Màster Universitari en Enginyeria Informàtica

Optimització de diagnòstics mitjançant arbres de classificació

Treball Final de Màster – Àrea d'Intel·ligència artificial

Memòria

Gener del 2015

Alumne: Pau Xiberta i Armengol
Director del treball: Samir Kanaan Izquierdo
Codirectors externs: Imma Boada i Esteve del Acebo

Agraïments

Trobo just començar l'apartat d'agraïments, sempre que tinc l'oportunitat de realitzar un treball d'aquestes característiques, agraint profundament als meus pares l'esforç majúscul que han fet per mi i la voluntat constant d'empènyer-me a formar-me acadèmicament. Puc relacionar tota la meva vida acadèmica, incloent el Màster Universitari en Enginyeria Informàtica que cloc amb aquest projecte, amb el consell i l'ajuda dels meus pares, als quals dec tot el que hagi pogut aconseguir.

D'altra banda, i de manera més específica, he d'agrair també al meu director del treball, en Samir Kanaan, i als meus codirectors externs, l'Imma Boada i l'Esteve del Acebo, el seu suport constant i els comentaris encertats a l'hora de corregir alguns dels aspectes del treball. Sense les seves idees, correccions, aportacions de millores i, sobretot, sense els seus ànims, hagués estat molt més complicat, per no dir impossible, realitzar aquest projecte còmodament i amb la motivació suficient.

També vull donar les gràcies als meus companys de despatx a la Universitat de Girona, els quals han sabut escoltar-me i ajudar-me quan els he plantejat algun problema d'aquest projecte, i han aportat idees que he manllevat per a enfortir alguns dels apartats que es descriuen en aquesta memòria.

Finalment, voldria estendre l'agraïment a tots els meus amics i família que sempre ajuden a mantenir aquest equilibri tan fantàstic entre el món acadèmic/professional i la vida personal.

A tots vosaltres, moltes gràcies!

Índex

Índex de figures

1. Introducció i objectius del projecte

1.1. Visió global

El projecte que volem desenvolupar té per objectiu optimitzar el procediment que utilitzen els professionals mèdics per a elaborar el diagnòstic de les patologies que pateixen els pacients. Els criteris que caldrà optimitzar poden ser variables, com ara minimitzar el cost de les exploracions necessàries i maximitzar l'eficiència, sempre a partir dels símptomes que acusi el pacient. Com a estratègia per a resoldre aquest problema utilitzarem els arbres de classificació.

1.2. Característiques del problema

El nostre projecte intentarà resoldre un problema que sorgeix a l'hora de seguir el procediment per a fer un diagnòstic.

Quan un pacient va a l'hospital i descriu al metge els símptomes que té, aquest últim ha de seguir un protocol per a realitzar un diagnòstic a partir d'aquests símptomes. Moltes vegades no n'hi ha prou amb la informació que ofereix el pacient, i cal fer algunes exploracions per a confirmar que els símptomes són reals. Per exemple, si un pacient manifesta que té dolor al cap, és possible que el metge demani una analítica, o potser una prova d'imatge com pot ser una ressonància magnètica o una tomografia computada.

Per a determinar quines proves cal fer es poden aplicar diferents tipus de criteris, com ara el temps d'espera del pacient, el cost mínim de l'exploració (per exemple, una ressonància magnètica comporta un cost més elevat que una tomografia computada), l'eficàcia de la prova, etc. Decidir què és millor en cada cas no sempre pot resultar fàcil. Una forma de resoldre aquest problema seria disposar de mètodes que ens indiquessin quin és el millor procediment en cada situació.

Cal tenir en compte, doncs, que no només ens interessa esbrinar quina patologia té el pacient a través dels símptomes que manifesta, sinó que sobretot ens volem centrar en quines són les exploracions més eficients en funció de determinats criteris per a identificar la patologia del pacient.

1.3. Propostes de solució

La nostra proposta és utilitzar els arbres de classificació com a estratègia per a optimitzar el diagnòstic d'un pacient, podent determinar quines són les exploracions que cal realitzar en funció de diferents paràmetres com poden ser el temps d'espera, el cost, etc.

Els arbres de classificació ens permeten dibuixar els diferents camins que pot seguir un metge a l'hora de fer el diagnòstic a partir d'un conjunt de símptomes. Cal tenir en compte que les entrades sobre les quals treballem són directament comparables amb la definició de les sortides que busquem; és a dir, els símptomes que patirà el pacient (descoberts a

través del mateix pacient o a través d'exploracions) seran les nostres entrades, i una patologia, que serà la sortida que busquem, vindrà definida també per un conjunt de símptomes. Per tant, els arbres de classificació s'adapten a la nostra problemàtica tenint en compte que l'elecció d'un camí o un altre es farà depenent de si es compleixen o no els diferents nodes, que correspondran als símptomes.

Tot i això, per a aconseguir escollir el procediment òptim a partir de diferents criteris, caldrà estendre l'algorisme bàsic dels arbres de decisió per a tenir en compte altres factors diferents de la capacitat de discriminació dels símptomes i de les proves, com ara el cost de les exploracions, per exemple.

També ens plantegem estudiar altres classificadors que permetin trobar la patologia més probable a partir d'uns determinats símptomes, de manera que puguem fer la comparació amb els arbres de decisió. De totes maneres, l'objectiu del projecte és sobretot implementar un algorisme que ens permeti trobar el procediment òptim de diagnòstic, i no només la patologia més probable; és a dir, ens interessa molt més el camí traçat en l'arbre de decisió que no pas el node terminal final. Per tant, i en aquest cas, els arbres de decisió ens permeten discernir molt més clarament quin és el camí que s'ha seguit per a obtenir el resultat final, a diferència d'altres classificadors com ara l'AdaBoost o les màquines de suport vectorial.

1.4. Objectius

L'objectiu principal del projecte és aconseguir un procediment que permeti optimitzar un diagnòstic a partir de determinats criteris fixats per l'usuari. Per a aconseguir-ho caldrà:

- Fer un estudi sobre els arbres de classificació per a determinar quin és el que s'adapta millor al nostre problema. També ens caldrà estudiar els diferents algorismes que es poden aplicar.
- Implementar els algorismes seleccionats a l'apartat anterior i aplicar-los sobre un entorn de proves que crearem. En aquest entorn simularem diferents patologies i símptomes.

També es farà una cerca per a trobar bases de dades públiques que relacionin patologies amb símptomes, però si no se'n troba cap d'adequada, i tenint en compte que no disposem del temps necessari per a recollir símptomes reals i crear una base de dades creïble, ens proposem crear una base de dades fictícia que simuli la relació entre diferents patologies i els símptomes que les defineixen, amb la possibilitat d'incorporar manualment algunes patologies i símptomes reals obtinguts a través de fonts fiables.

De totes maneres, aquesta base de dades servirà per a experimentar amb l'algorisme escollit i per a comprovar que s'aconsegueix seguir el procediment òptim per a elaborar el diagnòstic. Per tant, l'objectiu a llarg termini serà aconseguir una base de dades completa i real, però l'objectiu a curt termini és trobar l'algorisme que optimitzi el procediment de diagnòstic.

2. Planificació

Abans de començar qualsevol projecte cal dur a terme la planificació inicial d'aquest, on s'especifiquen tots els passos que, en un desenvolupament normal, s'haurien de dur a terme.

En aquest apartat, doncs, s'exposarà aquesta planificació inicial del projecte tot indicant els dies clau del calendari i diferents fases que caldrà anar superant per a completar el projecte.

2.1. Calendari de fites

Les fites són parts clau del projecte que ens permeten tenir una aproximació del progrés d'aquest. Acostumen a coincidir amb parts del projecte que cal entregar o amb els terminis de les fases, i és possible indicar una data d'assoliment de cada fita.

A continuació, a la Taula 1, es representa el calendari de fites, que marcarà els punts clau del calendari en què s'ha planificat acabar una determinada fase o lliurar una determinada part del projecte, sempre especificant la data en què caldria assolir la fita.

Fita	Data
1) Elecció del tema del Treball Final de Màster	**21 de setembre**
2) Aprovació de la planificació inicial del projecte	**2 d'octubre**
3) Disponibilitat de l'entorn de desenvolupament i de proves	**12 d'octubre**
4) Aprovació de l'algorisme a utilitzar	**19 d'octubre**
5) Obtenció dels primers resultats	**30 d'octubre**
6) Aprovació i realització de proves	**9 de novembre**
7) Realització de la comparativa entre diferents algorismes	**16 de novembre**
8) Obtenció i anàlisi dels resultats finals	**27 de novembre**
9) Aprovació final del projecte	**7 de desembre**
10) Finalització de la redacció de la memòria	**14 de desembre**
11) Realització de la presentació i preparació de la defensa	**4 de gener**

Taula 1. Calendari de fites.

2.2. Planificació inicial detallada

Tot i que les paraules "planificació inicial" i "detall" semblen un oxímoron, el que es proposa aquí és construir una planificació inicial setmana a setmana d'aquelles parts del projecte que s'haurien d'anar completant.

La Taula 2 mostra aquesta planificació indicant, per a cada setmana de desenvolupament del projecte, les tasques que caldria realitzar.

Setmana	Tasques
Setmana 1 (15/09 – 21/09)	Elecció del tema del treball, elaboració de la descripció general del projecte i estudi dels mòduls de l'assignatura corresponent al Treball Final de Màster.
Setmana 2 (22/09 – 28/09)	Descripció més detallada del projecte, redacció dels objectius, elaboració de la planificació i cerca d'informació i de l'estat de l'art.
Setmana 3 (29/09 – 05/10)	**Lliurament de la primera entrega**
Setmana 4 (06/10 – 12/10)	Estudi de la diversitat de classificadors i dels seus avantatges respecte al nostre problema, i cerca d'informació sobre les eines disponibles.
Setmana 5 (13/10 – 19/10)	Aprofundiment sobre l'eina escollida per a fer el desenvolupament (adaptació) i elaboració d'esbossos dels algorismes.
Setmana 6 (20/10 – 26/10)	Aplicació de l'algorisme escollit i obtenció dels primers resultats.
Setmana 7 (27/10 – 02/11)	**Lliurament de la segona entrega**
Setmana 8 (03/11 – 09/11)	Anàlisi dels resultats i elaboració i execució de proves.
Setmana 9 (10/11 – 16/11)	Aplicació d'altres algorismes i comparació dels resultats.
Setmana 10 (17/11 – 23/11)	Descripció detallada de l'algorisme utilitzat, desglossant els avantatges respecte a altres algorismes.
Setmana 11 (24/11 – 30/11)	**Lliurament de la tercera entrega**
Setmana 12 (01/12 – 07/12)	Correcció d'errors i aplicació de millores al projecte.
Setmana 13 (08/12 – 14/12)	Finalització de la redacció de la memòria.
Setmana 14 (15/12 – 21/12)	Preparació de la presentació.
Setmana 15 (22/12 – 28/12)	Repàs general del projecte i preparació de la defensa.
Setmana 16 (29/12 – 04/01)	**Lliurament de la quarta i última entrega**

Taula 2. Planificació inicial setmana a setmana.

2.3. Diagrama de Gantt

Una altra manera de representar la planificació inicial és fer-ho de forma gràfica. El diagrama de Gantt permet veure la planificació de l'evolució del projecte en una línia temporal i observar la durada de cada tasca.

A continuació, a la Figura 1, es representa el diagrama de Gantt de la planificació inicial d'aquest projecte, marcant adequadament les fases principals.

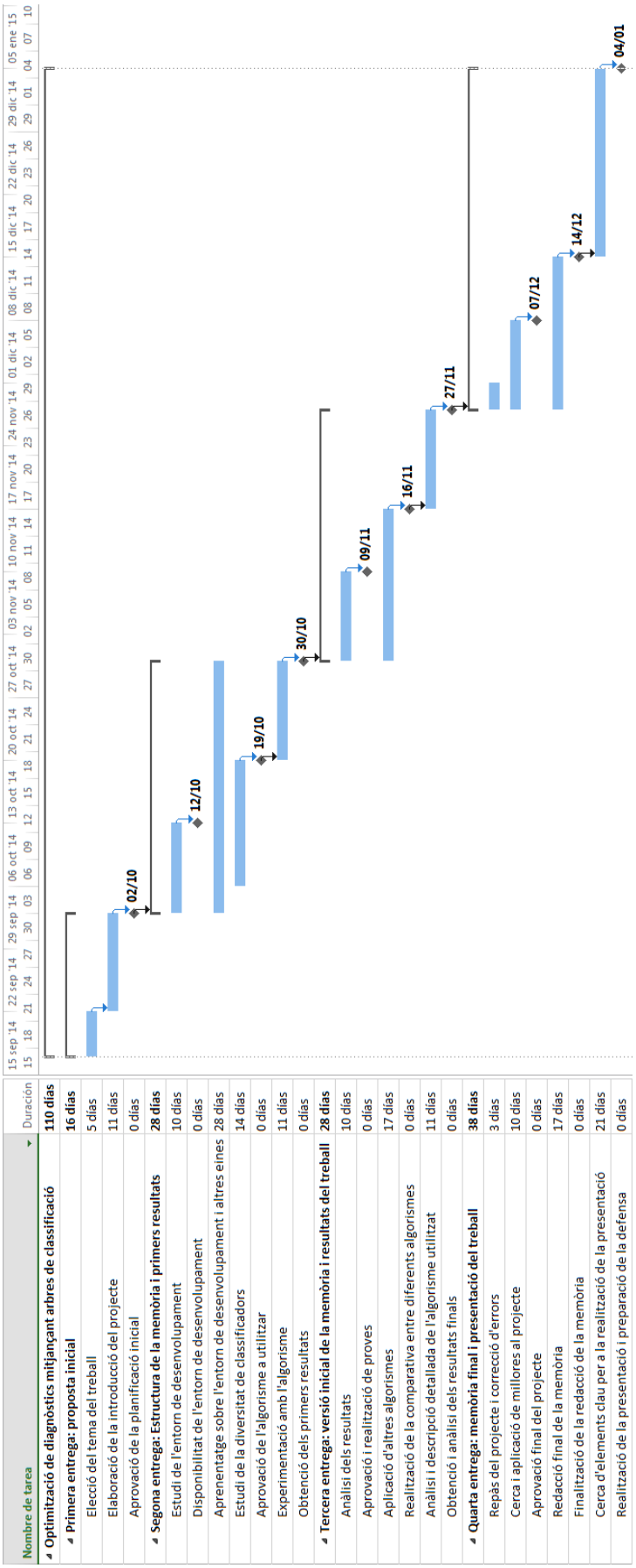

Figura 1. Diagrama de Gantt de la planificació inicial.

3. Estat de l'art

Per a desenvolupar una nova proposta per a un problema concret, primer cal estudiar quina és la situació actual d'aquest problema i de les solucions que s'han aportat fins ara. Per tant, en aquest apartat s'exposarà el marc de treball i els conceptes previs necessaris per a comprendre millor el projecte. Això inclou, doncs, un estudi sobre la situació actual del mètode que es pretén utilitzar per a solucionar el problema que planteja el projecte, tant el mètode en general com el seu ús en l'àmbit en el qual s'aplicarà en aquest projecte.

A continuació es donaran indicacions generals sobre la situació actual dels arbres de decisió i de les seves característiques, així com de les aplicacions d'aquests arbres sobre temes relacionats amb l'àmbit mèdic.

3.1. Conceptes previs

En els següents apartats d'aquest projecte es farà al·lusió a diferents conceptes que estan molt relacionats amb la problemàtica que es vol resoldre. Alguns dels conceptes són específics d'aquesta problemàtica, i alguns altres són conceptes que s'han creat per a entendre millor la proposta de solució que s'aportarà. Per tant, és important que abans de desenvolupar el contingut del projecte es dediqui un apartat a explicar el significat d'aquests conceptes clau que ajudaran a comprendre la globalitat del projecte.

3.1.1. Arbre de decisió i arbre de classificació

En aquest projecte es parlarà indistintament d'arbres de decisió i arbres de classificació. En realitat no es tracta del mateix concepte, sinó que els arbres de classificació són un tipus dels arbres de decisió. Quan el resultat predit d'un arbre de decisió correspon a una classe o etiqueta, és a dir, a un conjunt finit de valors, llavors es tracta d'un arbre de classificació; en canvi, quan el resultat predit correspon a un valor continu, com ara un nombre real, llavors es tracta d'un arbre de regressió.

En aquest projecte, com que els resultats predits de l'arbre de decisió sempre seran classes o etiquetes, o sigui, noms de patologies, els arbres sempre seran de classificació. Per tant, sigui quin sigui el concepte que s'utilitzi en aquest projecte, sempre s'estarà fent referència al mateix: a l'arbre de decisió generat per a predir una patologia a través d'un procediment òptim.

3.1.2. Nodes interns i nodes terminals

Els arbres de decisió tenen una terminologia pròpia que cal tenir clara per a comprendre els següents apartats. Quan es parla d'un node de l'arbre es fa referència a un atribut que separa les dades en dos o més subconjunts depenent del valor que pugui prendre aquest atribut. Quan aquest node es troba al final de l'arbre i ja no separa les dades s'anomena "node terminal" o "fulla", i no s'associa a cap atribut, sinó a una classe (en el nostre cas, a una patologia). En canvi, quan no es troba al final de l'arbre s'anomena "node intern".

D'altra banda, cadascuna de les opcions d'un node intern, és a dir, cadascun dels camins que es pot escollir des d'un node intern i a partir d'un valor concret de l'atribut, rep el nom de "branca" o "aresta". El primer node de l'arbre, per la seva banda, rep el nom de "node arrel". Com es pot observar, les referències a un arbre són clares, i és fàcil entendre'n el significat si es fa la comparació amb la realitat.

3.1.3. Mètode de discriminació

Durant tot el projecte es parlarà molt sovint dels mètodes de discriminació, sobretot quan es faci referència als diversos arbres de decisió que es puguin generar. Quan es construeix un arbre de decisió, el primer que cal fer és esbrinar quin dels atributs té la capacitat de separar millor les dades, és a dir, quin atribut discrimina millor. Sabent això, aquest atribut es pot col·locar en nivells superiors de l'arbre i separar abans les dades en diferents branques per acabar arribant a la classe que millor predigui un exemple.

En el nostre cas, quan una exploració determini molt clarament que un exemple concret pot o no pot pertànyer a una patologia concreta, serà important que aquesta exploració es realitzi el més aviat possible, així es podran descartar de seguida algunes patologies.

El mètode que s'utilitza per a esbrinar quin dels atributs és el que separa millor les dades s'anomena mètode de discriminació. En aquest projecte se n'estudiaran quatre d'existents i es proposarà una millora d'optimització que es pot aplicar a tots, però n'existeixen molts més.

3.1.4. Patologia i malaltia

Tal com es podrà observar més endavant, el terme que s'utilitzarà en tot aquest projecte quan es vulgui fer referència a allò que pateix un malalt serà el terme "patologia". En llenguatge general, els termes patologia i malaltia es poden utilitzar com a sinònims i es pot suposar que tenen el mateix significat. De fet, l'exposició que es fa aquí del projecte es podria entendre igualment si el terme utilitzat fos "malaltia".

De totes maneres, però, el llenguatge mèdic diferencia aquests dos termes en un punt molt concret. S'utilitza el terme "malaltia" quan es vol referir al procés i a les conseqüències d'aquell mal que pateix un pacient, mentre que s'utilitza el terme "patologia" per a referir-se a l'estudi del conjunt de símptomes que defineixen una malaltia.

Tenint en compte que aquest projecte es basa en l'estudi de diferents símptomes per a predir quin mal pateix un pacient, s'ha cregut convenient utilitzar el terme "patologia" perquè fa referència precisament a això, a l'estudi d'aquests símptomes.

3.1.5. Símptoma, síndrome, trastorn, patologia i malaltia

A part de la diferència entre patologia i malaltia, és important explicar la diferència que hi ha també entre una malaltia i un trastorn. Una malaltia ha de tenir una causa coneguda, un diagnòstic i un tractament, mentre que un trastorn és simplement una descripció. Per tant, es pot observar un trastorn d'algun àmbit concret sense que això signifiqui que es tracta d'una malaltia (per exemple, un trastorn de la conducta alimentària), de la mateixa manera

que es pot observar un trastorn que acabi resultant ser una malaltia (per exemple, un trastorn de la visió pot resultar ser esclerosi múltiple). Com que en aquest projecte només es tindran en compte patologies, el terme "trastorn" no s'utilitzarà.

D'altra banda, un símptoma fa referència a aquell fenomen que es pot percebre en el pacient i que ha sigut causat per una malaltia. Per tant, una malaltia pot causar diversos símptomes, l'estudi dels quals es recull a la patologia. Sovint també s'utilitza el terme "síndrome" per a referir-se a les causes d'una malaltia. Un síndrome és, doncs, el conjunt de símptomes que defineixen una malaltia. En aquest projecte no es farà ús d'aquest terme perquè sovint interessarà desglossar aquests símptomes i estudiar-los de forma separada; de fet, l'arbre de decisió requereix separar-los i esbrinar quin d'ells discrimina millor les dades, de la mateixa manera que cal saber quin d'ells té associada una exploració més eficient.

3.1.6. Símptoma i exploració

Tot i que es tracta de dos termes amb diferències clares de significat, en aquest projecte sovint s'utilitzaran indistintament. Un símptoma, tal com s'ha explicat, és un fenomen causat per una malaltia, mentre que una exploració és aquella acció que cal dur a terme per a esbrinar si un pacient pateix un determinat símptoma.

En aquest projecte, quan es generi un arbre de decisió per a ajudar en el diagnòstic d'una patologia, els nodes interns de l'arbre faran referència tant al símptoma com a l'exploració. Primer caldrà dur a terme l'exploració per a obtenir uns resultats, i a partir d'aquests resultats caldrà escollir una branca de l'arbre de decisió. De fet, l'elecció d'una determinada branca es pot entendre com el fet de patir o no patir el símptoma que està relacionat amb l'exploració que s'ha realitzat.

En aquest sentit, un símptoma sempre estarà lligat a una exploració, i a l'inrevés. En el cas que un símptoma es reconegui a través de múltiples exploracions, es consideraran símptomes diferents per a no complicar la generació de l'arbre de decisió. En general, en aquest projecte s'utilitzarà el terme "símptoma" quan es parli de predir correctament una patologia, i s'utilitzarà el terme "exploració" quan es parli d'obtenir el procediment òptim. Això és així perquè la predicció d'una patologia es basa únicament en el conjunt de símptomes que la defineixen, mentre que el procediment òptim, és a dir, el camí de l'arbre, s'obtindrà en part tenint en compte l'eficiència de les exploracions.

3.1.7. Paràmetre d'eficiència i criteri

El terme "paràmetre d'eficiència" o "criteri" s'utilitzarà, en aquest projecte, per a definir algun dels factors que influeixen en l'eficiència d'una exploració. Els dos termes seran utilitzats indistintament i agruparan un conjunt de factors que s'utilitzaran per a descriure l'eficiència de totes les exploracions.

Així, l'eficiència d'una exploració es podrà descriure, en aquest projecte, pel seu cost, pel temps que tarda en realitzar-se, pel nivell d'invasió que suposa per al pacient, i pel nivell de fiabilitat que té. Evidentment, podrien existir molts altres paràmetres d'eficiència, però en aquest projecte es tindran en compte aquests quatre. Per exemple, una tomografia computada és costosa en diners, pot tardar un temps significatiu en realitzar-se i és molt

invasiva, però alhora és molt fiable. Per tant, depenent de la importància que s'assigni a cadascun d'aquests paràmetres d'eficiència o criteris, es pot acabar definint de diverses maneres l'eficiència d'una exploració.

La definició d'aquests criteris, de fet, constitueix la innovació que es presenta en aquest projecte per a obtenir el procediment òptim a l'hora de realitzar el diagnòstic. No només es tindrà en compte la capacitat de discriminació de cada símptoma, sinó també l'eficiència de les exploracions que es requereixen per a obtenir els resultats que han de determinar si el pacient pateix o no el símptoma. Aquesta definició serà subjectiva i podrà ser modificada en tot moment.

3.2. Arbres de decisió

Els arbres de decisió han estat àmpliament utilitzats per a construir models de classificació que siguin semblants a la manera de raonar dels éssers humans i que siguin fàcils d'entendre. En aquest apartat, doncs, i tenint en compte que els arbres de decisió seran la base sobre la qual es desenvoluparà aquest projecte, s'anomenaran els problemes bàsics dels arbres de decisió i els punts de recerca actuals [Kotsiantis 2011; Rokach i Maimon 2005a; Rokach i Maimon 2005b; Murthy 1998; Safavian i Landgrebe 1991].

En primer lloc, cal destacar l'avantatge crucial que tenen els arbres de decisió sobre la majoria dels altres mètodes, i és que es poden generar models que no són de "caixa negra", és a dir, que els models generats amb els arbres de decisió són molt més fàcils de comprendre. En aquest sentit, s'han desenvolupat diversos algorismes per als arbres de decisió, com ara els algorismes C4.5, CART, SPRINT o SLIQ.

La recerca actual també diferencia dues fases clares durant la generació dels arbres de decisió: la fase de creixement i la fase de poda. La primera fase se centra en la creació de l'arbre a partir de la discriminació de dades i l'elecció d'atributs més òptims, mentre que la segona fase se centra en l'optimització de l'arbre eliminant aquelles parts més irrellevants i millorant la precisió global de l'arbre. Dins d'aquesta segona fase també es diferencia entre el que s'anomena la "prepoda" (*pre-pruning*), és a dir, la poda durant la fase de creixement de l'arbre; i la "postpoda" (*post-pruning*), és a dir, la poda després de la generació de l'arbre.

En realitat, un arbre de decisió no deixa de ser l'agrupament de diferents "subarbres" que s'uneixen mitjançant un node de nivell superior. Aquests subconjunts s'anomenen diferent depenent de les classes dels nodes terminals. Així, un subconjunt serà "pur" quan tots els nodes terminals facin referència a la mateixa classe, mentre que serà "impur" quan contingui nodes terminals que facin referència a classes diferents.

Un dels punts en els quals s'està centrant més la recerca actual sobre els arbres de decisió és en els diferents mètodes de discriminació que existeixen [Buntine i Niblett 1992; Fayyad i Irani 1992]. Alguns dels exemples són els que es tractaran en aquest projecte, i que discriminen per guany d'informació, per ràtio de guany, per índex de Gini o per exactitud. De fet, es pot considerar que en aquest projecte també s'innovarà en els mètodes de discriminació, ja que es proposarà una millora dels mètodes existents.

També s'estan estudiant temes relacionats amb la complexitat dels arbres, és a dir, amb la importància d'obtenir arbres petits per tal de no afectar massa la topologia dels arbres quan es produeixen petits canvis en les dades. Aquesta complexitat es pot controlar a través dels criteris d'acabament i dels mètodes de poda.

Existeixen diversos criteris d'acabament, i aquesta també és una línia de recerca actual. Per exemple, es pot decidir acabar la generació de l'arbre quan totes les instàncies del conjunt d'entrenament corresponen a la mateixa classe, quan s'arriba a una profunditat màxima establerta, o quan el millor atribut decidit pel mètode de discriminació no supera un determinat llindar ja establert.

Un altre objectiu de la recerca actual és evitar el sobreentrenament (*overfitting*) de l'arbre de decisió. Això ocorre quan l'arbre de decisió es genera tenint massa en compte característiques que són més aviat irrellevants. Algunes de les tècniques que existeixen per a evitar aquest sobreentrenament són les tècniques de poda i el preprocessament de dades (escollir els atributs més significatius).

Finalment, l'ús actual dels arbres de decisió se centra en intentar solucionar problemes estesos com ara la gestió de grans conjunts de dades, la gestió de problemes sensibles al cost de les dades [Drummond i Holte 2000], la gestió de dades incertes o absents, o la gestió de dades que es poden classificar en més d'una classe, per exemple. Algunes de les solucions que s'han aportat per a resoldre aquests problemes tenen en compte els models híbrids [Sheng i Ling 2005; Scott i Nowak 2004], és a dir, solucions que intenten incrementar la precisió dels models obtinguts combinant diferents mètodes d'aprenentatge supervisats.

Com a conclusió, doncs, es pot dir que la recerca actual en els arbres de decisió es pot enfocar en dos grans procediments que també constitueixen el mateix concepte dels arbres de decisió: en el procediment de construcció de l'arbre (o inducció) i en el procediment de classificació de nous exemples utilitzant l'arbre generat (o inferència).

3.3. Aplicacions dels arbres a medicina

Una vegada vistes les principals línies de recerca pel que fa als arbres de decisió, és fàcil veure com aquest mètode pot tenir diverses aplicacions en l'àmbit mèdic. En el nostre projecte els arbres de decisió han d'ajudar a obtenir no tan sols una predicció acurada sobre la patologia que pateix un pacient, sinó també un procediment òptim de diagnòstic. Tot i això, la idea d'aplicar els arbres de decisió a l'àmbit mèdic ja ha sigut àmpliament explotada pels investigadors, i existeixen multitud de casos en els quals els arbres de decisió (o la combinació d'aquests amb altres mètodes d'aprenentatge) han ajudat a solucionar problemes d'aquest àmbit.

Per posar alguns exemples d'aquests casos, a continuació es llisten algunes de les aplicacions que han sigut estudiades per altres investigadors i que han donat bons resultats:

- L'ús de mètodes de mineria de dades i d'aprenentatge per a la classificació i la predicció de malalties, desenvolupant un cas concret de classificació de subtipus de fallades cardíaques [Austin et al. 2013]. De fet, la tasca desenvolupada consisteix en

millorar les limitacions dels arbres de decisió per a millorar la precisió de les prediccions. En aquest sentit, té relació amb el nostre projecte en el fet que s'utilitzen mètodes d'aprenentatge per a solucionar un problema de classificació de malalties, i també que s'intenten superar les limitacions que pateixen els arbres de decisió.

- L'ús de la combinació de tecnologia de processament d'imatge amb un algorisme d'arbre de decisió en la identificació automàtica de malalties comunes de les cordes vocals en un estroboscopi de vídeos de laringe [Kuo et al. 2013]. En aquest cas, l'estudi intenta aprofitar les imatges capturades per un estroboscopi per a classificar algunes malalties de les cordes vocals a través d'un arbre de decisió amb una precisió superior al 92% que, amb algunes millores, asseguren que pot arribar gairebé al 99%. En relació amb el nostre projecte, aquest treball també utilitza els avantatges dels arbres de decisió per a classificar malalties.

- Un arbre de classificació per a la predicció de la malaltia benigna en la gestió de les masses renals tot ajudant al procés de pensament del metge [Rendon et al. 2011]. Aquest treball se centra sobretot en la predicció d'una sola malaltia, és a dir, a través de diverses entrades i amb un arbre de classificació es prediu si la malaltia és o no és benigna, havent-hi tan sols dues classes a l'arbre. El punt comú amb el nostre projecte és l'ús de l'arbre de decisió per a predir, a partir d'alguns símptomes, si la malaltia és benigna; la diferència és que només se centra en una malaltia, i no en diverses.

- Classificació de malalties de l'arròs utilitzant tècniques de selecció de característiques i de generació de regles [Phadikar et al. 2013]. Aquí l'estudi es refereix a la classificació de malalties en plantes infectades a partir de l'opinió d'experts, de símptomes de la planta i de característiques rellevants com ara el color, la forma o la posició de la part infectada. A diferència del nostre projecte, en aquest cas no s'utilitza explícitament un arbre de decisió (tot i que s'utilitzen estructures en forma d'arbre), però sí que té en comú l'ús de tècniques de selecció d'atributs i de generació de regles.

D'altra banda, per a fer estudis com el que es proposa en aquest projecte i com els que s'han exposat en els exemples anteriors és necessari disposar d'una base de dades que contingui exemples suficients per a obtenir resultats fiables. Les bases de dades actuals sobre aquests temes són moltes i diverses, però no se n'ha trobat cap d'accés lliure que complís els requeriments del nostre projecte, és a dir, que relacionés de forma clara algunes patologies amb els seus símptomes.

En alguns casos s'han trobat pàgines web que feien aquesta relació [Mayo Clinic; HazMap®; Diseases Database], però sense aclarir realment quins símptomes tenia cada patologia i, sobretot, sense disposar d'una manera senzilla d'utilitzar aquestes dades que no fos introduint-les manualment en un format adequat. De manera semblant, s'han trobat pàgines web de patologies més específiques [CTD; Gemina; Pagon et al. 1993], però totes mantenien el problema de l'exportació de les dades. Finalment, existeix un projecte d'una base de dades oberta i pública sobre relacions entre patologies i símptomes [Webb 2012], però es tracta d'un projecte encara en desenvolupament i la base de dades actual no és accessible.

Finalment, també existeixen alguns programaris que fan la funció de classificar les possibles patologies que pot patir un pacient en funció dels símptomes d'aquest, és a dir, softwares de diagnòstic. En cap dels casos, però, es posa l'accent en el procediment que cal seguir per a fer el diagnòstic, tal com es vol aconseguir en aquest projecte. Alguns dels softwares que existeixen són els següents:

- Software de suport a les decisions de diagnòstic anomenat SimulConsult®, gratuït però amb registre previ.
- Sistema de suport a les decisions de diagnòstic per a internistes (especialistes en les malalties internes no quirúrgiques) anomenat QMR (Quick Medical Reference), propietat d'OpenClinical i actualment no disponible.
- Projecte de suport a les decisions de diagnòstic anomenat Promedas, comercial, i que proveeix un esquema que mostra les malalties més probables a partir d'un conjunt de símptomes.

Després d'haver repassat alguns dels conceptes clau que s'utilitzaran en aquest projecte, i després d'haver vist l'estat de l'art actual sobre els arbres de decisió i la seva aplicació en l'àmbit mèdic, ja s'està en condicions d'explicar la proposta de solució que s'ha escollit per a solucionar el problema que planteja aquest projecte.

4. Proposta

Una vegada plantejat el problema i fet l'estudi sobre la situació actual de les possibles solucions, en aquest apartat es presentarà la proposta de solució escollida per a resoldre el problema que planteja aquest projecte, descrivint l'algorisme que s'utilitzarà i cadascun dels punts que es creguin adients per a entendre'l.

Primer de tot es farà una descripció general de la proposta i s'explicaran els detalls de la base de dades que caldrà utilitzar, per acabar explicant el funcionament d'alguns tipus d'arbres de decisió que seran útils per al problema i la innovació que s'hi afegirà per a resoldre el problema concret del projecte.

4.1. Descripció de la proposta

Abans d'entrar en detall, és bo explicar de manera general en què consisteix la proposta de solució escollida, els motius pels quals s'ha escollit i quins resultats es podran obtenir.

La proposta que s'ha escollit, doncs, consisteix en utilitzar els mecanismes que ofereixen els arbres de decisió per a intentar trobar el camí òptim a l'hora de fer un diagnòstic sobre una patologia. S'entén per "camí òptim" aquell que diagnostica la patologia d'una forma més eficient, ja sigui pel cost associat a les exploracions o per la comoditat dels pacients. En aquest sentit, els arbres de decisió permetrien que cada node de l'arbre fes referència a un símptoma específic (o a una exploració concreta relacionada amb aquest símptoma) de manera que avançar per l'arbre significaria anar descartant possibles patologies fins que només en quedés una, la més probable.

Com s'ha dit, cada node intern de l'arbre faria referència a un símptoma o a una exploració. Evidentment, un símptoma i una exploració són dos conceptes diferents, però van estretament relacionats. En la proposta que s'ha escollit, es pot dir que l'interès que generen cadascun dels dos conceptes és el mateix i es pot utilitzar de la mateixa manera. És a dir, conèixer els símptomes és necessari per a definir la patologia que pateix el pacient, però per a saber si un pacient té o no té un determinat símptoma sempre cal fer una exploració. Així, per exemple, per a saber si un pacient té febre sempre caldrà fer-li una exploració com ara prendre-li la temperatura amb un termòmetre. Per tant, en un arbre de decisió que dibuixi camins cap a una possible patologia, té sentit que els nodes interns facin referència a símptomes i a exploracions a la vegada; per una banda, el símptoma servirà per a orientar el camí cap a la patologia més probable (i descartar les menys probables), mentre que l'exploració associada servirà per a trobar el camí més eficient tenint en compte que cada exploració porta associats uns paràmetres com ara el temps o el cost, entre d'altres. Per continuar amb l'exemple, un node intern podria anomenar-se "febre/termòmetre", on "febre" seria el símptoma que podria ajudar a descobrir la patologia i "termòmetre" seria l'exploració que podria fer variar aquest camí depenent dels paràmetres d'eficiència d'aquesta exploració.

D'altra banda, els nodes finals o fulles farien referència a les patologies, i seria on acabarien tots els camins possibles de l'arbre, indicant finalment quina patologia és la més probable.

Si només es tingués en compte la part dels nodes interns que fa referència als símptomes, es podria considerar que la proposta de solució escollida no és massa innovadora, ja que predir una patologia a través dels seus símptomes i utilitzant arbres de decisió no és estrany. La part que fa més atractiva aquesta proposta és, doncs, el fet d'associar a cada símptoma una exploració i, a cada exploració, uns paràmetres d'eficiència. Per tant, el node terminal de l'arbre deixa de tenir tota la importància, i és el camí escollit de l'arbre el que s'acaba prioritzant. A diferència d'altres propostes, en aquesta no és indiferent el camí escollit de l'arbre per arribar al node final; al contrari, es busca que el camí de l'arbre passi per nodes interns que tinguin paràmetres d'eficiència elevats.

Tota aquesta explicació porta a definir els motius principals pels quals s'ha escollit aquesta proposta de solució i no una altra. En primer lloc, s'ha utilitzat un mètode d'aprenentatge basat en regles perquè la base de dades de relació entre símptomes i patologies anirà creixent i caldrà refer l'arbre de decisió a mesura que hi hagi més exemples per a discriminar. A més a més, el mètode d'aprenentatge basat en regles permet reorientar les prediccions depenent de la tendència del conjunt de regles, de manera que es pot construir una estructura totalment diferent depenent del conjunt de regles, que en aquest cas correspondrien a exemples de relacions entre un conjunt de símptomes i una patologia.

D'altra banda, i d'entre tots els mètodes d'aprenentatge basats en regles (i fins i tot els que no estan basats en regles), s'ha escollit el mètode d'arbres de decisió per un motiu força evident. La majoria de sistemes intel·ligents que s'han comparat són capaços de predir el resultat aprenent dels diferents exemples del conjunt de dades, i també donen un resultat més o menys fiable. Ara bé, tal com s'ha comentat, en aquest cas no interessa només el resultat, és a dir, no interessa només saber quina patologia pateix el pacient, sinó que sobretot interessa saber quin camí d'exploracions òptimes ha de seguir el metge per a esbrinar la patologia del pacient. Per tant, un dels mètodes que millor il·lustra el camí escollit per a arribar al resultat final és, sens dubte, el mètode dels arbres de decisió. Aquesta és la principal raó, doncs, per la qual s'ha escollit aquest mètode i no un altre.

Finalment, i pel que fa als resultats que es podran obtenir mitjançant la proposta de solució escollida, és evident que es podran obtenir els resultats que interessen al projecte; és a dir, es podrà obtenir la patologia predita pel mètode després de discriminar entre els diferents exemples del conjunt de dades, alhora que es podrà determinar fàcilment el camí que s'ha escollit per arribar al resultat final. A més a més, el resultat es podrà representar en forma d'arbre de tal manera que el professional de l'àmbit mèdic que l'utilitzi serà molt més capaç de comprendre'l, si fos necessari, i de seguir el camí adequat depenent dels símptomes i dels paràmetres d'eficiència associats a cada exploració. Els resultats són fàcilment comparables amb altres resultats obtinguts, i és senzill observar-ne les diferències. Evidentment, la representació gràfica de l'arbre no serà útil en si mateixa per a esbrinar el camí òptim del diagnòstic, ja que l'arbre pot arribar a prendre una mida considerable. Per aquest motiu existeixen eines d'intel·ligència artificial que són capaces de trobar el camí adequat a partir d'un exemple nou.

4.2. Base de dades

Com a qualsevol mètode d'aprenentatge basat en regles, els arbres de decisió requereixen un conjunt de dades d'exemples a partir dels quals elaborar un algorisme que permeti predir exemples nous.

Cadascun d'aquests exemples ha de seguir la mateixa estructura de tal manera que el mètode d'aprenentatge sigui capaç de reconèixer-los i utilitzar-los per a fer millors prediccions. A continuació, doncs, s'explicarà quina ha de ser aquesta estructura, així com també s'explicaran els motius per a crear una base de dades de prova, el seu contingut i el manteniment de la base de dades final una vegada es posi en funcionament la solució proposada en aquest projecte. En definitiva, s'explicarà el funcionament de la base de dades que servirà per a poder predir amb més exactitud nous casos a mesura que sigui més extensa.

4.2.1. Estructura de la base de dades

Abans d'explicar d'on s'obtindran els exemples que poblaran la base de dades i com es mantindrà aquesta, és imprescindible explicar quines taules es guardaran, en quin format i el motiu pel qual es farà així. Tot això és important no tan sols per a comprendre millor el comportament de l'algorisme que es dissenyarà, sinó també per a facilitar la lectura de les dades per part d'aquest tot mantenint el mateix format per a cada exemple.

La primera taula que es descriurà és la més corrent en una problemàtica d'aquest tipus; és a dir, la que fa referència a la relació entre les diverses patologies i els diferents símptomes que les defineixen. En aquest sentit, l'estructura serà la següent: les columnes faran referència als diferents símptomes i les files faran referència a cadascun dels exemples relacionats amb una sola patologia.

Tenint en compte que els camins d'un arbre de decisió han d'arribar sempre a un node terminal, referent a una patologia, és important que l'estructura de cada exemple de la base de dades defineixi clarament quin camp descriu la patologia, és a dir, és necessari que cada exemple tingui un apartat on s'anomeni la classe o l'etiqueta a la qual pertany aquell exemple, entenent que les classes o etiquetes seran les diverses patologies tractades.

És possible tenir diverses files de la taula que facin referència a la mateixa patologia. Tot i que a priori sembli que una patologia sempre tindrà el mateix conjunt de símptomes, alguns d'aquests símptomes poden prendre valors d'un rang força ampli. Per exemple, si una patologia conté com a símptoma la febre, és possible que una fila de la taula relacioni una temperatura de 38 graus a aquesta patologia, de la mateixa manera que una altra fila pot relacionar amb la mateixa patologia una temperatura de 39 graus; en els dos casos es considera que hi ha febre, però les mesures poden ser diferents. Això constituiria dues files de la base de dades que farien referència a la mateixa patologia.

La Taula 3 mostra un exemple de l'estructura d'aquesta taula de la base de dades. Es pot observar com la primera columna fa referència a les diverses patologies (classes o etiquetes de cada exemple), mentre que la resta de columnes fan referència als diferents símptomes. Cada fila, per la seva banda, correspon a un exemple diferent. També es pot observar que en aquest cas només es tenen en compte dues patologies diferents, i que cada

símptoma pot tenir valors de diferents tipus, és a dir, que el Símptoma 1 pot tenir valors de tipus numèric, per exemple, mentre que el Símptoma 2 pot tenir valors de tipus nominal; això sí, un mateix símptoma sempre presentarà valors del mateix tipus a tots els exemples.

Patologia	Símptoma 1	Símptoma 2	Símptoma 3	Símptoma 4	Símptoma 5
Patologia 1	Valor S1	Valor S2	Valor S3	Valor S4	Valor S5
Patologia 1	Valor S1	Valor S2	Valor S3	Valor S4	Valor S5
Patologia 1	Valor S1	Valor S2	Valor S3	Valor S4	Valor S5
Patologia 2	Valor S1	Valor S2	Valor S3	Valor S4	Valor S5
Patologia 2	Valor S1	Valor S2	Valor S3	Valor S4	Valor S5

Taula 3. Estructura de la taula de relació entre símptomes i patologies.

Fins aquí l'estructura de la base de dades és senzilla i, fins i tot, típica per a un problema d'aquest tipus. Tenint en compte la nova aportació que es fa en aquest projecte per a trobar el diagnòstic òptim, cal definir altres taules que aportin informació nova que pugui modificar els resultats típics obtinguts dels arbres de decisió. Així, la segona taula que es descriurà fa referència als paràmetres d'eficiència que s'associen a cadascuna de les exploracions. Com ja s'ha explicat, cada símptoma està relacionat amb una exploració, que a la vegada està relacionada amb diferents paràmetres que avaluen la seva eficiència, com ara el temps o el cost, entre d'altres. L'estructura d'aquesta taula, doncs, serà la següent: les columnes faran referència als diferents paràmetres d'eficiència i les files faran referència a les diferents exploracions que es tinguin en compte.

En aquest cas cada fila farà referència a una sola exploració, i no hi podran haver dues files que facin referència a la mateixa exploració. D'altra banda, cada paràmetre d'eficiència pot prendre un rang de valors diferent, però un mateix paràmetre sempre prendrà un valor del mateix rang. Per exemple, el paràmetre "cost" podrà prendre un rang de valors que siguin iguals o superiors a 0 euros, mentre que el paràmetre "temps" podria prendre un rang de valors que siguin iguals o superiors a 0 dies; és a dir, el tipus dels valors pot ser diferent per cada paràmetre, però mai per al mateix. A diferència de la taula anterior, en aquest cas el tipus dels valors haurà de ser sempre numèric. Això és així perquè l'algorisme que es dissenyarà haurà d'interpretar si el valor donat a una exploració per a un paràmetre concret és alt o baix, i ampliar o reduir la importància d'aquella exploració en funció d'aquest valor. De fet, l'única opció que hi hauria per a permetre tipus nominals seria el fet d'utilitzar valors d'escala ("baix", "mitjà" i "alt", per exemple), que al cap i a la fi caldria transformar en valors numèrics ("1", "2" i "3", per exemple).

La Taula 4 mostra un exemple de l'estructura d'aquesta segona taula de la base de dades. Es pot observar com la primera columna fa referència a les diverses exploracions, mentre que la resta de columnes fan referència als diferents paràmetres d'eficiència. Cada fila, per la seva banda, correspon a un únic conjunt de paràmetres d'eficiència per a una exploració concreta. També es pot observar que no es repeteix cap exploració (hi hauria duplicitat d'informació), i que cada paràmetre d'eficiència pot tenir valors de diferents mesures (euros, dies, valors d'escala, etc.).

	Cost	Temps	Invasió	Fiabilitat
Exploració 1	Valor en euros	Valor en dies	Valor d'1 a 5	Valor d'1 a 5
Exploració 2	Valor en euros	Valor en dies	Valor d'1 a 5	Valor d'1 a 5
Exploració 3	Valor en euros	Valor en dies	Valor d'1 a 5	Valor d'1 a 5
Exploració 4	Valor en euros	Valor en dies	Valor d'1 a 5	Valor d'1 a 5
Exploració 5	Valor en euros	Valor en dies	Valor d'1 a 5	Valor d'1 a 5

Taula 4. Estructura de la taula de relació entre exploracions i paràmetres d'eficiència.

Per acabar, i per afegir més flexibilitat a la solució escollida, es guardarà una tercera taula que fa referència a la importància que es dóna a cadascun dels paràmetres d'eficiència definits. Tal com s'havia definit fins ara la base de dades, cada paràmetre d'eficiència gaudia de la mateixa importància i podia influir de la mateixa manera al resultat final. Això, òbviament, pot no ser desitjable, i per aquest motiu cal definir una taula que s'encarregui de determinar quina prioritat té cadascun dels paràmetres d'eficiència respecte als altres. L'estructura d'aquesta taula serà la següent: la primera columna farà referència als diferents paràmetres d'eficiència o criteris, la segona columna farà referència al factor d'importància associat a un criteri determinat, i la tercera columna indicarà si el criteri és més eficient quan el seu valor és baix o alt, mentre que les files faran referència als diferents paràmetres d'eficiència definits.

De la mateixa manera que abans, cada fila farà referència a un sol paràmetre d'eficiència, i no hi podran haver dues files que facin referència al mateix paràmetre. D'altra banda, el factor d'importància haurà d'estar comprès sempre entre 0 i 1, per evitar problemes d'escala. L'última columna és important per saber com tractar el paràmetre d'eficiència a l'algorisme que es dissenyarà. Per exemple, la fiabilitat és millor quan és alta, de manera que l'algorisme haurà d'incrementar la influència de la fiabilitat quan el factor d'importància sigui alt tot ponderant el resultat final pel valor de la fiabilitat; en canvi, el cost és millor quan és baix, de manera que l'algorisme també haurà d'incrementar la influència del cost quan el factor d'importància sigui alt, però en aquest cas no ho podrà fer ponderant el resultat final pel valor del cost, ja que si el cost és molt baix reduiria el valor del resultat final.

Per tant, cal avisar a l'algorisme d'aquest fet per tal que respongui adequadament; si considerem que el valor del criteri pot prendre un valor entre 0 i 1, l'algorisme pot utilitzar directament el valor quan el criteri sigui millor quan és alt, mentre que caldrà restar el valor d'1 (el màxim valor, en aquest cas) quan el criteri sigui millor quan és baix. Els valors que pot prendre l'última columna poden ser "alt" o "baix", fent referència a quan és millor el criteri, quan el seu valor és alt o quan és baix.

La Taula 5 mostra un exemple de l'estructura d'aquesta tercera taula de la base de dades. Es pot observar com la primera columna fa referència als diversos criteris (paràmetres d'eficiència), la segona al factor d'importància i la tercera a la referència sobre quan és millor cada criteri. Cada fila, per la seva banda, correspon a un únic criteri. També es pot observar que no es repeteix cap criteri (hi hauria duplicitat d'informació), i que el factor d'importància sempre està comprès entre 0 i 1 i la referència sobre quan és millor cada criteri només pot contenir els valors "alt" o "baix".

Criteri	Factor	Millor quan
Cost	Valor dins el rang [0..1]	"alt" o "baix"
Temps	Valor dins el rang [0..1]	"alt" o "baix"
Invasió	Valor dins el rang [0..1]	"alt" o "baix"
Fiabilitat	Valor dins el rang [0..1]	"alt" o "baix"

Taula 5. Estructura de la taula de criteris, importància i referències sobre quan són millors.

En aquests moments, doncs, es pot dir que ja s'ha descrit l'estructura de la base de dades que s'utilitzarà. La primera taula que s'ha descrit és la més susceptible de patir variacions, ja que per a cada nou exemple caldrà afegir una nova fila. La segona taula caldrà modificar-la quan es registrin noves exploracions que no s'havien tingut en compte fins ara, mentre que la tercera taula s'haurà de modificar quan es vulguin tenir en compte nous paràmetres d'eficiència. Per tant, es preveu que a llarg termini les dues últimes taules variïn poc o, almenys, amb poca freqüència.

4.2.2. Base de dades de prova

Per a fer les primeres proves amb els nous mètodes de discriminació d'arbres de decisió que es dissenyaran cal crear una base de dades de símptomes i exploracions relacionades amb cada patologia. Tot seguit s'explicarà quina serà aquesta base de dades que haurà de servir per a obtenir els resultats del projecte, i els motius pels quals s'utilitzarà aquesta i no una base de dades final i completa.

La base de dades de prova no ha de ser, en un principi, forçosament real; és a dir, no ha d'ajustar-se necessàriament a la realitat. El més important és que aquesta base de dades de prova intenti contenir la màxima varietat d'exemples possible, sobretot aquells que poden comportar més problemes a l'hora d'elaborar un diagnòstic. D'aquesta manera provem la robustesa de l'algorisme i evitem errors futurs.

Aquesta base de dades també ha de contenir atributs de diferents tipus per demostrar que l'arbre de decisió es pot construir igualment. Per exemple, ha de contenir atributs de tipus nominal, com ara l'existència d'un determinat patró en una radiografia, que podria prendre els valors "sí" o "no"; de tipus numèric, com ara la febre, que podria prendre els valors numèrics equivalents als graus de temperatura del pacient; de valors d'escala, com ara la percepció d'un pacient sobre un determinat dolor, que podria prendre els valors "baix", "mitjà" o "alt", o simplement els valors "1", "2" o "3"; etc.

També cal tenir en compte que totes les patologies han d'estar relacionades amb tots els símptomes o exploracions, perquè la generació d'arbres de decisió amb valors absents, si bé és possible [Kusters et al. 2009], implica que el resultat i les prediccions siguin confuses, sobretot quan manquen tots els valors d'un símptoma concret per a una patologia determinada. És a dir, l'arbre de decisió que es genera presenta algunes branques amb incògnites, sense determinar quin valor ha de prendre l'atribut per escollir una branca o una altra. Això, evidentment, no serveix per a elaborar un diagnòstic òptim.

Per tant, el grup de patologies que englobi la base de dades haurà de tenir certes característiques en comú i poder-se definir amb el mateix conjunt de símptomes. D'aquesta manera s'aconsegueix un diagnòstic més específic segons el tipus de patologia que es vulgui predir. Si s'intentés obtenir un diagnòstic específic d'una base de dades que inclogués totes les possibles patologies, caldria gestionar aquells símptomes que no estan relacionats amb cap patologia; per exemple, no cal fer cap anàlisi d'orina si un pacient manifesta mal de cap. En aquest sentit, en haver-hi tipus tan diferents de patologies dins la mateixa base de dades, segurament la predicció de la patologia seria menys precisa, i probablement els arbres de decisió no serien la millor estratègia. La base de dades de prova, doncs, haurà d'incloure aquells símptomes o exploracions que realment aportin informació sobre el diagnòstic de les patologies.

També és cert que hi ha la possibilitat que manquin els valors d'alguns símptomes en alguns exemples, no perquè no tinguin relació amb la patologia assignada a l'exemple, sinó perquè accidentalment s'han perdut o no s'han calculat. En aquests casos cal establir una estratègia per a gestionar aquests valors absents [Saar-Tsechansky i Provost 2007], tenint en compte que els arbres de decisió agraeixen disposar de tots els valors.

La manera més simple de tractar aquests valors absents és eliminar de la base de dades (almenys durant la generació dels arbres de decisió) aquells exemples que no tinguin tots els valors. Tot i això, hi ha una manera més elaborada que tracta de recuperar aquests valors absents a partir d'estimacions. En el nostre cas, es considera que és preferible la segona manera perquè el petit error que es pugui produir en l'estimació compensa en escreix la pèrdua d'informació que suposa eliminar un exemple (per un sol valor absent, es perdria la resta de valors que són correctes).

L'estratègia que s'ha escollit per a recuperar aquests valors absents depèn del tipus de valor que prengui cada símptoma, és a dir, depèn de si el tipus és nominal (o valor d'escala) o numèric:

- Si el tipus és nominal, primer de tot cal obtenir la distribució de probabilitat de cada possible valor dins el conjunt d'exemples que tinguin la mateixa patologia assignada. Així, per exemple, si en un exemple que té assignada la patologia referent a l'Alzheimer hi manca el valor d'un símptoma que correspon a observar un determinat patró en una tomografia computada (per exemple, observar una determinada part del cervell), i que pot prendre els valors "sí" o "no", caldrà prendre tots els exemples de la base de dades que tinguin assignada la patologia referent a l'Alzheimer i obtenir la freqüència de valors equivalents a "sí" i a "no" per a aquest símptoma concret. Si hi ha un 40% de valors equivalents a "sí" i un 60% de valors equivalents a "no", llavors caldrà obtenir un valor aleatori que tingui un 60% de possibilitats d'equivaldre a "no". Si el tipus de valor del símptoma és un valor d'escala ("baix", "mitjà" i "alt", o "1", "2" i "3", per exemple), es pot utilitzar la mateixa estratègia.

- Si el tipus és numèric, la distribució de probabilitat s'ha d'ajustar a una distribució estadística (per exemple, una distribució normal, una distribució exponencial, etc.). També s'utilitzarà només aquell conjunt d'exemples que tinguin assignada la mateixa patologia, i caldrà fer ús d'alguna eina estadística per a obtenir la distribució.

Tot i que, com s'ha dit, la base de dades de prova no ha de ser forçosament real, tampoc és aconsellable que els valors siguin totalment aleatoris. Per exemple, si una determinada patologia es diagnostica quan el pacient té febre alta, entre altres coses, la majoria d'exemples del conjunt de dades que tinguin com a classe aquesta patologia hauran de tenir valors força alts de l'atribut referent al símptoma "febre". Per tant, la base de dades de prova ha de reflectir, per a cada patologia, la uniformitat de valors per a determinats atributs; alguns altres poden presentar més varietat, de manera que aquests últims es puguin considerar atributs més irrellevants pel que fa a aquella patologia. D'aquesta manera el mètode de discriminació corresponent també podrà determinar quins atributs ajuden a decidir més clarament de quina patologia es tracta.

De moment, i tenint en compte que és una base de dades de prova, es guardarà en un fitxer CSV (*Comma-separated values*, o valors separats per comes), ja que l'eina que s'utilitzarà per fer les proves, l'eina RapidMiner, accepta fitxers d'aquest format per a importar-los com a bases de dades.

També es guardaran en aquest format dos fitxers més que seran importants per a calcular la importància de cada atribut, tal com s'ha explicat en l'apartat de l'estructura de la base de dades. El primer fitxer emmagatzemarà les diferents exploracions i els valors que prenen per a cadascuna d'aquestes els diferents criteris especificats, com ara el cost, el temps, la invasió i la fiabilitat. Així, per exemple, una entrada d'aquest fitxer podrà detallar que una radiografia costa 50€, que tarda una setmana, que és poc invasiva (2 sobre 5, per exemple) i que és mitjanament fiable (3 sobre 5, per exemple). Aquests valors s'utilitzaran per saber si un atribut, a part de la seva capacitat de discriminació de dades respecte als altres atributs, també surt a compte pel que fa a altres criteris.

El segon fitxer emmagatzemarà la importància que s'assigna a cada criteri i una referència per saber si el criteri és millor quan el seu valor és baix o alt. Amb això s'aconsegueix reduir l'impacte d'alguns criteris que poden no ser massa importants; per exemple, si la fiabilitat fos menys important que el cost, aquest últim podria tenir un factor d'importància equivalent a 1 i la fiabilitat un factor equivalent a 0.8. D'altra banda, la referència és important per modificar la fórmula del mètode de discriminació depenent de si el criteri és millor quan el seu valor és més baix (per exemple, el cost) o més alt (per exemple, la fiabilitat). En comptes de capgirar els valors, això es fa així per intentar mantenir al màxim la comprensió dels valors per a cada criteri, intentant incrementar la comoditat de l'usuari (podríem dir que un valor petit significa un cost baix, en euros, però una fiabilitat alta, en valors d'una escala de l'1 al 5, però això podria confondre l'usuari).

A la Taula 6 es pot veure un extracte de la base de dades de prova que s'ha creat, introduint tots els aspectes que s'acaben de comentar. La primera columna fa referència a les patologies, on "P1" significa "Patologia 1", "P2" significa "Patologia 2", i "P3" significa "Patologia 3". D'altra banda, la resta de columnes fan referència a nou símptomes diferents, on cadascun d'ells es representa pels símbols "S1", "S2", etc. Es poden observar valors numèrics i nominals, i alguns d'ells formen part de valors d'escala. A més a més, no hi ha valors absents perquè es considera que tots els símptomes tenen relació amb les patologies, evitant així problemes amb la generació dels arbres de decisió. Finalment, també es pot observar la coherència entre els valors d'algun símptoma concret per a una patologia determinada, mantenint valors semblants.

Patologia	S1	S2	S3	S4	S5	S6	S7	S8	S9
P1	35.0	no	1	baix	primer	4	no	no	1
P1	35.0	no	1	baix	primer	4	no	no	1
P1	35.6	no	1	baix	primer	3	no	no	1
P1	37.0	no	2	baix	primer	4	no	no	1
P1	37.0	no	2	baix	primer	4	no	no	1
P1	36.7	no	1	baix	primer	5	sí	sí	1
P1	35.0	no	2	baix	primer	4	sí	no	1
P1	38.3	no	1	moderat	primer	4	sí	no	1
P1	36.2	no	1	baix	primer	3	sí	no	1
P1	36.5	no	1	baix	primer	4	sí	no	1
P2	38.4	sí	6	baix	primer	1	no	no	4
P2	37.4	sí	5	baix	primer	1	sí	no	4
P2	36.9	sí	5	moderat	primer	2	sí	sí	5
P2	38.5	sí	5	considerable	primer	1	sí	no	4
P2	35.5	no	5	baix	primer	1	sí	no	5
P2	38.0	sí	3	baix	primer	1	sí	no	4
P2	38.5	sí	4	baix	primer	1	sí	sí	4
P2	38.1	sí	6	baix	primer	1	sí	no	5
P2	38.2	sí	5	baix	primer	2	sí	no	5
P2	38.8	no	7	moderat	primer	1	sí	no	4
P3	40.0	no	1	alt	tercer	1	no	no	1
P3	38.5	no	1	alt	tercer	1	no	no	1
P3	39.6	sí	1	considerable	tercer	1	no	no	2
P3	38.0	no	1	alt	tercer	1	no	no	1
P3	39.7	no	1	alt	segon	1	no	sí	2
P3	38.8	no	2	alt	tercer	1	sí	no	2
P3	37.3	sí	1	extrem	segon	2	no	no	3
P3	38.2	no	2	alt	segon	1	no	no	1
P3	38.4	no	1	considerable	tercer	1	no	no	2
P3	39.9	no	1	considerable	tercer	1	no	no	2

Taula 6. Extracte de la base de dades de prova.

De la mateixa manera, la Taula 7 i la Taula 8 mostren un exemple dels dos fitxers relacionats amb els criteris que s'han explicat abans, és a dir, el fitxer que desa les exploracions i els valors que pren cada criteri per a cadascuna d'aquestes, i el fitxer que desa la importància que l'usuari dóna a cada criteri juntament amb la referència indicant si el criteri és millor quan és alt o quan és baix.

	Cost	Temps	Invasió	Fiabilitat
S1	0.05	1	1	2
S2	10	1	2	3
S3	0	1	1	1
S4	2	1	2	2
S5	30	5	3	4
S6	0	1	1	1
S7	50	7	4	4
S8	130	14	5	5
S9	0	1	1	5

Taula 7. Exemple de fitxer que desa les exploracions i els valors per a cada criteri.

Criteri	Factor	Millor quan és...
Cost	0.5	baix
Temps	0.3	baix
Invasió	0.5	baix
Fiabilitat	0.6	alt

Taula 8. Exemple de fitxer que desa el factor d'importància dels criteris i indica quan és millor.

Finalment, després de crear la base de dades de prova, també serà necessari implementar algunes funcions específiques que llegeixin aquests dos últims fitxers i calculin, a través de la fórmula del mètode de discriminació que s'exposarà més endavant, el factor que cal multiplicar al benefici original de cada atribut (entenent benefici com la probabilitat d'un atribut o símptoma de ser escollit per a ser representat primer a l'arbre de decisió) per a obtenir el benefici final del nostre algorisme.

4.2.3. Manteniment de la base de dades

Tot i haver creat la base de dades de prova i haver explicat els motius pels quals és necessària a l'inici del projecte, l'objectiu a llarg termini és, evidentment, que la base de dades que s'utilitzi estigui composta d'exemples reals i que vagi creixent a mesura que els professionals de l'àmbit mèdic vagin diagnosticant nous casos. Per aquest motiu, doncs, cobra importància el fet de mantenir adequadament aquesta base de dades de tal manera que no perjudiqui greument el rendiment i que sigui fàcil i còmode de mantenir pels usuaris.

Com s'ha explicat en l'apartat de l'estructura de la base de dades, la taula que és més susceptible de canviar freqüentment és la que relaciona les diferents patologies amb els seus símptomes, ja que la idea és que es vagin afegint nous exemples constantment per a millorar la predicció de l'arbre de decisió que es generi. Les altres dues taules, si bé poden patir canvis al principi (sobretot la que relaciona les diferents exploracions amb els criteris d'eficiència a mesura que apareguin nous símptomes o exploracions), a llarg termini es

preveu que es modifiquin poc sovint. Per tant, l'estudi del manteniment s'ha de fer sobretot sobre la primera taula, que serà també la que contindrà més informació.

Per començar, i per facilitar la comprensió dels usuaris, cal que les taules siguin entenedores i que cada camp estigui ben especificat. En el cas de la base de dades de prova que s'ha explicat a l'apartat anterior, això ja s'ha tingut en compte i s'ha intentat que qualsevol usuari pugui comprendre la taula sense gaire esforç. Si s'aconsegueix això, doncs, és molt més fàcil per a un usuari introduir nous exemples a la base de dades, o molt més fàcil per a algun programa d'interpretar la taula.

D'altra banda, i per garantir el bon rendiment de la base de dades a mesura que creixi, també caldria estudiar algun sistema d'emmagatzematge prou potent. En aquests moments, i tal com s'ha explicat, la base de dades es guarda en el format CSV perquè és un dels formats acceptats per l'eina que s'utilitza per fer les proves (RapidMiner) i perquè, de moment, les dades que comprèn la base de dades de prova es poden considerar escasses (tot i que suficients per fer les proves). El sistema d'emmagatzematge per a la base de dades final, doncs, dependrà de l'entorn on es faci servir l'aplicació final (per exemple, es podria fer ús de sistemes com ara MySQL).

Deixant de banda el rendiment i la comoditat de l'usuari, també cal explicar com podria ser el procediment per a mantenir la base de dades. És lògic que en un principi la base de dades real s'inicïi omplint-se d'exemples reals extrets a partir d'experiències de professionals de l'àmbit mèdic o de bases de dades ja existents. En aquest cas, es tracta de fer una injecció de dades inicial que pobli la base de dades amb uns quants exemples que siguin suficients per fer funcionar l'aplicació final. Ara bé, a partir d'aquí cal establir algun procediment que permeti incrementar el volum de la base de dades de mica en mica a mesura que es van afegint nous exemples.

Un possible procediment podria consistir en fer injeccions periòdiques de conjunts de dades a la base de dades a partir d'experiències de professionals de l'àmbit mèdic. Aquest procediment, però, requereix que els professionals dediquin una part del seu temps a augmentar el volum de la base de dades i que recordin algunes experiències passades.

Un altre procediment possible, i probablement el millor, seria que en el mateix moment d'elaborar un diagnòstic amb l'aplicació final, la mateixa aplicació registrés a la base de dades aquest nou cas amb la patologia final detectada. És a dir, quan un professional utilitzés l'aplicació per a fer un diagnòstic (a través de l'arbre de decisió generat amb les dades actuals), la mateixa aplicació prepararia un exemple nou de la base de dades recollint tots els símptomes que introdueixi el professional mèdic per a fer la predicció de la patologia. Amb aquesta informació, l'únic que caldria contrastar finalment seria que la patologia predita sigui finalment la patologia que pateix el pacient. Aquesta confirmació es podria fer en una segona etapa, i la mateixa aplicació podria guardar els diferents exemples que resten a l'espera de confirmar la patologia. Un cop sabut tot això, la mateixa aplicació s'encarregaria de guardar un nou exemple a la base de dades i contribuir així a millors prediccions futures. D'aquesta manera s'aconsegueix que el professional de l'àmbit mèdic no dediqui més temps a introduir nous exemples (s'aprofita el mateix temps que dedica per a fer la predicció del diagnòstic) i que no hagi de recordar els diagnòstics que ja ha fet. De totes maneres, es podria mantenir l'opció de modificar la base de dades manualment.

4.3. Arbres de decisió

Per entendre la proposta de solució que es presenta en aquest projecte per a resoldre el problema plantejat, primer cal entendre sobre quina base es treballarà. Tal com s'ha explicat a la proposta de solució, per a resoldre el problema s'utilitzaran arbres de decisió modificats convenientment per a satisfer les condicions del problema. Per tant, i per a situarnos en el context, en primer lloc cal explicar de manera breu el funcionament dels arbres de decisió i com es poden aplicar al problema que s'intenta resoldre. Més endavant s'explicaran les diferents variants que es tindran en compte en el desenvolupament d'aquest projecte.

Un arbre de decisió és un graf o un model en forma d'arbre. Es podria dir que més aviat és un arbre invertit perquè té l'arrel a dalt de tot i creix cap a baix. Comparada amb altres tipologies, aquesta representació de les dades té l'avantatge que manté plenament el significat i que és fàcil d'interpretar.

L'objectiu dels arbres de decisió és crear un model de classificació que predigui el valor d'un "atribut objectiu" (sovint anomenat "classe" o "etiqueta") a partir de diversos atributs d'entrada d'un conjunt d'exemples. Cadascun dels nodes interns d'un arbre de decisió correspon a un dels atributs d'entrada, que poden ser de tipus nominal o de tipus numèric. El nombre de branques d'un node interior nominal és equivalent al nombre de possibles valors del corresponent atribut d'entrada, mentre que les branques que surten d'un node interior numèric s'etiqueten amb rangs disjunts (per exemple, una branca podria contenir els valors més petits que 5 i una segona branca la resta de valors, és a dir, els iguals o superiors a 5). D'altra banda, cadascun dels nodes terminals o fulles representen un valor de l'atribut objectiu (classe o etiqueta), al qual s'arriba a través dels valors dels atributs d'entrada representats pel camí des de l'arrel fins a la fulla.

La Figura 2 mostra un exemple d'un arbre de decisió on s'intenta predir, a partir de diferents paràmetres meteorològics, si és possible o no jugar a l'esport del golf.

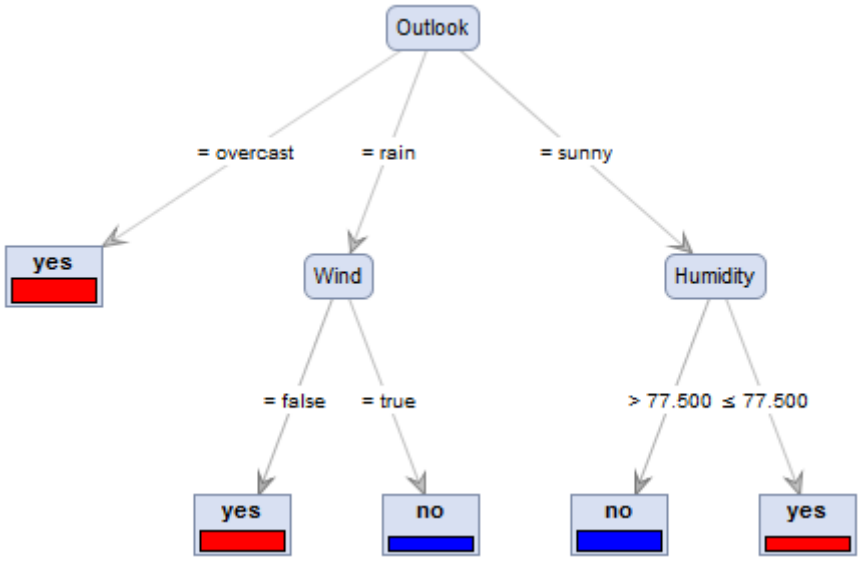

Figura 2. Exemple d'arbre de decisió.

Com es pot observar, hi ha diversos camins possibles depenent dels valors dels atributs d'entrada, però en qualsevol cas sempre s'arriba a un node terminal que indica si és possible jugar al golf o no.

Per entendre una mica més el funcionament dels arbres de decisió, cal explicar també quin procediment se segueix per a construir-los. Els arbres de decisió es generen a partir de particions recursives, és a dir, partint de manera repetida els valors dels atributs. A cada iteració l'algorisme segueix els següents passos:

- Se selecciona un atribut *A* per a partir-se. L'elecció de bons atributs per a fer les particions a cada etapa és crucial per a generar un arbre útil, i la manera d'escollir aquests atributs depèn del mètode de discriminació d'atributs que es prefereixi i que es detallaran més endavant.

- Els exemples del conjunt d'exemples s'ordenen en subconjunts, un per a cada valor de l'atribut *A* en el cas d'un atribut nominal. En el cas d'atributs numèrics, els subconjunts es formen per rangs disjunts de valors de l'atribut.

- Es retorna un arbre amb una branca per a cada subconjunt. Cada branca té un "sub-arbre" descendent o un valor de l'atribut objectiu (classe o etiqueta) que s'ha generat aplicant el mateix algorisme de forma recursiva.

Generalment, les iteracions s'aturen quan tots els exemples tenen el mateix valor d'etiqueta, és a dir, quan el subconjunt és pur. Les iteracions també es poden aturar si la majoria dels exemples tenen el mateix valor d'etiqueta. Aquest últim cas és una generalització del primer cas amb un llindar d'error. Tanmateix, hi ha altres condicions que poden fer aturar les iteracions, com ara:

- Que hi hagi menys d'un nombre determinat d'exemples en el subconjunt actual, on aquest nombre determinat correspondria a un paràmetre ajustable.

- Que no hi hagi cap atribut que assoleixi un determinat llindar de discriminació, és a dir, que cap atribut sigui prou bo, on aquest llindar també correspondria a un paràmetre ajustable.

- Que s'assoleixi la màxima profunditat de l'arbre indicada (nombre de nivells de l'arbre, és a dir, nombre de subconjunts encadenats), on aquesta profunditat també correspondria a un paràmetre ajustable.

D'altra banda, existeix una tècnica anomenada "poda d'arbres" l'ús de la qual implica que els nodes terminals que no tenen influència en el poder de discriminació de l'arbre de decisió s'eliminen. Això es fa per convertir un arbre massa específic o massa carregat en un arbre amb una forma més general que permeti millorar el seu poder de predicció en conjunts de dades nous. Altres tècniques més específiques són la "prepoda d'arbres", que és un tipus de poda que es realitza de forma paral·lela al procés de creació de l'arbre, i la "postpoda d'arbres", que es fa després que el procés de creació de l'arbre s'hagi completat.

Fins aquí, doncs, queda clar com es construeix un arbre i com es pot utilitzar per a fer prediccions. A més a més, també queda clara l'aplicació dels arbres de decisió al problema que es planteja en aquest projecte, ja que els nodes interiors de l'arbre es corresponen a la

perfecció als diferents símptomes que pot tenir un pacient, i els nodes terminals es corresponen a les diferents patologies que es poden diagnosticar. D'altra banda, i tenint en compte que el que ens interessa sobretot en aquest projecte és quina seqüència d'exploracions és l'òptima a l'hora d'elaborar un diagnòstic, també és evident que els arbres de decisió permeten de forma molt senzilla esbrinar quin camí cal seguir per arribar al node terminal corresponent, coincidint els nodes interiors amb cadascuna de les exploracions de la seqüència òptima predita. Ara bé, el que diferencia un arbre de decisió d'un altre és, bàsicament, el mètode de discriminació d'atributs que utilitza, és a dir, el criteri que fa servir per a escollir un atribut o un altre.

L'objectiu del nostre projecte és millorar els mètodes de discriminació d'atributs dels arbres de decisió per tal que contemplin el procediment òptim per a l'elaboració de diagnòstics. Per tant, no es tracta de crear un nou mètode de discriminació d'atributs, sinó de modificar-ne alguns dels existents per a què tinguin en compte la importància de cada atribut a partir de factors externs i subjectius.

Els mètodes de discriminació d'atributs seleccionen aquell atribut que cal comprovar abans en un arbre de decisió basant-se en el benefici que se'n traurà. És a dir, si en un conjunt de dades existeixen tres atributs, per exemple, un mètode de discriminació estudiarà quin d'aquests atributs pot separar millor les dades i descartar-ne més. Aquest atribut, doncs, serà l'arrel de l'arbre i serà el primer que es comprovarà. A continuació es farà el següent pels altres dos atributs de manera que l'arbre es construirà seguint un camí d'atributs ordenats de més discriminadors a menys discriminadors. El que varia d'un mètode de discriminació a un altre, per tant, és la manera de considerar quin dels atributs separa millor les dades.

En aquest sentit, el primer pas que cal fer és cercar quins mètodes habituals de discriminació d'atributs existeixen i estudiar-ne el seu funcionament. En aquest projecte se n'analitzaran quatre: per guany d'informació, per ràtio de guany, per índex de Gini i per exactitud.

4.3.1. Guany d'informació

Aquest mètode de discriminació es basa en el càlcul de l'entropia de tots els atributs. Llavors, se selecciona l'atribut amb l'entropia més baixa i s'afegeix a l'arbre de decisió, de manera que els primers atributs que es comproven a l'arbre de decisió són sempre aquells que tenen l'entropia més baixa, ja que equivalen a aquells atributs que més bé separen les dades.

De fet, si X correspon a un conjunt d'exemples i a a un atribut concret, la fórmula del guany d'informació per a aquest atribut és la que es troba representada a l'Equació 1, on $H()$ és la funció que calcula l'entropia.

$$GI(X, a) = H(X) - H(X|a)$$

Equació 1. Definició del guany d'informació.

Si es desenvolupa una mica més, es transforma en la fórmula representada a l'Equació 2.

$$GI(X,a) = H(X) - \sum_{v \in valors(a)} \left(\frac{|\{x \in X | valor(x,a) = v\}|}{|X|} \cdot H(\{x \in X | valor(x,a) = v\}) \right)$$

Equació 2. Definició desenvolupada del guany d'informació.

Això indica que el guany d'informació és la diferència d'entropia entre el conjunt d'exemples original i el conjunt d'exemples coneixent l'atribut a. És a dir, que es preferiran aquells atributs que maximitzin el guany d'informació perquè significarà que, coneixent aquest atribut, l'entropia del conjunt d'exemples es redueix significativament, i això significa a la vegada que hi ha menys incertesa i que les dades queden més ben classificades.

Aquest mètode, però, té un biaix que fa que se seleccionin aquells atributs que poden prendre un nombre més elevat de valors diferents, ja que l'entropia per a aquests sempre serà més baixa perquè es produeix menys incertesa (quan es poden prendre més valors, les dades queden més ben separades al llarg de tots aquests valors; en canvi, quan se'n poden prendre menys, les dades es confonen més).

4.3.2. Ràtio de guany

Es tracta d'una variant del mètode de discriminació per guany d'informació exposat abans. En aquest cas se soluciona el problema del biaix ajustant el guany d'informació per a cada atribut per permetre l'amplitud i la uniformitat dels valors dels atributs [Harris 2002].

Això s'aconsegueix calculant el valor intrínsec d'un atribut en el conjunt d'exemples, i dividint el guany d'informació per aquest valor. El valor intrínsec fa referència a la quantitat d'informació que es necessita per a explicar a quina branca de l'arbre correspon un exemple, i es pot resumir amb la fórmula representada a l'Equació 3.

$$VI(X,a) = -\sum \frac{|X_i|}{|X|} \cdot \log_2 \frac{|X_i|}{|X|}$$

Equació 3. Definició del valor intrínsec.

Si es desenvolupa una mica més, es transforma en la fórmula representada en l'Equació 4.

$$VI(X,a) = -\sum_{v \in valors(a)} \frac{|\{x \in X | valor(x,a) = v\}|}{|X|} \cdot \log_2 \left(\frac{|\{x \in X | valor(x,a) = v\}|}{|X|} \right)$$

Equació 4. Definició desenvolupada del valor intrínsec.

Així, el biaix del guany d'informació es pot corregir a través de la fórmula representada en l'Equació 5, que aplica la divisió entre el guany d'informació i el valor intrínsec.

$$RG(X, a) = \frac{GI(X, a)}{VI(X, a)}$$

Equació 5. Definició de la ràtio de guany.

De totes maneres, aquest ajust pot acabar conferint massa importància a aquells atributs que tenen un rang petit de valors, i això pot continuar sent igualment injust.

4.3.3. Índex de Gini

L'índex de Gini és, en realitat, una mesura de la impuresa d'un conjunt de dades. Així, quan se separa el conjunt de dades a partir d'un atribut escollit (aquell que tingui l'índex de Gini més baix) es redueix l'índex de Gini mitjà dels conjunts de dades restants.

La impuresa de Gini mesura la freqüència en què un exemple del conjunt de dades escollit aleatòriament es classificaria erròniament si fos classificat aleatòriament segons la distribució d'etiquetes del subconjunt de dades. La impuresa de Gini es pot calcular sumant la probabilitat que té cada exemple de ser escollit multiplicada per la probabilitat que es produeixi un error en classificar aquest exemple. La fórmula representada en l'Equació 6 resumeix aquesta mesura tenint en compte que f_i és la fracció d'exemples classificats amb valor i en el conjunt, i que n és el nombre de classes que pot prendre un exemple (en el nostre cas, el nombre de patologies).

$$Gini(X) = 1 - \sum_{i=1}^{n} {f_i}^2$$

Equació 6. Definició de l'índex de Gini.

Els millors atributs, és a dir, els que se seleccionaran abans, sempre seran aquells que tinguin un índex de Gini més baix, perquè significarà que les dades queden més ben classificades (menys impuresa).

Aquesta mesura sol comparar-se sovint amb la mesura del guany d'informació o de la ràtio de guany [Raileanu i Stoffel 2004].

4.3.4. Exactitud

Aquest mètode de discriminació intenta seleccionar un atribut per a separar les dades que maximitzi l'exactitud de l'arbre de decisió sencer. És a dir, es calcula l'exactitud per a cada valor de l'atribut, i se sumen tots els resultats, obtenint així l'exactitud de l'arbre de decisió sencer.

L'exactitud d'un valor, per la seva banda, correspon al nombre d'exemples que tenen aquest valor i que estan classificats a la classe majoritària dividit pel nombre total

d'exemples que tenen aquest valor. La fórmula representada a l'Equació 7 resumeix aquesta mesura.

$$Exactitud(X,a) = \sum_{v \in valors(a)} \frac{\max_{c \in classes(X)} |\{x \in X | valor(x,a) = v \wedge classe(x,a) = c\}|}{\sum_{c \in classes(X)} |\{x \in X | valor(x,a) = v\}|}$$

Equació 7. Definició de l'exactitud.

Els atributs que se seleccionaran abans seran aquells que impliquin una exactitud més gran de tot l'arbre de decisió.

4.4. Ponderació dels mètodes de discriminació

En aquest punt és on s'exposarà la innovació que es pretén inserir en els arbres de decisió per tal que ajudin a resoldre el problema específic que planteja el projecte, és a dir, la modificació dels arbres de decisió que caldrà dur a terme per a orientar el mètode als objectius del projecte.

Deixant de banda el comportament específic de cada mètode de discriminació, en aquest projecte ens interessa ponderar el benefici obtingut de cada mètode per a cada atribut amb un factor que representi la importància d'aquest atribut a partir de criteris com ara el cost de l'exploració, el temps que requereix, etc. Per a calcular aquest factor i multiplicar-lo al benefici original, cal construir una fórmula adient que permeti obtenir el benefici final per a un determinat atribut.

Aquesta fórmula es detalla a continuació, juntament amb un exemple pràctic que mostrarà l'aplicació de la fórmula en un context real, i també un fragment de pseudocodi que farà referència a les indicacions de programació necessàries per a fer els càlculs que requereix la fórmula. Amb tot això s'intentarà que la fórmula es comprengui.

4.4.1. Fórmula general

Després de la introducció feta, en aquest apartat es mostrarà la fórmula general per a obtenir el factor que representarà la importància de cada atribut respecte a criteris propis que s'hauran definit prèviament, a part de la pròpia importància que es desprendrà del mètode de discriminació propi dels arbres de decisió.

També s'explicaran cadascuna de les variables i les funcions que hi apareixen, i quina relació tenen entre elles.

Aquesta fórmula es desglossa a l'Equació 8, on un *atribut* fa referència a una determinada exploració (allò que cal fer per a esbrinar si un pacient té un símptoma concret o no) i un *criteri* fa referència a alguna de les variables que es tenen en compte per a valorar cada exploració, com podria ser el cost, el temps, la invasió o la fiabilitat de l'exploració. La primera fórmula és la general, i la resta són sub-fórmules que permeten entendre la fórmula principal.

$$\text{Benefici}(Atribut) = \text{Pes}(Atribut) * \text{Benefici\_original}(Atribut)$$

$$\text{Pes}(Atribut) = \frac{\text{Pes\_criteris}(Atribut)}{Num\_criteris}$$

$$\text{Pes\_criteris}(Atribut) = \sum_{c=1}^{Num\_criteris} (\text{Factor}(c) * \text{Pes\_total\_criteri}(Atribut, c))$$

$$\text{Pes\_total\_criteri}(Atribut, c) = 1 - \frac{\text{Pes\_criteri}(Atribut, c) - \text{Pes\_minim\_criteri}(c)}{\text{Pes\_maxim\_criteri}(c) - \text{Pes\_minim\_criteri}(c)}$$

Equació 8. Fórmula general per a la ponderació dels mètodes de discriminació.

Les diferents funcions i variables s'expliquen a continuació:

- ***Atribut***. És la variable que fa referència a un atribut del conjunt de dades. En el nostre cas, un atribut farà referència a alguna de les diferents exploracions o símptomes (per exemple: febre, part específica d'una analítica, ressonància, etc.).
- **Benefici**(***Atribut***). És la fórmula general que calcula el benefici per a un atribut concret. A partir dels resultats d'aquesta fórmula que s'obtinguin amb tots els atributs, es determinarà quin d'aquests atributs és l'òptim per aparèixer abans a l'arbre de decisió, és a dir, l'exploració que cal fer abans per seguir un procediment òptim.
- **Pes**(***Atribut***). És la funció que calcula el pes que caldrà multiplicar al benefici original d'un atribut concret, és a dir, la ponderació que modifica el benefici original segons els criteris que haguem definit (cost, temps, etc.) i la seva importància.
- **Benefici_original**(***Atribut***). És la funció que calcula el benefici original d'un atribut segons el tipus d'arbre de decisió que s'hagi escollit, és a dir, el benefici total que s'obtindria d'un atribut concret si no apliquéssim el pes obtingut dels criteris per al procediment òptim (cost, temps, etc.). Per exemple, el "benefici original" d'un atribut podria correspondre a l'entropia.
- **Pes_criteris**(***Atribut***). És la funció que calcula la suma dels pesos que s'obtenen per a cada criteri definit (cost, temps, etc.) per a un atribut concret.
- ***Num_criteris***. És la variable que fa referència al nombre total de criteris que haguem definit.
- ***c***. És la variable que fa referència a un criteri concret. Per exemple, $c = 1$ podria fer referència al cost, $c = 2$ podria fer referència al temps, etc.

- **Factor**($\boldsymbol{c}$). És la funció que ens retorna la importància assignada a un criteri concret. El valor retornat estarà comprès entre 0 i 1, on 0 significarà que el criteri no té cap importància per a escollir un atribut concret, i 1 significarà que té la màxima importància.

- **Pes_total_criteri**($\boldsymbol{Atribut, c}$). És la funció que ens retorna el pes total d'un criteri per a un atribut concret tenint en compte el conjunt de valors que pren el mateix criteri per a la resta d'atributs.

- **Pes_criteri**($\boldsymbol{Atribut, c}$). És la funció que ens retorna el valor assignat a un atribut per a un criteri concret. Per exemple, podria retornar un valor de 250€ per a l'atribut "ressonància" i per al criteri "cost".

- **Pes_minim_criteri**($\boldsymbol{c}$). És la funció que ens retorna el valor mínim assignat al conjunt d'atributs per a un criteri concret.

- **Pes_maxim_concret**($\boldsymbol{c}$). És la funció que ens retorna el valor màxim assignat al conjunt d'atributs per a un criteri concret.

Pel que fa a la funció **Pes_total_criteri**($\boldsymbol{Atribut, c}$), és convenient fer una consideració: la divisió final es resta d'1 només en el cas que **Pes_criteri**($\boldsymbol{Atribut, c}$) sigui millor com més baix sigui el seu valor (per exemple, el cost serà millor quan sigui més baix); en cas contrari, es pren únicament el resultat de la divisió, és a dir, no cal restar-la d'1 (per exemple, la fiabilitat serà millor quan sigui més alta).

4.4.2. Exemple d'aplicació de la fórmula

Per a representar millor la fórmula escollida, en aquest apartat es mostraran dues taules amb exploracions, criteris, valors i factors ficticis, i un exemple d'aplicació de la fórmula on es calcula el benefici per a un atribut concret.

Així, la Taula 9 mostra els valors assignats a cada criteri per a cada atribut o exploració existent en el conjunt de dades d'exemple.

VALORS CRITERIS – ATRIBUTS				
Explor./Criteris	**Cost**	**Temps**	**Invasió**	**Fiabilitat**
Exploració 1	0.05	1	1	2
Exploració 2	10	1	2	3
Exploració 3	0	1	1	1
Exploració 4	2	1	2	2
Exploració 5	30	5	3	4
Exploració 6	0	1	1	1
Exploració 7	50	7	4	4
Exploració 8	130	14	5	5

Taula 9. Valors dels criteris per a cada atribut o exploració.

En aquest cas, el cost es mesura en euros, el temps en dies, i la invasió i la fiabilitat en nombres d'un rang entre 1 i 5.

La Taula 10, per la seva banda, mostra els factors d'importància i les referències sobre quan són millors els criteris per a l'exemple que s'està desenvolupant.

FACTORS D'IMPORTÀNCIA DELS CRITERIS		
Criteri	**Factor**	**Millor quan és...**
Cost	0.5	baix
Temps	0.3	baix
Invasió	0.5	baix
Fiabilitat	0.6	alt

Taula 10. Factors d'importància dels criteris i referència sobre quan són millors.

A partir d'aquestes dades d'exemple, doncs, l'Equació 9 desenvolupa la fórmula per a calcular el benefici d'un atribut concret, en aquest cas de l'Exploració 5 (E5). Es pot observar la fórmula general i algunes sub-fórmules que deriven d'ella.

$$\text{Benefici(E5)} = \text{Pes(E5)} * \text{Benefici\_original(E5)}$$

$$\text{Pes(E5)} = \frac{\text{Pes\_criteris(E5)}}{4}$$

$$\text{Pes\_criteris(E5)} = \sum_{c=1}^{4} (\text{Factor}(c) * \text{Pes\_total\_criteri(E5}, c))$$

Equació 9. Desenvolupament de la fórmula per a calcular el benefici de l'Exploració 5.

Per acabar de calcular el pes que caldrà multiplicar al benefici original de l'Exploració 5, cal estendre el sumatori de l'última fórmula generada. Amb el resultat del sumatori es podrà calcular la suma dels pesos de cada criteri i es podrà aconseguir el pes mitjà dels criteris per a l'Exploració 5, que és el resultat que interessa. Aquests càlculs queden representats a l'Equació 10.

$$\text{Factor(cost)} = 0.5 \qquad \text{Factor(temps)} = 0.3$$
$$\text{Factor(invasió)} = 0.5 \qquad \text{Factor(fiabilitat)} = 0.6$$

$$\text{Pes\_total\_criteri(E5, cost)} = 1 - \frac{30 - 0}{130 - 0} = 0.769$$

$$\text{Pes\_total\_criteri}(\text{E5}, \text{temps}) = 1 - \frac{5-1}{14-1} = 0.692$$

$$\text{Pes\_total\_criteri}(\text{E5}, \text{invasió}) = 1 - \frac{3-1}{5-1} = 0.5$$

$$\text{Pes\_total\_criteri}(\text{E5}, \text{fiabilitat}) = \frac{4-1}{5-1} = 0.75$$

Equació 10. Extensió del sumatori per a calcular els pesos dels criteris per a l'Exploració 5.

Finalment, i després d'haver obtingut els pesos per a cada criteri, es pot recuperar la fórmula principal, calcular el pes mitjà dels criteris i obtenir el benefici final per a l'Exploració 5 en funció del benefici original que s'hagués obtingut aplicant el mètode de discriminació corresponent. L'Equació 11 acaba de desenvolupar aquests càlculs i ofereix la fórmula final que s'haurà d'aplicar per a obtenir el benefici de l'Exploració 5.

$$\text{Pes\_criteris}(\text{E5}) = 0.5 * 0.769 + 0.3 * 0.692 + 0.5 * 0.5 + 0.6 * 0.75 = 1.292$$

$$\text{Pes}(\text{E5}) = \frac{1.292}{4} = 0.323$$

$$\textbf{Benefici(E5)} = \mathbf{0.323} * \textbf{Benefici\_original(E5)}$$

Equació 11. Càlcul del benefici final per a l'Exploració 5.

El resultat final, doncs, i per a aquest exemple, correspon a multiplicar un pes de 0.323 al benefici original que s'obté aplicant el mecanisme propi de l'arbre de decisió escollit, és a dir, aplicant el mètode de discriminació corresponent. Evidentment, els resultats podran ser diferents depenent del mètode de discriminació escollit, però això no està relacionat amb els càlculs de la ponderació que s'ha proposat, que sempre seguirà la mateixa estructura i sempre obtindrà el mateix pes.

4.4.3. Pseudocodi de la fórmula

Després d'haver explicat la fórmula general de ponderació dels mètodes de discriminació, i després d'haver proposat un exemple per a comprendre'n el seu funcionament, en aquest últim apartat es vol presentar aquesta fórmula des del punt de vista de la programació, de manera que quedi representada d'una forma més esquemàtica.

Sovint, veure quin és el recorregut de cada variable i observar les diferents iteracions de l'algorisme pot ajudar a entendre el funcionament de la fórmula i l'obtenció del seu resul-

tat. Per aquest motiu, a l'Algorisme 1 es representa el pseudocodi d'aquesta fórmula tal com si es programés en una funció.

```
PER cada exploració "e" FER
   suma_pesos := 0

   PER cada criteri "c" FER
      SI c és millor quan és baix LLAVORS
         suma_pesos := suma_pesos + (Factor(c)
            * (1 – (Pes_criteri(e,c) – Pes_minim_criteri(c))
               / (Pes_maxim_criteri(c) – Pes_minim_criteri(c))))
      ALTRAMENT
         suma_pesos := suma_pesos + (Factor(c)
            * ((Pes_criteri(e,c) – Pes_minim_criteri(c))
               / (Pes_maxim_criteri(c) – Pes_minim_criteri(c))))
      FSI
   FPER

   pesos[e] := suma_pesos / num_criteris
FPER
```

Algorisme 1. Pseudocodi de la fórmula de ponderació dels mètodes de discriminació.

5. Implementació

Amb el problema plantejat i la proposta de solució definida, en aquest apartat s'exposaran els requeriments d'implementació necessaris per a desenvolupar el projecte, i també es deixarà constància dels problemes que haurà calgut resoldre.

El que s'ha explicat fins ara forma part sobretot de la vessant més teòrica de la solució que es proposa. A partir d'ara, doncs, cal implementar aquesta solució per a comprovar si els resultats que es preveuen són realment els que s'obtenen, i per a comprovar sobretot si són resultats útils per a resoldre el problema.

Així, es posarà en pràctica el que s'ha estudiat fins ara indicant quins són els passos previs a la implementació, les eines que caldrà utilitzar i els comentaris pertinents que caldrà destacar durant la implementació.

5.1. Requeriments de l'entorn

Abans de començar la implementació d'una solució, és necessari preveure què requerirà aquesta solució, qui hi intervindrà, a qui va dirigida, què ha de resoldre exactament i com ho ha de mostrar. A part de tot això, també és necessari preveure des del principi quin manteniment s'haurà de fer per a què la solució sigui escalable.

Per tant, caldrà tenir en compte quins usuaris o actors interactuaran amb aquesta solució, com s'aconseguirà que la solució ajudi a la predicció o al suport al diagnòstic, com es representaran els resultats que han d'ajudar als professionals de l'àmbit mèdic, i quin serà el procediment per a entrar noves dades que ajudin a millorar les prediccions.

5.1.1. Tipus d'usuaris o actors

Els actors que intervindran en una aplicació que implementi la solució que es proposa en aquest projecte són força evidents. Cal tenir en compte que l'aplicació la utilitzaran professionals de l'àmbit mèdic que, a partir d'uns paràmetres que ells mateixos introduiran, obtindran un resultat que es generarà a partir d'un càlcul fet pel propi sistema.

Observant tot això, es pot dir qui haurà fins a tres tipus d'actors diferenciats:

- El professional de l'àmbit mèdic que faci ús de la solució.
- El sistema, que s'encarregarà de traduir els paràmetres d'entrada en resultats.
- L'administrador, que seria l'encarregat d'actualitzar la solució si es requerís.

El primer actor és el més evident; els professionals de l'àmbit mèdic o, en qualsevol cas, l'usuari que utilitzi l'aplicació que implementa la solució, és l'actor més visible de tots tres. Aquest usuari serà l'encarregat d'interaccionar amb el sistema introduint dades desconegudes per aquest últim (per exemple, símptomes que pateixi un pacient, o nous valors dels paràmetres d'eficiència), i també serà l'encarregat d'interpretar el resultat que li oferirà el

sistema. No es diferencia entre altres tipus d'usuaris perquè tots tindran el mateix objectiu: seguir el procediment òptim per a diagnosticar una patologia. Per tant, tots els usuaris que facin ús d'aquesta solució per a aconseguir aquest procediment interactuaran de la mateixa manera amb el sistema, siguin professionals de l'àmbit mèdic o no.

El segon actor també és interessant destacar-lo, ja que juga un rol molt important en la interacció necessària per a aconseguir el resultat final. Després que l'usuari hagi introduït els paràmetres d'entrada, és el sistema qui s'encarrega de traduir aquestes dades i, a partir de l'arbre de decisió que genera el mateix sistema amb les dades ja existents, ofereix un resultat que prediu la patologia del pacient i el procediment òptim. Per tant, tota la part de predicció la fa el sistema, mentre que és l'usuari qui l'ajuda proporcionant-li dades.

El tercer actor simplement compleix les funcions de millorar o actualitzar el sistema amb noves dades o creant noves versions. És l'encarregat de desenvolupar-lo i mantenir-lo; tot i que les noves dades les aniran introduint els usuaris, l'administrador ha de ser qui vetlli perquè el sistema sigui capaç d'interpretar aquestes dades i oferir el resultat que els usuaris esperen. Si calgués millorar la fórmula de ponderació dels mètodes de discriminació, o simplement canviar o afegir nous mètodes de discriminació, seria l'administrador qui hauria de fer-ho.

5.1.2. Predicció o suport al diagnòstic

La solució que es proposa en aquest projecte sempre intentarà ajudar a la predicció de diagnòstics, és a dir, sempre donarà suport als professionals mèdics. Evidentment, l'última responsabilitat serà la d'aquests professionals, que seran qui finalment decideixin quin ha de ser el procediment per a diagnosticar una patologia. Per tant, cal deixar clar que la solució serà útil per a ajudar a decidir, no per decidir definitivament.

En aquest sentit, doncs, cal implementar la solució de tal manera que s'aportin noves dades o nous punts de vista que el professional de l'àmbit mèdic valori positivament. En aquest cas, és important que el sistema sigui capaç de predir quina patologia té el pacient, però aquesta informació pot ser poc útil per als professionals si no s'explica o no es té l'oportunitat d'esbrinar com s'ha arribat a aquesta conclusió. Per tant, i com s'ha anat dient fins ara, no només és important el resultat final, sinó el camí que s'ha seguit per a arribar-hi.

Mitjançant els arbres de decisió és fàcil saber com s'ha arribat a la predicció final, però un arbre molt complet també pot arribar a ser força complicat d'entendre. O sigui, pot ser senzill entendre com s'arriba al node terminal (patologia), però no el motiu pel qual hi ha uns determinats nodes interiors (exploracions) i no uns altres. Generalment, si la fórmula de ponderació funciona correctament, els primers nodes interiors que es trobaran coincidiran amb aquelles exploracions que siguin més eficients. Per tant, el professional de l'àmbit mèdic ja tindrà una primera referència per a entendre el motiu pel qual apareix primer una determinada exploració. Quan es comparin dues exploracions d'eficiència més o menys equivalent, llavors caldrà confiar amb el mètode de discriminació que utilitzi el sistema i creure que s'ha escollit una determinada exploració perquè separava millor el conjunt de dades.

Finalment, una aplicació que implementés la solució que es proposa en aquest projecte hauria de tenir en compte dues parts diferenciades:

- La predicció d'una patologia i l'elecció d'un procediment òptim a partir d'un exemple nou introduït per l'usuari.
- La inclusió d'aquest nou exemple al conjunt de dades del sistema per a generar un nou arbre de decisió més precís.

La primera part faria referència als resultats que espera obtenir l'usuari en entrar les dades d'un exemple nou, és a dir, les dades que demana el sistema per a seguir un procediment òptim i diagnosticar una patologia. Aquest és el resultat final que interessa al professional de l'àmbit mèdic, i l'objectiu final de la solució.

La segona part, en canvi, faria referència únicament al manteniment i a l'actualització del sistema. És una part també molt important per garantir que la precisió de les prediccions del sistema és cada vegada més gran, ja que disposa de més exemples reals.

Per tant, el sistema ha de ser capaç d'utilitzar un arbre de decisió que ell mateix ha generat per a predir una patologia a partir d'un conjunt de símptomes, però també ha de ser capaç de generar nous arbres de decisió a mesura que creixi el conjunt de dades amb nous exemples introduïts pels usuaris. El sistema pot guardar un únic arbre de decisió que correspongui a l'últim arbre generat amb les dades més actualitzades.

5.1.3. Representació dels resultats

Si la solució que es proposa en aquest projecte no fos una eina de suport al diagnòstic, és a dir, si fos una eina definitiva que els professionals de l'àmbit mèdic haguessin d'obeir, llavors els resultats que s'haurien d'obtenir serien únicament els resultats finals, és a dir, la patologia que s'ha predit. Com que aquest no és l'objectiu de la solució, però, i com que el que es proposa és que aquesta solució simplement ajudi als professionals, el més important serà donar nous punts de vista i nous arguments als professionals per tal que elaborin el seu diagnòstic.

Per tot això, és important saber quina ha de ser la representació dels resultats que es mostrin als professionals de l'àmbit mèdic, ja que aquesta representació ha d'ajudar a convèncer-los d'un determinat diagnòstic.

Existeixen molts altres mètodes d'aprenentatge basats en regles que podrien haver fet la mateixa funció que els arbres de decisió i obtenir resultats finals molt semblants. Tot i això, la majoria d'aquests mètodes tenen l'inconvenient que encapsulen el camí que han seguit per arribar a aquests resultats finals, i llavors és més complicat convèncer als professionals perquè hi manca transparència. Els arbres de decisió, en canvi, si bé pot ser difícil entendre com s'han generat (no pas més difícil que altres mètodes, tanmateix), sí que és cert que deixen molt clar el camí que s'ha seguit per arribar al resultat final. Per tant, costa d'imaginar algun altre mètode on aquest camí sigui més fàcil d'observar.

La representació dels resultats, doncs, que pot consistir en la representació de l'arbre i del camí que s'ha escollit per arribar a diagnosticar una determinada patologia, sembla molt adient per a donar nous arguments als professionals de l'àmbit mèdic.

A més a més, també existeixen diferents maneres de representar els arbres de decisió. Una d'aquestes maneres que pot ser força útil és la inclusió, a cada node terminal, del percentatge d'èxit d'aquella predicció. És a dir, segons el conjunt d'exemples de què es disposa, i agafant de referència tots aquells que es classificarien en un determinat node terminal, quants d'aquests últims serien classificats correctament i encertarien la predicció. Això pot ajudar al professional de l'àmbit mèdic a saber si un node terminal (en realitat, una determinada predicció) és estable o inestable; o sigui, si una predicció és més o menys segura.

5.1.4. Entrada de dades noves

Una de les claus del manteniment de l'aplicació que implementi la solució proposada és el fet que actualitzi les dades de què disposa cada vegada que s'introdueix un nou exemple per a fer una predicció. Això possibilita que el conjunt d'exemples vagi creixent i que, conseqüentment, les noves prediccions siguin cada vegada més precises.

L'alternativa a aquesta manera d'actualitzar la base de dades seria introduir nous conjunts d'exemples de forma periòdica, però és evident que, a part de ser menys eficient, requereix recordar els nous casos i tots els seus detalls.

Per tant, el que es proposa és que, quan un usuari utilitzi l'aplicació per a predir una patologia seguint un procediment òptim, el sistema guardi les dades que introduirà aquest usuari com a un nou exemple, almenys el conjunt de símptomes. És a dir, que el sistema guardi totes les dades menys la patologia que hagi predit el sistema, que encara no s'ha confirmat que sigui certa. Aquests exemples incomplets es podrien guardar temporalment fins que el professional de l'àmbit mèdic diagnostiqui finalment, a partir de totes les eines que tingui al seu abast, entre elles la mateixa predicció del sistema, la patologia que pateix el pacient. Un cop confirmada aquesta patologia, quan se sàpiga realment que és aquesta i no una altra, llavors seria el moment de completar l'exemple incomplet corresponent que haurà guardat la base de dades. En aquest moment, l'exemple passaria a formar part del conjunt d'exemples de tal manera que els nous arbres de decisió que es generessin serien més precisos (a més dades, més precisió).

No és necessari que es generi un nou arbre de decisió per a cada nou exemple introduït, ja que segurament la variació no serà significativa. De fet, a mesura que hi hagi més dades, un nou exemple cada vegada modificarà menys la precisió total de l'arbre de decisió. Una alternativa seria generar un nou arbre de decisió quan s'hagi introduït un determinat nombre d'exemples nous, o cada cert temps.

Es proposen aquestes alternatives perquè, a mesura que creixi el conjunt d'exemples, pot ser més costós generar l'arbre de decisió. Un dels inconvenients que té aquest mètode és que el temps que triga a generar-se l'arbre és proporcional al volum del conjunt de dades de què es disposi. Tot i això, no es considera un problema perquè la predicció, que és el que realment importarà en el moment d'introduir les dades, sí que es pot considerar instantània. La generació de l'arbre es pot fer periòdicament quan es cregui necessari.

5.2. Detalls d'implementació

Una vegada especificats els requeriments, cal escollir les eines que s'utilitzaran per a implementar la solució. En aquest apartat es detallarà l'entorn de desenvolupament utilitzat, així com d'altres eines, llibreries o complements que hagin ajudat a la implementació.

També es farà un recull dels problemes d'implementació que s'hagin trobat, o dels comentaris que es creguin pertinents per a fer entendre l'evolució del desenvolupament de la solució.

5.2.1. Entorn de desenvolupament i eines utilitzades

En primer lloc, i després d'estudiar el funcionament dels arbres de decisió i la característica addicional que hi volem implementar, s'ha decidit que s'utilitzaria l'eina RapidMiner (RapidMiner Studio 6.0.008) per a fer algunes proves inicials [Akthar i Hahne 2012; RapidMiner 2010]. Aquesta eina permet fer proves immediates amb diferents tipus d'arbres de decisió i bases de dades d'exemple que ofereix el mateix software. Per tant, és una bona manera d'estudiar el comportament dels diferents mètodes de discriminació d'atributs de què disposa l'eina per als arbres de decisió. Evidentment, l'eina és molt més completa i té múltiples funcionalitats relacionades amb la Intel·ligència Artificial, però en aquest projecte caldrà centrar-se en la part dels arbres de decisió.

A més a més, l'eina RapidMiner té una versió limitada i gratuïta, i es permet la descàrrega del seu codi. Per tant, aprofitant aquesta opció, s'ha descarregat el codi de l'aplicació (en aquest cas la versió RapidMiner 5.3.009) i s'ha importat com a projecte de l'entorn de desenvolupament Eclipse mitjançant l'extensió d'aquest entorn que permet descarregar projectes des de SVN. D'aquesta manera s'ha pogut accedir al codi i estudiar com funcionen els algorismes.

RapidMiner també disposa de maneres per estendre alguna de les seves funcionalitats sense haver de modificar directament el codi, és a dir, creant un projecte d'Eclipse nou que interaccioni amb el projecte de RapidMiner. De totes maneres, aquesta opció s'ha descartat per a implementar els nostres mètodes de discriminació, ja que presenta força dificultats. S'ha optat, doncs, per crear noves classes dins el mateix projecte de RapidMiner i sobreescriure'n alguna d'existent per a poder incloure els nous mètodes de discriminació.

A part de l'entorn de desenvolupament Eclipse i l'eina RapidMiner, que són les dues eines principals que s'han utilitzat, hi ha altres eines o complements que s'han utilitzat per a dur a terme la implementació:

- L'extensió d'Eclipse per a descarregar projectes des de SVN, que ha servit per a poder descarregar el projecte de RapidMiner a l'Eclipse i modificar el seu codi convenientment per a adaptar-lo a la nostra solució.

- El format o tipus de document CSV (*Comma-separated values*), que s'utilitza per a representar dades en forma de taula de tal manera que les columnes se separen per comes i les files per salts de línia. S'ha utilitzat per a representar la base de dades inicial, ja que el format CSV és molt senzill i és un dels que accepta com a vàlids

l'eina RapidMiner, a part que també és molt senzill crear funcions que llegeixin aquests fitxers des del codi.

- El programa Microsoft Excel, que s'ha utilitzat per a editar còmodament els fitxers de la base de dades en format CSV. Tot i que els fitxers d'aquest format es poden editar amb gairebé qualsevol editor de text, el programa Microsoft Excel ha sigut útil per a representar les diferents taules gràficament en forma de cel·les. Un cop guardat el resultat, però, cal tenir en compte que en aquest cas les columnes queden separades per un punt i coma (";") en comptes d'una coma, però això no comporta cap problema d'interpretació per a l'eina RapidMiner, que permet especificar el tipus de separador.

Amb totes aquestes eines ha sigut possible dur a terme la implementació de la solució i comprovar-ne els resultats.

5.2.2. Extensió del projecte

Un dels problemes inicials que s'han trobat a l'hora de desenvolupar la solució proposada ha sigut la impossibilitat de crear una extensió per al projecte de RapidMiner. Tal com diu la documentació de la mateixa eina [Land 2012], s'ofereix la possibilitat de descarregar el projecte de RapidMiner a l'entorn de desenvolupament Eclipse i, alhora, crear un nou projecte d'Eclipse independent que interaccioni amb el projecte de RapidMiner. Llavors, aquest nou projecte pot incloure modificacions del codi de RapidMiner que s'adaptin a les necessitats de la solució i incorporar-les de tal manera que no afecti al codi original de l'eina.

Evidentment, aquesta manera de treballar és molt bona per a provar diferents canvis a l'eina sense sobreescriure el codi original. De fet, existeix un manual de RapidMiner expressament dedicat a crear aquest tipus d'extensions. Tot i això, però, hi ha manca de coherència entre aquest manual i l'eina real, i falten alguns fitxers o no hi ha coherència entre versions. Això implica que algun dels passos del manual sigui difícil de dur a terme i s'hagin de buscar alternatives. Després de provar-ho de diverses maneres, i després d'aconseguir executar l'eina amb l'extensió creada, s'ha comprovat com les modificacions fetes a l'extensió no quedaven reflectides a l'execució de l'eina. Per tant, en veure que no s'aconseguia el resultat esperat, s'ha abandonat l'opció de crear una extensió.

L'alternativa, doncs, ha sigut modificar el mateix codi de l'eina. La modificació del codi, però, s'ha intentat fer evitant al màxim sobreescriure el codi ja existent, és a dir, s'ha intentat crear noves classes que s'encarreguin d'implementar la solució i incloure-les al projecte, modificant només aquell codi original que sigui imprescindible.

5.2.3. Addició dels nous mètodes de discriminació

Una vegada decidit el mètode d'adaptació del codi, cal modificar-lo per a introduir la fórmula de ponderació que ha d'ajudar a optimitzar el diagnòstic. Tenint en compte que es disposa del projecte RapidMiner a l'entorn de desenvolupament Eclipse, tal com s'ha dit, és possible modificar el codi font d'algunes classes per tal de modificar el comportament d'alguns algorismes.

En el nostre cas, després d'estudiar l'estructura de classes del projecte, s'ha aconseguit determinar quines fan referència als arbres de decisió i, més concretament, quines fan referència a cadascun dels mètodes de discriminació que s'analitzen. Com que la intenció és poder comparar els nous resultats que s'obtindran aplicant la nostra fórmula amb els resultats que s'obtindrien sense aplicar-la, no es pot sobreescriure el codi existent perquè es perdria la informació dels mètodes originals. Per tant, la solució passa per crear noves classes que incorporin els nous mètodes que apliquen la fórmula, sempre de forma complementària als mètodes originals. L'única classe que cal sobreescriure, doncs, és aquella on s'indica el conjunt de mètodes que es poden escollir, ampliant el conjunt i relacionant els nous mètodes amb les seves classes.

D'aquesta manera s'ha aconseguit crear un nou mètode de discriminació per a cadascun dels que s'analitzen, conformant així un conjunt de vuit mètodes (els quatre mètodes originals i els quatre mètodes que apliquen la fórmula).

A la interfície gràfica també s'ha aconseguit ampliar la llista de mètodes amb els nous algorismes creats, de manera que l'usuari pot escollir qualsevol dels vuit mètodes existents. Així és molt més fàcil fer proves entre els diferents mètodes i observar les diferències entre ells. La Figura 3 mostra aquesta interfície gràfica.

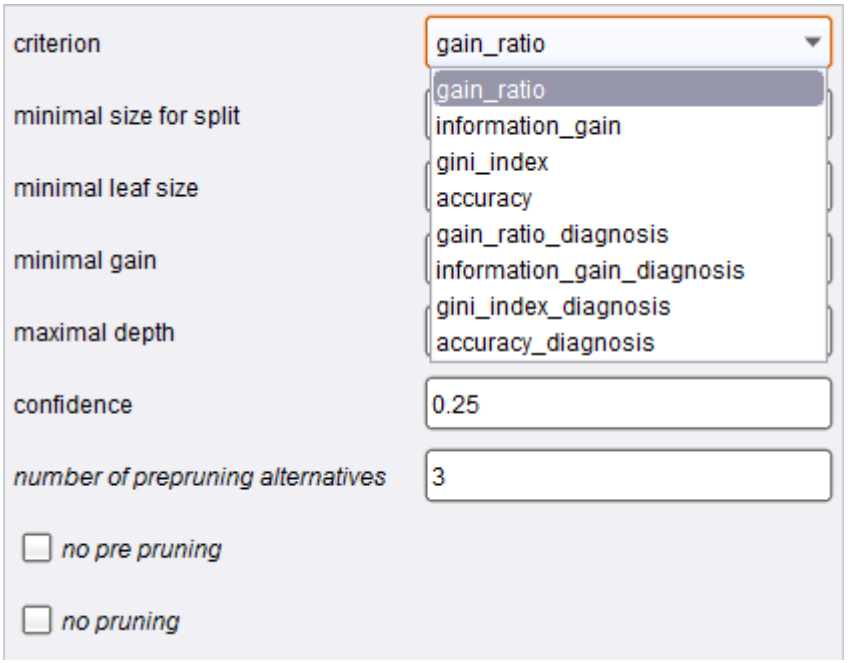

Figura 3. Interfície gràfica de l'eina RapidMiner per a escollir el mètode de discriminació.

Els nous mètodes de discriminació implementats apareixen amb el sufix "diagnosis", mentre que els originals apareixen sense el sufix.

Per a incloure la fórmula de ponderació que s'ha definit per a optimitzar els diagnòstics, la millor estratègia ha consistit en implementar aquesta fórmula una sola vegada i executar-la cada vegada que ho requereixin els diferents mètodes de discriminació. Per a aconseguir-ho, existeix una classe abstracta que és implementada per a cadascuna de les classes

referents als mètodes de discriminació. La fórmula de ponderació, doncs, s'ha implementat en aquesta classe abstracta en forma de diversos mètodes i atributs, de manera que cada classe específica que la implementa ja disposa d'aquests mètodes i atributs. L'Algorisme 2 mostra la part del codi referent a l'obtenció del benefici d'un atribut nominal per a un mètode de discriminació sense aplicar la fórmula.

```
@Override
public double getNominalBenefit(ExampleSet exampleSet, Attribute attribute) {
       double[][] weightCounts =
          frequencyCalculator.getNominalWeightCounts(exampleSet, attribute);

       return getBenefit(weightCounts);
}
```

Algorisme 2. Codi Java de l'obtenció del benefici d'un atribut nominal sense aplicar la fórmula.

En canvi, i per a poder observar la diferència essencial en el codi, l'Algorisme 3 mostra la part del codi referent a l'obtenció del benefici d'un atribut nominal per a un mètode de discriminació però, en aquest cas, aplicant la fórmula.

```
@Override
public double getNominalBenefit(ExampleSet exampleSet, Attribute attribute) {
       double[][] weightCounts =
          frequencyCalculator.getNominalWeightCounts(exampleSet, attribute);

       return getBenefit(weightCounts)
          * explorationWeights.get(attribute.getName());
}
```

Algorisme 3. Codi Java de l'obtenció del benefici d'un atribut nominal aplicant la fórmula.

Com es pot observar, en aquest cas es retorna el benefici original obtingut a partir del mètode de discriminació multiplicat pel pes que retorni la fórmula de ponderació que s'ha definit, que es guarda en una estructura de dades funcional on la clau és el nom de l'atribut (és a dir, del símptoma o l'exploració) i el valor és el pes que té assignat. El codi Java de la fórmula es pot observar a l'Annex 1.

5.2.4. Obtenció dels resultats referents a l'aplicació de la fórmula

Tal com es veurà a l'apartat de resultats, per a observar les diferències que apareixen quan s'aplica la fórmula o quan no s'aplica, una de les maneres és comparar els arbres de decisió generats i veure en què difereixen per a esbrinar quin d'ells és més òptim. Com que això pot ser complicat de concloure a simple vista, una de les maneres és comptar quants exemples requereixen dur a terme una determinada exploració per a predir la patologia, és a dir, quants exemples contenen un determinat node intern de l'arbre (o més d'un, si correspon a la mateixa exploració) en el seu camí cap al node terminal.

D'aquesta manera es pot saber numèricament si una exploració deixa de tenir presència en l'arbre de decisió quan els paràmetres d'eficiència indiquen que és poc eficient. Per a comptar aquests exemples, però, primer de tot ha calgut generar l'arbre de decisió a partir d'un determinat conjunt d'exemples i després demanar a l'eina RapidMiner que fes la predicció del mateix conjunt.

També ha calgut esbrinar quina de les classes del projecte era la responsable de fer les prediccions, i modificar-la per tal d'implementar un comptador que indiqui el nombre de vegades que el camí predit a l'arbre inclou una determinada exploració (un determinat node intern). Ha sigut necessari tenir en compte dos aspectes:

- En primer lloc, mantenir el comptador anterior cada vegada que es prediu un nou exemple, ja que el recorregut d'exemples es produeix en una altra classe diferent.
- En segon lloc, no comptar dues vegades el mateix node intern si apareix dues o més vegades al mateix camí de l'arbre.

Un cop contemplats aquests requeriments, s'ha dut a terme la implementació d'aquest comptador i s'ha creat un procés de l'eina RapidMiner per a obtenir els resultats. Aquest procés es mostra a la Figura 4.

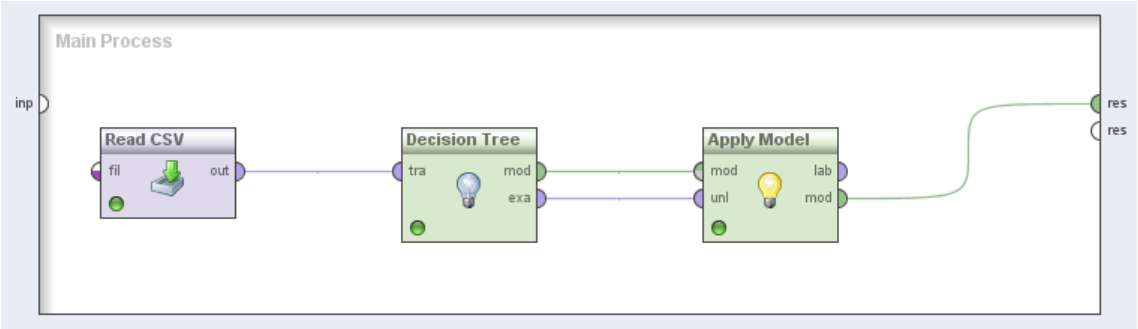

Figura 4. Procés de RapidMiner per a obtenir els resultats del comptador.

Com es pot veure, l'arbre de decisió es genera amb un conjunt d'exemples d'entrenament que s'obtenen llegint un fitxer CSV (sortida anomenada "out" que es connecta amb l'entrada "tra", referent al conjunt d'entrenament o *training*), i és aquest mateix conjunt d'exemples el que s'utilitza per a fer les prediccions (sortida "exa", referent al conjunt d'exemples o *exampleSet*, que es connecta amb l'entrada "unl", referent a les dades no etiquetades o *unlabelled data*), ja que interessa saber quants d'ells requereixen una determinada exploració per a predir la patologia. Les entrades i sortides anomenades "mod" fan referència al model, és a dir, a l'arbre de decisió, que finalment es connecten a l'entrada "res" que fa referència al resultat, de manera que s'obté la representació gràfica de l'arbre.

5.2.5. Obtenció dels resultats referents als mètodes de discriminació

Una altra part de l'apartat de resultats fa referència a l'estudi dels diferents mètodes de discriminació. Tal com es veurà, la principal mesura que s'utilitza per a esbrinar quin dels mètodes de discriminació és millor és la precisió de l'arbre de decisió. Per a fer-ho

s'utilitza la validació creuada, i l'eina RapidMiner ofereix la possibilitat de fer-ho d'aquesta manera i, a més a més, de calcular diverses mesures de rendiment com ara la precisió. La Figura 5 mostra el procés de RapidMiner que s'ha creat per a obtenir aquests resultats.

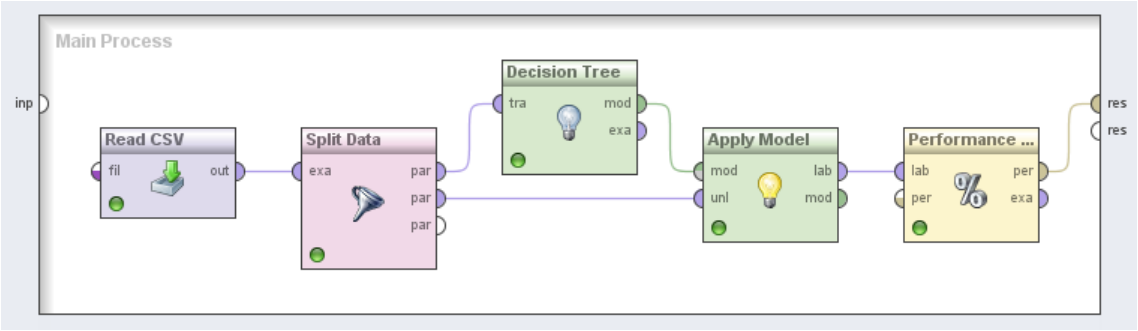

Figura 5. Procés de RapidMiner per a obtenir els resultats de la precisió.

Es pot observar com se separen les dades del conjunt d'exemples original (sortida "out", que es connecta a l'entrada "exa") en dos subconjunts, un per a l'entrenament de l'arbre de decisió (primera sortida "par", referent a la primera partició, que es connecta a l'entrada "tra" de l'arbre de decisió) i l'altre per a fer les proves (segona sortida "par", referent a la segona partició, que es connecta a l'entrada "unl"), eliminant així el biaix. Després de generar l'arbre de decisió amb el subconjunt d'entrenament, s'aplica el model obtingut (és a dir, l'arbre de decisió) al subconjunt de proves, i els resultats s'envien a l'operador corresponent (sortida "lab", referent a les dades ja etiquetades o *labelled data*, que es connecta a l'entrada amb el mateix nom) per a què calculi el rendiment (en aquest cas, la precisió de l'arbre de decisió). La sortida anomenada "per", referent al rendiment o *performance*, és la que es connecta amb l'entrada "res" per a obtenir el resultat final, un exemple del qual es mostra a la Figura 6.

accuracy: 85.83%

	true D1	true D2	true D3	true D4	true D5	true D6	true D7	true D8	true D9	true D10	class precis
pred. D1	12	0	0	0	0	0	0	0	0	0	100.00%
pred. D2	0	12	0	0	0	0	0	0	0	0	100.00%
pred. D3	0	0	9	0	1	0	0	0	0	0	90.00%
pred. D4	0	0	0	12	0	1	5	0	0	0	66.67%
pred. D5	0	0	3	0	11	0	0	0	1	0	73.33%
pred. D6	0	0	0	0	0	11	0	0	2	0	84.62%
pred. D7	0	0	0	0	0	0	7	0	0	0	100.00%
pred. D8	0	0	0	0	0	0	0	9	0	1	90.00%
pred. D9	0	0	0	0	0	0	0	1	9	0	90.00%
pred. D10	0	0	0	0	0	0	0	2	0	11	84.62%
class recall	100.00%	100.00%	75.00%	100.00%	91.67%	91.67%	58.33%	75.00%	75.00%	91.67%	

Figura 6. Exemple dels resultats obtinguts en mesurar la precisió.

S'observa com, a part de la precisió total, l'eina RapidMiner també ofereix la precisió per a cada patologia, tant la que fa referència al nombre de vegades que s'ha predit, com la que fa referència al nombre de vegades que s'ha encertat.

5.2.6. Tractament dels valors absents en la base de dades

Un altre problema que s'ha trobat durant l'obtenció dels resultats ha sigut el tractament dels valors absents de la base de dades. Tot i que ja s'ha explicat en apartats anteriors el procediment que s'ha seguit per a fer aquest tractament, a l'hora d'obtenir els primers resultats no es va aplicar aquest procediment.

Per tant, els primers resultats obtinguts consistien en arbres de decisió que contenien nodes amb valors discriminatoris desconeguts. Amb aquests arbres, quan s'intentava fer una predicció a partir d'un exemple nou i es trobava un node que no especificava els valors que separaven una branca de l'altra, era impossible continuar teixint el camí de l'arbre i arribar al node terminal. A part d'això, la generació de l'arbre era menys precisa perquè tenia en compte alguns valors absents dels atributs.

La solució, doncs, ha sigut establir un procediment de tractament d'aquests valors basat sobretot en estimacions. D'aquesta manera s'aconsegueix tenir una base de dades completa i uns resultats de predicció més precisos.

5.2.7. Implementació dels tests estadístics

Tal com es podrà veure a l'anàlisi dels resultats, per a comparar els diversos mètodes de discriminació s'utilitzaran alguns tests estadístics. Concretament, el test de Friedman, per a comparar diversos mètodes de discriminació; una millora d'aquest test no tan conservadora; i el test de Student per a comparar parelles de mètodes de discriminació.

Per a implementar aquests tests s'ha fet ús del llenguatge de programació Python, i el codi es pot observar a l'Annex 2.

6. Resultats

Després d'haver descrit la proposta de solució, d'haver plantejat els reptes per a la seva implementació i, finalment, d'haver implementat la solució, és imprescindible obtenir resultats i fer-ne una anàlisi completa per a esbrinar si la proposta de solució escollida és útil i permet optimitzar els diagnòstics.

Per a fer-ho, es dividirà aquest apartat en tres parts ben diferenciades: la primera part s'encarregarà de descriure les proves que s'han dissenyat, la segona descriurà els resultats obtinguts d'aquestes proves, i la tercera compararà i analitzarà els resultats per a extreure'n les conclusions oportunes.

6.1. Proves realitzades

Per a esbrinar si la solució escollida és adequada per a solucionar el problema que planteja el projecte, cal dissenyar conjunts de proves que permetin obtenir resultats comparables i analitzables i demostrar així la idoneïtat de la solució.

En aquest projecte hi ha almenys dues característiques que cal estudiar i analitzar:

- Les diferències en els resultats quan s'utilitza la fórmula de ponderació descrita anteriorment (respecte a la no utilització d'aquesta fórmula).
- Les diferències en els resultats quan s'utilitza un determinat mètode de discriminació (respecte a la utilització d'altres mètodes).

La fórmula de ponderació implica tenir en compte altres criteris que no formen part exclusivament del mètode de discriminació, de manera que aquesta ponderació és independent del mètode que s'utilitzi. Per tant, cal comprovar que realment sigui útil l'ús d'aquesta fórmula i que es tinguin en compte els paràmetres d'eficiència més subjectius, però després també cal estudiar quin mètode de discriminació és més adequat per a obtenir procediments de diagnòstic òptims i una predicció de patologies també òptima.

6.1.1. Proves respecte a la utilització de la fórmula

Per a comprovar les diferències entre els arbres de decisió generats quan s'utilitza la fórmula o quan no s'utilitza, s'ha dissenyat el següent conjunt de proves:

- Generació de dos arbres de decisió amb un extracte de la base de dades de prova: un aplicant la fórmula i l'altre sense aplicar la fórmula. Aquest extracte constarà de 250 exemples, 50 per a cadascuna de les 5 patologies que s'analitzen en aquest extracte, que estan definides per 20 símptomes diferents. Serà útil per a comprovar que realment l'aplicació de la fórmula té sentit, i serà molt més fàcil veure'n les diferències perquè el conjunt de dades serà més petit.
- Generació d'un arbre de decisió amb la base de dades de prova sencera sense tenir en compte la fórmula. La base de dades sencera consta de 500 exemples, 50 per a

cadascuna de les 10 patologies, que estan definides per 20 símptomes diferents. Aquest arbre servirà per fer les comparacions amb la resta de proves.

- Per a cada criteri o paràmetre d'eficiència (en aquest cas, per al cost, el temps, la invasió i la fiabilitat), generació de diferents arbres de decisió amb la base de dades de prova sencera:
    - Generació d'un arbre de decisió incrementant el valor del criteri per a una exploració concreta i mantenint els factors d'importància de tots els criteris a un valor mitjà (0.5, per exemple).
    - Generació d'un arbre de decisió incrementant el valor del criteri per a una exploració concreta i reduint el factor d'importància d'aquest criteri.
    - Generació d'un arbre de decisió incrementant el valor del criteri per a una exploració concreta i augmentant el factor d'importància d'aquest criteri.
    - Generació d'un arbre de decisió reduint el valor del criteri per a una exploració concreta i mantenint els factors d'importància de tots els criteris a un valor mitjà (0.5, per exemple).
    - Generació d'un arbre de decisió reduint el valor del criteri per a una exploració concreta i reduint el factor d'importància d'aquest criteri.
    - Generació d'un arbre de decisió reduint el valor del criteri per a una exploració concreta i augmentant el factor d'importància d'aquest criteri.
- Generació de diferents arbres de decisió amb la base de dades de prova sencera fent combinacions amb els valors i els factors de dos criteris, com ara el temps i el cost. Es tractaria de 36 combinacions diferents: augment o reducció del valor del criteri per a una exploració concreta, i augment, reducció o manteniment del factor d'importància; és a dir, 4 combinacions pel que fa als valors, i 9 combinacions pel que fa als factors, de manera que multiplicant els nombres apareixen 36 combinacions diferents. Això servirà per veure l'efecte en els arbres de decisió de la modificació de dos criteris alhora, en comptes de considerar-ne només un.

El conjunt d'aquestes proves ha de servir per veure si hi ha diferències clares en els arbres de decisió quan s'aplica o no s'aplica la fórmula.

6.1.2. Proves respecte als mètodes de discriminació

En aquest cas, i suposant que els resultats obtinguts aplicant la fórmula són millors que sense aplicar-la, cal comprovar les diferències dels resultats entre diferents mètodes de discriminació aplicant la fórmula.

Tenint en compte que el conjunt de proves de l'apartat anterior és força complet i engloba molts casos diferents, es pot utilitzar el mateix conjunt però executant-lo per a cada mètode de discriminació que s'analitzi.

Per a comprovar la precisió dels resultats sense biaix s'utilitzarà la validació creuada, és a dir, es dividirà el conjunt de dades en dues parts: una part servirà per a entrenar l'arbre de decisió i generar-lo, i l'altra part servirà per a fer les proves de predicció. D'aquesta mane-

ra s'evita que els exemples que es prediuen hagin estat utilitzats per a generar l'arbre i que aquest tingui un biaix respecte a aquests exemples. La validació creuada serà estratificada (es prendrà la mateixa proporció d'exemples de cada patologia per a configurar les dues parts), i la part destinada a l'entrenament correspondrà a un 75% del conjunt de dades total (el 25% restant es destinarà a la part de proves).

6.2. Resultats obtinguts de les proves

Una vegada desglossat el conjunt de proves que es pretenen realitzar, en aquest apartat s'exposaran els resultats obtinguts de cadascuna d'aquestes proves. Si algun dels resultats presenta problemes d'interpretació, també s'explicarà.

La intenció, doncs, és mostrar tots els resultats i deixar-ho tot preparat per a fer-ne l'anàlisi corresponent i extreure'n les conclusions oportunes.

6.2.1. Resultats sobre la utilització de la fórmula

Seguint el conjunt de proves explicat, la primera prova que es realitzarà és sobre un extracte de la base de dades de prova. D'aquesta manera es podrà comprovar la utilitat de la fórmula observant les diferències del conjunt de dades reduït.

Així, prenent de referència l'extracte de la base de dades de prova que s'ha creat amb la meitat d'exemples de la base de dades de prova sencera, modificant l'exploració "S3" per tal que el seu cost sigui de 100€, i modificant també el factor d'importància del criteri "cost" a 0.5, la Figura 7 mostra els arbres de decisió que s'obtenen aplicant el mètode de discriminació per ràtio de guany original i el nou mètode de discriminació per ràtio de guany que s'ha creat i que aplica la fórmula.

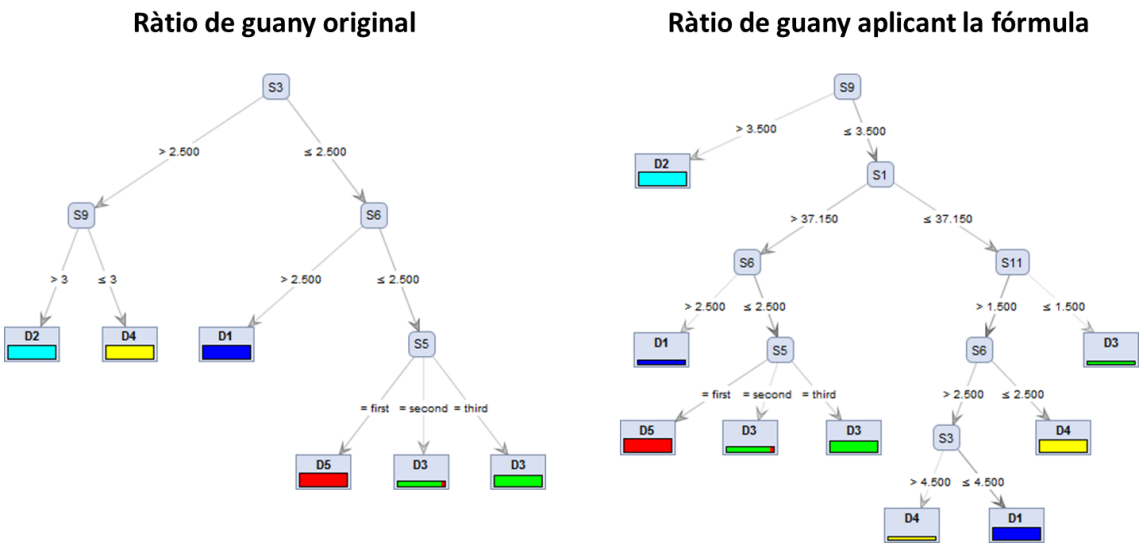

Figura 7. Arbres de decisió diferents en aplicar la fórmula.

És fàcil veure com en el mètode original l'atribut "S3" es troba a l'arrel de l'arbre perquè és el que discrimina millor, mentre que en el nou mètode l'atribut "S3" no es troba fins molt més tard i només en una sola branca i cap al final d'aquesta. Això és així perquè el nou mètode no només té en compte la capacitat de discriminació, sinó també la importància del cost de l'exploració; en aquest cas, l'exploració "S3" és de les més cares i el factor d'importància del criteri "cost" no és menyspreable (factor equivalent a 0.5). Per tant, és lògic que s'eviti aquesta exploració des d'un principi, optimitzant així el diagnòstic i fent ús d'altres exploracions més barates que poden resultar en conclusions semblants.

La segona prova que cal fer correspon a la generació d'un arbre de decisió amb la base de dades de prova sencera i sense aplicar la fórmula. Aquest arbre servirà per a fer les comparacions amb els arbres de decisió de les proves següents que sí que es generaran aplicant la fórmula. La Figura 8 mostra aquest arbre de decisió utilitzant el mètode de discriminació per ràtio de guany.

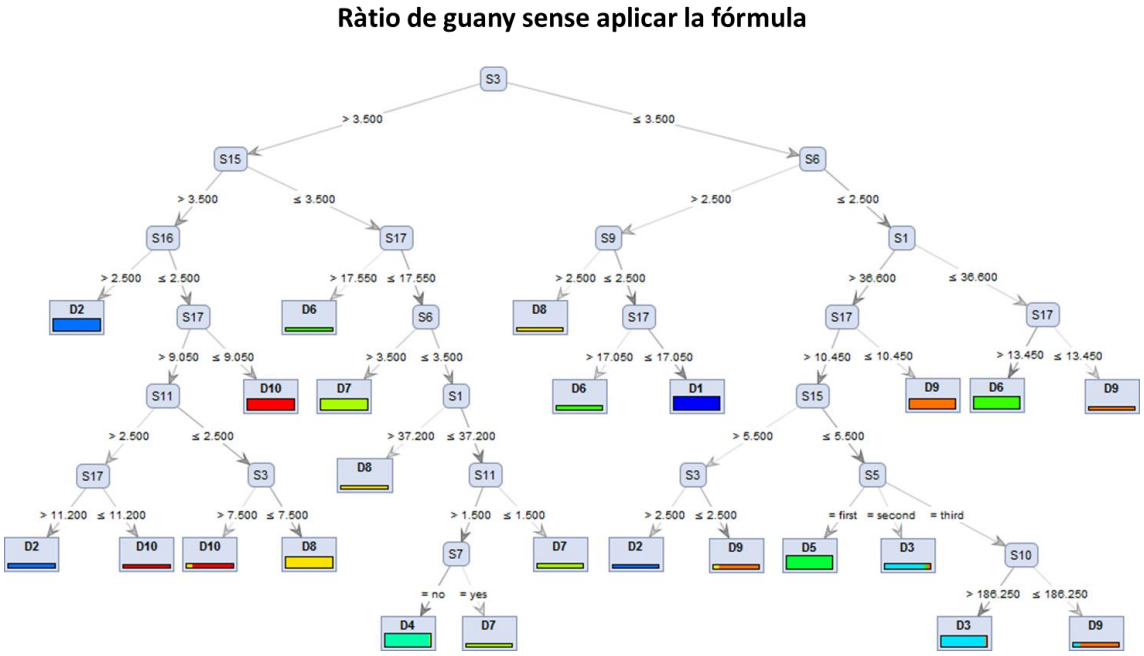

Figura 8. Arbre de decisió sencer sense aplicar la fórmula.

En aquest cas es pot observar l'arbre de decisió original per ràtio de guany, de manera que els primers atributs que es comproven corresponen directament a aquells que millor discriminen el conjunt de dades. També es poden observar les 10 patologies diferents ("D1", "D2", etc.) i alguns dels 20 símptomes que les defineixen ("S1", "S2", etc.). Cada patologia té assignat un color, i cada color es representa amb una barra que és més o menys gran segons el nombre d'exemples utilitzats en l'entrenament que van a parar en aquest node terminal. També és possible que un mateix node terminal tingui una barra amb més d'un color associat; això significa que en aquest node hi van a parar exemples de diferents patologies, de manera que l'arbre està exposat a predir una patologia equivocada. La probabilitat de predir una patologia o una altra es representa amb el volum de la barra que ocupa cada color, i el node terminal pren el nom de la patologia amb més probabilitat de ser predita.

La tercera prova inclou un conjunt d'arbres que s'obtenen modificant els valors dels diferents criteris per a una exploració concreta, i alhora modificant els factors d'importància d'aquests criteris. Per a l'obtenció d'aquests resultats, l'exploració concreta correspondrà a l'exploració "S17", ja que apareix freqüentment a l'arbre de decisió generat sense aplicar la fórmula. De fet, dels 500 exemples de la base de dades de prova sencera, 455 requereixen comprovar el resultat de l'exploració "S17" per a predir la patologia, és a dir, un 91% dels exemples.

El primer criteri que es modificarà serà el cost, i es començarà incrementant el valor del cost de l'exploració "S17". Amb el valor del cost augmentat (de 0.5 a 100 euros), es generaran tres arbres de decisió: un mantenint els factors d'importància de tots els criteris a 0.5, un altre reduint el factor d'importància del cost a 0.1 i mantenint la resta a 0.5, i un últim arbre augmentant el factor d'importància del cost a 1 i mantenint la resta a 0.5.

La Figura 9 mostra el primer d'aquests tres arbres, que manté la mateixa importància per a tots els criteris.

Ràtio de guany aplicant la fórmula: inc. del cost per a S17 i mateixos factors d'importància

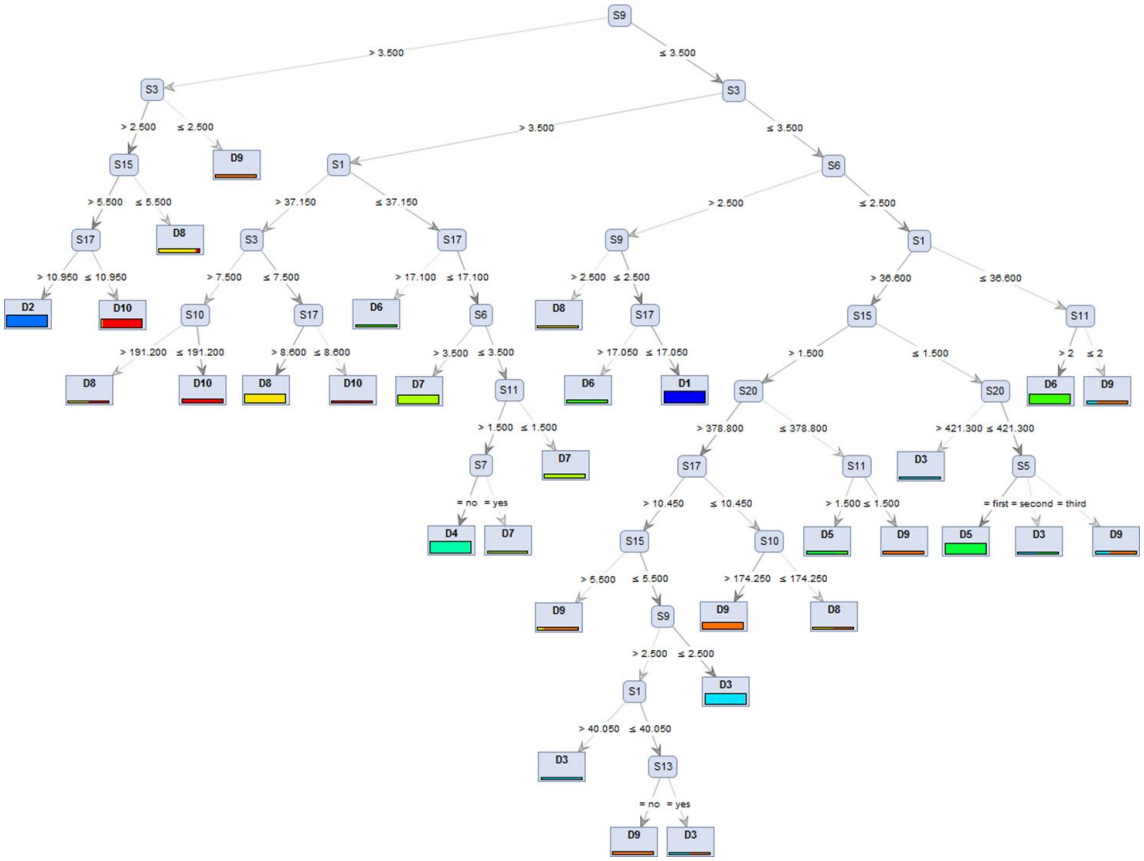

Figura 9. Increment del valor del cost per a "S17" i mateixos factors d'importància.

En aquest cas, tot i que el símptoma "S17" continua apareixent força freqüentment, es pot observar com apareix a nivells inferiors de l'arbre, prioritzant abans altres exploracions més barates. És a dir, com a conseqüència de l'increment del cost de l'exploració "S17",

aquesta apareix més tard a l'arbre de decisió. Dels 500 exemples, 364 requereixen el resultat de l'exploració "S17" (72.8%).

La Figura 10 mostra el segon arbre de decisió després d'incrementar el cost per a l'exploració "S17", en aquest cas reduint el factor d'importància del cost a 0.1 i mantenint la resta a 0.5.

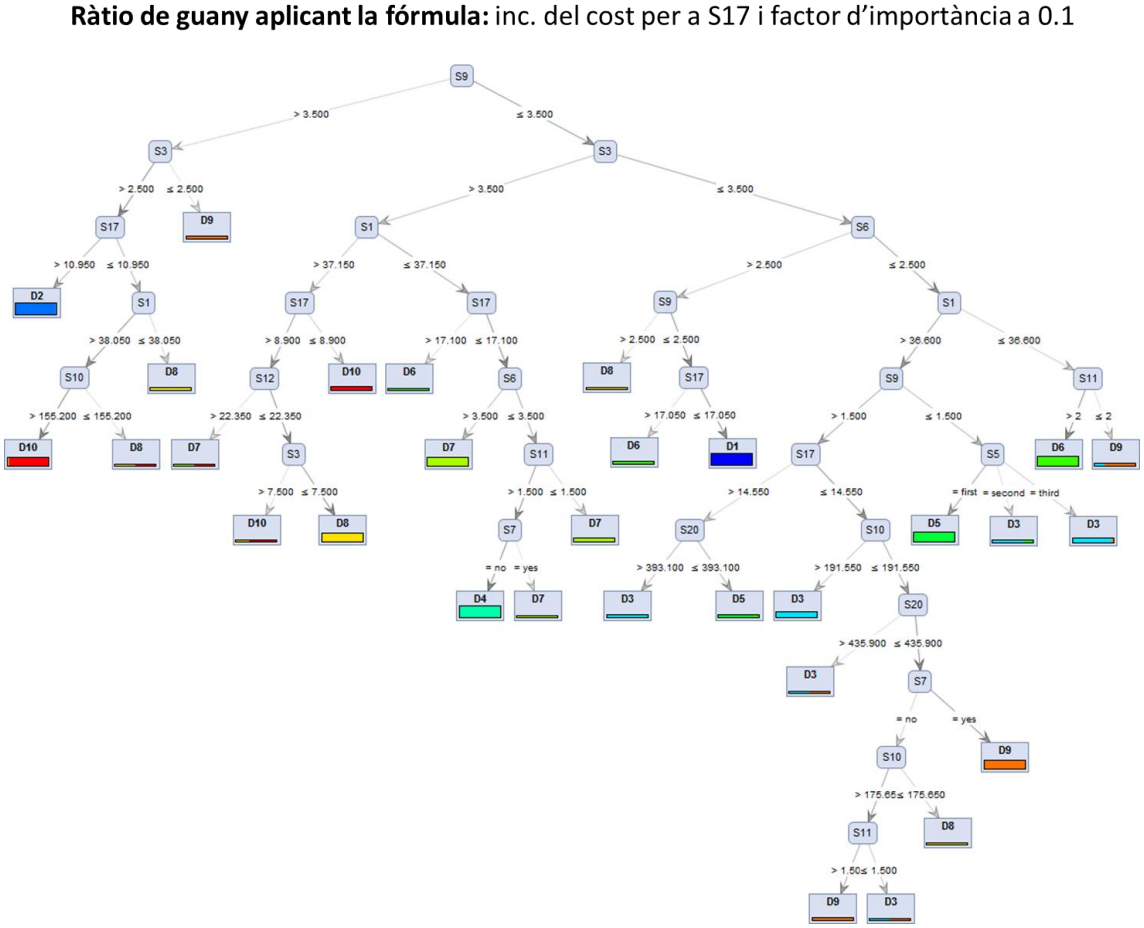

Figura 10. Increment del valor del cost per a "S17" i reducció del factor d'importància a 0.1.

Després de reduir el factor d'importància del cost a 0.1, la diferència respecte a l'anterior arbre de decisió no és gaire destacada, però sí que es pot observar com els nodes interiors referents al símptoma "S17" tornen a pujar alguns nivells a l'arbre de decisió, de manera que són comprovats abans que altres. Per tant, la reducció de la importància del criteri "cost" implica que aquest criteri no es tingui tant en compte a l'hora de situar les diverses exploracions a l'arbre. Dels 500 exemples, 385 requereixen el resultat de l'exploració "S17" (77%).

L'últim dels tres arbres de decisió després d'incrementar el cost per a l'exploració "S17" es mostra a la Figura 11. En aquest cas s'ha augmentat el factor d'importància del cost a 1 (el màxim valor possible) i s'ha mantingut la resta a 0.5.

Ràtio de guany aplicant la fórmula: inc. del cost per a S17 i factor d'importància a 1

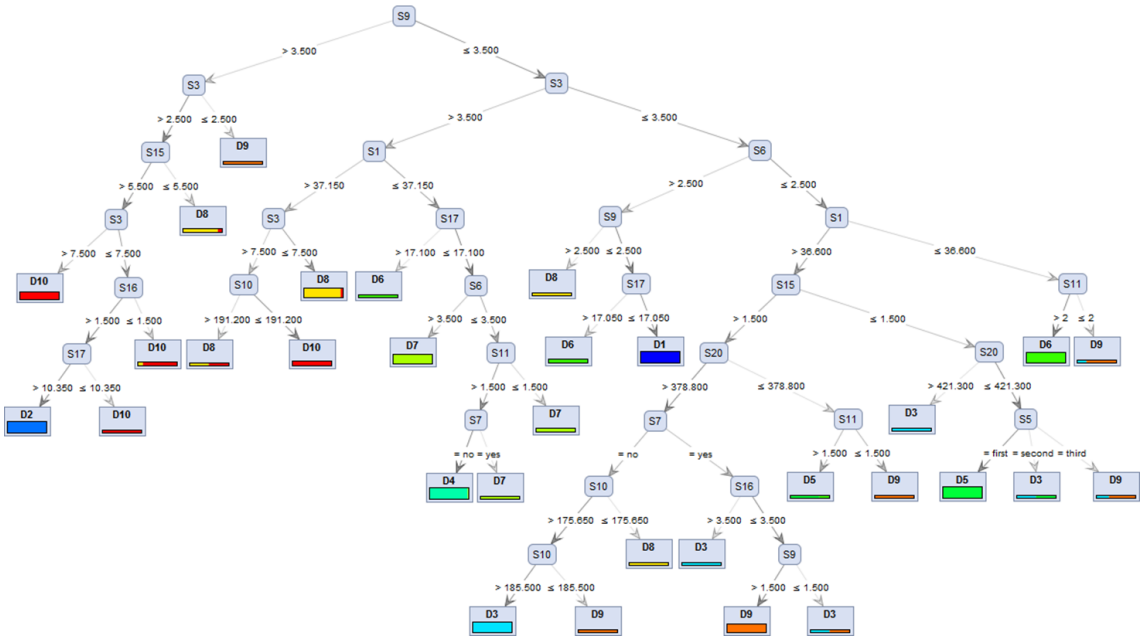

Figura 11. Increment del valor del cost per a "S17" i augment del factor d'importància a 1.

De manera semblant als altres casos, i seguint la tendència, en aquest cas s'observa que els nodes interns referents al símptoma "S17" se situen a nivells més baixos de l'arbre, perquè la importància del criteri "cost" ha augmentat al màxim i el cost de l'exploració "S17" és força elevat. Dels 500 exemples, 209 requereixen el resultat de l'exploració "S17" (41.8%).

Ràtio de guany aplicant la fórmula: red. del cost per a S17 i mateixos factors d'importància

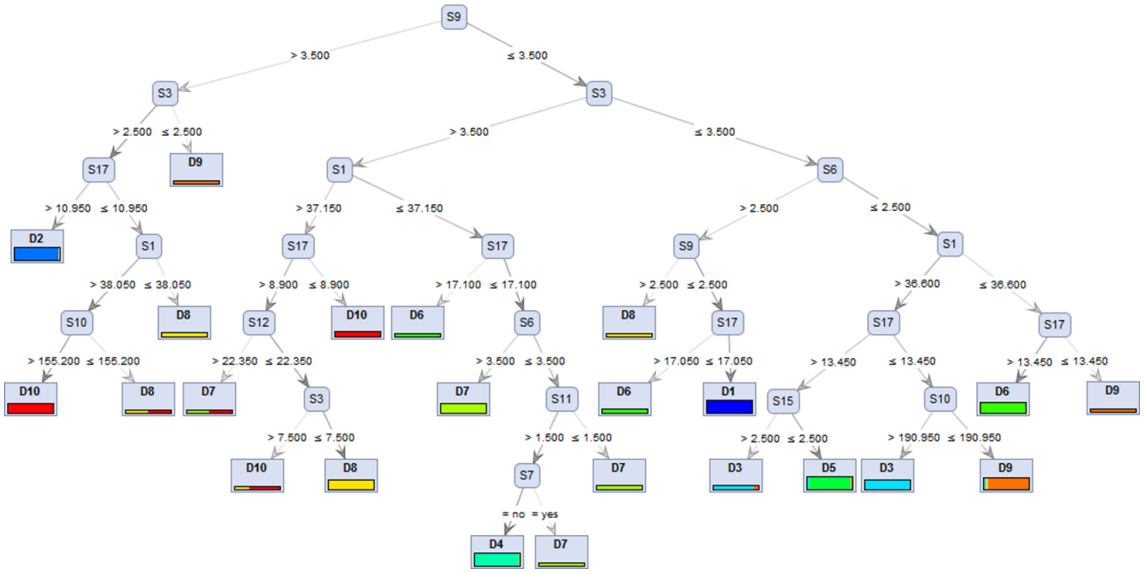

Figura 12. Reducció del valor del cost per a "S17" i mateixos factors d'importància.

Seguint amb el conjunt de proves descrit anteriorment, i després de descriure els resultats obtinguts en incrementar el valor del cost per a l'exploració "S17", a continuació es faran les mateixes proves però, en aquest cas, reduint el valor del cost per a la mateixa exploració (el cost serà de 0 euros). Per tant, també es generaran tres arbres diferents segons es mantingui, es redueixi o s'augmenti el factor d'importància del cost.

La Figura 12 mostra el primer d'aquests tres nous arbres, mantenint la mateixa importància per a tots els criteris.

Es pot observar com, tot i que el criteri "cost" té un factor d'importància de 0.5, els nodes interiors referents al símptoma "S17" es troben a nivells força superiors de l'arbre. A diferència de l'arbre generat abans on també es mantenia a 0.5 el factor d'importància del cost, en aquest cas el valor del cost per a l'exploració "S17" és nul, de manera que encara que se li assigni importància, el pes es rebaixa perquè l'exploració no té cost. Dels 500 exemples, 494 requereixen el resultat de l'exploració "S17" (98.8%).

El segon arbre de decisió que s'ha generat reduint el valor del cost per a l'exploració "S17" també té en compte que s'ha reduït el factor d'importància del cost a 0.1, tot mantenint la resta a 0.5. Aquest arbre es mostra a la Figura 13.

Ràtio de guany aplicant la fórmula: red. del cost per a S17 i factor d'importància a 0.1

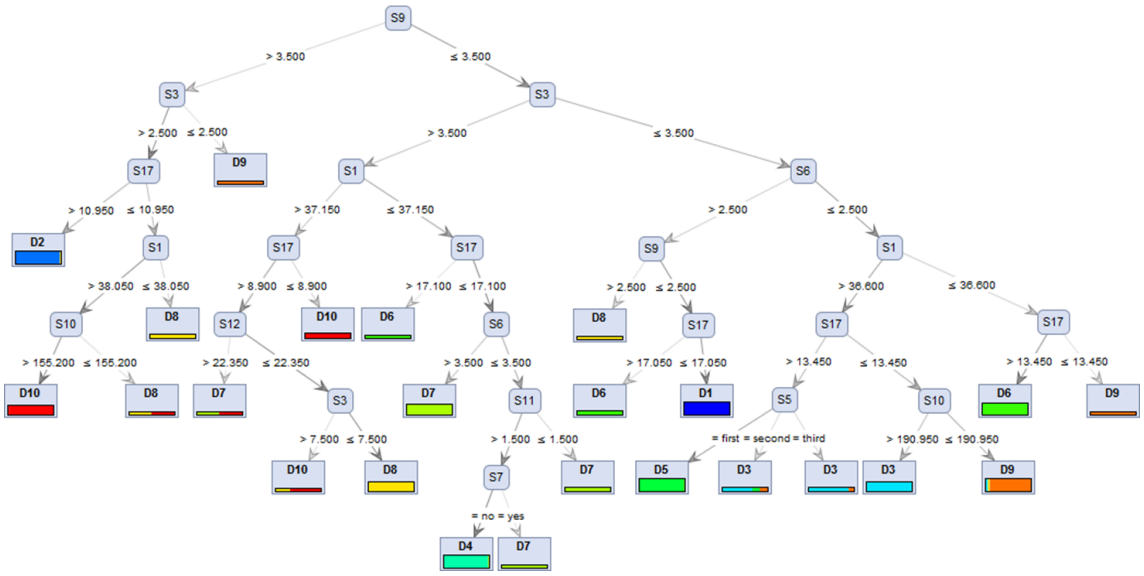

Figura 13. Reducció del valor del cost per a "S17" i reducció del factor d'importància a 0.1.

En aquest segon cas s'observa com l'arbre de decisió generat és pràcticament el mateix que l'anterior; només varia una petita part dels nodes terminals. Per tant, la reducció del factor d'importància del cost no afecta a la posició dels nodes interns referents a l'exploració "S17" quan el cost és molt baix o nul. Com abans, dels 500 exemples, 494 requereixen el resultat de l'exploració "S17" (98.8%).

Finalment, el tercer i últim arbre de decisió d'aquest grup, on s'ha reduït el valor del cost per a l'exploració "S17", s'ha generat després d'augmentar el factor d'importància del cost

a 1, tot mantenint la resta de factors a 0.5. L'arbre de decisió resultant coincideix exactament amb l'arbre representat a la Figura 12.

Per tant, tenint en compte que el cost de l'exploració "S17" és nul, el factor d'importància del cost no influeix gens en la posició dels nodes interiors referents a aquesta exploració. Com en els dos casos anteriors, doncs, dels 500 exemples, 494 requereixen el resultat de l'exploració "S17" (98.8%).

Després d'haver obtingut els resultats pertinents en modificar el criteri "cost", ara cal fer el mateix per a la resta de criteris per a comprovar que tots tenen la capacitat d'influir en l'arbre de decisió que es genera finalment.

Com que ja s'ha pogut comprovar la relació que hi ha entre els arbres de decisió gràfics que s'han mostrat i la quantitat d'exemples que requereixen el resultat d'una determinada exploració, la resta de criteris s'avaluaran tenint en compte només aquesta última mesura. O sigui, per a avaluar la influència de la resta de criteris es comptarà el nombre d'exemples que requereixen fer una determinada exploració per a predir la patologia. Si aquest nombre és gran, i suposant que l'aplicació de la fórmula funciona, significarà que l'exploració que s'hagi escollit és eficient i no perjudica l'obtenció d'un diagnòstic òptim; en canvi, si aquest nombre és petit significarà que l'exploració és poc eficient i que és millor prescindir-ne. Com en el cas del criteri "cost", també es prendrà de referència l'exploració "S17" perquè apareix freqüentment a l'arbre de decisió generat sense aplicar la fórmula, així com també s'utilitzarà el mètode de discriminació per ràtio de guany.

El següent criteri a avaluar, doncs, és el temps. Es començarà incrementant el valor del temps de l'exploració "S17" i, amb el valor augmentat (d'1 a 14 dies), es generaran, com abans, tres arbres de decisió: un mantenint els factors d'importància de tots els criteris a 0.5, un altre reduint el factor d'importància del temps a 0.1 i mantenint la resta a 0.5, i un últim arbre augmentant aquest factor d'importància a 1 i mantenint la resta a 0.5. Després es repetirà el procés generant tres arbres més però, en aquest cas, reduint el valor del temps de l'exploració "S17" a un sol dia.

La Taula 11 mostra els resultats desglossant cadascuna de les proves i mostrant el nombre d'exemples que han requerit el resultat de l'exploració "S17" per a predir la patologia, el nombre d'exemples total de la base de dades de prova, i el percentatge corresponent.

Criteri "temps"	Tots els FI a 0.5	FI temps a 0.1	FI temps a 1
Inc. temps S17	280 / 500 (56%)	385 / 500 (77%)	106 / 500 (21.2%)
Red. temps S17	449 / 500 (89.8%)	449 / 500 (89.8%)	449 / 500 (89.8%)

Taula 11. Resultats de les proves per al criteri "temps" (FI = Factor d'importància).

De manera molt similar als resultats del criteri "cost", en aquest cas es pot observar com la importància del criteri "temps" és inversament proporcional al nombre d'exemples que requereixen una determinada exploració amb un temps elevat per a predir la patologia.

Quan el temps és baix o gairebé nul, llavors la importància d'aquest criteri no afecta al resultat final perquè el mateix valor del temps ja és òptim.

El tercer criteri que cal avaluar és la invasió. Com abans, es començarà incrementant el valor de la invasió de l'exploració "S17" i, amb el valor augmentat (de 2 a 5 punts), es generaran els mateixos arbres de decisió que s'han generat en avaluar els criteris anteriors. Després es tornarà a repetir el procés reduint el valor de la invasió de l'exploració "S17" (de 2 a 1 punt).

La Taula 12 mostra els resultats desglossant cadascuna de les proves i mostrant el nombre d'exemples que han requerit el resultat de l'exploració "S17" per a predir la patologia, el nombre d'exemples total de la base de dades de prova, i el percentatge corresponent.

Criteri "invasió"	Tots els FI a 0.5	FI invasió a 0.1	FI invasió a 1
Inc. invasió S17	364 / 500 (72.8%)	441 / 500 (88.2%)	175 / 500 (35%)
Red. invasió S17	498 / 500 (99.6%)	498 / 500 (99.6%)	498 / 500 (99.6%)

Taula 12. Resultats de les proves per al criteri "invasió" (FI = Factor d'importància).

Seguint la mateixa tendència que els criteris anteriors, també es pot veure com la importància del criteri "invasió" és inversament proporcional al nombre d'exemples que requereixen una determinada exploració amb alta invasió per a predir la patologia. Quan la invasió és baixa o gairebé nul·la, llavors la importància d'aquest criteri, com en els casos anteriors, no afecta al resultat final perquè el mateix valor de la invasió ja és òptim.

Finalment, l'últim criteri que s'ha d'avaluar és la fiabilitat. També es començarà incrementant el valor de la fiabilitat de l'exploració "S17" i, amb aquest valor augmentat (de 2 a 5 punts), es generaran altre cop els mateixos arbres de decisió que s'han generat en avaluar els criteris anteriors. Per acabar, es tornarà a repetir el mateix procés reduint el valor de la fiabilitat de l'exploració "S17" (de 2 a 1 punt).

La Taula 13 mostra els resultats desglossant cadascuna de les proves i mostrant el nombre d'exemples que han requerit el resultat de l'exploració "S17" per a predir la patologia, el nombre d'exemples total de la base de dades de prova, i el percentatge corresponent.

Criteri "fiabilitat"	Tots els FI a 0.5	FI fiabilitat a 0.1	FI fiabilitat a 1
Inc. fiabilitat S17	500 / 500 (100%)	409 / 500 (81.8%)	500 / 500 (100%)
Red. fiabilitat S17	388 / 500 (77.6%)	362 / 500 (72.4%)	386 / 500 (77.2%)

Taula 13. Resultats de les proves per al criteri "fiabilitat" (FI = Factor d'importància).

En aquest últim cas, es pot observar com la importància del criteri "fiabilitat" és directament proporcional al nombre d'exemples que requereixen una determinada exploració amb alta fiabilitat per a predir la patologia. Aquest cas és diferent de la resta de criteris

perquè la fiabilitat és millor si és alta. D'altra banda, quan la fiabilitat és baixa o gairebé nul·la, sembla que els resultats no presentin massa diferències quan es modifica el factor d'importància. Això ocorre perquè la majoria d'exploracions tenen una fiabilitat baixa, i les que tenen alta fiabilitat presenten altres problemes com ara un cost o un temps elevat. Si es rebaixen els factors d'importància de la resta de criteris o s'apuja la fiabilitat d'alguna exploració que tingui un cost i un temps reduït, llavors els resultats sí que reflecteixen un nombre més reduït d'exemples que requereixen l'exploració "S17" per a predir la patologia. Per tant, la fiabilitat també es comporta com la resta de criteris.

Per acabar aquest apartat de proves, la quarta i última prova que s'ha definit en apartats anteriors correspon a generar diferents arbres de decisió amb la base de dades de prova sencera fent combinacions amb els valors i els factors d'importància de dos criteris, com ara el cost i el temps. L'objectiu és fer 36 combinacions diferents tot augmentant o reduint els valors dels dos criteris per a una exploració concreta (4 possibles combinacions), i augmentant, reduint o mantenint els seus factors d'importància (9 possibles combinacions per a cadascuna de les 4 anteriors, és a dir, un total de 36).

Per a mostrar els resultats s'utilitzarà la mateixa mesura que s'ha utilitzat en les proves anteriors, o sigui, el nombre d'exemples que requereixen el resultat d'una exploració concreta per a predir la patologia. Aquesta exploració, com abans, correspondrà a l'exploració "S17", que apareix freqüentment a l'arbre de decisió generat sense aplicar la fórmula.

	Valor cost	**reduït**	**reduït**	**augmentat**	**augmentat**
	Valor temps	**reduït**	**augmentat**	**reduït**	**augmentat**
FI cost	**FI temps**	*Nombre d'exemples que requereixen l'exploració "S17"*			
0.1	**0.1**	**393** (78.6%)	**375** (75%)	**388** (77.6%)	**363** (72.6%)
0.1	**0.5**	**494** (98.8%)	**106** (21.2%)	**385** (77%)	**106** (21.2%)
0.1	**1**	**494** (98.8%)	**106** (21.2%)	**449** (89.8%)	**54** (10.8%)
0.5	**0.1**	**494** (98.8%)	**385** (77%)	**279** (55.8%)	**143** (28.6%)
0.5	**0.5**	**494** (98.8%)	**280** (56%)	**364** (72.8%)	**106** (21.2%)
0.5	**1**	**494** (98.8%)	**106** (21.2%)	**374** (74.8%)	**54** (10.8%)
1	**0.1**	**494** (98.8%)	**388** (77.6%)	**106** (21.2%)	**106** (21.2%)
1	**0.5**	**494** (98.8%)	**364** (72.8%)	**209** (41.8%)	**106** (21.2%)
1	**1**	**454** (90.8%)	**157** (31.4%)	**241** (48.2%)	**54** (10.8%)

Taula 14. Resultats de les proves per als criteris "cost" i "temps" (FI = Factor d'importància).

La Taula 14 mostra aquests resultats separant els valors dels criteris per a l'exploració "S17" i els factors d'importància d'aquests. Cada columna fa referència a una combinació d'augment o reducció dels valors dels criteris, mentre que cada fila fa referència a una combinació dels factors d'importància dels criteris. El resultat correspon al nombre

d'exemples que necessiten l'exploració "S17" per a predir la patologia, i entre parèntesis apareix el percentatge respecte al nombre total d'exemples de la base de dades de prova, que correspon a 500 exemples.

Els resultats mostren com, quan els factors d'importància són molt petits, o quan els valors dels criteris són gairebé òptims (molt baixos), l'ús de l'exploració "S17" és més corrent, sense que es notin les diferències entre els diferents casos. En canvi, és evident que, quan un dels criteris té un valor elevat, l'ús de l'exploració "S17" disminueix, i aquest efecte encara s'accentua més quan el factor d'importància del criteri que té un valor alt també és elevat. Depenent dels valors i els factors d'importància assignats als criteris, el nombre d'exemples que requereix l'exploració "S17" per a predir la patologia varia d'un 10.8% a un 98.8%, mostrant l'alta flexibilitat que permet l'ús de la fórmula a l'hora de generar arbres de decisió.

6.2.2. Resultats sobre els mètodes de discriminació

Partint del mateix conjunt de proves que s'ha utilitzat a l'apartat anterior, s'han repetit aquestes proves per a cada mètode de discriminació per a avaluar-ne les diferències. Abans que res, però, s'han fet proves amb altres bases de dades reduïdes per a exemplificar les diferències existents entre els diferents mètodes.

La Figura 14 mostra aquestes diferències a través dels arbres de decisió generats a partir de la base de dades "Golf" estesa convenientment. Aquesta base de dades conté exemples que estudien la idoneïtat de jugar a aquest joc segons algunes característiques com ara el temps meteorològic, la temperatura, la humitat o el vent.

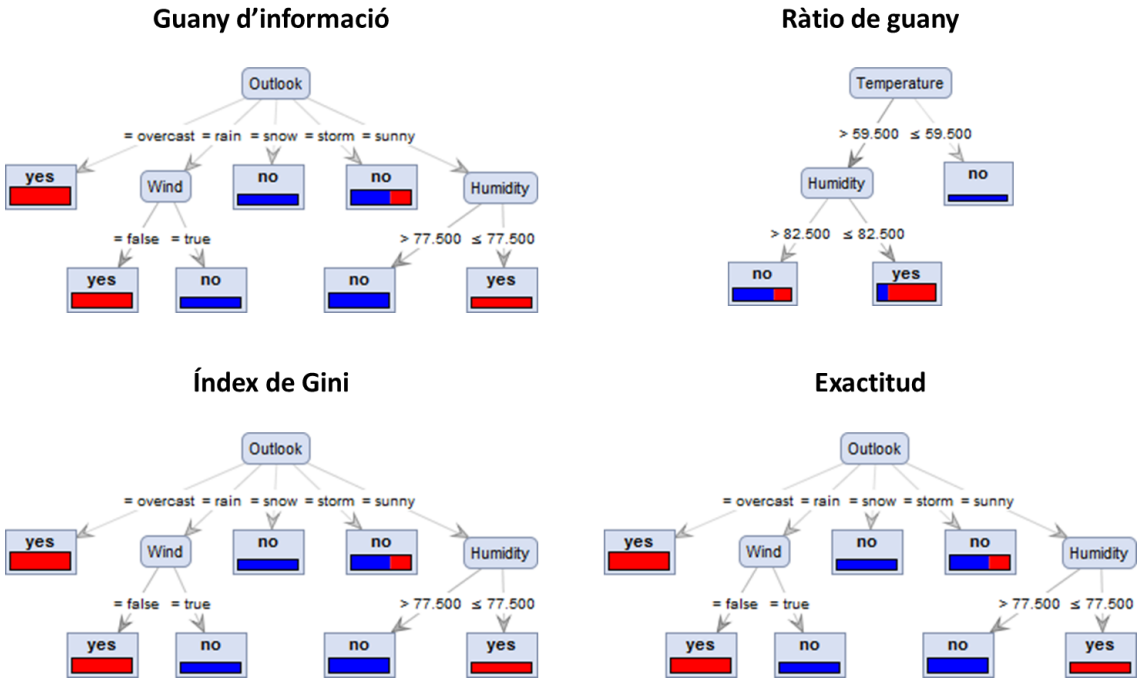

Figura 14. Arbres de decisió amb diferents mètodes de discriminació (BdD: Golf estesa).

Tot i que els arbres de decisió obtinguts aplicant els mètodes de discriminació per guany d'informació, per índex de Gini i per exactitud són idèntics, sí que es poden observar diferències respecte al mètode de discriminació per ràtio de guany.

Per demostrar que tots els mètodes de discriminació són diferents, la Figura 15 mostra altra vegada diferents arbres de decisió però sobre una base de dades diferent. Aquesta base de dades s'anomena "Labor-Negotiations" i fa referència a condicions de treball.

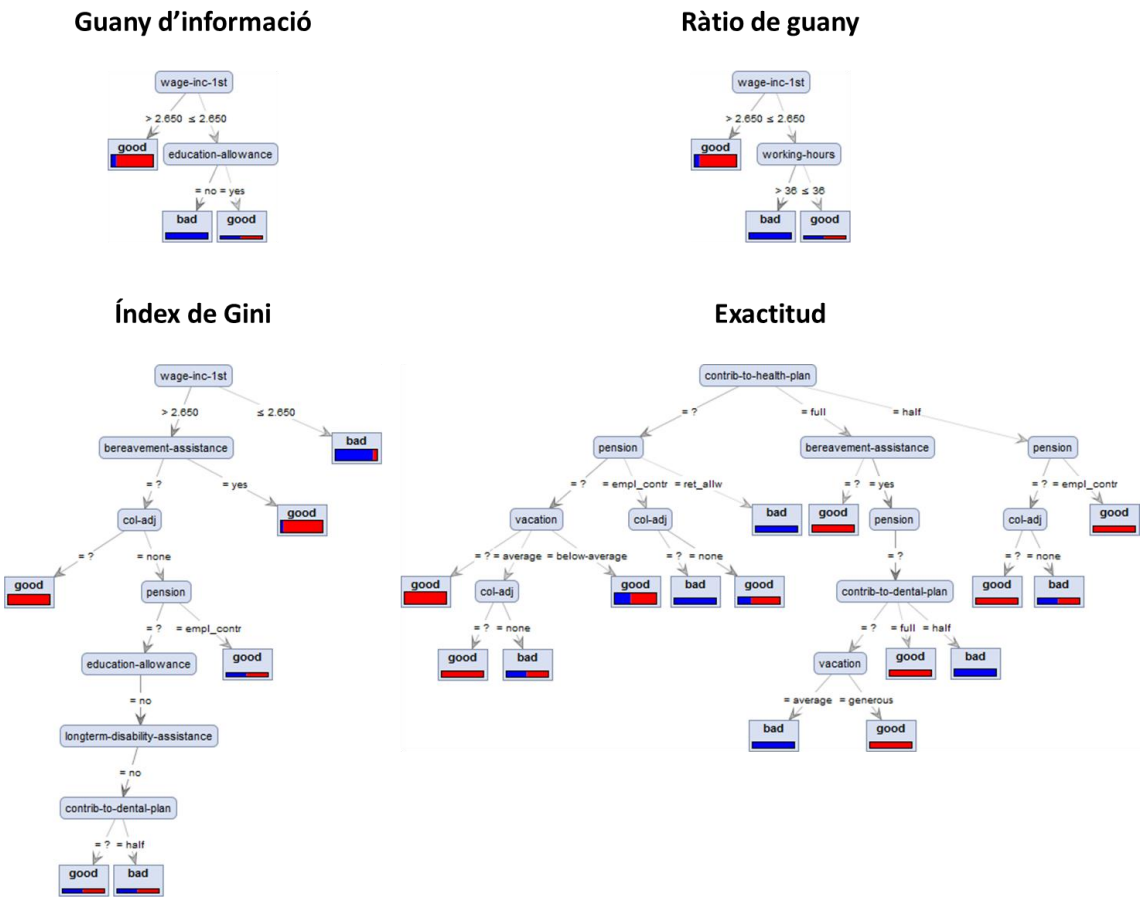

Figura 15. Varietat d'arbres segons mètode de discriminació (BdD: Labor-Negotiations).

En aquest cas, doncs, és evident que cada arbre és diferent. Els que presenten més semblances són els arbres generats amb els mètodes de discriminació per guany d'informació i per ràtio de guany, però es pot observar com un dels nodes interns varia.

Amb aquests resultats es pretén demostrar que el mètode de discriminació influeix en la generació de l'arbre de decisió final. Cada mètode utilitza una mesura diferent per a considerar quin dels atributs de l'arbre discrimina millor les dades, de manera que els primers atributs que se seleccionen poden ser diferents.

En el nostre projecte, tot i que la part que s'ha introduït és l'aplicació d'una fórmula que té en compte altres criteris subjectius a l'hora de decidir quina exploració cal fer abans o, el que és el mateix, quin atribut cal seleccionar primer, és evident que també és molt important saber quin símptoma discrimina millor les dades per a obtenir finalment la patologia

que pateix un pacient. En aquest sentit, cal esbrinar quin mètode de discriminació és el més òptim.

Fins ara s'han mostrat els arbres de decisió que es generaven en aplicar la fórmula o sense aplicar-la, i també s'ha mostrat el nombre d'exemples que requerien una determinada exploració per a predir la patologia. Tots aquests resultats serviran per a saber si l'aplicació de la fórmula té sentit o no. A part, però, cal comprovar quin dels mètodes de discriminació és millor, i això es pot saber comprovant el nombre d'exemples que es classifiquen correctament després de fer una predicció. Per tant, en aquest apartat serà més important saber la precisió dels arbres de decisió depenent del mètode de discriminació escollit.

En aquest sentit, els resultats s'obtindran a partir del mateix conjunt de proves que s'ha utilitzat per a comprovar la idoneïtat de l'aplicació de la fórmula, ja que es considera que és un conjunt força complet que inclou una gran varietat de combinacions. A més a més, com ja s'ha dit, es farà ús de la validació creuada per a evitar biaixos a l'hora de fer les prediccions, és a dir, s'evitarà classificar o predir la patologia d'un exemple que ja s'hagi utilitzat per a entrenar l'arbre de decisió. Per tant, el conjunt d'exemples es dividirà en dues parts:

- La primera, corresponent al 75% dels exemples (aproximadament 375 exemples dels 500 disponibles), s'utilitzarà per a entrenar l'arbre de decisió.
- La segona, corresponent al 25% dels exemples (aproximadament 125 exemples dels 500 disponibles), s'utilitzarà per a fer la predicció i obtenir la precisió de l'arbre.

La separació de les dades serà estratificada, o sigui, es prendrà la mateixa proporció d'exemples de cada patologia tant per al conjunt d'entrenament com per al conjunt de proves.

En primer lloc, doncs, la primera prova consistia en generar dos arbres de decisió amb un extracte de la base de dades de prova, un aplicant la fórmula i l'altre sense aplicar-la. La Taula 15 mostra els diferents resultats de la precisió obtinguda depenent del mètode de discriminació.

Precisió	Aplicant la fórmula	Sense aplicar la fórmula
Guany d'informació	98.33%	98.33%
Ràtio de guany	98.33%	98.33%
Índex de Gini	93.33%	96.67%
Exactitud	91.67%	95.00%

Taula 15. Precisió segons el mètode de discriminació per a un extracte de la base de dades.

Es pot observar com, en l'extracte de la base de dades, els mètodes de discriminació que presenten més bons resultats són el guany d'informació i la ràtio de guany, ja sigui aplicant

la fórmula o sense aplicar-la. Tot i això, la resta de mètodes també presenten un percentatge força elevat, de manera que caldrà fer proves amb la base de dades sencera per a obtenir resultats més fiables.

D'altra banda, tot i que pugui semblar que l'aplicació de la fórmula no afecta al resultat perquè la precisió és la mateixa, els arbres de decisió que es generen són diferents. Un exemple d'això es mostra a la Figura 16, en què es representen els dos arbres de decisió generats utilitzant el mètode de discriminació per guany d'informació a partir de l'extracte de la base de dades de prova.

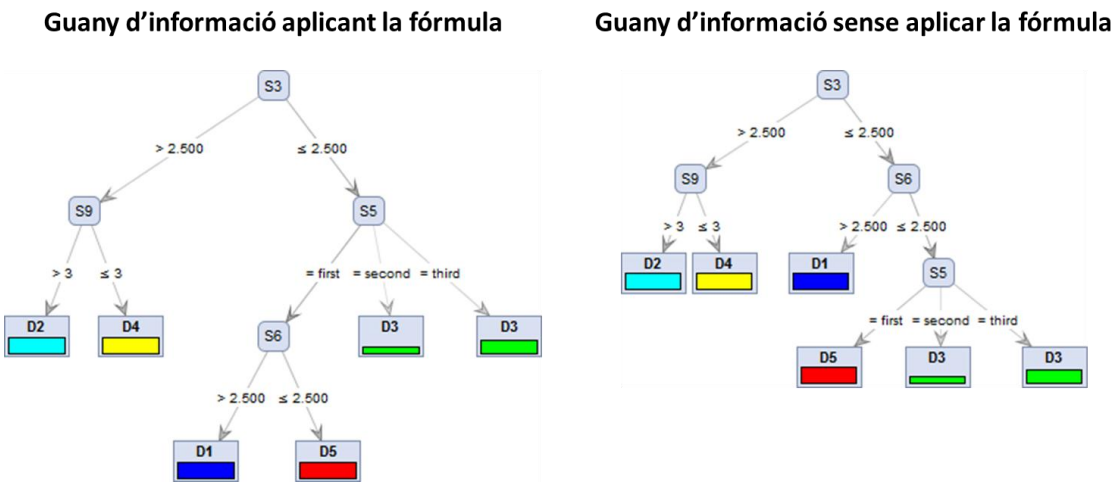

Figura 16. Arbres amb la mateixa precisió però forma diferent segons l'aplicació de la fórmula.

Tot i que la precisió obtinguda és la mateixa, l'exploració "S5" es comprova abans en l'arbre generat utilitzant la fórmula, mentre que es comprova més tard en l'arbre generat sense aplicar-la. Això significa que l'exploració "S5" és més eficient segons els criteris assignats a cada exploració, i que els resultats obtinguts pel que fa a la precisió són els mateixos, de manera que un arbre més eficient en referència a diagnòstics òptims té la mateixa probabilitat d'encertar la predicció d'una patologia que un arbre menys eficient. Aquest, en realitat, és l'objectiu del projecte.

La segona prova correspon a generar un arbre de decisió amb la base de dades de prova sencera sense tenir en compte la fórmula. La Taula 16 mostra els resultats segons els diferents mètodes de discriminació.

Mètodes de discriminació	Precisió sense aplicar la fórmula
Guany d'informació	94.17%
Ràtio de guany	85.83%
Índex de Gini	83.33%
Exactitud	71.67%

Taula 16. Precisió segons el mètode de discriminació sense aplicar la fórmula.

En aquest cas es pot veure com el mètode de discriminació per guany d'informació presenta una precisió més elevada que la resta de mètodes, mentre que el mètode de discriminació per exactitud presenta la precisió més baixa. De totes maneres, cal veure quina precisió s'obté en els arbres on s'aplica la fórmula.

En aquest sentit, la tercera prova servirà per a obtenir la precisió dels arbres generats aplicant la fórmula alhora que es varien els valors dels criteris i els seus factors d'importància. Com ja s'ha fet en la primera execució del conjunt de proves, els valors dels criteris que es modificaran seran els de l'exploració "S17", que apareix freqüentment a l'arbre de decisió generat sense aplicar la fórmula.

La Taula 17 mostra, per al criteri "cost", els diferents resultats obtinguts, o sigui, la precisió dels arbres de decisió generats, per a cada mètode de discriminació. Cal recordar que, en totes les taules on es mostraran els resultats per a tots els criteris, "FI" fa referència al factor d'importància.

Precisió modificant el criteri "cost"		**Inc. cost S17**	**Red. cost S17**
Guany d'informació	**Tots els FI a 0.5**	85.83%	85.00%
	FI cost a 0.1	85.00%	85.00%
	FI cost a 1	82.50%	85.00%
Ràtio de guany	**Tots els FI a 0.5**	92.50%	90.00%
	FI cost a 0.1	90.83%	90.00%
	FI cost a 1	91.67%	90.83%
Índex de Gini	**Tots els FI a 0.5**	81.67%	80.00%
	FI cost a 0.1	77.50%	80.00%
	FI cost a 1	84.17%	80.83%
Exactitud	**Tots els FI a 0.5**	60.83%	61.67%
	FI cost a 0.1	61.67%	62.50%
	FI cost a 1	66.67%	65.83%

Taula 17. Precisió segons el mètode de discriminació modificant el criteri "cost".

Observant els resultats sembla evident que el mètode de discriminació per exactitud presenta els pitjors resultats. Els altres tres, en canvi, presenten resultats força semblants, encara que els mètodes de discriminació per guany d'informació i per ràtio de guany són els que obtenen una millor precisió, sent aquest últim l'únic que presenta en totes les combinacions comprovades una precisió igual o superior al 90%.

Seguint amb el següent criteri, la Taula 18 mostra els mateixos resultats, és a dir, la precisió dels arbres de decisió generats a partir de diverses combinacions, però en aquest cas modificant el criteri "temps".

Precisió modificant el criteri "temps"		Inc. temps S17	Red. temps S17
Guany d'informació	**Tots els FI a 0.5**	85.00%	85.00%
	FI temps a 0.1	92.50%	92.50%
	FI temps a 1	84.17%	87.50%
Ràtio de guany	**Tots els FI a 0.5**	92.50%	90.00%
	FI temps a 0.1	90.83%	90.00%
	FI temps a 1	93.33%	90.83%
Índex de Gini	**Tots els FI a 0.5**	80.00%	80.00%
	FI temps a 0.1	76.67%	76.67%
	FI temps a 1	85.00%	84.17%
Exactitud	**Tots els FI a 0.5**	60.83%	61.67%
	FI temps a 0.1	62.50%	62.50%
	FI temps a 1	65.00%	64.17%

Taula 18. Precisió segons el mètode de discriminació modificant el criteri "temps".

De manera molt semblant al criteri "cost", els resultats obtinguts en modificar el criteri "temps" també presenten una precisió més alta quan s'utilitzen els mètodes de discriminació per guany d'informació i per ràtio de guany, sent encara aquest últim el millor.

Precisió modificant el criteri "invasió"		Inc. invasió S17	Red. invasió S17
Guany d'informació	**Tots els FI a 0.5**	85.83%	87.50%
	FI invasió a 0.1	92.50%	92.50%
	FI invasió a 1	88.33%	95.83%
Ràtio de guany	**Tots els FI a 0.5**	92.50%	90.00%
	FI invasió a 0.1	88.33%	90.00%
	FI invasió a 1	93.33%	90.83%
Índex de Gini	**Tots els FI a 0.5**	81.67%	80.83%
	FI invasió a 0.1	79.17%	80.83%
	FI invasió a 1	85.00%	80.83%
Exactitud	**Tots els FI a 0.5**	60.83%	62.50%
	FI invasió a 0.1	60.83%	59.17%
	FI invasió a 1	65.83%	60.83%

Taula 19. Precisió segons el mètode de discriminació modificant el criteri "invasió".

Continuant amb el pròxim criteri, la Taula 19 mostra els mateixos resultats en modificar els valors i el factor d'importància del criteri "invasió".

Es pot observar com la tendència és la mateixa, tot i que la diferència entre la precisió obtinguda utilitzant el mètode de discriminació per guany d'informació i utilitzant el mètode de discriminació per ràtio de guany és més confusa.

Finalment, els resultats obtinguts modificant l'últim criteri que falta, el criteri "fiabilitat", es mostren a la Taula 20.

Precisió modificant el criteri "fiabilitat"		Inc. fiabilitat S17	Red. fiabilitat S17
Guany d'informació	**Tots els FI a 0.5**	89.17%	85.00%
	FI fiabilitat a 0.1	90.00%	90.83%
	FI fiabilitat a 1	84.17%	90.00%
Ràtio de guany	**Tots els FI a 0.5**	90.00%	90.00%
	FI fiabilitat a 0.1	90.83%	90.83%
	FI fiabilitat a 1	88.33%	88.33%
Índex de Gini	**Tots els FI a 0.5**	79.17%	79.17%
	FI fiabilitat a 0.1	82.50%	84.17%
	FI fiabilitat a 1	72.50%	73.33%
Exactitud	**Tots els FI a 0.5**	65.83%	60.83%
	FI fiabilitat a 0.1	69.17%	66.67%
	FI fiabilitat a 1	63.33%	59.17%

Taula 20. Precisió segons el mètode de discriminació modificant el criteri "fiabilitat".

Repassant els resultats, doncs, es confirma que, per a totes les variacions dels criteris comprovades, els mètodes de discriminació que aporten una precisió més elevada són els que discriminen per guany d'informació i per ràtio de guany. Aquest últim, fins i tot, es podria considerar una mica millor que el que discrimina per guany d'informació.

Per acabar l'execució del conjunt de proves, la quarta i última prova fa referència a la generació de diferents arbres de decisió fent combinacions amb els valors i factors d'importància de dos criteris, com ara el cost i el temps. Com ja s'ha fet en apartats anteriors, per a cada mètode de discriminació s'obtindran els resultats de 36 combinacions diferents, i els valors dels criteris que es modificaran correspondran a l'exploració "S17".

En primer lloc, doncs, la Taula 21 mostra aquests resultats per al mètode de discriminació per guany d'informació separant els valors dels criteris per a l'exploració "S17" i els factors d'importància d'aquests. En aquest cas, els resultats corresponen a la precisió obtinguda de l'arbre de decisió.

Guany d'informació	Valor cost	reduït	reduït	augmentat	augmentat
	Valor temps	reduït	augmentat	reduït	augmentat
FI cost	FI temps	*Precisió de l'arbre de decisió*			
0.1	0.1	90.83%	90.83%	90.83%	90.83%
0.1	0.5	85.00%	85.83%	85.00%	85.83%
0.1	1	87.50%	84.17%	86.67%	84.17%
0.5	0.1	92.50%	92.50%	93.33%	93.33%
0.5	0.5	85.00%	85.00%	85.83%	85.00%
0.5	1	87.50%	84.17%	84.17%	84.17%
1	0.1	85.83%	85.83%	83.33%	83.33%
1	0.5	85.00%	82.50%	82.50%	82.50%
1	1	86.67%	83.33%	83.33%	83.33%

Taula 21. Precisió del mètode de discriminació per guany d'informació modificant dos criteris.

Veient els resultats, sembla un mètode de discriminació prou adequat, ja que la precisió acostuma a ser, per a totes les combinacions provades, d'un 85%, aproximadament.

La Taula 22 mostra els mateixos resultats per al segon mètode de discriminació, és a dir, el que fa referència a la ràtio de guany.

Ràtio de guany	Valor cost	reduït	reduït	augmentat	augmentat
	Valor temps	reduït	augmentat	reduït	augmentat
FI cost	FI temps	*Precisió de l'arbre de decisió*			
0.1	0.1	88.33%	85.83%	87.50%	85.00%
0.1	0.5	90.00%	91.67%	90.83%	91.67%
0.1	1	90.83%	87.50%	90.83%	87.50%
0.5	0.1	90.00%	90.83%	92.50%	91.67%
0.5	0.5	90.00%	92.50%	92.50%	90.83%
0.5	1	90.83%	93.33%	90.00%	93.33%
1	0.1	90.83%	90.83%	88.33%	88.33%
1	0.5	90.83%	90.83%	91.67%	91.67%
1	1	90.83%	93.33%	93.33%	93.33%

Taula 22. Precisió del mètode de discriminació per ràtio de guany modificant dos criteris.

En aquest cas, la precisió mitjana encara és més elevada, i s'acosta a un 90% en gairebé totes les combinacions provades.

Continuant amb el següent mètode de discriminació, que utilitza l'índex de Gini, la Taula 23 mostra la precisió per a les mateixes combinacions provades fins ara.

Índex de Gini	**Valor cost**	**reduït**	**reduït**	**augmentat**	**augmentat**
	Valor temps	**reduït**	**augmentat**	**reduït**	**augmentat**
FI cost	**FI temps**	*Precisió de l'arbre de decisió*			
0.1	**0.1**	75.00%	71.67%	72.50%	71.67%
0.1	**0.5**	80.00%	74.17%	77.50%	74.17%
0.1	**1**	83.33%	82.50%	84.17%	82.50%
0.5	**0.1**	79.17%	76.67%	74.17%	74.17%
0.5	**0.5**	80.00%	80.00%	81.67%	80.00%
0.5	**1**	84.17%	85.00%	85.00%	85.00%
1	**0.1**	80.83%	80.83%	81.67%	81.67%
1	**0.5**	80.83%	84.17%	84.17%	84.17%
1	**1**	84.17%	85.00%	85.00%	85.00%

Taula 23. Precisió del mètode de discriminació per índex de Gini modificant dos criteris.

Els resultats que s'obtenen en aquest cas són una mica més baixos que els anteriors, aproximant-se a una mitjana de precisió del 80%. És clar que també és un bon mètode de discriminació i que la precisió obtinguda no és menyspreable, però els mètodes de discriminació anteriors, per a les mateixes combinacions, presenten resultats millors, sobretot el que utilitza la ràtio de guany.

Finalment, la Taula 24 mostra els resultats obtinguts per a les mateixes combinacions provades fins ara utilitzant el mètode de discriminació per exactitud.

Segons els resultats, aquest mètode de discriminació és el que obté una precisió més dolenta, aproximant-se de mitjana al 60 o al 65%. Fins i tot sense comparar aquest mètode amb la resta es pot veure com no és un mètode prou adequat si es pretén diagnosticar una patologia a partir de diferents símptomes. A més a més, la comparació amb la resta de mètodes ratifica aquesta afirmació i el col·loca a l'última posició de mètodes de discriminació més adequats.

Fins aquí, doncs, tots els resultats que s'han obtingut després d'executar el conjunt de proves definit en apartats anteriors han servit per a avaluar els diferents mètodes de discriminació i obtenir mesures concretes que ajudin a decidir quin d'ells és més adequat per a ser utilitzat en el diagnòstic òptim de patologies.

Exactitud	Valor cost	reduït	reduït	augmentat	augmentat
	Valor temps	reduït	augmentat	reduït	augmentat
FI cost	FI temps	*Precisió de l'arbre de decisió*			
0.1	0.1	60.00%	60.83%	60.83%	60.83%
0.1	0.5	62.50%	61.67%	61.67%	61.67%
0.1	1	63.33%	63.33%	62.50%	63.33%
0.5	0.1	62.50%	62.50%	62.50%	62.50%
0.5	0.5	61.67%	60.83%	60.83%	60.83%
0.5	1	64.17%	65.00%	65.00%	65.00%
1	0.1	65.00%	65.83%	65.83%	65.83%
1	0.5	65.83%	66.67%	66.67%	66.67%
1	1	65.00%	65.83%	65.83%	65.83%

Taula 24. Precisió del mètode de discriminació per exactitud modificant dos criteris.

També es considera que el volum de proves incloses en el conjunt que s'ha executat és suficient per a extreure bones conclusions. Les combinacions que s'han definit han intentat representar al màxim la varietat de possibilitats que permet la modificació dels valors dels criteris i dels factors d'importància de cadascun d'ells.

6.3. Anàlisi dels resultats

Després d'observar els resultats obtinguts a les proves, és important analitzar-los per saber si la fórmula que s'aplica als mètodes de discriminació és adequada i es compleixen les expectatives, així com també per saber quin dels mètodes de discriminació que s'han provat són millors.

Cal tenir en compte que, quan es comparen els resultats obtinguts quan s'ha aplicat la fórmula o quan no s'ha aplicat, el més important és veure si l'estructura de l'arbre de decisió ha canviat. La fórmula s'ha ideat per a generar un arbre de decisió òptim que indiqui el millor camí possible per a fer un diagnòstic molt més eficient. Per tant, la diferència que hi hauria d'haver entre un arbre de decisió generat amb la fórmula i un altre arbre de decisió generat sense la fórmula és bàsicament la localització dels nodes interiors, és a dir, les exploracions. Cal veure si l'aplicació de la fórmula implica que unes determinades exploracions més eficients apareguin abans que unes altres de menys eficients.

En canvi, quan es comparen els resultats obtinguts a partir de diferents mètodes de discriminació, el més important és veure si el resultat final és més precís o no. És a dir, si la patologia predita és correcta o s'ha fet una mala classificació. Per a comprovar això, el millor és calcular la precisió (*accuracy*) amb una validació creuada. D'aquesta manera es pot

contrastar el nombre d'exemples ben classificats i determinar quin mètode de discriminació és més precís.

6.3.1. Anàlisi dels resultats sobre la utilització de la fórmula

Veient els resultats obtinguts de les proves sobre l'extracte de la base de dades de prova, s'han pogut observar diferents fenòmens que fan pensar que els mètodes de discriminació que apliquen la fórmula funcionen correctament. Així, es pot veure com algunes exploracions desapareixen de l'arbre de decisió (o apareixen més tard) quan s'utilitza el mètode de discriminació que aplica la fórmula. Per exemple, si el criteri "cost" té un factor d'importància elevat, es pot veure com desapareix (almenys inicialment) de l'arbre alguna exploració que sigui molt cara, com podria ser una ressonància magnètica.

Aquests fenòmens que fan variar l'arbre depenent dels valors i la importància dels criteris indiquen que el mètode utilitzat pot funcionar i tenir aplicació pràctica, però cal fer proves més extenses i fiables amb la base de dades de prova sencera.

D'altra banda, sí que es pot comprovar com, només modificant els fitxers que emmagatzemen els valors dels criteris per a cada atribut i la importància de cada criteri, els arbres de decisió varien; és a dir, no cal modificar cap part dels algorismes (codi font), sinó només els valors dels fitxers.

La segona prova que s'ha executat corresponia a generar un arbre de decisió amb la base de dades de prova sencera i sense aplicar la fórmula. A l'arbre obtingut es poden observar els diferents camins de l'arbre, i també que l'exploració "S17" és un dels nodes interns que apareix de forma més freqüent a l'arbre, segurament perquè discrimina força bé el conjunt de dades. Per aquest motiu s'ha utilitzat aquesta exploració per a fer les proves, de manera que es pugui observar la seva freqüència a l'arbre en modificar els valors i els factors d'importància dels criteris.

A partir de l'arbre obtingut sense aplicar la fórmula, la tercera prova inclou la generació d'un conjunt d'arbres després de modificar convenientment els criteris. Els resultats mostren que quan s'incrementa el valor d'un criteri negatiu per a l'exploració "S17", com ara el cost, el temps o la invasió, el nombre d'exemples que requereixen aquesta exploració per a predir la patologia es redueixen. Ocorre el mateix quan s'augmenta el factor d'importància d'algun d'aquests criteris. En el primer cas, en augmentar el valor del criteri negatiu per a una exploració, l'algorisme reconeix que aquella exploració és poc eficient i en prioritza d'altres a l'hora d'aparèixer abans a l'arbre de decisió. En el segon cas, quan a més a més d'incrementar el valor del criteri negatiu, també augmenta el factor d'importància d'aquest criteri, llavors l'algorisme encara prioritza més aquelles exploracions que tinguin un valor més reduït d'aquell criteri. Per tant, quan un criteri negatiu té un valor molt baix, el factor d'importància és molt poc rellevant, ja que el mateix valor del criteri ja indica que aquella exploració és eficient pel que fa a aquest criteri, és a dir, el factor d'importància no penalitza l'elecció de l'exploració.

En canvi, es produeix l'efecte contrari en els criteris positius, com ara la fiabilitat. Quan el valor d'aquests criteris augmenta per a una exploració concreta, també augmenta la presència d'aquesta exploració en nivells superiors de l'arbre, perquè l'algorisme considera

que és una exploració més eficient. Això també s'aprecia quan s'augmenta el factor d'importància d'aquests criteris; com més alt és el factor, més presència hi ha de les exploracions que tinguin un valor alt d'aquest criteri. De manera semblant, quan el valor d'aquests criteris és baix per a una exploració concreta, llavors la presència d'aquesta exploració a l'arbre també es veu reduïda.

Finalment, la quarta prova indica que es produeix el mateix efecte quan es fan combinacions entre dos criteris. En aquest cas, els resultats obtinguts en fer combinacions entre els criteris "cost" i "temps" són evidents. Es pot observar, per exemple, la diferència entre els resultats quan els valors dels criteris són tots dos reduïts, o quan ambdós són augmentats. Aquesta diferència es fa més evident quan els factors d'importància dels dos criteris són alts, i es pot observar com el nombre d'exemples que requereixen l'exploració "S17" oscil·la entre els 54 exemples (10.8%) quan els valors dels criteris són alts, i els 494 exemples (98.8%) quan els valors dels criteris són baixos.

Tots aquests resultats, doncs, indiquen de forma òbvia que l'aplicació de la fórmula té sentit i que els arbres de decisió generats són més òptims en funció dels valors i dels factors d'importància assignats a cada criteri i per a cada exploració.

6.3.2. Anàlisi dels resultats sobre els mètodes de discriminació

Començant pels resultats obtinguts de la generació d'arbres de decisió a partir de bases de dades diferents, la conclusió principal és que els mètodes de discriminació presenten diferències significants. En el primer cas (base de dades "Golf" estesa) potser són menys visibles, però igualment observables.

Així, entre el mètode de discriminació per guany d'informació i la seva variant, la que discrimina per ràtio de guany, es pot veure com el primer atribut (l'arrel de l'arbre) varia. Això es pot deure al fet que l'atribut "Outlook" té un rang de valors molt gran (fins a 5 possibles valors: "sunny", "overcast", "rain", "storm" i "snow") i el mètode de discriminació per guany d'informació l'afavoreix. En canvi, el mètode de discriminació per ràtio de guany corregeix aquest biaix i prioritza l'atribut "Temperature", que segurament té un rang de valors més petit.

Pel que fa al segon cas (base de dades "Labor-Negotiations"), la diferència és més evident, demostrant així que cada mètode de discriminació també està implementat de manera diferent, seguint diferents criteris per a seleccionar un atribut o un altre.

Aquests resultats, doncs, ens permeten saber que els diferents mètodes actuen de forma diferent i que té sentit investigar quin d'ells és millor per al nostre cas. En aquest sentit, i tal com s'ha exposat a l'apartat de resultats, s'ha executat altra vegada el conjunt de proves definit per a calcular la precisió dels arbres de decisió generats depenent del mètode de discriminació utilitzat.

Pel que fa a la primera prova, que consistia en generar dos arbres, un aplicant la fórmula i l'altre sense aplicar-la, a partir d'un extracte de la base de dades, els resultats han estat poc concloents. Segons aquests resultats, els quatre mètodes de discriminació obtenen una precisió molt semblant. De fet, els mètodes de discriminació per guany d'informació i per ràtio de guany obtenen una precisió idèntica.

Per aquest motiu, la resta de proves tenen en compte la base de dades de prova sencera que inclou més exemples i, per tant, és més fiable.

A partir d'aquí, doncs, també s'ha obtingut la precisió de l'arbre de decisió generat amb la base de dades de prova sencera i sense aplicar la fórmula, tal com demana la segona prova. En aquest cas, els resultats són més diferents i el mètode de discriminació per ràtio de guany sembla obtenir una millor precisió, però es continua sense tenir massa exemples per a fer les comparacions.

La tercera i la quarta prova, doncs, permeten obtenir una gran quantitat d'exemples de precisió dels arbres de decisió que s'han generat a partir de les diferents combinacions definides, ja sigui modificant cada criteri per separat (tercera prova) o dos criteris a la vegada (quarta prova). Observant el conjunt d'aquests resultats, sembla obvi que el mètode de discriminació per exactitud s'ha de descartar, ja que sempre obté una precisió més baixa que la resta de mètodes. Dels altres tres mètodes, el que discrimina per ràtio de guany sembla el que obté més bona precisió, però això només es pot comprovar realment duent a terme algun test estadístic. D'aquesta manera es pot assegurar, a un determinat nivell de confiança, si els diferents mètodes de discriminació presenten o no diferències estadísticament significatives.

En primer lloc es comprovarà si el conjunt de tots els mètodes presenten resultats similars, és a dir, se suposarà que la hipòtesi nul·la és considerar que els mètodes no són diferents i que ofereixen el mateix resultat. En aquest sentit, el test de Friedman és una bona opció per a comparar diferents classificadors, tot i que requereix també diferents conjunts de dades. En el nostre cas, com que només disposem d'un conjunt de dades (la base de dades de prova), es pot considerar que cadascuna de les diferents combinacions que s'han provat és un conjunt de dades diferent (no és la millor opció, però és útil).

El test de Frideman segueix la fórmula que es mostra a l'Equació 12, on n és el nombre de conjunts de dades (en el nostre cas, cadascuna de les combinacions provades, és a dir, 63), k és el nombre de mètodes de discriminació (en el nostre cas, 4) i R_j és la mitjana dels rànquings dels resultats de cada mètode de discriminació.

$$\chi_F^2 = \frac{12n}{k(k+1)}\left(\sum_j R_j^2 - \frac{k(k+1)^2}{4}\right)$$

Equació 12. Fórmula del test de Friedman.

Concretament, el valor R_j es calcula tal com es mostra a l'Equació 13, on r_i^j es el rànquing del mètode de discriminació j sobre el conjunt de dades (combinació) i.

$$R_j = \frac{1}{n}\sum_i r_i^j$$

Equació 13. Fórmula de la mitjana dels rànquings de cada resultat.

Per exemple, la Taula 25 mostra els rànquings que s'obtindrien per a cada mètode de discriminació en un conjunt de dades concret a partir dels resultats obtinguts.

Mètode de discriminació	Precisió	Rànquing
Guany d'informació	85%	**2**
Ràtio de guany	90%	**1**
Índex de Gini	80%	**3**
Exactitud	65%	**4**

Taula 25. Rànquings assignats a cada mètode de discriminació segons la seva precisió.

Si dos mètodes presenten el mateix resultat, llavors cal fer la mitjana dels rànquings. Així, i seguint l'exemple anterior, si el mètode de discriminació per exactitud tingués una precisió del 80%, llavors tindria el mateix resultat que el mètode de discriminació per índex de Gini; en aquest cas, doncs, com que els rànquings dels mètodes serien 3 i 4, caldria fer la mitjana i s'obtindria un rànquing de 3.5 per ambdós.

Un cop implementat el test, el resultat obtingut pels quatre mètodes de discriminació ha sigut de 160.033. Com que el test de Friedman en aquest cas particular segueix una distribució χ_F^2 amb 3 graus de llibertat (és a dir, $k-1$, on k és el nombre de mètodes de discriminació, que equival a 4), llavors es pot rebutjar la hipòtesi nul·la si el valor obtingut és major que 9.348 (amb un nivell de confiança del 97.5%), tot i que per valors de k petits aquest valor crític no és tan fiable. Tenint en compte que 160.333 és major que 9.348, es pot rebutjar la hipòtesi nul·la i es pot dir que els mètodes de discriminació presenten diferències en els resultats que són estadísticament significatives.

Existeix també una millora de l'estadístic del test de Friedman proposat per Iman i Davenport que segueix la fórmula que es mostra a l'Equació 14, on n és el nombre de conjunts de dades (combinacions executades en el conjunt de proves, que equivalen a 63), k és el nombre de mètodes de discriminació (que equival a 4) i χ_f^2 és el resultat obtingut del test de Friedman.

$$F_F = \frac{(n-1)\chi_f^2}{n(k-1)-\chi_f^2}$$

Equació 14. Millora de l'estadístic del test de Friedman.

Fent aquest càlcul, doncs, el resultat obtingut és de 342.534. Com que aquesta millora, en el nostre cas, segueix una distribució F amb 3 ($k-1$) i 186 ($k-1$ multiplicat per $n-1$) graus de llibertat, llavors es pot rebutjar la hipòtesi nul·la si el valor obtingut és major que 3.187 (amb un nivell de confiança del 97.5%). Com que 342.534 és major que 3.187, en aquest cas també es pot rebutjar la hipòtesi nul·la i es pot confirmar que les diferències entre els mètodes de discriminació són estadísticament significatives.

Tot i que el test de Friedman i la seva millora ja han conclòs que els diferents mètodes de discriminació presenten diferències estadísticament significatives, s'ha cregut convenient comparar també cada parella de mètodes de discriminació utilitzant el test de Student. En aquest cas, però, en comptes d'utilitzar les diferents particions que ofereix la validació creuada, es considerarà que cada partició és una de les combinacions del conjunt de proves; de fet, cada combinació és una manera d'interpretar el mateix conjunt de dades.

Així, la hipòtesi nul·la consisteix en considerar que la parella de mètodes de discriminació presenta resultats semblants i que aquests no són estadísticament diferents. Donats dos mètodes de discriminació F i G, el test de Student segueix la fórmula mostrada a l'Equació 15, on n és el nombre de particions (en aquest cas equivalent a les combinacions, o sigui, 63), $\bar{p}$ és la mitjana dels diferents $p(i)$, $p(i)$ equival a $p_F(i) - p_G(i)$, i $p_X(i)$ és el nombre d'exemples mal classificats per l'arbre de decisió generat amb el mètode de discriminació X (en aquest cas, $X = F$ o $X = G$).

$$\frac{\bar{p} \times \sqrt{n}}{\sqrt{\frac{\sum_{i=1}^{n}(p(i) - \bar{p})^2}{n-1}}}$$

Equació 15. Fórmula del test de Student.

De forma més concreta, i per a entendre-ho millor, el valor de $\bar{p}$ es representa a la fórmula que es mostra a l'Equació 16.

$$\bar{p} = \frac{1}{n}\sum_{i=1}^{n} p(i)$$

Equació 16. Desenvolupament de la fórmula per a $\bar{p}$.

A partir d'aquí, s'ha fet aquest càlcul amb cada possible parella de mètodes de discriminació, i els resultats obtinguts es mostren a la Taula 26.

Mètode de discriminació 1	Mètode de discriminació 2	Test de Student
Guany d'informació	**Ràtio de guany**	5.954
Guany d'informació	**Índex de Gini**	8.285
Guany d'informació	**Exactitud**	31.543
Ràtio de guany	**Índex de Gini**	21.88
Ràtio de guany	**Exactitud**	42.817
Índex de Gini	**Exactitud**	31.804

Taula 26. Resultats del test de Student per a cada parella de mètodes de discriminació.

Com que el test de Student, en el nostre cas, segueix una distribució t de Student amb 62 graus de llibertat ($n - 1$, on n equival al nombre de particions o combinacions que, en el nostre cas, correspon a 63), llavors es pot rebutjar la hipòtesi nul·la per a un parell de mètodes de discriminació si el resultat obtingut és major que 1.999. Per tant, i veient els resultats anteriors, es pot assumir que totes les parelles de mètodes de discriminació presenten diferències estadísticament significatives, és a dir, que no es pot considerar que cap mètode tingui la mateixa precisió que un altre.

De fet, observant els resultats obtinguts de les proves ja es pot veure com el mètode de discriminació per ràtio de guany acostuma a obtenir una precisió més alta que la resta de mètodes, però ha sigut necessari comprovar-ho a través de tests estadístics per a obtenir-ne la conclusió final.

Com a conclusió d'aquesta anàlisi, doncs, es pot considerar que el millor mètode de discriminació és el que discrimina per ràtio de guany, seguit pels que discriminen per guany d'informació, per índex de Gini i per exactitud, en aquest ordre. Evidentment, aquest rànquing només s'aplica al problema que s'estudia en aquest projecte, i sempre suposant que la base de dades de prova que s'ha utilitzat és representativa.

7. Conclusions

Amb la proposta descrita, els detalls d'implementació explicats i els resultats presentats, en aquest apartat es vol concloure el projecte comparant les idees inicials amb els resultats finals que s'han obtingut.

Per tant, en aquest apartat s'avaluarà si s'han complert els objectius que s'havien proposat inicialment i s'exposarà la planificació final del projecte per a què es pugui comparar amb la planificació inicial. Finalment, es farà una valoració personal del projecte.

7.1. Assoliment dels objectius

Es pot determinar si un projecte ha sigut exitós o satisfactori avaluant l'assoliment dels objectius inicials. Així, si els objectius que es proposaven en fases inicials del projecte s'han anat complint, es pot dir que el projecte ha seguit el seu curs i que, en aquest sentit, ha sigut exitós.

Per a fer aquesta avaluació dels objectius, a continuació es tornen a llistar els objectius inicials i es discuteix el seu assoliment en la fase final del projecte.

1. Fer un estudi sobre els arbres de classificació per a determinar quin és el que s'adapta millor al nostre problema.

Tal com s'ha descrit al capítol referent a la proposta de solució, s'ha investigat el funcionament dels arbres de classificació i s'han considerat arbres diferents aquells que es generaven a partir de diferents mètodes de discriminació de les dades. Per tant, l'estudi dels diferents arbres de classificació ha acabat resultant en l'estudi d'alguns dels mètodes de discriminació existents, que s'han presentat convenientment dins l'explicació de la proposta de solució.

A més a més, els resultats d'aquest estudi s'han mostrat al capítol referent als resultats, on s'ha pogut observar que l'aplicació del mètode de discriminació per ràtio de guany generava arbres de classificació amb una millor precisió respecte a l'aplicació dels altres mètodes analitzats. Amb tot, es pot considerar que aquest objectiu inicial s'ha assolit satisfactòriament.

2. Estudiar diferents algorismes que es puguin aplicar als arbres de decisió per a optimitzar els diagnòstics.

En aquest projecte no tan sols era important generar arbres de decisió amb una bona precisió a l'hora de predir una patologia a partir d'un conjunt de símptomes, sinó que també era molt important elaborar diagnòstics òptims seleccionant primer aquelles exploracions més eficients. En aquest sentit, aquest objectiu inicial reclamava estudiar diferents algorismes que permetessin elaborar aquests diagnòstics, i això s'ha aconseguit reunint tots els càlculs en la fórmula de ponderació que s'ha exposat al capítol referent a la proposta de solució.

A través d'aquesta fórmula s'ha demostrat que els arbres de decisió generats són més eficients en el sentit que prioritzen aquelles exploracions més eficients. Per tant, es pot considerar que aquest objectiu inicial també s'ha assolit satisfactòriament.

3. Implementar els algorismes seleccionats per a optimitzar els diagnòstics.

Una vegada descrita la fórmula i després de desglossar-la i explicar-ne cadascuna de les seves parts, un altre dels objectius consistia en implementar-la; és a dir, modificar l'algorisme de generació dels arbres de decisió per tal d'incloure-hi la fórmula de ponderació i esbiaixar els resultats segons els paràmetres d'eficiència, que corresponien a valors subjectius introduïts per l'usuari.

Així, tal com es descriu al capítol referent a la implementació, l'addició d'aquesta fórmula a l'algorisme de generació dels arbres s'ha dut a terme correctament, i s'han generat quatre nous algorismes que aplicaven la fórmula (un per a cada mètode de discriminació analitzat). Aquest objectiu inicial, doncs, també s'ha assolit satisfactòriament.

4. Aplicar els algorismes implementats en un entorn de proves.

Amb la teoria estudiada i els algorismes implementats, el següent pas que requeria aquest objectiu era posar a prova la solució creada i avaluar si els resultats eren positius. A través de l'eina RapidMiner, que no només ofereix la possibilitat de modificar el codi, sinó que també ofereix un entorn de proves sofisticat, s'ha pogut provar la solució tot aplicant els algorismes implementats en un conjunt de dades.

El capítol referent als resultats és la prova d'aplicació d'aquests algorismes, i els resultats obtinguts demostren que l'aplicació de la fórmula permet obtenir arbres de decisió més eficients. Per tant, aquest quart objectiu inicial també s'ha assolit satisfactòriament.

5. Crear una base de dades de prova de símptomes i patologies.

Finalment, tot i ja haver estudiat, implementat i aplicat els algorismes, i haver vist que s'obtenien resultats positius, l'últim objectiu inicial proposava crear una base de dades de prova de símptomes i patologies prou gran i robusta com per a assegurar que els resultats obtinguts siguin fiables.

Tal com s'explica al capítol referent a la proposta de solució, la base de dades que s'ha creat compta amb uns 500 exemples, on es defineixen 10 patologies a partir de 20 símptomes diferents. A més a més, els valors que poden prendre aquests símptomes són ben variats, com ara valors numèrics reals, valors nominals o valors d'escala. D'altra banda, els valors no s'han escollit de forma absolutament aleatòria, sinó que cada símptoma tenia una distribució de probabilitat diferent depenent de la patologia a la qual estigués assignat un exemple concret.

Tenint en compte el seu volum i la seva robustesa, doncs, es considera que la base de dades és suficient per a obtenir resultats fiables i assegurar que l'aplicació de la fórmula genera arbres més eficients. Així, aquest últim objectiu inicial també s'ha assolit satisfactòriament.

7.2. Planificació final

Com és sabut, una de les parts més complicades a l'hora de planificar un projecte és establir la seva línia temporal, és a dir, saber el temps que requerirà cadascuna de les parts del projecte. Tot i això, aquesta planificació inicial és necessària per a fer una estimació i descartar, si fos el cas, projectes massa ambiciosos.

En aquest apartat, doncs, i després de completar el projecte, es mostra la planificació final que s'ha dut a terme, de manera que es pugui comparar amb la planificació inicial, veure'n les diferències, i recollir l'experiència per a millorar futures planificacions d'altres projectes. Seguint la metodologia de la planificació inicial, la Taula 27 mostra la planificació final setmana a setmana.

Setmana	Tasques
Setmana 1 (15/09 – 21/09)	Elecció del tema del treball, elaboració de la descripció general del projecte i estudi dels mòduls de l'assignatura corresponent al Treball Final de Màster.
Setmana 2 (22/09 – 28/09)	Descripció més detallada del projecte, redacció dels objectius, elaboració de la planificació i cerca d'informació i de l'estat de l'art (sobretot en bases de dades públiques i softwares de suport a diagnòstics).
Setmana 3 (29/09 – 05/10)	Millores en la descripció del projecte, i instal·lació i realització de proves amb l'eina RapidMiner (després de fer un estudi sobre possibles eines), sobretot pel que fa als seus algorismes d'arbres de decisió. **Lliurament de la primera entrega.** Estudi del mecanisme d'extensió d'operadors de l'eina RapidMiner.
Setmana 4 (06/10 – 12/10)	Estudi de la diversitat de classificadors i dels seus avantatges respecte al nostre problema, importació de fitxers CSV a RapidMiner, estudi mitjançant Eclipse del funcionament dels diversos mètodes de discriminació, creació d'una primera base de dades de prova, i intent sense èxit de crear una extensió del projecte.
Setmana 5 (13/10 – 19/10)	Estudi, elaboració i aprovació de la fórmula de ponderació (detallant-la formalment), i aprofundiment sobre l'eina escollida per a fer la implementació final dels algorismes, incloent les opcions de la interfície gràfica.
Setmana 6 (20/10 – 26/10)	Aplicació de l'algorisme escollit, obtenció dels primers resultats, i inici de l'elaboració de la segona entrega (redacció de la planificació i de l'estructura de la memòria).
Setmana 7 (27/10 – 02/11)	Elaboració final de la segona entrega (redacció dels primers resultats). **Lliurament de la segona entrega.**

Setmana 8 (03/11 – 09/11)	Agrupament d'entregues anteriors per a confeccionar la primera versió de la memòria final amb els continguts disponibles i amb la descripció de la proposta de solució i del funcionament de la base de dades.
Setmana 9 (10/11 – 16/11)	Descripció dels arbres de decisió i les seves característiques, explicació formal de la ponderació dels mètodes de discriminació, i descripció dels requeriments de l'entorn i dels detalls d'implementació.
Setmana 10 (17/11 – 23/11)	Ampliació de la base de dades, elaboració i execució de proves, i comparació i anàlisi de resultats. Observació de problemes amb els valors absents de la base de dades, i anul·lació dels resultats obtinguts.
Setmana 11 (24/11 – 30/11)	Elaboració de la tercera entrega tenint en compte tan sols una petita part dels resultats obtinguts. **Lliurament de la tercera entrega.** Discussió sobre el tractament dels valors absents i redacció de la nova proposta.
Setmana 12 (01/12 – 07/12)	Nova execució de les proves i obtenció dels nous resultats (aquest cop sense problemes per valors absents), anàlisi d'aquests juntament amb tests estadístics, i descripció de la implementació.
Setmana 13 (08/12 – 14/12)	Organització i redacció final de l'apartat d'estat de l'art, confecció d'annexos, índexs i conclusions de la memòria final.
Setmana 14 (15/12 – 21/12)	Redacció final de la memòria, repàs general del seu contingut, i preparació de la presentació.
Setmana 15 (22/12 – 28/12)	Finalització de la presentació i preparació de la defensa.
Setmana 16 (29/12 – 04/01)	**Lliurament de la quarta i última entrega.**

Taula 27. Planificació final setmana a setmana.

En general, doncs, es pot dir que la planificació final del projecte ha seguit en gran mesura el camí que s'havia traçat en la planificació inicial. Si s'ha de destacar alguna desviació important, potser caldria fer referència a l'obtenció dels resultats finals. A la planificació inicial es preveia obtenir aquests resultats per al lliurament de la tercera entrega, però en el moment d'obtenir-los va aparèixer un problema relacionat amb els valors absents de la base de dades. Abans de continuar amb l'obtenció dels resultats, doncs, va ser necessari estudiar i debatre el tractament d'aquests valors absents i, una vegada decidit això, tornar a obtenir els resultats sense problemes. Per tant, l'obtenció d'aquests resultats es va retardar un parell de setmanes. A part d'això, la resta de modificacions no té cap gran efecte en el desenvolupament del projecte.

Finalment, i tal com s'ha mostrat en la planificació inicial, la Figura 17 mostra el diagrama de Gantt de la planificació final per tal de representar-ho de forma més gràfica i poder fer una comparació més visual de les dues planificacions. En aquest diagrama també es pot veure la data final de les fites, algunes de les quals s'han desplaçat.

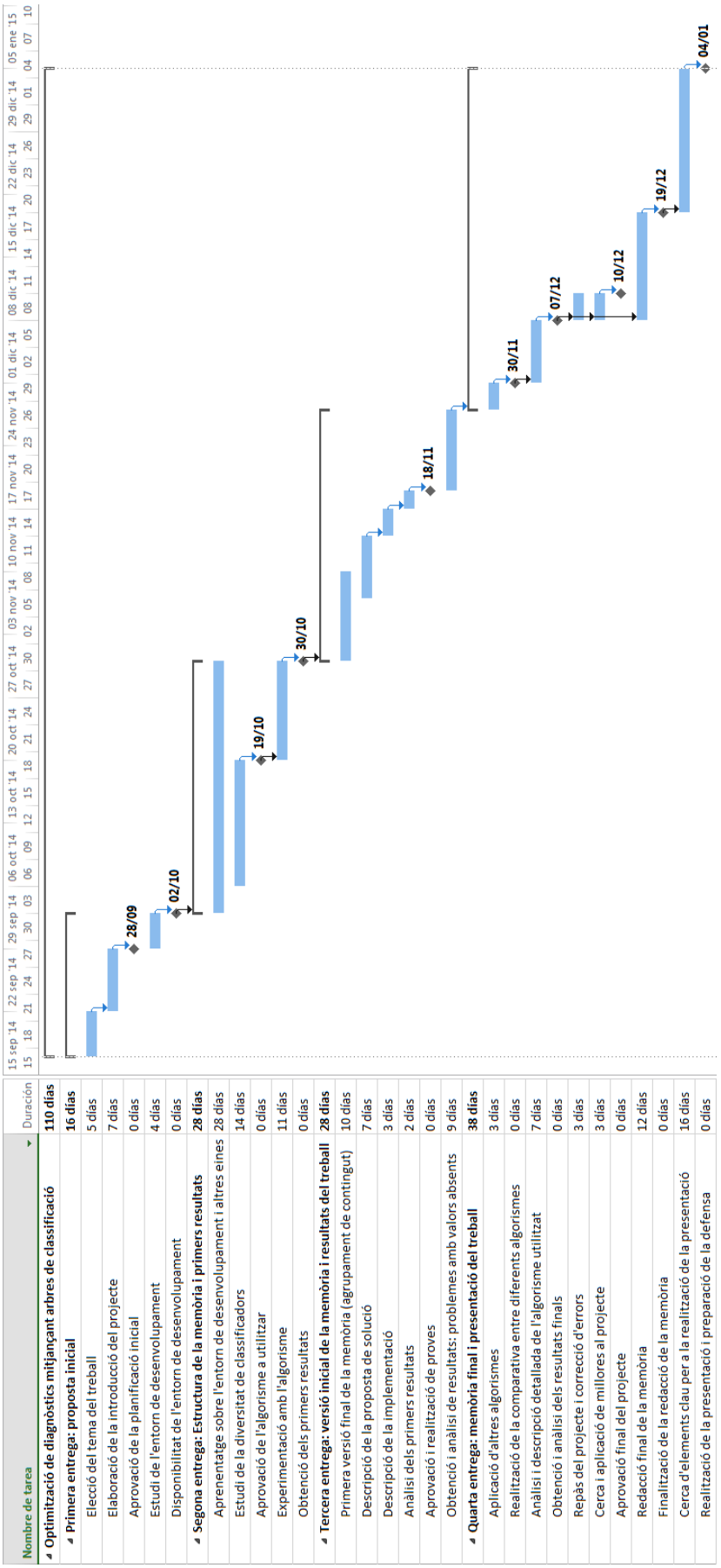

Figura 17. Diagrama de Gantt de la planificació final.

7.3. Valoració personal del projecte

En primer lloc, puc dir que estic molt content d'haver tingut l'oportunitat de barrejar dos àmbits que es compenetren molt bé: la imatge mèdica, àmbit en el qual treballo normalment, i la intel·ligència artificial. Generalment treballo en projectes d'imatge mèdica des d'una perspectiva diferent, sobretot enfocada a l'*e-learning*, però aquest projecte m'ha ajudat a incorporar aquesta nova visió de la intel·ligència artificial. De fet, aquest àmbit m'ha ajudat a veure la potència que tenen alguns mètodes per a fer prediccions a partir de conjunts de dades ja existents. El fet d'endinsar-me en altres àmbits de la informàtica és una pràctica que em resulta molt satisfactòria, perquè tinc el convenciment que sempre es poden aprofitar alguns aspectes per al desenvolupament de nous projectes, siguin de l'àmbit que siguin.

En aquest sentit, un dels aspectes d'aquest projecte que més m'ha sorprès és el coneixement i l'aprofundiment sobre l'eina RapidMiner, una eina que amb tota probabilitat utilitzaré més endavant quan m'enfronti a algun altre problema d'intel·ligència artificial. Es tracta d'una eina senzilla d'utilitzar però que és capaç d'executar operacions molt complexes. En el nostre cas, aquesta eina ha sigut imprescindible per a avaluar els arbres de decisió generats.

A més a més, i en aquest cas, he pogut aprofitar en gran mesura els conceptes apresos a l'assignatura d'Intel·ligència Artificial Avançada impartida en el Màster Universitari en Enginyeria Informàtica. Aquests conceptes els he pogut utilitzar a l'hora d'entendre i implementar els algorismes que requeria el projecte, sobretot tot allò que fa referència als arbres de decisió. Haver repassat el concepte inicial i haver aprofundit en aquest mètode en l'assignatura m'ha ajudat a comprendre molt millor l'enfocament que havia de prendre en iniciar el projecte.

D'altra banda, i deixant una mica de banda els conceptes més tècnics del projecte, sempre és una motivació addicional dur a terme un projecte que té una aplicació pràctica evident. Tot i que l'ús dels arbres de classificació per a predir patologies a partir de símptomes no és cap innovació, certament, sí que considero que ho és el fet d'utilitzar una fórmula de ponderació ideada expressament per a l'ocasió per a aconseguir elaborar diagnòstics òptims. Quan un projecte pot ajudar i optimitzar un procés actual, aleshores això es converteix en un motiu potent i en una motivació més per a realitzar aquest projecte.

Finalment, i després d'haver realitzat altres projectes durant les carreres universitàries d'Enginyeria Tècnica en Informàtica de Gestió i d'Enginyeria Informàtica, tinc la sensació d'haver recollit l'experiència d'aquests projectes per a dur a terme l'organització d'aquest, marcant clarament les fases d'elaboració d'idees, d'implementació, de proves, d'anàlisi de resultats i de redacció de la memòria. Sens dubte, en aquesta organització també hi ha influït clarament la metodologia d'avaluació contínua de la Universitat Oberta de Catalunya; no tan sols el lliurament d'entregues periòdiques del Treball Final de Màster, sinó també la metodologia emprada en la resta d'assignatures del Màster. Crec que l'habilitat de dividir la feina i de fer la planificació temporal ha sorgit principalment d'aquesta experiència.

Resumint, doncs, considero que ha estat una gran experiència que m'ha permès posar en pràctica coneixements adquirits al llarg del Màster, utilitzar-los per a desenvolupar un projecte amb aplicació pràctica, i fer-me adonar que al llarg de la meva experiència acadèmica no tan sols he après conceptes de diferents assignatures, sinó que també he après a desenvolupar-me en una metodologia concreta i a utilitzar-la per a realitzar projectes en una planificació més o menys acurada.

8. Treball futur

Un projecte sempre té possibilitats de millora, ja sigui per les limitacions temporals sota les quals s'ha hagut de planificar i, per tant, deixant de planificar el desenvolupament d'algunes característiques, o perquè durant el desenvolupament del projecte sorgeixen noves idees o noves maneres d'enfocar el problema que no es poden adaptar a la planificació actual.

Per tant, en aquest apartat es descriuen les possibles ampliacions, millores o treballs futurs que es podrien aplicar en aquest projecte:

- La base de dades actual, tot i ser suficientment extensa i robusta, continua sent una base de dades creada per a l'ocasió. Per tant, una de les possibles millores seria construir una base de dades amb exemples reals, tot i que per a fer-ho caldria l'assessorament de professionals de l'àmbit mèdic o bé la introducció d'exemples a mesura que els pacients són consultats pels professionals.

- Tot i que no és l'objectiu d'aquest projecte, una possible ampliació seria construir una aplicació (un software) que pogués ser utilitzada pels professionals de l'àmbit mèdic per a ajudar a predir una patologia a partir de símptomes i ajudar també a elaborar un diagnòstic òptim. En aquest projecte s'ha estudiat si aquesta solució era possible, però no s'ha desenvolupat cap aplicació específica.

- En aquest projecte s'han estudiat fins a quatre mètodes de discriminació diferents. Una possible millora seria ampliar l'estudi amb altres mètodes de discriminació que poguessin obtenir una precisió total més alta. Per exemple, es podrien estudiar mètodes de discriminació que consistissin en modificacions dels mètodes de discriminació que ja s'han analitzat en aquest projecte.

- S'ha demostrat que la fórmula de ponderació utilitzada per a tenir en compte els paràmetres d'eficiència és correcta i serveix per a obtenir diagnòstics més eficients. Ara bé, això no significa que es tracti de la millor fórmula possible, de manera que una possible millora seria estudiar noves fórmules que poguessin optimitzar els diagnòstics.

- Els arbres de decisió ajuden a traçar un camí d'exploracions que sigui òptim. Tot i això, l'experiència del professional que fa ús de l'arbre és la més valuosa i aquest professional serà qui finalment tindrà l'última paraula. En aquest sentit, si un metge, pels motius que siguin, escull no fer una determinada exploració de l'arbre per a un cas concret, llavors caldria oferir-li una alternativa que no fes ús d'aquesta exploració. D'altra banda, una exploració també pot proporcionar resultats de més d'un símptoma; per exemple, una anàlisi pot donar resultats de glòbuls blancs i de colesterol, de manera que no hauria de ser necessari tornar a fer l'exploració quan es requereixi un símptoma ja explorat. Tot això potser podria solucionar-se si s'enfoqués el problema com a un problema de restriccions. Això, però, modifica tota l'estructura de la solució proposada, i caldria plantejar-lo d'una altra manera que potser seria més difícil de comprendre pels professionals de l'àmbit mèdic.

9. Bibliografia

Akthar, F., i Hahne, C. (2012). *RapidMiner 5: Operator Reference*.

Austin, P. C., Tu, J. V., Ho, J. E., Levy, D., i Lee, D. S. (2013). Using methods from the data-mining and machine-learning literature for disease classification and prediction: a case study examining classification of heart failure subtypes. *Journal of Clinical Epidemiology*, 66(4), 398-407.

Buntine, W., i Niblett, T. (1992). A Further Comparison of Splitting Rules for Decision-Tree Induction. *Machine Learning*, 8(1), 75-85.

Comparative Toxicogenomics Database (CTD). (2014, novembre). *Data Downloads*. Recuperat 11 desembre 2014, des de http://ctdbase.org/downloads/

Diseases Database. (2014, desembre). *Diseases Database Search*. Recuperat 11 desembre 2014, des de http://www.diseasesdatabase.com/begin.asp

Drummond, C., i Holte, R. C. (2000, juny). Exploiting the Cost (In)sensitivity of Decision Tree Splitting Criteria. Dins *ICML* (p. 239-246).

Fayyad, U. M., i Irani, K. B. (1992, juliol). The Attribute Selection Problem in Decision Tree Generation. Dins *AAAI* (p. 104-110).

Gemina. (2013, abril). *Description*. Recuperat 11 desembre 2014, des de http://sourceforge.net/projects/gemina/

Harris, E. (2002). Information Gain Versus Gain Ratio: A Study of Split Method Biases. Dins *AMAI*.

Haz-Map®. (2014, agost). *Symptoms/Findings*. Recuperat 11 desembre 2014, des de http://hazmap.nlm.nih.gov/findings

Kotsiantis, S. B. (2013). Decision trees: a recent overview. *Artificial Intelligence Review*, 39(4), 261-283.

Kuo, C. F. J., Wang, P. C., Chu, Y. H., Wang, H. W., i Lai, C. Y. (2013). Using image processing technology combined with decision tree algorithm in laryngeal video stroboscope automatic identification of common vocal fold diseases. *Computer methods and programs in biomedicine*, 112(1), 228-236.

Kusters, V. J. J., Leemans, S. J. J., Schunselaar, D. M. M., i Staals, F. (2009). A practical way of handling missing values in combination with tree-based learners. Dins *BNAIC 2009 Benelux Conference on Artificial Intelligence*.

Laboratory for biophysics of the University of Nijmegen and the University Medical Center Utrecht, The Netherlands. *Promedas*. Recuperat 11 desembre 2014, des de http://www.promedas.nl

Land, S. (2012). *How to Extend RapidMiner 5*.

Mayo Clinic. (2014). *Diseases and Conditions*. Recuperat 11 desembre 2014, des de http://www.mayoclinic.org/diseases-conditions/

Murthy, S. K. (1998). Automatic Construction of Decision Trees from Data: A Multi-Disciplinary Survey. *Data Mining and Knowledge Discovery*, 2(4), 345-389.

Northwestern University: Northwestern University Biomedical Informatics Center (NUBIC). (2013, febrer). *Disease Ontology Wiki*. Recuperat 11 desembre 2014, des de http://do-wiki.nubic.northwestern.edu/do-wiki/index.php/Main_Page

OpenClinical. (2013, juliol). *Quick Medical Reference (QMR)*. Recuperat 11 desembre 2014, des de http://www.openclinical.org/aisp_qmr.html

Pagon, R. A., Adam, M. P., Ardinger, H. H., Bird, T. D., Dolan, C. R., Fong, C. T., Smith, R. J., i Stephens, K. (1993). *GeneReviews*. Recuperat 11 desembre 2014, des de http://www.ncbi.nlm.nih.gov/books/NBK1116/

Phadikar, S., Sil, J., i Das, A. K. (2013). Rice diseases classification using feature selection and rule generation techniques. *Computers and Electronics in Agriculture*, 90, 76-85.

Raileanu, L. E., i Stoffel, K. (2004). Theoretical comparison between the Gini Index and Information Gain criteria. *Annals of Mathematics and Artificial Intelligence*, 41(1), 77-93.

Rapid-I GmbH. (2010). *RapidMiner 5.0: Manual*.

Rendon, R. A., Mason, R. J., Kirkland, S., Lawen, J. G., i Abdolell, M. (2011, abril). A Classification Tree for the Prediction of Benign Disease in the Management of Renal Masses: Aiding the Clinician's Thought Process. *The Journal of Urology*, 185(4), e280-e281.

Rokach, L., i Maimon, O. (2005a). Decision Trees. Dins *Data Mining and Knowledge Discovery Handbook* (p. 165-192). Springer US.

Rokach, L., i Maimon, O. (2005b). Top-Down Induction of Decision Trees Classifiers – A Survey. *IEEE Transactions on Systems, Man, and Cybernetics – Part C: Applications and Reviews*, 35(4), 476-487.

Saar-Tsechansky, M., i Provost, F. (2007). Handling Missing Values when Applying Classification Models. *Journal of Machine Learning Research*, 8, 1625-1657.

Safavian, S. R., i Landgrebe, D. (1991, maig). A Survey of Decision Tree Classifier Methodology. *IEEE Transactions on Systems, Man, and Cybernetics*, 21(3), 660-674.

Scott, C., i Nowak, R. (2004). On the Adaptive Properties of Decision Trees. Dins *Advances in Neural Information Processing Systems* (p. 1225-1232).

Sheng, S., i Ling, C. X. (2005). Hybrid Cost-sensitive Decision Tree. Dins *Knowledge Discovery in Databases: PKDD 2005* (p. 274-284). Springer Berlin Heidelberg.

SimulConsult®. (2014). *SimulConsult®: A Simultaneous Consult On Your Patient's Diagnosis*. Recuperat 11 desembre 2014, des de http://simulconsult.com

The Open Biological and Biomedical. *Human disease ontology*. Recuperat 11 desembre 2014, des de http://obofoundry.org/cgi-bin/detail.cgi?id=disease_ontology

University of Maryland: Institute for Genome Sciences (School of Medicine). (2009, octubre). *Symptom Ontology Wiki*. Recuperat 11 desembre 2014, des de http://symptomontologywiki.igs.umaryland.edu/mediawiki/index.php/Main_Page

Webb, C. (2012, setembre). *Open disease x symptom data (OpenDDx)*. Recuperat 11 desembre 2014, des de http://openddx.net

Annex 1: Codi de la fórmula

En aquest annex es presenta el codi Java de la fórmula de ponderació utilitzada per a escollir, juntament amb el corresponent mètode de discriminació de l'arbre de decisió, el millor atribut del conjunt de dades, és a dir, aquell atribut que sigui més eficient i que separi millor les dades. El codi de la fórmula es presenta a l'Algorisme 4.

```
// Funcio per a inicialitzar els pesos de cada atribut segons els valors i
// els factors d'importancia dels parametres d'eficiencia o criteris
protected void initializeWeights() {
    // Variable que desara la relacio entre cada atribut i el seu pes
    explorationWeights = new HashMap<String, Double>();

    // Noms dels fitxers on es troben els factors d'importancia i els
    // valors dels criteris per a cada atribut
    String explorationCriteriaFactorsFile = "explorationCriteriaFactors.csv";
    String explorationWeightsFile = " explorationWeights.csv";

    // Variables per a llegir els fitxers CSV
    BufferedReader br = null;
    String line = "";
    String csvSplitBy = ";";

    // Variables per a emmagatzemar els factors d'importancia dels criteris i
    // la referencia que indica si es millor quan es alt o quan es baix
    Map<String, Double> explorationCriteriaFactors =
      new HashMap<String, Double>();
    Map<String, String> explorationCriteriaLimits =
      new HashMap<String, String>();

    try {
        // Iniciem la lectura del fitxer amb els factors d'importancia
        br = new BufferedReader(new
          FileReader(explorationCriteriaFactorsFile));

        // Llegim cada linia del fitxer i emmagatzemem els valors a les
        // variables (la primera linia nomes conte titols)
        if ((line = br.readLine()) != null) {
            while ((line = br.readLine()) != null) {
                String[] factor = line.split(csvSplitBy);
                explorationCriteriaFactors.put(factor[0],
                  Double.parseDouble(factor[1]));
                explorationCriteriaLimits.put(factor[0], factor[2]);
            }
        }

        // Iniciem la lectura del fitxer amb els valors dels criteris per a
        // cada atribut
        br = new BufferedReader(new FileReader(explorationWeightsFile));
```

```
        // Comprovem que es pugui llegir la primera linia on consta el nom de
        // cada criteri
        if ((line = br.readLine()) != null) {
            // Definim variables per a emmagatzemar els diferents valors per
            // a cada atribut i els seus minims i maxims
            Map<String, Map<String, Double>> explorationWeightsTable =
              new HashMap<String, Map<String, Double>>();
            Map<String, double[]> minMax = new HashMap<String, double[]>();

            // Omplim les variables amb el nom dels criteris (la primera
            // posicio es en blanc) i amb el valor per defecte dels minims i
            // els maxims, que corresponen a infinit positiu i negatiu,
            // respectivament
            String[] criteria_tmp = line.split(csvSplitBy);
            String[] criteria = new String[criteria_tmp.length - 1];
            for (int c = 1; c < criteria_tmp.length; c++) {
                double[] singleMinMax = {Double.POSITIVE_INFINITY,
                  Double.NEGATIVE_INFINITY};
                criteria[c - 1] = criteria_tmp[c];
                minMax.put(criteria[c - 1], singleMinMax);
            }

            // Anem llegint cadascuna de les linies del fitxer per a coneixer
            // els valors dels criteris per a cada atribut
            while ((line = br.readLine()) != null) {
                // Definim les variables per a separar els diferents valors
                String[] explorationWeightsString = line.split(csvSplitBy);
                Map<String, Double> singleExplorationWeights =
                  new HashMap<String, Double>();

                // Guardem cada valor existent a una de les variables i
                // comprovem si es tracta del valor minim o maxim
                for (int c = 0; c < criteria.length; c++) {
                    String criterion = criteria[c];
                    double value = Double.parseDouble(
                      explorationWeightsString[c + 1]);
                    singleExplorationWeights.put(criterion, value);

                    if (value < minMax.get(criterion)[0])
                        minMax.get(criterion)[0] = value;
                    if (value > minMax.get(criterion)[1])
                        minMax.get(criterion)[1] = value;
                }

                // Guardem a la taula de valors general els valors obtinguts
                // d'aquest atribut concret
                explorationWeightsTable.put(explorationWeightsString[0],
                  singleExplorationWeights);
            }
```

```
            // Una vegada emmagatzemats tots els valors dels criteris a la
            // taula, cal calcular la formula per a cada atribut
            for (String e : explorationWeightsTable.keySet()) {
                double weights_sum = 0;

                // Per a cadascun dels criteris, calculem el resultat de la
                // formula i l'afegim al comptador global
                for (String c : criteria) {
                    // Si el criteri es millor quan es baix, cal restar el
                    // resultat final d'1 per a obtenir valors mes alts quan
                    // el resultat sigui baix
                    if (explorationCriteriaLimits.get(c).equals("low")) {
                        weights_sum += explorationCriteriaFactors.get(c)
                          * (1 - (explorationWeightsTable.get(e).get(c)
                            - minMax.get(c)[0]) / (minMax.get(c)[1]
                              - minMax.get(c)[0]));
                    }
                    else if (explorationCriteriaLimits.get(c).
                      equals("high")) {
                        weights_sum += explorationCriteriaFactors.get(c)
                          * ((explorationWeightsTable.get(e).get(c)
                            - minMax.get(c)[0]) / (minMax.get(c)[1]
                              - minMax.get(c)[0]));
                    }
                }

                // Obtenim el pes de l'atribut dividint la suma total pel
                // nombre de criteris existents
                explorationWeights.put(e, weights_sum / criteria.length);
            }
        }

    } catch (FileNotFoundException e) {
        e.printStackTrace();
    } catch (IOException e) {
        e.printStackTrace();
    } finally {
        if (br != null) {
            try {
                br.close();
            } catch (IOException e) {
                e.printStackTrace();
            }
        }
    }
}
```

Algorisme 4. Codi Java de la fórmula.

Annex 2: Codi dels tests estadístics

En aquest annex es presenta el codi Python utilitzat per a calcular els tests estadístics necessaris per a comprovar les diferències existents entre els arbres de decisió generats a partir de diferents mètodes de discriminació. Es mostra el codi per a calcular el test de Friedman, una millora d'aquest i, finalment, el test de Student per a cada parella de mètodes. El codi dels tests es presenta a l'Algorisme 5.

```
# Per a cada metode de discriminacio, omplim els resultats obtinguts a la
# llista corresponent, on:
# - IG = Information Gain (guany d'informacio)
# - GR = Gain Ratio (ratio de guany)
# - GI = Gini Index (index de Gini)
# - AC = Accuracy (precisio)
IG = [98.33, 98.33, 94.17, 85.83, 85.00, 82.50, 85.00, 85.00, 85.00, 85.00,
  92.50, 84.17, 85.00, 92.50, 87.50, 85.83, 92.50, 88.33, 87.50, 92.50,
  95.83, 89.17, 90.00, 84.17, 85.00, 90.83, 90.00, 90.83, 85.00, 87.50,
  92.50, 85.00, 87.50, 85.83, 85.00, 86.67, 90.83, 85.83, 84.17, 92.50,
  85.00, 84.17, 85.83, 82.50, 83.33, 90.83, 85.00, 86.67, 93.33, 85.83,
  84.17, 83.33, 82.50, 83.33, 90.83, 85.83, 84.17, 93.33, 85.00, 84.17,
  83.33, 82.50, 83.33]
GR = [98.33, 98.33, 85.83, 92.50, 90.83, 91.67, 90.00, 90.00, 90.83, 92.50,
  90.83, 93.33, 90.00, 90.00, 90.83, 92.50, 88.33, 93.33, 90.00, 90.00,
  90.83, 90.00, 90.83, 88.33, 90.00, 90.83, 88.33, 88.33, 90.00, 90.83,
  90.00, 90.00, 90.83, 90.83, 90.83, 90.83, 85.83, 91.67, 87.50, 90.83,
  92.50, 93.33, 90.83, 90.83, 93.33, 87.50, 90.83, 90.83, 92.50, 92.50,
  90.00, 88.33, 91.67, 93.33, 85.00, 91.67, 87.50, 91.67, 90.83, 93.33,
  88.33, 91.67, 93.33]
GI = [93.33, 96.67, 83.33, 81.67, 77.50, 84.17, 80.00, 80.00, 80.83, 80.00,
  76.67, 85.00, 80.00, 76.67, 84.17, 81.67, 79.17, 85.00, 80.83, 80.83,
  80.83, 79.17, 82.50, 72.50, 79.17, 84.17, 73.33, 75.00, 80.00, 83.33,
  79.17, 80.00, 84.17, 80.83, 80.83, 84.17, 71.67, 74.17, 82.50, 76.67,
  80.00, 85.00, 80.83, 84.17, 85.00, 72.50, 77.50, 84.17, 74.17, 81.67,
  85.00, 81.67, 84.17, 85.00, 71.67, 74.17, 82.50, 74.17, 80.00, 85.00,
  81.67, 84.17, 85.00]
AC = [91.67, 95.00, 71.67, 60.83, 61.67, 66.67, 61.67, 62.50, 65.83, 60.83,
  62.50, 65.00, 61.67, 62.50, 64.17, 60.83, 60.83, 65.83, 62.50, 59.17,
  60.83, 65.83, 69.17, 63.33, 60.83, 66.67, 59.17, 60.00, 62.50, 63.33,
  62.50, 61.67, 64.17, 65.00, 65.83, 65.00, 60.83, 61.67, 63.33, 62.50,
  60.83, 65.00, 65.83, 66.67, 65.83, 60.83, 61.67, 62.50, 62.50, 60.83,
  65.00, 65.83, 66.67, 65.83, 60.83, 61.67, 63.33, 62.50, 60.83, 65.00,
  65.83, 66.67, 65.83]

# Obtenim el nombre de proves o combinacions que s'han fet
partitions = len(IG)

# Creem una sola llista amb tots els resultats
classifiers = [IG, GR, GI, AC]
```

```
# Associem un numero representatiu de cada metode a cadascun dels resultats,
# de manera que tots els resultats d'un mateix metode tindran el mateix
# numero (que fa la funcio d'etiqueta) assignat
tagged_classifiers = [list(zip(*[classifiers[i], [i] * len(classifiers[i])]))
  for i in range(len(classifiers))]

# Obtenim la transposada de la llista etiquetada anterior, de manera que a
# cada fila tindrem els resultats de tots els metodes per a un mateix conjunt
# de dades
unordered_classifiers = list(zip(*tagged_classifiers))

# Ordenem de forma descendent cada fila de la llista anterior per situar
# primer els metodes amb resultats mes bons
ordered_classifiers = [sorted(unordered_classifiers[i], reverse = True)
  for i in range(len(unordered_classifiers))]

# Assignem un "ranquing" provisional a cada metode per a cada conjunt de
# dades
tagged_ranks = [[[ordered_classifiers[i][j], j + 1]
  for j in range(len(ordered_classifiers[i]))]
    for i in range(len(ordered_classifiers))]

# Actualitzem el ranquing per a equilibrar posicions repetides; per exemple,
# si dos valors de dos metodes son iguals i ocupen les posicions 2 i 3 del
# ranquing, actualitzem aquests ranquings a 2.5 i 2.5, respectivament
for i in range(len(tagged_ranks)):
    lim = 1

    # Mentre no superi el limit, encara queden ranquings per a actualitzar
    while lim <= len(tagged_ranks[i]):
        # "init" sera el primer ranquing a actualitzar, "index" sera la
        # posicio que s'anira recorrent per a comprovar valors iguals, "rank"
        # sera la suma dels ranquings amb valors seguits iguals, i "count"
        # sera el nombre de valors seguits iguals
        init = lim - 1
        index = lim - 1
        rank = lim
        count = 1

        # Mentre es continui dins el limit de posicions i el valor seguent
        # sigui igual a l'actual, continuem
        while (lim < len(tagged_ranks[i])) and (tagged_ranks[i][index][0][0]
          == tagged_ranks[i][index + 1][0][0]):
            # Incrementem l'"index" per a recorrer una posicio mes en la
            # seguent iteracio, afegim el nou ranquing a la suma de ranquings
            # ("rank"), incrementem el limit actual "lim" per a vigilar no
            # passar-se, i incrementem el comptador "count" per a indicar un
            # nou valor igual
            index += 1
            rank += index + 1
            lim += 1
            count += 1
```

```
        # Des del primer valor comprovat fins a l'ultim, actualitzem els
        # ranquings per a amitjanar-los si son iguals
        for j in range(init, lim):
            tagged_ranks[i][j][1] = float(rank) / float(count)

        # Incrementem el limit actual "lim" per a passar a la seguent posicio
        lim += 1

# Ordenem de forma ascendent la llista amb els ranquings finals segons
# l'etiqueta referent a cada metode, de manera que els resultats continuen
# mantenint el ranquing pero cada fila conte el mateix metode a la mateixa
# columna
ordered_ranked_classifiers = [sorted(tagged_ranks[i],
  key = lambda x: x[0][1]) for i in range(len(tagged_ranks))]

# Obtenim altra vegada la transposada de la llista per a tenir a cada fila
# els resultats, les etiquetes i els ranquings d'un mateix metode per a tots
# els conjunts de dades
dataset_ranks = list(zip(*ordered_ranked_classifiers))

# Calculem el ranquing mitja per a cada metode prenent de referencia tots els
# conjunts de dades
average_ranks = [sum([dataset_ranks[i][j][1]
  for j in range(len(dataset_ranks[i]))]) / len(dataset_ranks[i])
    for i in range(len(dataset_ranks))]

# Calculem el Test de Friedman a partir de la seva formula
nc = len(classifiers)
friedman = ((12 * partitions) / (nc * (nc + 1))) * (sum(list(map(lambda x:
  x ** 2, average_ranks))) - ((nc * ((nc + 1) ** 2)) / 4))

# Calculem una millora del test de Friedman (estadistic no tan conservador)
friedman_improve = ((partitions - 1) * friedman) /
  ((partitions * (nc - 1)) - friedman)

# Mostrem els resultats dels dos testos calculats, aixi com el valor critic
# per a cada distribucio
# Test de Friedman: Distribucio Khi-Quadrat, amb alfa = 0.025 i 3 graus de
#  llibertat (k-1, on k es el nombre de metodes de discriminacio)
# Millora del test de Friedman: Distribucio F, amb alfa = 0.025, i 3 i 186
#  graus de llibertat (k-1 i (k-1)(n-1), on k es el nombre de metodes de
#  discriminacio i n es el nombre de conjunts de dades)
print("\n<<< TESTOS ESTADISTICS >>>")
print("(alfa = 0.025, graus de llibertat calculats a partir de 63 conjunts de
  dades i 4 metodes de discriminacio)")
print("(d.e.s. = diferencia estadisticament significativa)\n")
print("Test de Friedman: d.e.s. si {} > 9.348".format(round(friedman, 3)))
print("Millora del test de Friedman: d.e.s. si {} > 3.187"
  .format(round(friedman_improve, 3)))
```

```
# Retorna el nom de l'algorisme codificat amb una etiqueta numerica
def classifierName(i):
    if i == 0: return "Information gain"
    if i == 1: return "Gain ratio"
    if i == 2: return "Gini index"
    if i == 3: return "Accuracy"
    return "Desconegut"

# Calculem el test de Student per a comparar cada parell de metodes de
# discriminacio en el mateix conjunt de dades
print("\nTest de Student per a cada parell de metodes:")
for i in range(len(classifiers) - 1):
    for j in range(i + 1, len(classifiers)):
        # Calculem la mitjana de les diferencies entre els dos metodes
        p_mitj = sum(list(map(lambda x, y: x - y, classifiers[i],
          classifiers[j]))) / partitions

        # Calculem el valor de la t de Student a partir de la seva formula
        student = abs((p_mitj * (partitions ** (1/2))) /
          ((sum(list(map(lambda x, y: (x - y - p_mitj) ** 2, classifiers[i],
            classifiers[j]))) / (partitions - 1)) ** (1/2)))

        # Mostrem el resultat del test calculat per al parell de metodes,
        # aixi com el valor critic de la distribucio
        # Test de Student: Distribucio t de Student, amb alfa = 0.025 i 62
        #  graus de llibertat (n-1, on n es el nombre de conjunts de dades)
        print("\t{} - {}: d.e.s. si {} > 1.999".format(classifierName(i),
          classifierName(j), round(student, 3)))
```

Algorisme 5. Codi Python dels tests estadístics.