Memòria TFC

Mineria de dades - 2012

.: Xavier Marín Gómez – Enginyeria Tècnica de Gestió

Consultor: Ramón Cahiuelas Quiles

UOC Universitat Oberta de Catalunya

Proposta d'un model predictiu del risc d'ingrés hospitalari amb l'ús d'algoritmes bayesians

Continguts

Índex de figures i taules

SUMARI

Aquest projecte pretén trobar una eina que permeti identificar els pacients que probablement ingressaran a un hospital segons un seguit d'atributs que obtindrem d'analitzar diferents apartats de l'assistència, com poden ser l'activitat urgent i les característiques pròpies.

No hi ha a l'atenció primària un model acceptat universalment que ens identifiqui el risc d'ingrés hospitalari, i, en general, tampoc no hi ha la cultura d'emprar la modelització de pacients, la classificació de case mix o de la complexitat dels pacients.

Els models de risc d'ingrés hospitalari, amb la seva càrrega de consum de recursos associada, la complexitat, s'estan avaluant, actualment, com a possibles eines per a l'ajust del model capitatiu (pressupost ajustat segons el consum de recursos per càpita al territori).

En el cas concret de l'atenció primària, el fet de poder identificar grups de pacients amb unes característiques o atributs similars i un consum de recursos associat, com és l'ingrés hospitalari, permet una visió addicional o un factor més d'ajust en diferents àmbits i pot marcar les polítiques preventives adreçades als pacients tributaris.

L'atenció primària és un banc de dades excel·lent, ja que no només es disposa d'informació dels problemes de salut dels pacients, sinó que, a més, es disposa de la seva activitat i dades demogràfiques, la qual cosa permet analitzar, o si més no intuir, possibles predictors de risc aplicables en el nostre àmbit.

Paraules clau

Data Ware, ETL, Ingressos hospitalaris, Atenció primària de salut, Oracle, Weka, Elvira, SQL, Model predictiu, Xarxa bayesiana, Magatzem de dades.

Àrea

Mineria de Dades

INTRODUCCIÓ I JUSTIFICACIÓ

El nostre Sistema Nacional de Salut garanteix a tot el poble uns serveis de salut, accessibles, gratuïts i amb la màxima qualitat, sobre la base d'un dels principis fonamentals, el que estipula que la salut de la població és responsabilitat de l'Estat i que els serveis de salut estan al abast de tot el poble. La xarxa assistencial actual permet portar els més recòndits llocs l'atenció mèdica. A la base de l'estructura del sistema, esquematitzada en forma de piràmide, es troben les unitats assistencials del primer nivell d'atenció, constituïdes pels hospitals rurals i els serveis atenció continuada i urgent de base territorial en l'àmbit de l'atenció primària.

Entre els serveis que ofereixen aquestes unitats assistencials es presenta amb relativa i variable freqüència la demanda d'una atenció d'urgència, que cobra major rellevància quan l'àrea de salut es troba allunyada dels centres assistencials hospitalaris o en poblacions rurals disperses, amb serveis d'urgències condicionats per aquest tipus d'atenció. Les atencions d'urgències en els serveis d'AP han augmentat amb el temps. L'objectiu d'aquest treball és mostrar un model predictiu dels pacients utilitzadors d'aquests serveis, per tal de detectar aquells usuaris que podrien ser abordats, de forma preventiva, per evitar que haguessin de recórrer a aquest servei urgent.

La tasca educativa i les activitats preventives sobre alguns col·lectius són molt important per evitar la sobrecàrrega dels serveis d'urgències, així com la prevenció de les complicacions de les malalties de base. Aquesta sobrecàrrega consumeix temps, recursos humans i materials que una detecció precoç podria identificar i corregir. D'altre banda, la intervenció precoç preventiva sobre aquests col·lectius, a mes de tenir un efecte beneficiós sobre el sobreconsum dels recursos sanitaris, produeix un clar benefici sobre la salut d'aquests usuaris, que poden veure resoltes les seves afeccions sense haver de recórrer a un servei urgent.

A diferència de moltes disciplines en les què ja s'ha aplicat la mineria de dades, la medicina té una metodologia d'investigació sòlida, establerta i acceptada, per el que l'aplicació de les tècniques de mineria de dades , tot just contemplades en el sector de la medicina, hauran de demostrar la seva vàlua per tal de ser preses seriosament.

El teorema de Bayes va ser desenvolupat en 1764 pel matemàtic i religiós Thomas Bayes amb el propòsit de calcular les probabilitats condicionals, és a dir la probabilitat que un esdeveniment passi (o no passi) a condició que un altre esdeveniment previ hagi ocorregut. [1] El Teorema de Bayes no ha perdut actualitat i, per contra, cada vegada s'ho proposa més com un mètode d'anàlisi en moltes àrees del coneixement científic. [2] En medicina clínica es recomana el seu ús, entre altres coses, per esbrinar la probabilitat d'un diagnòstic a condició que un signe o símptoma sigui present. [3 i 4]

Emprarem la lògica del Teorema per buscar la relació existent entre la presència de patologies cròniques o politractaments i la necessitat d'ingressar finalment per un usuari de la salut és molt diferent si la estudiem en la població general, entre els que consulten un centre d'atenció urgent, o entre els que ja han ingressat prèviament en un hospital. La població de malalts i no malalts, amb la qual el metge treballa, és realment una mostra representativa de la població, que ens permetrà interpretar els resultats d'un test aplicat a un pacient individual. Amb aquest projecte tractarem de comprendre la lògica que explica aquest fenomen.

PLANIFICACIÓ DEL PROJECTE (Project Plan)

Per poder realitzar adequadament la planificació del projecte, cal conèixer en primer lloc les tasques a realitzar, distribuir-les correctament en el temps i adaptar-les a les dates de desenvolupament proposades en el pla de treball del projecte.

En aquest apartat, prenent com a base les fases de la metodologia CRISP per a la implantació d'un projecte de mineria de dades, es comparen i combinen aquestes amb les fases proposades en el programa de l'assignatura i finalment, s'adapten i integren les fases resultants a les dates claus previstes en el del pla docent. Finalment es realitza una distribució de la càrrega de treball dins de cada fase per aconseguir una planificació equilibrada.

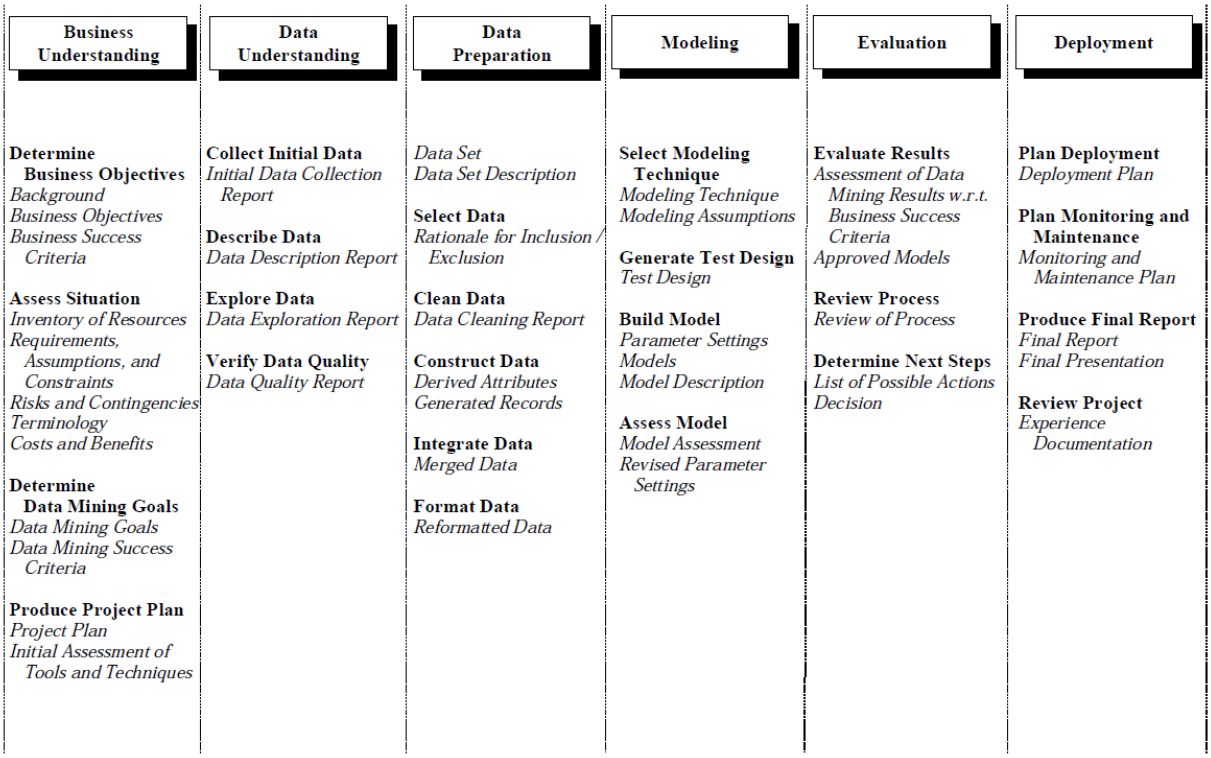


Figura 1. Les tasques genèriques (negreta) i sortides (cursiva) del model de referència CRISP-DM

Fases del projecte

INICI DEL PROJECTE:

Descàrrega i lectura del material del TFC.

1: Pla de Treball i objectius (Business understanding):

Elaboració del Pla de Treball, on s'indicarà la planificació estimada de les diferents tasques a realitzar. També es descriuran els objectius generals i del projecte.

A causa de la particularitat d'aquest TFC, la instal·lació del suport de programari s'ha dut a terme al llarg de diferents etapes. En aquesta fase, s'instal·laran aplicacions com OpenProj, Toad, MySQL, MySQL Workbench, MS Excel i Word.

2: Contextualització i estat de l'art (Data understanding).

Elaboració d'un document que descrigui l'estat actual dels sistemes de gestió del coneixement i la Mineria de dades. També es recollirà documentació sobre les aplicacions en el terreny de la salut i l'ús que es farà d'aquets processos en el projecte escollit.

3: Disseny i Implementació (Data preparation & Modeling).

Aquesta fase constarà de les següents tasques:

- Depuració i tria de dades: base de dades, càrregues, preparació de les dades, etc.
- Instal·lació de l'eina d'explotació de les dades (WEKA, Elvira).
- Construcció dels informes i anàlisi de la informació.

Les eines emprades, que s'hauran d'instal·lar seran del tipus opensource. El paquet Pentaho que inclou, entre d'altres, WEKA i el programa ELVIRA.

MEMORIA y PRESENTACIÓN VIRTUAL (Evaluation).

Entrega de la memòria final del projecte.
Elaboració de la documentació de la memòria final així com d'una presentació virtual del projecte.

i finalment **DEBAT**.

OBJECTIUS i PLA DE TREBALL (Business Understanding)

Objectius

En aquest apartat es descriuen els objectius del projecte entès com a treball final de carrera i en relació amb els objectius generals de l'assignatura, així com els objectius específics del projecte com a tal.

Objectius del TFC

L'objectiu fonamental d'aquest TFC és la realització d'un projecte de mineria de dades partint de la informació procedent de bases de dades transaccionals.

Aquest objectiu general del projecte es pot descompondre en una sèrie d'objectius puntuals més concrets que són:

- Consolidar els coneixements prèviament adquirits sobre com elaborar un document d'especificacions basat en els requeriments d'usuari
- Obtenir experiència en la planificació i seqüenciació d'un projecte de mineria de dades
- Adquirir coneixements sobre el disseny i maneig de dades
- Obtenir experiència en l'explotació de bases de dades
- Entendre la importància de l'ús de llenguatges com PL / SQL en la implantació, gestió i explotació d'un magatzem de dades
- Obtenir experiència en l'ús de les eines que faciliten l'explotació de les dades i la generació d'informes

Objectius concrets del projecte

L'objectiu principal d'aquesta investigació és desenvolupar un projecte de mineria de dades utilitzant tècniques de modelatge, en el nostre cas una xarxa bayesiana del tipus Naive Bayés, per tal de trobar un patró o model predictiu dels usuaris que acudeixen a urgències d'atenció primària.

Els objectius d'aquest treball són els següents:

1. Implementar un model predictiu dels usuaris que es visiten a l'atenció primària de salut i acabaran ingressant en un hospital.

2. Agrupar els pacients que acudeixen a atenció primària per les seves necessitats per tal de discriminar aquells que podien beneficiar-se d'una actuació preventiva.

3. Detectar variables i factors diferencials entre els usuaris que acudeixen a urgències i aquells que no ho fan, presentant altres característiques comuns, per tal d'avaluar els trets diferencials.

4. Esbossar algunes recomanacions per descobrir coneixement en les dades electròniques d'aquest usuaris, a través de mineria de dades.

5. Trobar correlacions amb altres sistemes d'agrupació/clusterització dels usuaris, així com les hospitalitzacions a un tercer nivell.

Diagrama de Gantt

La planificació indicada a la taula anterior es pot veure representada gràficament en forma de diagrama de Gantt en la gràfica següent:

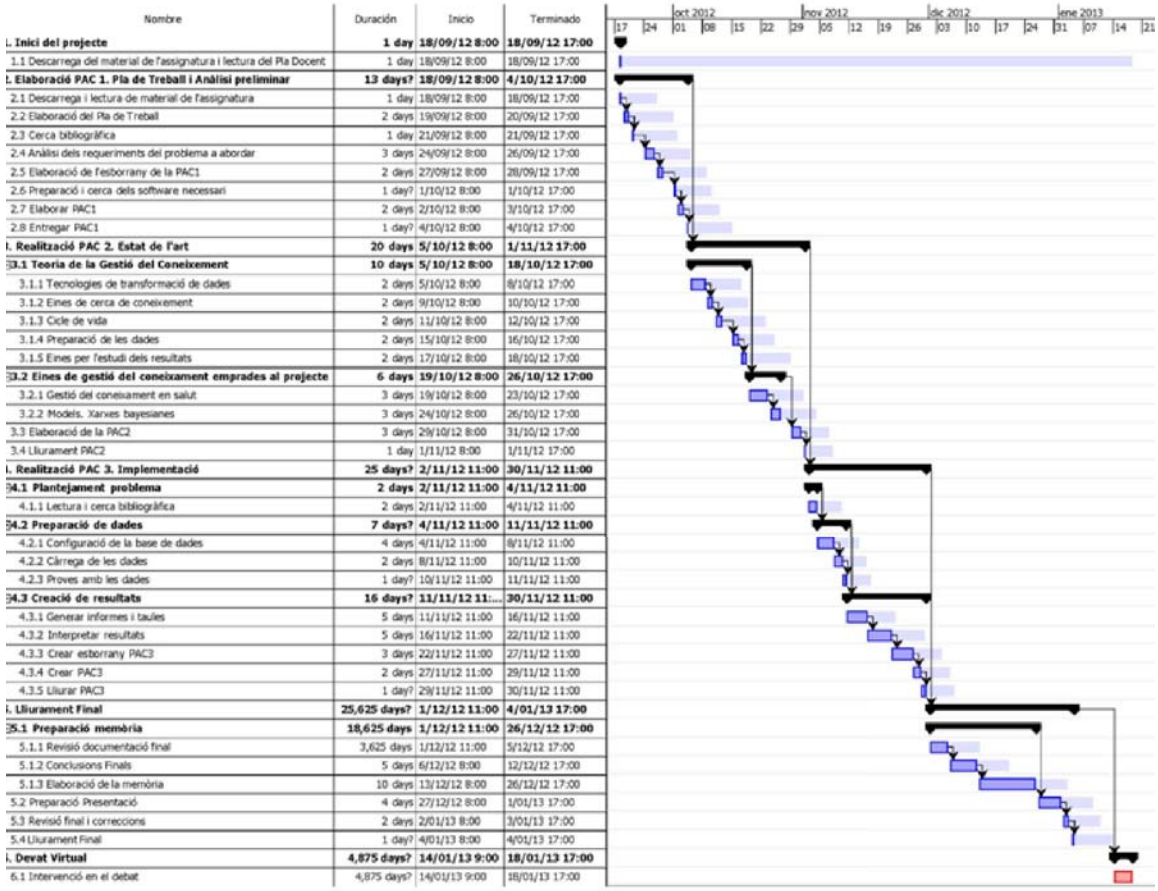


Figura 2. Diagrama de Gantt del projecte

CONTEXTUALITZACIÓ I ESTAT DE L'ART (Data Understanding).

Mineria de dades en Salut

En aquest treball es presenta un projecte per la recerca de predictors, a través de tècniques d'agrupament, portades a terme amb programari lliure. La investigació es basa en un cas d'estudi, des d'on, a través de programari lliure (Weka i Elvira) es presenten predictors en l'àmbit de la salut. A través del nostre cas d'estudi no només es proven algoritmes d'agrupament sinó que també es fan propostes per trobar metes addicionals.

En l'última dècada el nombre d'estudis sobre els factors que poden intervenir en l'avaluació i mesura dels recursos necessaris en l'atenció sanitària d'una població ha crescut de manera exponencial, de manera que l'interès per millorar el procés d'assignació de recursos ha estat cada vegada més àmpliament tractat per els experts en els sectors i els investigadors científics. Les diferències de comportament dels usuaris de la sanitat que determinen les necessitats en tota mena de recursos que necessita un sector, no només es pot entendre com a causa de l'envelliment poblacional o del tipus de cobertura, sinó que s'ha d'entendre, també, com una conseqüència de la mala previsió d'unes necessitats que ens venen marcades per variables més o menys perdibles i per tant, més o menys influenciables a priori. Per tant, la millora dels processos d'aprenentatge sobre les variables que influencien en l'ús dels recursos sanitaris ha de desenvolupar-se a partir d'un punt de partida que requerirà de l'aportació de tanta informació com sigui possible sobre aquests usuaris (totes les característiques que poden tenir un efecte potencial sobre les necessitats dels usuaris en una forma o una altra): d'una part les característiques personals (sexe, edat, etc.) i d'altre el seu comportament quan a l'ús dels serveis realitzats per l'usuari. L'anàlisi d'aquesta potencial quantitat d'informació ens condueix a la necessitat d'utilitzar tècniques de mineria de dades.

La mineria de dades és també coneguda com el descobriment de coneixement en bases de dades (KDD) i és una branca del camp de l'anàlisi genèric de dades, que te com a principal objectiu extreure coneixement a partir d'una gran quantitat de dades (grans conjunts de dades) continguda en tot tipus de bases de dades. El procediment experimental adaptat a un procés de mineria de dades típic inclou les següents fases: l'estat del problema i formulació de la hipòtesi, recol·lecció de dades, pre-processar les dades, estimar el model i, finalment, interpretar, el model i treure'n les conclusions.

En aquest treball es presenta un cas d'estudi en l'àmbit de la salut, per a això s'apliquen les diferents fases esmentades en un procés de mineria de dades, a partir de conjunts de dades que inclouen les dades dissociades d'usuaris/pacients, així com els ingressos i visites a urgències dels mateixos amb l'objectiu principal d'aconseguir una conclusió que ens permetin millorar el procés d'assignació de recursos i de planificació preventiva.

Breu història de la Mineria de dades en l'àmbit de la salut

La pràctica d'utilitzar dades concretes i evidències per recolzar les decisions mèdiques (també conegut com medicina basada en l'evidència o MBE) ha existit durant segles. John Snow, considerat com el pare de l'epidemiologia moderna, va utilitzar mapes amb les primeres formes de gràfics de barres en 1854 per descobrir la font de còlera i provar que va ser transmesa a través del subministrament d'aigua, per sota (Tufte 1997).

Snow va comptar el nombre de morts i es representen les adreces de la víctima al mapa com negre bars. Es va descobrir que la majoria de les morts agrupades cap a una bomba d'aigua específica a Londres (Centre del cercle vermell al mapa).

Florence Nightingale va inventar els diagrames d'àrea en 1855 (a baix) per mostrar que l'increment de la mortalitat es podia deure a les pràctiques insalubres dels clínics i, per tant evitables. Utilitzant els diagrames va convèncer els responsables polítics per implementar les reformes i reduir el nombre de morts (Audain 2007). (Diagrama de Nightingale 1858.)

Snow i Nigthtingale van ser capaços de recollir, garbellar i analitzar les dades per que en el seu temps el volum d'informació era manejable. Avui dia, la mida del les poblacions estudiades, la quantitat de dades electròniques recollides, juntament amb la globalització i la velocitat de brots de malalties, fan que sigui gairebé impossible d'aconseguir el que van fer aquests pioners.

Aquí és on la mineria de dades arriba a ser útil a la salut. S'ha fet poc a poc, però cada vegada s'està aplicant mes per afrontar els diversos problemes de descobriment de coneixement en salut.

La mineria de dades i la seva aplicació a la medicina i la salut pública és un camp relativament jove de estudi. El 2003, Wilson et al van començar a estudiar els casos en què les tècniques de KDD i mineria de dades s'aplicaven en bases de dades de salut. Van trobar força confusió ; "Alguns autors es refereixen a la mineria de dades com el procés d'adquisició d'informació, mentre que altres es refereixen a la mineria de dades com la utilització de tècniques estadístiques en el descobriment de coneixement procés ". (Wilson et al. 2003)

La definició generalment acceptada avui en dia sobre la mineria, és el conjunt de procediments i tècniques per descobrir i descriure els patrons de les tendències en les dades (Witten i Frank, 2005).

Problemes i desafiaments

L'aplicació de la mineria de dades en el camp de la medicina és una tasca molt difícil a causa de la idiosincràsia de la professió mèdica. Treballs de Shillabeer i Roddick (2007) citen diversos conflictes inherents als diferents enfocaments metodològics entre les metodologies tradicionals en medicina i la mineria de dades.

En la investigació mèdica, la mineria de dades comença amb una hipòtesi i després els resultats s'ajusten per encaixar en la hipòtesi. Això s'aparta de la pràctica estàndard de la mineria de dades, en que simplement s'inicia amb el conjunt de dades, sense una hipòtesi prèvia evident.

A més, mentre que la mineria de dades tradicional es preocupa sobre els patrons i les tendències en sèries grans de dades, la mineria de dades en medicina està més interessada en aquella minoria que no

s'ajusta als patrons i les tendències. El que augmenta aquesta diferència d'enfocament és el fet que l'extracció de dades en la majoria de mineries de dades es preocupa sobretot de la descripció però no d'explicar els patrons i tendències. Per contra, en medicina són necessàries aquestes explicacions perquè una petita diferència podria canviar l'equilibri entre la vida o la mort.

Per exemple, l'àntrax i la grip comparteixen els mateixos símptomes de problemes respiratoris. Baixant el senyal de llindar en un experiment de mineria de dades podríem donar l'alarma de l'existència d'un àntrax quan només es tracta d'un brot de grip. El contrari és encara més greu: un brot de grip percebut resulta ser una epidèmia d'àntrax (Wong et al 2005). No és casualitat que trobem que, en la majoria de les publicacions de mineria de dades sobre la malaltia i el tractament, les conclusions van ser gairebé sempre vagues i cauteloses. Molts reporten resultats encoratjadors però es recomana un estudi addicional. El fet de no ser concloent indica l'actual manca de credibilitat de les dades la mineria en aquests nínxols particulars de salut.

La confusió sobre la definició de la mineria de dades també complica l'assumpte. Per exemple, trobem un parell de publicacions amb les paraules clau "mineria de dades" en els seus títols, però resulta ser una simple utilització de gràfics. Shillabeer (2009) va dir que aquest malentès és freqüent en l'existència relativament jove de la mineria de dades en l'assistència sanitària. Fins i tot si els resultats de mineria de dades són creïbles, convèncer els professionals de la salut per canviar els seus hàbits basats en l'evidència pot ser un problema major. Ayres (2008) informa d'un parell de casos on els metges de l'hospital es van negar a canviar la política de l'hospital, fins i tot quan s'enfrontaven a les proves.

En un cas, es va demostrar que els metges que sortien de les autòpsies sense rentar-se les mans provocaven una alta probabilitat de mort entre els pacients tractats a posteriori. Tot i presentar aquesta evidència, els metges encara es negaven a canviar els seus hàbits.

Shillabeer (2009) també van reportar que la majoria dels metges (almenys a Austràlia) prefereixen escoltar un respectat líder d'opinió en la professió mèdica, en lloc de resultats demostrats procedents de la mineria de dades.

La privacitat dels registres i l'ús ètic de la informació del pacient és també un gran obstacle per a la mineria de dades en l'assistència sanitària. Per a la mineria de dades per ser més exactes, es necessita una quantitat considerable de registres reals. Els registres sanitaris són informació privada i confidencial, no obstant això, l'ús d'aquests registres privats poden ajudar a aturar i prevenir algunes malalties actualment greus o mortals.

Models Gràfics Probabilístics

Podem dir que la Intel·ligència Artificial té com a objectiu la creació de programes que realitzin tasques que requereixin un comportament intel·ligent.

A la nostra manera de relacionar-nos amb el món utilitzem, gairebé sense adonar-nos, incertesa, és a dir, no tenim un coneixement exacte del que passa al nostre voltant. Per exemple, no ens recordem literalment el que deia el noticiari del dia anterior però, si li prestem una mica d'atenció, sabem quines van ser les notícies més rellevants del dia. Per tant, si volem que un programa imiti el comportament de l'ésser humà, desenvolupant amb problemes del món real, hauria de ser capaç de manejar la incertesa

existent sobre el problema que està tractant. Des de fa uns anys la Intel·ligència Artificial ha dedicat un considerable esforç al tractament de la incertesa. D'entre tots els mètodes que s'han proposat, la Teoria de la Probabilitat és la més clàssica i la més coneguda, sobretot la òptica bayesiana. Les xarxes bayesianes són una eina que ha demostrat la seva capacitat com a model de representació del coneixement amb incertesa en Intel·ligència Artificial, sent capaces d'adaptar amb èxit un gran nombre d'aplicacions pràctiques.

Les xarxes bayesianes permeten representar el coneixement de manera gràfica i compacta, usant els conceptes de probabilitat i causalitat o independència entre les variables d'un problema, de manera molt semblant al mateix l'ésser humà. Però a més de ser fàcilment interpretables, tenen la capacitat d'obtenir explicacions de la informació que representen i s'adapten amb facilitat davant l'arribada de nova informació.

Sistemes experts probabilístics.

Un sistema expert el podem definir com un sistema informàtic que simula els experts humans en una àrea d'especialització determinada[19]. Els sistemes experts que tracten la informació d'una manera determinista són poc realistes, ja que el coneixement humà és majoritàriament heurístic, és a dir, aproximat. Si volem construir un sistema expert ho haurem de dotar de la capacitat per raonar de manera aproximada, o el que ve a ser el mateix amb incertesa.

Els sistemes experts probabilístics utilitzen la probabilitat com a mesura d'incertesa en els seus raonaments. No obstant això, en els primers sistemes experts es van utilitzar factors de certesa [20] com a mesura per tractar la incertesa però requerien molta informació i uns càlculs massa complexos per poder resoldre problemes reals en què intervinguessin un gran nombre de variables.

Amb l'aparició dels primers models gràfics probabilístics [21] (entre els quals destaquem les xarxes de Markov i les xarxes bayesianes) es va comprovar que les dificultats en l'ús de la probabilitat eren superables, de fet, actualment la probabilitat és la mesura d'incertesa més acceptada.

La idea bàsica que rau en els models gràfics probabilístics és codificar el coneixement de manera que no sigui necessari utilitzar informació irrellevant i, per tant, en treballar amb una complexitat menor, disminuir la complexitat dels càlculs. El que es fa és aprofitar les relacions de dependència i independència entre les variables d'un problema -codificades de manera gràfica en el model-, abans d'especificar i calcular els valors numèric de les probabilitats. Aquestes relacions es representen mitjançant models gràfics, habitualment grafs dirigits acíclics.

Teorema de Bayes: aquest teorema ens permet representar la probabilitat condicionada *p (y\x)* mitjançant la següent expressió:

$$p(y|x) = \frac{p(x|y)p(y)}{p(x)}$$

Si $p(x) = \sum_{y \in \Omega_Y} p(x,y)$ $p(x,y) = p(x|y)p(y)$, podem representar el teorema de Bayes utilitzant la següent expressió:

$$p(y|x) = \frac{p(x|y)p(y)}{\sum_{y \in \Omega_Y} p(x|y)p(y)}$$

On distingim:

■ La probabilitat *p(y)* s'anomena probabilitat marginal, a priori, o inicial de Y = y ja que pot ser obtinguda abans de conèixer l'evidència, és a dir, no té en compte cap informació de X = x.

■ La probabilitat *p (y \ x)* és la probabilitat posterior, a posteriori, o condicional de Y ja que s'obté després de conèixer l'evidència, és a dir, depèn del valor x.
■ La probabilitat *p (x \ y)* se l'anomena versemblança i és la probabilitat de l'observació X = x donat Y = y.

Xarxes bayesianes.

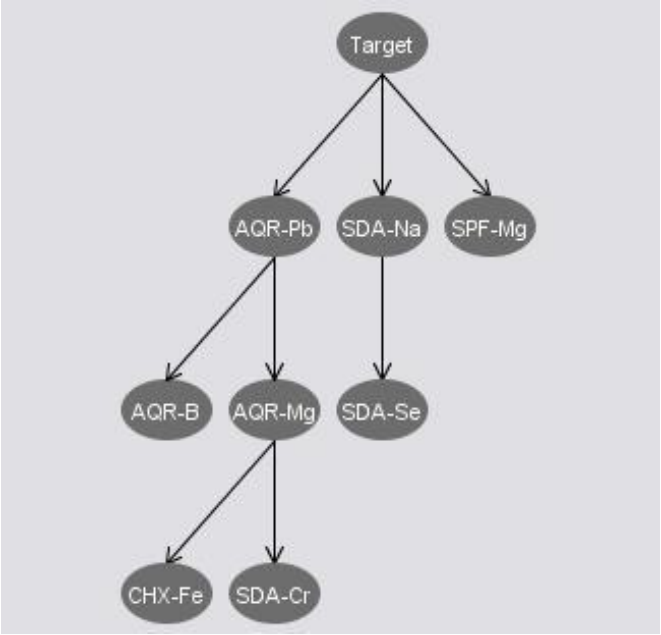
L'avanç important que ha tingut el camp de la història clínica informatitzada ha portat com a conseqüència un augment significatiu en la quantitat de dades que s'emmagatzemen en moltes ocasions en diferents formats. La mineria de dades és un mecanisme que ens permet facilitar la recerca d'informació valuosa en grans volums de dades. Aquest treball té com a objectiu aplicar una tècnica predictiva al camp de la salut com ho és el de la prevenció.

Les xarxes bayesianes és una tècnica que pertany al grup de les tècniques de classificació i consisteix en un model gràfic que utilitza arcs per formar una gràfica acíclica i és aplicat en aquelles situacions en què la incertesa s'associa amb un resultat que es pot expressar en termes de probabilitat. És a dir els nodes del graf representen variables, els arcs representen dependència condicional i una distribució de probabilitat.

En un començament, aquests models eren construïts a mà basats en un coneixement expert, però en l'actualitat s'han investigat tècniques per aprendre d'aquestes dades, tant l'estructura com els paràmetres associats al model [22].

Aquesta tècnica busca determinar relacions causals que expliquin un fenomen i és aplicat en aquells casos que són de caràcter predictiu.

És a dir el raonament probabilístic o propagació de probabilitats consisteix en difondre els efectes de l'evidència per mitjà de la xarxa per conèixer la probabilitat a posteriori de les variables. És a dir a determinades variables (conegudes) se'ls atorga una probabilitat i en base a això s'obté una probabilitat posterior[23].



Hi ha certs conceptes bàsics que estan associats a aquesta tècnica com:

- **Graf**: Parell de conjunts G = (X, L) on X és un conjunt finit d'elements (nodes) i L és un conjunt d'arcs.
- **Arc**: subconjunt de parells ordenats.
- **Graf Dirigit**: Parell ordenat G = (X, L) on X és el conjunt de nodes i L conjunt d'arcs.
- **Graf acíclic**: graf que no té cicles.

Una xarxa bayesiana G defineix una distribució de probabilitat conjunta única sobre U donada per:

$$PB(X1, X2, ..., Xn) = \pi PB(X i \mid \pi X i)$$

Qualsevol sistema de classificació de patrons es basa en el següent: donat un conjunt de dades (que dividirem en dos conjunts d'entrenament i de test) representats per parells <atribut, valor>, el problema consisteix en trobar una funció f (x) (anomenada hipòtesi) que classifiqui aquests exemples.

La idea d'usar el teorema de Bayes en qualsevol problema d'aprenentatge automàtic és que podem estimar les probabilitats a posteriori de qualsevol hipòtesi consistent amb el conjunt de dades d'entrenament per així seleccionar la hipòtesi més probable. Per estimar aquestes probabilitats s'han proposat nombrosos classificadors bayesians.

Un classificador en general subministra una funció que classifica una instància especificada per una sèrie de característiques o atributs, en una o en diferents classes predefinides. En general els classificadors bayesians són àmpliament utilitzats pel fet que presenten certes característiques:

- Són simples de construir i comprendre.
- El procés d'inducció a partir dels mateixos és ràpid i senzill.
- Robust quant considera atributs irrellevants.
- Considera una important quantitat d'atributs per generar la predicció final.

Un classificador bayesià pot ser un cas particular d'una xarxa bayesiana, on hi ha una variable que compleix el rol de la classe i els altres variables són considerades atributs. L'estructura dependrà fonamentalment del tipus de classificador.

Els classificadors de xarxes bayesianes són:

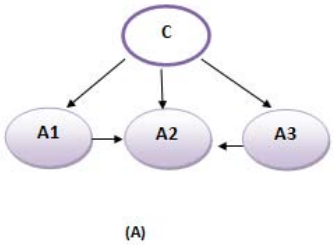
Classificador bayesià Simple (naïves Bayes classifier, NBC):

Permet obtenir la probabilitat posterior de cada classe Ci, usant la regla de Bayes. Aquest classificador assumeix que els atributs són independents entre si donada la classe, així que la probabilitat es pot obtenir pel producte de les probabilitats condicionals individuals de cada atribut.

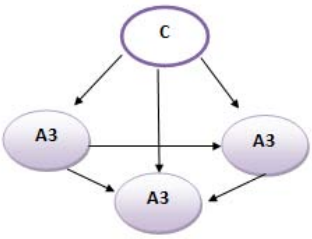
La representació gràfica pot donar-se com una xarxa bayesiana en forma d'estrella. És a dir un node arrel, que representa la variable de la classe i que està connectada als atributs. Aquest classificador bayesià té extensions i el fonament del seu ús és quan es disposen d'atributs que són dependents.

Una manera de considerar aquest dependència és estenent l'estructura bàsica de NBC, incorporant arcs a aquests nodes. Les possibilitats bàsiques són:

- TAN: classificador bayesià simple augmentat amb un arbre (A).
- BAN: classificador bayesià simple augmentat amb una xarxa (B).



(A)



(B)

Xarxes bayesianes dinàmiques:

Aquest classificador permet representar l'estat de les variables en un cert moment de temps. En el cas d'existir la necessitat de representar aquests processos dinàmics existeix aquesta extensió coneguda com xarxa bayesiana dinàmica (RBD)[24].

Per a les xarxes bayesianes dinàmiques, generalment es fan les següents suposicions:

- *Procés Markovià*: l'estat actual només depèn de l'estat anterior.
- *Procés estacionari en el temps*: les probabilitats condicionals en el model no s'alteren amb el temps.

L'aprenentatge de les xarxes bayesianes consisteix a inferir un model, estructures i paràmetres, a partir de les dades, que pot ser agrupat en dos aspectes: Aprenentatge Estructural: obté l'estructura de la xarxa bayesiana (o topologia de xarxa) prenent com a punt de partida una base de dades, és a dir les relacions de dependència entre les variables involucrades.

D'acord al tipus d'estructura, podem dividir els mètodes d'aprenentatge estructural en: Aprenentatge d'arbres, Aprenentatge poliarbres, Aprenentatge de xarxes multiconnectades.

Aprenentatge Paramètric: donada una estructura, obté les probabilitats associades. El requisit fonamental per dur a terme la tasca d'aprenentatge de xarxes bayesianes és disposar d'una base de dades en què aquest detallat el valor de cada variable en cada un dels casos.

En el cas de la nostra situació problemàtica considerada l'elecció d'aquesta tècnica de mineria de dades es fonamenta en que la base d'una xarxa ja construïda, i atesos els valors concrets d'algunes variables d'una instància, podrien tractar d'estimar els valors d'altres variables de la mateixa instància aplicant raonament probabilístic.

PREPARACIÓ DE LES DADES (Data preparation)

Punts de partida del nostre cas d'estudi

El nostre cas d'estudi pretén abordar els problemes relacionats amb algun tipus de predictors de domini o hipòtesis sobre el problema de la predicció de l'ingrés final hospitalari de l'usuari sobre la base de les seves atencions al servei d'urgències i sobre els seus antecedents. D'aquesta manera, el nostre enfocament ens ofereix l'establiment de predictors que ens permetran agrupar els usuaris pel seu comportament esperat, sempre que es demostri que heterogeneïtat dels grups d'usuaris.

Font de les dades

Les dades emprades en l'estudi s'han obtingut de les taules que recullen de forma anonimitzada i dissociada dels registres provinents de l'activitat assistencial diària que fan els professionals d'atenció primària.

Dades demogràfiques, visites urgent a primària i tractaments crònics

Per fer aquest estudi, ens hem centrat, únicament, en els resultats obtinguts a partir de la informació provinent de l'atenció primària i la del fitxer històric de problemes de salut, les visites i algunes dades demogràfiques notificats a l'HCAP d'atenció primària.

Aquesta informació s'emmagatzema en diferents servidors i s'anonimitza i selecciona en funció de les necessitats d'explotació de dades.

Per tal de garantir la confidencialitat de les dades es realitza un procés d'encriptació i dissociació d'aquelles variables que ens poden conduir a la identificació dels usuaris.

D'aquesta font de dades s'extraurà la informació corresponent a la Població assignada i atesa durant els darrers 12 mesos als centres d'atenció primària estudiats, agrupada segons les variables d'estudi.

Ingressos Hospitalaris

L'activitat hospitalària d'aquesta població, s'avaluarà segons la informació de la mateixa. D'aquesta s'ha exclòs l'activitat dels hospitals privats, els ingressos urgents mèdics amb estades de 0 a 1 dia per tal d'homogeneïtzar la informació que els diferents centres han notificat al registre. La informació en forma de nombre d'ingressos hospitalaris els darrers 12 mesos, s'ha calculat dels contactes categoritzats com hospitalització convencional i corresponents a pacients residents a Catalunya.

L'obtenció dels agrupadors de malalties de cada usuari també prové de les taules d'explotació comentades i prové d'una clusterificació prèvia segons l'algoritme de l'aplicatiu emprat.

Grups de risc

L'agrupador es fonamenta, sobretot, en les dades procedents de l'atenció primària, que és la principal font d'enregistrament de la patologia crònica poblacional, perquè és on es diagnostica o perquè és on es controla. Per definició, la patologia crònica, un cop diagnosticada, acompanya el pacient al llarg de la seva vida, i, en condiciona el tractament i els costos de l'atenció, independentment del motiu de la visita. La patologia crònica ha de quedar enregistrada en cada un dels contactes del pacient per a una correcta identificació de la morbiditat de la població. L'estudi de diferents algoritmes de preagrupació està motivat, d'una banda, per la certesa de manca d'exhaustivitat en la font de les dades provinents de l'atenció primària, i, de l'altra, per la mateixa estructura o el model de dades, la qual cosa condiciona la posterior extracció d'informació.

En aquest sentit, a la bibliografia sobre el funcionament i desenvolupament dels agrupadors, s'especifica que s'analitza la patologia crònica "en tractament actiu", i s'ha de suposar que un pacient que pateix una patologia crònica, si fa un contacte assistencial per qualsevol altre motiu, continua "en tractament actiu" dels seus problemes crònics. Aquest concepte de "patologia activa" (present) s'ha d'entendre des d'un perspectiva clínica, no s'ha de confondre amb "activa" informàticament (notificada al registre del contacte).

Per aquests motius, i com ja s'ha dit, es considera que, per a una correcta identificació de la morbiditat de la població, és fonamental recollir: la patologia que origina el contacte, i la patologia crònica que pateix el pacient, la qual posa de manifest tota la complexitat de la població atesa a l'atenció primària.

Pel que fa a la distribució per estats de salut de la població assignada, es pot dir que, aproximadament, la meitat de la població està sana, mentre que una tercera part pateix alguna malaltia crònica important, com ara, diabetis mellitus (DM), insuficiència cardíaca congestiva (ICC), malaltia pulmonar obstructiva crònica (MPOC), hipertensió arterial (HTA), asma, malaltia cerebrovascular, insuficiència renal crònica (IRC), etc.

Pel que fa a les diferents patologies cròniques analitzades diabetis, insuficiència cardíaca i hipertensió, s'observa que tots els pacients queden classificats associats a patologies cròniques en qüestió i, en particular, en grups clínicament coherents.

La identificació dels problemes crònics es duu a terme a partir del Chronic Condition Index de la Health Cost and Utilization Project (HCUP)de l'Agency for Health Research and Quality (AHRQ), que és una taula de codis en versió ICD-9 on es classifiquen els problemes com a crònics o no crònics.

Tècniques de mineria de dades que s'han d'aplicar

La mineria de dades és un procés que té com a objectiu descobrir i extreure informació rellevant de base de dades o d'altres fonts d'emmagatzematge de dades, facilitant la identificació de patrons, tendències com així també desvetllar fets anormals que poden estar succeint.

Aquest treball té com a objectiu l'aplicació d'una tècnica de mineria de dades dins del grup de tècniques de classificació com ho és les xarxes bayesianes que serà aplicada a una problemàtica pertanyent a l'àmbit de la salut, en una problemàtica amb molt d'interès com ho és el tema de la predicció amb un interès correctiu, dels factors involucrats en l'ingrés hospitalari. La problemàtica expressada és la preocupació per l'increment de les assistències urgents i els ingressos hospitalaris degut a una falta d'activitats preventives sobre els grups més utilitzadors. Per desvetllar quines poden ser les causes d'aquest comportament, com ja s'ha exposat aplicarem la tècnica de xarxes bayesianes i el classificador seleccionat és NaivesBayes usant el programari Weka i Elvira. L'eix de l'estudi i anàlisi es centra a determinar quins són els possibles factors que tenen una major incidència en els ingressos hospitalaris.

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

Eines i mètodes

Per a les tasques d'anàlisi prèvia i l'estadística descriptiva es va recórrer a les eines IBM SPSS Statistics™, versió 20.0, de IBM i el JMP® 8.0 de SAS.

L'extracció de dades i la fase prèvia de discretització es va fer a través de consultes sobre les taules d'oracle 9.1, amb les eines de Toad™ for Oracle Xpert, versió 9.6.0.27, una eina de l'empresa Quest Software. En local es crea un DWH amb MySQL Server 5.5 i MySQL , que es gestiona amb MySQL Workbench 5.2 CE.

Per tal de poder preprocessar les dades s'utilitza un driver de MySQL per Weka i finalment es fa el preprocessat d'atributs amb el programari de codi obert Weka, versió 3.6.8, una eina desenvolupada per la Universitat de Waikato que conjuga mineria de dades amb eines de visualització gràfica i ofereix, alhora, algunes eines de discretització i d'anàlisi de dades.

Com s'ha dit més amunt, es pretén emprar un algoritme basat en xarxa bayesiana, per tal de poder predir els ingressos hospitalaris a través de les dades d'atenció primària.

La proposta metodològica de treball consisteix, bàsicament, en l'ús d'una eina per a l'aprenentatge automàtic anomenada Weka [25] i del programari Elvira. Aquestes eines són de distribució lliure, desenvolupades en Java i permeten la implementació de tècniques de classificació, agrupament i associació, així com realitzar tasques de preprocessat i filtratge de dades.

En aquest treball s'empraran algoritmes de Classificadors, Naïve Bayes, Hill-Climber, K2, TAN i KDB. Es farà emprant els programes Elvira i Weka, per arribar a obtenir una xarxa bayesiana amb aquests classificadors. Aquesta xarxa variarà depenent de l'algorisme classificador aplicat, i de la combinació d'aquest amb algun algorisme d'inducció d'arbres de decisió. Finalment es mostrarà una comparació de resultats que permeti analitzar les diferències entre els diferents classificadors i la influència que hi genera els algoritmes generadors d'arbres de decisió en el supòsit plantejat del problema de la detecció d'ingressos hospitalaris.

El programa Elvira és fruit d'un projecte de recerca finançat per la CICYT i el Ministeri de Ciència i Tecnologia espanyol, en el qual participen investigadors de diverses universitats espanyoles i d'altres centres. Està destinat a l'edició i avaluació de models gràfics probabilistes, concretament xarxes bayesianes i diagrames d'influència. Elvira té un format propi per a la codificació dels models, un lector intèrpret per als models codificats, una interfície gràfica per a la construcció de xarxes, amb opcions específiques per models canònics (portes OR, AND, MAX, etc.), algoritmes exactes i aproximats (estocàstics) de raonament tant per variables discretes com contínues, mètodes d'explicació del raonament, algorismes de presa de decisions, aprenentatge de models a partir de bases de dades, fusió de xarxes, etc. Elvira està escrit i compilat en el llenguatge Java, la qual cosa permet que funcioni en diferents plataformes i sistemes operatius (MSDOS / Windows, Linux, Solaris, etc.).

D'altre banda Weka és una extensa col·lecció d'algorismes de Màquines de coneixement desenvolupats per la universitat de Waikato (Nova Zelanda) implementats en Java, útils per ser aplicats sobre dades mitjançant les interfícies que ofereix o per encabir-los dins de qualsevol aplicació. Weka conté les eines necessàries per realitzar transformacions sobre les dades, tasques de classificació, regressió, clustering, associació i visualització. Està dissenyat com una eina orientada a l'extensibilitat de manera que afegir noves funcionalitats és una tasca senzilla ". És un programari que ha estat desenvolupat sota llicència GPL4 la qual cosa ha impulsat que sigui una de les suites més utilitzades en l'àrea en els últims anys [23], així mateix si es pren a [24], és un programari per al aprenentatge automàtic o mineria de dades.

Per ser GPL la llicència, aquest programa és de lliure distribució i difusió, a més és independent de l'arquitectura, ja que funciona en qualsevol plataforma sobre la qual hi hagi una màquina virtual Java disponible [25]. A continuació es detallen les etapes dutes a terme per a l'anàlisi i estudi de la problemàtica exposada anteriorment.

Anàlisi preliminar i preparació de les dades

La Preparació dels inputs de dades per a una investigació de mineria de dades consumeix la major part de l'esforç invertit en el procés de mineria de dades. L'amarga experiència demostra que les dades reals no acostumen a tenir la qualitat suficient i requereixen d'un acurat procés previ de "prepocessament"- que es coneix en el mon anglosaxó com data-cleaning.

En començar a treballar en un projecte de mineria de dades, primer cal transformar totes les dades en un conjunt homogeni d'instàncies. D'altre banda, la integració de dades de diferents fonts sol presentar problemes. Caldrà fer servir diferents estils de manteniment de registres, diferents convencions diferents, diferents períodes de temps, diferents graus d'agregació de dades, diferents claus principals, i la detecció de diferents tipus d'error. Caldrà muntar les dades, integrar-les i netejar els seus atributs. La idea de la integració de bases de dades es coneix com a data-warehousing (DWH). Aquests magatzems

de dades proporcionen un únic punt d'accés consistent a les dades i són el lloc on abocarem les dades des de la seva procedència original per tal d'analitzar-les.

El fet que hi hagin qüestions tan diferents involucrades, pot generar un procés lent de creació amb diferents aproximacions a l'objectiu final. Aquesta és la raó per la qual la recopilació de dades, la integració, la neteja, agregació i preparació de les dades acostumen a trigar tant. En treballar amb dades del món real (no aquells que són generats de forma sintètica) ens trobem amb una sèrie de problemes en usar xarxes bayesianes, problemes que haurem d'afrontar abans d'intentar emprar aquestes xarxes a partir de les dades. A continuació es detallen les etapes dutes a terme per a l'anàlisi i estudi de la problemàtica exposada anteriorment:

Anàlisi de les dades (data set)

Selecció de les variables significatives

En la gran majoria d'estudis en els que s'apliquen tècniques de mineria de dades, les característiques o atributs disponibles no són sempre realment rellevants per obtenir un model de coneixement. El nostre cas no és l'excepció; el nostre problema s'enfoca en saber quina és la probabilitat que un usuari acabi ingressant en un hospital donades certes condicions (variables) i també hem d'escollir aquelles variables que ens poden aportar informació d'utilitat per obtenir un bon model. Per aconseguir-ho ens ajudarem, bàsicament, de les eines que ens ofereix el programari Weka usant els filtres establerts en la secció de Preprocessament.

Per a la realització d'aquest model probabilístic considerarem les dades i antecedents de les persones que empren els serveis d'atenció primària d'un territori, per tal de determinar el nivell de risc que té aquesta mateixa persona d'acabar ingressant en un hospital en els propers 12 mesos. Per aquest model es disposa d'una base de coneixement amb un domini que té sis atributs de classe.

Atribut	Descripció
INGRÉS	Aquest atribut correspon a la classe i ens permetrà conèixer els atributs que incideixen en la probabilitat d' un ingrés hospitalari en els darrers 12 mesos. El tipus de dades és booleà.
SEXE	Representa el sexe de la persona. El tipus de dades és categòrica.
GRUP EDAT	Indica en quin grup d'edat es troba l'usuari. Aquest tipus de dada és nominal, categòrica.
ESTAT	Indica l'estat que s'ha assignat al pacient segons les seves malalties i variant segons els CRG (case-mix) final obtingut. El tipus de dada és nominal.
URGÈNCIES	Indica si l'usuari ha acudit a un servei d'urgències de primària abans o durant el període d'estudi. Aquest tipus de dada és booleana.
MES DE 10 TTS	Indica el nombre de medicaments diferents que pren de forma crònica l'usuari. Aquest tipus de dada és nominal.

Taula 1. **Variables considerades per a l'aplicació dels algorismes.**

Es recull un total inicial de 512.614 usuaris diferents que han estat atesos, al menys en una ocasió, en alguna agenda d'atenció primària del territori de la comarca d'Osona durant els darrers 12 mesos. D'aquest s'acaben escollint 464.800 casos finals.

Selecció i neteja de dades

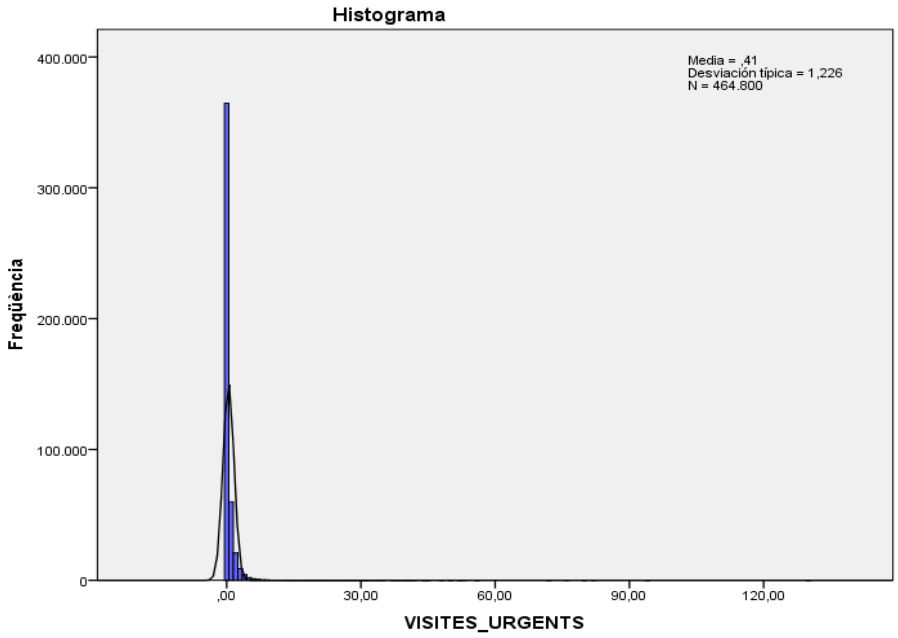
Les dades originals es presenten en les següents taules segons el tipus de variable que veiem a continuació:

Variables numèriques:

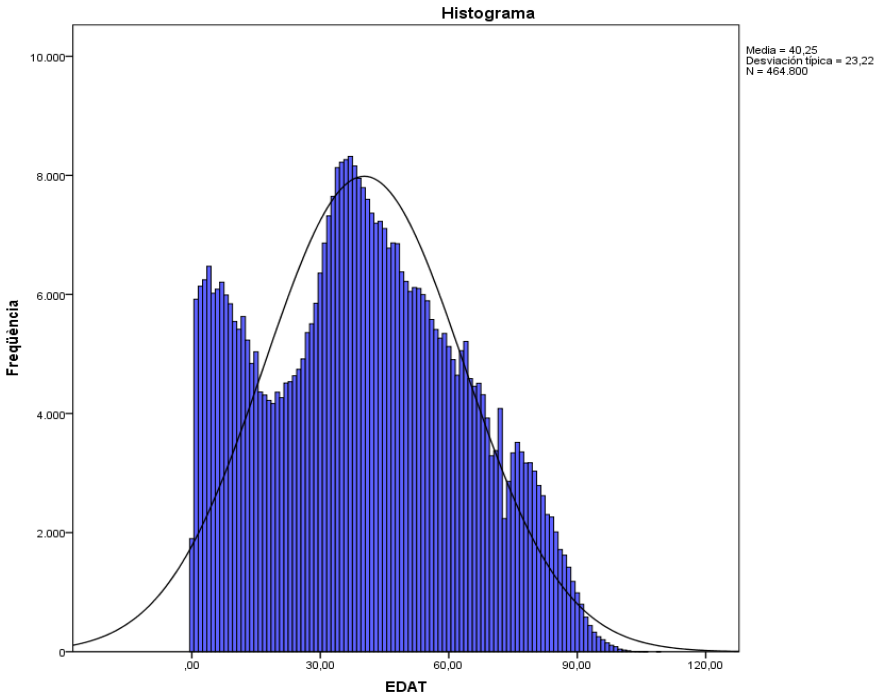
Estadístics descriptius

	N	Mínim	Màxim	Mitjana	Desv. típ.
EDAT	464800	,00	109,00	40,2455	23,22017
VISITES_URGENTS	464800	,00	130,00	,4102	1,22605
VISITES	464800	1,00	466,00	13,1681	13,60843
TRACTAMENTS_CRONICS	464800	,00	28,00	1,4000	2,62349
N vàlid (segons llista)	464800				

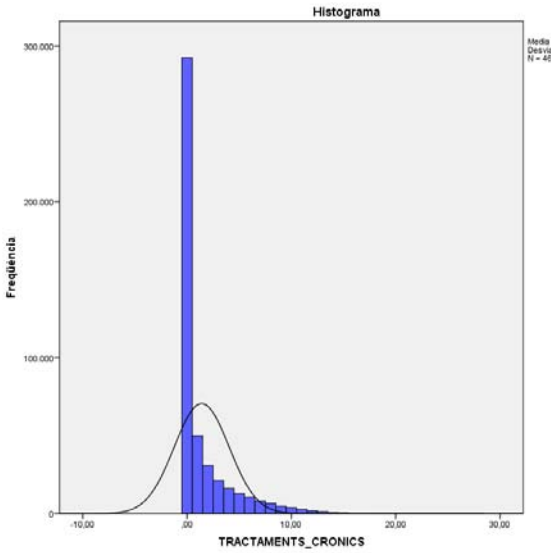
La distribució del nombre de visites urgents demanades per els usuaris s'agrupa entorn a 0 o 1 visita, amb una moda de 0 i una mitjana de 0,4 visites. La majoria d'usuaris no han acudit mai a urgències, tot i que existeixen outliers que caldrà valorar si cal depurar.

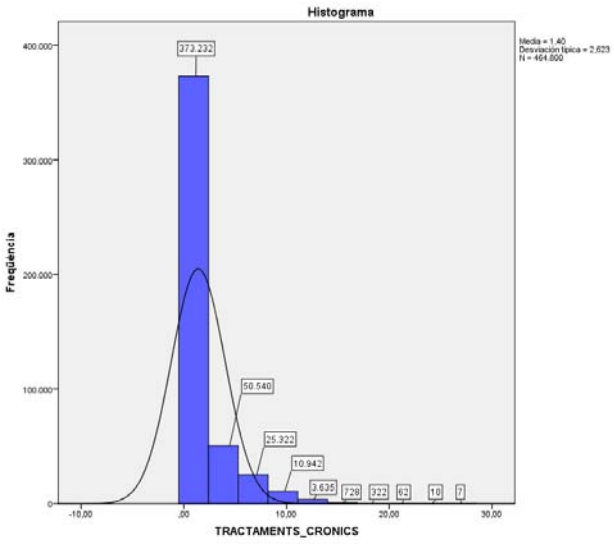


Les edats mostren dos pics màxims en les primeres edats de la vida i durant la mitjana edat, entre els 40 i 50 anys (major nombre de població). Destaca el major consum de visites efectuat per la població pediàtrica (corba bimodal).



En el cas dels tractaments crònics que segueixen els pacients del domini, veiem que predominantment no estan fent cap tractament crònic, però la mitjana es situa en més d'un tractament crònic

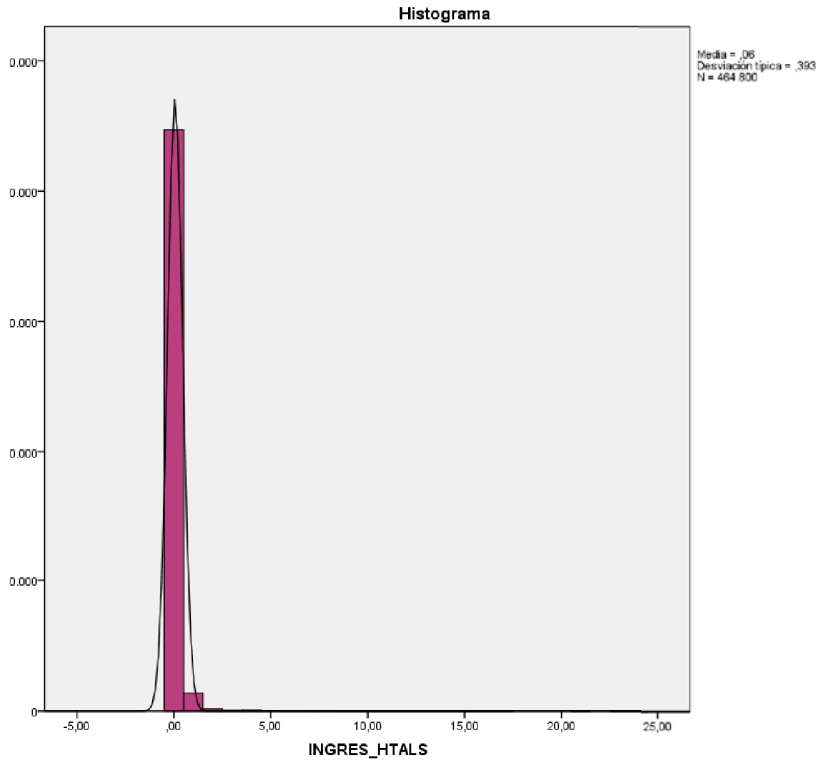




El nombre d'ingressos hospitalaris no mostra els valors nuls, que es correspondrien als pacients que no han fet cap ingrés hospitalari i que serien la majoria 446.930 (96% de la mostra).

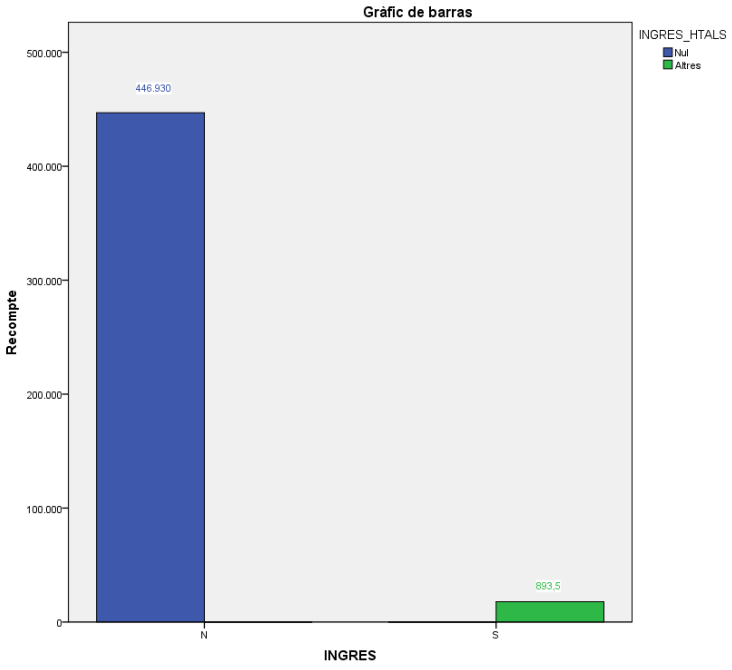
INGRES_HTALS

		Frecuencia	Porcentaje	Porcentaje válido	Porcentaje acumulado
Válidos	,00	446930	96,2	96,2	96,2
	1,00	13868	3,0	3,0	99,1
	2,00	1702	,4	,4	99,5
	3,00	742	,2	,2	99,7
	4,00	889	,2	,2	99,9
	5,00	259	,1	,1	99,9
	6,00	186	,0	,0	100,0
	7,00	77	,0	,0	100,0
	8,00	65	,0	,0	100,0
	9,00	31	,0	,0	100,0
	10,00	15	,0	,0	100,0
	11,00	12	,0	,0	100,0
	12,00	7	,0	,0	100,0
	13,00	7	,0	,0	100,0
	14,00	3	,0	,0	100,0
	15,00	2	,0	,0	100,0
	16,00	1	,0	,0	100,0
	17,00	1	,0	,0	100,0
	21,00	1	,0	,0	100,0
	23,00	1	,0	,0	100,0
	24,00	1	,0	,0	100,0
	Total	464800	100,0	100,0	



Entre els que han efectuat un ingrés hospitalari predominen els que han fet un sol ingrés.

	INGRES				URGENTS			
	N		S		N		S	
SEXE	N	% of Total	N	% of Total	N	% of Total	N	% of Total
D	230509	49,59%	10490	2,26%	188835	40,63%	52164	11,22%
H	216421	46,56%	7380	1,59%	175781	37,82%	48020	10,33%
GRUP_EDAT								
0-10	60823	13,09%	1549	0,33%	44156	9,50%	18216	3,92%
101-110	49	0,01%	9	0,00%	41	0,01%	17	0,00%
11-20	46992	10,11%	588	0,13%	37150	7,99%	10430	2,24%
21-30	48677	10,47%	2011	0,43%	39185	8,43%	11503	2,47%
31-40	75358	16,21%	3313	0,71%	62705	13,49%	15966	3,44%
41-50	68245	14,68%	1345	0,29%	56246	12,10%	13344	2,87%
51-60	55435	11,93%	1454	0,31%	46340	9,97%	10549	2,27%
61-70	42977	9,25%	1922	0,41%	36660	7,89%	8239	1,77%
71-80	29611	6,37%	2535	0,55%	25392	5,46%	6754	1,45%
81-90	16306	3,51%	2613	0,56%	14485	3,12%	4434	0,95%
91-100	2457	0,53%	531	0,11%	2256	0,49%	732	0,16%
ESTAT								
Estat1	196384	42,25%	4671	1,00%	159342	34,28%	41713	8,97%
Estat2	2412	0,52%	475	0,10%	2093	0,45%	794	0,17%
Estat3	58749	12,64%	1315	0,28%	47382	10,19%	12682	2,73%
Estat4	19996	4,30%	480	0,10%	15953	3,43%	4523	0,97%
Estat5	98344	21,16%	3485	0,75%	80060	17,22%	21769	4,68%
Estat6	64710	13,92%	5656	1,22%	53731	11,56%	16635	3,58%
Estat7	3902	0,84%	1176	0,25%	3786	0,81%	1292	0,28%
Estat8	1698	0,37%	470	0,10%	1607	0,35%	561	0,12%
Estat9	735	0,16%	142	0,03%	662	0,14%	215	0,05%
MES10TT								
0	284398	61,19%	7733	1,66%	228577	49,18%	63554	13,67%
1-10	156874	33,75%	8457	1,82%	130789	28,14%	34542	7,43%
mes de 10	5658	1,22%	1680	0,36%	5250	1,13%	2088	0,45%



Variables nominals:

Resumen del procesamiento de los casos

	Casos					
	Incluidos		Excluidos		Total	
	N	Porcentaje	N	Porcentaje	N	Porcentaje
SEXE	464800	100,0%	0	0,0%	464800	100,0%
ESTAT	464800	100,0%	0	0,0%	464800	100,0%

Resúmenes de casos

	SEXE	ESTAT
N	464800	464800
Mínimo	D	Estat1
Máximo	H	Estat9
% del total de N	100,0%	100,0%

De la tria inicial d'atributs es va descartar l'ús de la nacionalitat, que al inici de l'estudi es va considerar. Teníem un gran ventall d'orígens, amb un total de 55 països diferents. Tot i que la majoria són usuaris nacionals (154.284 espanyols), destaquen per ordre de freqüència els 17.833 marroquins, 5.300 equatorians, 4.198 bolivians, 3.076 senegalesos, 2.583 romanesos, 2.079 colombians, 1.433 xinesos.

NACIONALITAT

		Frecuencia	Porcentaje	Porcentaje válido	Porcentaje acumulado
Válidos		301129	58,7	58,7	58,7
	AFGANISTAN	31	,0	,0	58,7
	ALBÀNIA	6	,0	,0	58,8
	ALEMANYA	260	,1	,1	58,8
	ALGÈRIA	148	,0	,0	58,8
	BRETANYA)				
	VIETNAM	4	,0	,0	99,6
	XILE	738	,1	,1	99,7
	XINA	1433	,3	,3	100,0
	Total	512614	100,0	100,0	

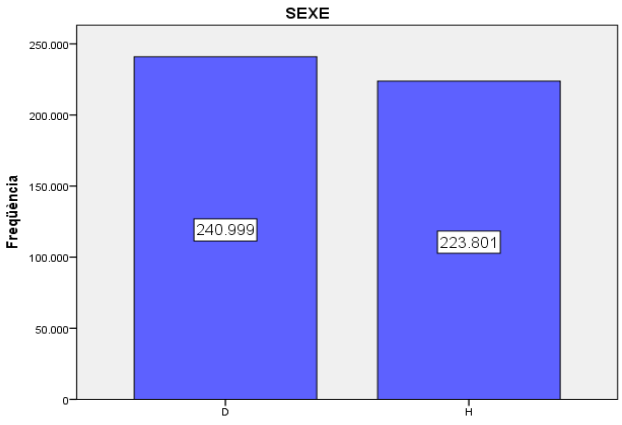
Malauradament el percentatge de vàlids és molt baix (58%) i a més s'associa a una gran dispersió de les dades. Ambdós fets fan que no sigui una bona variable candidata per a l'estudi.

Sexe

Entre les variables categòriques podem veure la distribució de sexes, amb una distribució equilibrada:

SEXE

		Frecuencia	Porcentaje	Porcentaje válido	Porcentaje acumulado
Válidos	D	240999	51,9	51,9	51,9
	H	223801	48,1	48,1	100,0
	Total	464800	100,0	100,0	

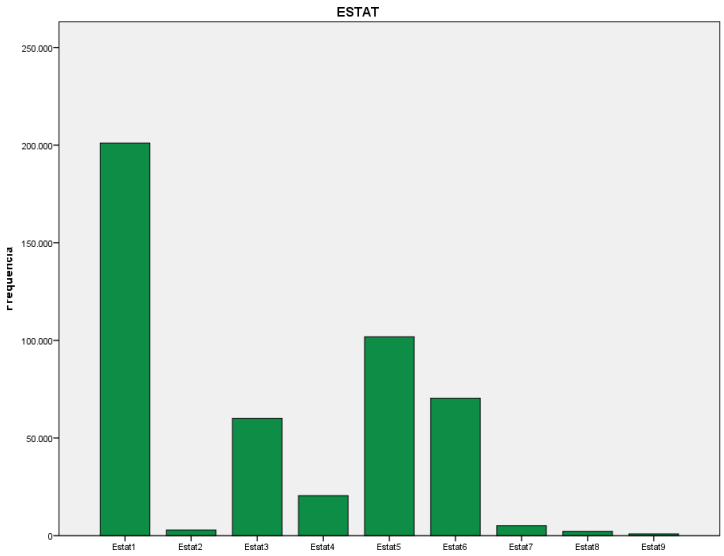


Estats

També destaquem la distribució dels Estatus, que es corresponen amb els grups de risc clínic als que pertanyen els usuaris [8] , amb la següent distribució:

ESTAT

		Frecuencia	Porcentaje	Porcentaje válido	Porcentaje acumulado
Válidos	Estat1	201055	43,3	43,3	43,3
	Estat2	2887	,6	,6	43,9
	Estat3	60064	12,9	12,9	56,8
	Estat4	20476	4,4	4,4	61,2
	Estat5	101829	21,9	21,9	83,1
	Estat6	70366	15,1	15,1	98,3
	Estat7	5078	1,1	1,1	99,3
	Estat8	2168	,5	,5	99,8
	Estat9	877	,2	,2	100,0
	Total	464800	100,0	100,0	



Les dades no mostren els 51.146 individus que no tenen l'estat assignat o nul (es van depurar a posteriori). Entre el total destaquen els usuaris en Estat 1 (sans o amb patologies agudes, no cròniques) i els usuaris de l'estat 5 que són aquells que tenen una patologia crònica severa o moderada.

Transformació de les dades en el format adequat

El gran volum de dades generada amb la consulta SQL (veure annexa amb l'script) a la base de dades oracle va impossibilitar la generació d'un sol fitxer .csv per poder treballar amb weka.

Això va fer que s'utilitzés un driver jdbc per mysql i així per poder treballar directament contra la base de dades (DWH) generada en MySQL. Per comoditat final es va realitzar una exportació de les dades a un format adequat d'arxiu per treballar amb weka (del tipus arff) i el treball amb Elvira es va fer important des de .csv. El format de .csv o arff és mol similar, tenint aquest darrer algunes característiques pròpies, tal com s'il·lustra:

(A) @relation ingressosHtals

(B) @attribute VISITA_URGENT numeric
@attribute VISITES_URGENTS_2012 numeric
@attribute SEXE {H,D}
@attribute EDAT numeric
@attribute MES10TT {0,1-10,mes10}
@attribute ESTAT {Estat1,Estat2,Estat3,Estat4,Estat5,Estat6,Estat7,Estat8,Estat9,DESC}
@attribute NUM_INGRESSOS_HTALS {0,1,2,3,4,5,6,7,8,9,10,11}

(C) @data
1.00,0.00,H,37.00,0,Estat6,0
0.00,0.00,D,7.00,0,Estat1,0
2.00,0.00,H,15.00,0,Estat1,0
1.00,0.00,D,15.00,0,Estat1,0
.....

(A): En aquesta secció es defineix el nom de la relació. (B): S'especifiquen els atributs de la relació i el tipus de dada. (C): És la secció de dades pròpiament dita.

Tractament dels valors absents en alguns atributs

És important, un cop importades les dades a l'eina, realitzar una anàlisi de la quantitat de valors perduts (missing) en les variables considerades, ja que és un factor important que pot arribar a influir en el model de predicció a obtenir.

Weka ens ofereix una eina per al tractament de valors perduts o absents, que consisteix a eliminar totes les instàncies amb valors nuls de les nostres dades. Weka ens permet aplicar una gran diversitat de filtres sobre les dades, per tal de poder realitzar tot tipus de transformacions sobre aquestes. En aquest cas hem seleccionat un filtre que es troba en la categoria d'atributs", la denominació és "ReplaceMissingValues", i ens permet reemplaçar tots els valors indefinits en el cas que sigui l'atribut nominal, com el nostre cas.

D'aquest conjunt de dades, eliminarem els 52.212 casos en els que mancava el valor de l'estatus al que pertany l'usuari, quedant el valors definitius.

ESTAT

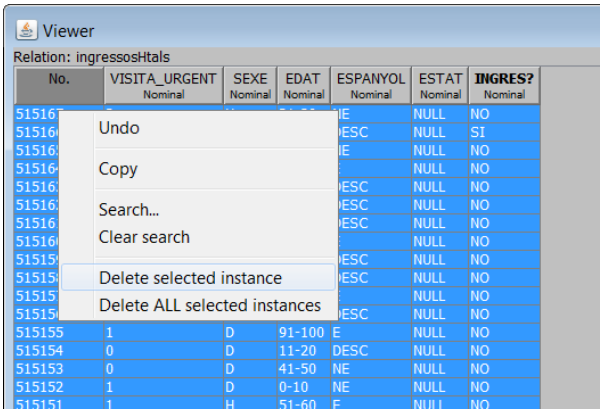
		Frecuencia	Porcentaje	Porcentaje válido	Porcentaje acumulado
Válidos	Estat1	201055	43,3	43,3	43,3
	Estat2	2887	,6	,6	43,9
	Estat3	60064	12,9	12,9	56,8
	Estat4	20476	4,4	4,4	61,2
	Estat5	101829	21,9	21,9	83,1
	Estat6	70366	15,1	15,1	98,3
	Estat7	5078	1,1	1,1	99,3
	Estat8	2168	,5	,5	99,8
	Estat9	877	,2	,2	100,0
	Total	464800	100,0	100,0	

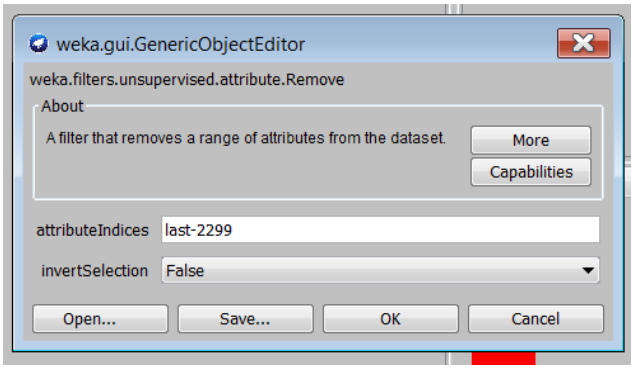
També es va plantejar la possibilitat d'eliminar els casos en que mancava la nacionalitat de l'usuari, però donada la gran dispersió de les dades i el baix percentatge de valors vàlids es va decidir descartar l'atribut. En total, van quedar 464.800 casos per efectuar la primera anàlisi.

L'atribut sexe tenia 16 registres sense valor que també es va plantejar eliminar, però finalment es va buscar aquest atribut a través de l'explotació mes fiable, recuperant aquesta falta de dades i comprovant que hi havia coincidència en la resta de casos.

Finalment, es va eliminar el registre que mostrava un outlier amb mes de 600 visites a urgències, que també es va eliminar, en tractar-se d'un usuari fictici de proves amb un valor fals de visites.

Per tal de reduir el conjunt d'atributs seleccionats hem fet servir un Filtre de Weka que es troba en la categoria de Atributs denominat "Remove", que permet esborrar un conjunt d'atributs especificats a l'arxiu de dades, amb el que s'han eliminat alguns atributs que no aportaven cap informació al projecte (origen de la visita, diagnòstic d'alta hospitalària, etc.) i dels que no mostrarem cap dada, però que es poden reconèixer en la select per extreure les dades de les taules originals





Discretització de variables contínues.

Un altre dels problemes que es presenta quan treballem amb xarxes bayesianes sobre dades reals, és que en moltes ocasions apareixen variables discretes (que tenen un nombre finit de possibles valors, com per exemple el sexe) i variables contínues (per exemple el pes en grams).

Les xarxes bayesianes només treballen amb variables discretes pel que necessitem fer alguna cosa amb les variables que són contínues. Això és degut a que la major part dels algoritmes de propagació no d'aprenentatge estan perfectament definits per variables discretes, però no per a variables contínues. Una solució a aquest problema passa per utilitzar variables de tipus continu usant distribucions gaussianes o exponencials de mesures truncades.

Una altra solució passa per discretitzar les variables contínues i tractar-les com si fossin discretes, és a dir, transformar els valors d'una variable contínua en un conjunt de valors discrets. Un algorisme de discretització divideix el domini de la variable contínua en un nombre finit d'intervals (que seran fet i fet els diferents estats discrets) i, posteriorment, a cada observació de la variable contínua se li assigna l'interval que inclou a aquesta observació. Per exemple, el mètode tot més senzill seria dividir el domini de la variable contínua en intervals d'igual mida (tants intervals com estats discrets es desitgin obtenir) i, finalment, assignar els valors con a integrants d'aquests intervals. La discretització, si bé és el mètode

més utilitzat en la literatura per al tractament de variables contínues té el problema que si el nombre d'intervals és massa petit, es perd precisions i si és massa gran, requereix una gran quantitat de dades per estimar les seves probabilitats. D'aquesta manera, podem considerar el problema de trobar el nombre d'intervals en un problema de recerca.

En el cas de classificació supervisada la tècnica de referència és el mètode de discretització per mínima entropia, presentat per Fayyad i Irani. Aquest mètode selecciona recursivament els punts de tall mitjançant un algorisme de minimització de la entropia entre cada atribut i la classe. Utilitza el principi de mínima longitud de descripció (MDN).

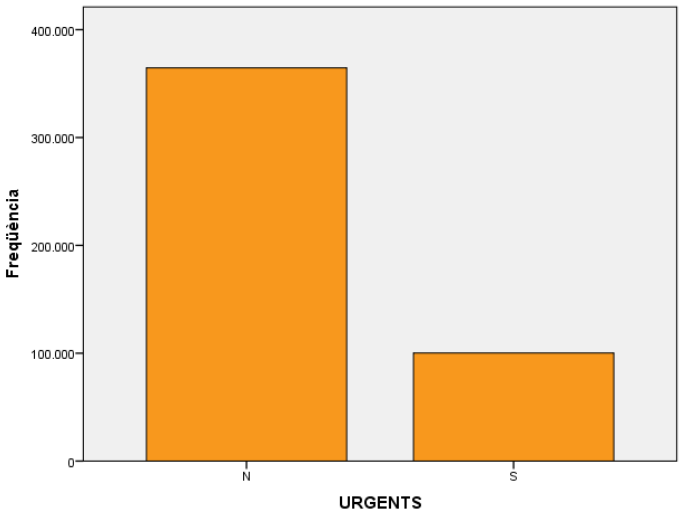
En el nostre domini hem discretitzat els atributs:

1. Visita_Urgent = {SI, NO}
2. Edat = {0-10, 11-20, 21-30, 31-40, 41-50, 51-60, 61-70, 81-90, 91-100, 101-110, altres}
3. MES10TT = {0, 1-10, mes10}
4. Estat = {Estat1, Estat2, Estat3, Estat4, Estat5, Estat6, Estat7, Estat8, Estat10}
5. Ingrés? = {SI, NO}

Visites Urgents

URGENTS

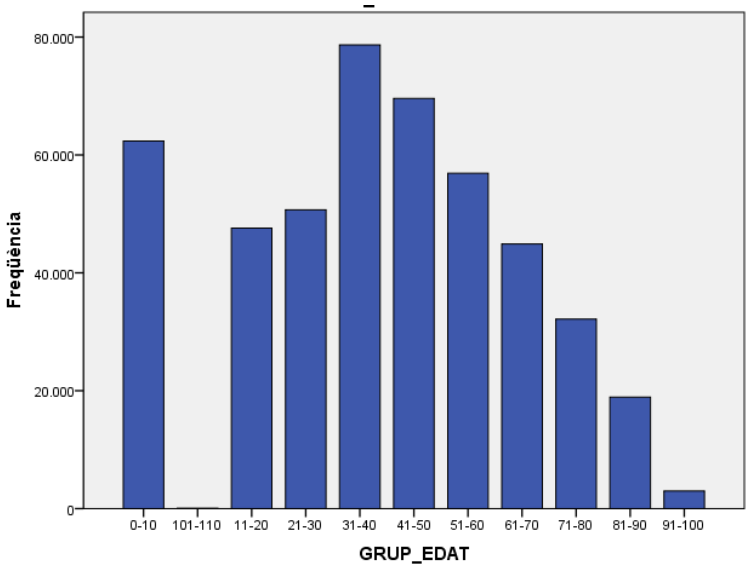
		Frecuencia	Porcentaje	Porcentaje válido	Porcentaje acumulado
Válidos	N	364616	78,4	78,4	78,4
	S	100184	21,6	21,6	100,0
	Total	464800	100,0	100,0	



Grups d'edat

GRUP_EDAT

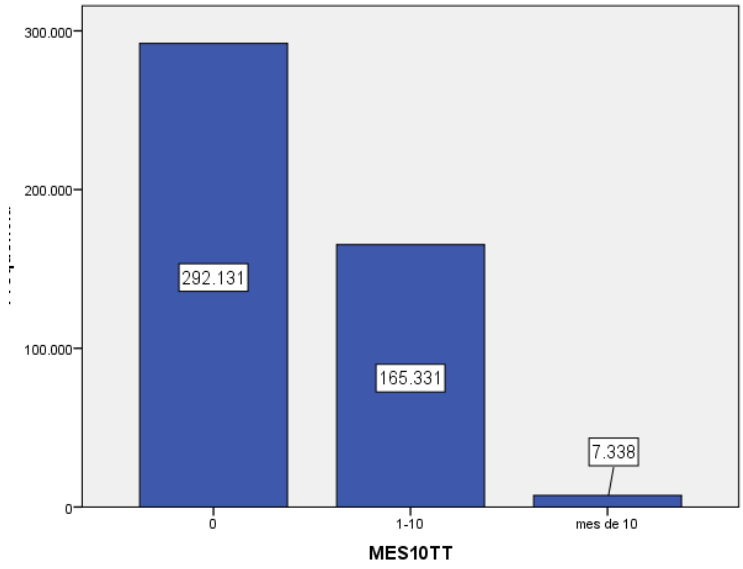
		Frecuencia	Porcentaje	Porcentaje válido	Porcentaje acumulado
Válidos	0-10	62372	13,4	13,4	13,4
	101-110	58	,0	,0	13,4
	11-20	47580	10,2	10,2	23,7
	21-30	50688	10,9	10,9	34,6
	31-40	78671	16,9	16,9	51,5
	41-50	69590	15,0	15,0	66,5
	51-60	56889	12,2	12,2	78,7
	61-70	44899	9,7	9,7	88,4
	71-80	32146	6,9	6,9	95,3
	81-90	18919	4,1	4,1	99,4
	91-100	2988	,6	,6	100,0
	Total	464800	100,0	100,0	



Mes de 10 Tractaments crònics

MES10TT

		Frecuencia	Porcentaje	Porcentaje válido	Porcentaje acumulado
Válidos	0	292131	62,9	62,9	62,9
	1-10	165331	35,6	35,6	98,4
	mes de 10	7338	1,6	1,6	100,0
	Total	464800	100,0	100,0	



DISSENY DEL MODEL (Modelling)

Selecció de l'algorisme a utilitzar

Les tasques que es duen a terme amb la mineria de dades es poden classificar com predictives i descriptives. En les tasques predictives cada element de la base de dades es caracteritza per tenir un paràmetre d'entrada i un paràmetre de sortida. L'objectiu és intentar predir el valor del paràmetre de sortida utilitzant la informació proporcionada pels paràmetres d'entrada (Individu -> Ingrés).

Dins de les tasques predictives existeixen també dos tipus: La classificació, on cada element de la base de dades té associat un valor discret (classe), amb l'objectiu de maximitzar el poder de predicció de la classificació de nous elements per als quals la classe és desconeguda; i la regressió, en que el valor associat a cada element és un nombre real i que no és el cas que ens ocupa. També distingim entre Classificadors supervisats i no supervisats; els primers (el nostre cas) requereixen d'una intervenció d'un expert coneixedor de la matèria que supervisi la determinació de les classes, l'elecció de les característiques discriminants, la selecció de la mostra, càlcul de funcions discriminants i test del classificador, mentre que en els no supervisats el procés serà automàtic.

Donades les característiques del projecte, proposem l'ús d'un algoritme d'aprenentatge supervisat de grafs de decisió probabilística orientada a la classificació (PDG). Donat que les xarxes bayesianes són un dels paradigmes més àmpliament usats per classificació supervisada, la nostre elecció s'ha decantat cap a l'ús d'aquest tipus d'algoritmes.

Una xarxa bayesiana és un graf acíclic dirigit en què cada node representa una variable i cada arc una dependència probabilística, són utilitzades per proveir: una forma compacta de representar el coneixement, i mètodes flexibles de raonament.

Depenent de les restriccions que li hem posat a la xarxa emprada obtenim diferents classificadors i seran aquestes restriccions les que ens determinaran l'ús dels classificadors escollits. En aquest treball cercarem l'algoritme òptim d'entre quatre algoritmes classificadors: Naïve Bayes, K2, TAN o Hill-Climber.

Amb aquests algoritmes es pretén trobar el classificador òptim que ens permeti maximitzar la possibilitat de que en una nova instància, que en el nostre cas seria un nou individu amb uns atributs de l'estudi determinats, es classifiqui correctament entre els individus ingressadors o no ingressadors d'un hospital. El mètode, d'entre els escollits, que obtingui les millors tasses d'error i veritat (Correctament classificats / incorrectament classificats) emprant el mateix espai d'hipòtesi i el mateix coneixement de partida a priori, serà el que ens marcarà el model predictiu d'ús òptim. Bàsicament per això emprarem el programa Weka (Llicència Pública - GPL).

Analitzarem les relacions entre nodes i alguns trets concrets del modelatge amb l'ajuda d'un altre programa GPL, l'Elvira. Finalment es mostrarà una comparació que permeti analitzar les diferències observades amb els diferents classificadors emprats.

El procés d'avaluació amb les xarxes bayesianes obtingudes a partir de les dades mèdiques emprades en l'estudi, es durà a terme a través de l'execució dels anteriors algorismes de classificació, la seva precisió

i rapidesa de processament ha estat clau en la tria d'aquest, essent els més utilitzats en aquest tipus de problemes.

Els algoritmes són els següents:

Naive Bayes

És un dels algorismes de classificació més efectiu i senzill. Aquest es construeix sota la suposició que totes les variables predictores són condicionalment independents donat el valor de la variable classe. Les seves principals qualitats són la simplicitat i precisió, i encara que la seva estructura sempre és fixa (la variable classe apuntant a cada node) ens mostra una gran precisió de classificació i un error mínim. En donar-se un nou cas, fa servir el teorema de Bayes per calcular la probabilitat condicional de cada node seleccionant del valor de la classe amb la probabilitat major obtinguda.[26]

Els altres classificadors Bayesians escollits relaxen la restricció d'independència condicionalment posada en el naive Bayes per així modelar relacions més complexes entre variables.

K2

Es tracta d'un algorisme de classificació que s'inicia amb la xarxa més simple possible, és a dir, una xarxa sense arcs, i suposa que els nodes estan ordenats. Per a cada variable del problema, l'algorisme afegeix al seu conjunt de pares el node amb menor probabilitat que condueix a un màxim increment de la qualitat, corresponent a la mesura de qualitat escollida en el procés de classificació. Aquest procés es repeteix fins que, o bé no s'incrementa la qualitat o s'ha arribat a una xarxa completa. [27]

TAN

És un algorisme conegut també com Xarxa Bayesiana Augmentada a Arbre, on es permet que les variables predictores formin entre elles una estructura d'arbre. Consisteix a construir un arbre de dependències entre les variables que s'han de predir i que al seu torn són filles de la variable classe. La probabilitat d'aquestes variables es calcula aplicant el teorema de Bayes en base a la probabilitat de la variable classe. [28]

Hill-Climber

Algorisme de classificació que s'inicia amb una xarxa generada de manera aleatòria. Per a cada node o variable, l'algoritme afegeix o esborra relacions de manera aleatòria, calculant així, a partir de la probabilitat conjunta de la variable classe, la probabilitat de cada node que forma la xarxa. Finalment l'algorisme tria la xarxa òptima, és a dir aquella que té la millor qualitat, eliminant aquelles que no assoleixen el seu nivell. [29]

KDB

Algoritme anomenat k Dependence Bayesian classifier (KDB) que va presentar per primer cop Sahami (1996), que possibilita travessar l'ampli espectre de dependències disponibles entre el model naïve Bayes i el model corresponent a una xarxa bayesiana completa. L'algorisme es fonamenta en el concepte de classificador bayesià K-dependent, el qual conté l'estructura del classificador naïve Bayes i permet a cada variable predictora tenir un màxim de K variables pars sense comptar la variable classe.

D'aquesta manera, el model naïve Bayes es correspon amb un classificador bayesià 0-dependent, el model TAN seria un classificador Bayesià 1-dependent i el classificador bayesià complet (en l'estructura no es reflecteix cap independència) correspondria a un classificador bayesià (n - 1) – dependent.

Així dons, l'elecció d'aquests classificadors busca poder variar la complexitat, a través de les combinacions i nombre de dependències de les variables predictores escollides per tal de trobar un millor model predictiu. Podrem contemplar, segons el classificador escollit, diferents nombres de parells de variables classe i variables predictores.

Per poder utilitzar la família dels classificadors bayesians, hem d'accedir des del programa weka a través de la ruta weka.classifiers.bayes.NaiveBayes.

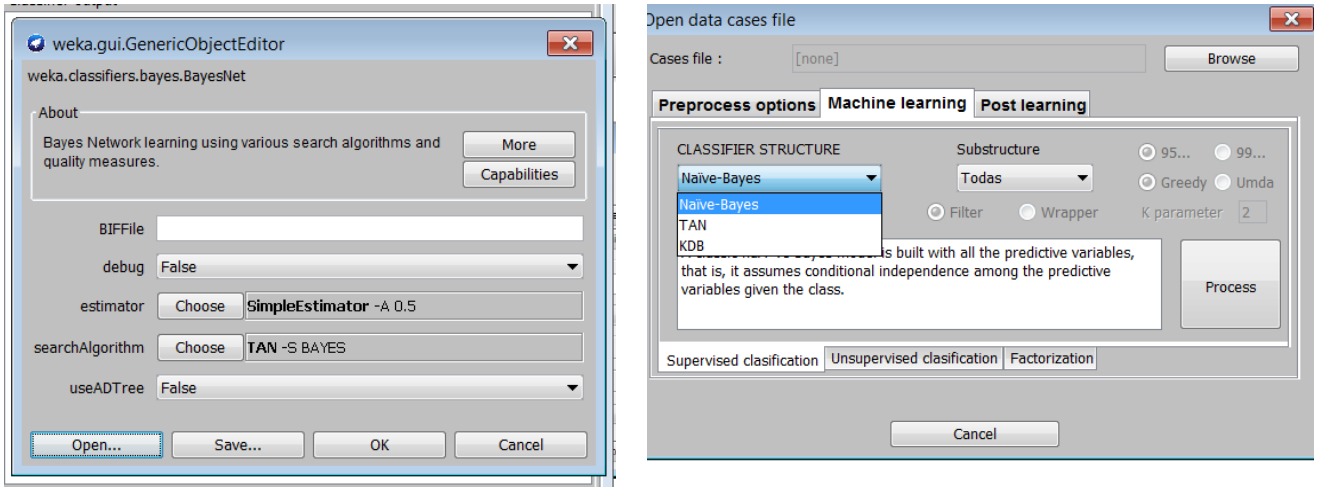


Figura 3. Paràmetres a configurar per a l'aplicació de l'algoritme en Weka i Elvira, respectivament.

Execució de classificadors

Els classificadors són algorismes adreçats a interpretar les dades provinents de l'arxiu d'entrenament per generar la xarxa Bayesiana. Com ja s'ha comentat només emprarem els suportats pel programa Weka i Elvira, que han estat descrits al punt anterior.

- **Obtenció d'una xarxa mitjançant el classificador Naïve Bayes**

El classificador a emprar en primera instància seria el més senzill, que cas de mostrar-nos els valors predictius desitjats, seria el més efectiu per al nostre cas d'estudi i suposaria que els atributs són condicionalment independents segons la classe.

Abans d'executar l'algorisme en Weka s'ha seleccionat una de les opcions de test: Cross-validation: això permet realitzar l'avaluació mitjançant la tècnica de validació creuada, permetent establir el nombre de mostres a utilitzar (per defecte 10). Després especifiquem del conjunt d'atributs seleccionats el que es considera classe principal, que serà Ingrés {si / no}. Quan s'executa l'algorisme es visualitza a la finestra de sortida amb la següent informació:

L'anàlisi dels resultats obtinguts, com es pot observar a continuació, mostra el nombre i percentatge de les instàncies classificades correcta i incorrectament.

```
=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances      443029               95.316  %
Incorrectly Classified Instances     21771                4.684  %
```

Podem observar en detall de cada classe i la matriu de confusió, com es mostra a continuació

```
=== Detailed Accuracy By Class ===

               TP Rate  FP Rate  Precision  Recall  F-Measure  MCC      ROC Area  PRC Area  Class
               0.987    0.895    0.965      0.987   0.976      0.139    0.722     0.982     N
               0.105    0.013    0.245      0.105   0.147      0.139    0.722     0.125     S
Weighted Avg.  0.953    0.861    0.937      0.953   0.944      0.139    0.722     0.949

=== Confusion Matrix ===

     a      b   <-- classified as
 441152   5778 |      a = N
  15993   1877 |      b = S
```

Figura 4. Finestra del classificador de sortida NB.

Weka també permet canviar les opcions per mostrar detalladament la predicció de les dades ingressades a la base del coneixement.

```
=== Predictions on test data ===

 inst#     actual  predicted error prediction
     1        1:N        1:N       0.988
     2        1:N        1:N       0.978          46465        2:S        1:N   +    0.97
     3        1:N        1:N       0.972          46466        2:S        1:N   +   0.983
     4        1:N        1:N       0.965          46467        2:S        1:N   +   0.946
     5        1:N        1:N       0.991          46468        2:S        1:N   +   0.992
     6        1:N        1:N       0.984          46469        2:S        1:N   +   0.618
     7        1:N        1:N       0.983          46470        2:S        2:S         0.612
     8        1:N        1:N       0.989          46471        2:S        1:N   +    0.97
     9        1:N        1:N       0.972          46472        2:S        1:N   +    0.96
    10        1:N        1:N       0.984          46473        2:S        1:N   +   0.614
    11        1:N        1:N       0.988          46474        2:S        1:N   +    0.91
    12        1:N        2:S   +   0.528          46475        2:S        1:N   +   0.946
    13        1:N        1:N       0.983          46476        2:S        1:N   +   0.974
    14        1:N        1:N       0.99           46477        2:S        1:N   +   0.946
    15        1:N        1:N       0.993          46478        2:S        1:N   +   0.984
    16        1:N        1:N       0.977          46479        2:S        1:N   +   0.885
    17        1:N        1:N       0.993          46480        2:S        1:N   +   0.965
```

Figura 5. Prediccions del test

Finalment analitzarem els errors presentats per aquest algorisme, com ens podem donar compte en les gràfiques; les instàncies del quadrant superior esquerre i inferior dret serien les instàncies no classificades, que es mostren com quadres vermells o blaus, mentre que les instàncies en forma d'aspa estan classificades correctament.

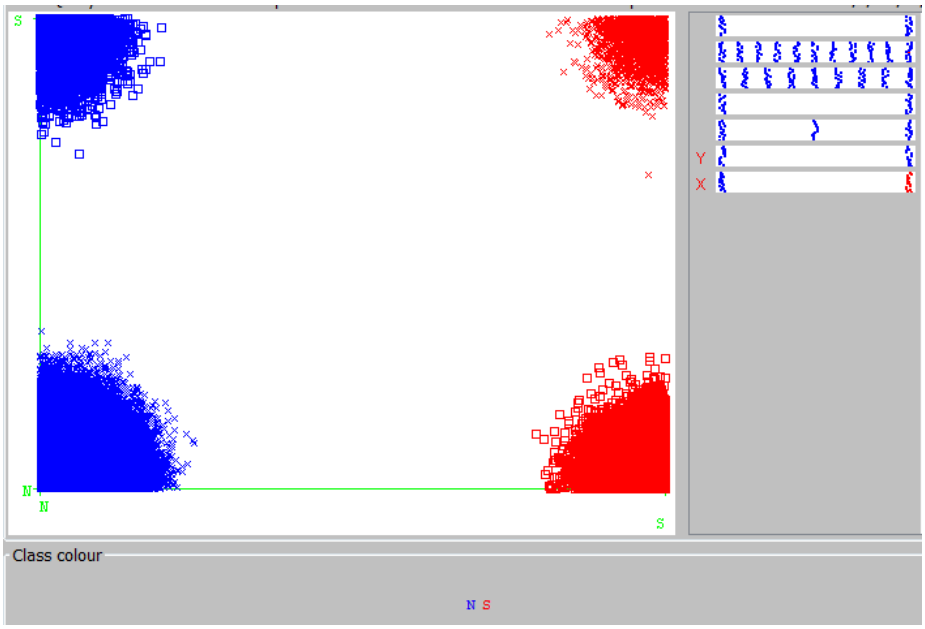


Figura 6. Gràfica d'errors de classificació NB

En cas d'emprar l'eina Elvira en seleccionar les eines d'aprenentatge automàtic es farà un preprocessat automàtic en que es reemplaçarà per zeros la manca de valor en alguna de les variables a importar.

Això es fa per tal de mostrar com opera el classificador sense influència d'altres algoritmes que el poden complementar en la generació de la xarxa Bayesiana. Per tant es processarà directament l'eina d'"Aprenentatge automàtic".

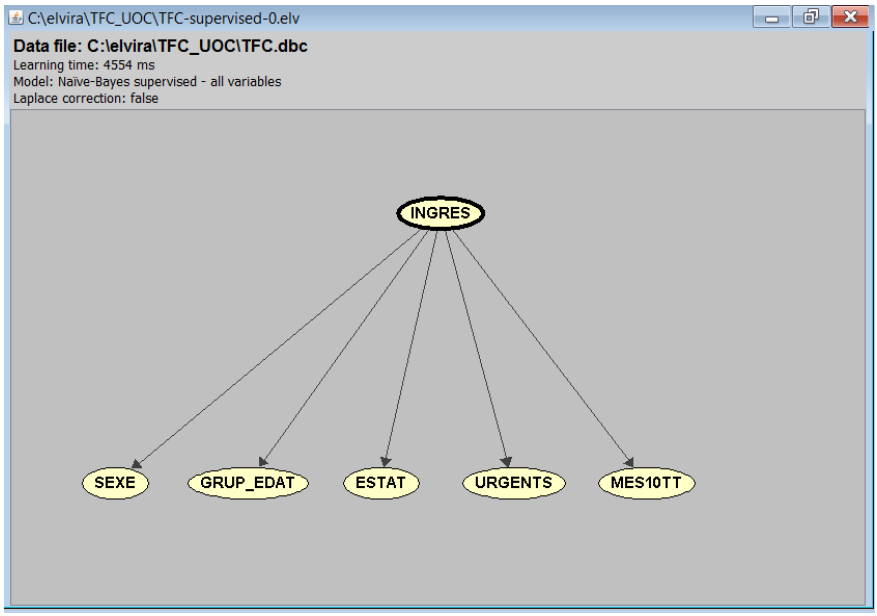


Figura 7. Finestra principal apareix la gràfica de la xarxa generada.

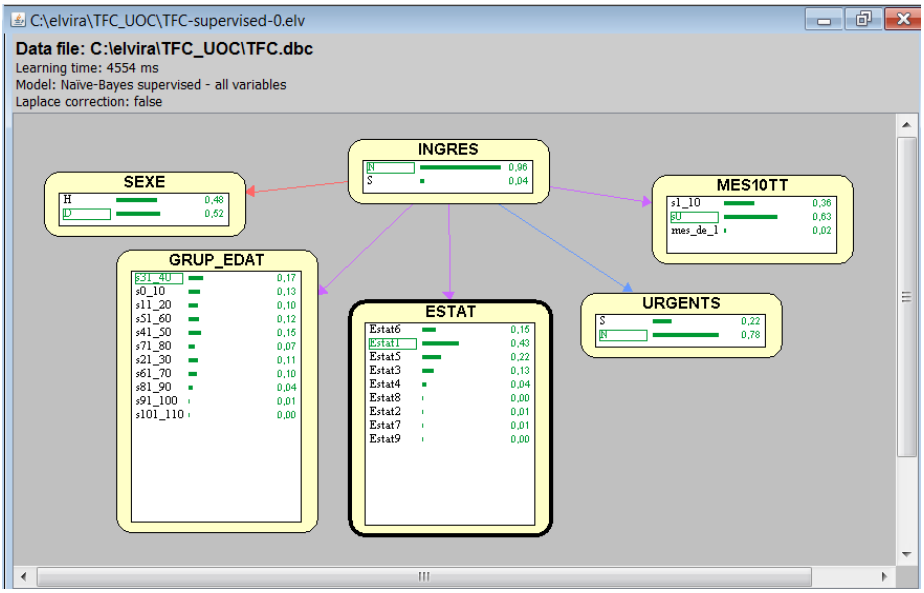


Figura 8. Xarxa Bayesiana en mode d'inferència

Podem veure les probabilitats de cadascuna de les diferents variables. Un cop analitzats tots els tipus de classificadors s'analitzaran amb més detall les opcions del menú d'Inferència. El detall dels valors de probabilitat associats a cada node en particular es detallen en els quadres corresponents del node i la classe. La gràfica que ens ofereix Elvira mostra els diferents tipus d'influències entre nodes calorejats de forma diferent. Així els enllaços amb un impacte positiu tenen un color vermell, els negatius estan en blau, els nuls en negre i els d'influència indefinida en violeta (xarxes causals en vermell). En el nostre gràfic sembla que detecta un impacte positiu sobre un ingrés del sexe femení i un impacte negatiu del fet d'haver patit una atenció urgent prèvia a atenció primària. Si visualitzem la funció "explain node" podem veure la relació d'odds ratio entre les diferents probabilitats de les categories dels paràmetres. Així en la figura 10 podem veure els Odds Ratio del node ESTAT, on podem veure les posicions que ocupen els diferents valors del paràmetre segons la seva Odds.

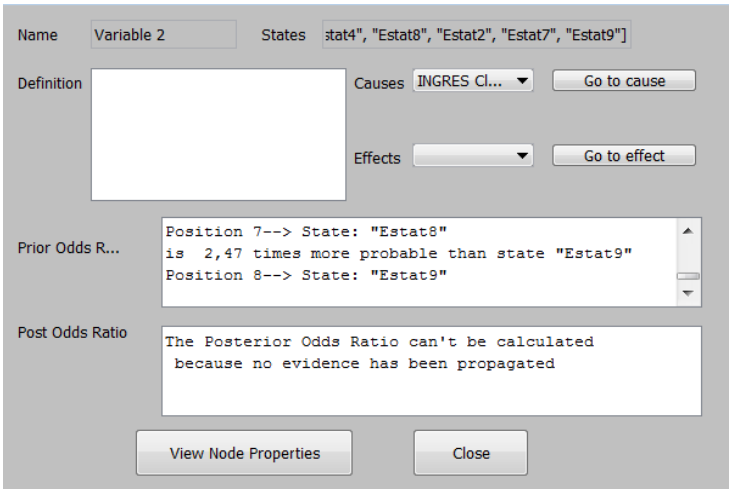


Figura 9. Odds ratio dels nodes i classes

Així podem veure les següents probabilitat ràtio respecte als ESTATs (la disposició dels valors d'acord a un ordre decreixent de probabilitats és):

Posició 0 -> Estat: "Estat1"

és 1,97 vegades més probable que l'estat "Estat5"
és 2,86 vegades més probable que l'estat "Estat6"
és 3,35 vegades més probable que l'estat "Estat3"
és 9,82 vegades més probable que l'estat "Estat4"
és 39,59 vegades més probable que l'estat "Estat7"
és 69,64 vegades més probable que l'estat "Estat2"
és 92,74 vegades més probable que l'estat "Estat8"
és 229,25 vegades més probable que l'estat "Estat9"

Posició 1 -> Estat: "Estat5"

és 1,45 vegades més probable que l'estat "Estat6"
és 1,70 vegades més probable que l'estat "Estat3"
és 4,97 vegades més probable que l'estat "Estat4"
és 20,05 vegades més probable que l'estat "Estat7"
és 35,27 vegades més probable que l'estat "Estat2"
és 46,97 vegades més probable que l'estat "Estat8"
és 116,11 vegades més probable que l'estat "Estat9"

Posició 2 -> Estat: "Estat6"

és 1,17 vegades més probable que l'estat "Estat3"
és 3,44 vegades més probable que l'estat "Estat4"
és 13,86 vegades més probable que l'estat "Estat7"
és 24,37 vegades més probable que l'estat "Estat2"
és 32,46 vegades més probable que l'estat "Estat8"
és 80,23 vegades més probable que l'estat "Estat9"

Posició 3 -> Estat: "Estat3"

és 2,93 vegades més probable que l'estat "Estat4"
és 11,83 vegades més probable que l'estat "Estat7"
és 20,80 vegades més probable que l'estat "Estat2"
és 27,70 vegades més probable que l'estat "Estat8"
és 68,49 vegades més probable que l'estat "Estat9"

Posició 4 -> Estat: "Estat4"

és 4,03 vegades més probable que l'estat "Estat7"
és 7,09 vegades més probable que l'estat "Estat2"
és 9,44 vegades més probable que l'estat "Estat8"
és 23,35 vegades més probable que l'estat "Estat9"

Posició 5 -> Estat: "Estat7"

és 1,76 vegades més probable que l'estat "Estat2"
és 2,34 vegades més probable que l'estat "Estat8"
és 5,79 vegades més probable que l'estat "Estat9"

Posició 6 -> Estat: "Estat2"

és 1,33 vegades més probable que l'estat "Estat8"
és 3,29 vegades més probable que l'estat "Estat9"

Posició 7 -> Estat: "Estat8"

és 2,47 vegades més probable que l'estat "Estat9"

Posició 8 -> Estat: "Estat9"

En el cas dels nodes dicotòmics [H, D] o [si, no] si la probabilitat ràtio és inferior a 1, implica que l'estat "D" o "SI", és més probable que l'estat "H" o "NO". En el cas del sexe D és 1,08 vegades més provable que H.

Els valors més provables es mostren amb requadre verd al voltant. Així NO ingressar, NO acudir a urgències, NO seguir cap tractament crònic, el grup d'edat entre 31-40 anys, l'Estat de CRGs 1i el sexe femení, són els valors més provables de cada node.

Valors de probabilitat associats a cada node en particular respecte a la classe "causa" (ingrés hospitalari):

N	0.9615532735655627
S	0.03844672643443725

Classe INGRES

INGRES Class	N	S
H	0.48424022607617767	0.41298265249020705
D	0.5157597739238223	0.587017347509793

Node SEXE

INGRES Class	N	S
s31_40	0.1686129116705338	0.18539451594851708
s0_10	0.13609096746910584	0.08668158925573587
s11_20	0.1051442175379087	0.032904308897593736
s51_60	0.12403536132137319	0.08136541689983212
s41_50	0.15269763206236336	0.0752658086177952
s71_80	0.06625437149972367	0.14185786233911585
s21_30	0.10891215383204043	0.11253497481813095
s61_70	0.09616068771549843	0.10755456071628428
s81_90	0.03648454228747743	0.14622271964185787
s91_100	0.005497517502780084	0.029714605484051484
s101_110	1.0963710119493708E-4	5.036373810856087E-4

Node GRUP_EDAT

INGRES Class	N	S
S	0.21126845651098944	0.3224398433128148
N	0.7887315434890105	0.6775601566871852

Node URGENT

INGRES Class	N	S
Estat6	0.14478809833329231	0.3165081141578064
Estat1	0.43940536416298787	0.26138780078343593
Estat5	0.22004389959031523	0.19501958589815332
Estat3	0.131450409349136	0.07358701734750979
Estat4	0.04474088725502261	0.026860660324566313
Estat8	0.003799261180187457	0.02630106323447118
Estat2	0.005396830369029533	0.026580861779518745
Estat7	0.008730693242103332	0.06580861779518747
Estat9	0.0016445565179258326	0.007946278679350605

Node ESTAT

INGRES Class	N	S
s1_10	0.35100429822186524	0.4732512590934527
s0	0.6363359728278989	0.4327364297705652
mes_de_10	0.012659728950235927	0.09401231113598207

Node MES_10_TT

Figura 10. Probabilitat associada a la classe INGRÉS de cada node

Dels valors de les taules es pot observar una lleugera major probabilitat d'ingressar del sexe femení. Una menor probabilitat d'ingrés en els rangs d'edat entre els 0 – 20 anys i entre els 41 – 60 anys, invertint la tendència entre els 21-40 i a partir dels 61 anys. Els Estats que s'associen més probablement a un ingrés són els E2 (malaltia aguda severa), E6 (Malalties cròniques lleus en múltiples òrgans), E7 (Malaltia crònica severa en 3 o més òrgans), E8 (càncers greus i metàstasi) i E9 (malalties catastròfiques).

La probabilitat d'ingressar també sembla superior entre els que s'han visitat alguna vegada a urgències de primària que entre els que no ho han fet.

Quan al nombre de tractaments crònics del pacient, sembla que la probabilitat d'ingrés augmenta entre els pacients que tenen algun tractament crònic, especialment entre els que prenen més de 10.

Però com ja hem comentat abans, NB només pot representar distribucions de probabilitat senzilles (arbre de profunditat 1) i sovint aquesta situació, a la pràctica no es satisfà i existeixen dependències entre els atributs, per el que emprarem els altres classificadors que poden representar situacions mes arbitràries i que ens poden mostrar millors resultats, donat que afegir profunditat a la xarxa sovint maximitza la probabilitat de les dades. Els següents classificadors ens permetran optimitzar els nodes afegint/eliminant arcs des d'altres nodes.

- **Obtenció d'una xarxa mitjançant el classificador K2 - BayesNet**

Aquest classificador és senzill i eficient i ens permet l'ús de qualsevol estructura (GDA). El K2 ens ha permès fer una ordenació dels diferents atributs (nodes) per fer un processament posterior segons aquest ordre, afegint arcs d'associació dels nodes ja processats anteriorment de forma continuada. Un cop que el node ja no es pugui optimitzar mes passa al següent. D'aquesta forma els diferents atributs es van associant de forma múltiple però ordena fins trobar el conjunt d'arcs que millor associa els nodes. Els primers resultats d'aquest algorisme mostren 17886 instàncies classificades incorrectament (3,85%) i mes del 96% correctament classificades, millorant el percentatge d'instàncies correctament classificades respecte al classificador anterior (NB).

```
=== Classifier model (full training set) ===

Bayes Network Classifier
Using ADTree
#attributes=6 #classindex=5
Network structure (nodes followed by parents)
SEXE(2): INGRES
GRUP_EDAT(11): INGRES SEXE
ESTAT(9): INGRES GRUP_EDAT SEXE
URGENTS(2): INGRES GRUP_EDAT ESTAT
MES_10_TTS(3): INGRES GRUP_EDAT ESTAT
INGRES(2):
LogScore Bayes: -2441725.7049031244
LogScore BDeu: -2447440.0222831126
LogScore MDL: -2446943.5652970527
LogScore ENTROPY: -2440490.6555484575
LogScore AIC: -2441479.6555484575
```

```
=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances      446914               96.1519 %
Incorrectly Classified Instances     17886                3.8481 %
Kappa statistic                          0.0036
Mean absolute error                      0.0696
Root mean squared error                  0.1867
Relative absolute error                 94.0761 %
Root relative squared error             97.1217 %
Coverage of cases (0.95 level)          98.5957 %
Mean rel. region size (0.95 level)      63.1987 %
Total Number of Instances           464800
```

```
=== Detailed Accuracy By Class ===

               TP Rate  FP Rate  Precision  Recall  F-Measure  MCC    ROC Area  PRC Area  Class
               1        0.998    0.962      1       0.98       0.027  0.762     0.985     N
               0.002    0        0.409      0.002   0.004      0.027  0.762     0.146     S
Weighted Avg.  0.962    0.96     0.94       0.962   0.943      0.027  0.762     0.953

=== Confusion Matrix ===

      a      b   <-- classified as
 446878     52 |      a = N
  17834     36 |      b = S
```

Figura 11. Finestra del classificador de sortida K2.

La figura 14 presenta la xarxa bayesiana obtinguda, on els nodes representen les diferents variables, com el sexe, el grup d'edat, l'estat, el nombre de tractaments crònics o les visites urgents que poden conduir a un ingrés hospitalari. Com ja s'ha comentat, la variable a la qual apunta un arc és dependent de la que està en l'origen d'aquest. Per exemple, urgent depèn d'edat.

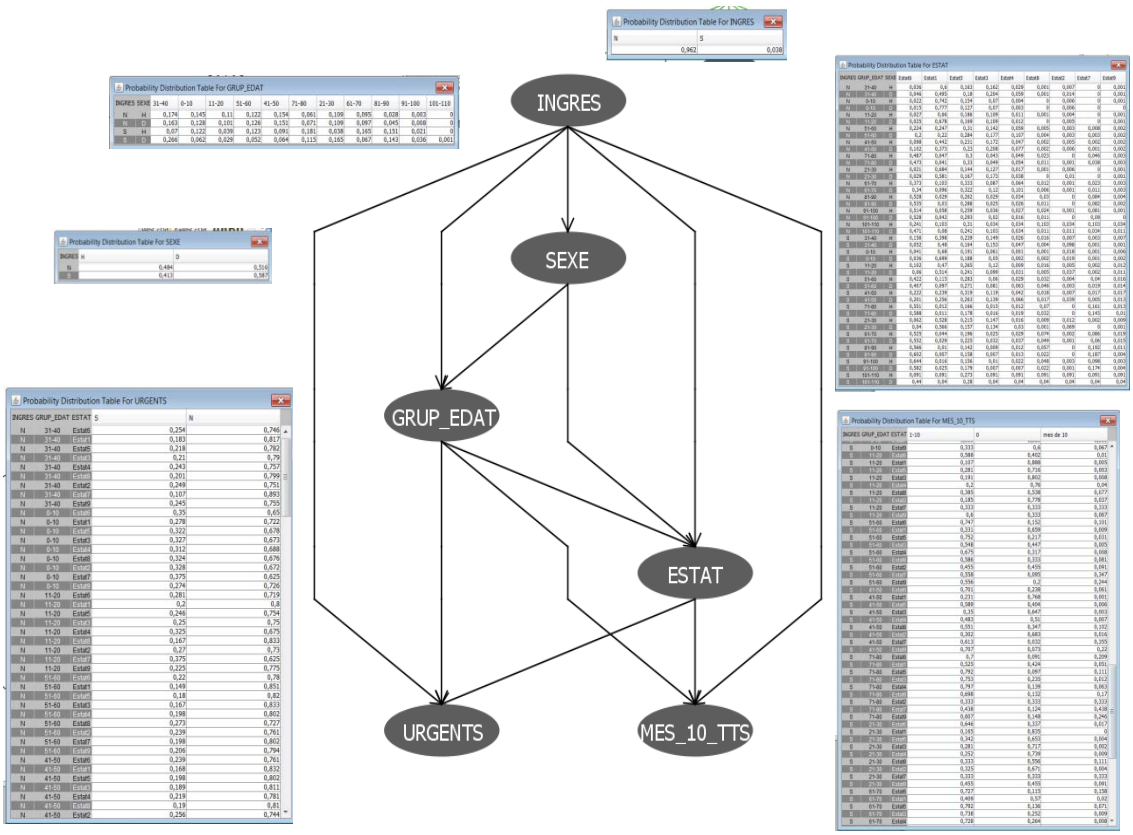


Figura 12. **Arbre de classificació generat per K2.**

En la gràfica podem determinar que en realitat la majoria d'instàncies classificada incorrectament es troben entre els predits com a ingrés (VPP), que es mostren com un quadre vermell. Weka ens ofereix la possibilitat de visualitzar el gràfic de l'arbre generat a partir de la classificació.

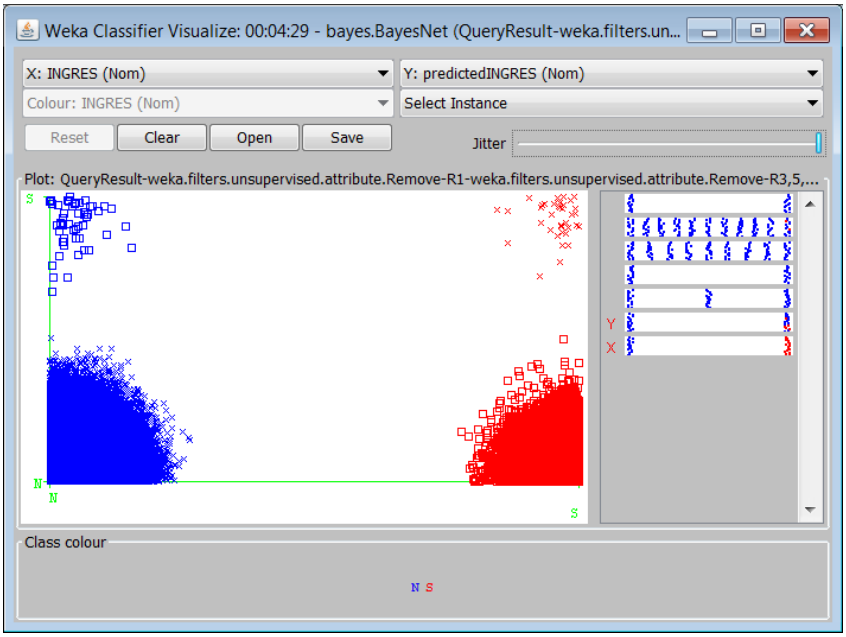


Figura 13. Gràfica d'errors de classificació K2

- **Obtenció d'una xarxa mitjançant el classificador TAN (Naïve Bayes Augmentat a Arbre)**

Amb el classificador TAN intentarem millorar les probabilitats dels nodes a través d'una extensió en arbre de l'NB. Amb aquest classificador començarem amb una estructura de NB, considerant la possibilitat d'afegir un pare addicional, evitant els cicles, per tal de trobar un algoritme més eficient que els anteriors (nou pare grup_edat, per exemple). Els primers resultats d'aquest algorisme mostren només un 3,9% d'instàncies classificades incorrectament (18120) i poc mes del 96% correctament classificades.

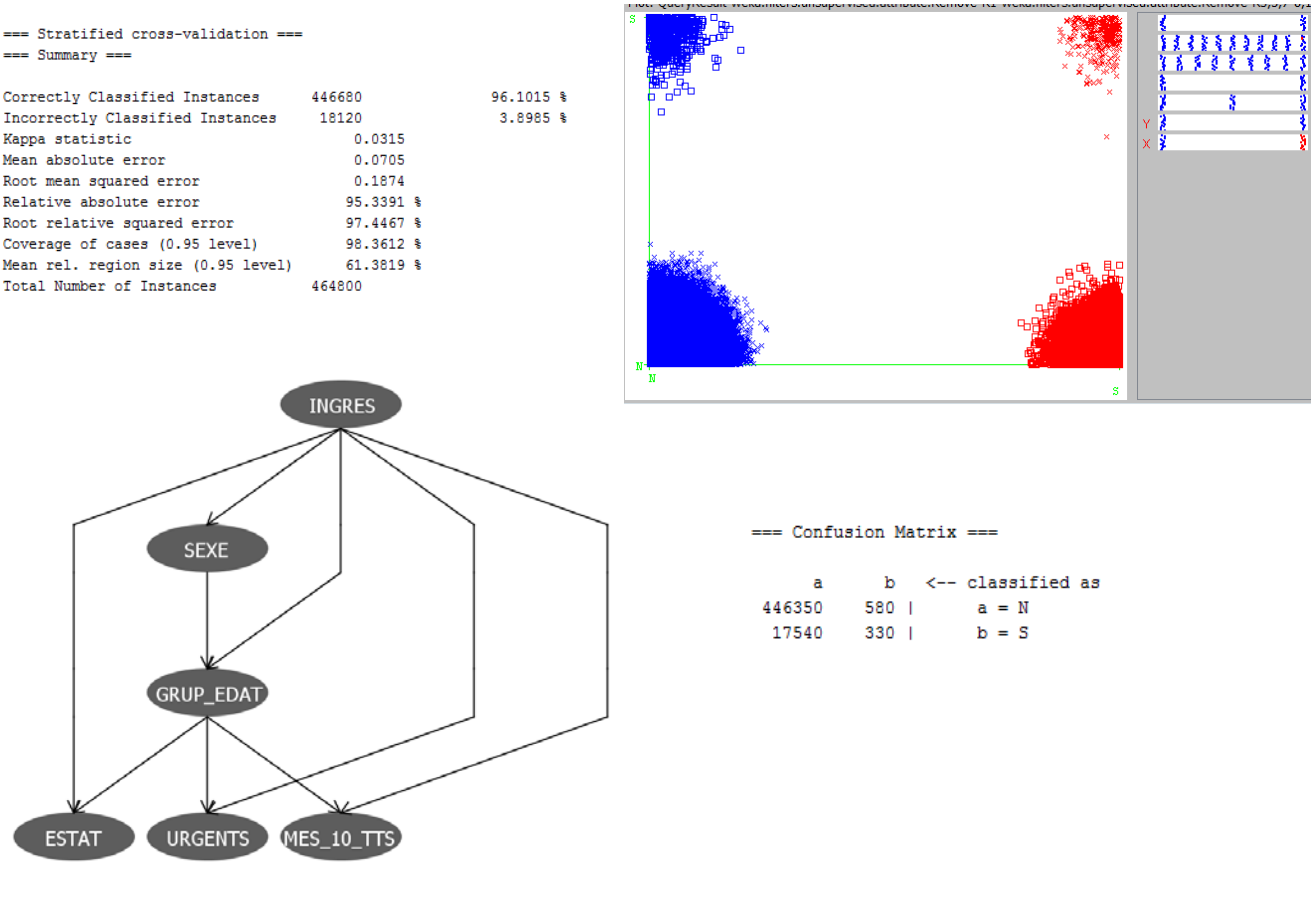


Figura 14. Classificador de sortida i gràfica d'errors i arbre generat per TAN.

En aquest cas, podem veure que els resultats són millors que amb el classificador NB, però lleugerament pitjors que amb el classificador K2, amb el que no ens aportaria avantatges sobre el K2.

Quan l'apliquem amb el programa Elvira, l'algorisme proposat per Friedman et al. 1996 construeix un model TAN, s'adapta a l'algorisme de Chow & Liu condicional, utilitzant la informació mútua de dues variables predictores donada la classe.

Aplicant Elvira ens apareix el mateix gràfic d'arbre que amb Weka, observant la relació múltiple del node GRUP_EDAT, que estableix valors de probabilitat amb la resta de nodes (SEXE, ESTAT, URGENTS i MES_10_TTS). Els valors de probabilitat associada a cada node en particular respecte a la classe "causa" (ingrés hospitalari i Grup d'edat) són els següents:

Variable 1	s31_40	s31_40	s0_10	s0_10	s11_20	s11_20	s51_60	s51_60	s41_50	s41_50	s71_80	s71_80	s21_30	s21_30	s61_70	s61_70	s81_90	s81_90	s91_100	s91_1...	s101...	s101_...
INGRES Cl...	N	S	N	S	N	S	N	S	N	S	N	S	N	S	N	S	N	S	N	S	N	S
H	0.500...	0.156...	0.514...	0.582...	0.506...	0.482...	0.474...	0.623...	0.488...	0.499...	0.445...	0.525...	0.482...	0.137...	0.478...	0.634...	0.369...	0.426...	0.261...	0.288...	0.204...	0.111...
D	0.499...	0.843...	0.485...	0.417...	0.493...	0.517...	0.525...	0.376...	0.511...	0.500...	0.554...	0.474...	0.517...	0.862...	0.521...	0.365...	0.630...	0.573...	0.738...	0.711...	0.795...	0.888...

Node SEXE

Variable 1	s31_40	s31_40	s0_10	s0_10	s11_20	s11_20	s51_60	s51_60	s41_50	s41_50	s71_80	s71_80	s21_30	s21_30	s61_70	s61_70	s81_90	s81_90	s91_1...	s91_100	s101_...	s101_110
INGRES Cl...	N	S	N	S	N	S	N	S	N	S	N	S	N	S	N	S	N	S	N	S	N	S
Estat6	0.040...	0.068...	0.018...	0.038...	0.026...	0.079...	0.211...	0.418...	0.099...	0.212...	0.479...	0.570...	0.025...	0.042...	0.355...	0.536...	0.532...	0.588...	0.525...	0.608...	0.469...	0.5555...
Estat1	0.547...	0.468...	0.759...	0.691...	0.669...	0.498...	0.232...	0.107...	0.406...	0.248...	0.043...	0.011...	0.630...	0.562...	0.099...	0.037...	0.029...	0.008...	0.045...	0.020...	0.081...	0.0
Estat5	0.171...	0.176...	0.140...	0.190...	0.177...	0.255...	0.296...	0.279...	0.230...	0.292...	0.316...	0.171...	0.155...	0.165...	0.327...	0.207...	0.278...	0.151...	0.284...	0.173...	0.285...	0.4444...
Estat3	0.183...	0.152...	0.070...	0.056...	0.109...	0.108...	0.160...	0.067...	0.190...	0.129...	0.046...	0.015...	0.150...	0.136...	0.104...	0.027...	0.026...	0.007...	0.023...	0.005...	0.081...	0.0
Estat4	0.044...	0.043...	0.003...	6.455...	0.011...	0.018...	0.084...	0.041...	0.061...	0.053...	0.052...	0.014...	0.027...	0.027...	0.083...	0.031...	0.028...	0.012...	0.018...	0.009...	0.020...	0.0
Estat8	9.421...	0.005...	2.630...	6.455...	4.256...	0.008...	0.004...	0.037...	0.001...	0.017...	0.016...	0.051...	5.136...	0.001...	0.008...	0.064...	0.018...	0.036...	0.014...	0.028...	0.020...	0.0
Estat2	0.010...	0.083...	0.006...	0.018...	0.004...	0.020...	0.002...	0.002...	0.005...	0.022...	3.714...	0.0	0.008...	0.061...	0.001...	0.001...	6.132...	0.0	0.0	0.0	0.0	0.0
Estat7	1.725...	0.001...	4.932...	0.0	6.384...	0.0	0.005...	0.031...	0.001...	0.010...	0.041...	0.153...	0.0	0.0	0.016...	0.076...	0.082...	0.189...	0.087...	0.152...	0.040...	0.0
Estat9	0.001...	0.001...	6.740...	0.003...	0.001...	0.010...	0.002...	0.014...	0.001...	0.014...	0.002...	0.011...	0.001...	0.001...	0.002...	0.017...	0.002...	0.006...	0.0	0.001...	0.0	0.0

Node ESTAT

Variable 3	S	S	N	N
INGRES Class	N	S	N	S
s31_40	0.1599097...	0.1504685...	0.1709441...	0.2020151...
s0_10	0.1862595...	0.1091634...	0.1226528...	0.0759828...
s11_20	0.1083751...	0.0341895...	0.1042787...	0.0322926...
s51_60	0.1065535...	0.0846928...	0.1287180...	0.0797819...
s41_50	0.1371396...	0.0685525...	0.1568649...	0.0784605...
s71_80	0.0620406...	0.1555015...	0.0673830...	0.1353650...
s21_30	0.1157251...	0.0999652...	0.1070872...	0.1185166...
s61_70	0.0803943...	0.1124609...	0.1003838...	0.1052196...
s81_90	0.0373641...	0.1572370...	0.0362489...	0.1409811...
s91_100	0.0061108...	0.0269003...	0.0053332...	0.0310538...
s101_110	1.2708902...	8.6775425...	1.0496245...	3.3036009...

Node SEXE

Variable 1	s31_40	s31_40	s0_10	s0_10	s11_20	s11_20	s51_60	s51_60	s41_50	s41_50	s71_80	s71_80	s21_30	s21_30	s61_70	s61_70	s81_90	s81_90	s91_100	s91_100	s101_110	s101_110
INGRES Cl...	N	S	N	S	N	S	N	S	N	S	N	S	N	S	N	S	N	S	N	S	N	S
s1_10	0.203...	0.296...	0.05...	0.08...	0.087...	0.209...	0.5768...	0.674...	0.347...	0.484...	0.812...	0.676...	0.146...	0.24...	0.761...	0.709...	0.784...	0.639...	0.7830...	0.6497...	0.6326...	0.55555...
s0	0.796...	0.702...	0.94...	0.91...	0.912...	0.789...	0.4152...	0.257...	0.650...	0.493...	0.121...	0.105...	0.853...	0.75...	0.214...	0.149...	0.104...	0.119...	0.1184...	0.1789...	0.3061...	0.44444...
mes_de_10	5.971...	0.001...	0.0	0.0	2.128...	0.001...	0.0079...	0.068...	0.001...	0.022...	0.066...	0.218...	2.054...	9.94...	0.023...	0.140...	0.110...	0.240...	0.0984...	0.1713...	0.0612...	0.0

Node MES_10_TT

Figura 15. Relació múltiple del node GRUP_EDAT amb la resta de nodes (SEXE, ESTAT, URGENTS i MES_10_TTS)

- ## Obtenció d'una xarxa mitjançant el classificador HILL-CLIMBER - BayesNet

Amb aquest classificador intentem optimitzar el classificador a través d'un algoritme que introdueix un mecanisme d'aleatorietat en la elecció de nodes i variables, de forma que durant una sèrie de successions definides, anirà provant relacions aleatòries quedant-se amb l'optima. Els primers resultats d'aquest algorisme igualen als resultats obtinguts amb el classificador K2, mostrant un 3,85%

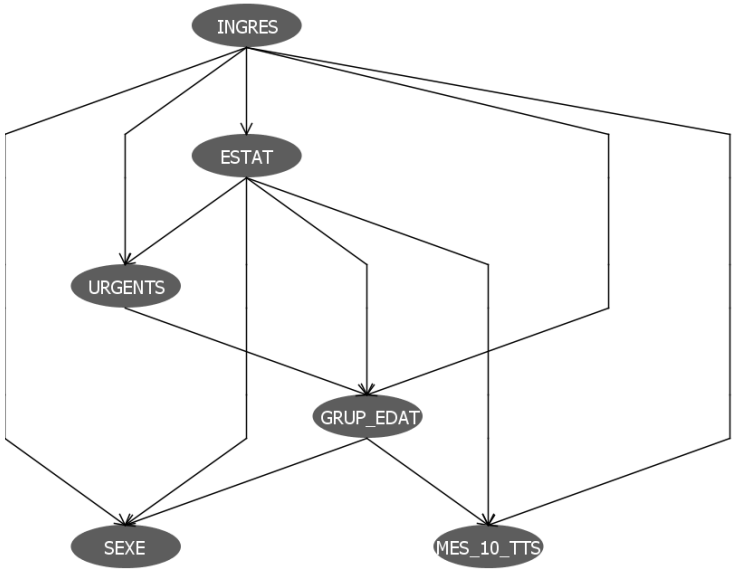
d'instàncies classificades incorrectament i mes del 96% correctament classificades. Amb valors molt similars als obtinguts amb l'algoritme K2

```
=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances      446914               96.1519 %
Incorrectly Classified Instances     17886                3.8481 %
Kappa statistic                          0.0035
Mean absolute error                      0.0696
Root mean squared error                  0.1867
Relative absolute error                 94.1177 %
Root relative squared error             97.1226 %
Coverage of cases (0.95 level)          98.596  %
Mean rel. region size (0.95 level)      63.1974 %
Total Number of Instances           464800


        === Confusion Matrix ===

           a      b   <-- classified as
      446879     51 |      a = N
       17835     35 |      b = S
```




```
=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall  F-Measure  MCC    ROC Area  PRC Area  Class
                 1        0.998    0.962      1       0.98       0.026  0.762     0.985     N
                 0.002    0        0.407      0.002   0.004      0.026  0.762     0.146     S
Weighted Avg.    0.962    0.96     0.94       0.962   0.943      0.026  0.762     0.953
```

Figura 16. Arbre de classificació generat per HC

- ## Possible obtenció d'una xarxa mitjançant el classificador KDB

El model NB i el model TAN són casos particulars d'un model més general, el model k-dependence (kdB) Bayesian network (KDB). Aquest classificador és una xarxa bayesiana en què cada node pot tenir un màxim de k variables característiques com pares, a part de la variable classe, que és un pare de tots els nodes. Així, el model Naive Bayes és un model 0dB, i el model TAN és un model 1. En el nostre estudi, aquest classificador ens permetrà realitzar el major nombre de dependències entre nodes per tal de trobar l'òptima.

k-dependence Bayesian classifier

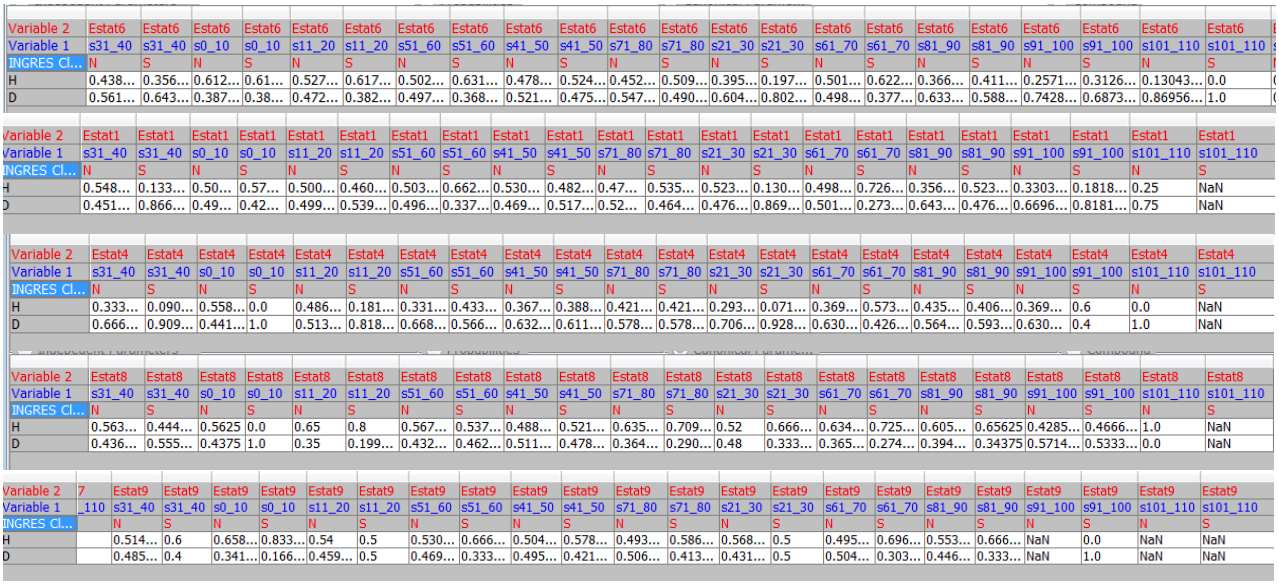
A continuació es detallen els valors de probabilitat associats a cada node en particular

N	0.9615532735655627
S	0.03844672643443725

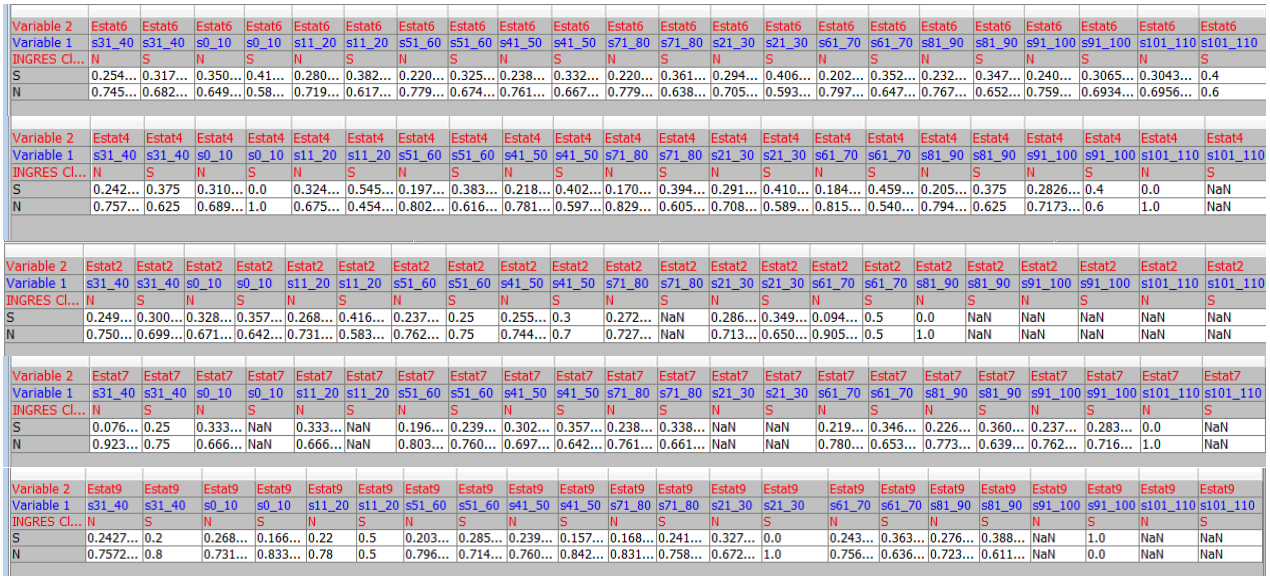
Taula 2. Node INGRES

Variable 2	Estat6	Estat6	Estat1	Estat1	Estat5	Estat5	Estat3	Estat3	Estat4	Estat4	Estat8	Estat8	Estat2	Estat2	Estat7	Estat7	Estat9	Estat9
INGRES Cl...	N	S	N	S	N	S	N	S	N	S	N	S	N	S	N	S	N	S
s31_40	0.047...	0.040...	0.210...	0.332...	0.131...	0.167...	0.235...	0.383...	0.16...	0.3	0.041...	0.038...	0.332...	0.581...	0.003...	0.003...	0.140...	0.0352...
s0_10	0.017...	0.010...	0.235...	0.229...	0.087...	0.084...	0.072...	0.066...	0.01...	0.002...	0.009...	0.002...	0.156...	0.058...	7.688...	0.0	0.055...	0.0422...
s11_20	0.019...	0.008...	0.160...	0.062...	0.084...	0.043...	0.087...	0.048...	0.02...	0.022...	0.011...	0.010...	0.087...	0.025...	7.688...	0.0	0.068...	0.0422...
s51_60	0.181...	0.107...	0.065...	0.033...	0.167...	0.116...	0.151...	0.074...	0.23...	0.125	0.143...	0.114...	0.067...	0.008...	0.073...	0.039...	0.153...	0.1478...
s41_50	0.105...	0.050...	0.141...	0.071...	0.160...	0.112...	0.220...	0.132...	0.21...	0.15	0.078...	0.048...	0.164...	0.063...	0.019...	0.011...	0.159...	0.1338...
s71_80	0.219...	0.255...	0.006...	0.005...	0.095...	0.124...	0.023...	0.029...	0.07...	0.079...	0.280...	0.278...	0.004...	0.0	0.318...	0.331...	0.112...	0.2042...
s21_30	0.018...	0.015...	0.156...	0.242...	0.077...	0.095...	0.124...	0.208...	0.06...	0.116...	0.014...	0.006...	0.163...	0.258...	0.0	0.0	0.078...	0.0281...
s61_70	0.236...	0.182...	0.021...	0.015...	0.143...	0.114...	0.076...	0.039...	0.17...	0.127...	0.225...	0.263...	0.021...	0.004...	0.182...	0.125	0.167...	0.2323...
s81_90	0.134...	0.271...	0.002...	0.004...	0.046...	0.113...	0.007...	0.015...	0.02...	0.066...	0.173...	0.204...	4.145...	0.0	0.346...	0.420...	0.063...	0.1267...
s91_100	0.019...	0.057...	5.703...	0.002...	0.007...	0.026...	9.872...	0.002...	0.00...	0.010...	0.020...	0.031...	0.0	0.0	0.055...	0.068...	0.0	0.0070...
s101_110	3.554...	8.840...	2.036...	2.220...	1.423...	0.001...	6.808...	0.0	5.00...	0.0	5.889...	1.110...	0.0	0.0	5.125...	1.110...	0.0	0.0

Taula 3. Node GRUP_EDAT



Taula 4. Node SEXE



Taula 5. Node URGENT

Variable 2	Estat2	Estat2	Estat2	Estat2	Estat2	Estat2	Estat2	Estat2	Estat2	Estat2	Estat2	Estat2	Estat2	Estat2	Estat2	Estat2	Estat2	Estat2	Estat2	Estat2	Estat2	Estat2
Variable 1	s31_40	s31_40	s0_10	s0_10	s11_20	s11_20	s51_60	s51_60	s41_50	s41_50	s71_80	s71_80	s21_30	s21_30	s61_70	s61_70	s81_90	s81_90	s91_100	s91_100	s101_110	s101_110
INGRES Cl...	N	S	N	S	N	S	N	S	N	S	N	S	N	S	N	S	N	S	N	S	N	S
s1_10	0.215...	0.307...	0.079...	0.035...	0.113...	0.166...	0.353...	0.5	0.232...	0.3	0.818...	NaN	0.170...	0.325...	0.547...	0.0	1.0	NaN	NaN	NaN	NaN	NaN
s0	0.784...	0.692...	0.920...	0.964...	0.886...	0.833...	0.646...	0.5	0.765...	0.7	0.181...	NaN	0.829...	0.674...	0.452...	1.0	0.0	NaN	NaN	NaN	NaN	NaN
mes_de_10	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.002...	0.0	0.0	NaN	0.0	0.0	0.0	0.0	0.0	NaN	NaN	NaN	NaN	NaN

Variable 2	Estat7	Estat7	Estat7	Estat7	Estat7	Estat7	Estat7	Estat7	Estat7	Estat7	Estat7	Estat7	Estat7	Estat7	Estat7	Estat7	Estat7	Estat7	Estat7	Estat7	Estat7	Estat7
Variable 1	s31_40	s31_40	s0_10	s0_10	s11_20	s11_20	s51_60	s51_60	s41_50	s41_50	s71_80	s71_80	s21_30	s21_30	s61_70	s61_70	s81_90	s81_90	s91_100	s91_100	s101_110	s101_110
INGRES Cl...	N	S	N	S	N	S	N	S	N	S	N	S	N	S	N	S	N	S	N	S	N	S
s1_10	0.538...	0.5	0.333...	NaN	0.333...	NaN	0.771...	0.565...	0.776...	0.642...	0.652...	0.438...	NaN	NaN	0.708...	0.564...	0.617...	0.491...	0.6465...	0.5432...	0.5	NaN
s0	0.461...	0.25	0.666...	NaN	0.666...	NaN	0.091...	0.086...	0.105...	0.0	0.073...	0.123...	NaN	NaN	0.094...	0.102...	0.084...	0.115...	0.1162...	0.1851...	0.0	NaN
mes_de_10	0.0	0.25	0.0	NaN	0.0	NaN	0.136...	0.347...	0.118...	0.357...	0.273...	0.438...	NaN	NaN	0.196...	0.333...	0.297...	0.392...	0.2372...	0.2716...	0.5	NaN

Variable 2		Estat9	Estat9	Estat9	Estat9	Estat9	Estat9	Estat9	Estat9	Estat9	Estat9	Estat9	Estat9	Estat9	Estat9	Estat9	Estat9	Estat9	Estat9	Estat9	Estat9	Estat9	Estat9
Variable 1	0	s31_40	s31_40	s0_10	s0_10	s11_20	s11_20	s51_60	s51_60	s41_50	s41_50	s71_80	s71_80	s21_30	s21_30	s61_70	s61_70	s81_90	s81_90	s91_100	s91_100	s101_110	s101_110
INGRES Cl...		N	S	N	S	N	S	N	S	N	S	N	S	N	S	N	S	N	S	N	S	N	S
s1_10		0.524...	1.0	0.268...	0.333...	0.36	0.666...	0.761...	0.571...	0.623...	0.736...	0.734...	0.620...	0.568...	0.5	0.796...	0.545...	0.765...	0.666...	NaN	0.0	NaN	NaN
s0		0.427...	0.0	0.731...	0.666...	0.64	0.333...	0.132...	0.190...	0.324...	0.052...	0.108...	0.137...	0.413...	0.5	0.089...	0.060...	0.085...	0.111...	NaN	0.0	NaN	NaN
mes_de_10		0.048...	0.0	0.0	0.0	0.0	0.0	0.106...	0.238...	0.051...	0.210...	0.156...	0.241...	0.017...	0.0	0.113...	0.393...	0.148...	0.222...	NaN	1.0	NaN	NaN

Taula 6. Node MES_10_TTS

INGRES Class	N	S
Estat6	0.144788098...	0.3165081141...
Estat1	0.439405364...	0.2613878007...
Estat5	0.220043899...	0.1950195858...
Estat3	0.131450409...	0.0735870173...
Estat4	0.044740887...	0.0268606603...
Estat8	0.003799261...	0.0263010632...
Estat2	0.005396830...	0.0265808617...
Estat7	0.008730693...	0.0658086177...
Estat9	0.001644556...	0.0079462786...

Taula 7. Node ESTAT

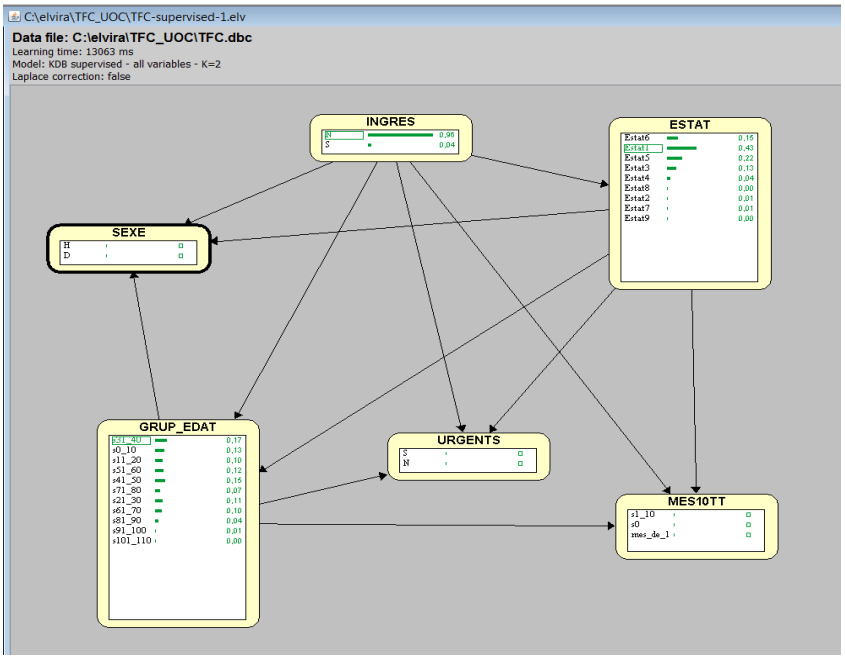


Figura 17. Xarxa Bayesiana KDB en mode d'inferencia

La idea dels algoritmes emprats és construir un PDG amb bosc de variables amb estructura lineal, on l'arrel és la variable classe i les característiques es disposen linealment en ordre topològico. Els nodes paramètricos es connecten de manera que hi hagi un camí entre cada un d'ells i la configuraci'on de les seves variables pare a la xarxa bayesiana original que es correspon amb la distribució continguda en aquest node. L'algorisme és vàlid per als classificadors esmentats anteriorment, excepte per alguns com els KDB amb subestructures més complexes de dificil representació en arbre. Aquest fet fa que no es processi en weka, descartant el seu anàlisi final.

Tot i l'ús dels diferents algoritmes classificadors amb la intenció de trobar un de més òptim, a mesura que aumentaven el nivell de complexitat, els resultats obtinguts no han estat molt diferents (NB amb CCI 94,3/ ICI 4,64 vs K2 amb CCI 96,1 / ICI 3,84), amb unes precisions mitjanes del 94%; per el que haurem de valorar altres paràmetres per avaluar diferències significatives.

RESULTATS (Evaluation)

Un cop obtingut l'arbre de classificació, és necessari poder avaluar la qualitat o nivell de confiança d'aquest. La mètrica Kappa Statistic és considerada per aquesta finalitat. El Kappa Statistic és una mesura que permet conèixer el nivell de predicció respecte a la variable considerada com a classe.

Hem de tenir en compte que per a cada arbre de decisió generat es presenten certes característiques importants que també comentarem i que ens permeten una millor conceptualització i entesa dels resultats.

Probabilitat condicional a posteriori

Quan disposem d'una xarxa bayesiana necessitem obtenir noves conclusions a mesura que anem obtenint nova informació, o evidència. El mecanisme per obtenir conclusions a partir de l'evidència es coneix com propagació de l'evidència o, de manera més simple, propagació [36, 37, 38], encara alguns autors utilitzen una altra terminologia com propagació de la incertesa o inferència probabilística. Podem llavors veure que la propagació consisteix en realitzar els càlculs necessaris per obtenir la probabilitat a posteriori d'una o diverses variables donats els valors assignats a altres variables en la xarxa bayesiana (evidència E), és a dir, el càlcul de P (Xi | E), on Xi i E són una única variable o un conjunt d'elles.

Per exemple, en els resultats obtinguts amb l'algoritme NB, que figuren en les taules 1 – 6, per a la variable sexe podem inferir que si el sexe és "D" (femení) la probabilitat que es produeixi un ingrés hospitalari és del 59% i en el cas de tractar d'una "H" (masculí) serà del 41%. Si seleccionem el valor ingrés hospitalari podem tenir aquesta relació en funció de l'ingrés [S]:

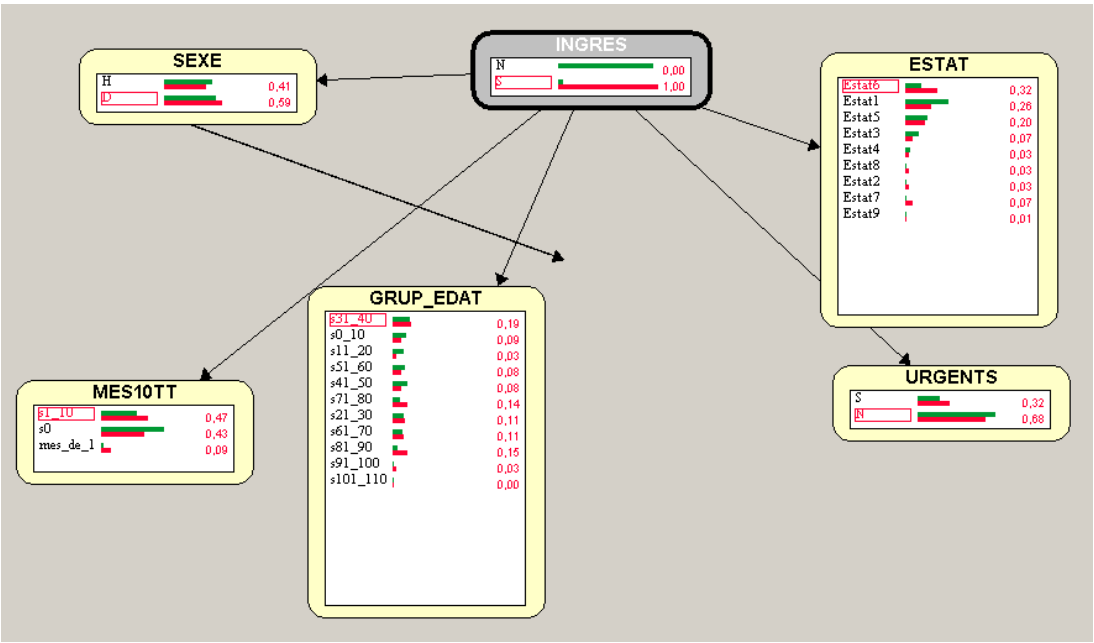


Figura 18. Xarxa Bayesiana TAN ajustada a ingrés hospitalari

Entre els usuaris que presenten un ingrés hospitalari predomina el sexe femení, els d'edat compresa entre els 31 i 40 anys, el usuaris en Estadi 6, aquells que no han acudit mai a urgències de primària de forma prèvia i aquells que prenen entre 1 i 10 medicaments crònics.

Si l'ajust el faig per ingressos hospitalaris i absència de visites a urgències de primària, observo que aquests comportaments es continuen mantenint, predominant el sexe femení, els grup d'edat d'entre 31 i 40 anys i els que prenen entre 1 i 10 medicaments crònics.

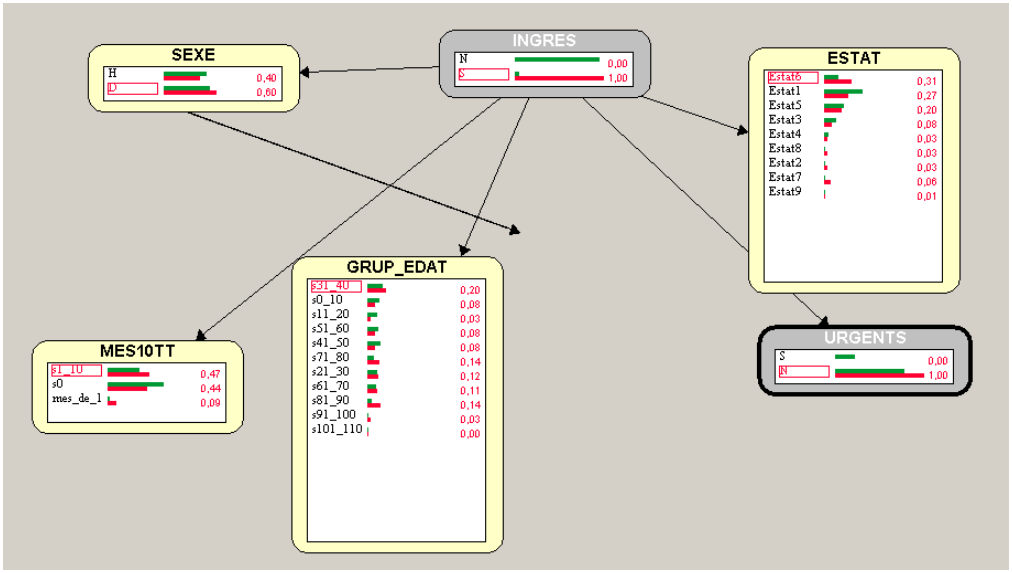


Figura 19. Xarxa Bayesiana TAN ajustada a ingrés hospitalari i absència de visita urgent a atenció primària

Si ara ho acoto als pacients en Estat 6 (més d'una malaltia crònica) ens varia el grup d'edat, passant a ser els d'entre 81 i 90 anys (27%)

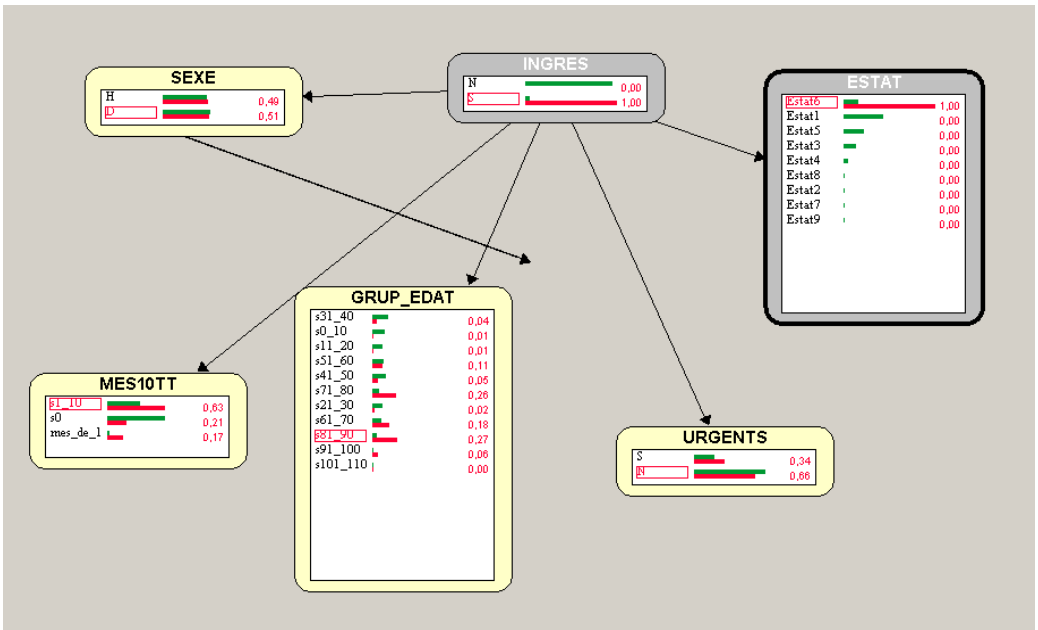


Figura 20. Xarxa Bayesiana TAN ajustada a ingrés hospitalari i E6

Curiosament si ajusto els resultats prenent la població ingressadora de sexe femení, l'estat més prevalent és l'Estat 1 (absència de malaltia o patologia aguda lleu) amb un 30%, les que estan entre 31-40 anys i les que no prenen cap medicament crònic.

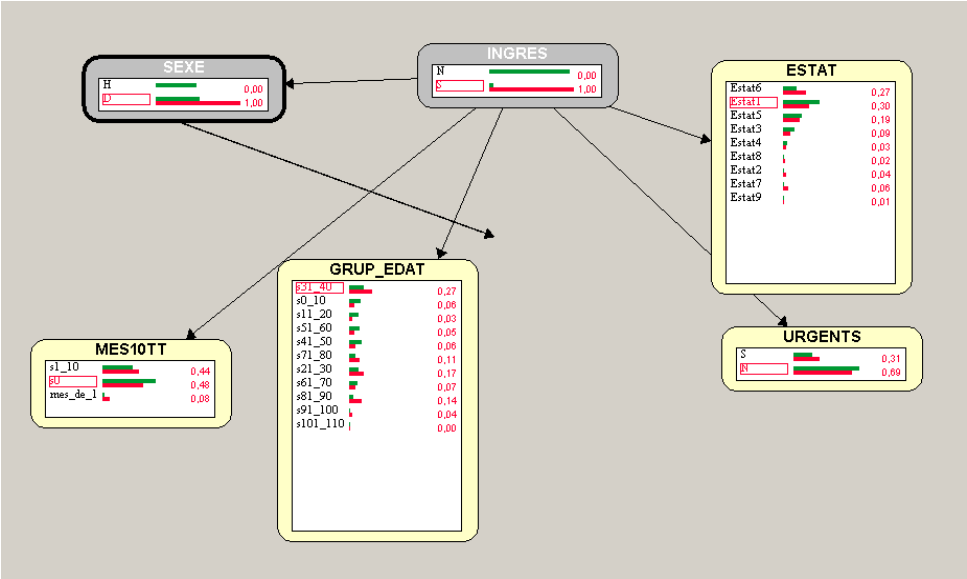


Figura 21. Xarxa Bayesiana TAN ajustada a ingrés hospitalari i Sexe femení

Si finalment ajusto per usuaris que prenen més de 10 medicaments crònics, la probabilitat en ambdos sexes s'iguala, es manté una alta probabilitat entre els que no han acudit mai a urgències i tornen a ser més prevalents els pacients en Estat 6 i entre 81 i 90 anys.

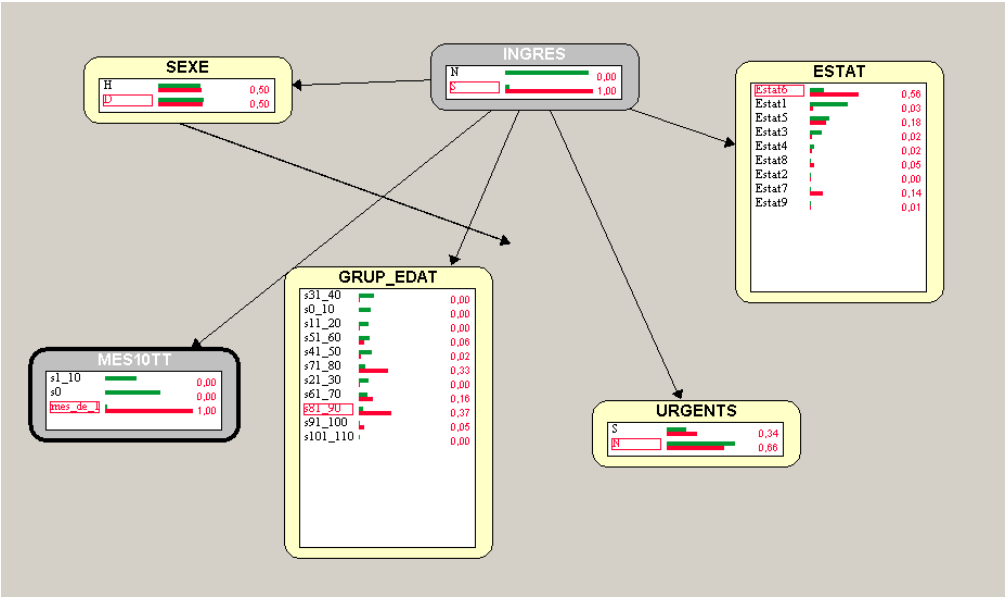


Figura 22. Xarxa Bayesiana TAN ajustada a ingrés hospitalari i més de 10 medicaments crònics

Sembla que hi ha una tendència que la major freqüència d'ingrés hospitalari s'origina entre els pacients que no han acudit a una visita prèvia a un servei d'urgències d'atenció primària i entre el sexe femení. Els altres paràmetres varien segons el sexe i el grup d'edat. Aquests resultats estan en consonància amb

el color que ens mostrava les relacions entre els nodes en l'algoritme de NB (vermell amb una relació positiva, blau en una relació inversa i violeta indeterminada)

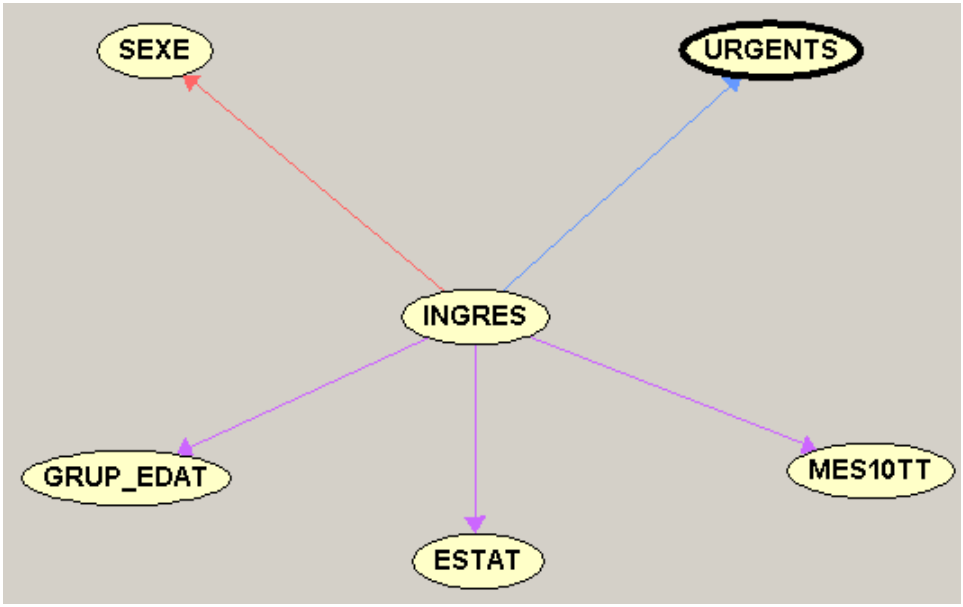


Figura 23. Relacions entre nodes en la Xarxa Bayesiana

Avaluació del classificador

Un detall important quan construïm classificadors és quantificar d'alguna manera el bons o dolents que són [39]. Per exemple, si fem servir un classificador per detecció de càncer no ens podem permetre el luxe que falli tres de cada quatre pacients. A l'hora d'avaluar un classificador haurem de tenint en compte diferents criteris, com pot ser el temps que triguem a construir-lo, la interpretabilitat del model obtingut, la senzillesa del model (com més senzill, major capacitat de abstracció) o diferències respecte al model original, però al qual més atenció se li presta és a la precisió del classificador (o, al revés, la taxa d'error) que posseeix. La precisió d'un classificador és la probabilitat amb la que classifica correctament un cas seleccionat a l'atzar [40], o també, el podem veure com el número de casos classificats correctament entre el número total d'elements.

precisió = número d'encerts / número de casos

A més de ser la mesura més acceptada per l'avaluació d'un classificador, la precisió és utilitzada en alguns procediments per guiar la construcció del classificador, per això exposarem diferents formes d'obtenir el seu valor. En el nostre cas emprarem la valoració creuada que ens ofereix Weka.

En els 4 algoritmes emprats observem una mitjana de precisió del 93,9% (93,7 – 94%).

Validació creuada de k-fulles (k-fold cross-validation):

Es pot veure com una generalització del criteri de remostratge. Fem k particions del conjunt de dades mútuament excloents i de la mateixa mida. k - 1 conjunts s'utilitzen per a construir el classificador i es valida amb el conjunt restant. Aquest pas es realitza k vegades i la estimació de la precisió del classificador s'obté com la mitjana de les k mesuraments realitzats.

En el nostre cas hem escollit la precisió amb una k = 10, que és la que per defecte ens ofereix el programari, obtenint una precisió dels algoritmes que en la seva mitjana (weighted average) es situa en el 94%, però que en el cas de ser avaluada per als valors negatius arriba al 97% però baixa al 40% per als valors positius.

Algunes vegades, com és el nostre cas, és interessant no només conèixer la precisió del classificador o la taxa d'error, sinó que és important el sentit en què s'equivoca.

Amb el nostre classificador correm el risc de considerar una persona amb alt risc d'ingrés, sense que aquesta ho sigui, mentre que som molt precisos en descartar una persona amb risc d'ingrés. Podríem clarament descartar una acció sobre el grup que difícilment acabarà ingressant en un hospital, però difícilment podrem actuar amb seguretat sobre aquells que si que acabaran ingressant, donat que la precisió de la detecció sobre aquest grup és baixa (que es coneix com un fals negatiu), potser a l'hora de solucionar l'error o intervenir preventivament ja sigui massa tard (perdem la possibilitat de tractar el pacient a temps).

Matriu de confusió

Quan distingir entre els diferents tipus d'errors és important, llavors es pot utilitzar una matriu de confusió (també anomenada taula de contingència) per mostrar els diferents tipus d'error. Si tenim un problema amb dues classes (per exemple, risc d'ingrés o no), com es pot veure a les matrius dels algoritmes estudiats, un classificador pot donar la sortida per a un nou cas: veritable positiu, prediu que el pacient té risc d'ingrés i és veritat, veritable negatiu si encerta que el pacient no té risc, fals positiu si pronostica de risc no tenint-ne i, finalment, fals negatiu, si prediu sense risc però el pacient en té.

A partir de la matriu de confusió podem construir algunes mesures que ens seran d'utilitat. La sensibilitat és la probabilitat de classificar correctament a un individu amb risc, per tant, és la capacitat del classificador per detectar la classe positiva (la que mes ens interessa, en l'estudi, que té risc), es defineix com:

Sensibilitat (S)

S= veritables positius / (veritables positius + falsos negatius)

d'altra banda, definim l'especificitat com la probabilitat de classificar correctament a un individu sense risc, en altres paraules, es pot definir la especificitat com la capacitat de classificar correctament a un individu que té realment un risc negatiu per a la prova que es fa (en el nostre exemple, que no té risc).

Es pot calcular a partir de la matriu de confusió com:

Especificitat (E)

E = veritables negatius / (veritable negatius + falsos positius)

```
=== Confusion Matrix ===

      a      b   <-- classified as
 441152   5778 |      a = N
  15993   1877 |      b = S
```

Figura 24. Matriu de confusió de NB

En els diferents algoritmes emprats obtenim una escassa sensibilitat al voltant del 10%, mentre que l'especificitat es troba en el 98% de promig.

Corbes ROC

En problemes on distingir el tipus d'error és important, es poden utilitzar les corbes ROC (del anglès, Receiver Operating Characteristics) [41]. Les corbes ROC encara tenen l'origen en la detecció de senyals de radar, s'usen habitualment en la presa de decisions mèdiques] i ens permeten avaluar de forma gràfica el funcionament d'un classificador. Són corbes en les que es presenta la sensibilitat en funció dels falsos positius (complementari de l'especificitat) per diferents resultats de un classificador.

Com més gran sigui l'àrea sota la corba ROC, millor serà el classificador. La millor predicció possible seria un mètode que passés per la cantonada superior esquerra, ja que representaria un 100% de sensibilitat (no hi ha falsos negatius) i un 100% d'especificitat (no hi ha falsos positius). Per tant, quan mes distant és la corba de la diagonal millor serà el classificador.

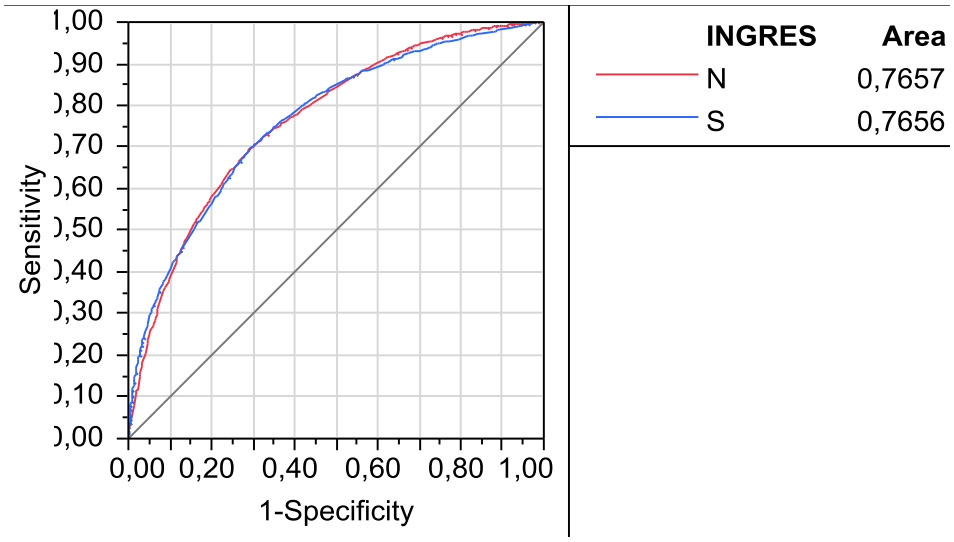
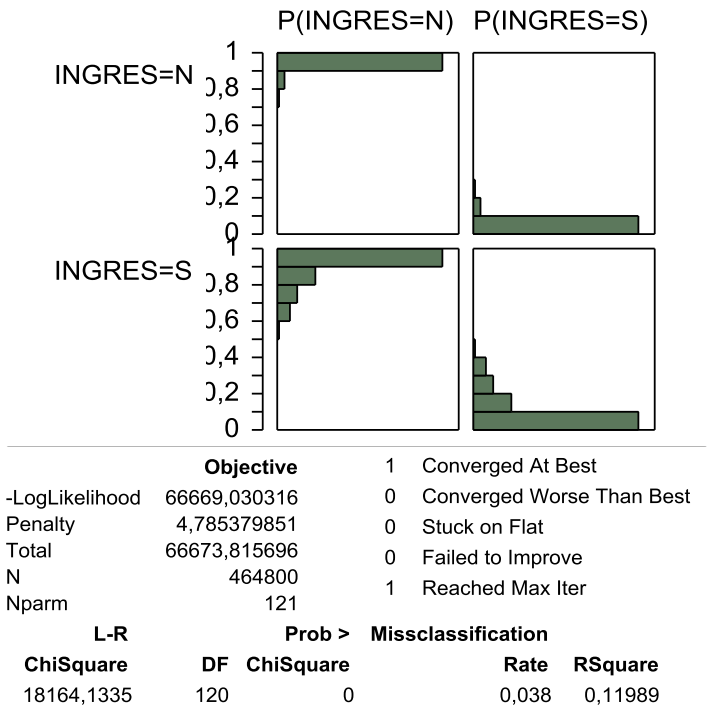


Figura 25. Corba ROC del nostre algoritme NB

Està clar que en el model de l'estudi els valors de la corba ROC són bons (76%), confirmant la precisió del model tot i la baixa sensibilitat. Si visualitzem els valors predictius en forma d'histograma podem veure les diferències entre els valors predits d'ingrés i els de no ingrés.



	Objective
-LogLikelihood	66669,030316
Penalty	4,785379851
Total	66673,815696
N	464800
Nparm	121

1	Converged At Best
0	Converged Worse Than Best
0	Stuck on Flat
0	Failed to Improve
1	Reached Max Iter

L-R ChiSquare	DF	Prob > ChiSquare	Missclassification Rate	RSquare
18164,1335	120	0	0,038	0,11989

Figura 26. Histograma del valor predictiu observat

Correlacions

Per tal de poder entendre millor el comportament dels diferents paràmetres a analitzar s'ha fet una petita aproximació a la relació existent entre els valors numèrics d'alguns paràmetres estudiats a través de la correlació existent entre ells. En aquest cas s'han estudiat l'edat, el nombre de visites urgents prèvies a l'ingrés, el nombre de tractaments crònics que pren el pacient i els ingressos hospitalàris, observant les següents correlacions

	EDAT	VISITES_URGENTS	TRACTAMENTS_CRONICS	INGRES_HTALS
EDAT	1,0000	-0,0142	0,5879	0,0965
VISITES_URGENTS	-0,0142	1,0000	0,0580	0,0712
TRACTAMENTS_CRONICS	0,5879	0,0580	1,0000	0,1699
INGRES_HTALS	0,0965	0,0712	0,1699	1,0000

The correlations are estimated by Pairwise method.

Figura 27. Correlació entre paràmetres observats

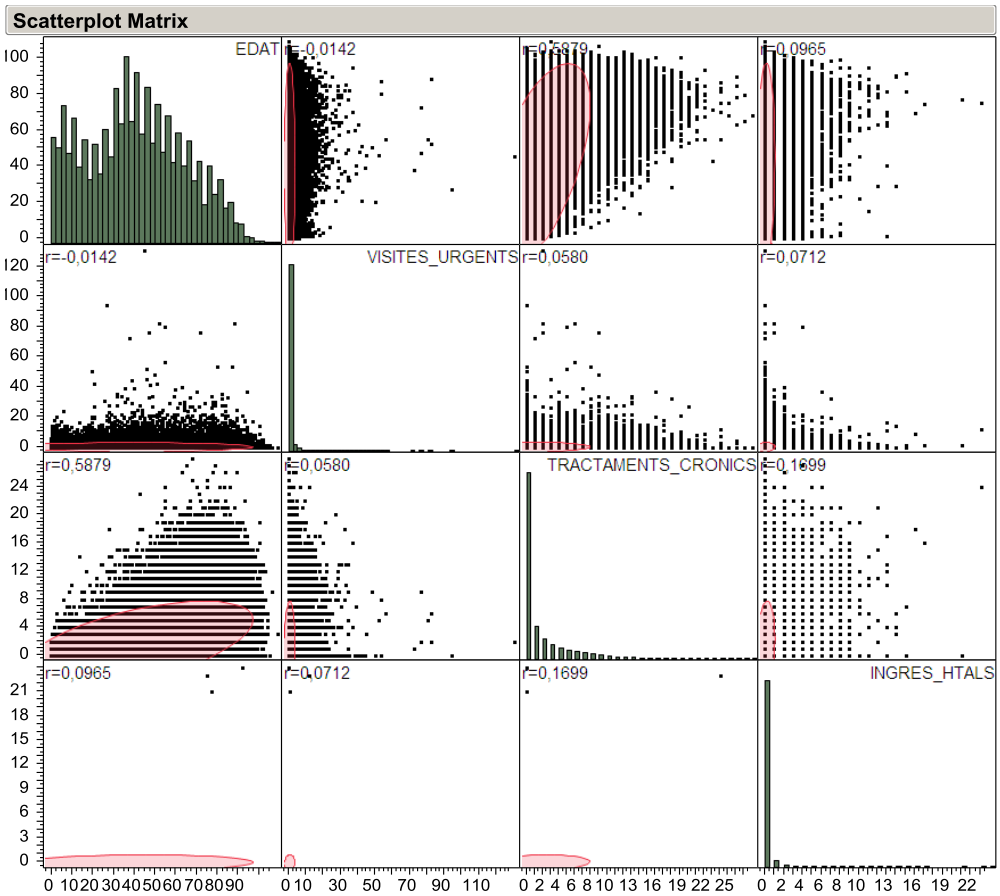
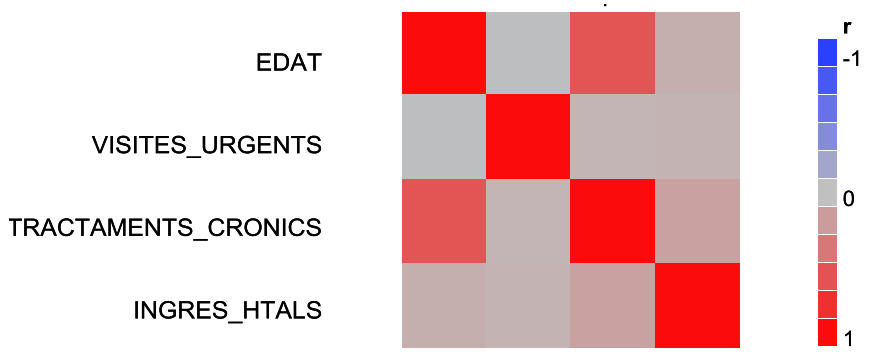


Figura 28. Scatterplot amb la correlació i les corbes de comportament

El comportament dels valors dels paràmetres respecte als ingressos hospitalaris sembla dèbilment correlacionat, però positivament correlacionat. Amb els valors observats podem deficnir el següent mapa de colors:



Caldrà valorar si aquests resultats junt amb la proposta d'altres paràmetres (consell d'experts) i la millora en la recollida de les fonts podrien millorar el model i la seva capacitat predictiva.

Propostes de millora del model (Next steps)

Com s'ha pogut veure en els resultats observats el model obtingut, malgrat tenir una alta sensibilitat i l'alta precsisió dels algoritmes, pateix d'una important manca d'especificitat i valor predictiu del risc que cercavem. La utilitat pràctica del model reclama que ens pugui servir per detectar aquells pacients amb risc, mes que descartar els que no en tenen. Cal dons trobar atributs que ens permetin arribar a la especificitat i detecció del risc dessitjada.

Darrerament s'han creat indicadors que ens podrien ajudar a modelar aquest risc i compten amb la opinió favorable dels experts en el tema. Estem parlant d'indicadors que no depenen directament de les dades clíniques dels usuaris, sinó dels nivell o qualitat de l'assitencia que reben i de factors que tradueixen el seu benestar social. En concret estem parlant de dos indicadors:

Indicadors de qualitat assistencial:

Aquests són indicadors sintètics que es composen de diferents indicadors que s'han escollit per un grup d'experts per avaluar la qualitat assistencial que ofereix un professional assistencial als seus usuaris. Aquest indicadors s'utilitza, en l'àmbit de l'atenció primària, per associar un incentiu econòmic al professional, però reflexa les politiques sanitaries que la salut pública exigeix per garantir el bon control dels pacients a la primària.

Un exemple entenidors seria per exemple el control de la TA en uns valors contemplats com a normal en tots aquells pacients que pateixen una malaltia cardiovascular crònica. Altres tractarien sobre el control de les glucèmies capil.lars en usuaris diabètics i així continuariem fins arribar als mes de 40 indicadors que el composen.

Resulta obvi deduir que aquells pacients que tinguin un professional assignat (que els controla) amb millors valors d'aquest indicador sistètic de qualitat assistencial, es puguin veure també beneficiats per un menor nombre de ingressos hospitalaris per complicacions de les seves patolohgies de base. Sembla, per tant, aquest atribut, un bon candidat a participar en la creació del model de risc d'ingrés hospitalari i per tan, podria ser un nou atribut a valorar en un posterior modelatge.

Indicadors socioeconòmics:

Aquests indicadors també han estat recentment introduïts com a un indicadors que actuen com una variable explicativa o d'ajust en l'avaluació d'indicadors clínics i de gestió, així com en estudis epidemiològics. Es tracta d'indicadors compostos que consideren com a unitat d'anàlisi la secció censal dels usuaris. Es calculen a partir d'altres indicadors desenvolupats a partir del cens del 2001 i poden contemplar la situació laboral i nivell d'instrucció dels usuaris majors de 15 anys.

Diversos estudis han demostrat que hi ha una relació entre la prevalença de certes malalties cròniques i els indicadors socioeconòmics de la població (benestar social), per el que també s'ha proposat com un possible atribut a contemplar a l'hora de predir el risc d'un usuari d'acabar ingressat en un hospital donat el seu major risc de patir malalties.

Hi ha altres factors que seria interessant analitzar com la distància que existeix entre els centres hospitalaris i el lloc de residència de l'usuari, que podrien influenciar l'accessibilitat que aquest tenen a l'hospital i per tan, el risc d'acabr ingressat en el mateix, però aquesdta i altres dades no estan contrastades i no són fàcils d'obtenir i potser requieririen d'un treball previ abans de poder-se contemplar en el modelatge.

Índex de qualitat de prevenció

Aquests indicadors s'han començat a utilitzar recentment i són indicadors de qualitat de la prevenció compostos d'una bateria de mesures calculades amb les dades notificades al informe de l'alta hospitalària, que han estat codificades mitjançant la Classificació Internacional de Malalties 9a revisió Modificació Clínica (CIM·9·MC), 7a edició, i recollides en el registre del Conjunt mínim bàsic de dades dels hospitals d'aguts (CMBD-HA). L'objectiu dels mateixos és identificar les hospitalitzacions potencialment evitables Ambulatory Care Sensitive Conditions -ACSC-. Les ACSC són patologies en les que una bona atenció ambulatòria pot, potencialment, prevenir la necessitat d'hospitalització, o en les que una ràpida intervenció pot prevenir complicacions o un empitjorament de la condició clínica del pacient.

Malgrat que aquests indicadors estan elaborats a partir de la informació recollida sobre els ingressos en els hospitals d'aguts, proporcionen informació sobre la qualitat del sistema sanitari en la seva globalitat, i especialment sobre la capacitat de l'assistència ambulatòria per prevenir complicacions mèdiques.

Aquests s'haurien d'analitzar conjuntament els diferents atributs proposats ja que es poden oferir "sinergies" en la interpretació de les dades.

Estan basats en els PQI (Preventive quality indicators) de l'AHRQ (Agency for Healthcare Research and Quality) [45] adaptats a les característiques del nostre entorn.

APLICACIÓ DEL MODEL (Deployment)

S'ha construït un model predictiu utilitzant un conjunt limitat de les variables que es van generar a partir de història clínica electrònica dels pacients. El model estima el risc d'ingrés a un hospital del Sistema Nacional de Salut de la zona d'influència d'un territori dins dels 12 mesos anteriors a la notificació de l'alta hospitalària. S'han seleccionat les variables que segons la opinió dels metges experts, es troben disponible a la història clínica dels pacients i es poden relacionar amb un ingrés hospitalari.

La capacitat predictiva d'encert del model és baixa, amb un índex kappa que oscil.la al voltant del 0,0035 en els diferents algoritmes emprats. El VPP és del 24,5%, arribant a un VPN del 96,5% i un àrea sota la corba ROC ('c-statistic') de 0,76. L'especificitat d'aquest model (98,7%) és alt, tot i que la sensibilitat de la model és bastant baixa amb només 10,5% de tots els pacients. El rendiment del model es podria millorar mitjançant la inclusió de més variables però això podria fer disminuir la utilitat pràctica del model. També cal destacar que els valors de precisió de la prova són, de mitjana, alts, amb un 97,3%, però en el cas dels valors negatius arriben fins al 24,5%. El coneixement del percentatge de risc dels pacients segons la puntuació de risc d'ingrés pot ser molt útil en la distribució de recursos i d'esforços, implicant que segons aquest indicador de risc podem distribuir de forma més justa i efectiva recursos i activitats preventives a fi de reduir l'ingrés final dels pacients amb alt risc. En el grup de més alt risc d'ingrés hospitalari el conformen els pacients predominantment de sexe femení, que no han acudit a visitar-se prèviament a urgències d'atenció primària i que depenent del sexe (dona o home) es troben, respectivament, entre els 31-40 anys o entre els 81-90 i tenen un Estat derivat del seu CRG, també respectivament, E1 (saludables o patologia aguda) i E6 (dos malalties còniques). El grup de tractament crònic, tot i que de forma dissociada s'associa a l'ingrés (bona correlació), quan és major de 10 medicaments crònics, en associar-lo diferents edats i sexes, varia substancialment.

Està clar que el nivell i tipus dels recursos assignats a pacients amb alt risc han de ser diferents dels assignats als pacients en els nivells de baix risc, amb el que el fet de tenir un model de gran especificitat i VPN ens pot ajudar a planificar la destinació de recursos i esforços de forma diferenciada. Els nivells i tipus d'intervenció d'aquests pacients ha de variar per banda risc i característiques dels pacients, però els metges i els gestors dels recursos poden utilitzar aquestes dades per seleccionar els llindars per a qualsevol intervenció preventiva. El model té les seves limitacions[42, 43]. Es va desenvolupar utilitzant les dades extretes de les històries clíniques i pot servir per complementar d'altres dades de gestió que ens ajudin en la planificació de l'atenció dels pacients des de l'atenció primària.

Cal dir que és possible que existeixin errors d'algunes dates dels ingressos, donat que les fonts d'origen de les dades són diferents i els ingressos fan referència al total efectuat durant un període de 12 mesos. El fet de garantir que les dades de tractament i visites urgents s'han donat exactament en un període previ i no solapat en el temps és difícil de controlar. D'altre banda, el nombre d'ingressos hospitalaris és una dada d'obtenció indirecta, que procedeix d'una font hospitalària. El fet de procedir de dues fonts prèvies pot dificultar la plena precisió de les dades. A més, hem de recordar que sovint i especialment en les dades extremes de la població (nou nats i avis) l'índex emprat per associar les dades, no és 100% fiable, per l'existència de codis provisionals i per la utilització d'indexacions pròpies, com els números d'història clínica (NHC) per part dels hospitals. No obstant això, les diferències en la indexació i recuperació en les dades dels pacients des del seu origen hospitalari són suficientment fiables i no posen en dubte la validesa del model.

La capacitat per identificar pacients amb alt risc d'ingrés constitueix el primer pas en qualsevol estratègia per millorar l'atenció i serveis als pacients susceptibles. L'objectiu final, però, és acoblar aquest procés de "recerca de casos" amb intervencions cost-efectives que mitiguin el risc d'ingrés, i idealment, poder concentrar esforços i recursos en finançar aquells amb major risc. En una revisió sistemàtica recent, Hansen et al [43] identifiquen una àmplia gamma d'estratègies que s'han emprat per estalviar els re-ingressos i que també podrien ser efectives en els ingressos un cop detectats els pacients de risc, que inclouen intervencions prealta (descàrrega millorada la planificació, l'educació del pacient, la reconciliació de medicaments, després de l'alta la següent cita, etc.), intervencions posteriors a l'alta del pacient (línies directes, telèfon recordatoris de cites, visites a domicili, etc.) i altres intervencions per salvar la transició de l'hospital la llar, com ara entrenament per la infermera. Hi ha encara pocs estudis, alguns són petits i no estan ben dissenyades, però la majoria són anglosaxons i provenen de dades hospitalàries, per el que cal impulsar la realització de mes estudis des de l'atenció primària i des del nostre territori. Està clar que les intervencions preventives estan més a l'abast de l'atenció primària i cal detectar aquest risc abans de que el pacient arribi a necessitar realment un ingrés hospitalari. Cinc de 16 assaigs controlats aleatoris documentats amb significança estadísticament unes reduccions en el risc absolut de readmissió, però cap intervenció o conjunt d'estratègies ha resultat encara exitosa en la reducció de riscos.

Si bé la planificació de l'alta aporta certes millores, com també ho fan la organització de visites de seguiment i telefòniques posteriors a l'alta, també intervencions com entrenament per la infermera i visites a domicili en els pacients amb criteris de risc poden arribar a ser molt efectives en la prevenció dels casos, arribant a estalviar el que resulta molt més costos, tan en salut com a nivell econòmic, que és l'ingrés final del pacient a l'hospital. Aquestes dades permeten l'orientació la realització d'intervencions dirigides, limitades a pacients amb més risc.

Als hospitals d'Anglaterra es comencen a oferir incentius financers que s'inclouen en el sistema marc operatiu 2011-2012 [44] sobre aquests camps en la prevenció dels ingressos i re-ingressos hospitalaris i no trigarem a veure que es prenen mesures en el nostre entorn en el marc econòmic que se'ns presenta, per el que és important recopilar evidència sobre quines intervencions són eficaces, en quins pacients i a quin cost. Les àrees per a la investigació futura podria incloure determinar quines són les intervencions necessàries i la seva major eficàcia d'acord amb el nivell de risc subjacent.

Pot ser que els pacients de menor o moderada risc d'ingrés tinguin condicions o circumstàncies on la intervenció és més probable que tingui èxit perquè pacients d'alt risc. Igualment, pot haver certs subgrups dels pacients dins d'una franja de risc especial que són més o menys susceptibles a l'atenció preventiva. L'ús de models predictius com la detecció de casos, els instruments per seleccionar la prevenció o les millors intervencions van guanyat acceptació entre la comunitat mèdica i de gestors sanitaris. Cal considerar la forma en que aquestes eines es poden utilitzar en l'entorn de l'atenció primària i hospitalària per tal de garantir la salut i supervivència del sistema sanitari i la dels propis pacients.

Propostes d'aplicació i monitoratge de l'aplicació del model proposat

Un cop establerta la validesa del model caldria habilitar-lo com a eina d'ajuda per detectar i intervenir sobre la població. Tractant-se d'un indicador que detecta majoritàriament la població amb menys risc d'ingressar caldria complementar-se amb altres indicadors, però es podria afegir com un marcador a

nivell de les històries dels pacients per tal que poguessin identificar-se els subjectes amb l'indicador de baix risc.

Aquest indicador, com s'ha vist, varia segons altres atributs com l'estat dels grups de risc o les assistències prèvies a urgències d'atenció primària, per el que es podrien complementar amb aquesta informació. La visualització conjunta a nivell de la història del pacient podria facilitar el nivell d'intervenció que els professionals sanitaris, especialment en àrees de medicina preventiva, poguessin fer.

Passat un període prudent d'aplicació del model en forma d'etiqueta de risc i havent donat temps per intervenir sobre els subjectes diana, caldria re-avaluar els resultats per confirmar que ha hagut una variació passat aquest temps d'intervenció i veure com pot haver variat els resultats del model aquesta intervenció efectuada. El període podria ser de no menys de 12 mesos per tal de poder avaluar un període similar al de l'estudi.

És obvi per els resultats obtinguts i la validesa del model proposat que caldria afegir nous atributs al model, per el que a mes d'assajar la inclusió dels atributs comentats en l'apartat anterior, podria ser molt interessant la recerca, segons els resultats obtinguts, de nous factors que ens puguin haver passat desapercebuts. Sempre resulta molt útil incloure la opinió al respecte d'un dels actors del procés estudiat. Fonamentalment ens referim als professionals sanitaris que un cop emprat el model de risc poden aportar informació molt valuosa sobre factors intercurrents que ens poden enriquir un posterior estudi amb nous atributs proposats per els professionals.

GLOSARI

Atribut: Es un tipus bàsic d'informació descriptiva d'una dimensió.

CIM10: És la 10 ª revisió de la Classificació Estadística Internacional de Malalties i Problemes Relacionats amb la Salut (CIM), una llista de classificació mèdica per l'Organització Mundial de la Salut (OMS). Codifica per malalties, signes i símptomes, troballes anormals, denúncies, circumstàncies socials i causes externes de lesions o malalties. [1] El conjunt de codis permet que més de 14.400 codis diferents i permet el seguiment dels molts nous diagnòstics.

Data Warehouse: Repositori central amb la informació més valuosa de l'empresa, on s'emmagatzemen les dades estratègiques, tècniques i operatives. Les dades emmagatzemades aquí han passat un procés de qualitat que assegura la seva consistència. A més el repositori esta construït per a que l'accés sigui lo mes ràpid possible. La seva construcció es fa per etapes que normalment es corresponen a les principals àrees operatives de l'empresa.

Data Mining: Procés que ajuda a descobrir els patrons i relacions que poden passar desapercebuts en l'anàlisi del negoci i els clients. Ha d'estar orientar a resoldre un problema de negocis i no ha de necessitar el ser un especialista en la matèria per a poder usar-lo.

Dimensió: Una estructura que representa una de las cares d'un cub.

ETL: Procés d'extracció, transformació i càrrega (Extraction, Transformation, and Loading).

Finançament capitatiu: Per permetre el control de la despesa sanitària, es rep un pressupost per proporcionar atenció sanitària a la població definida. Aquests pressupostos es estableixen la base de la capitació, entesa aquesta com la quantitat de finançament sanitària que s'assigna perquè una persona rebi l'atenció sanitària especificada durant un període de temps determinat.

Gestió del coneixement: procés continu d'adquisició, distribució i anàlisi de la informació que es mou en l'entorn de l'organització per a fer més intel·ligent als seus treballadors (que rendeixin més), i ser més precisos en la presa de decisions, donar una resposta més ràpida a les necessitats del mercat i ser més competitiu en aquest entorn tan canviant.

GPL o GNU: La Llicència pública general de GNU o més coneguda pel seu nom en anglès GPL és la llicència que s'utilitza àmpliament al món del programari i garanteix als usuaris finals (persones, organitzacions , companyies) la llibertat d'utilitzar, estudiar, compartir (copiar) i modificar el programari.

Knoweldge Discovery in Databases (KDD): és el procés complet d'extracció d'informació, que s'encarrega a més de la preparació de les dades i la interpretació dels resultats obtinguts.

Model Predictiu: És el procés pel qual es crea un model o és escollit per intentar predir millor la probabilitat d'un resultat. En molts casos, el model es tria en base a la teoria de detecció per intentar endevinar la probabilitat d'un resultat donat una quantitat fixa de dades d'entrada, per exemple, donat

un correu electrònic per determinar la probabilitat que sigui correu no desitjat. Els models poden utilitzar un o més classificadors per intentar determinar la probabilitat d'un conjunt de dades que pertanyen a un altre grup, per exemple el correu brossa o "pernil".

MySQL Workbench: Es tracta d'una eina de disseny visual de base de dades SQL que integra el desenvolupament, administració, disseny de bases de dades, creació i manteniment en un únic entorn de desenvolupament integrat per al sistema de base de dades MySQL. És el successor de DBDesigner 4 de fabFORCE.net, i substitueix l'anterior paquet de programari, MySQL GUI Bundle eines.

Naïve Bayes: classificador probabilístic basat en l'aplicació teorema de Bayes basat en supòsits d'independència (ingenu).

Oracle: Sistema de gestió de bases de dades relacional, desenvolupat por Oracle Corporation.

PDG: (Provabilistic Decision Graft). Algorismes d'aprenentatge supervisats de grafs de decisió probabilística.

PL/SQL: Llenguatge de programació que suporta totes les consultes i manipulació de dades que s'utilitzen en SQL, però que inclou característiques com el maneig de variables, les estructures modulars , les de control de fluix y el control d'excepcions.

SQL*Loader: Utilitat d'importació de dades que possibilita la carrega automàtica de dades externes (residents en fitxers) en taules de las bases de dades. La informació pot carregar-se en una o vàries taules prèviament creades i que poden contenir dades prèvies. Les noves dades podran substituir a les que ja existien o be afegir-se com a nous registres.

Xarxa bayesiana: És un model probabilístic gràfic (un tipus de model estadístic) que representa un conjunt de variables aleatòries i les seves dependències condicionals a través d'un gràfic acíclic dirigit (DAG).

BIBLIOGRAFIA

[1] E. Castillo, J. M. Gutiérrez y A. S. Hadi. "Sistemas expertos y modelos de redes probabilísticas". Academia de Ingenieríia (1997).

[2] P. Cheeseman. In defense of probability. En "Proceedings of the 9th International Joint Conference on Artificial Intelligence (IJCAI-85)", pags. 1002–1009. Morgan Kaufmann Publishers (1985).

[3] S. L. Lauritzen y D. J. Spiegelhalter. Local computations with probabilities on graphical structures and their application to expert systems (with discussion). Journal of the Royal Statistical Society,
Series B 50, 157–224 (1988).

[4] Basilio Serra; Araujo.: Aprendizaje Automático: conceptos básicos y avanzados. Aspectos prácticos utilizando el software Weka; Prentice Hall (2006)

[5] Luque Malagón; Constantino.: Clasificadores Bayesianos. El algoritmo de Naives Bayes. (2003)

[6] Villanueva Velasco; David.: Redes Bayesianas. Inteligencia Artificial II. (2007)

[7] Weka 3. Data Mining Software in Weka, http://www.cs.waikato.ac.nz/ml/weka/

[8] Hughes JS, Averill RF, Eisenhandler J, Goldfield NI, Muldoon J, Neff JM, et al. Clinical Risk Groups (CRGs): a classification system for risk-adjusted capitation-based payment and health care management. Medical Care 2004;42(1):81–90.

[9] Audain, C. (2007). Florence Nightingale.
Online:http://www.scottlan.edu/lriddle/women/nitegale.htm. Accessed 30 July 2009.
Ayres, I (2008). Super Crunchers. New York: Bantam Books.

[10] Bailey-Kellog, C. Ramakrishnan, N. And Marathe, M. Spatial Data Mining to Support Pandemic Preparedness. SIGKDD Explorations (8) 1, 80-82.

[11] Cheng, T.H., Wei, C.P., Tseng, V.S. (2006) Feature Selection for Medical Data Mining: Comparisons of Expert Judgment and Automatic Approaches. Proceedings of the 19th IEEE Symposium on Computer-Based Medical Systems (CBMS'06).

[12] Kou, Y., Lu, C.-T., Sirwongwattana, S., and Huang, Y.-P. (2004). Survey of fraud detection techniques. In Networking, Sensing and Control, 2004 IEEE International Conference on Networking, Sensing and Control. (2) 749-754.

[13] Nightingale, F (1858). Notes on Matters Affecting the Health, Efficiency and Hospital Administration of the British Army.

[14] Shillabeer, A (29 July 2009). Lecture on Data Mining in the Health Care Industry. Carnegie Mellon University Australia.

[15] Thangavel, K., Jaganathan, P.P. and Easmi, P.O. Data Mining Approach to Cervical Cancer Patients Analysis Using Clustering Technique. Asian Journal of Information Technology (5) 4, 413-417.

[16] Tufte, E. (1997). Visual Explanations. Images and Quantities, Evidence and Narrative. Connecticut: Graphics Press.

[17] Wong, W.K., Moore, A., Cooper, G. And Wagner, M (2005). What's Strange About Recent Events (WSARE): An Algorithm for the Early Detection of Disease Outbreaks. Journal of Machine Learning Research. 6, 1961- 1998.

[18] Witten, I. H. and Frank, E. (2005). Data mining : practical machine learning tools and techniques. Morgan Kaufmann series in data management systems. Morgan Kaufman.

[19] Basilio Serra; Araujo.: Aprendizaje Automático: conceptos básicos y avanzados. Aspectos prácticos utilizando el software Weka; Prentice Hall (2006)

[20] Luque Malagón; Constantino.: Clasificadores Bayesianos. El algoritmo de Naives Bayes. (2003)

[21] Villanueva Velasco; David.: Redes Bayesianas. Inteligencia Artificial II. (2007)

[22] Weka 3. Data Mining Software in Weka, http://www.cs.waikato.ac.nz/ml/weka/

[23] J. Hernández. and C. Ferri. "Practica de Minería de Datos, Introducción al WEKA, Curso de Doctorado Extracción Automática de Conocimiento en Bases de Datos e Ingeniería de Software", Universidad de Valencia, 2006, pp. 2-15
[24] E. Frank. "Machine Learning with WEKA". Department of computer science; University of Waikato, New Zeland. pp.1

[25] "Programa de Doctorado Tecnologías Industriales. Aplicaciones de la inteligencia robótica. Practica 1: Entorno de WEKA de aprendizaje automático y data mining", pp. 6-9

[26] Cruz-Ramírez N, Acosta-Mesa HG, Barrientos- Martínez RE and Nava-Fernández LA. How Good are Bayesian Information Criterion and the Minimun Description Lenght Principle for Model Selection? A Bayesian Networks Analysis. Advances in Artificial Intelligence. Vol. 4293, 2006; 494-504.

[27] Pérez A, Larrañaga P and Inza I. Modelos gráficos probabilísticos para la clasificación supervisada empleando la estimación basada en kernels Gaussianos esféricos. III Taller Nacional de Minería de Datos y Aprendizaje. 2005; 125-134.

[28] Cruz-Ramírez N, Acosta-Mesa HG, Carrillo-Calvet H, Nava-Fernández LA and Barrientos-Martínez RE. Diagnosis of Breast Cancer using Bayesian Networks: A case study. Computers in Biology and Medicine. Vol. 37, 2007; 1553-1564.

[29] Rocío Erandi Barrientos Martínez, Nicandro Cruz Ramírez, Héctor Gabriel Acosta Mesa, Ivonne Rabatte Suárez. Evaluation of the Potential of Bayesian Networks on the Classification of Medical Data. Revista Médica de la Universidad Veracruzana / Vol. 8 núm. 1, Enero - Junio 2008.

[30] J. Chiang. "Agreement between categorical measurements: Kappa Statistics"
http://www.dmi.columbia.edu/homepages/chuangj/kappa/

[31] "Mean Absolute Error (MAE) and Root Mean Squared Error (RMSE)"
http://www.eumetcal.org.uk/eumetcal/verification/www/english/msg/ver_cont_var/uos3/uos3_ko1.htm

[32] gepsot. "Analyzing APS Models Statistically. Root mean squared error"
http://www.gepsoft.com/Gepsoft/APS3KB/Chapter09/Section3/SS04.htm

[33] gepsot. “Analyzing APS Models Statistically. Relative Absolute Error”. http://www.gepsoft.com/gxpt4kb/Chapter10/Section1/SS08.htm

[34] gepsot. “Analyzing APS Models Statistically. Relative Absolute Error”. http://www.gepsoft.com/

[35] J. Pearl. “Probabilistic reasoning in intelligent systems: Networks of plausible inference”. Morgan Kaufmann Publishers Inc. (1988).

[36] P. P. Shenoy y G. Shafer. Axioms for probability and belieffunction propagation. En “Uncertainty in Artificial Intelligence”, p´ags. 169–198. Morgan Kaufmann (1990).

[37] P. P. Shenoy y G. Shafer. Probability propagation. En “Annals of Mathematics and Artificial Intelligence”, vol. 2, p´ags. 327–351 (1990).

[38] S. M. Weiss y C. A. Kulikowski. “Computer systems that learn: Classification and prediction. Methods from statistics, neural nets, Machine learning and expert systems”, cap. Chapter 2: How to estimate the True Performance of a Learning System, p´ags. 17–49. Morgan Kaufmann Publishers Inc. (1991).

[39] R. Kohavi. “Wrappers for performance enhancement and oblivious decision graphs”. Tesis Doctoral, Stanford University, Stanford, CA, USA (1996).

[40] T. Fawcett. ROC graphs: Notes and practical considerations for researchers. Informe técnico HPL-2003-4, HP Labs, Palo Alto, USA (2004).

[41] Bellón JA, Lardelli P, de Dios LJ, et al. Validity of self reported utilisation of primary health care services in an urban population in Spain. J Epidemiol Community Health 2000;54:544–51.

[42] Cleary PD, Jette AM. The validity of self-reported physician utilization measures. Med Care 1984;22:796–803.

[43] Hansen LO, Young RS, Hinami K, et al. Interventions to reduce 30-day rehospitalisation: a systematic review. Ann Intern Med 2011;155:520–8.

[44] Billings J, Blunt I, Steventon A, et al. BMJ Open 2012;00:e001667. doi:10.1136/bmjopen-2012-001667.

[45] Prevention Quality Indicators Overview. AHRQ Quality Indicators. July 2004. Agency for Healthcare Research and Quality, Rockville, MD. http://qualityindicators.ahrq.gov/pqi_overview.htm

ANNEXES

Planificació de tasques

La següent taula mostra la distribució de tasques i la planificació del treball en fases temporals, per tal de marcar el calendari del projecte. Aquesta planificació resulta del compendi de les fases descrites en la metodologia de treball de mineria de dades escollida (CRISP), combinades amb les fites i dates claus de l'assignatura.

ID	Nom de la tasca	Durada	Inici	Fi	Pred.
1	**1. Inici del projecte**	**1 dia**	**18/09/2012 8:00**	**18/09/2012 17:00**	
2	1.1 Descarrega del material i lectura del Pla Docent	1 dia	18/09/2012 8:00	18/09/2012 17:00	
3	**2. Elaboració PAC 1. Pla de Treball i Anàlisi preliminar**	**13 dies**	**18/09/2012 8:00**	**04/10/2012 17:00**	
4	2.1 Descarrega i lectura de material de l'assignatura	1 dia	18/09/2012 8:00	18/09/2012 17:00	
5	2.2 Elaboració del Pla de Treball	2 dies	19/09/2012 8:00	20/09/2012 17:00	4
6	2.3 Cerca bibliogràfica	1 dia	21/09/2012 8:00	21/09/2012 17:00	5
7	2.4 Anàlisi dels requeriments del problema a abordar	3 dies	24/09/2012 8:00	26/09/2012 17:00	6
8	2.5 Elaboració de l'esborrany de la PAC1	2 dies	27/09/2012 8:00	28/09/2012 17:00	7
9	2.6 Preparació i cerca dels software necessari	1 dia	01/10/2012 8:00	01/10/2012 17:00	8
10	2.7 Elaborar PAC1	2 dies	02/10/2012 8:00	03/10/2012 17:00	9
11	2.8 Entregar PAC1	1 dia	04/10/2012 8:00	04/10/2012 17:00	10
12	**3. Realització PAC 2. Estat de l'art**	**20 dies**	**05/10/2012 8:00**	**01/11/2012 17:00**	**3**
13	3.1 Introducció a la Mineria de Dades	10 dies	05/10/2012 8:00	18/10/2012 17:00	
14	3.1.1 Tecnologies de transformació de dades	2 dies	05/10/2012 8:00	08/10/2012 17:00	
15	3.1.2 Eines de cerca de coneixement	2 dies	09/10/2012 8:00	10/10/2012 17:00	14
16	3.1.3 CRIPS. Metodologia del projecte	2 dies	11/10/2012 8:00	12/10/2012 17:00	15
17	3.1.4 Preparació de les dades	2 dies	15/10/2012 8:00	16/10/2012 17:00	16
18	3.1.5 Algoritmes i eines per l'estudi dels resultats	2 dies	17/10/2012 8:00	18/10/2012 17:00	17
19	3.2 Eines de gestió del coneixement emprades al projecte	6 dies	19/10/2012 8:00	26/10/2012 17:00	13
20	3.2.1 Mineria de dades aplicada a la Salut	3 dies	19/10/2012 8:00	23/10/2012 17:00	
21	3.2.2 Models i Projectes. Xarxes bayesianes	3 dies	24/10/2012 8:00	26/10/2012 17:00	20
22	3.3 Elaboració de la PAC2	3 dies	29/10/2012 8:00	31/10/2012 17:00	19
23	3.4 Lliurament PAC2	1 dia	01/11/2012 8:00	01/11/2012 17:00	22
24	**4. Realització PAC 3. Implementació**	**25 dies**	**02/11/2012 8:00**	**30/11/2012 9:00**	**12**
25	4.1 Plantejament problema	2 dies	02/11/2012 8:00	03/11/2012 17:00	
26	4.1.1 Lectura i cerca bibliogràfica	2 dies	02/11/2012 8:00	03/11/2012 17:00	
27	4.2 Preparació de dades	7 dies	04/11/2012 9:00	11/11/2012 9:00	25

28	4.2.1 Configuració de la base de dades	4 dies	04/11/2012 9:00	08/11/2012 9:00	
29	4.2.2 Càrrega de les dades	2 dies	08/11/2012 9:00	10/11/2012 9:00	28
30	4.2.3 Proves amb les dades	1 dia	10/11/2012 9:00	11/11/2012 9:00	29
31	4.3 Creació de resultats	16 dies	11/11/2012 9:00	30/11/2012 9:00	27
32	4.3.1 Generar informes i taules	5 dies	11/11/2012 9:00	16/11/2012 9:00	
33	4.3.2 Interpretar resultats	5 dies	16/11/2012 9:00	22/11/2012 9:00	32
34	4.3.3 Crear esborrany PAC3	3 dies	22/11/2012 9:00	27/11/2012 9:00	33
35	4.3.4 Crear PAC3	2 dies	27/11/2012 9:00	29/11/2012 9:00	34
36	4.3.5 Lliurar PAC3	1 dia?	29/11/2012 9:00	30/11/2012 9:00	35
37	**5. Lliurament Final**	**26 dies**	**01/12/2012 8:00**	**04/01/2013 17:00**	**24**
38	5.1 Preparació memòria	19 dies	01/12/2012 8:00	26/12/2012 17:00	
39	5.1.1 Revisió documentació final	4 dies	01/12/2012 8:00	05/12/2012 17:00	
40	5.1.2 Conclusions Finals	5 dies	06/12/2012 8:00	12/12/2012 17:00	39
41	5.1.3 Elaboració de la memòria	10 dies	13/12/2012 8:00	26/12/2012 17:00	40
42	5.2 Preparació Presentació	4 dies	27/12/2012 8:00	01/01/2013 17:00	38
43	5.3 Revisió final i correccions	2 dies	02/01/2013 8:00	03/01/2013 17:00	42
44	5.4 Lliurament Final	1 dia	04/01/2013 8:00	04/01/2013 17:00	43
45	**6. Debat Virtual**	**4 dies**	**14/01/2013 9:00**	**18/01/2013 17:00**	**37**
46	6.1 Intervenció en el debat	4 dies	14/01/2013 9:00	18/01/2013 17:00	

Programari Lliure per l'anàlisi de dades i Knowledge-Discovery (KDD)

Hi ha un gran nombre d'eines de programari lliure relacionades amb tot tipus d'anàlisi de dades. Des de programes amb complexes aplicacions per a l'estadística clàssica, fins a eines de programari d'objectius molt específics (per a un conjunt concret de mètodes de mineria de dades o algoritmes, com la classificació o clustering i d'altres tècniques KDD). Podem trobar una exhaustiva llista d'eines de programari, de codi obert o no, que contemplen des d'àmplies suites fins a aplicacions concretes per a la mineria de dades, l'estadística clàssica, l'emmagatzematge de dades o solucions d'intel·ligència de negocis.

La mineria de dades és un camp ampli i emergent en el que s'han fet grans treballs d'investigació i es segueixen fent avui en dia. És una de les àrees de més ràpid creixement en informàtica i ofereix una gran quantitat d'aplicacions, així com eines de gran abast per analitzar les grans bases de dades utilitzades en diferents dominis específics, com en els negocis, la ciència o la indústria.

La mineria de dades, també coneguda com Knowledge Discovery Dates (KDD) o "descobriment de coneixement en les bases de dades", és una branca del camp de l'anàlisi de les dades més genèrica, que te com a objectiu principal l'extracció de coneixement a partir d'una gran quantitat de dades, continguda en tot tipus de bases de dades .El procediment experimental general adaptat a un procés de mineria de dades típic inclou les següents fases: l'estat del problema i formulació de la hipòtesi, recol·lecció de les dades, pre-processament de les dades, estimació del model i, finalment, interpretació del model i extracció de conclusions.

Podem plantejar un exemple d'un problema que s'ha modelat en el domini de la salut i: Pot ser interessant, per diferents raons, per trobar grups homogenis de pacients en funció dels seus antecedents personals i problemes. D'aquesta manera, el repartiment dels pressupostos/recursos o 'capitas' assignades a les unitats d'atenció d'aquests pacients podrien ser proporcionals a la composició d'aquests grups de pacients. Formulem aquí una hipòtesi relacionada amb el domini d'estudi (el camp de l'àmbit de la salut), és a dir, que l'assignació de recursos segons als agrupaments esmentats millorarà l'efectivitat de l'aprofitament del recursos en front a un repartiment a l'atzar. Una vegada que el procés de mineria de dades ha acabat i després d'una avaluació de l'efectivitat del nou repartiment de recursos requerida en els nous grups heterogenis, és possible establir, almenys en un cert grau, si les hipòtesis coincideix amb la realitat (això és simplement una aplicació del mètode científic, on una hipòtesi pot aconseguir l'estat de la teoria a través dels resultats experimentals i de prova). Però no hem de confondre aquesta hipòtesi específica amb una hipòtesi de mineria de dades: no hi ha tals hipòtesis per formular en un procés de mineria de dades, perquè, simplement, no es pot afirmar res abans de que el procés de mineria de dades hagi acabat; el procés de mineria de dades és només un procés de descobriment de coneixement, de manera que el coneixement de qualsevol manera no està disponible en aquesta primera fase. En altres paraules, l'ésser humà no té res a dir i hem de deixar que els conjunts de dades "parlin per si mateixos".

La recol·lecció de dades és la següent fase, i podem distingir aquí dos tipus de mètodes de recol·lecció de dades, en relació amb el grau d'implicació de l'usuari en aquestes: la forma intrusiva (participació activa de l'usuari) o la no intrusiva (usuari passiu de contribució posterior). Malgrat aquesta classificació

dels mètodes de recollida de dades, hi ha altres classificacions possibles si es consideren altres factors (sistemes d'informació utilitzats, tipus de domini, observacional versus enfocament experimental, etc.)

Les dades poden ser recollides d'una gran varietat de fonts de dades, que inclou cada tipus diferent de bases de dades (relacional, orientada a l'objecte, processament analític en línia multidimensional i, deductiu, paral·lel, distribuïdes, etc.), i també d'altres tipus de fonts que no estiguin específicament en bases de dades, però que poden exercir el paper d'un origen de dades implícita, com ara els registre d'ús dels arxius d'un servidor web, els continguts de text i multimèdia (mineria de contingut web) i HTML, les etiquetes XML per extreure estructures DOM (mineria web estructura).

Aquesta heterogeneïtat de les fonts de dades, a més d'altres desavantatges en els processos de recollida de dades, com per exemple, les característiques o atributs amb valors coherents en blanc o no, o la necessitat d'adaptar la informació disponible a la forma que en que millor s'ajusta al nostre problema de mineria de dades, poden explicar perquè les dades han de ser pre-processades sempre abans d'aplicar-se la tècnica de mineria de dades triada. En mineria s'han de fer tasques molt diferents de preprocessament sobre les dades genèriques (detecció de valors atípics, selecció de característiques, atributs numèrics de discretització, escalament o normalització i característiques de codificació, la reducció de la seva dimensió en conjunts grans de dades, etc.).

El filtrat de dades i la seva transformació són dues de les etapes de pre-processament de dades dels arxius de registres. Les dades que no són d'utilitat en el nostre estudi han de ser eliminades, així com qualsevol altre informació sense utilitat en l'estudi. Un exemple de transformació de dades també s'explica quan s'aborden els problemes de privacitat: hi ha registres que porten la informació identificativa de l'usuari, que es substitueixen amb un nou identificador dissociat que no ens permet associar les dades amb els usuaris originals.

Un cop que les dades s'han pre-processat, com s'ha indicat anteriorment, la següent fase es la de l'estimació del model. L'objectiu d'aquesta fase és aplicar les tècniques de mineria de dades escollides, els algoritmes o mètodes que aplicarem a les dades prèviament processades, per construir un model a partir del qual podrem extreure algunes conclusions sobre el problema plantejat o les hipòtesis formulades en el nostre domini o camp d'estudi (en la primera fase). Així, el model ha de ser interpretable per als nostres propòsits, és a dir, oferir un coneixement útil sobre els objectius dels nostres dominis. Per exemple, si el nostre estudi cerca establir una relació, si n'hi ha, entre les característiques dels pacients i els ingressos hospitalaris futurs, necessitarem, en primer lloc, construir un model de mineria de dades descriptiu per tal de trobar patrons en les característiques dels usuaris, i després aplicarem un altre model de mineria de dades descriptives per establir la dependència entre els patrons trobats i els ingressos (en aquests casos, s'aplicaran les regles d'associació, però també es poden emprar altres models descriptius, com per exemple l'agrupació sense supervisió). Però, d'altra banda, si el nostre problema consisteix en la formulació d'una hipòtesi, llavors haurem de construir un model de mineria de dades predictiu per avaluar-la (per exemple arbres, xarxes, classificació i regressió). Un exemple podria ser la sospita que un usuari que ha estat visitat repetidament a urgències de primària per episodis aguts i amb unes característiques pròpies, pugui tenir major probabilitat d'acabar ingressant a l'hospital de referència (el cas del nostre estudi). Així, el model de predicció de mineria de dades ha de predir aquest fet sobre la base de les dades prèvies i classificar correctament als pacients que han visitat prèviament un servei d'urgències i que pertanyen al grup de pacients que ingressa a l'hospital.

A part de la construcció d'un model per a l'obtenció de nous coneixements o confirmar una hipòtesi relacionades amb el nostre problema, un altre dels objectius que es presenta generalment en una investigació de mineria de dades és la comparació de l'eficàcia i l'eficiència de les tècniques de mineria de dades o algoritmes diferents. De fet, l'àrea de la Intel·ligència artificial sovint està relacionada, perquè la comunitat científica està sempre a la recerca de nous algoritmes i nous enfocaments per augmentar la seva aplicabilitat i eficàcia; i la Intel·ligència Artificial proporciona una bona base teòrica al respecte, en oferir l'estudi teòric detallat sobre com algun tipus d'algorismes bayesians inductius es poden aplicar per inferir i predir comportaments.

Per tant, l'estat de l'art en el camp de la mineria de dades en l'àrea de la salut no només tractarà la investigació de nous factors relacionats amb els problemes d'estudi i el suggeriment de possibles hipòtesis, sinó que també cercarà la millora de les tècniques de mineria de dades, algorismes i mètodes aplicats específicament a aquest camp.

Importància i usos de la mineria de dades en Medicina i Salut Pública

Malgrat les diferències en els enfocaments, el sector salut té cada cop més necessitat de la minera de dades. Hi ha diversos arguments que podrien recolzar l'ús de la mineria al sector de la salut, que abasten no només els interessos de la salut pública, sinó també del sector privat de la salut (que, de fet, com es pot demostrar més endavant, són també interessats en la salut pública).

Sobrecarrega de Dades. Hi ha una gran quantitat de coneixement que s'obtindrà a partir dels registres de salut informatitzats. No obstant això, la immensa majoria de les dades emmagatzemades en aquestes bases de dades fa que sigui extremadament difícil, si no impossible, per als éssers humans tamisar a través d'ells i descobrir el coneixement (Cheng, et al 2006). De fet, alguns experts creuen que els avenços mèdics han disminuït, atribuint aquest fet a l'escala i la complexitat prohibitiva de l'actual informació mèdica. Els ordinadors i la mineria de dades són més adequats per a aquest propòsit. (Shillabeer Roddick i 2007).

Medicina basada en l'evidència i la prevenció dels errors hospitalaris. Quan les institucions mèdiques s'apliquen la mineria de dades en les dades existents, es poden descobrir coneixements nous, útils i que pot salvar vides, que d'una altra manera haurien estat inútils en les seves bases de dades. Per exemple, un curs estudi sobre els hospitals i la seguretat va trobar que al voltant del 87% de les morts hospitalàries en els Estats Units s'haurien pogut evitar, si el personal de l'hospital (incloent metges) haguessin tingut més cura per tal d'evitar errors (HealthGrades Estudi Hospitals 2007). La utilitat de la mineria de dades amb els registres hospitalaris, amb qüestions com la seguretat, poden ser identificats i redirigits per la direcció de l'hospital i els reguladors del govern.

La formulació de polítiques en matèria de salut pública. Lavrac et al. (2007) van combinar els SIG i la mineria de dades utilitzant entre d'altres, Weka J48 (gratuït, de codi obert, eines de mineria de dades basades en Java), per analitzar similituds entre els centres de salut comunitaris a Eslovènia. L'ús de mineria de dades, va permetre descobrir patrons entre els centres de salut que van conduir a les noves recomanacions de Salut Pública. Van arribar a la conclusió que "la mineria de dades i els mètodes de suport a les decisions, incloent nous mètodes de visualització, poden conduir a un millor acompliment en la presa de decisions ".

Els factors anteriors ens recorden un incident a la Filipines en el Centre Mèdic de Rizal a Pasig City l'octubre de 2006. El no tenir implementades mesures estrictes de sanejament i esterilització a l'hospital, va contribuir a la mort de diversos nens nounats a causa de sèpsia neonatal (infecció bacteriana). En realitat, ningú sabia el que estava passant fins que la mort es van fer més freqüents.

En examinar els registres hospitalaris, el Departament de Salut (DOH) va trobar que 12 dels 28 nadons nascuts a l'octubre 4, per exemple, van morir de sèpsia (Tandoc 2006). Amb una base de dades integrada i l'aplicació de la mineria de dades el Departament de Salut va poder detectar aquests esdeveniments inusuals i reduir-los abans que empitjoressin.

Més valor pels seus diners i l'estalvi de costos. La mineria de dades permet a organitzacions i institucions treure més profit de les actuals dades a un cost addicional mínim. KDD i mineria de dades s'han aplicat a descobrir el frau en les targetes de crèdit i reclamacions d'assegurances (Kou et al. 2004). Per extensió, aquests tècniques també podrien usar-se per detectar patrons anòmals a les reclamacions

d'assegurances de salut, especialment aquells que funcionen per PhilHealth, el sistema nacional d'assegurança de salut per les Filipines.

La detecció primerenca i / o prevenció de malalties. Cheng, et al fa ús d'algorismes de classificació per ajudar en la detecció primerenca de la malaltia cardíaca, un problema important de salut pública a tot el món. Cao et al (2008) descriu l'ús de mineria de dades com una eina per ajudar en la monitorització de tendències en els assajos clínics de vacunes contra el càncer. Mitjançant l'ús de la mineria de dades metge experts podria trobar patrons i anomalies molt millor que simplement mirant a un conjunt de dades tabulades.

La detecció primerenca i tractament de les malalties pandèmiques i la formulació de polítiques de salut pública. Els experts en salut han començat a buscar la manera d'aplicar la mineria de dades per a la detecció primerenca i gestió de pandèmies. Kellogg et al. (2006) indica que combinen tècniques de modelatge espacial, simulació i la mineria de dades espacials per trobar característiques interessants del brot de la malaltia. L'anàlisi resultant de la mineria de dades en l'entorn simulat podria llavors utilitzar-se per la formulació de polítiques sanitàries i detectar i tractar els brots de malalties.

Wong et al. (2005) va introduir WSARE, un algorisme per detectar brots en les seves primeres etapes. WSARE, que és l'abreviatura de "Esdeveniments recents estranys", es basa en l'associació de normes i xarxes bayesianes. Aplicant-lo en models de simulació han donat lloc a prediccions relativament precises dels brots de malalties simulades. Per descomptat, aquests tipus de reclamacions sempre vénen amb advertències de tenir precaució en aplicar aquests models en la vida real.

Diagnòstic no invasiu i suport de decisions. Alguns dels procediments de diagnòstic i de laboratori són invasius, costosos i dolorosos per al pacient. Un exemple d'això és la realització d'una biòpsia en dones per detectar el càncer cervical. Thangavel et al (2006) utilitza l'algorisme d'agrupament K-means per l'anàlisi dels pacients amb càncer de coll uterí i es va trobar que l'agrupació trobada millorava els resultats predictius existents sobre la opinió mèdica. Van trobar un conjunt d'atributs interessants, que podrien ser utilitzats pels metges com a suport addicional sobre la conveniència o no de recomanar una biòpsia d'un pacient sospitós de que té el càncer de coll uterí.

Gorunescu (2009) descriu com diagnòstic assistit per ordinador (CAD) la ultrasonografia endoscòpica elastografia (Eusebi) realçades per la mineria de dades per crear un nou sistema no-invasiu de detecció de càncer. En l'enfocament tradicional, els metges analitzen la imatge ecogràfica per decidir si un pacient serà sotmès a una biòpsia.

El judici del metge és principalment subjectiu, depenent sobretot de la interpretació ecogràfica (veure captura de pantalla mostra de vídeo, pàgina següent). Gorunescu aborda aquest problema d'una manera diferent, utilitzant la mineria de dades. No va estudiar la demografia dels pacients. En lloc d'això es va centrar en les ecografies. Primer es va capacitar un algorisme de classificació usant un perceptró multicapa (MLP) en casos coneguts de tumors malignes i benignes.

El model analitza els píxels RGB i el seu contingut per trobar patrons suficients per distingir entre tumors malignes i benignes. A continuació, el seu equip va aplicar el model resultant a altres casos. Van trobar que el seu model presentava una alta precisió en el diagnòstic amb només una petita desviació sobre l'estàndard.

Els esdeveniments adversos de medicaments (EAM). Alguns medicaments i productes químics que han estat aprovats com inofensius en els éssers humans, més tard s'ha descobert tenien efectes nocius després del seu ús públic a llarg termini.

Wilson et al. (2003) va revelar que els EUA Food and Drug Administration a través de la mineria de dades permetia descobrir els efectes secundaris dels medicaments . Aquest algoritme anomenat MGPS o Multi-ítem Gamma Poisson Shrinker va ser capaç de trobar amb èxit el 67% dels EAM cinc anys abans que es van detectar utilitzant mètodes tradicionals.

Hem vist com les aplicacions de mineria de dades podrien ser utilitzats en la detecció primerenca de malalties, la prevenció de les morts, la millora dels diagnòstics i fins i tot detectar reclamacions fraudulentes de salut. No obstant això, hi ha advertències per l'ús de la mineria de dades en l'assistència sanitària.

Valors estadístics dels algoritmes

1. **Estadístic de Kappa o Kappa Statistics**: Prenent el concepte de [30] Kappa pel seu nom grec és un índex que compara l'encert o acostament entre el que s'ha d'esperar per realitzar o tenir un canvi d'acord a certes característiques i paràmetres plantejats. Pot ser pensat com un canvi correcte proporcional a l'acostament que es desitja; els possibles valors van des d'un rang de +1 (acord o apropament perfecte), 0 (cap acord per sobre del que s'esperava) i -1 (total desacord).

Per realitzar el càlcul emprarem la fórmula:

Kappa = (Encert observat-Canvi en l'esperat) / (1-canvi en l'esperat)

Si valorem els algoritmes que ens que tenim menys instàncies classificades incorrectament (K2 i HC) observem un kappa d'entre 0,0035 i 0,0036 respectivament, mentre que amb TAN el kappa millora a 0,0315. Tot i que els valors són baixos, es troben en un rang positiu, denotant acord o apropament al resultat positiu buscat, que en aquest cas seria la predicció d'un ingrés hospitalari. Independentment dels altres resultats dels algoritmes sembla que mostra un baix però positiu, nivell de predicció respecte a variable classe "ingrés hospitalari", considerada en l'estudi.

2. **Mean absolute error o error mitjà absolut**: Seguint els conceptes de [31], el MAE mesura la magnitud mitjana dels errors en un conjunt de càlculs, sense tenir en compte la seva direcció. Això dóna la mesura de precisió per a les variables contínues. en altres paraules, el MAE és la mitjana de la mostra de verificació dels valors absoluts de les diferències entre els càlculs i la corresponent observació. El MAE és un resultat lineal, el que significa que totes les diferències individuals es ponderen per igual a la mitjana.

La funció aquesta donada per:

$$\mathrm{MAE} = \frac{1}{n}\sum_{i=1}^{n}|f_i - y_i| = \frac{1}{n}\sum_{i=1}^{n}|e_i|\,.$$

On fi és la predicció i yi és el valor vertader.

Tots els algoritmes emprats en l'estudi mostren un valor similar d'entre 0,06 i 0,07 d'error mitjà.

3. **Root Mean Squared Error o error quadràtic mitjà**: El RMSE donat per [32] és una regla que mesura la magnitud mitjana de l'error. Això és, la diferència entre el pronosticat i els corresponents valors observats al quadrat perquè després sigui una mitjana a al llarg de la mostra. Finalment, es pren l'arrel quadrada de la mitjana. Atès que els errors són al quadrat, l'RMSE dóna un pes relativament alt als grans errors. Això significa que el RMSE és més útil quan els grans errors són particularment indesitjables. Tingueu en compte que el RMSE serà sempre major o igual a la MAE (error mitjà absolut), la gran diferència entre ells, és en els errors individuals de la mostra. Si el RMSE = MAE, llavors tots els errors són de la mateixa magnitud. Tots dos poden anar de 0 a ∞ i són orientats a què els valors més baixos són els millors.

$$E_i = \sqrt{\frac{1}{n}\sum_{j=1}^{n}\left(P_{(ij)} - T_j\right)^2}$$

On P (ij) és el valor que s'ha predit individualment per al programa i del cas j, i Tj és el valor objectiu per al cas j. És així que P (ij) = Tj i Ei són els rangs dels índexs de 0 a infinit, on 0 correspon a l'ideal [34].

En el nostre cas els resultats mostren un valor homogeni de 0,18, essent un valor baix d'error mitjà dels resultats

4. **Relative absolute error o error relatiu absolut**: Donant com concepte el pres per [34], és aquell que ajuda a predir un valor relatiu, que no és més que la mitjana dels valors reals. Això vol dir, l'error no és més que el total absolut de l'error més no és el total de l'error al quadrat. Per tant, l'error absolut relatiu pren el total i absolut error que es normalitza dividint pel total d'error absolut de la predicció simple.

Matemàticament, l'error relatiu absolut Ei d'un individu i és avaluat per l'equació:

$$E_i = \frac{\sum_{j=1}^{n}\left|P_{(ij)} - T_j\right|}{\sum_{j=1}^{n}\left|T_j - \overline{T}\right|}$$

On P (ij) és el valor que s'ha esperat per al valor i del cas j (fora dels n casos simples); Tj és el valor objectiu per al cas j; i T esta donat per la fórmula:

$$\overline{T} = \frac{1}{n}\sum_{j=1}^{n} T_j$$

Si el numerador és igual a 0 ia Ei = 0. Per tant el índex Ei va des de 0 fins a infinit, on 0 correspon a l'ideal.

En l'estudi plantejat trobem valors massa elevats per ser desitjables, tractant-se d'un elevar valor relatiu.

5. **Root relative squared error o arrel quadrada d'error relatiu:** Citant [35], aquesta formula simple no és més que la mitjana dels valors reals. D'aquesta manera, la relativa d'error al quadrat presa el total d'errors al quadrat i es normalitza dividint pel total d'errors simples al quadrat. En prendre l'arrel quadrada del valor relatiu, l'error es redueix a les mateixes dimensions que la quantitat prevista.

Matemàticament, l'arrel quadrada d'error relatiu Ei d'un individu i és avaluat per la següent equació:

$$E_i = \sqrt{\frac{\sum_{j=1}^{n}\left(P_{(ij)} - T_j\right)^2}{\sum_{j=1}^{n}\left(T_j - \overline{T}\right)^2}}$$

On P (ij) és el valor que s'ha esperat per al valor i del cas j (fora dels n casos simples); Tj és el valor objectiu per al cas j; i T esta donat per la fórmula:

$$\overline{T} = \frac{1}{n}\sum_{j=1}^{n} T_j$$

Per donar explicació als conceptes que segueixen (6 a 11), es prengués com a base la següent matriu de confusió:

Classify as ->	C1	C2	...	Cz
C1	N11	N12	...	N1z
C2	N21	N22	...	N2z
...	...	...	...	...
Cz	Nz1	Nz2	...	Nzz

Els valors del treball es mostren similars a l'error relatiu absolut

6. **TP Rate o True Positive Rate o Recall**: Aquesta mesura està definida pel quocient entre el nombre d'exemples que classifiquen correctament per a una classe i el nombre total d'exemples per a la classe estudiada. En altres paraules és la proporció de elements que estan classificats dins de la classe Ci, d'entre tots els elements que realment són de la classe Ci. A la matriu de confusió és l'element diagonal dividit per la suma de tots els elements de la fila. Quan les sensibilitats pertinents per cada exemple de classe botiga a 1, la matriu de confusió tendirà a ser una matriu diagonal.

TP Rate = TP / (TP + FN)
TP Rate (C1) = N11 / (N11 + N12 + ... + N1z)
TP Rate (C2) = N22 / (N21 + N22 + ... + N2z)
TP Rate (CZ) = NZZ / (Nz1 + Nz2 + ... + NZZ)

7. **FP Rate o False Positive Rate**: És la proporció d'exemples que han estat classificats dins de la classe Ci, però pertanyen a una classe diferent. A la matriu de confusió és la suma de la columna de la classe Ci menys l'element diagonal dividit la suma de les files de la resta de les classes.

TP Rate = FP / (FP + TN)
TP Rate (C1) = N21 / (N31 + ... + N1z) / [(N21+...+N2Z) + (N31+...+N3Z) + (NZ1+...NZZ)]

Cal destacar aquest indicadors en els resultats observats, donat que es mostren uns valors alts de TP rates i FP rates per al grup de casos que no han ingressat (pròxims a 1), però per als casos que si han ingressat i que ens interessa especialment tenir detectats per el model, ofereix uns valors molt baixos (pròxims a 0). Sembla que tindríem un model d'alt valor predictiu negatiu, però escàs valor predictiu positiu d'ingrés.

8. **Precisió:** Proporció d'exemples que realment tenen classe Ci d'entre tots els elements que s'han classificat dins de la classe Ci. A la matriu de confusió és l'element diagonal dividit per la suma de la columna en la qual s'estableixen.

Prec (Model) = (N11 + N22 + ... + NZZ) / Total exemples
Prec (C1) = N11 / (N11 + N21 + ... + Nz1)
Prec (C2) = N22 / (N12 + N22 + ... + Nz2)
Prec (CZ) = NZZ / (N1z + N2z + ... + NZZ)

En general els algoritmes mostren dades força precises, amb valors d'aquest índex propers a 1

9. **F-Measure**: És una mesura que combina la precisió amb el Recall o TPR per a la classe Ci.

F-Measure = (2 * Precisió * Recall) / (Precisió + Recall)

10. **False Negative Rate:** És la proporció d'elements que no classifiquen per a la classe Ci, d'entre tots els elements que realment són de la classe Ci. A la matriu de confusió és la suma de tots els elements de la fila excloent a la diagonal dividida per la suma de tots els elements de la fila.

FN Rate = 1 - TPR = 1 - [TP / (TP + FN)] = FN / (FN + TP)
FN Rate (C1) = [(N11 + ... + N1z) - N11] / (N11 + N12 + ... + N1z)
FN Rate (C2) = [(N21 + ... + N2z) - N22] / (N21 + N22 + ... + N2z)
FN Rate (CZ) = [(Nz1 + ... + NZZ) - NZZ] / (Nz1 + Nz2 + ... + NZZ)

11. **True Negative Rate o Especificitat**: És la proporció de exemples que han estat classificats dins de les altres classes diferent a la classe Ci. A la matriu de confusió és la suma de les diagonals menys l'element de la classe Ci dividit la suma de les files de la resta de les classes.

TN Rate = 1 – FPR = 1 – [FP / (FP + TN)] = TN / (TN + FP)
TP Rate (C1) = (N22 + N33 + ...+ NZZ) / [(N21+...+N2Z) + (N31+...+N3Z) + (NZ1+...NZZ)]

Tal com hem comentat anteriorment els valors dels algoritmes mostren un model amb bona especificitat, capaç de detectar els verdaders negatius, però escassament efectiu en detectar els realment positius.