UNIVERSITAT POLITÈCNICA DE CATALUNYA
BARCELONATECH
Escola d'Enginyeria de Barcelona Est

TREBALL FI DE GRAU

Grau en Enginyeria mecànica

APLICACIÓ DE TÈCNIQUES D'ANÀLISI DE DADES A UN CAS EMPRESARIAL

Memòria i Annexos

Autor: Gerard Pérez Garcia
Director: Joan Martínez Sánchez

Convocatòria: Juny2018

Resum

Aquest treball aplica tècniques de mineria de dades a la presa de decisions empresarials. Concretament, s'analitzen diversos sectors industrials, econòmics i geogràfics, i quines variables tenen una influència significativa, i en quina direcció o rangs de valors, fan més probable l'èxit d'aquesta. Això ens servirà com a instrument per il·lustrar una decisió d'inversió o de creació de nous negocis, ja sigui per la diversificació d'una empresa existent amb un excedent de tresoreria o per a il·lustrar la decisió d'on i com invertir en sectors i entorns geogràfics estudiats.

El projecte descriu, breument, el software *RapidMiner*, un dels grans referents en la mineria de dades i l'anàlisi de dades en general, i posteriorment, s'aplica a l'anàlisi de dades d'unes 10 mil empreses de diferents sectors econòmics, principalment industrials o amb una vessant d'enginyeria important, de Catalunya, País Valencià, Navarra i País Basc, creades en els últims 10 anys. Les dades utilitzades provenen de la base de dades SABI i s'han inclòs numerables variables de cada empresa, de les quals, s'ha estudiat la seva influència.

Resumen

Este trabajo aplica técnicas de minería de datos a la toma de decisiones empresariales. En concreto, se analizan diversos sectores industriales, económicos y geográficos, y qué variables tienen una influencia significativa, y en qué dirección o rango de valores, hacen más probable el éxito de una empresa. Esto será un instrumento para ilustrar una decisión de inversión o de creación de nuevos negocios, ya sea para la diversificación de una empresa existente con un excedente de tesorería o para ilustrar la decisión de dónde o cómo invertir en los sectores y entornos geográficos estudiados.

El proyecto describe brevemente el software *RapidMiner*, uno de los grandes referentes en la minería de datos, la analítica predictiva y del análisis de datos en general, y posteriormente se aplica al análisis de datos de unas diez mil empresas de diversos sectores económicos, principalmente industriales o con una componente de ingeniería importante, de Catalunya, País Valenciano, Navarra y País vasco, creadas en los últimos 10 años. Los datos utilizados provienen de la base de datos SABI y se han incluido numerosas variables de cada empresa cuya influencia se ha estudiado.

Abstract

This project applies data mining technics on business decisions. Specifically, analyses the different industrial, economic and geographic sectors, and which variables have a significant influence, and in which way or rank of values make more probable its success. This will help us as a tool to illustrate a decision of inversion or creation of new businesses, not just for the diversification of an existent enterprise with lots of possessions but also to illustrate the decision of where and how to invest in sectors and studied geographic environments.

The research describes, briefly, the RapidMiner software, one of the biggest referents on data mining field and afterwards applies on the data analysis of ten thousand enterprises of different economic sectors, mainly industrial or engineering, based on Catalonia, Valencian Country, Navarre and Basque Country, created in the last 10 years. The data used comes from the SABI database, and a whole bunch of variables from each company, whose influence has been studied, have been included.

Agraïments

M'agradaria fer una menció especial al tutor del meu treball, Joan Martínez Sánchez, per la constant ajuda i motivació proporcionada al llarg del desenvolupament del projecte, així com per facilitar-me totes les eines que han sigut bàsiques per a la realització d'aquest.

També m'agradaria donar un espai dels agraïments a totes les persones del meu entorn que m'han ajudat a realitzar aquest treball de la manera que han pogut.

Gràcies a tots ells.

VIST I PLAU D'AUTORIZACIÓ DE DEFENSA DE TREBALL FI DE GRAU

Jo, Joan Martínez Sánchez Director/a del TFG/TFM dut a terme per l'estudiant/a:

Nom : Gerard

Cognoms : Pérez Garcia

DNI : 48103314X

Grau en Enginyeria : Mecànica

A C R E D I T O:

Que l'estudiant/a es troba en condicions de realitzar, en la present convocatòria, la defensa del treball de fi de Grau/de Màster que a continuació es relaciona:

Títol del TFG/TFM: Aplicació de Tècniques d'Anàlisi de Dades a un Cas Empresarial

INFORMACIÓ NECESSÀRIA		
Codirector (en cas que n'hi hagi)		
Empresa externa (en cas de modalitat B o D)	Nom de l'Empresa	
	Codirector de l'empresa	

I perquè consti, a petició de l'interessat i als efectes d'autorització de defensa de TFG/TFM, signo el present vist i plau.

Barcelona a, 29 de Maig de 2018

El/la Director/a del TFG/TFM

Signatura:

Índex

Índex de figures

1. Introducció

Un dels aspectes que més podrien caracteritzar o representar la societat en la qual vivim des de fa relativament pocs anys és la necessitat que tenim per estar informats a temps real de tots els àmbits que ens interessen, vingui d'on vingui aquesta informació. La majoria d'aquesta informació la rebem de manera digital, sigui mitjançant diaris digitals o xarxes socials entre d'altres i, per tant, tot aquest contingut generat queda emmagatzemat en servidors que tenen com a única funció conservar les grandíssimes quantitats de textos produïts. Aquest mateix aspecte el podem aplicar a una empresa en la qual tota la informació que genera la guarda en els mateixos ordinadors que l'acumulen diàriament en grans quantitats, o inclús en l'àmbit domèstic on també emmagatzemem contínuament coses que generem i per una cosa o altra decidim conservar.

La pregunta que genera tot això és: com podem buscar, organitzar, classificar i utilitzar tota aquesta informació que generem constantment? Doncs bé, aquí entre en joc el *Data Mining*, un concepte realment modern que pretén ajudar-nos a trobar-li una funcionalitat a tot això.

1.1. Objectius del treball

L'objectiu prioritari del treball és endinsar-nos en el món de l'anàlisi de dades i conscienciar-nos del gran potencial que té aquest àmbit, tant en gestió comercial, esportiva o mèdica entre d'altres. S'aprofitarà també per conèixer *RapidMiner*, un dels softwares que tenen com a funció ajudar-nos en la mineria de dades. Donat que aquest tipus de softwares tenen tots una estructura similar ens permetrà familiaritzar-nos amb la interfície que ens trobaríem en qualsevol altra eina.

L'objectiu final serà, amb els coneixements necessaris adquirits, posar-los a prova en un exemple real amb les dades que ens haurà facilitat una empresa i veure si els resultats són fiables com per aplicar-los en cas necessari.

1.2. Motivació

El que ha motivat a entrar i interessar-se per aquest concepte és el gran potencial que té, ja que com s'ha comentat anteriorment, la diferència entre empreses que podrien ser rivals en un sector és saber tractar tota la informació que es té i aplicar-la en benefici propi per atraure més clients, gestionar millor els recursos o explotar àmbits que no hi ha competència.

A més a més la possibilitat de conèixer una eina informàtica nova sempre permet enriquir-se i ampliar els coneixents informàtics, que no cal dir que des de fa temps és bàsic per treballar en qualsevol sector.

1.3. Abast del treball

Aquest projecte, com hem dit a la introducció, és una primera presa de contacte amb la mineria de dades i, per tant, no pretén arribar a desenvolupar un model de dificultat màxima ni dominar *RapidMiner* al complet. Ens centrarem principalment a seguir les indicacions del llibre *Data Mining for the Masses* i de les informacions extretes d'altres fons adjuntes a la bibliografia, així com realitzar els tutorials del propi programa i, amb això, ser capaços d'entendre què és el *Data Mining* i algunes de les possibilitats que ens obre.

Gràcies a les nostres fonts d'aprenentatge, finalment, ens posarem a prova intentant crear un model per a satisfer algun objectiu que ens marquem utilitzant les eines que hem anat veient durant el procés.

2. *Data Mining*

2.1. Concepte i formació de la mineria de dades

La mineria de dades (*Data Mining*) és una metodologia que té com a principal funció trobar patrons que es repeteixen dintre d'un conjunt de grans quantitats de dades, la qual cosa ens permetrà localitzar i identificar aquests patrons per tal d'estar millor informats i ajudar-nos a prendre millors decisions. Tot i que la mineria de dades té com a base altres ciències com la intel·ligència artificial, estadística aplicada, lògica o bé el processament de dades, no cal estar format en totes per poder aplicar-la, ja que com veurem més endavant tenim eines que ens permeten, d'una manera relativament senzilla, treballar-la.

Durant el final dels anys 80 i principis dels 90 el concepte de la mineria de dades es trobava en els seus inicis, existint una idea però sense cap mena de reglamentació ni guia que la regís. No va ser fins a l'any 1999 quan un grup d'empreses -entre les quals es troben Daimler-Benz del sector automobilístic, NCR corp. dedicat a la producció de software i hardware o SPSS de l'àmbit de software estadístic-, van reunir-se per formalitzar i estandarditzar per primera vegada un concepte que fos pròxim al *Data Mining*. El resultat d'aquesta trobada va ser el que es coneix com a CRISP-DM, que és l'abreviació de *Cross-Industry Standard Process for Data Mining*, un procés de 6 fases aplicable a qualsevol eina pensada per la mineria de dades o qualsevol tipus de dades. El fet que sigui independent a qualsevol situació i software permet que diverses investigacions i desenvolupaments treballin en paral·lel provocant un creixement més ràpid i ampli.

Així doncs, les 6 fases del CRISP-DM són:

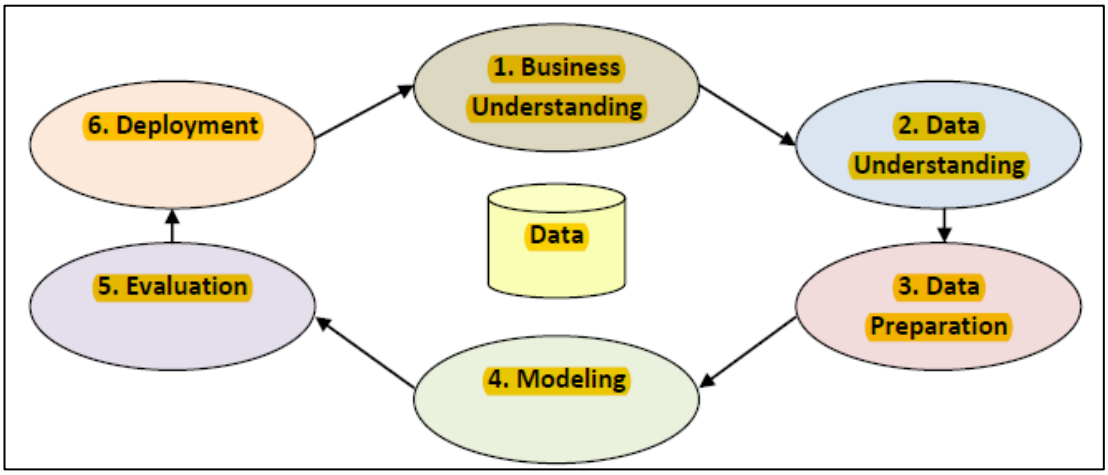


Figura 1. Esquema de CRISP-DM.

2.1.1. Comprensió de negoci

Encara que el títol d'aquesta fase contingui la paraula negoci, cal deixar clar -com s'ha dit anteriorment- que no es limita només a temes empresarials, ja que aquesta disciplina és aplicable a molts altres àmbits.

Aquesta primera etapa és, bàsicament, on deixarem clar què és el que busquem realment. Sembla un pas molt obvi però, moltes vegades, ens el saltem i comencem a treballar i a buscar solucions sense saber realment quin és el problema que volem solucionar. Així doncs, un cop tenim clar quin objectiu perseguim, podem passar a la següent etapa.

2.1.2. Enteniment o comprensió de les dades

Aquesta és una fase preparatòria la qual té un pes molt important, ja que la qualitat dels resultats que obtindrem vindrà donat, en gran part, per la qualitat de les dades que utilitzem i, per tant, hem de tenir clar quina informació necessitem i si és suficientment veraç. El format de les dades el podem classificar en 3 grans grups detallats a continuació:

- *Databases*: és un conjunt d'informació que s'ha organitzat seguint una estructura concreta, normalment, en taules. Les taules que s'utilitzen són, normalment, de relació, amb l'objectiu d'emparellar una dada d'una amb la mateixa d'una altra per evitar redundàncies. Aquest procés rep el nom de *Normalization*. En el cas de les *databases,* les files reben el nom d'arxius i les columnes de camps.

Owner_ID	Owner Name
1	Jim
2	Joan

Pet_ID	Pet_Name	Owner_ID
1	Fifi	2
2	Butch	1
3	Clover	2
4	Animal	1
5	Tank	1

Figura 2. Representació d'una *Normalization*.

- *Data warehouse*: és un tipus de *Database* de gran mida que combina, intencionadament, diferents taules en una sola, encara que això pugui crear duplicats en aquesta. Aquest procés de fusionar-les rep el nom de D*enormalitzation*. En aquest cas anomenarem les files com observacions o exemple mentre que les columnes són les variables o atributs.

Cal vigilar el fet de copiar arxius de dades en taules i anar actualitzant aquests arxius, ja que la taula pot quedar dessincronitzada i, per tant, les dades estaran desactualitzades i poden ser errònies.

Pet_ID	Pet_Name	Owner_Name
1	Fifi	Joan
2	Butch	Jim
3	Clover	Joan
4	Animal	Jim
5	Tank	Jim

Figura 3. Exemple de *Denormalization*.

- Data sets: són un subgrup tant de *Databases* com de *Data warehouse* que normalment trobarem de-normalitzat i, per tant, serà una sola taula on trobarem petits matisos com simplificacions o alguna correcció que ens pot ajudar a ajustar-nos al format de dades que estem utilitzant. Per exemple podrem trobar una data com 14-Juny-1993 simplificat a 14/06/93.

A més a més de poder-nos trobar diferents formats en les bases de dades també podem diferenciar, bàsicament, 2 grans grups d'informació:

- Dades operacionals: vindria a ser tota aquella informació que generem amb sistema transaccional en qualsevol activitat diària. Podria ser ficar gasolina, comprar un vol per internet, utilitzar la targeta client en un supermercat...
- Dades organitzacionals: són aquelles dades que s'han tractat per tenir un cert factor de privacitat. Des de propis governs a organitzacions sense benefici es dediquen a protegir la privacitat dels ciutadans retallant certa informació de caràcter més íntim, però deixant al descobert la resta d'informació perquè sigui utilitzada, per exemple, en la mineria de dades.

2.1.3. Preparació de les dades

Seguidament a la comprensió de dades passarem a la preparació d'aquestes, un treball, si cap, encara més important que l'anterior. La preparació inclou una gran varietat d'activitats, entre d'altres, ajuntar bases de dades, reduir atributs que no interessen, corregir anomalies en les dades, canviar el format d'una base...

Entrarem més en matèria quan comencem a parlar de *RapidMiner*, ja que explicarem detalladament les oportunitats que ens ofereix aquest software per la preparació de dades. Cert és, que cada software té els seus mètodes o maneres de funcionar i, per tant, d'afrontar la preparació de dades,

però sempre ens trobarem amb els mateixos problemes, i les solucions a aplicar seran les mateixes encara que de manera diferent.

2.1.4. Modelatge

Un model, referit a la minera de dades, consisteix en una representació computeritzada d'observacions del món real. Podria ser, per exemple, el cas d'una gràfica que representa les vendes anuals d'un producte. En altres paraules, un model, és l'aplicació d'algoritmes que busquen, identifiquen i mostren qualsevol patró en les bases de dades.

Bàsicament podem diferenciar dos tipus de models: Els que classifiquen, o bé, els que fan una predicció. També existeixen certs models com el cas de l'arbre de decisions, que és un model predictiu que localitza quin atribut té més pes sobre una variable, però a la vegada aquesta variable la dóna classificada en un cert grup.

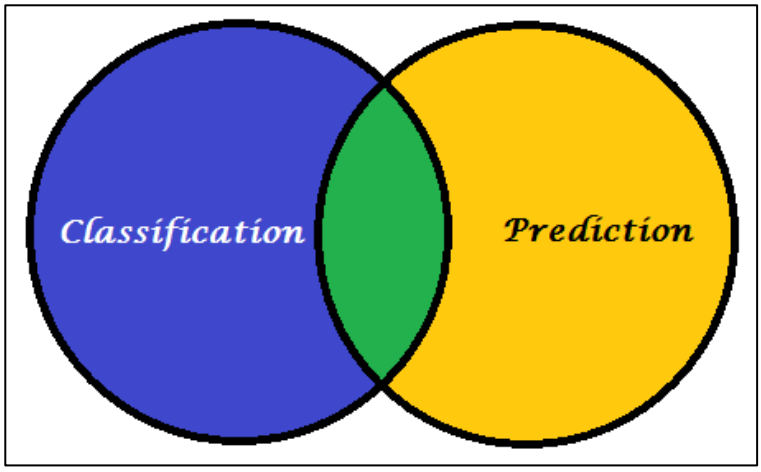


Figura 4. Classificació dels models.

2.1.5. Avaluació

Un cop obtinguts els primers resultats passem a l'avaluació d'aquets. Com en tots els anàlisis es poden donar resultats erronis o fals positius, per això, és necessària que aquesta fase s'apliqui de manera lenta i minuciosa.

Podem comprovar els valors obtinguts aplicant tècniques matemàtiques i lògiques, però en aquest cas, el factor més determinant és el factor humà. L'experiència d'una persona en un àmbit tindrà un pes molt gran a l'hora de donar per vàlids uns resultats, per això, és molt recomanable no només guiar-se amb el que digui el programa sinó ajuntar les dues parts per prendre una decisió.

Quan tinguem clar què fer amb els resultats seguirem endavant a l'última etapa, o bé, tornarem endarrere, segons el que haguem decidit.

2.1.6. Desplegament o aplicació dels resultats

Hem arribat a l'etapa final on aplicarem els resultats obtinguts. Aquesta etapa pot anar des de la presentació d'un projecte amb l'estudi que hem fet, canviar el sistema de funcionament de l'empresa, canviar el públic al qual va destinat un producte...

Cal ser pacients amb tots els canvis que apliquem o la nova tecnologia que presentem, ja que molt rarament, s'obtenen resultats o es veuen les millores de manera immediata, és per això, que si estem segurs del que hem fet i els resultats obtinguts hem de ser persistents en la idea que volem mostrar.

2.2. Ètica en el Data Mining

Abans d'explicar el per què s'ha decidit fer menció de l'ètica en la mineria de dades, definirem breument què entenem per ètica. És un conjunt de codis morals, més enllà dels mínims legals, que un individu utilitza per prendre decisions correctes i respectuoses. Això vol dir, que per molt que estiguem actuant de manera legal no vol dir que estiguem fent el correcte o èticament respectuós.

Aquest apartat s'ha afegit per recordar que quan tractem amb informació sigui del tipus que sigui, darrere aquesta informació hi ha persones i, per tant, no podem utilitzar-la com vulguem. Es pot donar el cas, per exemple, que estiguem treballant en algun tema de salut i, com a tal, necessitem buscar entre dades mèdiques de les persones. De la mateixa manera que no podem trobar certes informacions sobre el tema penjades per internet, nosaltres tampoc podem publicar informació d'aquest estil sense donar anonimat a les persones que apareixen. Això és aplicable a molts altres casos com direccions de casa, números de telèfon, temes legals... i, per tant, hem de respectar sempre la privacitat de les persones.

Per poder establir uns límits per situar què és ètic i que deixa de ser-ho el professor de dret Lawrence Lessig va proposar 4 mecanismes que ens ajuden a emmarcar les activitats que fem dintre uns límits raonables:

- Lleis: són estatuts escrits i aplicats pels governs. Si aquestes lleis són violades, un jutge serà l'encarregat de dictaminar el càstig que cal aplicar.

- Mercats: si accions no ètiques no obtenen benefici i, en canvi, les que estan dins un marc ètic sí en tenen, els mercats forçaran a empreses i organitzacions a romandre dins d'aquests límits.

- Codi: podem escriure un codi de conducte, conegut com a política d'ús, que dictarà el que poden fer els usuaris. Tot i no ser una llei i, per tant, no poder aplicar un judici sobre el seu incompliment, si podem treure privilegis dels usuaris en els nostres serveis si incompleixen aquest codi.

- Normes socials: aquest mecanisme es basa en el que està acceptat en la societat. Si nosaltres mateixos estem intentant ocultar el que fem de cara a la gent, vol dir que estem fent alguna cosa que no ètica.

Amb aquests 4 mecanismes hauríem de ser capaços de saber si allò amb el que estem treballant o el que estem utilitzant, inclús publicant és èticament correcte.

3. RapidMiner

3.1. Iniciació al software

RapidMiner és un software dissenyat per a l'anisis i mineria de dades. S'utilitza tant en empreses com en recerca o en educació, ja que permet un ampli ventall d'aplicacions, amb un dificultat que es pot adequar al nivell de l'usuari. Inicialment aquest software es coneixia amb el nom de YALE (*Yet Another Learning Enviroment*) i es va començar a desenvolupar pel departament d'intel·ligència artificial de la Universitat Tècnica de Dortmund l'any 2001. Posteriorment, l'any 2006 va seguir el desenvolupament una empresa anomenada *Rapid-I*, propietat de dos dels membres que formaven part del departament d'intel·ligència artificial nomenat anteriorment. Finalment, l'any 2007 tant el software com l'empresa van canviar el nom a *RapidMiner*, amb el qual se'l coneix actualment.

Com ja hem comentat anteriorment, existeixen altres softwares que compleixen les funcions de *RapidMiner* però hem decidit utilitzar-lo per aquest treball per les següents raons:

- *RapidMiner* és fàcil d'instal·lar i no requereix un ordinador realment potent per treure-li tot el rendiment.
- *RapidMiner* té una versió gratuïta, que tot i estar limitada en alguns aspectes, permet treballar pràcticament tot el necessari per aprendre a utilitzar-lo.
- *RapidMiner* té una interfície realment intuïtiva, així com ajudes a l'usuari que ens permetran seguir investigant i coneixent totes les possibilitats que ens ofereix.

Ara que ja tenim clar què és *RapidMiner* i el perquè hem decidit treballar amb ell, comencem a explicar com funciona.

Com en la majoria de programes, el primer que demanarà és si volem carregar un document ja començat o començar amb un en blanc. En el nostre cas, com és el primer cop que l'obrim seleccionarem *Blank*.

Aprofitem per recordar que no existeix una versió en espanyol i que el programa està íntegrament en anglès.

La pantalla inicial que trobarem un cop fet aquest pas és la següent:

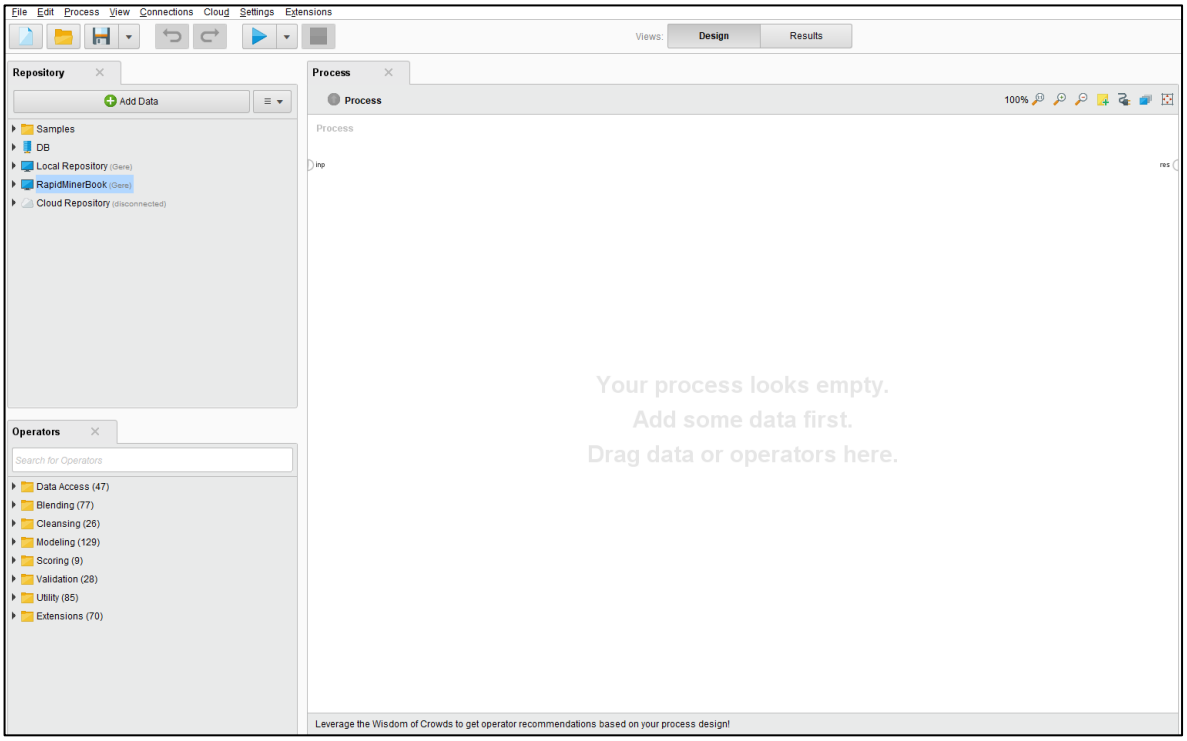


Figura 5. Pantalla inicial de *RapidMiner*.

Com observem en la figura 3.1.1. veiem 4 àrees clarament diferenciades que explicarem a continuació:

- *Repository*: és l'apartat on tindrem guardats tots els documents que anem fent i les bases de dades que hem importat. Per exemple, si avui hem començat un document nou i necessitem utilitzar dades o documents que hem fet anteriorment els trobarem guardat en aquest *Repository* i només l'haurem d'agafar.

- *Operators*: segurament és la part més important i complexa de *RapidMiner* i, és per això, que dedicarem un capítol exclusivament a ells. Dit de manera resumida són totes les eines que ens proporciona el programa per treballar i construir els models que desitgem.

- *Process*: és la zona on connectarem i importarem els operadors per tal de construir visualment el nostre model.

- *Results*: en aquesta pestanya és on trobarem tots els resultats un cop hagués iniciat el nostre model, ja sigui en forma de taules, gràfics o agrupacions de textos depenent dels operadors utilitzats.

A banda d'aquestes 4 àrees principals tenim una gran quantitat d'icones que ens permetran guardar el model, iniciar-lo, fer zoom, importar dades... però són molt intuïtives i, si se'ns presenta qualsevol dubte podrem, ràpidament, relacionar la funció que té.

3.2. Operadors

Com hem comentat anteriorment, un Operador és l'eina bàsica de *RapidMiner*. Existeixen moltíssims, i cadascun d'ells té una finalitat diferent que ens pot resultar útil en algun moment. Per la creació d'un model ajuntarem diferents Operadors mitjançant 'splines', que combinats entre ells, ens permetran obtenir una sèrie de resultats.

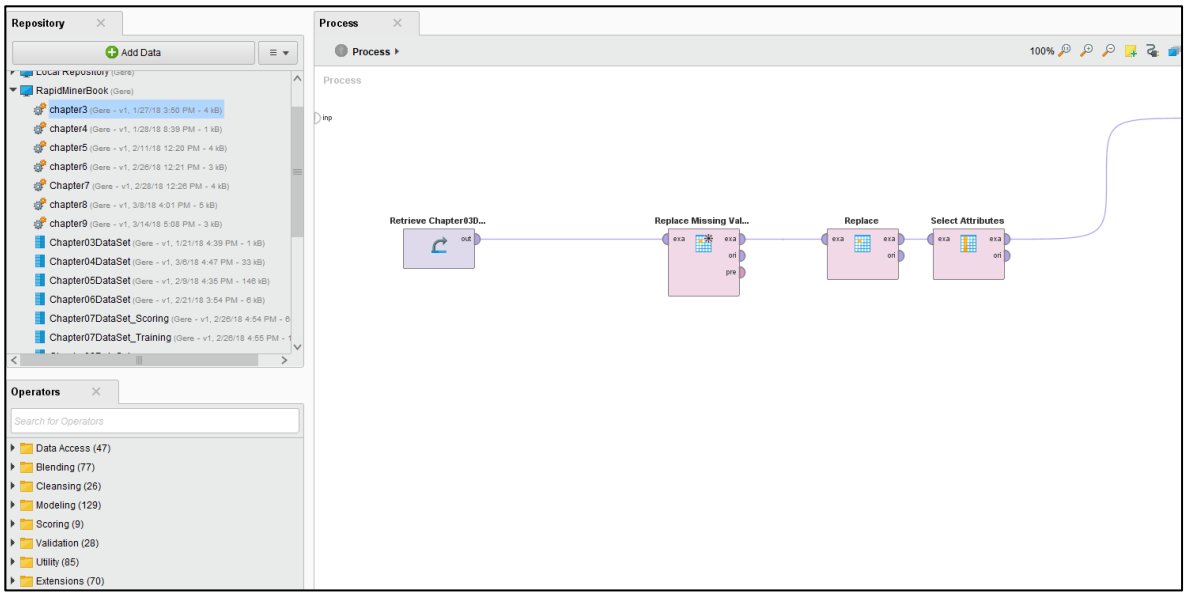


Figura 6. Model creat amb 3 operadors i dades importades.

Donat el grandíssim número d'Operadors que existeixen, és impossible poder veure tots, així que a continuació, mostrarem els que més hem utilitzat al llarg d'aprenentatge amb *RapidMiner* i els que més útils han estat.

3.2.1. Operadors de preparació de dades

Dins d'aquesta categoria que hem creat, trobarem tots aquells operadors que tenen una funció en la preparació de les dades, perquè posteriorment, aquestes dades ja preparades se'ls apliqui els operadors que hàgim col·locat a continuació.

- *Select Attributes* : Aquest operador ens permet, dins de tots els atributs que importem, seleccionar amb quins volem treballar, o vist d'una altra manera, quins no necessitem i, per tant, podem descartar.

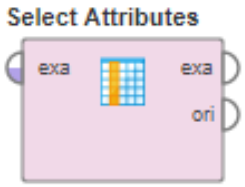


Figura 7. Operador *Select Attributes*.

Tots els operadors tenen el que s'anomenen ports, inputs si són d'entrada i outputs de sortida, que és per on reben i retornen la informació. L 'input que espera aquest operador és un *example set*, referint-se a una taula de dades les quals seran tractades. Els outputs que retorna són un mateix *example set* ja tractat i, si volem també, pot retornar una branca amb el document original (ori) que ha entrat per l'input.

- *Set role*: En aquest cas, l'operador ens permet donar-li un rol als atributs importats. Sense tocar res, el rol predeterminat és el regular, però nosaltres podrem canviar això en la mesura que ens faci falta, per exemple, si tenim una variable que simplement és un identificador i no té cap pes a l'hora de calcular res, li donarem la funció ID. De la mateixa manera si tenim un altre atribut que volem que sigui la variable que depèn dels altres atributs li donarem el rol de 'label'.

Figura 8. Operador *Set Role*.

Els inputs que ha de rebre i els outputs que retorna, com podem veure, són exactament iguals que els explicats anteriorment.

- *Replace* i *Replace missing values*: Aquests dos operadors tenen funcions iguals o molt semblants. El seu funcionament es base en substituir un caràcter, ja sigui una lletra, un numero, un espai en blanc o qualsevol altre per un de nou que nosaltres escollirem. Això ens pot ser de molta utilitat, ja que moltes vegades, en bases de dades molt gran, podem trobar

camps que no estan complets o caràcters que no esperem i ens podrien modificar els resultats que obtindrem a posteriori o donar-nos un error.

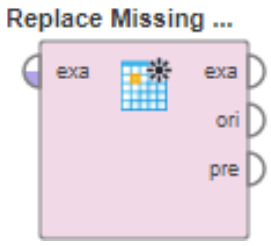


Figura 9. Operador *Replace Missing Values*.

Com podem veure apareix un nou output que rep el nom *de Preprocessing model (pre)* que ens pot ser útil si volem utilitzar la mateixa substitució de valors en un altre *example set*.

- *Numerical to Binomial*: Operadors que canviïn el tipus dels atributs seleccionats existeixen molts, no únicament el de numèric a binomial, però ens servirà d'exemple, ja que tots funcionen de la mateixa manera. Aquest operador ens serà útil com a preparador de dades per a operadors que només treballen amb un sol tipus d'atribut. Per exemple, si tenim un operador que només treballa amb binomis, i tenim algunes dades que són números, utilitzarem el *Numerical to Binomial* abans de fer entrar les dades a l'operador amb el qual estem treballant per adaptar-les.

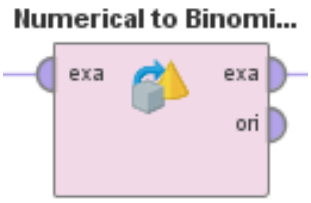


Figura 10. Operador *Numeric to Binomial*.

L'input i outputs d'aquest operador funcionen igual que, per exemple, els de *Set role*

- *Filter Example Range*: Ens permet establir uns rangs, dintre els quals, és on treballarem. Per exemple, si en les nostres dades tenim un atribut que és la nota mitjana dels estudiants, sabem que el valor anirà del 0 al 10. En el cas que alguna dada sigui errònia i estigui fora del marge que nosaltres hem establert com a correcte (0-10), aquest atribut serà eliminat.

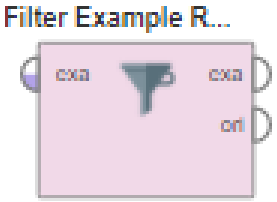


Figura 11. Operador *Filter Example Range*.

Com en la majoria dels operadors que hem comentat disposa de 3 ports de connexió i són iguals que en la resta.

Ara, ja entrarem a treballar amb operadors que ja són purament de càlcul i, no de preparació de dades, com els que hem vist fins ara.

3.2.2. Matriu de correlació

La correlació és una eina matemàtica que ens defineix com és de forta la relació entre dos atributs en una base de dades. Aquesta relació ens la definirà un valor que va de -1 fins a 1. Si la nostra correlació té un valor entre -1 i 0 s'anomena correlació negativa, en el cas contrari que estigui entre 0 i 1 rebrà el nom de correlació positiva. Quan la nostra correlació entre dos atributs és de valor -1 o pròxim, vol dir que els nostres atributs estan relacionats inversament, és a dir, quan un dels dos creix, l'altre decreixerà i al revés quan un decreixi l'altre creixerà. Per altra banda, si el valor és pròxim a 1, vol dir, que la relació entre els atributs és directament proporcional i quan un creixi, l'altre també, o bé, quan un decreixi l'altre farà el mateix. Quan tenim una correlació que és pròxima a 0, per la part positiva o per la negativa, significa que aquests dos atributs no tenen cap relació i varien de manera independent.

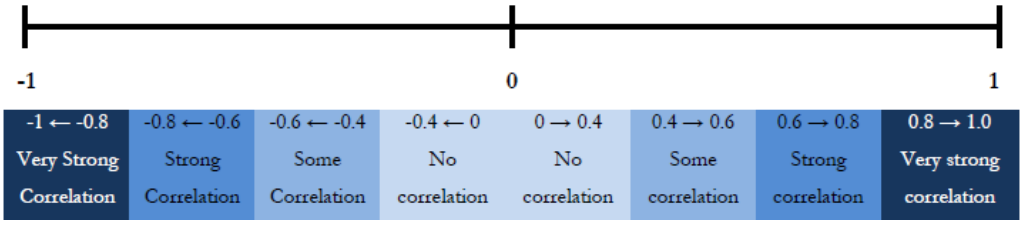


Figura 12. Gràfic de correlació.

Així doncs, quan nosaltres utilitzem aquest operador per saber quina relació hi ha entre els atributs que tenim, el programa ens retornarà una matriu, de tantes files i columnes com atributs tinguem, amb el coeficient de correlació entre cada atribut. A continuació tenim una matriu de correlació de 6 files i columnes d'un exemple que hem treballat.

Attribut...	Insulation	Temper...	Heating...	Num_O...	Avg_Age	Home_...
Insulation	1	-0.794	0.736	-0.013	0.643	0.201
Tempera...	-0.794	1	-0.774	0.013	-0.673	-0.214
Heating_...	0.736	-0.774	1	-0.042	0.848	0.381
Num_Oc...	-0.013	0.013	-0.042	1	-0.048	-0.023
Avg_Age	0.643	-0.673	0.848	-0.048	1	0.307
Home_S...	0.201	-0.214	0.381	-0.023	0.307	1

Figura 13. Matriu de correlació.

Com podem veure la diagonal sempre serà 1, ja que és la correlació d'un atribut amb si mateix i, per tant, és lògic que tingui la màxima relació. També podem veure que és una matriu simètrica, donat que els atributs en la fila i en la columna estan col·locats en el mateix ordre i, la correlació entre un atribut A i un B serà la mateixa sempre independentment de l'ordre de quin atribut mirem primer.

3.2.3. FP Growth

L'operador *FP Growth (frequency pattern)* calcula la freqüència en què diferents atributs apareixen de manera conjunta en les bases de dades i els classifica en grups. Aquesta tècnica està dintre del grup anomenat regles d'associació, que en *RapidMiner*, només treballa amb el tipus de data 'binomial'. Quan una associació entre atributs és molt freqüent la podem considerar com una norma. Per poder dictaminar passa això tenim, bàsicament, dos factors principals:

- Percentatge de confiança: És el grau de confiança que tenim en què si trobem un atribut, trobarem l'altre atribut amb el que està associat. Per poder explicar-ho bé utilitzarem un exemple molt clarificador: Donem per cas que en un supermercat tenim 10 cistelles de compra, en 7 d'elles trobarem llet i en 4 d'elles trobarem galetes. Aquests dos atributs coincideixen en 3 cistelles, per tant, la confiança en la relació galetes $\longrightarrow$ llet és de 75% perquè dels 4 cops que podríem trobar-los junts (només tenim 4 cistelles amb galetes) passa 3 vegades (3/4=75%). Si ara mirem la relació llet $\longrightarrow$ galetes veurem que la confiança passa a ser de 43% perquè dels 7 cops que apareix la llet només coincideixen en 3 (3/7=43%). Com podem veure el coeficient de confiança entre dos atributs varia depenent de quin atribut és la premissa i quin és la conclusió.

- Coeficient de suport: Conceptualment és molt més simple que l'altre, ja que simplement agafa el numero de vegades que s'ha produït la norma, dividida entre el nombre total d'observacions. En el cas del supermercat seria un 30% (coincideixen 3 vegades de les 10 cistelles que tenim)

Així doncs, utilitzant els dos coeficients podem marcar on està el límit entre el que nosaltres considerem una norma i el que no i, per tant, anar provant amb l'operador del *RapidMiner* quines normes es creen en funció del valor que donem nosaltres als coeficients.

Premises	Conclusion	Support	Confidence
Religious	Hobbies	0.239	0.571
Family	Religious	0.225	0.576
Hobbies	Religious	0.239	0.796

Figura 14. Taula amb coeficients de suport i confiança.

3.2.4. Clustering (K-means)

Aquest mètode, com en el cas anterior, també agrupa les dades, però aquest cop, observant els valors individuals de cada atribut i anar-los comparant amb la mitjana d'altres atributs o grups potencials que s'hagin anat formant, per finalment, anar conformant grups definitius amb atributs de característiques similars. La K que dóna nom a *K-means* és el valor que nosaltres posarem depenent del nombre de grups que vulguem formar. Per exemple, si volem separar les dades en 2 grups, la K prendrà el valor de 2.

Attribute	cluster_0	cluster_1	cluster_2	cluster_3
Weight	127.726	106.850	152.093	184.318
Cholesterol	154.385	119.536	185.907	218.916
Gender	0.459	0.543	0.441	0.591

Figura 15. Exemple de model de clúster.

Com veiem a la figura 3.2.9. es tracta d'un exemple amb una k=4 i, per això, tenim 4 grups. Aquest exemple es tractava de classificar pacients tenint en compte el gènere, el pes i el colesterol per saber quins tenien més risc de patir un infart. L'operador doncs, ha anat calculant les mitjanes d'aquests valors individualment, per posteriorment, posar en cada clúster aquells que tenen valors similars. Per exemple, en el clúster 0 estan aquells amb un pes al voltant de 127,726 i un colesterol pròxim a 154,385. Que a gènere doni el valor 0,459 vol dir que en aquest grup hi ha més dones que homes, ja que la dona prenia el valor 0 i el home 1.

3.2.5. LDA (Linear Discriminant Analysis)

LDA és un mètode molt semblant al *K-means*, ja que una de les seves funcions és classificar casos amb atributs de semblant valor, però en aquest cas, no només és capaç de classificar, sinó també, de predir. Això doncs, podríem dir que aquest operador classifica d'una manera predictiva i això és gràcies a un atribut que tindrem ara que sabem que pot ser útil per predir el mateix valor per altres casos. LDA treballa només amb atributs numèrics, excepte en el cas de la variable a predir, que pot ser de tipus diferents.

Quan apareix un factor predictiu en els nostres models necessitem incorporar dos tipus de dades:

- *Training Data*: Són 'data sets' que ja tenen les prediccions dels atributs (també anomenat variable o *label* en *RapidMiner*) fetes, i per tant, ens serveix com a referencia de cara a calcular el valor pels que no el tenen encara.

- *Scoring Data*: És l'altre 'Data Set' que té tots els atributs iguals que en *Training Data* amb l'excepció de la variable, que és el que intentarem predir.

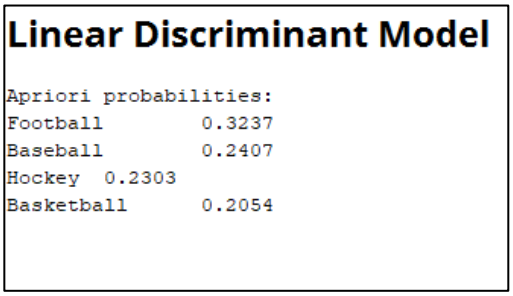


Figura 16. Exemple de LDA

La figura que tenim aquí dalt es un exemple que tractava de predir quin era la probabilitat de què un jugador es dediques a un d'aquests 4 esports tenint en compte una sèrie d'atributs i les variables que ja teníem del *Training Data*. Com podem veure, ens dóna una probabilitat per a cadascuna de les opcions, i en total sempre sumaran 1.

Realment aquest operador sol, veiem que no ens és de gran utilitat, i el seu potencial l'adquireix quan l'ajuntem amb l'operador *Apply Model* que explicarem a continuació.

3.2.6. Apply Model

Aquest operador permet ajuntar dues línies d'operadors i fer que surti només un procés d'ells. Això, bàsicament, ho utilitzarem en el cas que estiguem fent prediccions i tinguem els dos data sets que

hem comentat anteriorment: *Training* i *Scoring*. Així doncs, l'operador rebrà l'entrada d'una sèrie d'atributs amb la seva variable ja calculada (*Training*) i l'utilitzarà de referència per poder calcular la variable dels altres valors que no la tenen (*Scoring*). És important que els dos Data sets que rep l'operador tinguin el mateix número d'atributs, en el mateix ordre i que els atributs siguin del mateix tipus sinó no funcionarà.

Com veiem a la figura a continuació, la qual és el mateix exemple anterior basat en quin esport escollirien els jugadors, ara és de molta més utilitat, ja que específica, en cada cas, quin esport hauria d'escollir tenint en compte els atributs que té i les variables que ens aporta el *Training Data*. En el model només s'ha incorporat l'operador *Apply Model* i les variables del *Scoring Data* que volem predir, i com podem observar, el canvi ja és molt significatiu.

Row No.	prediction(P...	Age	Strength	Quickness	Injury	Vision	Endurance	Agility	Decision_M...
1	Basketball	18.500	5	1	1	0	5	33	61
2	Baseball	13.300	1	2	1	3	5	18	59
3	Football	13.400	2	1	0	2	5	40	11
4	Baseball	16.300	3	1	0	2	5	32	35
5	Football	15.700	1	1	0	2	3	43	37
6	Baseball	17	3	2	0	3	5	21	41
7	Football	16.300	3	1	0	1	1	41	29
8	Baseball	15.700	1	2	1	3	5	17	45
9	Football	16.500	3	2	0	1	3	46	40
10	Football	18.900	5	1	1	2	5	41	6

Figura 17. Taula de resultat de l'exemple amb *Apply Model*.

En la taula, tenim la primera columna que és només l'atribut identificador. Les altres columnes en blanc són els atributs de cada cas a estudiar, que junts amb el *Training Data*, ens han permès predir el valor de la variable, que en aquest cas, és la columna verda i ens mostra l'esport que hauria d'escollir cada cas.

3.2.7. Regressió Lineal

Aquest és un altre mètode predictiu que treballa amb el mateix model que LDA, és a dir, amb dues bases de dades (*Training Data* i *Scoring Data*) i amb el suport d'*Apply Model*. En aquest cas, és molt important que els rangs dels atributs de les dues bases de dades coincideixin exactament igual, i si fes falta, podríem utilitzar un dels operadors que hem explicat de preparació de dades per aconseguir que això sigui així.

Per poder entendre com treballa la regressió lineal en *RapidMiner* veurem la taula que ens ha retornat un exemple i la analitzarem.

Attribute	Coefficient	Std. Error	Std. Coefficient	Tolerance
Insulation	3.323	0.420	0.164	0.431
Temperature	-0.869	0.071	-0.262	0.405
Avg_Age	1.968	0.065	0.527	0.491
Home_Size	3.173	0.311	0.131	0.914
(Intercept)	134.511	7.589	?	?

Figura 18. Exemple de taula de regressió lineal.

Cal recordar que la regressió lineal es regeix per la fórmula de la recta y=m·x+b. Així doncs, necessitem saber que és cada incògnita en la taula que tenim com a resultat per poder utilitzar la fórmula que ja hem comentat. En primer lloc, sabem que la 'y' és la variable o incògnita ('label' en *RapidMiner*), que en aquest exemple era la quantitat de combustible que necessitaria cada casa per escalfar-se. La 'm' és la variable independent, que en aquest cas, seria el valor que nosaltres li donem a cadascuna de les 4 variables que tenim (*insulation, temperature, Avg_Age, Home_size*). Ara sí que necessitem mirar la taula per trobar l'x', equivalent al coeficient de cada atribut, i és el valor de la segona columna que apareix al costat de cada atribut. Finalment la 'b' és el valor que posa també a la segona columna última fila, que rep el nom d'*intercept*.

Ara que ja sabem que és cada cosa farem un exemple perquè quedi més clar. Posant els següents valors als 4 atributs que tenim:

Insulation: 6

Temperature: 67

Avg_Age: 35,4

Home_Size: 5

La fórmula y=m·x + m·x + m·x + m·x...+b serà:

$$Y=(6 \cdot 3{,}323)+(67 \cdot -0{,}869)+(35{,}4 \cdot 1{,}968)+(5 \cdot 3{,}173) + 134{,}511 = 181{,}75 = 182 \text{ unitats}$$

Per tant, amb els valors que nosaltres hem donat als atributs i utilitzant els coeficients que ha calculat *RapidMiner* hem pogut predir les unitats necessàries per satisfer les nostres necessitats.

3.2.8. Arbre de decisions

L'arbre de decisions és un altre mètode predictiu i a la vegada classificador amb un alt contingut visual que permet entendre quin és el camí a seguir per arribar al valor de la nostra variable. Aquest operador, a l'igual que els que estem presentant més recentment, espera les dues branques de

dades, tan les de *Training* com la de *Scoring*. En aquest cas no necessitem treballar amb un tipus d'atribut concret, ja que és capaç de treballar amb tots.

L'arbre de decisions esta formada per 3 parts:

- Nodes: és la part on trobarem els atributs que tenen un influencia important sobre la variable que estem calculant. Normalment es representa dins d'un rectangle o una circumferència.
- Branques: són les línies que uneixen els nodes.
- Fulles: és una barra, normalment d'un o diferents colors, que trobarem al final del recorregut de cada branca i representa la distribució de les categories de la nostra variable al final de cada branca.

Com que és complicat d'entendre sense veure un exemple passem a analitzar-ne un per tenir clar com és el seu funcionament i quina és cada part:

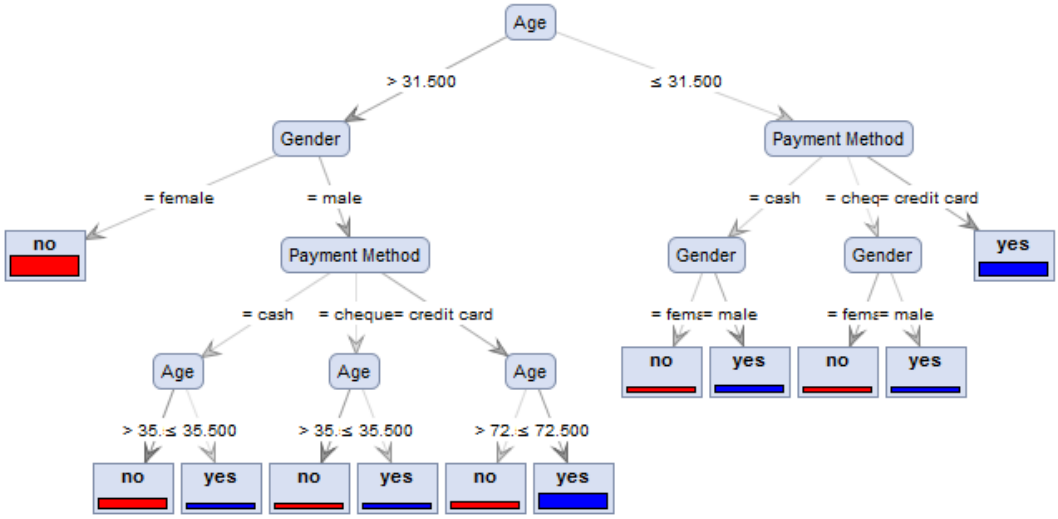


Figura 19. Exemple d'arbre de decisions.

En aquest exemple podem veure que els nostres nodes són; edat, gènere i mètode de pagament. Estan lligats per les branques que hem comentat anteriorment, i a final de cada branca, trobem Sí o No, que és la nostra variable. Si, per exemple, jo vull saber que passa en el meu cas he de començar per l'atribut de sobre de tot i anar baixant. Com jo tinc 24 anys agafaria la branca de la dreta, ja que 24<= 31,5. Seguidament demana quin mètode de pagament faria, i agafaria la branca de l'esquerra que és efectiu. Finalment la branca em porta al últim atribut, gènere, i com sóc home, agafaria la branca de la dreta que em porta a la variable Sí. Així doncs, en el meu cas o d'algú amb els mateixos atributs la predicció serà Sí.

Com podem veure és un mètode molt visual i intuïtiu, tot i que en alguns casos es pot complicar i fer-se molt gran i tediós. Per exemple, en aquest cas la variable només podia ser Sí o No però en altres casos tindrem més de dues i al final de cada branca veurem que les fulles tenen diferents colors representant la distribució que toca en cada cas i de la mateixa manera aquí només teníem 3 variables però si n'haguéssim tingut 10, hauria sigut un arbre molt més llarg.

3.2.9. Neural Network

És un operador molt semblant l'arbre de decisions, capaç de classificar, predir i donar percentatges de seguretat, però també ens permet trobar el grau de connexió entre els atributs. Per primer cop, a més a més, ens trobem un operador amb el qual no cal que els rangs de *Scoring* i *Training* coincideixin, ja que utilitza un concepte anomenat *fuzzy logic* que, d'acord a la probabilitat de connexió entre variables, ens permet inferir la connexió.

Aquest mètode també és molt gràfic i, per tant, hem de saber com interpretar-lo. Per tal d'explicar-lo, utilitzarem un exemple, ja que no és molt intuïtiu d'interpretar sense ajuda visual.

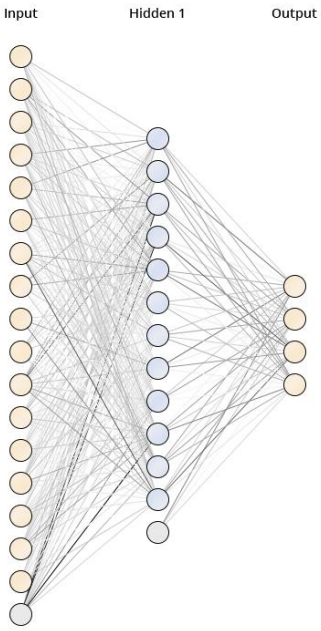


Figura 20. Exemple de *Neural Network.*

Els cercles que veiem en el gràfic són el que anomenem nodes, i les branques que els uneixen són les neurones. L'esquema es comença a llegir des de l'esquerra on trobem una columna de nodes que representa cadascun un atribut. La columna del mig, és la comparació entre atributs, com més gruixuda i negre sigui la neurona des de l'atribut al node de comparació, major pes tindrà. Finalment la columna de la dreta de tot són les 4 possibles variables d'aquest exemple, i de la mateixa manera,

depenent de la neurona, cada variable tindrà més o menys pes. Com és molt difícil guiar-se només pel gruix de cada neurona si posem el ratolí sobre cada variable, ens dirà quin pes té cada atribut.

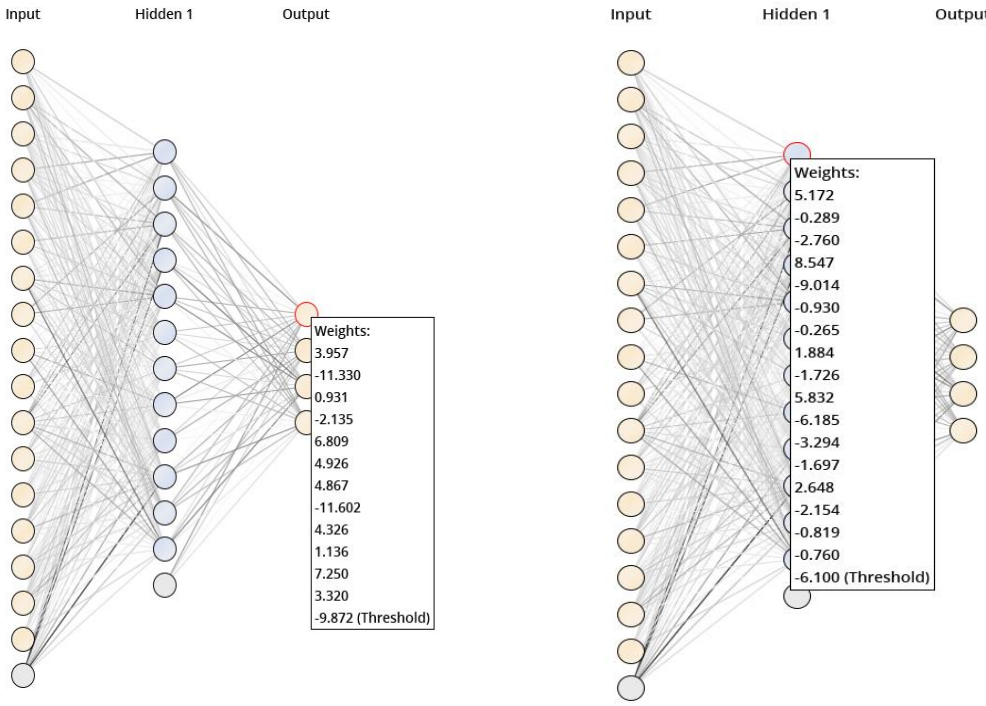


Figura 21. Pes dels atributs en l'exemple.

Com veiem en la figura 3.2.14 a cada node que posem el cursor ens surt el pes que té cadascuna dels atributs de la columna que té just a l'esquerra i, per últim, la mitjana total d'aquests atributs. Així doncs, amb aquest mètode podem arribar a predir les variables i veure quines han tingut més influència gràcies a l'esquema que genera l'operador.

3.2.10. Processador de documents

Aquest operador, tot i que té com a objectiu agrupar, i funciona bé treballant amb alguns dels altres operadors que hem explicat anteriorment, treballa amb dades molt diferents. Principalment, el que fa, és analitzar com estan tractats els textos que importem. Amb aquests textos podem, des de contar quants cops s'ha repetit cada paraula a veure quines paraules apareixen juntes normalment o buscar les paraules que més importància tenen en un text.

Com hem dit anteriorment, el *Process documents*, treballa amb altres operadors en sèrie com el *K-means* però també treballa amb sub-procesos. Aquests sub-procesos són els que realment utilitzarem per tractar les dades. Per veure que ens permeten fer aquests operadors explicarem uns exemples:

- *Tokenize: é*s l'operador encarregat d'agrupar i contar les paraules. Sense aquesta informació numèrica és difícil que el software pugui entendre i tractar les dades.
- *Filter Stopword: q*uan nosaltres escrivim, utilitzem paraules com alguns articles o conjuncions per tal que el text sigui llegible, però que a l'hora de la veritat, no tenen cap pes sobre el text. Així doncs, aquest operador s'encarrega de passar un filtre i eliminar aquestes paraules de més.
- *Transform cases:* per nosaltres 'Dada' i 'dada' és la mateixa paraula, però analitzant dades, això podria ser interpretat com dues paraules diferents i, per tant, necessitem que totes les paraules estiguin en el mateix format, sigui majúscula o minúscula, per no provocar males interpretacions. Per tal d'aconseguir-ho, utilitzarem l'operador *transform cases.*
- *Stemming*: és un operador molt útil que permet agrupa paraules de la mateixa família com si fossin una de sola. Per exemple, 'Amèrica', 'americà' o 'americans' són tres paraules que, en el fons, en mineria de dades, les podem considerar que fan referència al mateix, i per tant, el fet d'agrupar-les ens permet una simplificació.

A part d'aquests 4 que hem explicat, tenim una grandíssima varietat de tractadors de textos que ens permeten fer infinitat de models, i en següent capítol, en veurem un d'ells per donar-nos una idea del potencial o utilitat que això té.

3.3. Exemple teòrics d'aplicació

Ara que ja coneixem uns quants operadors i les funcions de cadascun d'aquests, passarem a resoldre uns exemples que ens proposa el llibre *'Data Mining For The Masses',* que com s'ha comentat anteriorment, ha sigut la base del nostre aprenentatge. Escollirem alguns exemples que siguin d'àmbits diferents perquè així puguem veure com d'útil ens pot resultar la mineria de dades en cada cas.

3.3.1. Exemple 1: Predicció Esportiva

Tenim una escola esportiva on formem atletes per tal de treure el seu màxim rendiment i, durant anys, han sigut sotmesos a proves esportives i de comportament. Amb les dades que hem recollit d'aquests alumnes i amb les que hem recopilat al llarg dels anys d'antics alumnes, volem ser capaços de predir en quin dels 4 esports (Futbol, basquet, hoquei, beisbol) s'hauria d'especialitzar cadascun per treure el màxim profit de les seves qualitats.

Amb les proves que han realitzat s'han recollit els resultats i tenim un total de 8 atributs:

- Edat: Estan entre 13-19 anys.
- Força: Valor comprés entre 0-10.
- Rapidesa: Valor entre 0-6.
- Resistència a les lesions: Valor 1 si ha tingut lesions de gravetat, 0 si són lesions lleus.
- Visió: Rang de 0-4.
- Resistència: Escala de 0-10.
- Agilitat: Valors entre 0-100.
- Presa de decisions: Valor comprés entre 0-100.

I la nostra variable és l'esport a especialitzar-se.

Row No.	Age	Strength	Quickness	Injury	Vision	Endurance	Agility	Decision_M...	Prime_Sport
1	15.100	3	2	1	2	3	29	4	Football
2	15.400	3	2	0	3	5	18	8	Baseball
3	13.600	5	5	0	2	5	27	28	Hockey
4	18.800	5	1	1	1	3	48	36	Hockey
5	16.100	3	1	0	3	3	38	29	Football
6	14.400	1	2	0	3	5	16	14	Football
7	13.700	3	1	1	3	5	20	32	Football
8	15.900	4	2	0	2	5	34	61	Basketball
9	17.300	3	3	0	3	5	19	41	Baseball
10	15.200	4	1	0	2	5	40	4	Hockey

Figura 22. Taula dels atributs registrats a *Training Data*

Com veiem en la figura anterior, totes les nostres variables són números, i com es tracta d'un model classificatori i a la vegada necessitem predir una variable, el millor operador que podem utilitzar és el LDA, que ja hem explicat anteriorment.

Com sempre que treballem amb mineria de dades, seguirem l'esquema CRISP-DM, que s'explica al segon capítol. El primer que cal doncs, és entendre quin problema volem resoldre i de quines dades disposem. Un cop tinguem això solucionat passarem a la tercera fase, la preparació de les dades.

Donat que volem treballar amb dades que tenim recopilades d'altres anys, de les quals ja tenim el resultat de la variable (*Training Data*), importarem primer aquesta base i començarem a treballar sobre ella. Com a recordatori, per tal d'importar dades, tenim l'opció *Add Data* que ens permetrà navegar pel nostre ordinador i seleccionar els arxius que vulguem utilitzar. Un cop afegits, tindrem l'opció de posar-los a la pantalla de processos on crearem el nostre model. Sabem que en la nostra *Training Data* hi ha alguns valors que sobrepassen els rangs que tenim marcats a cada variable i, per tant, és el primer que hem d'arreglar. Utilitzarem l'operador *Filter Example* per eliminar aquests valors i seguidament afegirem el *Set Role* per donar-li el rol de label a la nostra variable. Per tal d'afegir-los, anem a la barra buscadora d'operadors que hi ha a la part esquerra, i el cerquem. Quan els tinguem localitzats només cal arrossegar-los fins a la pantalla blanca de processos. Un cop ja tenim les dades preparades podem afegir l'operador LDA.

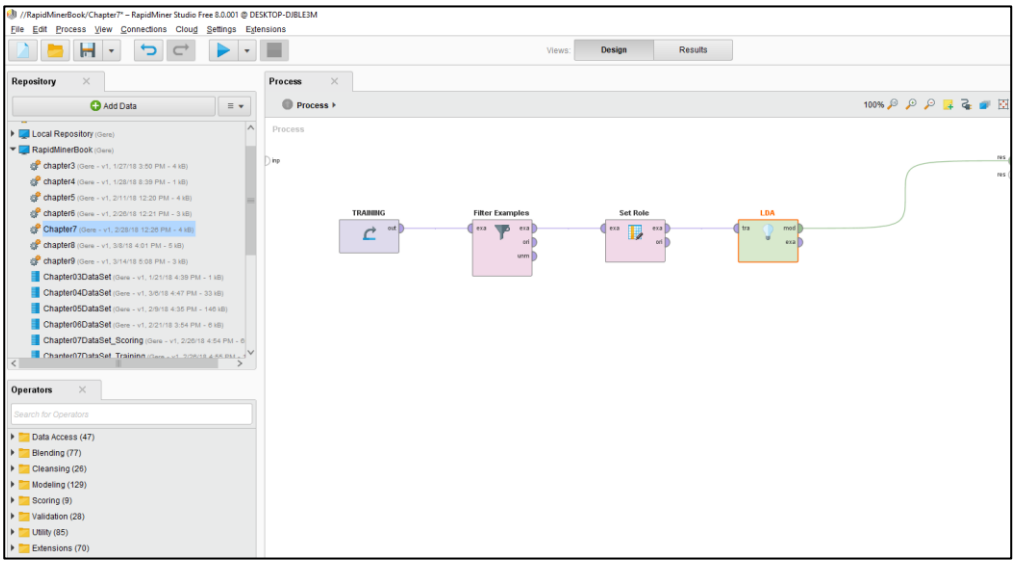


Figura 23. Model provisional de l'exemple 1.

El model que tenim actualment no ens servirà de res, ja que simplement hem afegit unes dades on ja tenim el valor de la variable i no podem predir res del que ens interessa. Així doncs, importem ara la nostra base *Scoring Data* que és on tenim els atributs dels alumnes actuals, dels quals volem predir

l'esport a escollir. Com en el cas anterior, cal que prèviament preparem les nostres dades. Es dóna el cas que coincideix la preparació necessària que hem fet amb el *Training,* però no sempre serà així.

Al capítol on expliquem els operadors, comentem que per tal d'ajuntar les dues *Data* que tenim i poder fer que treballin juntes, necessitem incorporar l'operador *Apply Model* i, per tant, l'importem en sèrie amb els altres que ja tenim. Ara mateix el nostre model és el que es mostra a la figura següent:

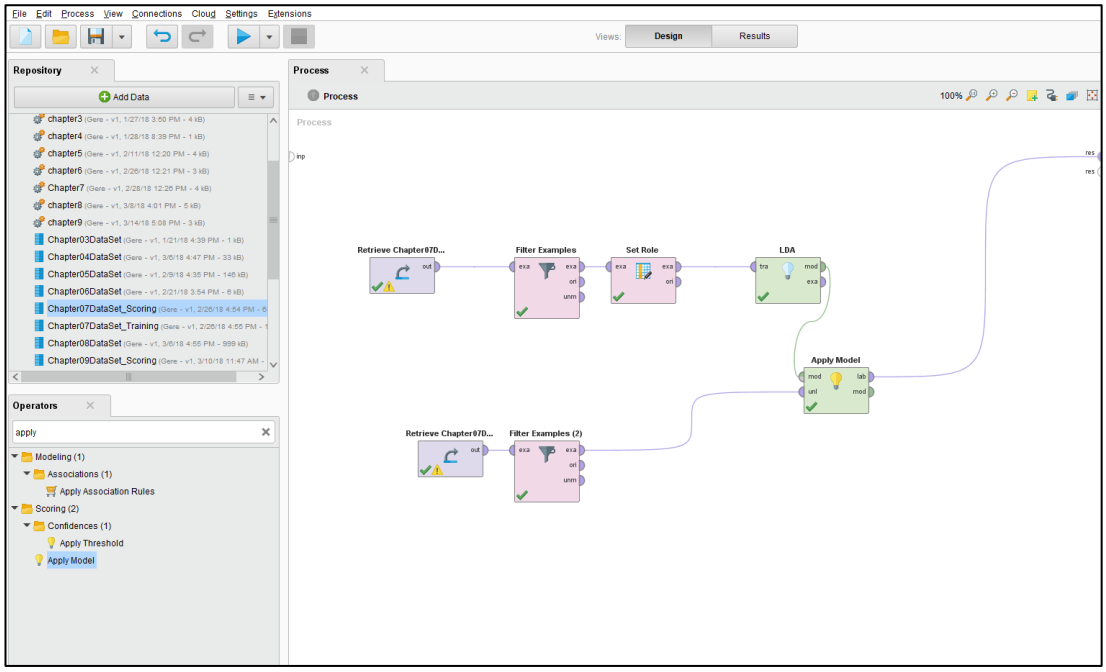


Figura 24. Model final de l'exemple 1.

Si donem *run* per obtenir els resultats veurem que, ara sí, ens retorna les nostres dades de *Scoring* amb la variable que ens ha calculat, en aquest cas, l'esport que cada alumne hauria d'escollir per treure màxim rendiment de les seves qualitats.

Row No.	prediction(P...	Age	Strength	Quickness	Injury	Vision	Endurance	Agility	Decision_M...
1	Basketball	18.500	5	1	1	0	5	33	61
2	Baseball	13.300	1	2	1	3	5	18	59
3	Football	13.400	2	1	0	2	5	40	11
4	Baseball	16.300	3	1	0	2	5	32	35
5	Football	15.700	1	1	0	2	3	43	37
6	Baseball	17	3	2	0	3	5	21	41
7	Football	16.300	3	1	0	1	1	41	29
8	Baseball	15.700	1	2	1	3	5	17	45
9	Football	16.500	3	2	0	1	3	46	40
10	Football	18.900	5	1	1	2	5	41	6

Figura 25. Resultat de l'exemple LDA.

La interpretació dels resultats, en aquest cas, és molt senzilla. Cada fila representa un dels nostres alumnes. La columna de color verd és la nostra variable i la resta els atributs que ha utilitzat *RapidMiner* per a calcular. Segon el model que hem construït, el que ha fet el software ha sigut, d'acord amb els valors que ja teníem d'altres anys, extrapolar-los per a calcular la variable.

Com podem veure, aquest operador, no dóna cap percentatge de confiança, però si es donés el cas que un alumne pot escollir dos esports o més, en els quals obtindria el mateix resultat, en la taula ens marcaria que la predicció té més de una opció.

3.3.2. Exemple 2: Botiga *online*

Tenim una petita botiga *online* on venem llibres, revistes, música i aparells electrònics. Estem a punt de posar a la venta el nou eReader i, per tal de maximitzar l'efectivitat de venda, volem saber quin tipus de client estarà disposat a comprar aquest aparell i en quant temps. Per tal de fer-ho, utilitzarem l'historial dels nostres clients a la nostra pàgina web per saber quins productes compren habitualment i el temps que tarden a adquirir-los d'ençà que surten a la venta.

Per a intentar predir quant tardarà cada client a comprar el nou eReader utilitzarem el mètode d'arbre de decisions. Donat que disposem de l'historial de compra de l'antiga generació de l'aparell i del temps de compra d'aquest, l'utilitzarem com a *Training Data.* La llista dels clients actuals amb els atributs que donarem a continuació serà la *Scoring Data:*

- *USER_ID*: identificació de l'usuari.
- Gènere: 'M' per home i 'F' per dona.
- Edat
- Estat civil: 'M' per als casats i 'S' per la resta.
- *Website_Activity:* té 3 categories depenent de l'assiduïtat*: Seldom, Regular, Frequent*.
- *Browsed_Electronics_12Mo:* Sí o No depenent si la persona ha buscat aparells electrònics a la web l'últim any.
- *Bought_Electronics_12Mo:* Sí o No depenent si la persona ha comprat aparells electrònics des de la web l'últim any
- *Bought_Digital_Media_18Mo:* Sí o No depenent si la persona ha comprat aparells digitals com un MP3 des de la web l'últim any i mig.
- *Bought_Digital_Books:* Sí o No depenent si mai ha comprat un llibre digital.
- Mètode de Pagament: tenim 4 possibles maneres: transferència bancaria, conta de pàgina web, targeta de crèdit i *monthly billing*.

La nostra variable serà el tipus de client segons el temps que tardi a adquirir el producte:

- *Innovator*: és el tipus de client que compra durant la primera setmana des de el llançament del producte.
- *Early Adopter*: són els clients que adquireixen el producte durant la segona i tercera setmana després de la posada a la venta.
- *Early Majority*: Són els que realitzen la compra entre la 3 setmana i els primers 2 mesos.
- *Late Majority*: aquests clients compren el producte a partir de 2 mesos des de el seu llançament.

Ara que ja tenim clar quin és el nostre objectiu i les dades de les que disposem, podem començar a treballar en el model. Com ja sabem, el primer pas a l'hora de modelar, és preparar les dades. En aquest cas, l'operador *Decision Tree* és capaç de tractar amb tot tipus de dades i, per tant, no necessitem cap convertidor. El que si caldrà, és que assignem un atribut com a label (variable) i, per fer-ho, utilitzarem l'operador *Set Role*.

El primer que farem doncs, serà importar el *Scoring i Training Data,* posats en paral·lel amb els corresponents operadors de Set Role en sèrie. Amb aquest últim, assignarem l'atribut *eReader_Adoption* com a label i *USER_ID* com a rol identificador. Seguidament, podem incorporar l'eina Desicion Tree al nostre model, sempre col·locat en sèrie amb el *training Data.*

Ara mateix tenim el nostre model com es mostra a la figura següent:

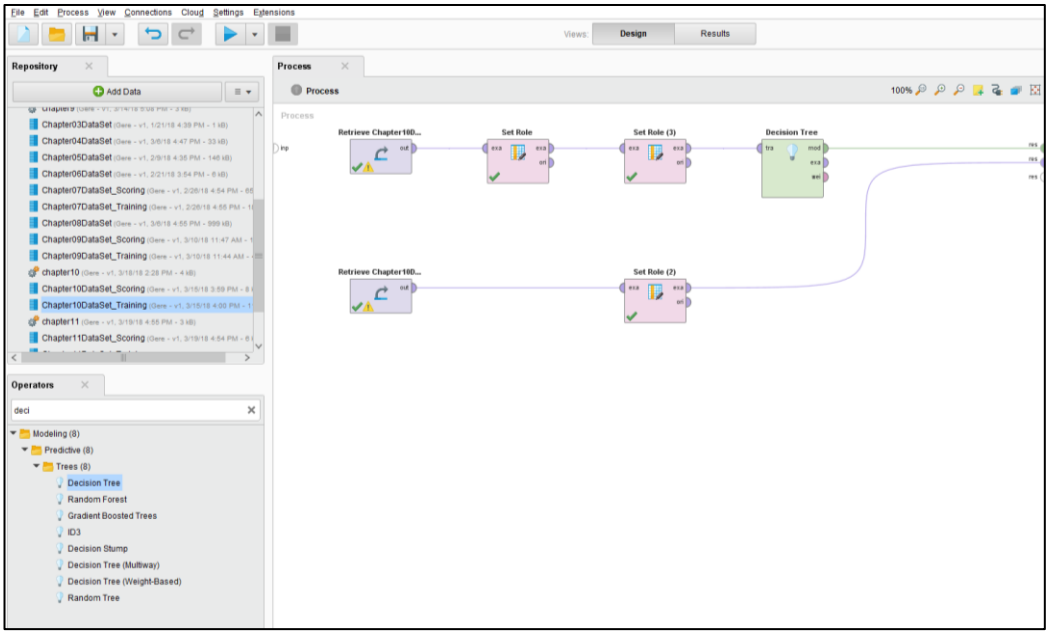


Figura 26. Model en la fase de preparació de l'exemple 2.

Com en l'exemple anterior, ens trobem en què no hem calculat res de valor pel fet que la *Scoring Data* no està interactuant amb l'operador principal del model. Per solucionar-ho, com sempre que treballem amb aquestes dues bases de dades en paral·lel, necessitarem utilitzar l'*Apply Model*. Recordem que cal unir les dues bases de dades l'operador, i connectar les dues sortides del mateix als resultats, sinó, no ens mostrarà tot el que esperem que retorni.

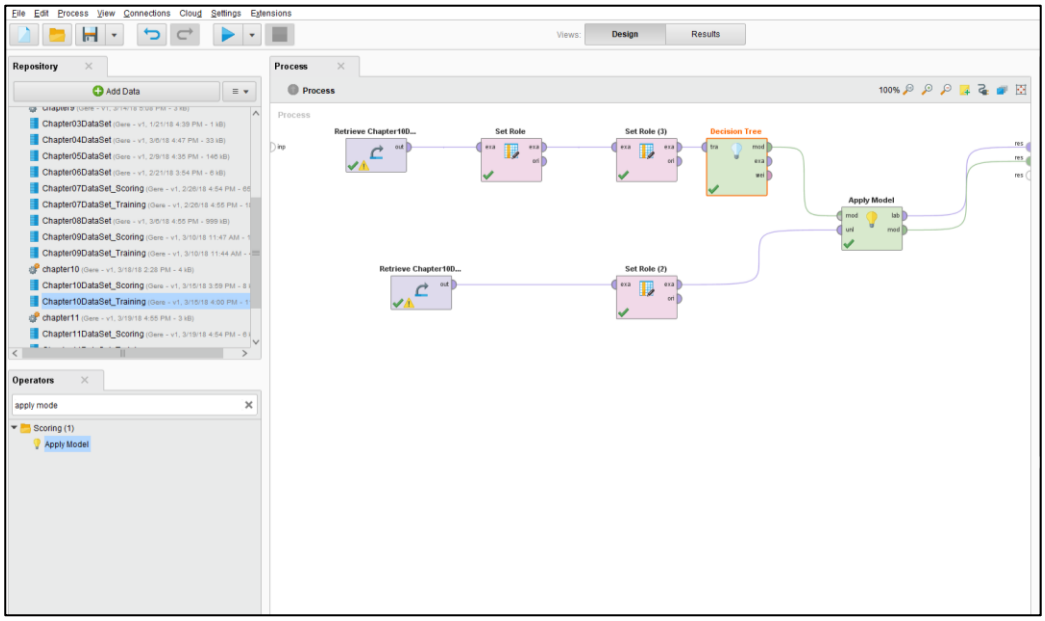


Figura 27. Model definitiu de l'exemple 2.

Ara si podem passar a analitzar els resultats. Abans per això, volem aclarir que tots els operadors que estem utilitzant presenten diferents opcions amb les que treballar i, nosaltres, estem escollint las que venen predeterminats perquè ja ens serveixen.

Tenim dues formes en les que se'ns retornen els resultats:

- La taula que ens proporciona l'*apply model* amb els càlculs del *Decision Tree*

- L'esquema visual de l'arbre de decisions

Ambdues ens seran d'utilitat, però passem a veure quin partit podem treure-li a cadascun.

ExampleSet (Apply Model)

ExampleSet (473 examples, 6 special attributes, 9 regular attributes)

Row No.	User_ID	prediction(e...	confidence(Late Majority)	confidence(Innovator)	confidence(Early Adopter)	confidence(Early Majority)
1	56031	Innovator	0.030	0.576	0.318	0.076
2	25913	Early Adopter	0	0	1	0
3	19396	Late Majority	0.751	0.021	0.053	0.175
4	93666	Early Majority	0.250	0	0	0.750
5	72282	Late Majority	0.751	0.021	0.053	0.175
6	64466	Early Majority	0.250	0	0	0.750
7	76655	Late Majority	0.751	0.021	0.053	0.175
8	48465	Late Majority	0.500	0.500	0	0
9	19889	Late Majority	0.500	0	0.500	0
10	63570	Early Majority	0	0	0	1

Figura 28. Resultat *d'Apply Model*.

En aquesta taula podem observar que tenim, a més a més de la variable (columna verda) i la ID (columna blava), unes columnes que ens donen el percentatge de confiança que té la variable en cada una de les possibles opcions. Per exemple, en el cas de l'usuari 56031 tenim com a resultat que serà un client *Innovator* perquè hi ha un 57,6% de possibilitats que ho sigui. No hem d'obviar que existeix un 31,8% de què sigui un *Early Adopter*, tot i això, com ens diu la lògica, el programa ens mostrarà sempre la que tingui més percentatge de confiança. Més complicat és el cas de l'usuari 19889, que té un 50% de ser *Late Majority* i un 50% de ser *Early Adopter*. El programa ens mostrarà el resultat de la columna, normalment, que primer tingui a la dreta de la variable, però és cosa nostra, decidir en quina categoria posem aquest usuari.

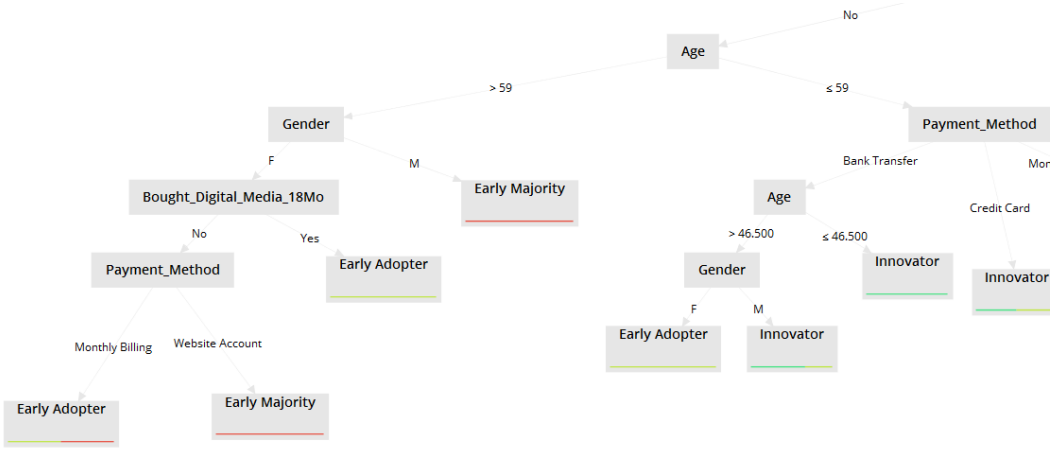


Figura 29. Part de l'arbre de decisions.

En el capítol 3.2. d'aquest treball expliquem com interpretar aquests esquemes i, per tal de no repetir-ho, simplement afegirem que mirant l'arbre de decisions podem veure quin tipus de client predomina, depenent de cada atribut i, per tant, el camí que seguim, al final de cada branca. Així doncs, amb aquest mètode veiem més el conjunt de persones que actua igual amb característiques semblants i, en la taula anterior, es mostra més individualment quin tipus de cient és cada usuari.

3.3.3. Exemple 3: Processament de documents

Al llarg del 1700 es van escriure una sèrie de documents *anomenats Federalist Papers*, que defensaven la ratificació de la constitució dels Estats Units d'Amèrica. Es desconeixen la majoria dels autors d'aquests documents, fins que amb la mort d'Alexander Hamilton es va saber que, ell junt amb James Madison i John Jay, eren autors d'alguns d'aquests papers. Concretament, se sap que els papers 3,4 i 5 van ser escrits per John Jay, el paper 14 per Madison i el 17 per Hamilton. Hi ha evidencies que el paper 18 el van escriure conjuntament Hamilton i Madison, i això, és el que volem confirmar.

Volem utilitzar *Rapidminer* i el seu operador *Process Docuemnts* per analitzar els documents que sabem segur el seu autor i, comparar-los amb el document 18, per tal de certificar si els autors són els que es creu o no hi ha evidencies d'això. El primer que cal fer, en aquest cas, és incorporar l'extensió de mineria de textos del nostre software. Per fer-ho, clicarem a la pestanya d'*Extensions* de *RapidMiner* i buscarem la de *text mining*. Un cop instal·lada, ens apareixerà l'operador *Read Documents* que serà l'encarregat de llegir els textos que importem. Com que volem importar 6 textos en total, haurem d'arrossegar 6 vegades l'operador, i en cadascun d'ella obrir un dels textos. Ara, afegim l'operador *Process Documents* i connectem els 6 textos.

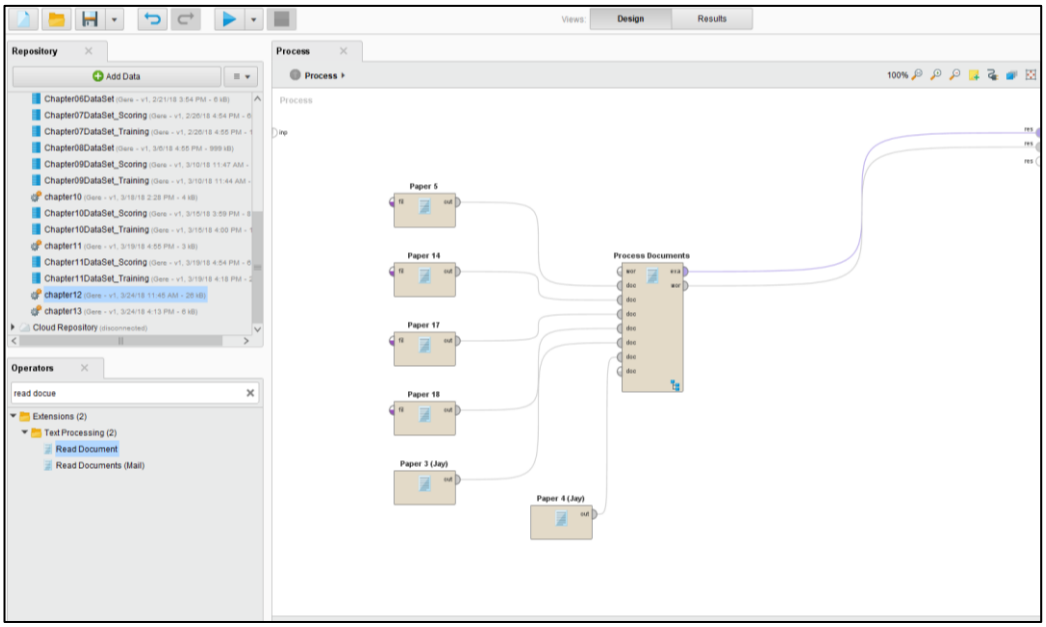


Figura 30. Model de preparació de mineria de textos.

Com ja vam comentar en l'apartat d'operadors, aquest últim conté un apartat de subprocessos on podem afegir altres operadors. En el nostre cas afegirem, justament, 3 operadors que vàrem explicar en l'apartat de *Process Documents: tokenize, filtre stopwords i Transform Cases*.

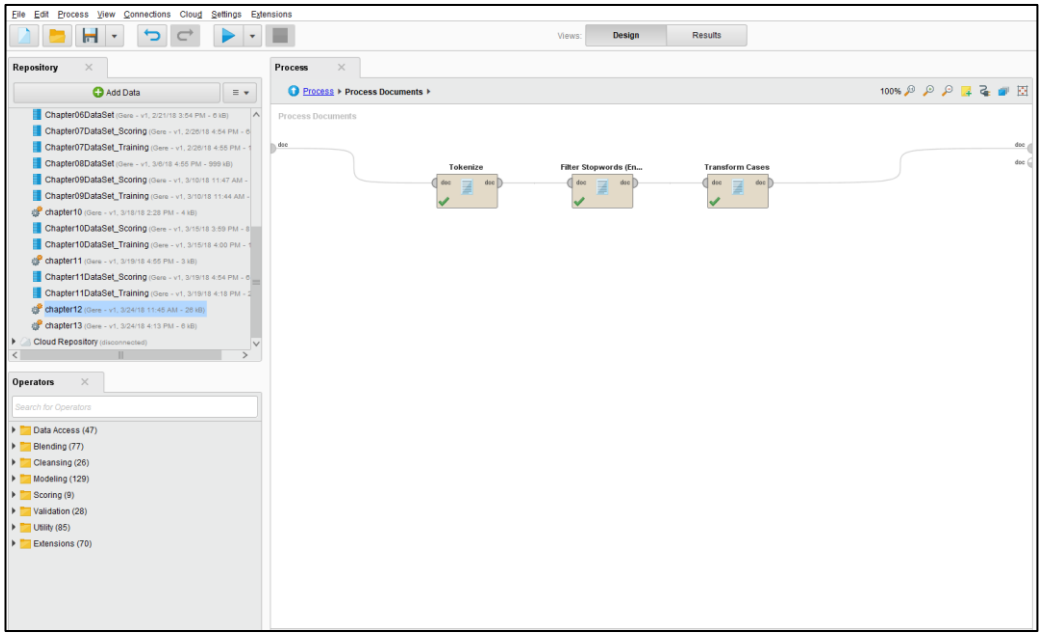


Figura 31. Subprocessos en el *Process Documents*.

Aquests 3 operadors ens permeten comptabilitzar les paraules rellevants, és a dir, traient conjuncions i articles, i saber en quants dels 6 textos, apareix cada paraula.

Word	Attribute Na...	Total Occurences	Document Occurences
abbe	abbe	2	1
abetted	abetted	1	1
abilities	abilities	1	1
abject	abject	1	1
able	able	5	3
ablest	ablest	1	1
abolish	abolish	1	1
abolished	abolished	1	1
abolition	abolition	1	1
abounds	abounds	1	1
abridgment	abridgment	1	1
abroad	abroad	3	3
absolute	absolute	3	2
absorb	absorb	1	1

Figura 32. Taula exemple de les paraules dels textos analitzats.

Es pot intuir, que amb el model que tenim fins ara, podem saber de què tracten els textos sense llegir-los, simplement, analitzant el número de cops que apareix cada paraula. Això però, no és el que estem buscant, ja que volem agrupar els diferents textos per autors segons les característiques de cadascun i, per tant, necessitem un operador capaç de fer-ho.

Utilitzarem el *K-Means,* operador que també està explicat en el corresponent capítol, i permetrà classificar els documents segons evidencies que els relacionin. Posarem un valor de K=2, ja que volem fer una divisió entre els textos escrits pel Jay i els escrits per Hamilton i Madison. Si es confirma que el document 18 va ser escrit per Aquests dos últim en el mateix clúster hauran d'aparèixer els documents 14,17 i 18 mentre que el 3,4,5 haurien d'aparèixer junts perquè són del mateix autor.

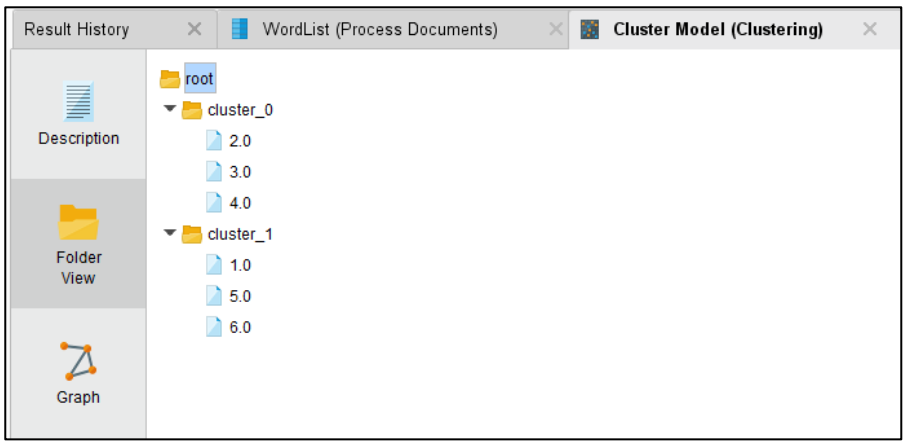


Figura 33. Resultat de *Clustering*

Els dos clústers que ha trobat el model son:

Clúster 0: Paper 14	Clúster 1: Paper 3
Paper 17	Paper 4
Paper 18	Paper 5

El fet que no coincideixin els números amb què els identifica l'operador (com veiem a la figura 3.3.11.), amb el número de document que són realment, és perquè el *K-means* els numera en l'ordre que han estat afegits en el model, així doncs, si sabem l'ordre en què els hem posat, sabem quin és cadascun. També, com mostrem a continuació, podem saber quin és cadascun si fem doble clic sobre els identificadors i veurem quin document és realment. En la figura següent podem veure que el text amb Id=1 és el Paper 5 de Jay (ho veiem en el nom de l'arxiu).

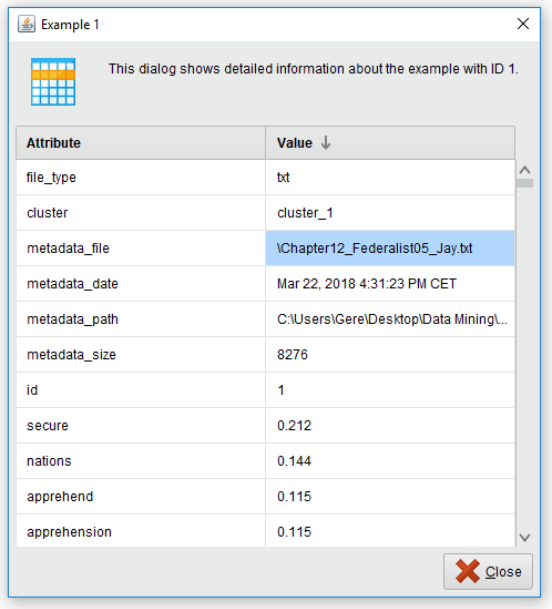


Figura 34. Identificador de *Clustering.*

Observant els resultats que hem obtingut podem afirmar que, com ja sospitàvem, Jay no té res a veure amb el document 18, i sí existeix una relació entre el document 18 amb el 14 i el 17. Sabent doncs, que el document 14 va estar escrit per Madison i el 17 per Hamilton, tenim bastantes raons per creure que si ens els ha agrupat junts, sigui perquè el 19 va estar escrit per ambdues persones.

4. Exemple Empresarial

Se'ns han facilitat les dades econòmiques d'aproximadament unes 10000 empreses espanyoles de diferents sectors. Entre les dades que hem rebut, trobem el capital de cada empresa, rendibilitat, endeutament, sector al qual pertanyen... i junt amb elles se'ns ha proposat analitzar-les i idear una situació amb la qual, tractant aquestes dades, obtinguem alguns resultats que ens puguin ser útils.

Abans d'entrar en l'exemple directament passarem a definir les variables que se'ns han proporcionat, ja que, majoritàriament, són paraules tècniques de comptabilitat i no està de més fer-hi un repàs:

- Capital social: és el valor del béns que posseeix l'empresa, així com les aportacions que realitzen els socis, siguin diners o altres formes. Aquest capital forma part dels fons propis de l'empresa i, per tant, està situat en el passiu del balanç.

- Actiu: conjunt de béns i drets que té l'empresa en un moment determinat. Representa la utilització, o bé, on estan invertits els recursos dels quals es disposen, per exemple, en maquinària, estocs, instal·lacions...

- Passiu: representen els deutes i obligacions que l'empresa ha contret en un moment determinat, sigui amb els propietaris d'aquesta o amb tercers.

- Rendibilitat econòmica (ROI): mesura la capacitat que tenen els actius de l'empresa de generar benefici sense importar com aquests hagin estat finançats. Aquests beneficis que tindrem en compte en aquest cas són previs a l'aplicació d'impostos i interessos. La fórmula per a calcular el ROI és la següent:

$$\text{Rendibilitat econòmica} = \frac{BAII}{Total\ Actiu}$$

Equació 1. Fórmula del ROI.

- Rendibilitat financera (ROE): com diu el propi terme, la rendibilitat és la capacitat de generar benefici i financera fa referencia a les finances, és a dir, als diners. Per tant, la rendibilitat financera són els beneficis obtinguts a causa d'invertir diners. Com veiem a la fórmula que s'adjunta a continuació, mesura el benefici net en relació a les inversions que s'han fet en la mateixa empresa:

$$\text{Rendibilitat financera} = \frac{BAI}{Recursos\ Propis}$$

Equació 2. Fórmula del ROE.

- Liquiditat: la liquiditat ens dóna informació sobre la capacitat que té l'empresa enfront a realitzar possibles pagaments en un curt termini.

Ratios de liquidez	Valores	Interpretación
$\boldsymbol{Liquidez} = \frac{Activo\ Corriente}{Pasivo\ Corriente}$	2 aprox.	Valor óptimo
	<1,5	Probabilidad suspensión pagos
	>>2	Posiblemente existe dinero ocioso. Esto disminuye la rentabilidad

Figura 35. Rati de liquiditat.

Com veiem a la taula, l'actiu ha de ser sempre major que el passiu i, si pot ser, pràcticament el doble. Quan vulguem millorar la situació actual, haurem de prendre mesures com:

  - Ampliar el capital.
  - Augmentar l'autofinançament.
  - Reconvertir el deute de curt a llarg termini.

- Resultat o Benefici: ens mostra l'evolució de l'empresa durant un termini de temps. S'obté de restar-li al benefici obtingut de les vendes, el cost que ha tingut produir-les i les altres despeses que s'han tingut durant aquest període.

  - Vendes: import dels productes entregats als clients, valorats cadascun amb el seu preu de venda.
  - Cost de les ventes: Preu que ha tingut produir els productes venuts.
  - Despeses: les que s'han tingut durant aquest termini de temps.

RESULTAT = VENDES – COS DE LES VENDES – DESPESES

Equació 3. Fórmula del resultat.

Si RESULTAT > 0 ⟶ BENEFICI

Si RESULTAT < 0 ⟶ PÈRDUES

- Endeutament: dóna informació sobre l'estructura de l'empresa.

Ratios de endeudamiento	Valores	Interpretación
$\boldsymbol{Endeudamiento} = \frac{Total\ Deudas}{Total\ Pasivo}$	0,4 a 0,6	Valor óptimo
	>> 0,6	Deuda alta, riesgo y dependencia de terceros
	< 0,4	Empresa demasiado capitalizada

Figura 36. Ratis d'endeutament.

Cal recordar que per a total de deutes es refereix a la suma de passiu corrent + passiu no corrent i que el total del passiu el formen els deutes + patrimoni net.

Ara que ja tenim les definicions fetes podem començar a treballar en l'exemple: som uns empresaris, que donat l'èxit que hem tingut amb les empreses que posseïm, volem invertir en nous sectors, però no tenim clar on fer-ho. Per això, decidim fer un estudi econòmic amb les dades que se'ns han facilitat observant diferents punts, per tal d'intentar decidir o encaminar cap a on volem expandir-nos. Així doncs, anirem creant models amb *RapidMiner* que puguin aportar-nos informació útil.

El primer que se'ns ha acudit pensar és, que depenent del capital que invertim, quin dels sectors ens assegura uns resultats de l'exercici més positius. Per fer-ho, necessitem crear un model que ens permeti filtrar els sectors que vulguem i aplicar una matriu de correlació a les variables que ens interessin (tots els operadors que utilitzarem estan explicats en l'apartat d'operadors d'aquest treball). Començarem important el nostre Excel amb les dades que tenim des de la pestanya d'*Add data* de *RapidMiner*:

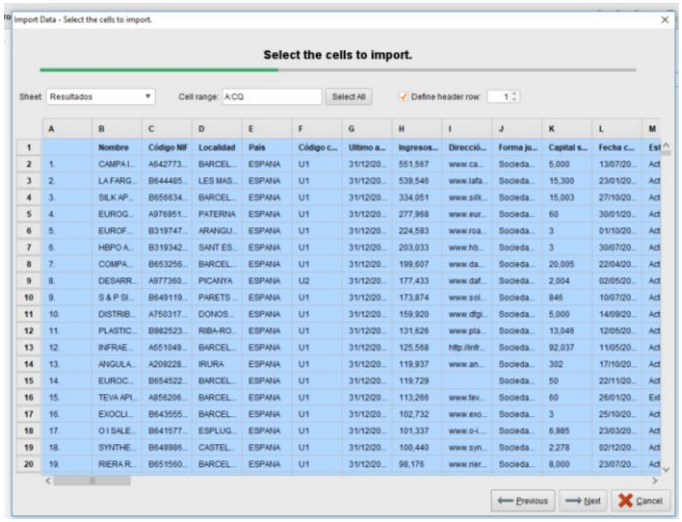


Figura 37. Importació de dades.

Ara que ja tenim les dades necessitem, poder filtrar per sectors i això és possible gràcies a la codificació CNAE que porten les nostres dades i, amb l'operador *Filter example* podem escollir quin sector és amb el que volem treballar:

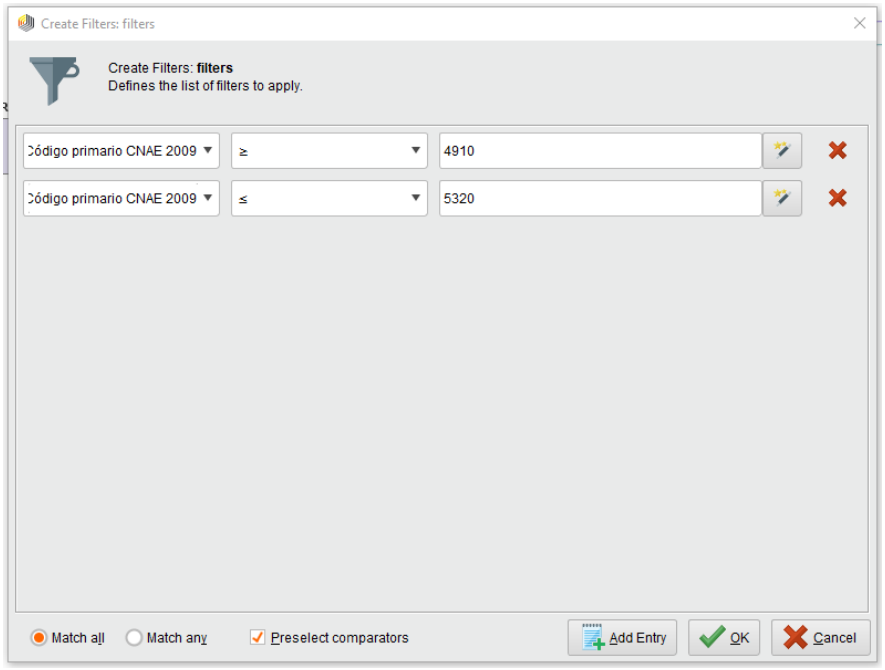


Figura 38. Filtració de dades per sector.

Un cop ja tenim el sector amb el qual volem treballar, necessitem seleccions, quines de totes les dades, utilitzarem en aquest cas. L'operador que ens ajudarà en això serà el *Select attributes*.

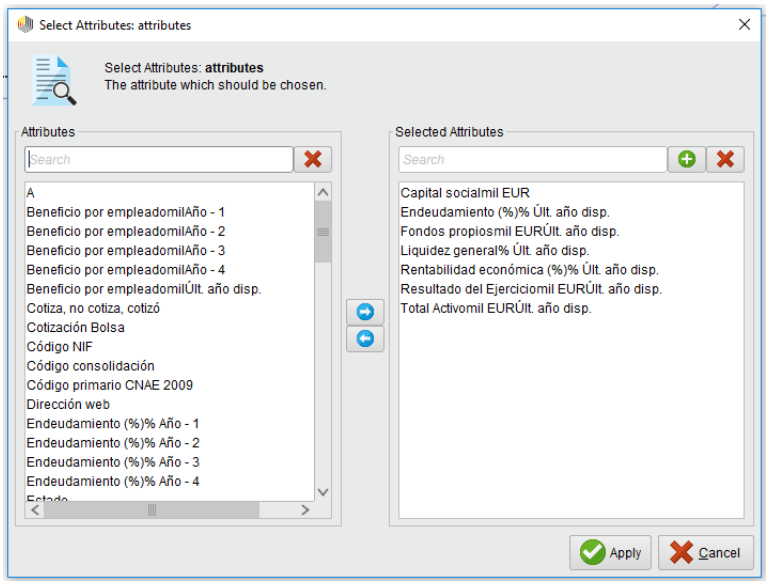


Figura 39. Selecció dels atributs necessaris.

Ara ja només queda afegir l'operador de matriu de correlació i ens queda el següent model:

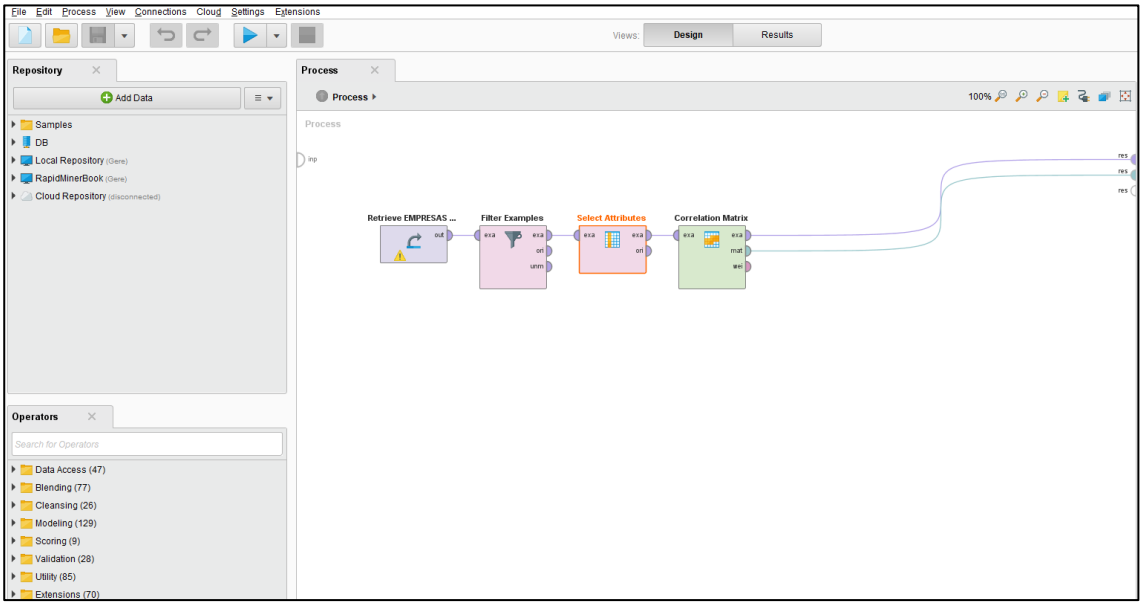


Figura 40. Model final de la correlació de variables.

Amb el model ja finalitzat, obtindrem resultats dels diferents sectors que ens interessen amb la seva respectiva matriu de correlació que analitzarem al final.

Attribut...	Capital ...	Resulta...	Total Ac...	Fondos ...	Rentabil...	Liquide...	Endeud...
Capital s...	1	0.532	0.548	0.558	0.009	0.026	-0.040
Resultad...	0.532	1	0.741	0.814	0.051	0.008	-0.067
Total Acti...	0.548	0.741	1	0.918	0.014	-0.008	-0.037
Fondos ...	0.558	0.814	0.918	1	0.018	0.021	-0.061
Rentabili...	0.009	0.051	0.014	0.018	1	0.039	-0.355
Liquidez ...	0.026	0.008	-0.008	0.021	0.039	1	-0.159
Endeuda...	-0.040	-0.067	-0.037	-0.061	-0.355	-0.159	1

Figura 41. Matriu de correlació de la indústria manufacturera.

Attribut...	Capital ...	Resulta...	Total Ac...	Fondos ...	Rentabil...	Liquide...	Endeud...
Capital s...	1	0.385	0.617	0.794	0.004	0.105	-0.012
Resultad...	0.385	1	0.775	0.716	0.021	-0.004	-0.021
Total Acti...	0.617	0.775	1	0.899	0.012	0.053	-0.015
Fondos ...	0.794	0.716	0.899	1	0.010	0.094	-0.019
Rentabili...	0.004	0.021	0.012	0.010	1	0.004	-0.972
Liquidez ...	0.105	-0.004	0.053	0.094	0.004	1	-0.013
Endeuda...	-0.012	-0.021	-0.015	-0.019	-0.972	-0.013	1

Figura 42. Matriu de correlació del sector de la construcció.

Attribut...	Capital ...	Resulta...	Total Ac...	Fondos ...	Rentabil...	Liquide...	Endeud...
Capital s...	1	0.288	0.435	0.591	0.002	0.003	-0.007
Resultad...	0.288	1	0.609	0.731	0.017	-0.006	-0.020
Total Acti...	0.435	0.609	1	0.647	0.005	-0.005	-0.004
Fondos ...	0.591	0.731	0.647	1	0.009	0.004	-0.018
Rentabili...	0.002	0.017	0.005	0.009	1	0.001	-0.998
Liquidez ...	0.003	-0.006	-0.005	0.004	0.001	1	-0.005
Endeuda...	-0.007	-0.020	-0.004	-0.018	-0.998	-0.005	1

Figura 43. Matriu de correlació de reparació de vehicles.

Attribut...	Capital ...	Resulta...	Total Ac...	Fondos ...	Rentabil...	Liquide...	Endeud...
Capital s...	1	0.782	0.970	0.806	0.015	0.130	-0.056
Resultad...	0.782	1	0.667	0.309	0.046	-0.020	-0.009
Total Acti...	0.970	0.667	1	0.910	0.016	0.163	-0.055
Fondos ...	0.806	0.309	0.910	1	0.005	0.236	-0.093
Rentabili...	0.015	0.046	0.016	0.005	1	0.102	-0.718
Liquidez ...	0.130	-0.020	0.163	0.236	0.102	1	-0.400
Endeuda...	-0.056	-0.009	-0.055	-0.093	-0.718	-0.400	1

Figura 44. Matriu de correlació de transport i emmagatzematge.

Attribut...	Capital ...	Resulta...	Total Ac...	Fondos ...	Rentabil...	Liquide...	Endeud...
Capital s...	1	0.209	0.508	0.701	-0.080	0.006	-0.275
Resultad...	0.209	1	0.666	0.679	0.106	0.129	-0.057
Total Acti...	0.508	0.666	1	0.777	-0.033	0.030	-0.042
Fondos ...	0.701	0.679	0.777	1	-0.032	0.075	-0.259
Rentabili...	-0.080	0.106	-0.033	-0.032	1	0.018	-0.199
Liquidez ...	0.006	0.129	0.030	0.075	0.018	1	-0.109
Endeuda...	-0.275	-0.057	-0.042	-0.259	-0.199	-0.109	1

Figura 45. Matriu de correlació d'activitats immobiliàries.

Com podem veure amb els resultats obtinguts, depèn del sector, hi ha una correlació major o menor entre el capital inicial i els resultats. Per exemple, en el cas del transport i emmagatzematge, veiem que és una relació forta, amb un coeficient de 0,782 (com més proper a 1 més forta és), el que significa que una bona inversió inicial ens pot garantir uns bons resultats. Pel que fa a l'indústria manufacturera, també existeix una relació però menys forta que la del cas anterior. En la resta de casos, però, podem dir que no existeix cap relació, a priori, entre aquestes dues variables, per tant, el capital inicial no és un factor que ens determini l'èxit de la nostra inversió.

També podem veure que en sectors com construcció, transport i reparació de vehicles, tenir una bona rendibilitat implica no tenir endeutament, ja que un coeficient negatiu ens indica que quan un creix l'altre decreix, cosa que no passa en els altres dos sectors. El que si podem veure en tots els casos, tot i que amb graus diferents, el fet de tenir un bon capital inicial ens garanteix una bona quantitat de fons.

Com a segon estudi se'ns ha acudit intentar predir quines variables, i amb quins valors podrien afectar que la nostra empresa hagués de patir un tancament. Per intentar resoldre aquesta qüestió hem decidit que un arbre de decisió ens ajudarà a veure de manera clara quin camí ens portaria, en cada cas, a dissoldre l'empresa. A l'igual que en el cas anterior, utilitzarem tot d'operadors que ja hem explicat.

El primer que farem serà preparar dos excels, un amb totes les variables (*Training Data*) i l'altre amb la variable d'estat (activa, extingida, dissolta...) sense completar (*Scoring Data*), ja que és la que haurem de predir. Un cop tinguem els dos excels importats, afegirem l'operador *Set role* per donar-li el paper de label a la variable estat i d'ID a la de nom de l'empresa.

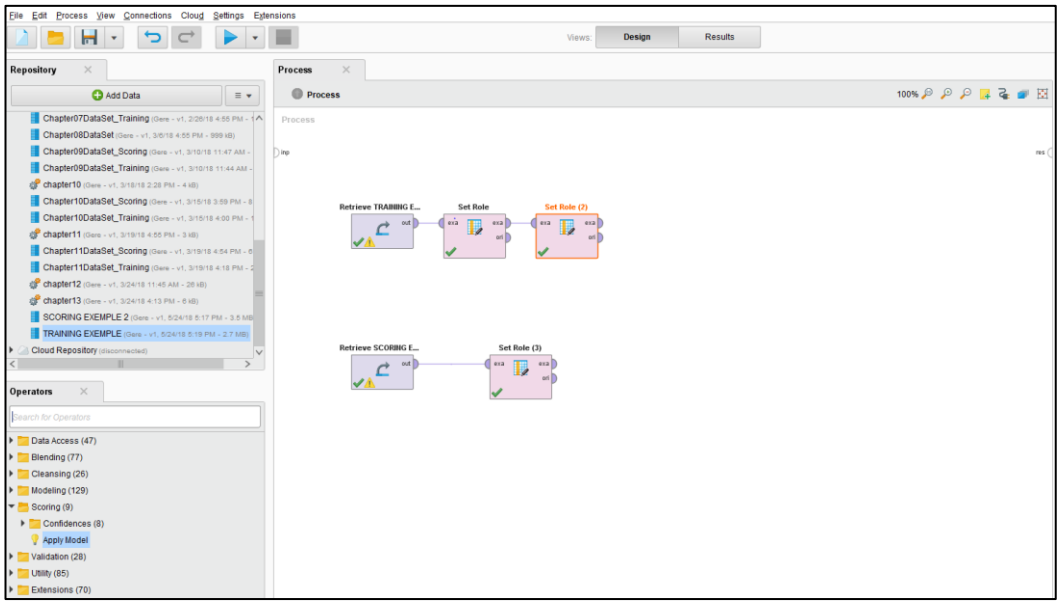


Figura 46. Preparació de l'arbre de decisions.

Ara que ja tenim els rols assignats, seleccionarem amb el *Select attributes* totes les variables amb les quals vulguem treballar. Finalment només ens queda afegir l'operador *Decision Tree* i *l'Apply model* per ajuntar les dues bases, amb la qual cosa ens queda el següent model:

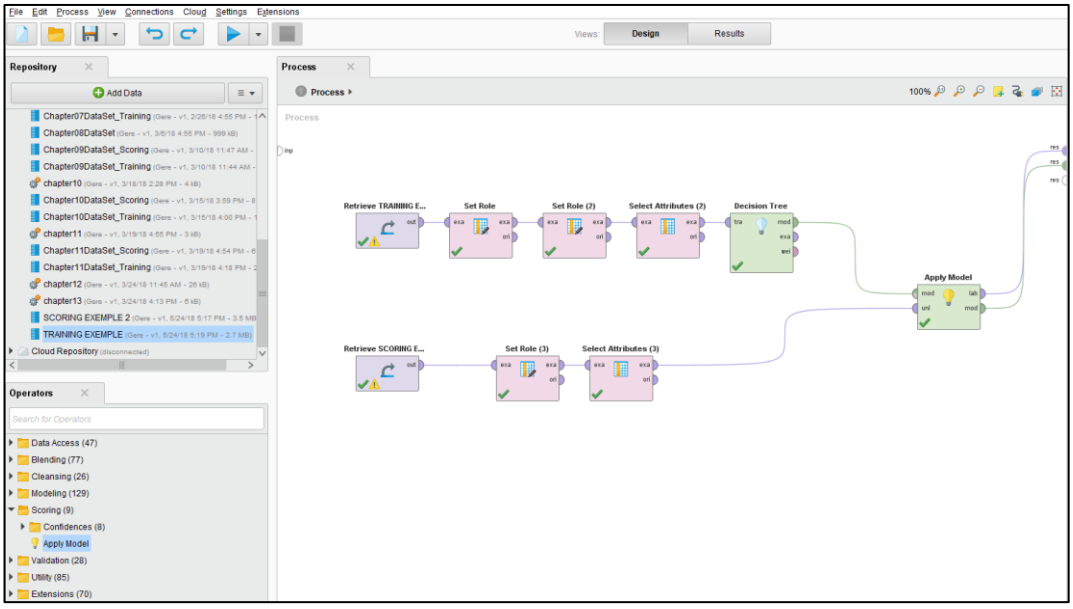


Figura 47. Model final del *Decision Tree.*

Ara ja només queda fer-li *Run* al model i analitzar els resultats.

Figura 48. Arbre de decisions.

El primer que podem tenir en compte d'aquest arbre de decisions que hem creat és que, de totes les variables que hem inclòs per tenir en compte, només n'apareixen 6 i, d'aquestes 6, el capital social i els ingressos d'explotació tenen un pes menor que la resta. Això vol dir, que variables com la localització de l'empresa, el sector o la forma jurídica no tenen molt influencia en l'estat de l'empresa en l'exemple que estem fent. Si comencem a analitzar-lo des dels primers punts, veurem que si acabem el primer any amb uns fons propis negatius de -193163 euros o més negatius és molt probable que la nostra empresa hagi de tancar, de la mateixa manera que ho farà si la rendibilitat és inferior al -478% o tenim un resultat de l'any de -5640132 euros. Fins aquí tot sembla molt evident, ja que tots els valors que hem vist són en situacions molt negatives i lògicament ens portarien a tancat l'empresa. Si seguíem baixant però, veurem que hi ha més casos on podria arribar a perillar la nostra

empresa. Per exemple, es dóna el cas que una empresa amb un fons propis negatius de -32000 euros, amb un endeutament del 110% i uns ingressos d'explotació inferiors a 979136 euros es va haver de tancar. Òbviament, el motiu de la dissolució no té per què ser únicament econòmic, però com nosaltres estem realitzant aquest estudi, només donarem per fet aquesta qüestió.

Així doncs, aquest model ens marca la pauta dels valors als quals no hem d'arribar si no volem haver de plegar, i per quin camí hem d'anar pel que fa a resultats anuals si volem mantenir activa la nostra empresa. Cal tornar a remarcar que en la vida real influirien altres factors que aquí no podem tenir en compte, però tot i això, ens ha servit per veure els límits econòmics en els quals ens hauríem de situar.

ExampleSet (4831 examples, 12 special attributes, 14 regular attributes)

Row No.	Nombre	prediction(E...	confidence(...	confidence(...	confidence(...	confidence(...	confidence(...	confidence(...
1	ALIMENTACI...	Activa	0.954	0.014	0.010	0.006	0.001	0.005
2	DAISA TECH...	Activa	0.954	0.014	0.010	0.006	0.001	0.005
3	SHANG HAI L...	Activa	0.954	0.014	0.010	0.006	0.001	0.005
4	ORMATIN SL	Activa	0.954	0.014	0.010	0.006	0.001	0.005
5	MAINCA MAQ...	Activa	0.954	0.014	0.010	0.006	0.001	0.005
6	MAGUS LOGI...	Activa	0.954	0.014	0.010	0.006	0.001	0.005
7	SODETIC CO...	Activa	0.954	0.014	0.010	0.006	0.001	0.005
8	KENKO MIRA...	Activa	0.954	0.014	0.010	0.006	0.001	0.005
9	INOXILAN SL	Activa	0.954	0.014	0.010	0.006	0.001	0.005
10	NEWCOSPR...	Activa	0.954	0.014	0.010	0.006	0.001	0.005

Figura 49. Taula de percentatges de confiança de l'arbre de decisions.

Com a últim estudi, hem decidit utilitzar *RapidMiner* per a treure tota la informació de les pàgines webs de les empreses i comprar-les entre elles. Amb això, podem intentar relacionar el fet de promocionar-se o vendre els teus serveis d'una determinada manera amb els resultats econòmics que té aquesta empresa a final d'any.

En aquest cas, necessitarem baixar-nos una extensió del software que es diu *Web Mining*, ja que no ve descarregada de manera predeterminada. Per fer-ho, clicarem a la pestanya d'extensions situada a dalt de la interfície, obrirem el mercat d'extensions de *RapidMiner* i buscarem la que volem amb el cercador i, un cop la tinguem seleccionada l'haurem de descarregar.

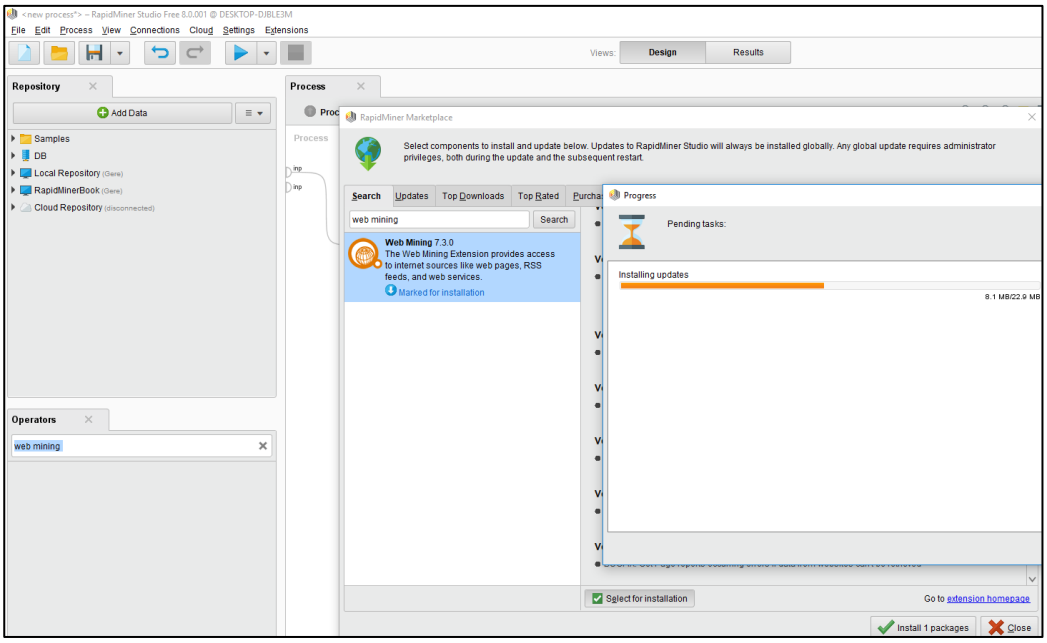


Figura 50. Descarregar extensions a *RapidMiner*.

Ara que ja tenim els operadors necessaris podem començar a construir el model. Com ja s'ha explicat anteriorment, importarem la nostra base de dades d'empreses. L'única diferencia respecte als cops anteriors, és que la columna de pàgines webs hem de marcar-la com a *file_path* perquè el programa sàpiga que això són direccions webs. És molt important que les direccions webs tinguin el format http://, ja que, sense això no podrà connectar-se a elles.

Ara que ja tenim les dades importades, farem un filtre per eliminar totes aquelles empreses que no tenen, o no tenim, una pàgina web amb el *Filter example*. Seguidament, podem eliminar, si volem, tots els atributs que no siguin direccions webs per tal d'alleugerir el processament de càlculs utilitzant el *Select Attributes*. Finalitzada la preparació de les dades, podem afegir l'operador *Get Pages*, de la nova extensió, que llegirà les webs i ens donarà informació d'elles, però no el seu contingut. Per tal de passar la data descarregada a un document llegible utilitzarem l'operador *Data to Documents* que, bàsicament, fa el que el seu nom ens indica.

Si analitzem el resultat tal com tenim el model, veurem que sí que tenim un document escrit, però són els codis amb els quals estan configurades les pàgines i, per tant, no podem treure res d'aquí. Ara necessitem, per exemple, comptabilitzar les paraules i veure quines coincideixen en diversos llocs per establir una relació. Sigui com sigui el que volem fer, necessitarem processar aquests documents i per fer-ho, importarem l'operador *Process Documents*, amb el que ja hem treballat anteriorment. Recordem que aquest operador conté subprocessos que són els que tractaran el nostre document. En aquest cas afegirem el *Tokenize* i el *Transform Cases*, també explicats anteriorment, dins d'aquests processos interns del *Process Documents* per tal de separar les paraules i posar-les totes en

minúscula. Finalment afegirem un operador anomenat *Data to Similarity* que compara el primer document amb la resta, seguidament el segon amb els posteriors i així fins que ha comparat tots entre ells i els atribueix un número del 0-1 depenent de com de similars són entre ells.

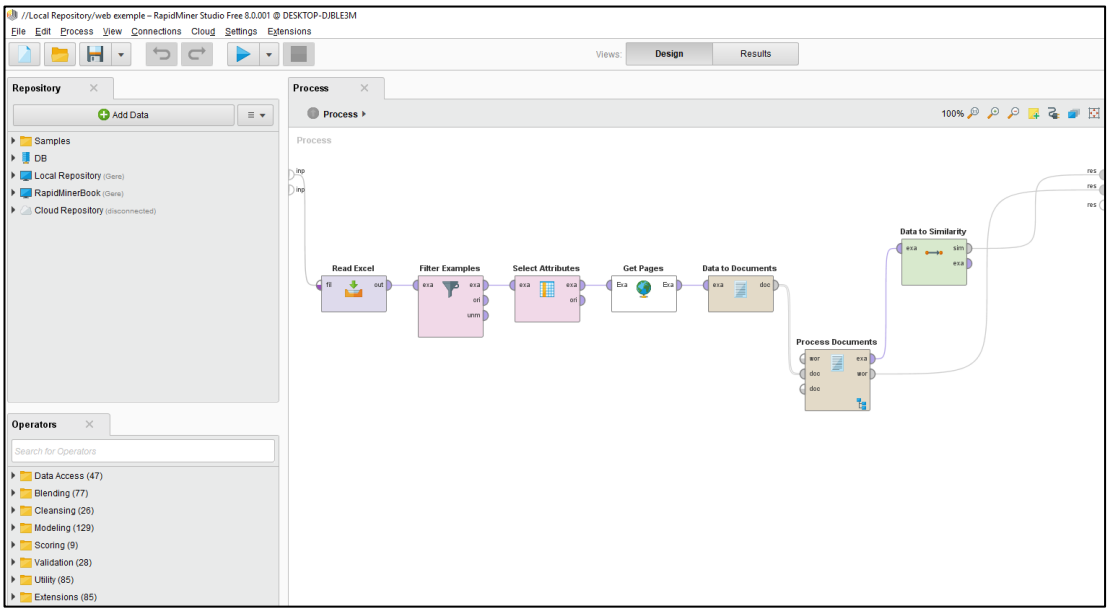
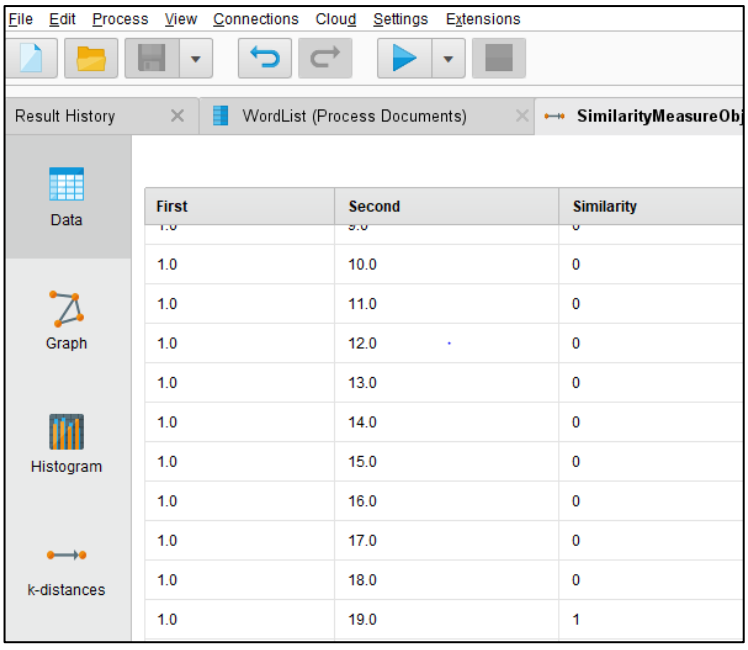


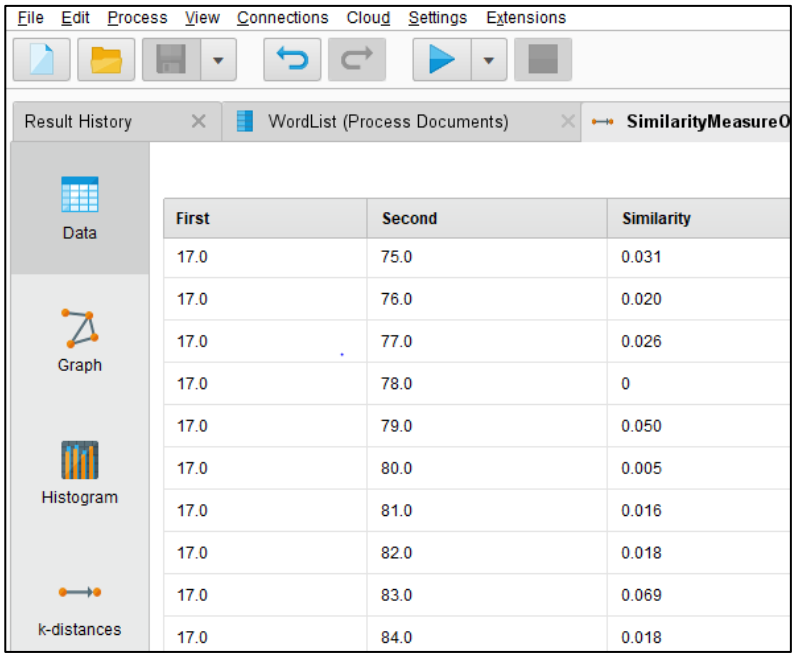
Figura 51. Model de pàgines web finalitzat

Ara que ja tenim el model finalitzat podem donar a *Run* i analitzar els resultats per intentar treure algunes conclusions.



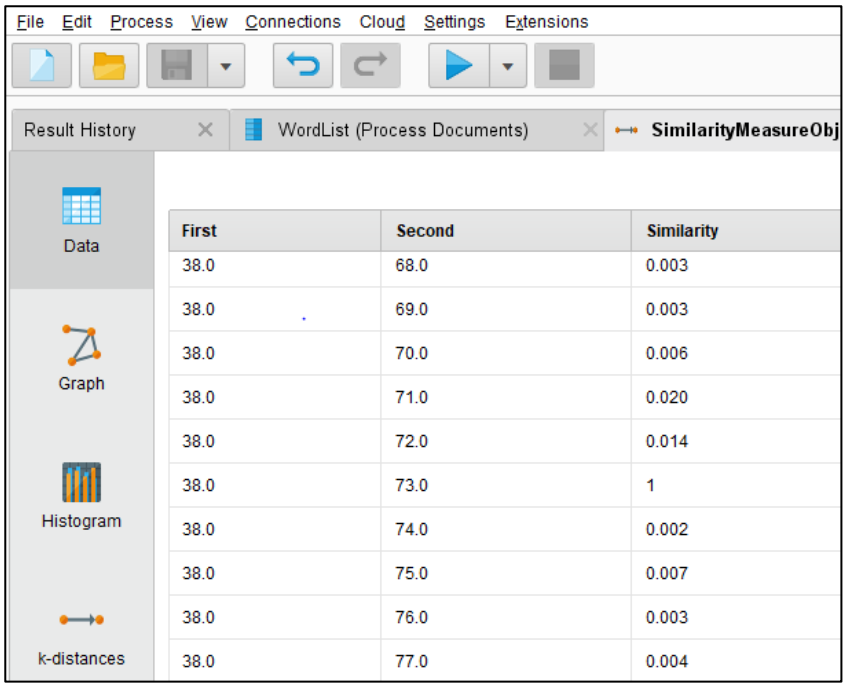
First	Second	Similarity
1.0	10.0	0
1.0	11.0	0
1.0	12.0	0
1.0	13.0	0
1.0	14.0	0
1.0	15.0	0
1.0	16.0	0
1.0	17.0	0
1.0	18.0	0
1.0	19.0	1

Figura 52. Taula de similitud 1.



First	Second	Similarity
17.0	75.0	0.031
17.0	76.0	0.020
17.0	77.0	0.026
17.0	78.0	0
17.0	79.0	0.050
17.0	80.0	0.005
17.0	81.0	0.016
17.0	82.0	0.018
17.0	83.0	0.069
17.0	84.0	0.018

Figura 53. Taula de similitud 2.



First	Second	Similarity
38.0	68.0	0.003
38.0	69.0	0.003
38.0	70.0	0.006
38.0	71.0	0.020
38.0	72.0	0.014
38.0	73.0	1
38.0	74.0	0.002
38.0	75.0	0.007
38.0	76.0	0.003
38.0	77.0	0.004

Figura 54. Taula de similitud 3.

Com podem veure en les diferents figures dels resultats, se'ns dona una taula amb 3 columnes: la primera conté el primer document, la segona és el document amb el qual se'l compara, i la tercera el grau de similitud. Observem que hi ha casos on la similitud és 1, la qual cosa vol dir que aquestes pàgines webs contenen informació molt relacionada, altres casos on és 0, volen dir que no tenen res a veure, i entremig, un munt de valors que com més pròxim són a 1 més paraules en comú tenen aquestes pàgines.

Hem decidit comparar algunes de les pàgines web que tenen una similitud total, per tal d'esbrinar si això implica algun tipus de relació amb alguna de les variables econòmiques. Tot i que, certament, algunes d'aquestes empreses que presenten unes pàgines web molt semblants, pertanyen al mateix sector i presenten unes variables econòmiques prou semblants, també hem vist d'altres que no només no són econòmicament semblants, sinó que no pertanyen al mateix sector i la similitud, per exemple, es deu al fet que presenten els seus diferents serveis molt enfocats a ser ecològics. Per tant, aquest model ens pot ajudar a trobar relació entre empreses i en com dirigeixen la creació de la seva pàgina web però no ens garanteix un impacte realment fort parlant econòmicament.

5. Anàlisi de l'impacte ambiental

Donat que es tracta d'una eina digital, l'impacte ambiental que provoca, prové majoritàriament de l'energia consumida pels ordinadors i pantalles, així com de la fusta talada per imprimir el llibre d'introducció a *RapidMiner*.

En el cas d'una única persona treballant en un ordinador durant una jornada laboral de 8 hores i utilitzen un llibre d'aproximadament 250 pàgines generarà el següent impacte:

- Utilitzant un ordinador de potència mitjana, per exemple, un HP 286 Pro G2 que és el que podríem trobar en bastantes empreses, el qual té un consum de 26,29 W quan està encès, 1,78 W en mode d'estalvi i 0,67W estan apagat i una pantalla que té un consum d'aproximadament 15 W encesa i 0,12 W apagada, ens donaria un consum total en acabar la jornada de (valors extrets de la taula adjunta):

  Consum Total= (26,29W · 8 hores + 0,67W · 16 hores) + (15W · 8 hores +

  0,12W · 16 hores) = 342,96 Wh =**0,343 KWh**

  Segon la Comissió Europea, partint de la base que utilitzem petroli per obtenir aquesta energia, aquest consum diari equivaldria a un total de **0,223 Kg de CO_2** diari.

  En el nostre cas, hem fet l'estimació que per a desenvolupar aquest treball hem destinat uns 5 mesos, la qual cosa implica unes emissions totals de 33,45 Kg de CO_2.

  Emissions totals= 0,223 Kg · 150 dies = **33,45 Kg de CO_2**

- Pel que fa a la desforestació, per l'obtenció dels llibres d'introducció partint dels estudis que diuen que d'un sol arbre es poden produir un promig de 12000 fulles de paper, necessitaríem la tala de **0,02 arbres** (250 fulles/12000), és a dir, un impacte realment petit que, a més a més, podem eliminar utilitzant la versió digital del llibre.

Model number *	*286 Pro G2*
Issue date *	*2/1/2016*

Product environmental attributes - Market requirements (continued)

Item

P9 **Energy consumption**

9.1 For the product the following power levels or energy consumptions are reported:

Energy mode *	Power level at **100** V AC	Power level at **115** V AC	Power level at **230** V AC
ENERGY STAR® On Mode * (System Short Idle)	*26.07* W	*26.30* W	*26.29* W
ENERGY STAR® On Mode * (System Long Idle)	*24.88* W	*24.53* W	*25.06* W
ENERGY STAR® Low Power Sleep Mode* (S3 - Windows *"Standby") With Wake On LAN (WOL) Enabled*	*1.71* W	*1.70* W	*1.78* W
ENERGY STAR® Low Power Sleep Mode* (S3 - Windows "Standby") With Wake On LAN (WOL) Disabled	*1.70* W	*1.69* W	*1.78* W
System Off/Apparent Off Mode (ACPI S5) With Wake On LAN(WOL) Enabled* (Test Unit connected to AC Mains, AC adapter connected to All-In-One PC, if applicable)	*0.62* W	*0.62* W	*0.67* W

Figura 55. Valors de consum d'un ordinador HP 286 Pro G2.

Conclusions

L'objectiu principal d'aquest treball ha estat la realització d'un estudi econòmic de diferents empreses mitjançant l'anàlisi de dades. Per a poder realitzar-lo hem optat per utilitzar el software *RapidMiner*, un dels més potents pel que fa a mineria de dades.

El primer que hem fet ha sigut investigar i profunditzar en els coneixements sobre el *Data Mining*, ja que en la nostra vida acadèmica no havíem tractat aquesta matèria. Gràcies a les diferents fonts d'informació consultades, hem observat que qualsevol problema que es vulgui solucionar amb aquesta metodologia segueix uns passos molt concrets que es converteixen en un cicle de 6 etapes i, per tant, és el camí que nosaltres hem seguit per fer el nostre estudi.

Seguidament hem hagut de familiaritzar-nos amb *RapidMiner*, tant amb la seva interfície com amb la seva manera de treballar. La interfície, en aquest cas, és molt intuïtiva i ens ha sigut fàcil adaptar-nos a ella. Pel que fa a la metodologia de treball, podem dir que pot arribar a ser tan complexa com un es proposi, ja que *RapidMiner* té una àmplia varietat d'operadors que permeten processar i analitzar les bases de dades de moltíssimes maneres. En el nostre cas, ens hem centrat a treballar amb les eines que més útils ens podien resultar pel nostre objectiu.

Un cop hem tingut clar com fer un anàlisi de dades i com utilitzar el programa que hem escollit com a eina de suport, hem començat a treballar en l'exemple. Tal com hem après anteriorment, el primer que hem fet ha sigut analitzar el problema que volíem resoldre. Com hem dit, hem volgut realitzar un estudi econòmic per a identificar quins sectors econòmics, quines variables econòmiques i financeres i quins valors d'aquestes variables podrien predir una major probabilitat d'èxit. Aquesta informació pot ser útil com a eina de suport a la decisió de noves inversions que disminueixin el risc. Quan ja hem tingut clar quin era l'objectiu, hem hagut d'entendre quines dades se'ns havien proporcionat i fins on ens permetien arribar. Amb tot això resolt, ja hem començat a preparar les dades per realitzar el model en *RapidMiner.*

Hem realitzat diferents models per tal d'esbrinar si les variables que nosaltres tenim mantenen alguna connexió entre elles que ens permeti predir en quin sector volem invertir i de quina manera. Un cop realitzats els estudis, hem observat que hi ha certs factors que hem de tenir en compte a l'hora de prendre la decisió. Primerament, el capital social no és un factor completament determinant que marqui els beneficis d'una empresa però, certament, en alguns sectors sí que hem trobat una correlació entre el capital invertit i el benefici a final d'any, per tant, depenent d'on decidim expandir-nos ho hem de tenir en compte. Posteriorment, hem pogut trobar quines de les variables tenen un pes més important a l'hora de marcar-nos quan una empresa arriba a la seva fallida. Òbviament, hi ha una gran quantitat de factors que nosaltres no hem pogut tenir en compte,

però amb els que sí tenim, hem pogut trobar els límits dins dels quals ens hem de moure si no volem tancar l'empresa. Finalment, hem analitzat les pàgines webs de les empreses i hem volgut veure quines tenien una certa semblança entre elles. Quan hem obtingut aquesta relació, hem analitzat si el fet de tenir pàgines webs semblants té alguna repercussió en les variables econòmiques i hem pogut veure que, en alguns casos, sí es compleix, però no amb suficient assiduïtat per a prendre'ns-ho com una norma.

Així doncs, podem dir que hem assolit l'objectiu de realitzar un estudi econòmic a petita escala i, a més a més, conèixer el funcionament de l'anàlisi de dades i com treballa un dels softwares més potents en aquest sector.

Pressupost i/o Anàlisi Econòmica

Com s'ha comentat en l'apartat d'impacte ambiental, aquest treball és purament digital i, per tant, tota la inversió estarà destinada a les eines necessàries per treballar, ja sigui l'ordinador i els accessoris d'aquest, l'energia per fer-lo funcionar o la llicència completa del software.

Podem desglossar el pressupost de la següent manera:

- Ordinador i perifèrics: Donat que hem calculat l'energia consumida en apartats anterior per un HP 286 Pro G2 donarem el pressupost també per aquest PC. Segons la pàgina oficial del fabricant aquest ordinador té un preu aproximat de 450 €. Si això sumem els 150 del monitor i els 50 dels accessoris sumen un total de **650 euros**. Tenint en compte que la vida útil d'un PC és d'aproximadament 3 anys i, que nosaltres, hem fet un ús d'aquest durant 5 mesos, ens surt un cost de **90,3 euros**.

- Energia: Com ja hem dit, al preu de l'ordinador cal sumar-li el cost que té fer-lo funcionar. En l'apartat de l'impacte ambiental, vam calcular quants kWh consumiríem en una jornada laboral de 8 hores, i ens va donar un valor de 0,343 KWh.
  L'any 2017 el KWh a Espanya va tenir un preu mig de 0,1216 €/kWh, per tant, seria un cost diari de 0,0417 euros. Si tenim en compte que en l'aprenentatge i preparació d'aquest treball hem dedicat 5 mesos, suposa un cost total de **6,25 euros**.

- Llicència: un dels punts positius que té *RapidMiner* és que disposa d'una versió gratuïta per estudiants, però té un límit de dades amb les que treballar i de temps d'utilització, és per això, que si volem treure el màxim rendiment del software, haurem d'adquirir una de les llicències que ens ofereix. Existeixen 3 tipus de llicència (petita, mitjana o gran) que, depenent de la quantitat de les dades amb les que et permet treballar, valen més o menys. En el nostre cas, agafarem la petita que permet treballar amb 100000 dades i té un preu de **2093 euros/l'any**.

- Llibre *Data Mining for the Masses*: ha sigut la base del nostre aprenentatge i és molt recomanable per a qualsevol persona que vulgui aprendre més sobre la mineria de dades. Adquirir-lo té un preu de **33,37 euros,** tot i que en el meu cas, se m'ha subministrat una versió online gratuïta pel fet de ser estudiant.

- Cost de personal: Tenint en compte que hem dedicat aproximadament 300 hores a un preu de 25 euros/hora ens dóna un valor de **7500 euros.**

Pressupost

Nº pressupost	1
Data del pressupost	17.05.2018
Total (EUR)	**11764,73**

Descripció	Quantitat	Preu (euros)	Import (euros)
Ordinador i perifèrics	1	90,3	90,3
Cost energètic	1	6,25	6,25
Llicència *RapidMiner*	1	2093	2093
Llibre: *Data Mining for the Masses*	1	33,37	33,37
Cost de personal	1	7500	7500

Taula 1. Pressupost

Subtotal sense IVA	**9722,92**
IVA 21% de 9722,92	**2041,81**
Total (EUR)	**11764,73**

Bibliografia

- CNAE. (2009). Listado completo de actividades de la CNAE 2009. 19/05/2018, de CNAE Sitio web: https://www.cnae.com.es/lista-actividades.php

- Comissió Europea. (2012). Como calcular CO2. 16/04/2018, de Arboliza Sitio web: http://arboliza.es/compensar-co2/calculo-co2.html

- Hofmann, Markus; Chisholm, Andrew (2016): Text mining and visualization. Case studies using open-source tools /edited by Markus Hofmann, Andrew Chisholm. 1st. Boca Raton: Chapman & Hall/CRC (Chapman & Hall/CRC data mining and knowledge discovery series).

- HP. (2016). THE ECO DECLARATION. 16/04/2018, de HP Sitio web: http://h22235.www2.hp.com/hpinfo/globalcitizenship/environment/productdata/Countries/_MultiCountry/iteco_deskto_20163242302845.pdf

- Kotu, Vijay; Deshpande, Balachandre (2015): Predictive analytics and data mining. Concepts and practice with RapidMiner. Amsterdam: Elsevier/Morgan Kaufmann Morgan Kaufmann is an imprint of Elsevier. Sitio web: http://londonmet.eblib.com/patron/FullRecord.aspx?p=1875324.

- Kumar, Ashish; Paul, Avinash (2016): Mastering text mining with R. Master text-taming techniques and build effective text-processing applications with R. Birmingham, UK: Packt Publishing. Sitio web: http://proquest.tech.safaribooksonline.de/9781783551811.

- Leskovec, Jure; Rajaraman, Anand; Ullman, Jeffrey David (2014): Mining of Massive Datasets. Cambridge: Cambridge University Press, Última comprobación el 02/06/2018.

- Markus Hofmann; Ralf Klinkenberg: RapidMiner: Data Mining Use Cases and Business Analytics Applications, Última comprobación el 02/06/2018.

- North, Matthew. (2012). Data Mining for the Masses. UTAH, USA: Global Text Project Book.

- North, Matthew (2016): Data mining for the masses. With implementations in RapidMiner and R. Second edition. [Verlagsort nicht ermittelbar], Wrocław, Poland: [Verlag nicht ermittelbar]; Amazon Fulfillment.

- Provost, Foster; Fawcett, Tom (2013): Data science for business. First edition. Sebastopol, Calif.: O'Reilly.

- RapidMiner (2016): Chapman and Hall/CRC, Última comprobación el 02/06/2018.

- RapidMiner. (2018). RapidMiner Pricing. 10/05/2018, de RapidMiner Sitio web: https://rapidminer.com/pricing/

- Rashid Al-Azmi, Abdul-Aziz (2013): Data, Text and Web Mining for Business Intelligence. A Survey. En: IJDKP 3 (2), pág. 1–21. DOI: 10.5121/ijdkp.2013.3201.

- Rokach, Lior; Maimon, Oded (2015): Data mining with decision trees. Theory and applications. Second edition. Hackensack New Jersey: World Scientific.

- Selectra. (2017). Precio del Kwh en España. 05/05/2018, de Selectra Sitio web: https://tarifasgasluz.com/faq/precio-kwh-espana-2017#que-es-kwh

- Sharda, Ramesh; Delen, Dursun; Turban, Efraim (2014): Business intelligence. A managerial perspective