joc.
- Nombre d'elements del inventari (objectes totals que ha comprat el jugador).
- Missatges al tauler (missatges deixats per l'usuari a un tauler públic que tos els usuaris poden llegir).
- Sancions rebudes per l'usuari per comportament inapropiat.
- Nombre de dies que l'usuari ha estat membre de pagament.
- Chats privats que ha rebut l'usuari.
- Chats privats que ha enviat l'usuari.
- Temps total que l'usuari ha estat al sistema.
- Nombre d'amics de l'usuari, és una acció que es fa quan dos perfils tenen relació. En aquest moment se'ls ofereix que siguin amics.

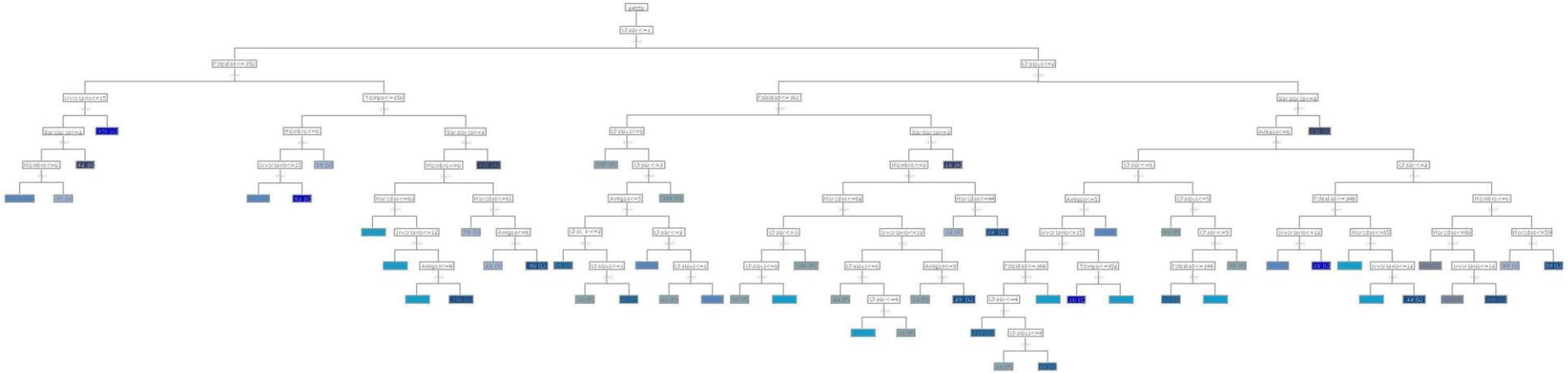
Arbre WEKA amb C4,5:

Es procedirà a provar un arbre amb 10 exemples de diferents classes que no han estat emprats per a la construcció del mateix.

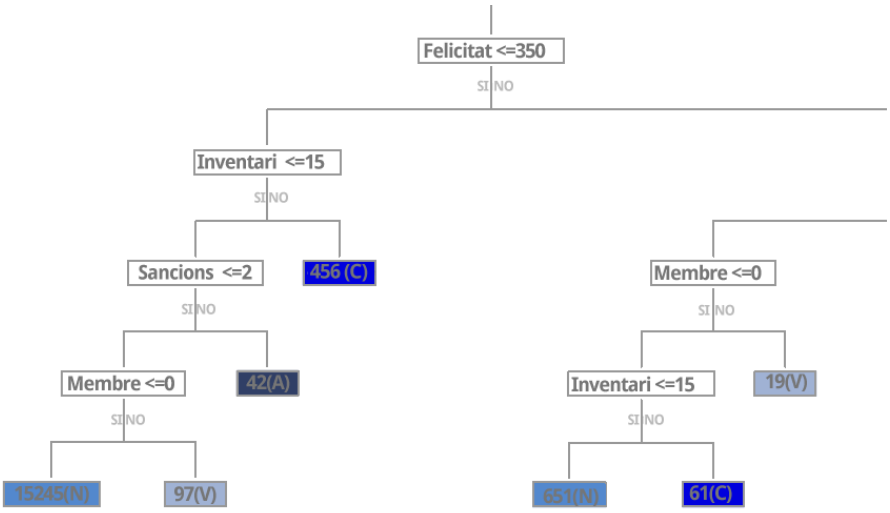
Cas: 35.460 – Classe: M -- Resultat: Correcte
Cas: 31.952 – Classe: T -- Resultat: Correcte
Cas: 7.742 – Classe: L – Resultat: Correcte
Cas: 20.919 – Classe:C – Resultat: Correcte
Cas: 8 – Classe: V – Resultat: Incorrecte (N)
Cas: 17.152 – Classe:A – Resultat: Correcte
Cas: 1.046 – Classe: P – Resultat: Correcte
Cas: 13.637 – Classe:S – Resultat: Correcte
Cas: 28.900 – Classe: N – Resultat: Correcte

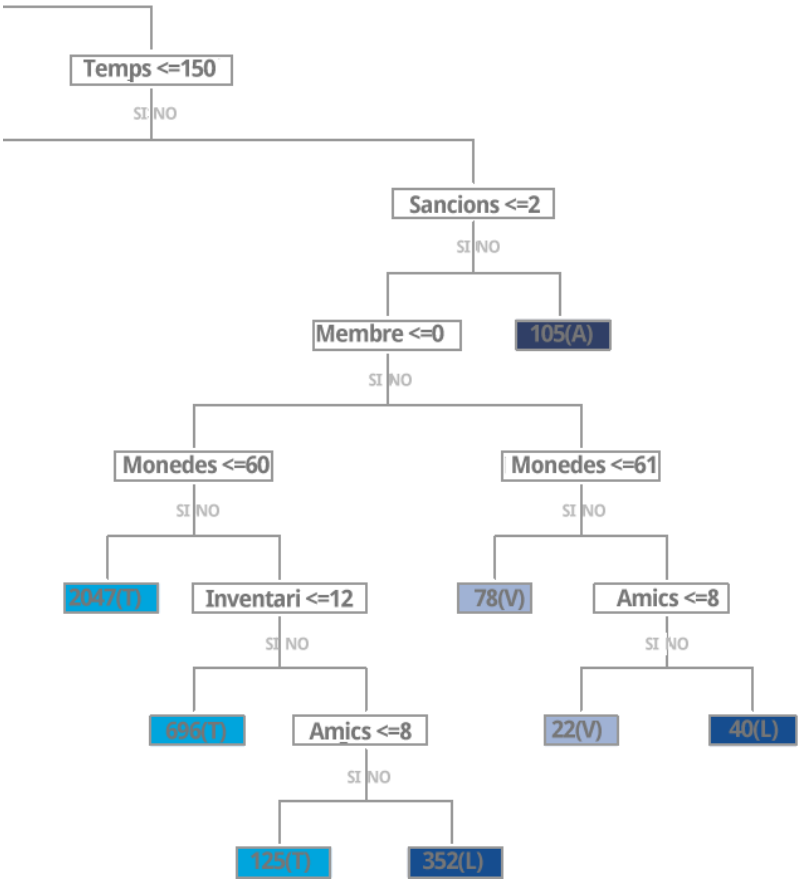
S'han classificat correctament el 90% dels exemples, tan sols ha hagut un error que s'ha produït a la classe V, classe que significa VIP i es deu a una confusió en la classificació típica als arbres. No és res alarmant donat que amb varis més provats de la classe V funciona a la perfecció. Encara i tot amb aquest petit error circumstancial un 90% d'encert és un percentatge molt alt i molt satisfactori que haurem de comparar amb l'encert del perceptró i la xarxa per la elecció del classificador final.

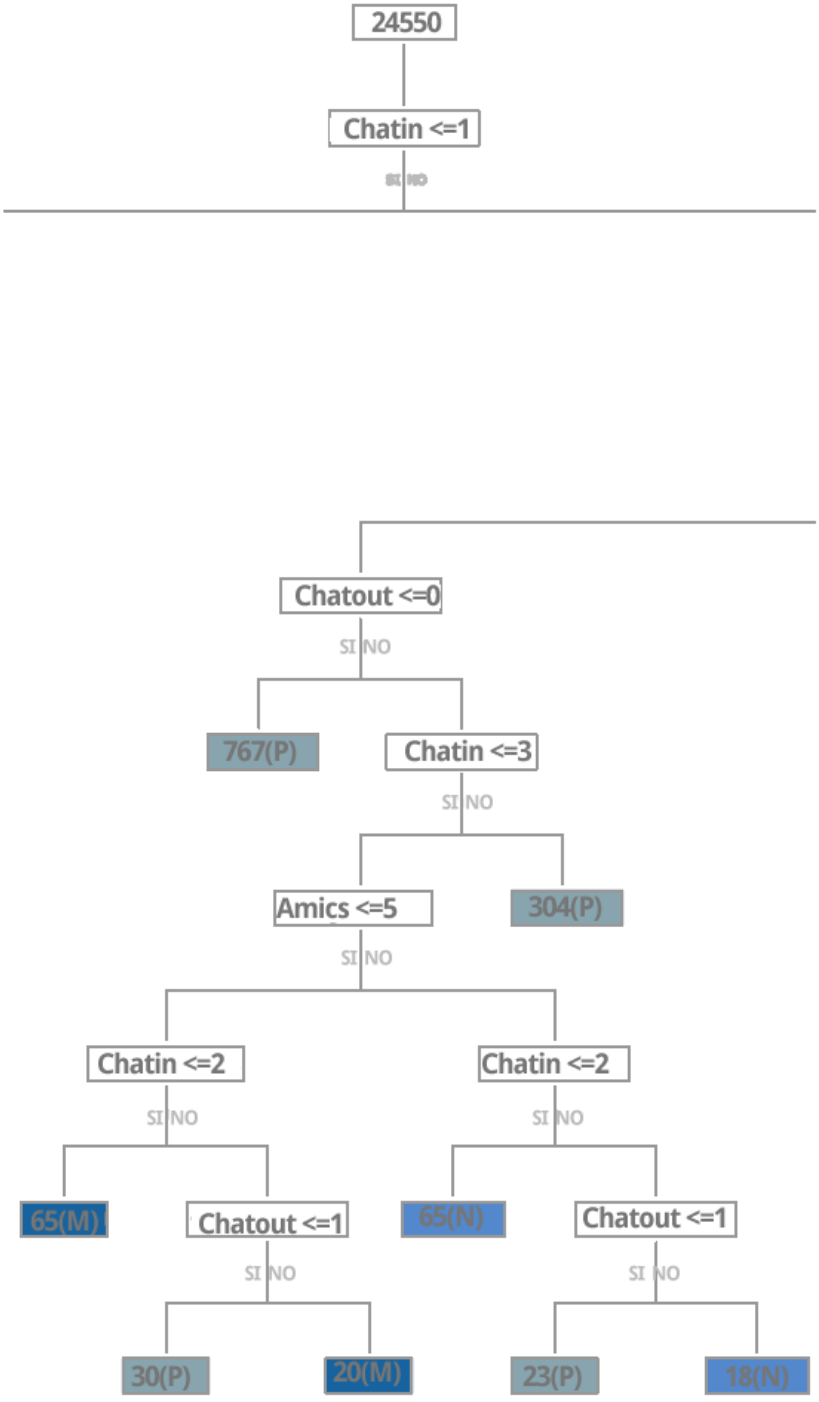
A continuació mostrarem l'arbre classificador en tota la seva extensió:

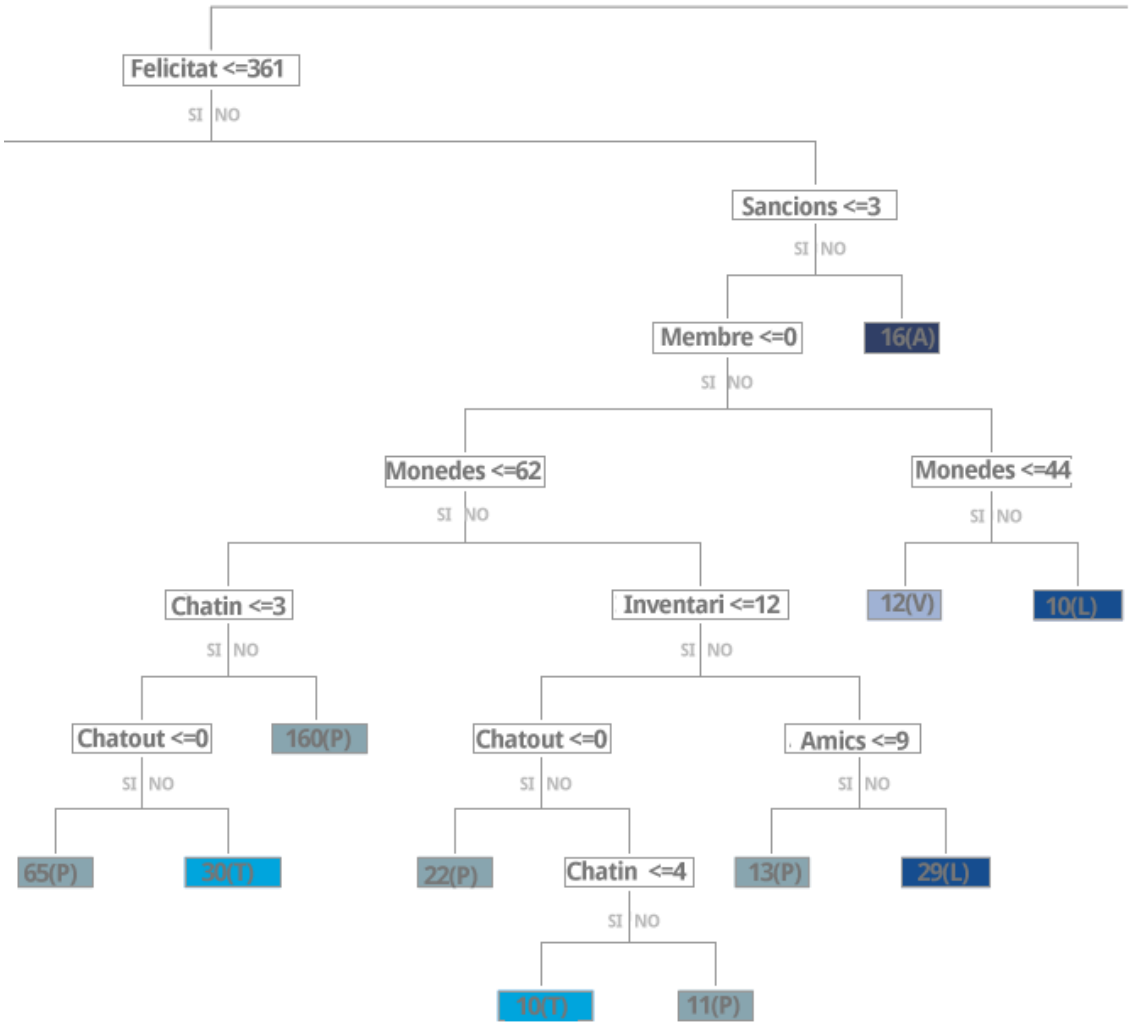


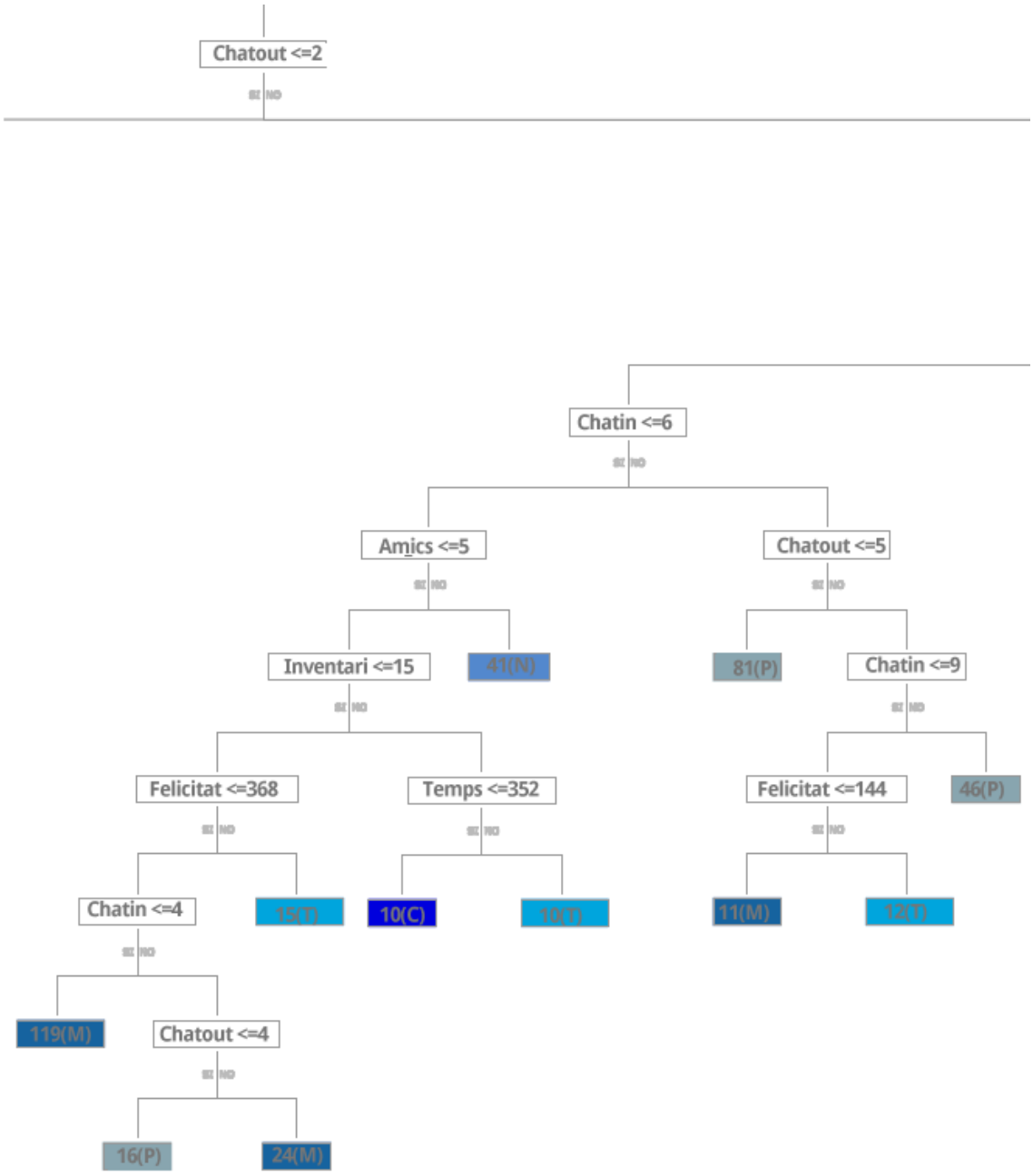
Ara que ja tenim clara la estructura general de l'arbre adjuntarem les diferents parts per veure amb més claredat les regles que hem obtingut. La ordenació serà la mateixa de la imatge de la pàgina anterior de esquerra a dreta.

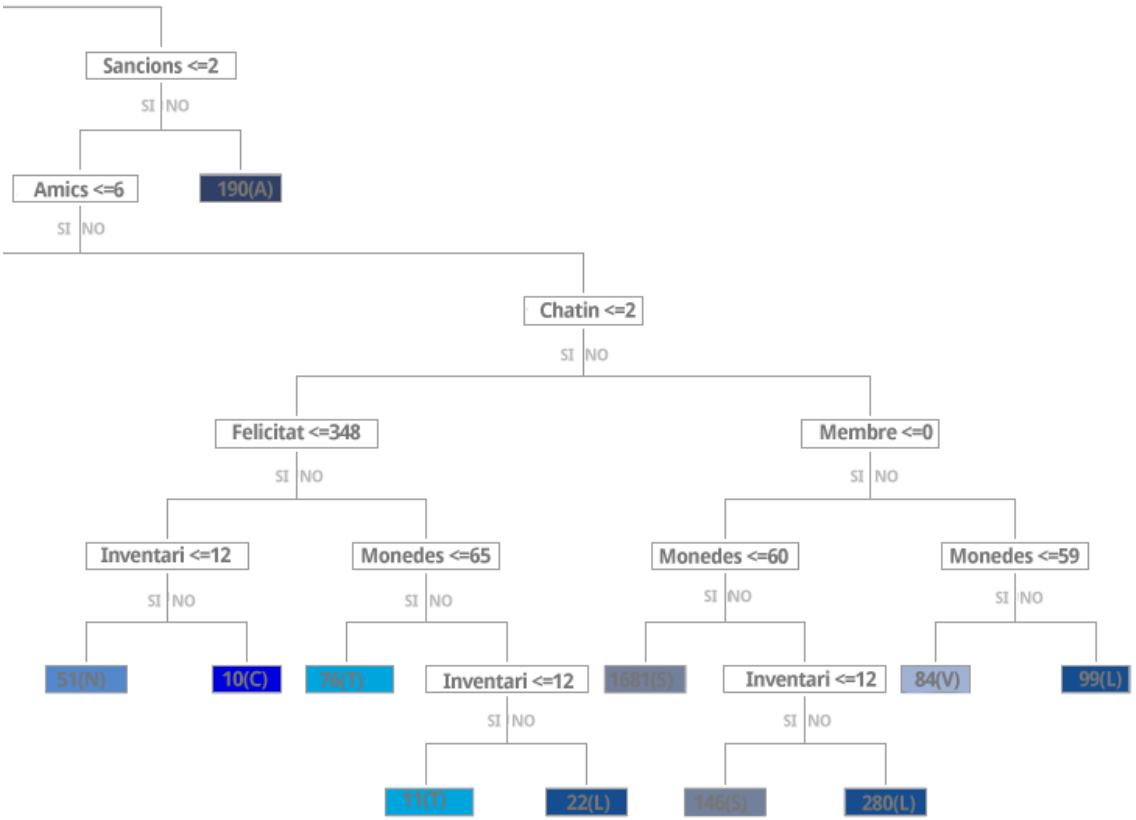



Temps <=150
SI NO
Sancions <=2
SI NO
Membre <=0
105(A)
SI NO
Monedes <=60
Monedes <=61
SI NO
SI NO
2047(T)
Inventari <=12
78(V)
Amics <=8
SI NO
SI NO
696(T)
Amics <=8
22(V)
40(L)
SI NO
125(T)
352(L)


24550
Chatin <=1
SI NO
Chatout <=0
SI NO
767(P)
Chatin <=3
SI NO
Amics <=5
304(P)
SI NO
Chatin <=2
Chatin <=2
SI NO
SI NO
65(M)
Chatout <=1
65(N)
Chatout <=1
SI NO
SI NO
30(P)
20(M)
23(P)
18(N)


Felicitat <=361
SI NO
Sancions <=3
SI NO
Membre <=0
16(A)
SI NO
Monedes <=62
Monedes <=44
SI NO
SI NO
Chatin <=3
Inventari <=12
12(V)
10(L)
SI NO
SI NO
Chatout <=0
160(P)
Chatout <=0
Amics <=9
SI NO
SI NO
SI NO
65(P)
30(T)
22(P)
Chatin <=4
13(P)
29(L)
SI NO
10(T)
11(P)


Chatout <=2
SI NO
Chatin <=6
SI NO
Amics <=5
Chatout <=5
SI NO
SI NO
Inventari <=15
41(N)
81(P)
Chatin <=9
SI NO
SI NO
Felicitat <=368
Temps <=352
Felicitat <=144
46(P)
SI NO
SI NO
SI NO
Chatin <=4
15(T)
10(C)
10(T)
11(M)
12(T)
SI NO
119(M)
Chatout <=4
SI NO
16(P)
24(M)



Un cop finalitzat el procés d'elaboració del sistema, sobre tot la part de validació de regles del sistema expert amb l'arbre és precís realitzar una altra validació amb un conjunt més gran amb les noves dades obtingudes.

La validació s'ha realitzat emprant les dades dels nous usuaris de la xarxa social que no es varen emprar en la realització de l'arbre. Un cop seleccionats els usuaris i eliminats els poc rellevants (seguint el mateix criteri comentat abans) ens queden uns 44000 usuaris (8000 més que a la fase d'entrenament).

La validació s'ha realitzat efectuant el mateix procés que es feia als apartats d'investigació de l'arbre. L'arbre elaborat en aquest procés és molt similar a l'anterior i tan sols hi ha dues regles noves i no ho són del tot. S'ha confirmat que totes les regles excepte dos són pràcticament semblants a les anteriors però canvia el llindar de decisió. És a dir, per exemple un usuari era sociable si enviava 2,7 vegades més missatges que la mitja i tenia 1,4 vegades més amics que la mitja. La redefinició d'aquesta regla és que ara per tenir aquesta categoria necessita enviar 4,4 vegades més que la mitja i tenir 1,8 vegades més amics. No canvia la regla, tan sols el llindar.

El resultat d'aquesta nova classificació com a validació ha estat molt positiu donat que ha indicat que el que havíem realitzat a la fase d'investigació va ser correcte i que el treball fet en aquell moment segueix vigent.

El sistema expert original tenia 56 regles i tan sols s'ha modificat 2, és a dir un 3,57%.

Per a finalitzar, comentar que els valors que discriminen per seleccionar una branca o l'altre de l'arbre són valors mitjans, per tant hauran de seguir un manteniment puntual modificant aquestes mitges per tal que la seva efectivitat a l'hora d'encertar a quina classe pertany es mantingui.

b) Estudi amb classificador mapa de Kohonen

Es realitzarà un mapa de Kohonen que serà capaç de classificar 10 classes. Aquest mapa es farà amb una configuració inicial que haurà d'anar modificant fins a aconseguir la millor classificació possible.

Són xarxes neuronals autoorganitzades capaces de codificar i aprendre una sèrie de associacions estímul-resposta on es reconeix de forma automàtica conjunts d'estímul de característiques similars i a ells se'ls associa una única resposta. Una vegada la xarxa ha aprés, és capaç de reconèixer nous estímuls i associar-los amb les respostes apreses. Serveixen per a reconèixer patrons.

El concepte de semblança entre estímul i patró que representa una neurona és fonamental en l'algoritme competitiu d'aquest tipus de xarxa neuronal per decidir qual serà la neurona guanyadora.

S'estableix una associació entre l'espai continu d'entrada i el conjunt discret de neurones.

La neurona guanyadora rep com a premi la possibilitat de modificar el seu pes (el seu patró) per apropar-se encara més a l'estímul rebut sense oblidar als casos que ja representa.

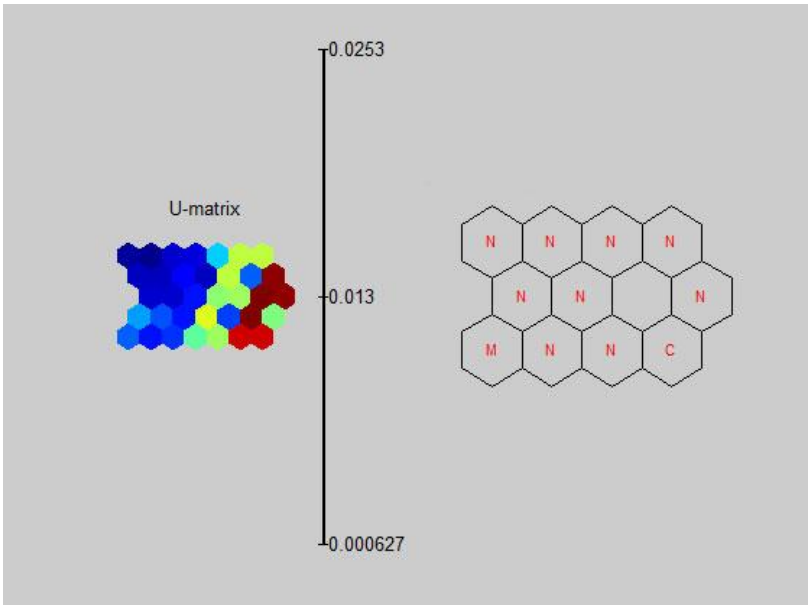
A efectes pràctics es defineix a voluntat del desenvolupador de la xarxa també una regió de veïnat de la neurona guanyadora en la que les neurones que estan incloses reben també el premi de poder modificar el seu pes.

Hem de tenir certa cura de triar el tamany de la regió de veïnat perquè només s'alterin part dels pesos de les neurones del mapa.

El paràmetre més important de una xarxa de Kohonen es el nombre d'elements que la conformen, es podria pensar que un nombre aproximat a les classes que s'han de classificar seria adequat, això es sap que no és així, però pareix un bon punt pel que començar, la primera configuració de la xarxa serà 3x4, donat que són 12 elements i s'han de classificar 10 classes.

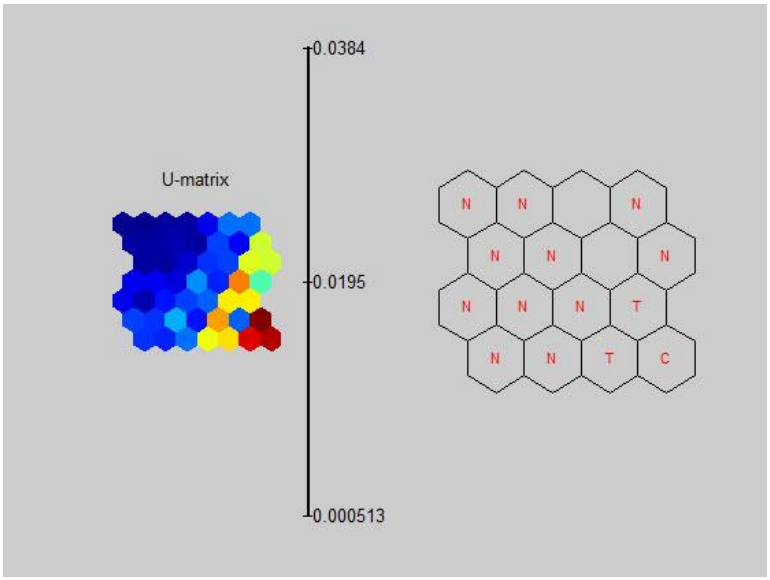
Cas1:

- Arquitectura: [3,4]
- Resultat: El resultat obtingut ha estat dolent, donat que de les 12 neurones, 10 queden com a classe N (classe predominant) una C i altre sense classificar, per la qual cosa sols reconeixeria 2 classes de 12, això no pot ésser i en la següent xarxa augmentarem el nombre de neurones.



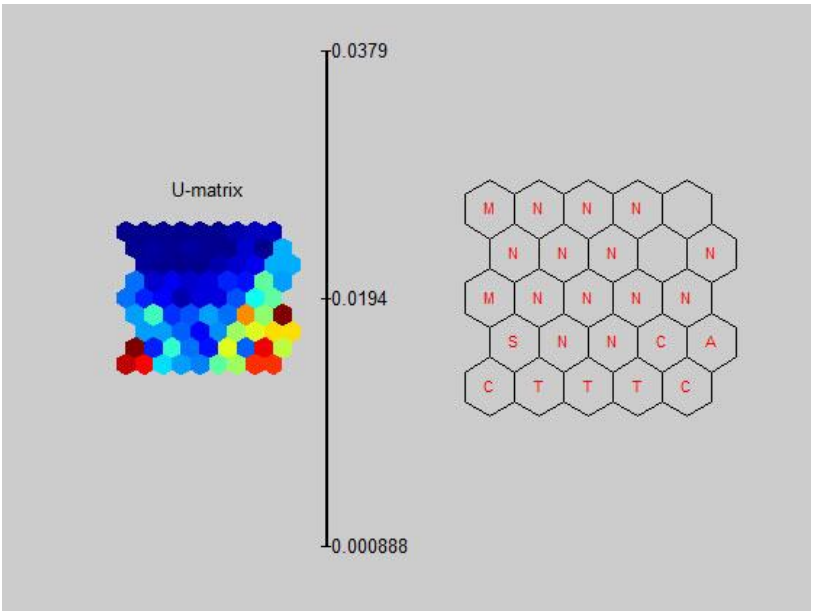
Cas2:

- Arquitectura: [4,4]
- Resultat: Nova iteració poc satisfactòria, però ha millorat donat que és capaç de reconèixer tres classes, millor que al cas 1, però no suficient, s'ha vist que la estratègia d'augmentar el nombre de neurones és positiva així que es seguirà fent.



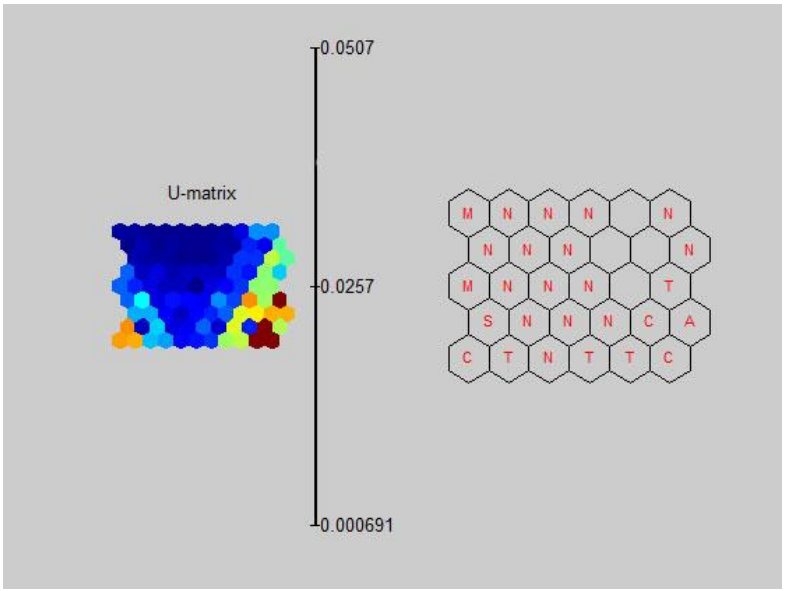
Cas3:

- Arquitectura: [5,5]
- Resultat: Una altra iteració que continua amb la tendència, millora la anterior i aquesta vegada més donat que ha passat de tres classes reconegudes a sis, això és fruit d'haver augmentat 9 neurones.



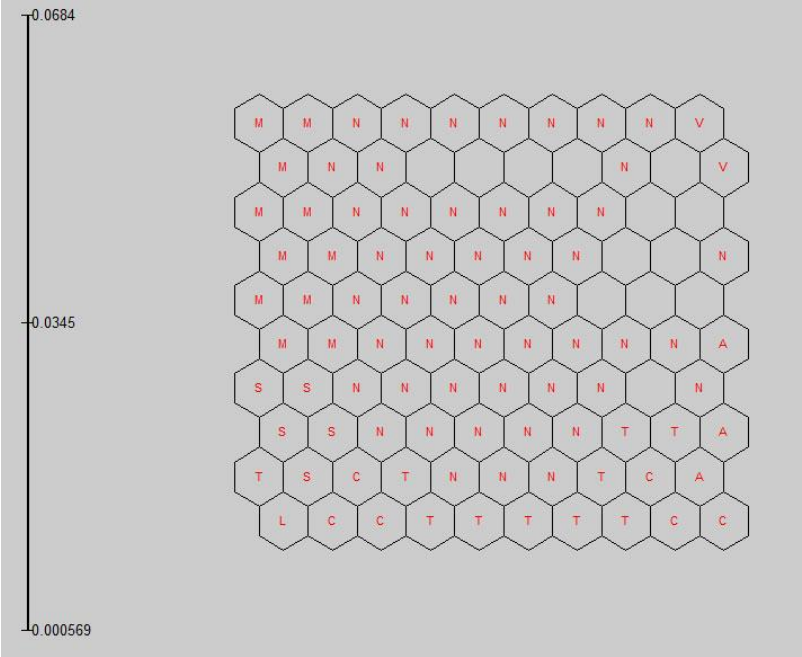
Cas4:

- Arquitectura: [5,6]
- Resultat: S'ha augmentat el nombre de neurones a cinc i segueix el mateix resultat, en aquest cas no hi ha millora, s'intentarà fer un augment més dràstic per veure si hi ha un gran salt en la millora i es planteja una xarxa de 10x10 a veure com respon.



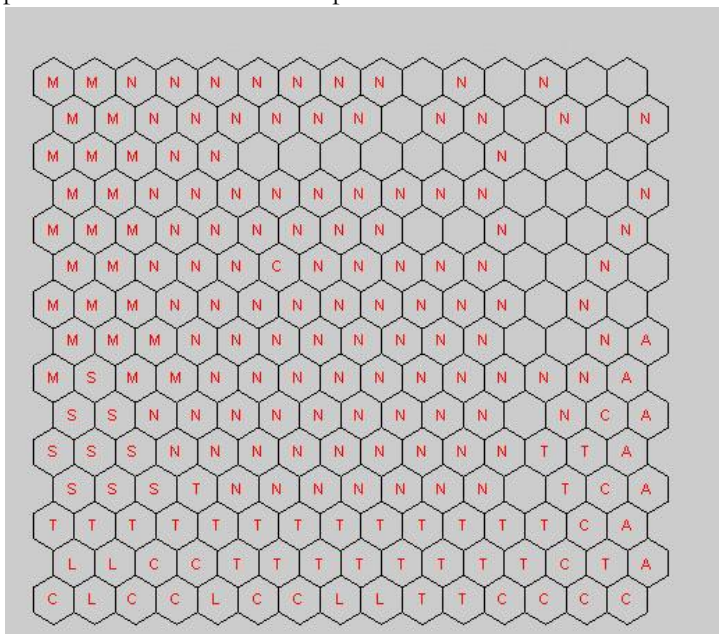
Cas5:

- Arquitectura: [10,10]
- Resultat: S'ha augmentat dràsticament el nombre de neurones i el resultat ha estat molt positiu, es reconeixen vuit classes de deu, només falten les classes R (usuari de xarxa social amb perfil rar) i P (usuari de xarxa social amb perfil popular). En aquest moment s'estan reconeixent les classes corresponents 35238 casos de 35542, només 304 no són reconeguts essent el percentatge d'encert reconeixement de classes del 99%. Això òbviament no significa que l'encert sigui del 99%, sinó que el 99% de les ocurrències tenen la possibilitat de esser classificades de manera encertada, mentre que el 1% està mal classificada segur. De totes formes s'intentarà arribar a que cada exemple tingui la possibilitat de caure a la seva classe.



Cas6:

- Arquitectura: [15,15]
- Resultat: S'ha augmentat dràsticament el nombre de neurones (més de 100) i segueix sense reconèixer-les, per tant és evident que no les pot reconèixer. Es continua intentant reconèixer les 8 amb el menor nombre de neurones, sabem que en 10x10 es reconeixen, però encara no sabem res respecte al mínim.

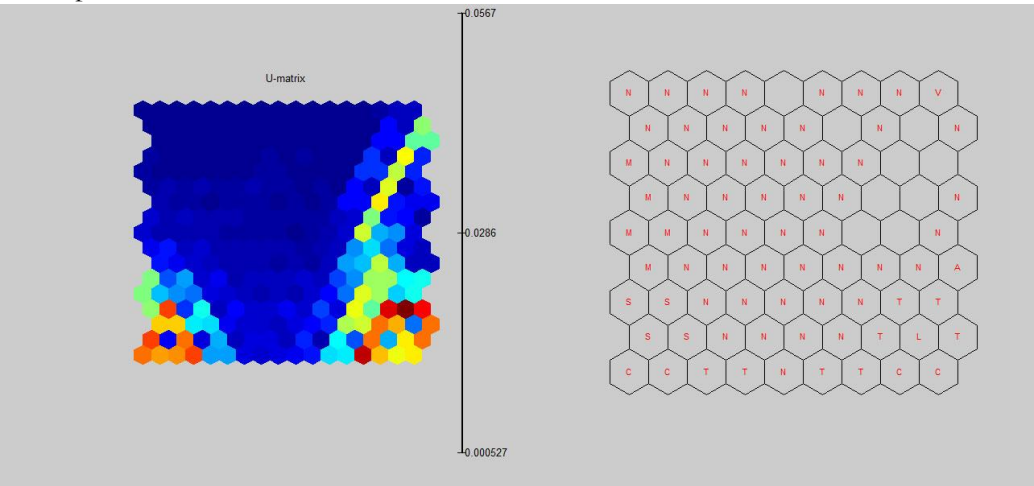


Característiques del mapa de Kohonen que s'utilitza com classificador

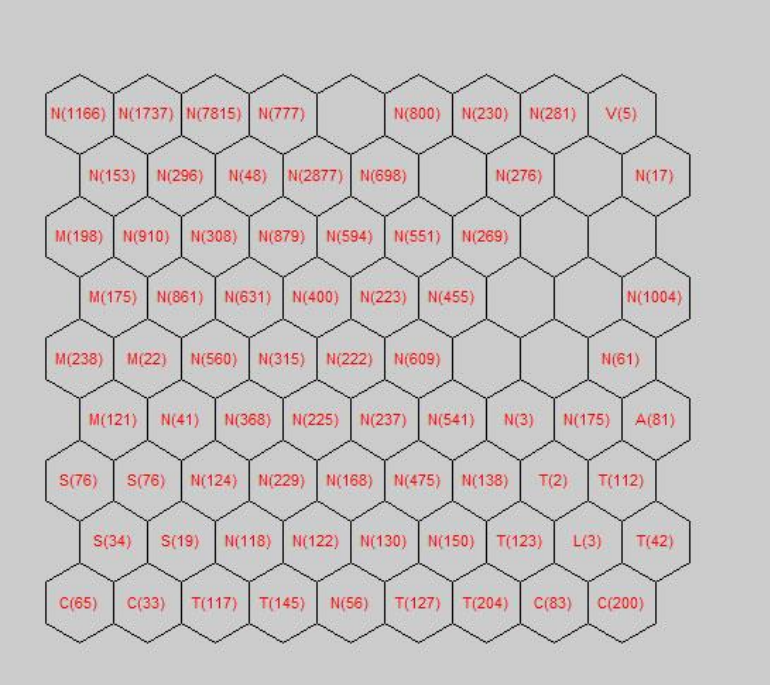
- Final quantization error: 0.007
- Final topographic error: 0.053

El quantization error és només de 7 mil·lèsimes i és l'error que es produeix per la conversió d'analògic a digital, és un bon valor.
Aquesta mesura considera la estructura del mapa. En un mapa que estigui retorçat de forma estranya, l'error tipogràfic és gran inclús si l'error de precisió és petit.
Una manera simple de calcular l'error topogràfic és: 1/N*SUMATORI(u(x)). En el cas el error topogràfic és petit amb la qual cosa es tracta d'un mapa simple.

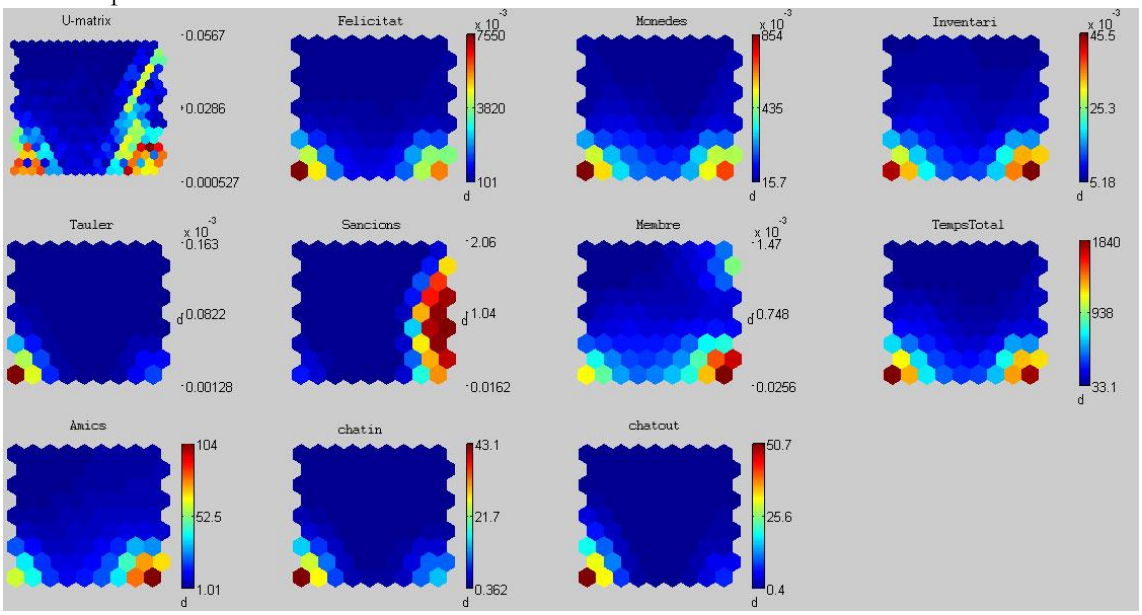
Classe predominant a cada neurona:



Nombre d'exemples per neurona:



U-matrix per atributs:

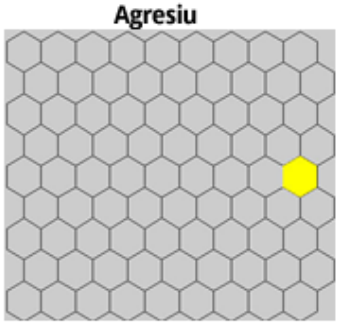


Les U-matrix visualitzen les distàncies entre les cèl·lules veïnes del mapa i ajuda a veure la estructura de grups del mapa. Els valors alts de la U de la matriu (colors vermell i groc) indiquen una separació alta. Els elements que són dels mateixos grups s'indiquen amb àrees uniformes de valors baixos (color blau).

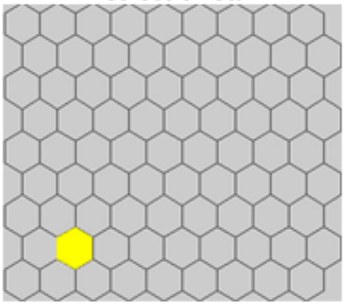
Proves dels perfils.

A continuació farem proves amb els diferents perfils que tenim a la nostra base de dades a fi de comprovar la validesa del mapa obtingut.

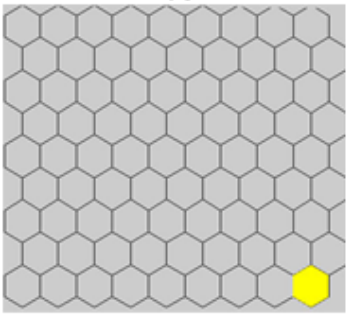
Agresiu



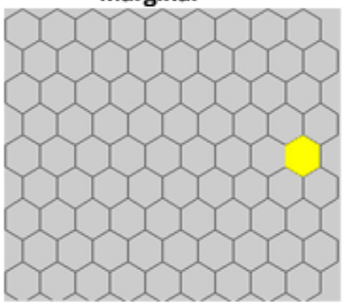
Coleccionista



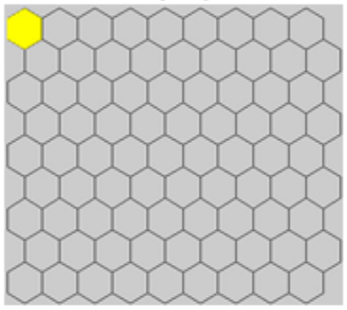
Lider



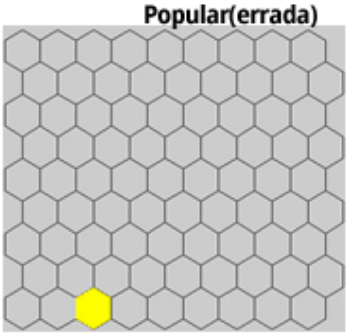
Marginal



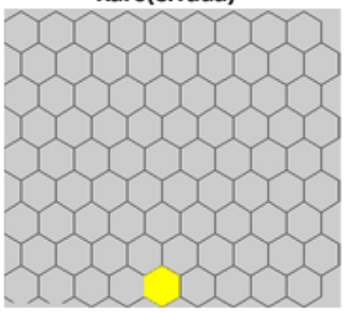
Normal



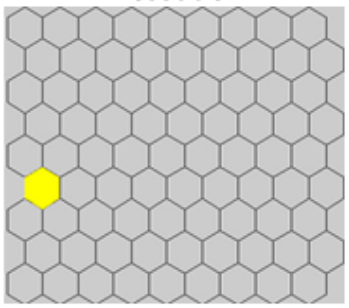
Popular(errada)



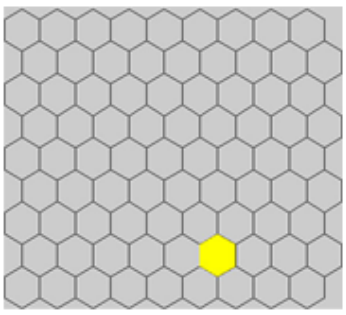
Raro(errada)



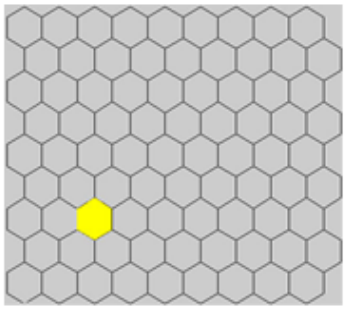
Sociable



Triomfador



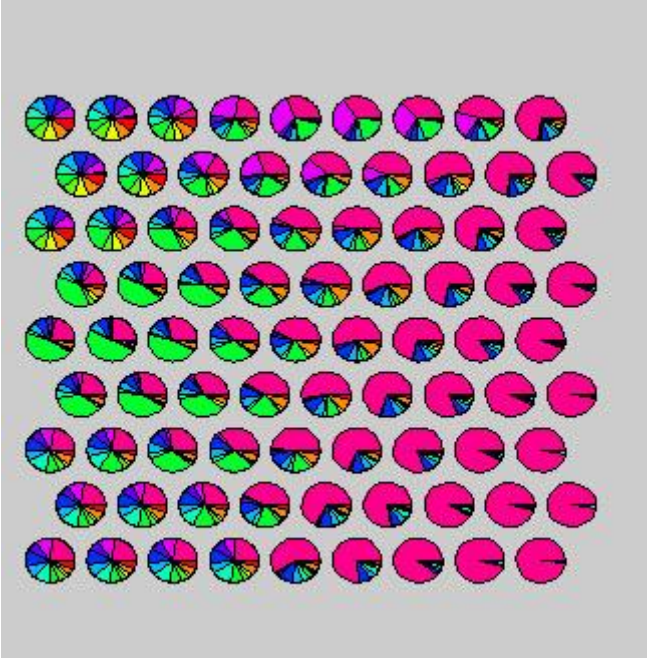
VIP



En aquestes gràfiques es mostren seleccionats uns exemples aleatoris de cada classe, a quina cèl·lula de la xarxa cauen. Als casos de P i R no poden caure a les seves cèl·lules donat que no en tenen, per la qual cosa l'errada està assegurada.

A continuació es presenten gràfiques que representen com són els exemples de cada neurona, es farà mitjançant "formatges", barres i representants. Cada color de les següents gràfiques representa una classe i el que pretenem veure és la distribució de les mateixes, és a dir, el grau de similitud entre si.

- Formatges

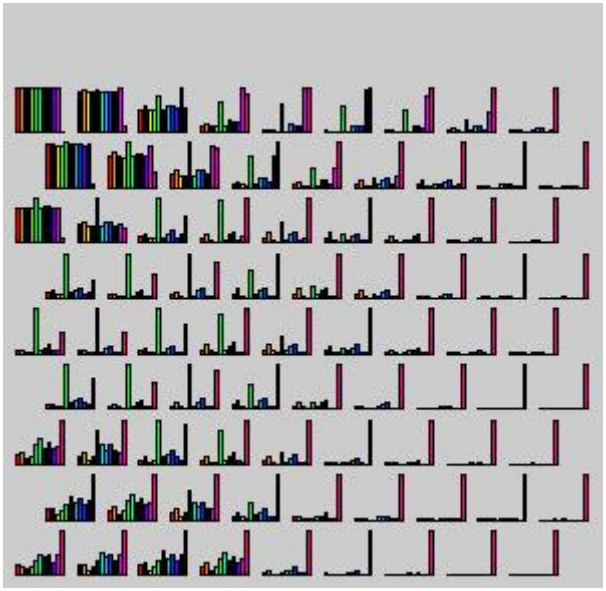


Als "formatges" es pot observar com estan distribuïdes les classes dintre una cèl·lula.

Es veuen algunes neurones que estan pràcticament copades per una sola classe (zona d' avall a la dreta), com altres que estan pràcticament igualades entre vàries classes (les d' amunt a la esquerra).

Quant més pures fossin les neurones, és a dir més majoria i menys varietat de classes tinguin, millor classificaran donat que en les que tenen moltes classes només una encerta, les demés són errors de classificació.

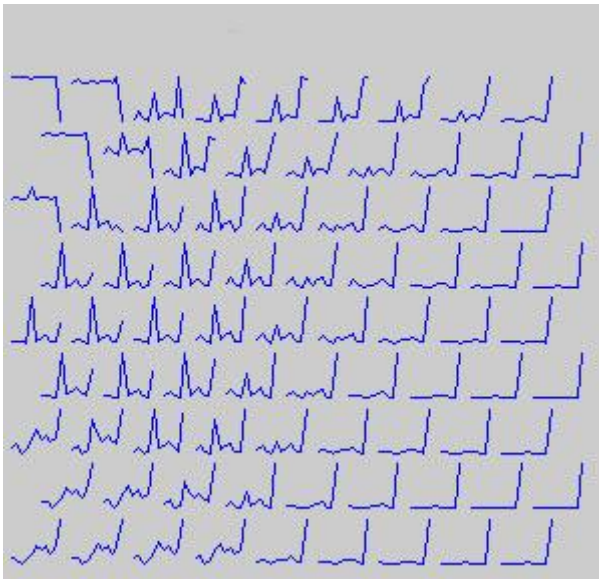
- Barres



Al diagrama de barres es pot veure alguna cosa molt similar, donat que les mateixes neurones que abans tenien poca varietat i un percentatge molt alt d'una classe, ho tornen a tenir amb una sola barra gran, mentre que les que eren més variades, tenen barres de quasi totes les classes (colors en aquest cas) i d'una altura semblant.

Això vol dir que hi ha neurones molt pures, que classifiquen molt bé i altres pitjors que tenen més varietat. Aquest diagrama i l'anterior expliquen el mateix.

- Representants



El diagrama de representants mostra exactament el mateix que els dos anteriors, mostren els vectors del prototipus, però en aquest cas en lloc de amb barres o formatges ho fan amb gràfiques.

Es pot observar exactament el mateix que als altres dos casos, neurones molt pures amb pràcticament exemples d'una sola classe avall a la dreta i neurones amb molta varietat de classes a la part superior esquerra.

c) Estudi amb perceptró multicapa

L'entrenament d'aquestes xarxes, es basa en la presentació successiva i de forma reiterada, de parells de vectors en les capes d'entrada i sortida (vectors entrada i sortida desitjada).

La xarxa crea un model a base d'ajustar els seus pesos en funció dels vectors d'entrenament, de forma que a mesura que es passen aquests patrons, per cada vector d'entrada la xarxa produirà un valor de sortida més similar al vector de sortida esperat.

Aquestes xarxes també es diuen de retropropagació (backpropagation), nom que ve donat pel tipus d'aprenentatge que utilitzen.

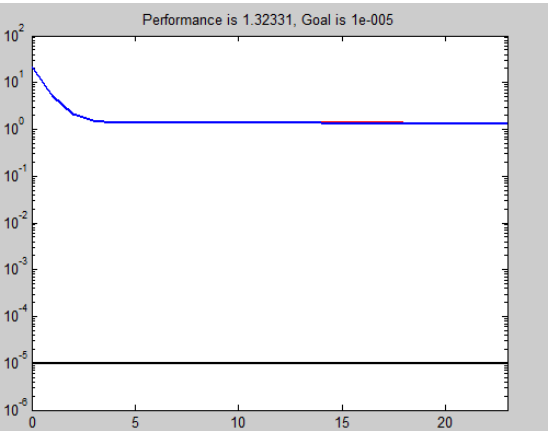
Els perceptrons multicapa amb aprenentatge de retropropagació són una variació del model ADALINE (Widrow et al., 1960) [1], que empra la regla Delta com una forma d'aprenentatge (aquesta regla d'aprenentatge, es fonamenta en la utilització de l'error entre la sortida real i esperada de la xarxa per modificar els pesos).

Aquestes xarxes accepten la regla Delta de tal forma, que es faciliti l'entrenament de totes les connexions entre els diferents nivells de la xarxa.

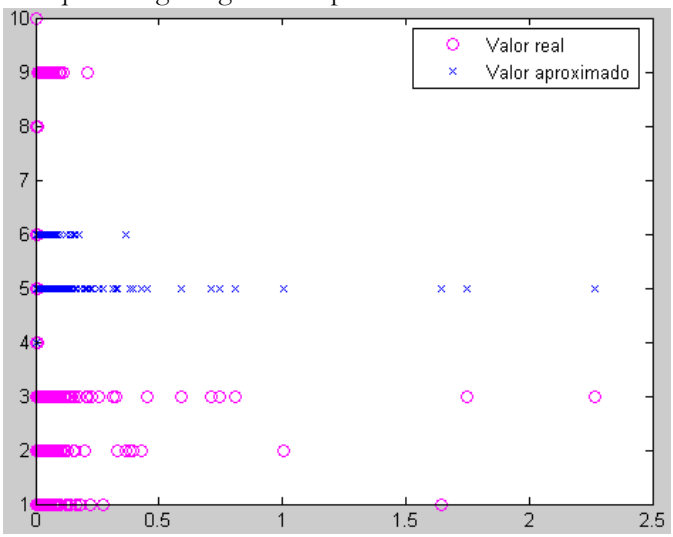
Estructura d'un cas:

- Arquitectura: [Neurones capa oculta, Neurones capa sortida]
- Iteracions: Nombre d'iteracions màximes de l'entrenament
- Factor d'aprenentatge: Factor d'aprenentatge emprat
- Error obtingut: Error total comés
- Resultat: El més rellevant de les gràfiques obtingudes si procedeix

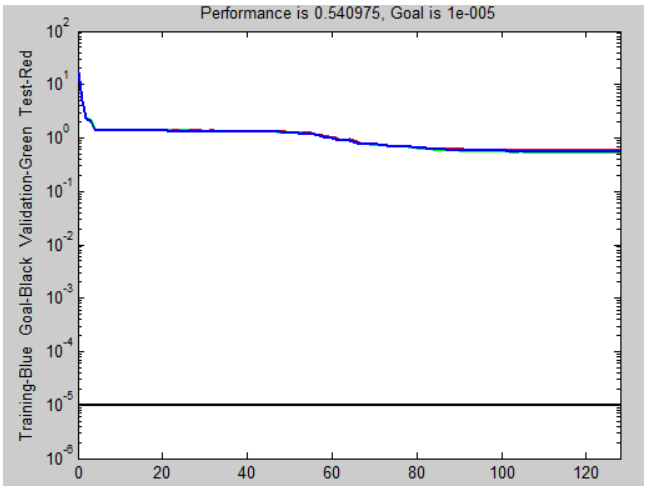
Cas1:
-Arquitectura: [20,10,1]
-Iteracions: 5.000
-Factor d'aprenentatge: 0.001
-Error obtingut: 1.32331
-Resultat: El resultat obtingut al ser d'una primera iteració es pot considerar com a vàlid, encara ha de millorar molt, donat que té un error alt i només classifica 2 classes de 10.



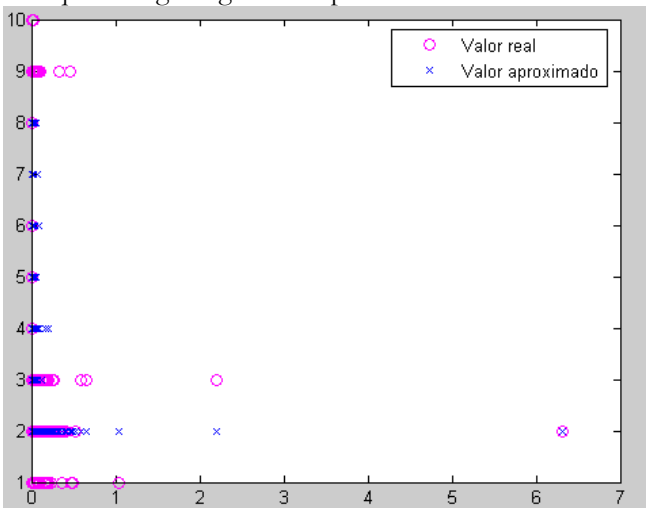
En aquesta segona gràfica es pot veure com només es classifiquen les classes 5i6:



Cas2:
-Arquitectura: [20,10,1]
-Iteracions: 5.000
-Factor d'aprenentatge: 0.005
-Error obtingut: 0.540975
-Resultat: El resultat obtingut ha estat molt millor que l'anterior reduint l'error a menys de la meitat i classificant 7 classes de 10, en lloc de 2 de 10.

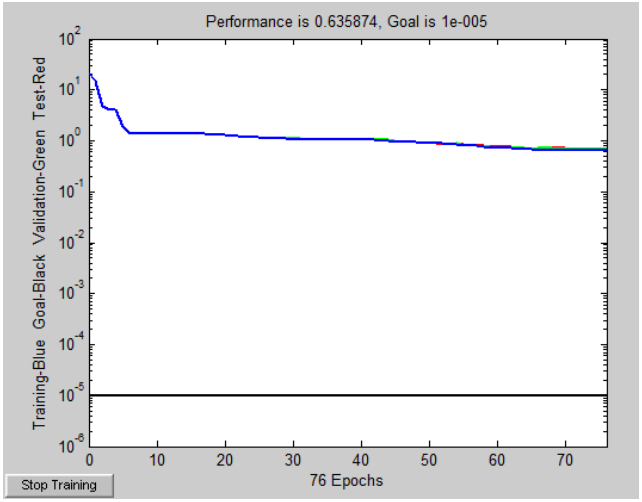


En aquesta segona gràfica es pot veure com la classificació de les classes 1,9 i 10 no es fa:



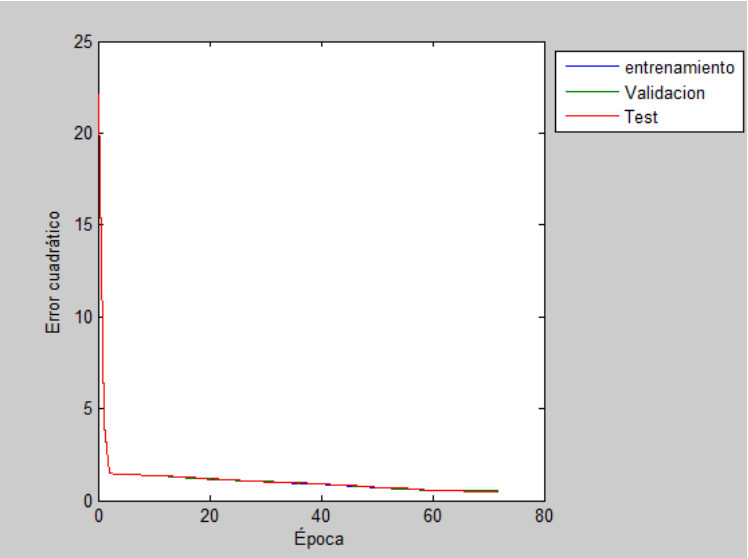
Cas3:

-Arquitectura: [20,10,1]
-Iteracions: 5.000
-Factor d'aprenentatge: 0.01
-Error obtingut: 0.635874
-Resultat: El resultat obtingut ha estat lleugerament pitjor que amb la iteració anterior, al haver augmentat el factor d'aprenentatge, amb el que es tornarà a la situació anterior (0.005) i s'intentarà modificar altres paràmetres.



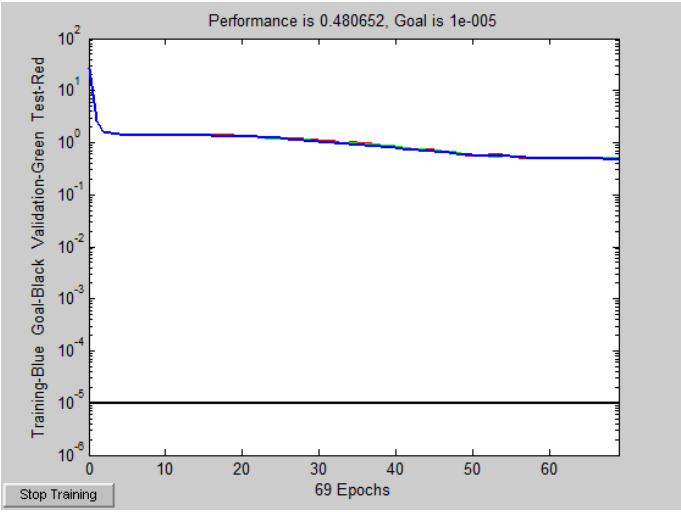
Es canviarà la estructura de neurones, de [20,10,1] a 30 en la primera capa oculta.

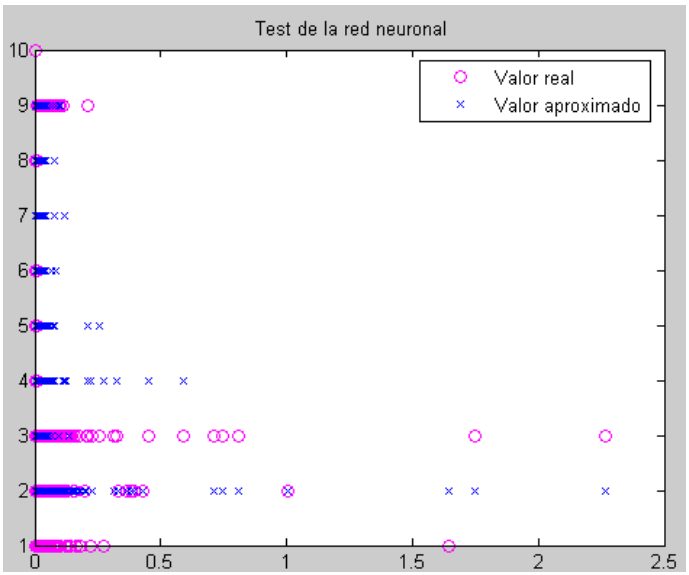
Cas4:
-Arquitectura: [30,10,1]
-Iteracions: 5.000
-Factor d'aprenentatge: 0.01
-Error obtingut: 0.505339
-Resultat: S'ha millorat molt lleugerament respecte al cas anterior, molt poc com es pot veure en el camp error obtingut. Donat que la decisió pareix haver estat bona es tornarà a modificar la distribució de les neurones. Es mostra la disminució de l'error quadràtic mitjà:



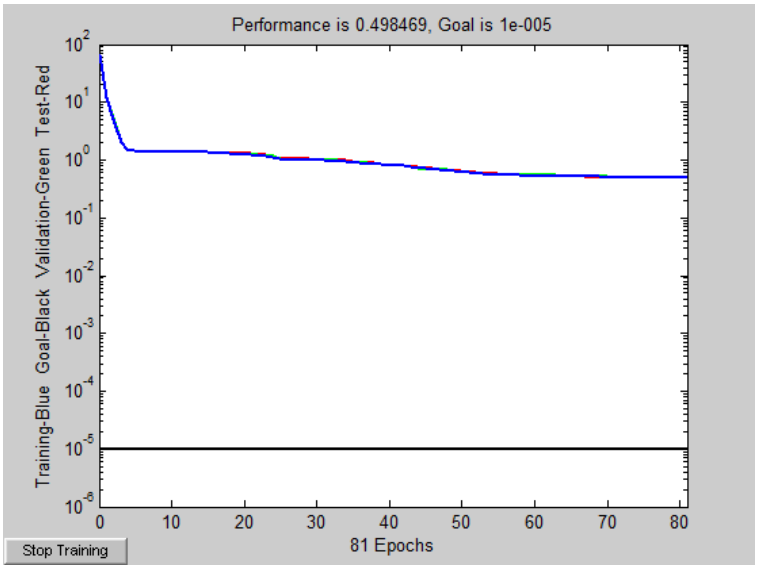
Es canviarà la estructura de neurones, de [30,10,1] a [40,20,1]

Cas5:
-Arquitectura: [40,20,1]
-Iteracions: 5.000
-Factor d'aprenentatge: 0.01
-Error obtingut: 0.480652
-Resultat: Aquesta iteració amb el PMC ha estat molt bona, no per l'error que ha disminuït però de manera molt lleugera, sinó perquè s'ha aconseguit classificar una classes més, la 9, això es molt important per l'estudi.



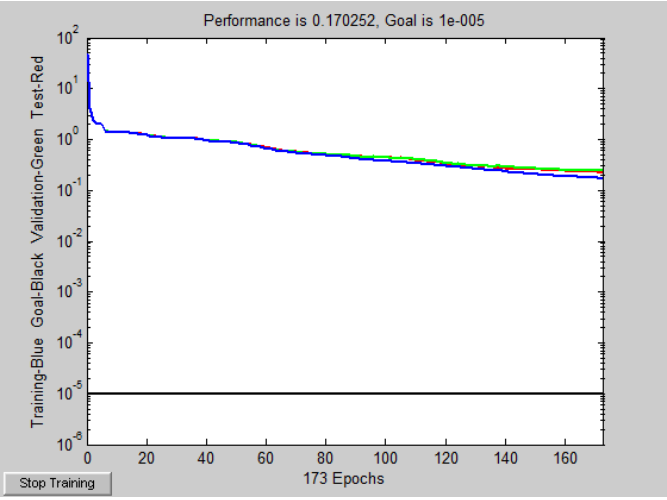


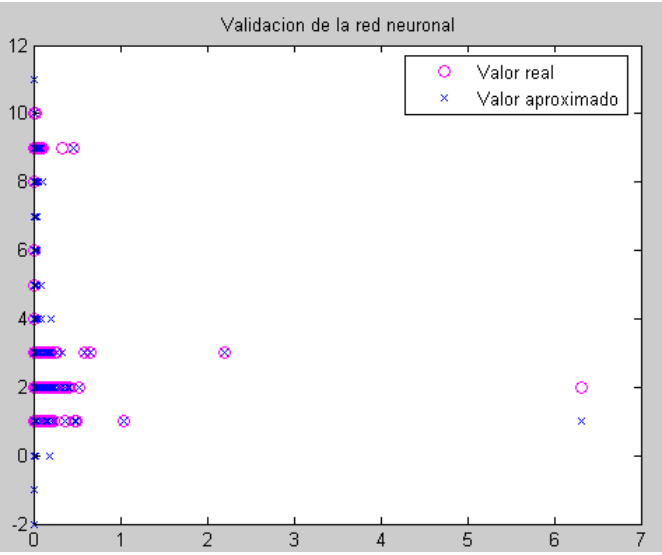
Cas6:
-Arquitectura: [50,25,1]
-Iteracions: 5.000
-Factor d'aprenentatge: 0.01
-Error obtingut: 0.498469
-Resultat: El resultat obtingut augmentant el nombre de neurones en les capes ocultes ha estat pràcticament igual però molt més costós en quant a temps de execució. Les classes són les mateixes així que la iteració ha estat pitjor que l'anterior.



S'intentarà en les següents iteracions modificar altres factors per veure si encara es pot millorar el resultat. Es començarà pel factor d'aprenentatge.

Cas7:
-Arquitectura: [40,20,1]
-Iteracions: 5.000
-Factor d'aprenentatge: 0.01
-Error obtingut: 0.170252
-Resultat: En aquesta iteració amb el PMC ha estat molt satisfactòria, s'ha disminuït l'error de aproximadament 0.48 a aprox. 0.17 Això ha estat per haver augmentat el factor d'aprenentatge de 0.01 a 0.1, i és curiós que aquest mateix moviment amb altres paràmetres en la configuració de la xarxa neuronal no havia estat positiu, però ara si, aquesta classificació ha estat la millor amb diferència, no només perquè l'error hagi disminuït sinó perquè aconsegueix classificar perfectament totes les classes com es pot veure en la gràfica que es veu a continuació.



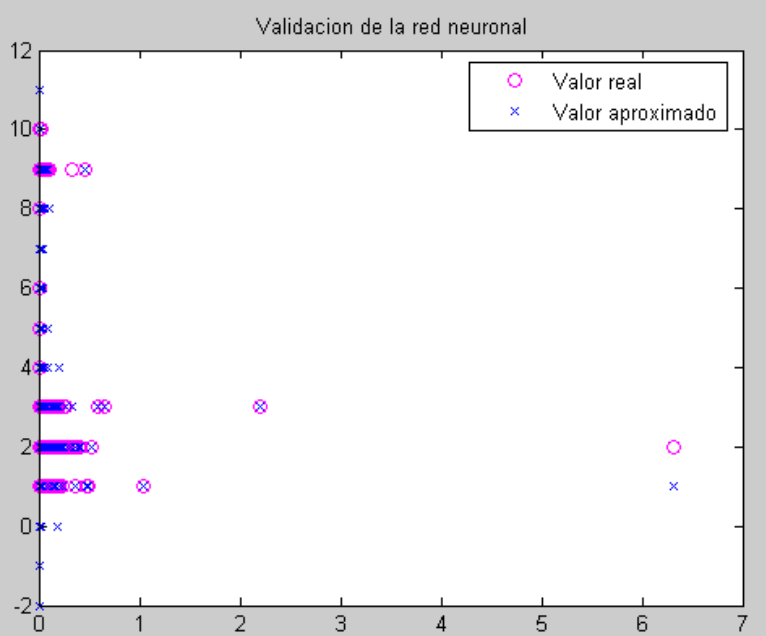


c) Comparacions, conclusions i elecció del millor

La classificació del PMC ha estat satisfactòria essent capaç de reconèixer totes les classes. L'error obtingut ha estat de 0.17 en el millor dels casos i és la millor classificació possible.

S'exposen a continuació el cas en qüestió i la classificació dels exemples:

-Arquitectura [40,20,1]
-Iteracions: 5.000
-Factor d'aprenentatge: 0.1
-Error obtingut: 0.170252



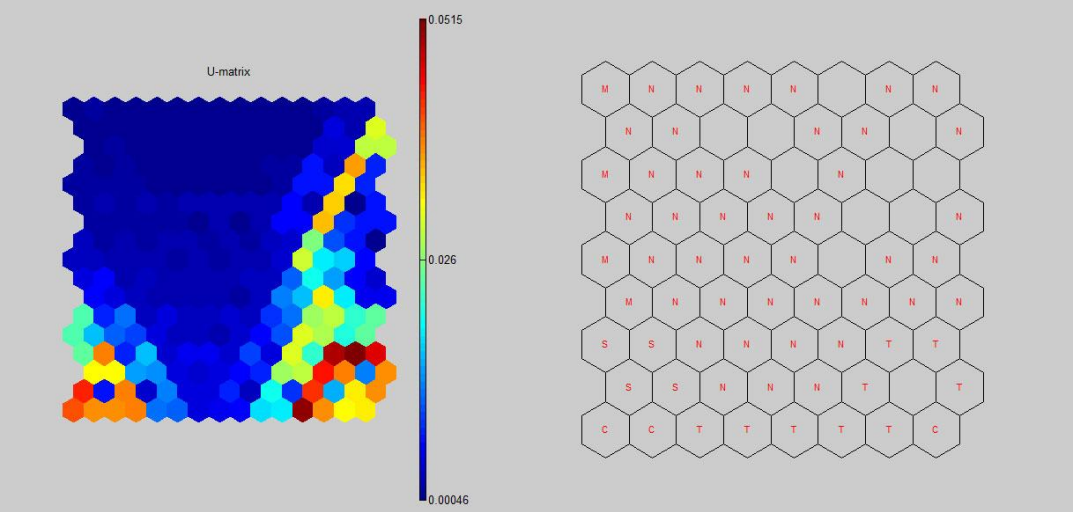
Totes les classes existents (1-10) estan representades a la classificació.

Comparació (PMC & Mapa de Kohonen) vs Arbre de decisió

És fonamental que totes les classes estiguin perfectament classificades i reflectides al classificador.

És per això que el mapa de Kohonen desenvolupat no és capaç de classificar dues classes (una d'elles sense molta importància donat que són els 'rars', 2 exemples entre 35.000 i que podria ésser ignorada, però l'altre, els usuaris populars si que és important pel sistema, i té una gran representació en els exemples) no seria en cap cas un classificador vàlid, és per això que encara tenint un percentatge d'encert alt no es considera bo. S'ha provat fins amb 225 neurones i no es classifiquen correctament, és possible que amb milers si ho faci, però el projecte treballarà a un entorn de producció i per tant la eficiència és molt important, i 225 pareix ja un nombre molt elevat, per això no s'ha provat amb més.

A continuació es mostra com el mapa no classifica les classes R i P.



Entre l'arbre i el perceptró, els quals són capaços de reconèixer totes les classes possibles, amb el que haurà d'entrar a valorar l'error d'entrenament, la eficiència que poden tenir al sistema i l'error de classificació amb exemples no emprats per l'entrenament.

Error al entrenament
Arbre: ~0.13
Perceptró: ~0.17

Nombre de nodes
Arbre: 23 nodes
Perceptró: [40,20,1] = 61 nodes

Error de classificació
Arbre: 10%
Perceptró: 40%

En tots i cada un dels paràmetres l'arbre és superior al perceptró, és més eficient i classifica millor tant als exemples d'entrenament com als de comprovació.

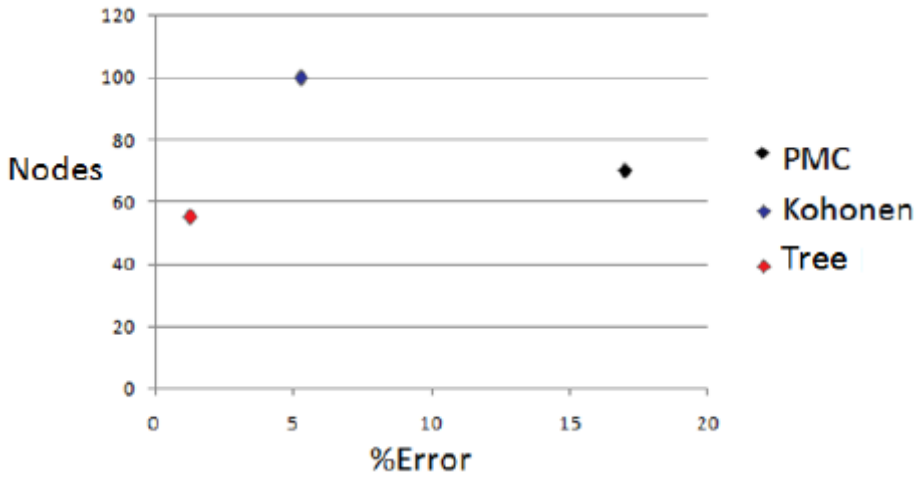
Es triarà l'arbre com millor classificador, ja quan es va fer va parèixer un classificador excel·lent i comparant-ho amb altres alternatives com s'ha fet es veu que és millor.

Destaquem per veure que l'arbre és excel·lent, que l'error de classificació després de la seva elaboració és només el 10% i a més de les proves que es feren amb les 10 classes, la que no reconeix és la R, que només té 2 elements de més de 35.000, motiu pel qual totes les classes rellevants estan perfectament reflectides.

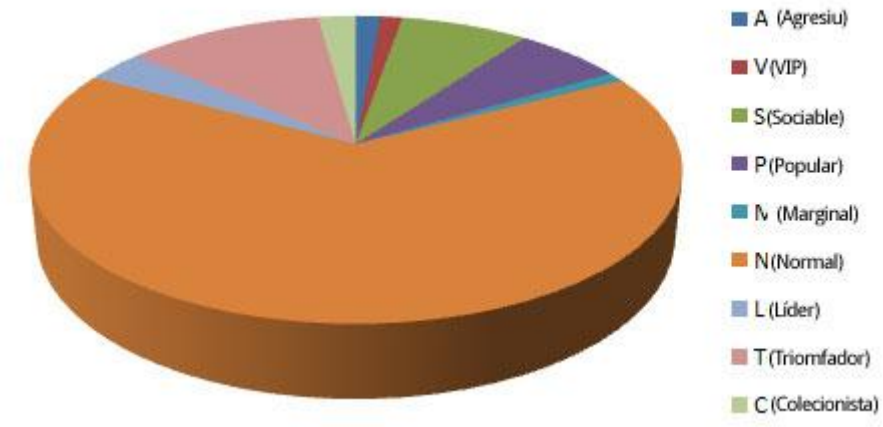
4. CONCLUSIONS FINALS

Els resultats més rellevants corresponen a la fase d'investigació a on hem avaluat tres alternatives molt completes.

Les alternatives a triar han donat molts bons resultats en general encara que l'arbre està molt per damunt de les demés com podem veure a continuació.



Per una altra banda, els resultats en quant a classificació d'usuaris dintre el sistema emprant l'arbre desenvolupat, han quedat de la següent forma.



El projecte ha estat un èxit tant per l'empresa contractant com per l'autor. Fruit d'aquest treball s'ha demostrat la efectivitat de les tècniques de Intel·ligència Artificial per a detectar patrons socials a una xarxa social.

Aquestes tècniques també han demostrat la seva eficàcia en problemes canviants donat que una xarxa social i molt més una infantil canvien constantment i és important que el fruit d'aquest treball tingui una durabilitat en el temps i quedi demostrat en la validació del sistema expert.

5. BIBLIOGRAFIA

[1] Widrow et al. "Adaptative Sampled-Data Systems – a statistical theory of adaptation" <http://www-isl.stanford.edu/~widrow/papers/c1959adaptivesampled.pdf>