Data Mining

En busca de patrons geogràfics per focalitzar recursos

Albert Salmerón Moreno
Grau en Enginyeria Informàtica

Humberto Andrés Sanz

13/01/2016

Aquesta obra està subjecta a una llicència de Reconeixement-NoComercial-SenseObraDerivada 3.0 Espanya de Creative Commons

FITXA DEL TREBALL FINAL

Títol del treball:	*DATA MINING – En busca de patrons geogràfics per focalitzar recursos*
Nom de l'autor:	*Albert Salmerón Moreno*
Nom del consultor:	*Humberto Andrés Sanz*
Data de lliurament (mm/aaaa):	*01/2016*
Àrea del Treball Final:	*Business Intelligence*
Titulació:	*Grau en Enginyeria Informàtica*

Resum del Treball (màxim 250 paraules):

Actualment, degut a la proliferació dels sistemes d'informació, dins de les organitzacions l'emmagatzematge i l'ús de tot tipus de dades està creixent de forma exponencial. Per tal de poder treure profit d'aquesta font de dades, moltes empreses comencen a implementar diverses tècniques amb l'objectiu de poder extreure'n coneixement.

Aquí és on el concepte de Intel·ligència de Negoci o *Business Intelligence* pren especial importància. Amb la finalitat de ser el màxim competitiu possible dins d'un entorn empresarial en constant evolució, al llarg d'aquest projecte s'implementaran diversos mecanismes de Mineria de Dades intentant trobar patrons de coneixement acceptables que permetin treure'n profit.

La situació inicial correspon a la d'una escola de negocis privada que pretén focalitzar els seus recursos de captació de nous estudiants en zones geogràfiques favorables. A l'hora d'accedir als estudis, cada candidat ha de superar diverses proves i entrevistes així que l'èxit d'aquests en una zona concreta serà l'indicador a avaluar en els anàlisis posteriors.

Prenent les dades operacionals i transaccionals de les campanyes de captació anteriors i els estudiants actuals, l'objectiu principal del projecte consistirà en cercar patrons i regles ocultes dins de les dades que permetin augmentar l'eficiència en la captació d'estudiants, geogràfica i econòmicament parlant.

Abstract (in English, 250 words or less):

Nowadays, due to the usage of System Information technologies, the storage and usage of all kind of data is growing exponentially in almost any organization. In order to take advantage of this information, it is well known that competitive enterprises are using several techniques to extract knowledge of this source.

Starting at the point that Business Intelligence wants to be as competitive and effective as possible, this project will implement Data Mining techniques looking for reasonable knowledge patterns in order to take profit of it.

In this project we will work in a real but test environment where we are representing a private Business School which wants to focus their resources on a specific geographical area looking for future students. To study in our school, it is necessary to pass several tests and assessments so it is very important to analyse this factor as a success indicator in a concrete zone.

Taking all transactional and operational data of previous recruitment campaigns, the main objective is to extract patterns and rules that should ensure more efficiency in the student recruitment process, geographically and economically speaking.

Paraules clau (entre 4 i 8):

Data Mining, Business Intelligence, R, Association rules, Classification rules

Índex

Data Mining 1

En busca de patrons geogràfics per focalitzar recursos 1

1. Introducció 1
- **1.1 Context i justificació del Treball** 1
- **1.2 Objectius del Treball** 1
- **1.3 Enfocament i mètode seguit** 2
- **1.4 Planificació del Treball** 3
- **1.5 Breu descripció dels altres capítols de la memòria** 4

2. El procés de mineria de dades 5
- **2.1 Primera fase – Preparació de l'entorn de treball** 5
- **2.2 Segona fase – Origen, obtenció, anàlisi i tractament de les dades** 5
  - 2.2.1. Origen i obtenció de les dades. 5
  - 2.2.2. Anàlisi preliminar de les dades. 6
  - 2.2.3. Tractament de les dades. 8
  - 2.2.4. Anàlisi descriptiu de les dades. "Data set". 10
- **2.3 Tercera fase – Generació dels models** 27
  - 2.3.1. Anàlisi de les campanyes anteriors. Algorismes associatiu i classificatori. 27
  - 2.3.2. Anàlisi dels alumnes actuals. Algorismes associatiu i classificatori. 36
- **2.4 Quarta fase – Avaluació i interpretació global** 41

3. Conclusions 45
- **3.1. Lliçons apreses del treball** 45
- **3.2. Assoliment d'objectius plantejats** 45
- **3.3. Seguiment de la planificació i metodologia** 46
- **3.4. Línies de treball futur** 47

4. Glossari 48

5. Bibliografia 49

1. Introducció

1.1 Context i justificació del Treball

Aquest projecte sorgeix del *feedback* rebut com a membre del departament de Sistemes d'Informació d'una escola de negocis. És una escola internacional, on la majoria d'estudiants són estrangers i que té un creixement constant des de fa bastants anys pel que fa a la quantitat d'alumnes matriculats. Aquesta evolució favorable és gràcies a una tasca exhaustiva de captació d'interessats i, que requereix doncs, una assignació de recursos molt important cap al departament de relacions exteriors.

Una de les principals accions a l'hora de registrar interessats és la de visites a fires, escoles, universitats, centres d'estudis diversos... d'arreu del món, d'on se n'extreu una gran quantitat de dades gràcies a una App que es sincronitza amb el CRM (on clarament el *customer* són els estudiants finals).

A més, l'escola exigeix una qualitat base de l'estudiant i, per això, d'entre tots els interessats tant sols una part podran accedir finalment als estudis, després d'un estricte procés d'entrada amb proves aptitudinals, d'idiomes, *assessments* i entrevistes personals.

Actualment, la captació és realitza de forma massiva i es fan viatges arreu del món sense gaire informació prèvia sobre si és un bon destí o no.

Així doncs, l'estratègia de negoci comença per mantenir un creixement constant tot optimitzant els recursos en les campanyes de màrqueting i captació de nous estudiants.

Per a dur a terme aquesta fita es vol analitzar les dades històriques de l'empresa i definir on i com cal aplicar els esforços, ja que també es vol conèixer si existeix alguna preferència per les fires, les visites o la captació online.

1.2 Objectius del Treball

L'objectiu principal d'aquest projecte correspon a posar en funcionament un projecte de Mineria de Dades en busca de models vàlids que permetin descriure grups geogràfics (destinacions) potencialment rentables pel que fa a possibles candidats així com el medi idoni (visites in situ mitjançant fires i visites o bé a través de contacte online).

Els objectius del projecte són doncs:

- Obtenció de les millors destinacions on fer campanyes de captació, entenent com a factor d'èxit els alumnes acceptats finalment (i no el volum de candidats inscrits).

- Establir el medi més òptim i segur pel que fa a la captació de candidats entre visites, fires i online.

- Avaluar si l'ús actual de recursos econòmics als diferents països potencials ha sigut l'adient.

D'aquesta forma, els sub-objectius que se'n deriven empresarialment parlant i que corresponen a l'abast del projecte són:

- Focalitzar esforços. Destinar recursos econòmics en aquelles zones i mitjans favorables ne la captació d'estudiants.

- Maximitzar els resultats.

    - Major quantitat de candidats/interessats.

    - Major taxa d'acceptació per accedir als estudis.

    - Derivat dels dos anteriors obtenir un major nombre d'estudiants.

1.3 Enfocament i mètode seguit

Per assolir els objectius d'aquest projecte es decideix executar un procés d'anàlisi de les dades actuals, tot seguint el cicle de vida complet d'un projecte de Mineria de Dades.

A nivell de processament i obtenció de resultats, la metodologia consisteix en emparar algorismes d'aprenentatge automàtic dins del programari estadístic i de càlcul R, tot executant processos d'associació i classificació.

Conseqüentment, el mètode de treball a seguir es distribueix en quatre fases principals:

- **Primera fase**

    Preparació de l'entorn d'anàlisi pel que fa al software i al hardware de computació.

- **Segona fase**

  Com l'objectiu del projecte ja està clarament definit, cal localitzar i establir l'origen de les dades, seleccionar aquelles importants com a *input* al sistema i preparar-les (neteja, estandardització...).

  En aquesta fase també s'estudiarà el contingut de les dades seleccionades.

- **Tercera fase**

  Procés d'anàlisi i de cerca de models mitjançant tècniques de mineria de dades. En aquest cas s'empararà el programari estadístic i de visualització R, tot aplicant algorismes concrets segons les dades obtingudes. Per la naturalesa d'aquestes i pels objectius definits els millors models a cercar són els associatius i els classificatoris.

- **Quarta fase**

  Avaluació i interpretació del el model obtingut.

  D'aquesta fase se n'esperen obtenir les destinacions més efectives pel que fa a la captació i acceptació de candidats alhora de definir quin és el medi més idoni.

  Es compararan els models entre ells i amb les dades estadístiques i descriptives dels atributs. També es farà una reflexió sobre si les inversions en els països potencials de destinació han sigut adients.

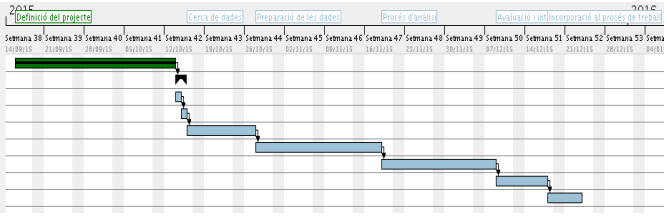
1.4 Planificació del Treball

Seguidament s'adjunta la planificació inicial amb la durada i procés de cadascuna de les etapes del projecte. Posteriorment de la finalització del projecte, esdevindrà la fase de construcció de la memòria final.

Nom	Inici	Finalització	Durada	Progrés
Definició del projecte	16/09/15	13/10/15	20	100
Instal·lació	14/10/15	15/10/15	2	0
Maquinari	14/10/15	14/10/15	1	0
Programari	15/10/15	15/10/15	1	0
Cerca de dades	16/10/15	27/10/15	8	0
Preparació de les dades	28/10/15	18/11/15	16	0
Procés d'anàlisi	19/11/15	08/12/15	14	0
Avaluació i interpretació	09/12/15	17/12/15	7	0
Incorporació al procés de treball	18/12/15	23/12/15	4	0

Cada etapa és dependent de l'anterior i per tant, el diagrama temporal de Gantt té la forma de la següent figura:

Nom	Inici	Finalització	Durada	Progrés
Definició del projecte	16/09/15	13/10/15	20	100
Instal·lació	14/10/15	15/10/15	2	0
Maquinari	14/10/15	14/10/15	1	0
Programari	15/10/15	15/10/15	1	0
Cerca de dades	16/10/15	27/10/15	8	0
Preparació de les dades	28/10/15	18/11/15	16	0
Procés d'anàlisi	19/11/15	08/12/15	14	0
Avaluació i interpretació	09/12/15	17/12/15	7	0
Incorporació al procés de treball	18/12/15	23/12/15	4	0



Molt probablement es produirà una iteració entre les etapes de preparació de dades, procés d'anàlisi i avaluació i interpretació, tot esdevenint en una execució cíclica d'aquestes.

1.5 Breu descripció dels altres capítols de la memòria

La resta de capítols corresponen al cicle de vida d'un procés de mineria de dades. Cadascun d'ells té una correlació directe amb les fases del projecte definides al punt 1.3 de mètode seguit i, per tant, la breu descripció del contingut ha sigut definida amb anterioritat.

Aquests capítols estan catalogats per a cada fase i, dins d'aquestes, es generen sub-apartats amb cascuna de les accions i procediments que s'han hagut d'executar.

2. El procés de mineria de dades

2.1 Primera fase – Preparació de l'entorn de treball

En aquesta primera fase, es prepara l'entorn de treball i les eines que seran emparades al llarg del projecte. Es necessitarà doncs alguna eina que permeti l'anàlisi de les dades a nivell descriptiu i estadístic, una altra per al processament i tractament de les dades i finalment una eina encarregada d'executar els algorismes de mineria de dades.

Per dur a terme les tasques prèviament esmentades, es prepara un entorn de treball amb les següents eines:

- **Programari R**. Mitjançant aquesta eina es realitzaran els anàlisis de les dades corresponents així com les execucions dels algorismes de mineria de dades. Per tal de treballar en un entorn més còmode, s'instal·la conjuntament R Studio i RCommander, que permeten interactuar amb el nucli d' R d'una forma més gràfica i ordenada. La versió concreta del nucli d' R emparat correspon a la 3.2.2 GUI 1.66 Mavericks build (6996).
    - Llibreries extra que s'han instal·lat i utilitzat:
        - Rcmdr (entorn gràfic per treballar amb les dades i aplicar funcions programades)
        - C50 (arbres decisionals)
        - arules (algorisme associatiu)
        - arulesViz (visualització dels resultats associatius)
        - class (generació de regles bayesianes)
        - e1071 (generació de regles bayesianes)
- **Excel 2016**. Aquesta serà l'encarregada del tractament de les dades, realització de filtres, creació de taules noves, etc.

Aquest programari és executat sobre un maquinari amb 16 GB de RAM i un Intel Core i7 a 2,8 GHz. El sistema operatiu base és MAC OS X 10.11.2.

2.2 Segona fase – Origen, obtenció, anàlisi i tractament de les dades

2.2.1. Origen i obtenció de les dades.

Les dades d'entrada al projecte es divideixen en tres grups o fitxers principals:

1. Dades de campanyes de captació anteriors (TOTES)
2. Dades d'estudiants actuals de primer curs provinents de campanyes de captació (es pot considerar un filtre de l'anterior i contindrà informació extra).

3. Despesa econòmica per país.

Les fonts de dades obtingudes en format .csv es convertiran i tractaran per acabar obtenint uns fitxers .xlsx, que posteriorment seran carregats al programari R.

Per motius de confidencialitat i privadesa, el conjunt de dades d'entrada ha sigut modificat i preparat amb dades fictícies, fet que podria condicionar el resultat del projecte. No obstant, tot i no correspondre's amb la realitat, de forma majoritària s'ha pogut relacionar cada valor concret amb un altre per tal que els resultats siguin semblants.

2.2.2. Anàlisi preliminar de les dades.

Les dades obtingudes i generades, sense discretització, normalització, neteja ni tractaments presenten la estructura que seguidament s'exposa. Entre [] està indicat el nom de l'atribut a la font de dades.

Dades de campanyes anteriors

S'han omès tot tipus d'atributs personals que no aporten informació classificatòria posterior, tals com Nom, Cognoms, Correu electrònic, DNI, etc.

- Lloc de contacte [lloc_contacte]: llista de llocs de contacte coneguts i codificats (Ex: Fira de l'Estudiant de Barcelona (1_23)).

- Ciutat [ciutat]: ciutat on s'ha realitzat el contacte.

- País [pais]: país on s'ha realitzat el contacte.

- Data de contacte [data_contacte]: data de contacte en format "YYYY-MM-DD hh:mm:ss".

- Medi de contacte [medi_contacte]: medi mitjançant el qual cada candidat va entrar en el procés d'interessat pels estudis. Pot prendre tres valors diferents [Feria, Visita, Online].

- Programa d'interès [programa_interes]: camp tancat dels programes pels quals un candidat pot estar interessat. [MSc Management, MSc Finnance, MBA].

- Centre educatiu d'origen [centre_origen]: camp lliure que especifica el centre del qual prové cada candidat. Al no ser un cap tancat, caldrà veure posteriorment si aporta informació o no.

- Tipus de centre educatiu d'origen [tipus_centre_origen]: aquest atribut pot prendre els valors [Públic, Privat, Concertat].

- Finalista [finalista]: valor que indica si un candidat va seguir interessat fins al final, és a dir, va mantenir la intenció d'estudiar en el centre fins a arribar a les proves d'accés o si pel contrari va abandonar en la intenció de matricular-se en els estudis. Pot prendre els valors de [Abandonament, Finalista].

- Seguiment personalitzat [seguiment_pers]: camp booleà que informa si sobre aquest candidat s'ha realitzat un seguiment personalitzat i pot prendre els valors [Si, No]. El seguiment personalitzat, s'acostuma a realitzar sobre candidats que ja mostren un interès molt elevat en un primer contacte o bé per insistència en la demanda d'informació.

- Resultat en les proves aptitudinals [res_aptitudinal]: camp numèric entre 0 i 100 que indica la nota d'aquest procés d'avaluació.

- Resultat en les proves lingüístiques [res_llengua]: camp numèric entre 0 i 100 que indica la nota d'aquest procés d'avaluació.

- Resultat en les proves aptitudinals [res_calcul]: camp numèric entre 0 i 100 que indica la nota d'aquest procés d'avaluació.

- Resultat de *l'assessement* [res_assessement]: camp booleà que indicarà si és un estudiant recomanat o no [Recomanat, No recomanat].

- Acceptat [acceptat]: camp booleà que indica si va ser acceptat o no. [Acceptat, No acceptat].

Dades d'estudiants actuals (primer curs)

- Prové de campanyes anteriors [captat]: camp booleà que informa si ha sigut captat per una campanya o bé es va presentar al centre per iniciativa pròpia (tot superant posteriorment les proves d'accés com tot els altres candidats). Valors [Si, No].

- Nombre d'idiomes que parla [nombre_idiomes]: nombre enter que representa la quantitat d'idiomes que parla l'estudiant.

- Llengua materna [llengua_natal]: idioma matern de l'estudiant.

- Altres llengües [altres_llengues]: llista separada per comes de les llengües que parla.

- Nacionalitat [nacionalitat]: nacionalitat de l'estudiant.

- Sexe [sexe]: indica el sexe de l'estudiant i pot prendre els valors [Home, Dona]

- Data de naixement [data_naixament]: camp amb la data de naixement de l'estudiant.

Dades de despesa econòmica per país

- País [pais]: país destí.

- Despesa [despesa]: despesa en € en campanyes/desplaçaments a un país concret.

Tenint en compte que som una escola de negocis mitjana/gran, hi ha una mitja de 400 estudiants matriculats anualment en el primer curs (contant tots els tipus d'estudi). Així doncs, aquesta xifra correspon directament a les dades d'estudiants actuals com a font d'entrada al projecte del segon grup de dades.

Pel que fa a dades de campanyes anteriors, s'obtenen com a entrada un total de 3600 registres aproximadament. Aquests corresponen a campanyes que comencen des del 2012 fins a l'actualitat.

La despesa econòmica per país és la suma del destinat en les diferents campanyes anteriors. Aquest conjunt de dades pot ser útil alhora de diferenciar/decidir si una zona geogràfica concreta ha sigut productiva i si té sentit continuar invertint allà.

2.2.3. Tractament de les dades.

Degut a que les dades provenen dels sistemes d'informació interns, i que han sigut obtingudes amb unes eines dissenyades especialment per aquest objectiu, es té l'avantatge que gran part dels atributs ja tenen el format desitjat. Tanmateix, s'ha fet una preselecció d'atributs excloent aquells que contenen informació personal mentre que s'han adaptat/generat les dades per a aquest projecte.

Com a conseqüència del punt anterior (origen controlat de les dades), no és necessari fer el procés de tractament de manca de dades ja que el conjunt d'atributs està complet.

Tot seguit s'exposen, agrupades per tipus de tractament, cadascuna de les modificacions que s'han s'han aplicat als atributs inicials de la font de dades original.

Discretització

- Atributs **res_aptitudinal, res_llengua** i **res_calcul**. La nota mínima per aprovar és 65 i, tenint en compte aquest llindar, s'agrupen o discretitzen les dades seguint els criteris següents.

Criteri	Nou valor
Si nota < 65	Suspens
Si 65 < nota < 75	Suficient
Si 75 < nota < 85	Desitjable
Si nota > 85	Excel·lent

D'aquesta manera l'atribut esdevé molt més explicatiu, tot permetent englobar un conjunt més elevat de candidats en un mateix grup. Això és especialment útil a l'hora de fer podes en algorismes decisionals i discriminar poblacions en l'anàlisi posterior.

Creació de nous atributs

- Atribut **data_naixament**. Es decideix crear un nou atribut **edat** a partir de la data de naixement que definirà l'edat de l'estudiant actual. Aquesta es calcula mitjançant la data actual.

- A partir de la **data_contacte** es genera un nou camp anomenat **campanya**, que correspondrà a l'any de la campanya en la qual es van recopilar les dades calculat a partir de la data de contacte. Aquest fet podrà englobar més registres dins d'un mateix grup fet que permetrà als algorismes predictius ser més eficients.

Reducció de dimensionalitat

- Supressió de l'atribut **centre_origen**. Després d'analitzar els valors d'aquest camp, s'observa clarament que existeixen masses valors dispars alhora que no compta amb una codificació explícita (fet que també seria pràcticament impossible degut a que el nombre d'instituts, escoles, centres educatius, etc. és gairebé il·limitat). Així doncs, la informació final que pugui aportar es veu clarament reduïda alhora de definir conjunts de candidats. Ha sigut una decisió complicada ja que inicialment semblava que podia interessar aquest atribut però els propis valors que pren aquest camp n'ha forçat la supressió.

- L'atribut **altres_llengues** tampoc sembla que pugui arribar a ser útil. Es pren la decisió de suprimir-lo tot posant èmfasi en el nombre d'idiomes que parla, atribut que ja es coneix.

- Per últim, la **data_contacte** ha sigut emparada per a generar un nou atribut que resultarà més útil posteriorment classificatòriament parlant. Així doncs, aquest atribut inicial ja no és necessari.

Com a observació a tenir en compte alhora de realitzar l'anàlisi tots aquells candidats que tenen estat "**No finalista**" en el camp **finalista** tindran conseqüentment buits els valors corresponents als resultats de les proves així com el de si ha sigut acceptat o no. És una puntualització molt important alhora de no distorsionar els resultats. Esdevé d'aquesta forma una categoria nova en els caps implicats igual a NA.

El procés de mineria de dades esdevé en un procés cíclic i part d'aquests tractaments han sigut conseqüència directa del procés de generació dels models.

Un altre tractament, no menys important, ha sigut l'adaptació de les dades a l'entrada al programari d'anàlisi R. Així doncs, s'ha tingut que adaptar la codificació del text, el tipus de camps (data, text, número), etc.

2.2.4. Anàlisi descriptiu de les dades. "Data set".

Dades de les campanyes anteriors

Estructura de les dades

Aquest apartat defineix els camps/atributs que conté i els nivells o valors diferents que pren cadascun d'aquests. Cal recordar que aquest fitxer ja ha sigut tractat a la fase anterior.

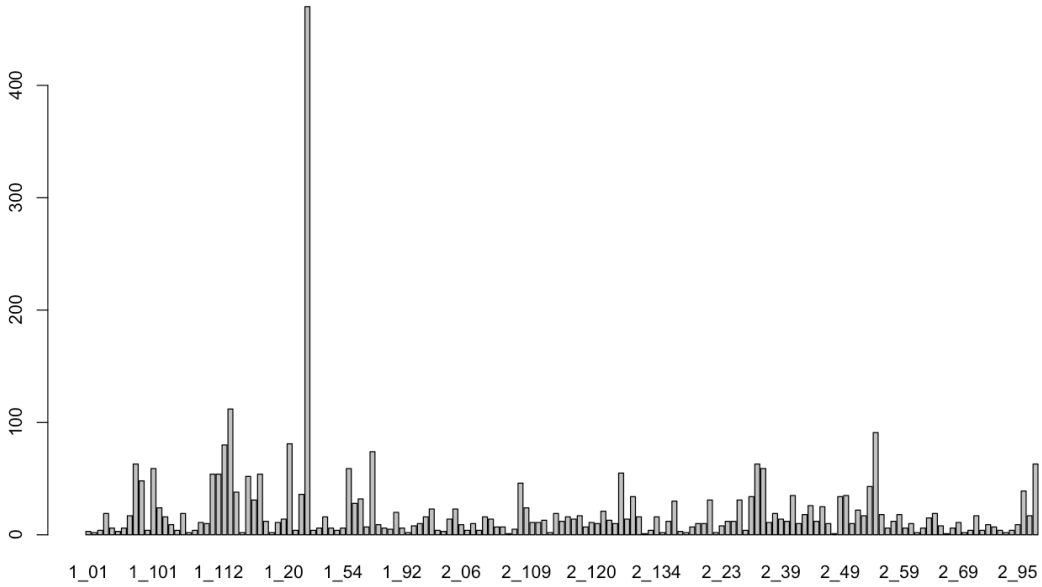
Es compon d'un total de **3041 observacions** i **14 atributs**. Cadascun dels atributs presenta la següent distribució (estructura).

- lloc_contacte: **161 valors** diferents
- ciutat: **87 valors** diferents
- pais: **43 valors** diferents
- campanya: **4 valors** diferents [2012-13, 2013-14, 2014-15, 2015-16]
- medi_contacte: **3 valors** diferents [Feria, Online, Visita]
- programa_interes**: 3 valors** diferents [MBA, MscF, MscM]
- tipus_centre_origen: **3 valors** diferents [Concertat, Privat, Public]
- finalista: **2 valors** diferents [0,1]
- seguiment_pers, res_assesement i acceptat: **3 valors** diferents [0,1,NA]
- res_aptitudinal, res_llengua i res_calcul : **5 valors** diferents [Suspens, Suficient, Desitjable, Excel·lent]

Anàlisi descriptiu

- **lloc_contacte**

Histograma de la freqüència d'aparició dels diferents llocs de contacte.



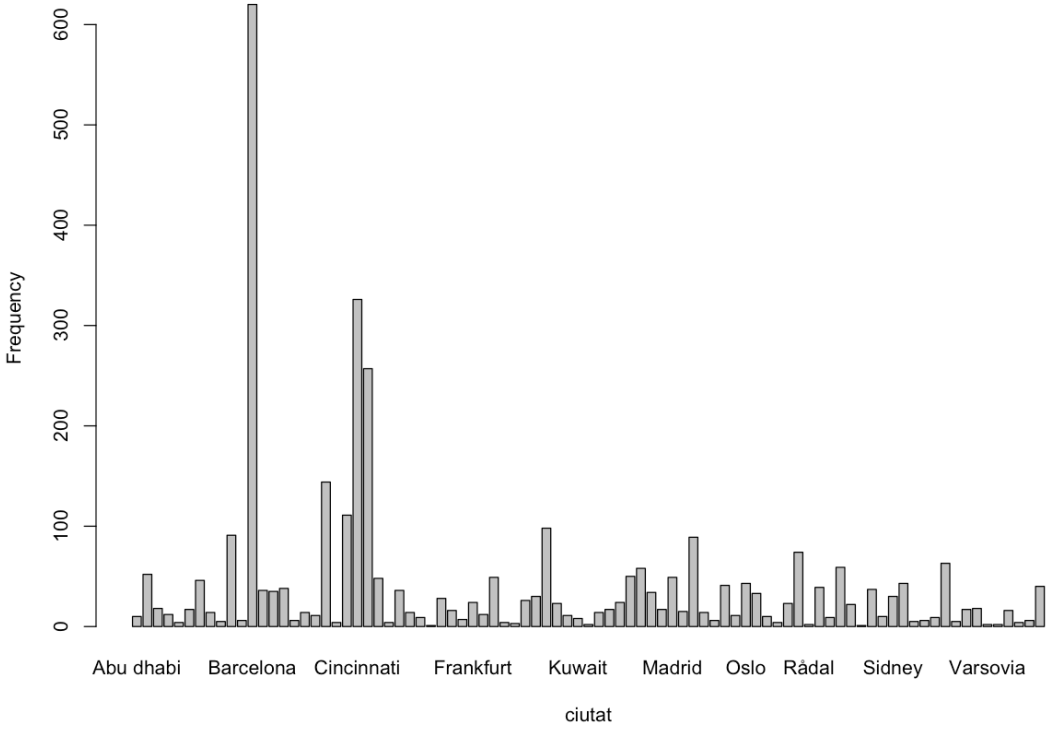
Com es pot observar, clarament i ha un lloc de contacte que predomina sobre la resta. Aquest és el 1_24, corresponent a un lloc de Barcelona → Espanya.

Pel que fa als 6 llocs de contacte més freqüents obtenim la següent llista.

```
lloc_contacte
1_23   : 470
1_114  : 112
2_55   :  91
1_200  :  81
1_113  :  80
1_74   :  74
```

Sembla doncs que un bon estimador pel futur anàlisi pot ser la quantitat d'acceptats provinents d'aquests llocs de contacte.

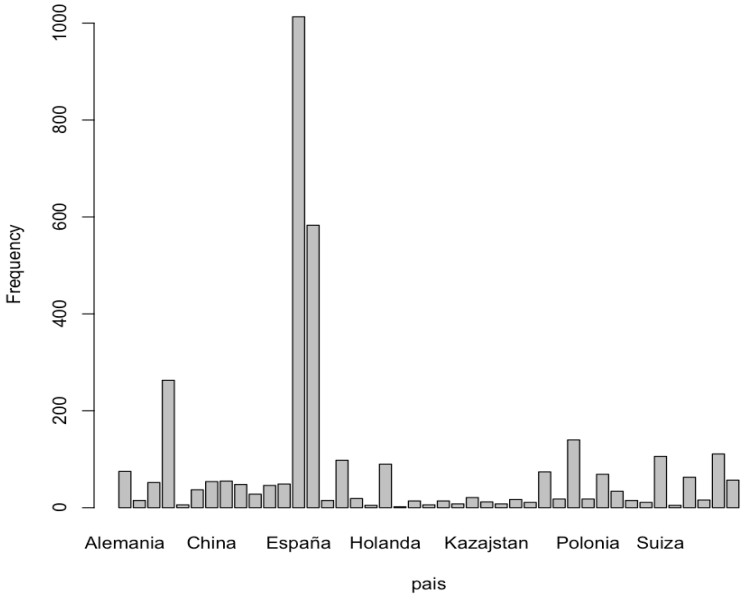
- **ciutat**



La ciutat amb major nombre de contactes és Barcelona, tot seguit de Cincinnati i Connecticut. Les 6 ciutats amb major nombre de candidats registrats són les següents:

```
        ciutat
Barcelona  : 620
Cincinnati : 326
Connecticut: 257
Canberra   : 144
Caracas    : 111
Helsinki   :  98
```

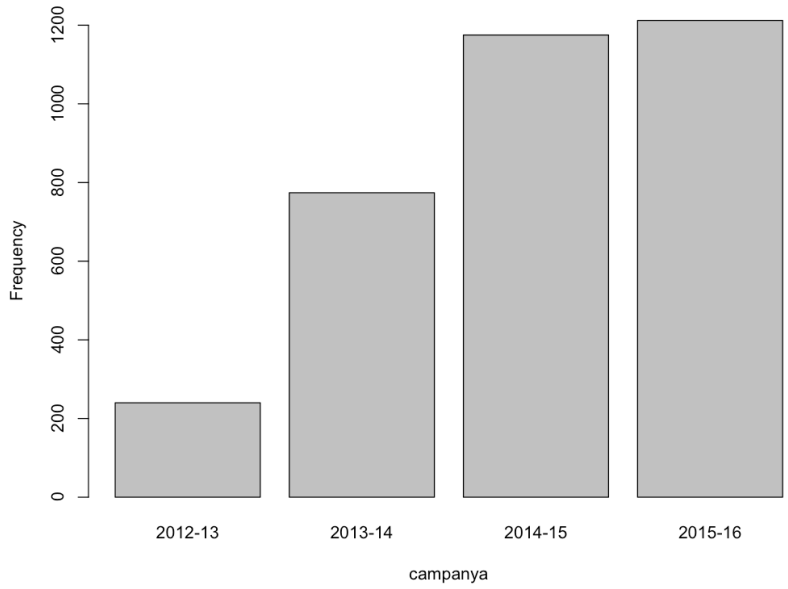
- **pais**



Previsiblement, derivat dels resultats anteriors, el país amb major nombre de candidats recaptats és Espanya. Les sis primeres posicions les conformen els següents països:

```
                pais
España        :1013
Estados Unidos: 583
Australia     : 263
Noruega       : 140
Venezuela     : 111
Suiza         : 106
```

- **campanya**

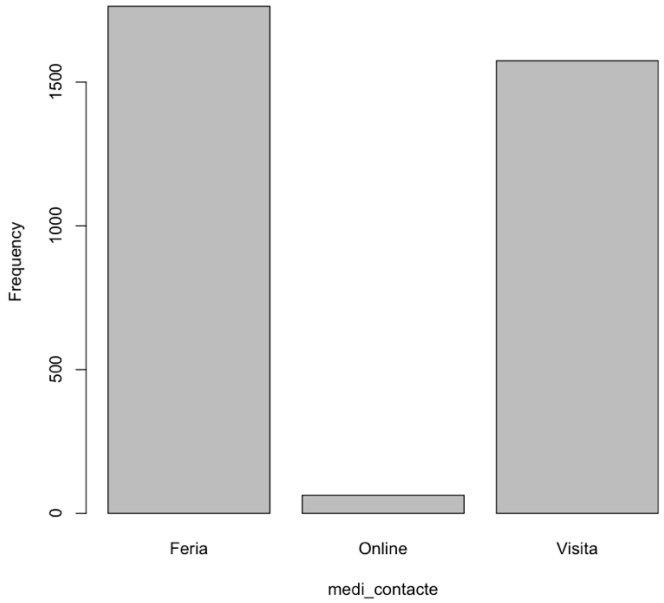


Aquests resultats deixen entreveure que l'activitat pel que fa a les campanyes de captació d'alumnes ha anat en augment.

El candidats captats en cada campanya són els següents:

```
   campanya
2012-13: 240
2013-14: 774
2014-15:1175
2015-16:1212
```

- **medi_contacte**

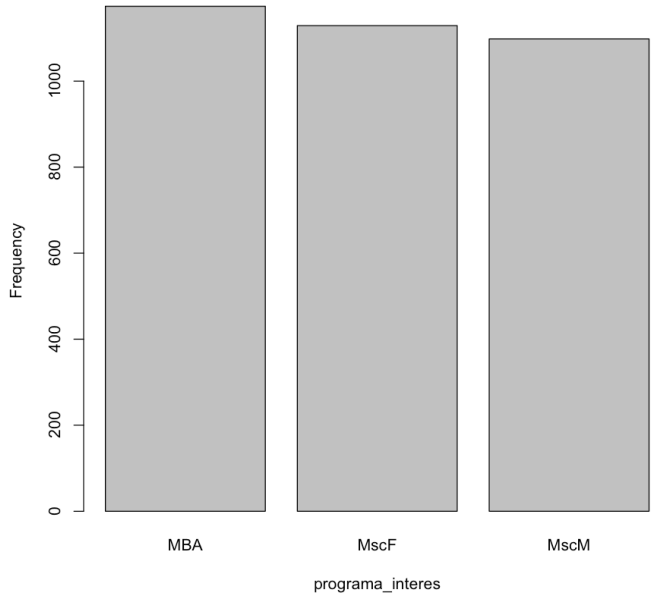


El medi de contacte amb els candidats clarament queda repartit entre Visites i Fires, sent la captació online una part poc significativa (a priori). Un supòsit en aquest punt és que el percentatge d'acceptats Online sigui molt elevat, definint doncs una possible estratègia a focalitzar esforços en aquest aspecte. Caldrà veure si els diferents models també arriben a aquesta conclusió.

El nombre de candidats provinents de cada medi són:

```
medi_contacte
Feria :1764
Online:  63
Visita:1574
```

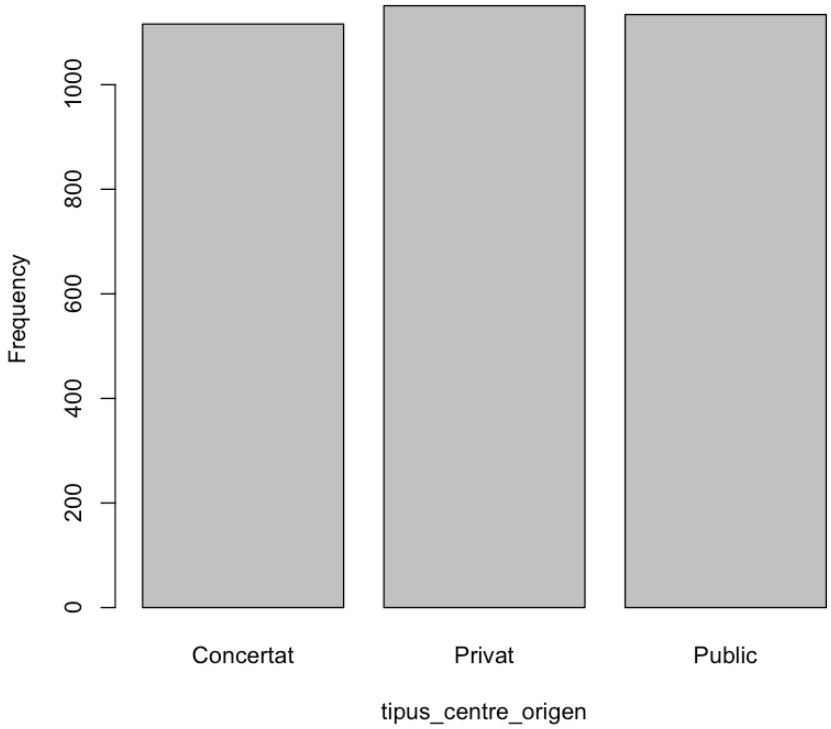
- **programa_interes**



Una possible interpretació d'aquest resultat és que el programa d'interès no és significatiu alhora de triar el nostre centre, ja que pràcticament tots tenen el mateix nombre d'interessats. Caldrà promocionar per igual els tres tipus de programes.

```
programa_interes
MBA :1174
MscF:1129
MscM:1098
```

- **tipus_centre_origen**

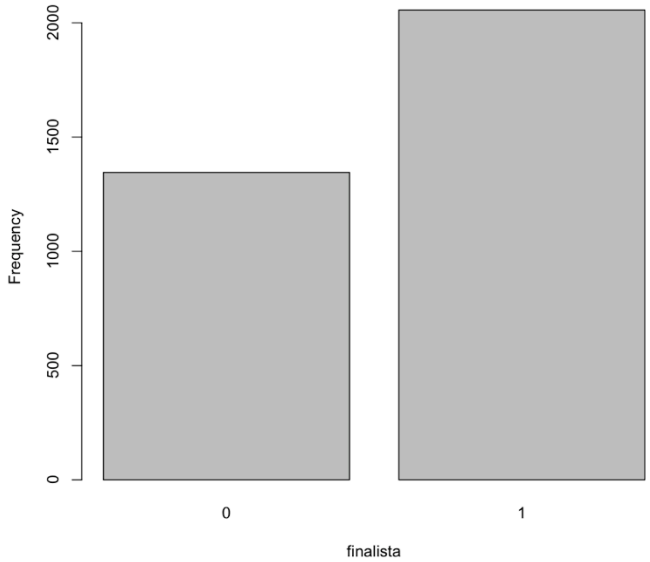


A aquest atribut li passa exactament el mateix que a l'anterior però, en aquest cas, si que és un atribut que hauria de participar en el procés de mineria.

Per exemple, podria succeir que establint com a output que un candidat sigui acceptat, aquests probablement provinguin d'un tipus de centre determinat. L'encarregat d'oferir-nos aquesta conclusió, si és que existeix, serà el model resultant.

```
tipus_centre_origen
Concertat:1116
Privat   :1151
Public   :1134
```

- **finalista**

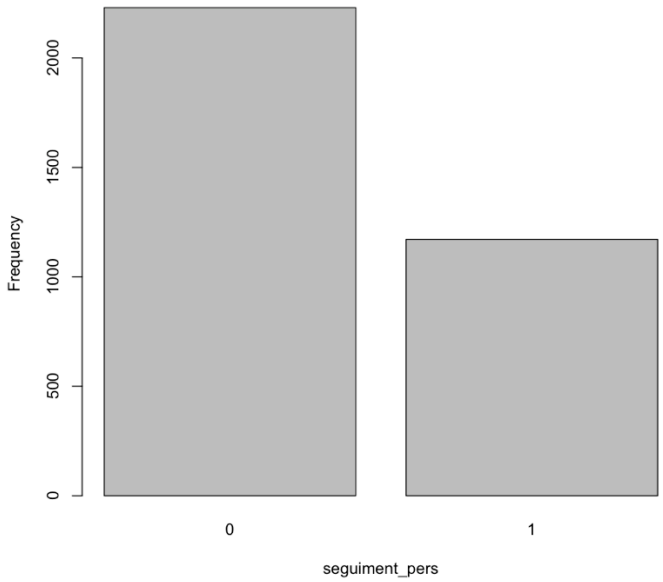


El resultat d'aquest atribut indica que la majoria de candidats que s'interessen pels nostres estudis continuen estant interessats fins a arribar a fer les proves de selecció.

La pròpia mitjana de l'atribut ja defineix aquest comportament.

```
  finalista
Min.   :0.0000
1st Qu.:0.0000
Median :1.0000
Mean   :0.6045
3rd Qu.:1.0000
Max.   :1.0000
```

- **seguiment_pers**



La majoria de candidats, gairebé 2/3, no tenen un seguiment personalitat degut a que molts d'aquests passen a estat "Finalista = 0" i per tant no

entren en el procés de selecció de candidats (valors NA en les proves i *l'assessement*).

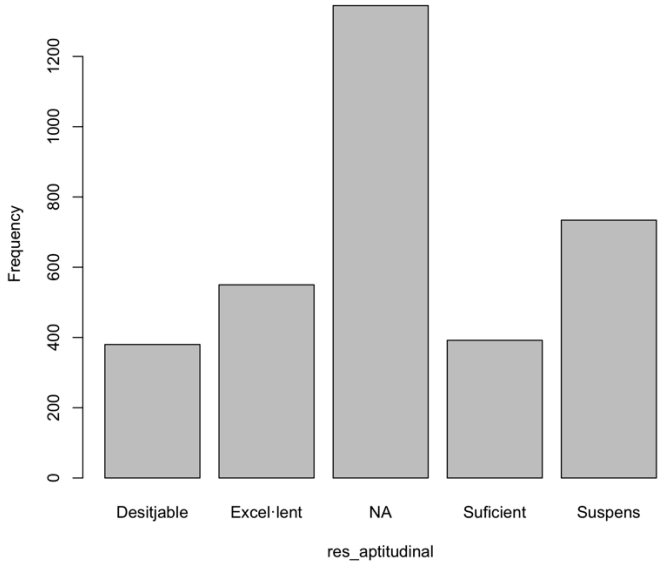
```
seguiment_pers
Min.   :0.0000
1st Qu.:0.0000
Median :0.0000
Mean   :0.3443
3rd Qu.:1.0000
Max.   :1.0000
```

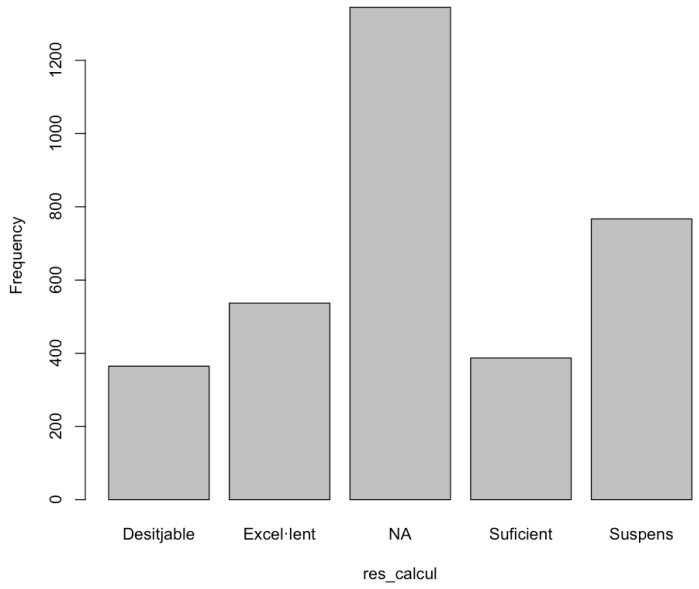
- **Resultats de les proves**: res_aptitudinal, res_calcul i res_llengua

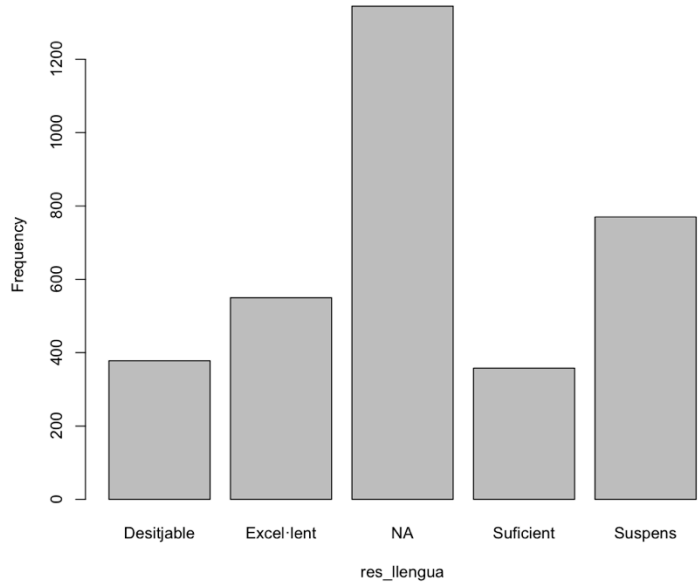
```
  res_aptitudinal     res_llengua        res_calcul
Desitjable: 380   Desitjable: 378   Desitjable: 365
Excel·lent: 550   Excel·lent: 550   Excel·lent: 537
NA         :1345  NA         :1345  NA         :1345
Suficient : 392   Suficient : 358   Suficient : 387
Suspens   : 734   Suspens   : 770   Suspens   : 767
```

Aquests tres atributs els s'estudien com un conjunt, ja que tant sols que un d'ells es suspengui automàticament els resultats dels altres no seran contemplats. Així doncs, és necessari aprovar les tres i *l'assessement* per a poder ser acceptat.

El resultat global defineix que la majoria de candidats o bé suspèn o bé te un nivell Excel·lent.

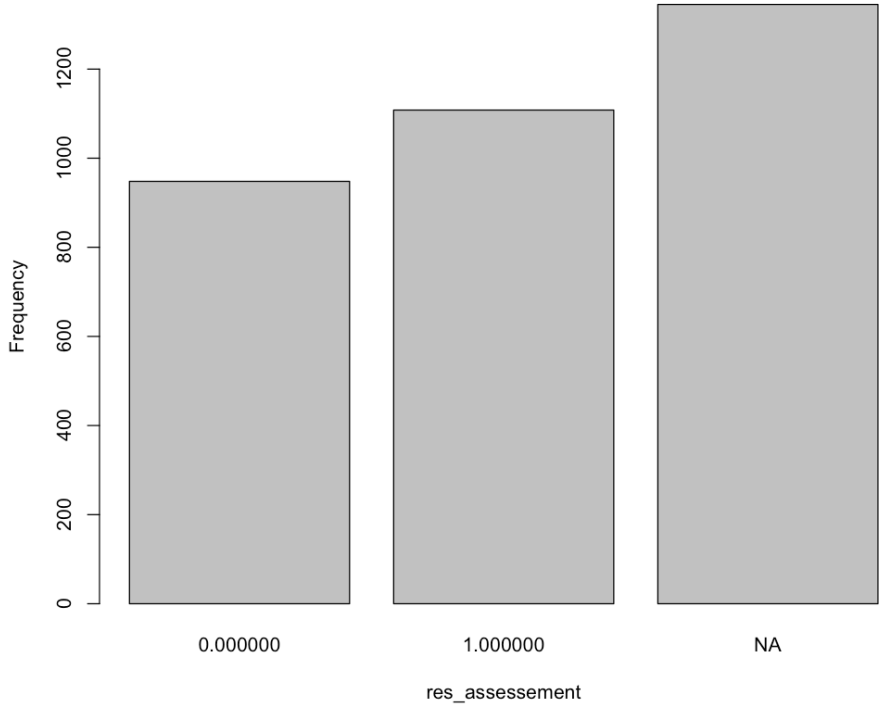






Els resultats NA corresponen a aquells que han abandonat (no finalistes).

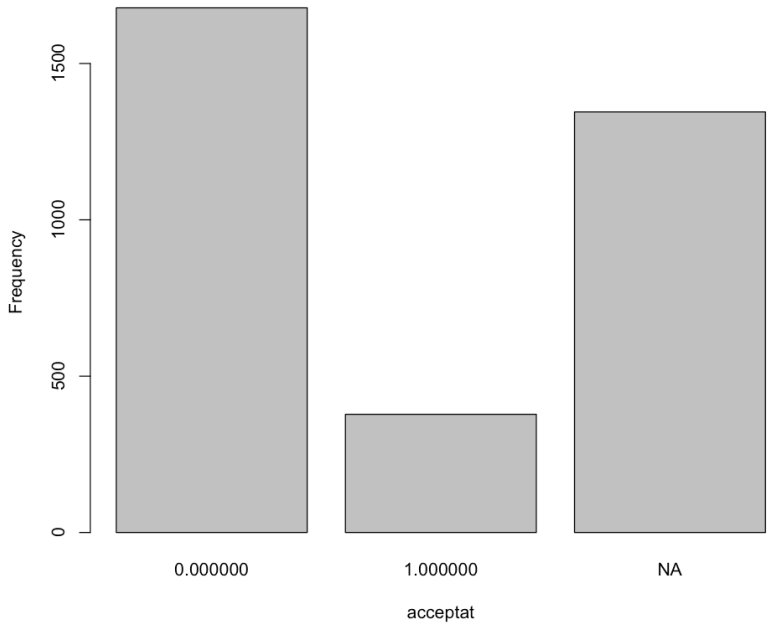
- **res_assessement**



La majoria d'alumnes que realitzen *l'assessement* l'aproven. No obstant, aquest atribut per si sol no és indicatiu de que siguin acceptats o no.

```
res_assessement
0.000000: 948
1.000000:1108
NA      :1345
```

- **acceptat**



Finalment, del total de candidats finalistes i que han realitzat les proves, tant sols el 18.27 % són acceptats.

```
     acceptat
0.000000:1678
1.000000: 378
NA      :1345
```

Aquest atribut, que és l'indicador clau a analitzar en el procés de mineria de dades, representa el valor de referència a l'hora d'avaluar la sortida dels algorismes, és a dir, en els models.

Dades alumnes actuals de primer

<u>Estructura de les dades</u>

En primera instància s'ha redefinit l'atribut "nacionalitat" per "pais_origen", degut a que el format de les dades precisament representava el segon (Ex: "Francia" i no "Francés").

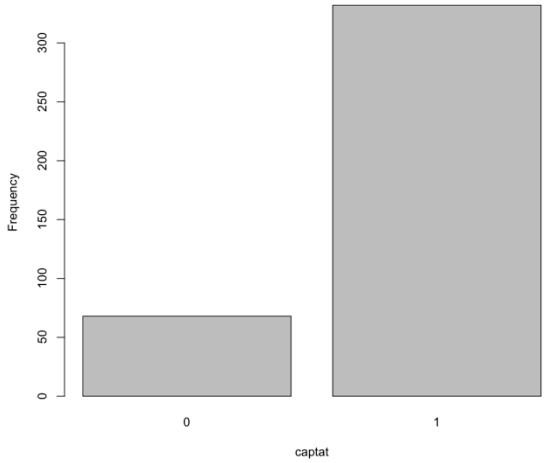
Al analitzar la llengua natal i el país d'origen esdevé una correlació directe entre aquest dos (a excepció dels països on disposen de diverses llengües oficials). No obstant, i de cara a focalitzar esforços geogràfics, es dedueix que l'atribut "llengua_natal" no aporta cap informació i és molt més interessant mantenir el país d'origen. Així doncs, aquest atribut ha sigut eliminat de la base de dades d'entrada per tal reduir-ne la dimensionalitat.

Aquesta font de dades es compon finalment d'un total de **400 observacions** i **5 atributs**. Cadascun dels atributs presenta la següent distribució (estructura).

- captat: **2 nivells/valors diferents**. [1,0] depenent si és cert o fals.
- nombre_idiomes: camp **enter lliure**.
- pais_origen: **34 nivells/valors diferents.**
- sexe: **2 nivells/valors diferents**. [1,0] depenent si és cert o fals.
- edat: camp **enter lliure**.

Anàlisi descriptiu

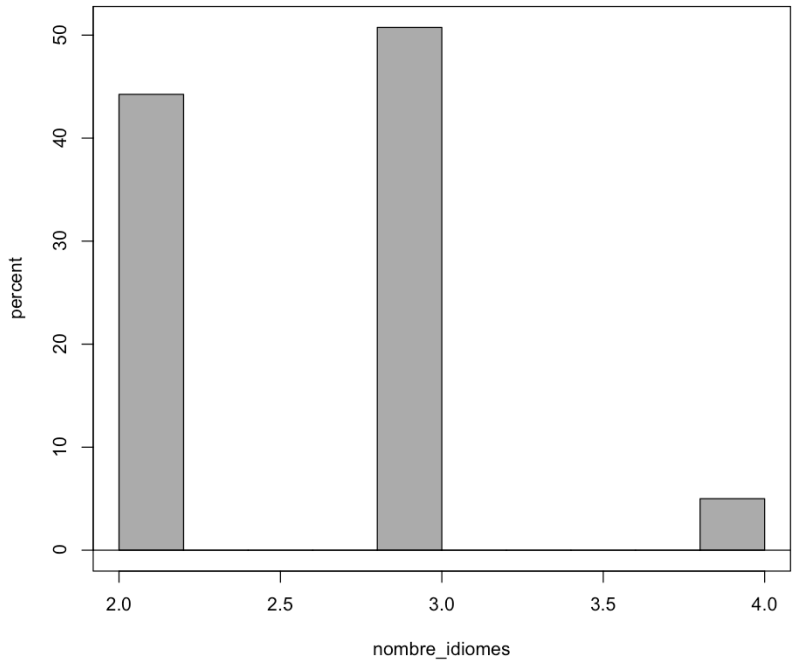
- **captat**



Aquest atribut indica si els actuals alumnes van ser captats en campanyes anteriors o per contra van accedir als estudis per iniciativa pròpia. S'observa que en el 83% dels casos proven de les campanyes, reforçant així la importància d'aquest procés alhora de seguir amb un bon creixement empresarial.

```
     captat
 Min.   :0.00
 1st Qu.:1.00
 Median :1.00
 Mean   :0.83
 3rd Qu.:1.00
 Max.   :1.00
```

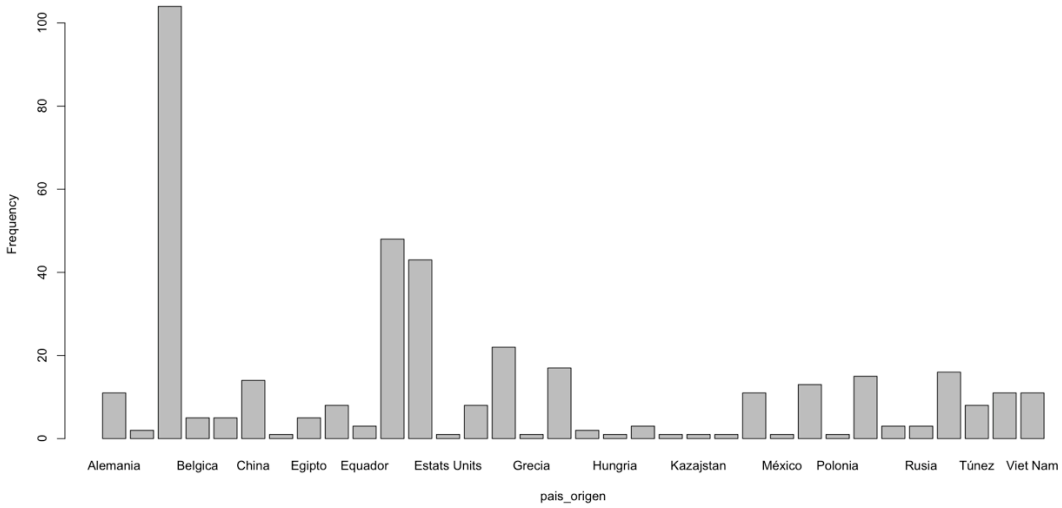
- **nombre_idiomes**



El perfil d'alumnes actual parla majoritàriament de 2 a 3 idiomes, sumant entre aquests més del 90% dels casos. El mínim d'idiomes que parlen és 2 i el màxim és 4.

```
nombre_idiomes
Min.   :2.000
1st Qu.:2.000
Median :3.000
Mean   :2.607
3rd Qu.:3.000
Max.   :4.000
```

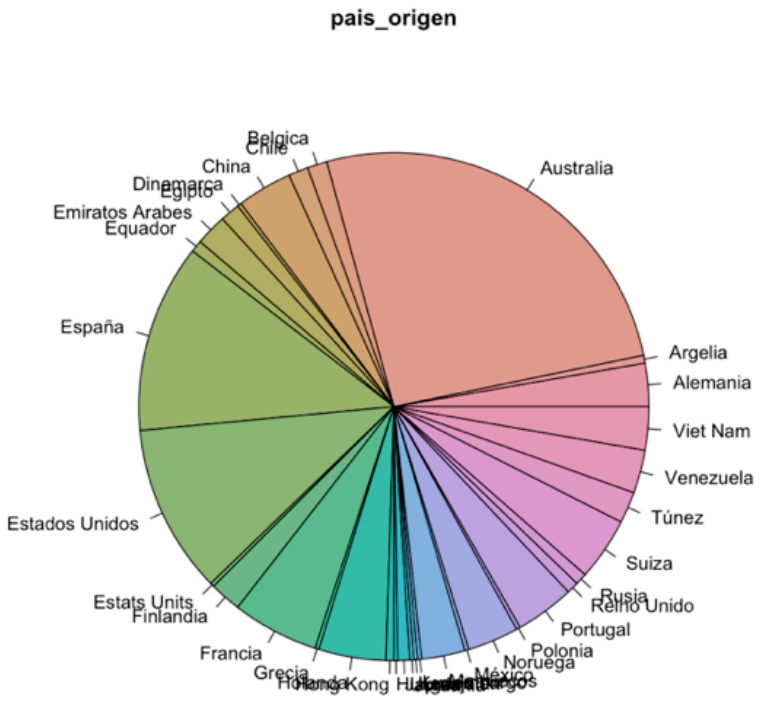
- **pais_origen**



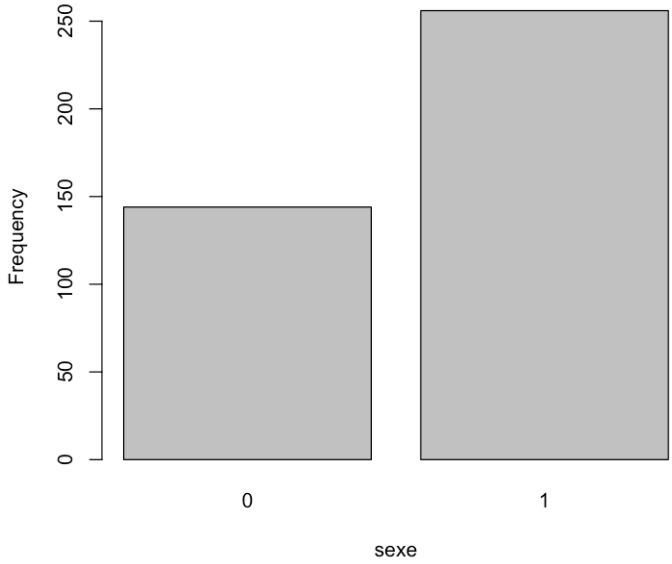
De la mateixa forma que amb els resultats de la campanya, hi ha un país que predomina sobre els altres. No obstant, succeeix un fet que caldrà analitzar, ja que aquest no correspon amb el país amb més nombre de candidats.

Així com en el fitxer de “campanyes_anteriors” el país predominant era “Espanya” tot seguit pels “Estats Units”, en aquest font de dades s’aprecia que és “Australia” el principal país d’on provenen els alumnes actuals.

```
        pais_origen
Australia     :104
Espa<U+00F1>a : 48
Estados Unidos: 43
Francia       : 22
Holanda       : 17
Suiza         : 16
```



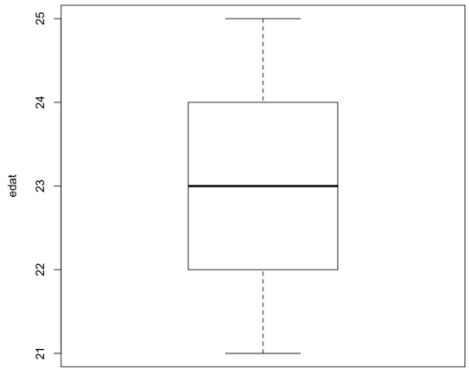
- sexe



El sexe predominant entre els estudiants actuals són els homes, amb un 64% de presència vers el 36% de dones.

```
      sexe
 Min.   :0.00
 1st Qu.:0.00
 Median :1.00
 Mean   :0.64
 3rd Qu.:1.00
 Max.   :1.00
```

- **edat**



La millor forma de representar aquest atribut és mitjançant una caixa de dispersió, on es poden observar de forma clara els màxims, els mínims i els quartils. Així doncs, s'oberva que la mitjana d'edat està sobre els 23 anys i, en conjunt, les edats presenten una distribució equilibrada. No s'aprecien valors *outliers* i defineix una forma molt simètrica.

```
     edat
Min.   :21.00
1st Qu.:22.00
Median :23.00
Mean   :23.11
3rd Qu.:24.00
Max.   :25.00
```

Dades de despesa per país

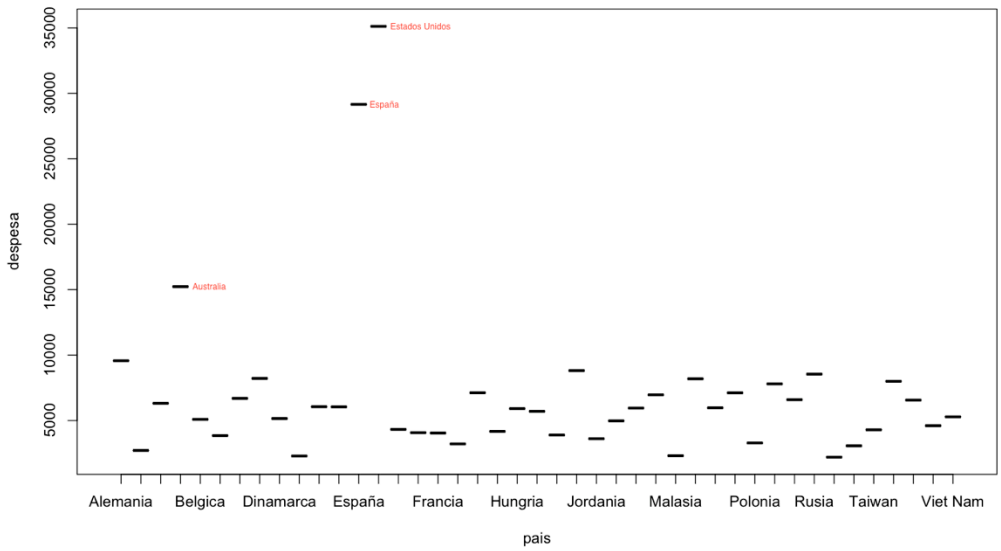
Es compon d'un total de **43 observacions** i **3 atributs**. Cadascun dels atributs presenta la següent distribució (estructura).

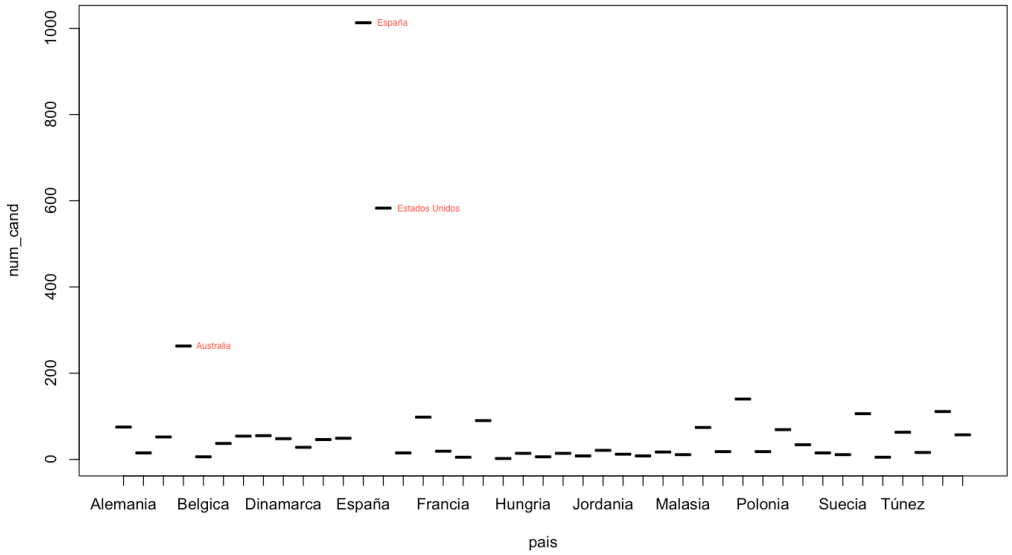
- pais: **43 nivells/valors diferents**, que corresponen directament als 43 països on s'han realitzat totes les campanyes de captació.
- despesa: camp **enter lliure** que indica la despesa en €.
- num_cand: camp **enter lliure** que indica el nombre de candidats total que han sigut captats en cada país.

<u>Anàlisi descriptiu</u>

- **pais**

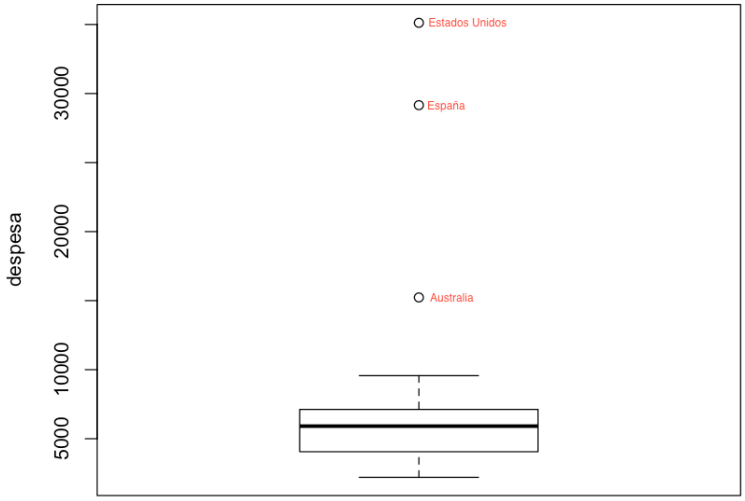
  Aquest atribut és de tipus únic, podent sent considerat com un ID de la taula. D'aquesta manera, els països seran descrits segons els atributs "despesa" i "num_cand".





Amb aquesta primera aproximació, es pot deduir que no és directament proporcional la despesa amb el nombre de candidats captats. Això té una explicació lògica i és que, per exemple, en el cas d'Espanya els desplaçaments són molt més econòmics que a la resta del món.

- **despesa**

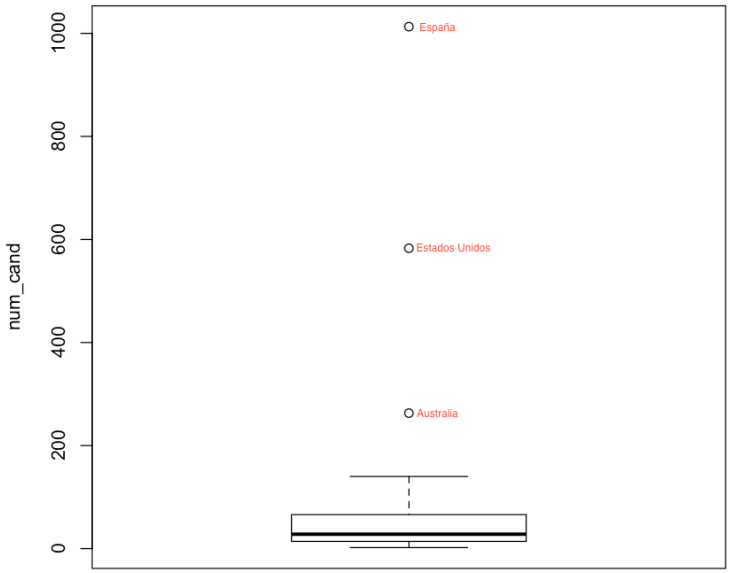


La informació que ens aporta la caixa de dispersió és que hi ha tres *outliers* (o valors que "trenquen" la tendència) entre tots els resultats. La mitjana està per sobre de la mediana, la despesa mínima és de 2202 € mentre que la màxima de 35123€. La dispersió d'aquests valors no presenta una estructura simètrica sinó que és asimètric cap a la dreta i caldrà analitzar amb compte els valors extrems.

Resum de l'atribut despesa

```
   despesa
Min.   : 2202
1st Qu.: 4058
Median : 5910
Mean   : 6931
3rd Qu.: 7120
Max.   :35123
```

- **num_cand**



En el nombre de candidats captats per país també existeixen 3 *outliers,* però en aquest cas els dos casos més extrems estan intercanviats amb els que apareixien a l'atribut despesa; mentre que en despesa el màxim corresponia als "Estats Units" en nombre de candidats aquest és "Espanya". La segona posició extrema com a màxim també té aquest comportament. El nombre màxim de candidats per país és de 1013 mentre que el mínim és de 2.

```
   num_cand
Min.   :   2.00
1st Qu.:  14.00
Median :  28.00
Mean   :  79.09
3rd Qu.:  66.00
Max.   :1013.00
```

2.3 Tercera fase – Generació dels models

2.3.1. Anàlisi de les campanyes anteriors. Algorismes associatiu i classificatori.

Algorisme associatiu. Apriori.

Per a les característiques pròpies d'aquesta font de dades es decideix emparar un algorisme associatiu; més concretament l'algorisme *apriori* de la llibreria *arules*. Aquest algorisme és especialment útil degut a que la majoria de dades són de tipus descriptiu i, per tant, no és possible definir patrons numèrics com es podria aconseguir mitjançant d'altres algorismes com els agregatius (ex: *kmeans*).

Gràcies a aquest mètode es poden generar models que descriguin una sèrie de regles que finalment esdevinguin en un objectiu concret. Tenint en compte que el factor d'èxit en tota campanya de captació d'estudiants és que aquests hagin estat finalment acceptats, es pretén definir la combinació d'atributs més freqüent que faci que un candidat sigui acceptat, no acceptat o abandoni.

La estructura que ofereix aquesta tècnica és del tipus X ➔Y, on X representa un conjunt d'esdeveniments que desencadenen de forma freqüent en un resultat Y.

De totes les proves realitzades, a continuació es mostren els resultats més aclaridors en quant a suport i confidència. Es defineixen tres casos interessants com a valor *rhs* de sortida, amb la finalitat de poder filtrar les regles resultats en "Acceptat", "No acceptat" i "Abandonament", que corresponen als valors [1,0,NA] respectivament. El nivell de confiança el situem al 0.8 per tal de tenir unes regles el més fiables possible i el suport entre 0,1 i 0,001 segons el conjunt de dades tractat.

El fet d'analitzar els casos que esdevenen en "No acceptat" o "Abandonament" no serveix per complir els objectius del projecte, però si per aplicar polítiques o estudiar amb profunditat què pot estar passant per tenir aquests resultats tant poc productius. D'aquests darrers es mostraran tant sols els resultats sobre el conjunt global de les dades, amb la finalitat de tenir una visió general sobre els fracassos.

Resultats amb el conjunt total de les dades

	lloc_contacte	ciutat	pais	campanya	medi_contacte	programa_interes	tipus_centre_origen	finalista	seguiment_pers	res_aptitudinal	res_llengua	res_calcul	res_assessement	acceptat
1	2_108	Lisboa	Portugal	2014-15	Visita	MscM	Privat	0	1	NA	NA	NA	NA	NA
2	1_114	Cincinnati	Estados Unidos	2015-16	Feria	MBA	Privat	0	1	NA	NA	NA	NA	NA
3	2_96	Rotterdam	Holanda	2015-16	Visita	MscM	Privat	1	1	Suspens	Suspens	Desitjable	1.000000	0.000000
4	2_06	Las Palmas	Espana	2014-15	Visita	MscM	Privat	0	1	NA	NA	NA	NA	NA
5	2_15	Guayaquil	Equador	2015-16	Visita	MscF	Concertat	0	1	NA	NA	NA	NA	NA
6	1_23	Barcelona	Espana	2013-14	Feria	MscF	Public	0	1	NA	NA	NA	NA	NA

*Captura dels 6 primers registres de la font de dades

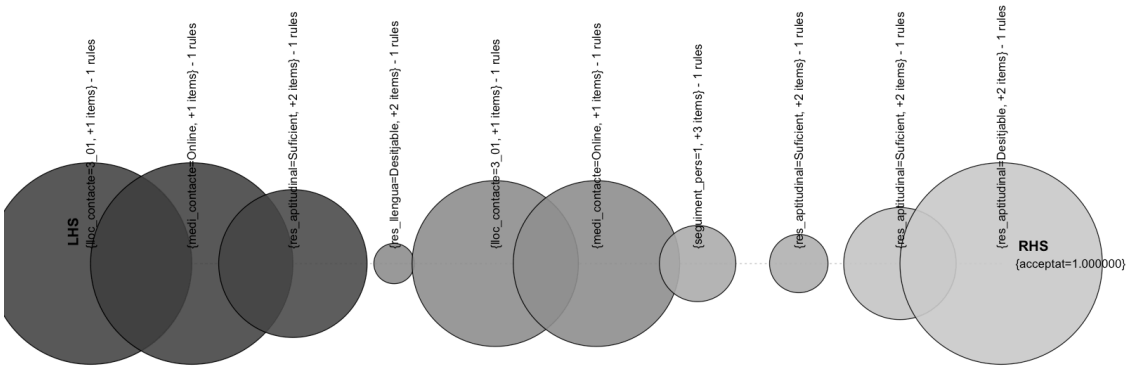
Esdevenen en candidat "Acceptat"

Resum del model generat

```
Apriori

Parameter specification:
 confidence minval smax arem  aval originalSupport support minlen maxlen target   ext
        0.8    0.1    1 none FALSE            TRUE    0.01      1     10  rules FALSE

Algorithmic control:
 filter tree heap memopt load sort verbose
    0.1 TRUE TRUE  FALSE TRUE    2    TRUE

Absolute minimum support count: 34

set item appearances ...[1 item(s)] done [0.00s].
set transactions ...[329 item(s), 3401 transaction(s)] done [0.00s].
sorting and recoding items ... [111 item(s)] done [0.00s].
creating transaction tree ... done [0.00s].
checking subsets of size 1 2 3 4 5 6 7 8 9 10 done [0.12s].
writing ... [35 rule(s)] done [0.01s].
creating S4 object  ... done [0.01s].
```

Regles resultants (15 primeres sense redundàncies i ordenades per confidència)

```
   lhs                                                                                                   rhs                 support    confidence lift
1  {lloc_contacte=3_01,res_assessement=1.000000}                                                       => {acceptat=1.000000} 0.01352543 0.9200000  7.612944
3  {medi_contacte=Online,res_assessement=1.000000}                                                     => {acceptat=1.000000} 0.01352543 0.9200000  7.612944
5  {lloc_contacte=3_01,medi_contacte=Online,res_assessement=1.000000}                                  => {acceptat=1.000000} 0.01352543 0.9200000  7.612944
7  {lloc_contacte=3_01,finalista=1,res_assessement=1.000000}                                           => {acceptat=1.000000} 0.01293737 0.9166667  7.585361
8  {medi_contacte=Online,finalista=1,res_assessement=1.000000}                                         => {acceptat=1.000000} 0.01293737 0.9166667  7.585361
21 {lloc_contacte=3_01,medi_contacte=Online,finalista=1,res_assessement=1.000000}                      => {acceptat=1.000000} 0.01293737 0.9166667  7.585361
18 {res_aptitudinal=Suficient,res_calcul=Excel·lent,res_assessement=1.000000}                          => {acceptat=1.000000} 0.01264334 0.9148936  7.570689
31 {finalista=1,res_aptitudinal=Suficient,res_calcul=Excel·lent,res_assessement=1.000000}              => {acceptat=1.000000} 0.01264334 0.9148936  7.570689
12 {res_llengua=Desitjable,res_calcul=Desitjable,res_assessement=1.000000}                             => {acceptat=1.000000} 0.01087915 0.8809524  7.289827
25 {finalista=1,res_llengua=Desitjable,res_calcul=Desitjable,res_assessement=1.000000}                 => {acceptat=1.000000} 0.01087915 0.8809524  7.289827
2  {lloc_contacte=3_01,finalista=1}                                                                    => {acceptat=1.000000} 0.01293737 0.8800000  7.281946
4  {medi_contacte=Online,finalista=1}                                                                  => {acceptat=1.000000} 0.01293737 0.8800000  7.281946
6  {lloc_contacte=3_01,medi_contacte=Online,finalista=1}                                               => {acceptat=1.000000} 0.01293737 0.8800000  7.281946
34 {seguiment_pers=1,res_aptitudinal=Excel·lent,res_llengua=Excel·lent,res_assessement=1.000000}       => {acceptat=1.000000} 0.01146722 0.8666667  7.171614
35 {finalista=1,seguiment_pers=1,res_aptitudinal=Excel·lent,res_llengua=Excel·lent,res_assessement=1.000000} => {acceptat=1.000000} 0.01146722 0.8666667  7.171614
```



*Matriu agrupada de les 10 primeres regles per ordre de confidència. Model resultant.

En aquest cas, es pot observar de forma clara que qualsevol zona geogràfica ha sigut omesa per l'algorisme associatiu. No obstant, si que hi ha un medi clarament diferenciador que esdevé en un candidat acceptat. Aquest és el medi "Online" o lloc de contacte "3_01".

La resta de paràmetres són conseqüència directe de ser acceptat. També es pot observar que el seguiment personalitzat és un atribut present en moltes regles i, per tant, és un bon indicador a l'hora d'establir estratègies de captació.

Esdevenen en candidat "No acceptat"

Resum del model generat

```
Apriori

Parameter specification:
 confidence minval smax arem  aval originalSupport support minlen maxlen target   ext
        0.8    0.1    1 none FALSE            TRUE    0.01      1     10  rules FALSE

Algorithmic control:
 filter tree heap memopt load sort verbose
    0.1 TRUE TRUE  FALSE TRUE    2    TRUE

Absolute minimum support count: 34

set item appearances ...[1 item(s)] done [0.00s].
set transactions ...[329 item(s), 3401 transaction(s)] done [0.00s].
sorting and recoding items ... [111 item(s)] done [0.00s].
creating transaction tree ... done [0.00s].
checking subsets of size 1 2 3 4 5 6 7 8 9 10 done [0.14s].
writing ... [8937 rule(s)] done [0.01s].
creating S4 object  ... done [0.01s].
```

En primera instància poden semblar moltes regles, però un cop eliminades les redundàncies esdevé un conjunt final de 442 regles.

```
set of 442 rules

rule length distribution (lhs + rhs):sizes
  2   3   4   5   6
  5  59 166 167  45

   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  2.000   4.000   4.000   4.425   5.000   6.000

summary of quality measures:
    support           confidence          lift
 Min.   :0.01029   Min.   :0.8000   Min.   :1.632
 1st Qu.:0.01264   1st Qu.:0.8355   1st Qu.:1.705
 Median :0.01691   Median :0.8696   Median :1.774
 Mean   :0.02977   Mean   :0.8764   Mean   :1.788
 3rd Qu.:0.02793   3rd Qu.:0.9211   3rd Qu.:1.879
 Max.   :0.49015   Max.   :1.0000   Max.   :2.040

mining info:
       data ntransactions support confidence
 camp_antOP          3401    0.01        0.8
```

Regles resultants (20 primeres sense redundàncies i ordenades per confidència)

```
     lhs                                                                                                  rhs                 support    confidence lift
1    {res_aptitudinal=Suspens}                                                                         => {acceptat=0.000000} 0.21464275 1.0000000  2.040192
2    {res_calcul=Suspens}                                                                              => {acceptat=0.000000} 0.22375772 1.0000000  2.040192
3    {res_llengua=Suspens}                                                                             => {acceptat=0.000000} 0.22581594 1.0000000  2.040192
4    {res_assessement=0.000000}                                                                        => {acceptat=0.000000} 0.27521317 1.0000000  2.040192
8011 {pais=Espana,campanya=2013-14,medi_contacte=Visita,finalista=1,seguiment_pers=0}                  => {acceptat=0.000000} 0.01146722 0.9750000  1.989187
5338 {pais=Espana,programa_interes=MscM,tipus_centre_origen=Privat,finalista=1}                        => {acceptat=0.000000} 0.01499559 0.9622642  1.963204
3501 {ciutat=Barcelona,campanya=2015-16,programa_interes=MscM,finalista=1}                             => {acceptat=0.000000} 0.01440753 0.9607843  1.960184
3266 {pais=Estados Unidos,tipus_centre_origen=Public,finalista=1,seguiment_pers=0}                     => {acceptat=0.000000} 0.01381946 0.9591837  1.956919
8196 {pais=Espana,campanya=2015-16,medi_contacte=Feria,programa_interes=MscM,finalista=1}              => {acceptat=0.000000} 0.01352543 0.9583333  1.955184
3506 {ciutat=Barcelona,campanya=2015-16,tipus_centre_origen=Concertat,finalista=1}                     => {acceptat=0.000000} 0.01293737 0.9565217  1.951488
5342 {pais=Espana,campanya=2015-16,programa_interes=MscM,finalista=1}                                  => {acceptat=0.000000} 0.01881799 0.9552239  1.948840
2629 {lloc_contacte=1_23,campanya=2015-16,programa_interes=MscM,finalista=1}                           => {acceptat=0.000000} 0.01234931 0.9545455  1.947456
8267 {medi_contacte=Feria,programa_interes=MscF,tipus_centre_origen=Concertat,finalista=1,seguiment_pers=1} => {acceptat=0.000000} 0.01852396 0.9545455  1.947456
1850 {ciutat=Connecticut,campanya=2014-15,medi_contacte=Visita,finalista=1}                            => {acceptat=0.000000} 0.01176125 0.9523810  1.943040
2001 {ciutat=Cincinnati,medi_contacte=Feria,programa_interes=MscF,finalista=1}                         => {acceptat=0.000000} 0.01176125 0.9523810  1.943040
3537 {ciutat=Barcelona,programa_interes=MBA,finalista=1,seguiment_pers=0}                              => {acceptat=0.000000} 0.01176125 0.9523810  1.943040
3545 {ciutat=Barcelona,campanya=2015-16,finalista=1,seguiment_pers=0}                                  => {acceptat=0.000000} 0.01764187 0.9523810  1.943040
5643 {ciutat=Cincinnati,campanya=2015-16,medi_contacte=Feria,finalista=1,seguiment_pers=1}             => {acceptat=0.000000} 0.01176125 0.9523810  1.943040
6426 {pais=Estados Unidos,campanya=2015-16,medi_contacte=Feria,finalista=1,seguiment_pers=1}           => {acceptat=0.000000} 0.01176125 0.9523810  1.943040
3259 {pais=Estados Unidos,medi_contacte=Feria,programa_interes=MscF,finalista=1}                       => {acceptat=0.000000} 0.01734784 0.9516129  1.941473
5353 {pais=Espana,campanya=2015-16,tipus_centre_origen=Concertat,finalista=1}                          => {acceptat=0.000000} 0.01705381 0.9508197  1.939855
8    {lloc_contacte=1_113,finalista=1}                                                                 => {acceptat=0.000000} 0.01087915 0.9487179  1.935567
2000 {ciutat=Cincinnati,campanya=2015-16,programa_interes=MscF,finalista=1}                            => {acceptat=0.000000} 0.01087915 0.9487179  1.935567
```

Amb les regles generades que esdevenen en no acceptació, clarament destaquen destinacions d'Espanya o dels Estats Units. Això és conseqüència directe de que són els països amb major nombre de candidats però una petita proporció d'aquests resulten acceptats.

Esdevenen en candidat "Abandonament"

Resum del model generat

```
Apriori

Parameter specification:
 confidence minval smax arem  aval originalSupport support minlen maxlen target   ext
        0.8    0.1    1 none FALSE            TRUE     0.1      1     10  rules FALSE

Algorithmic control:
 filter tree heap memopt load sort verbose
    0.1 TRUE TRUE  FALSE TRUE    2    TRUE

Absolute minimum support count: 340

set item appearances ...[1 item(s)] done [0.00s].
set transactions ...[329 item(s), 3401 transaction(s)] done [0.00s].
sorting and recoding items ... [40 item(s)] done [0.00s].
creating transaction tree ... done [0.00s].
checking subsets of size 1 2 3 4 5 6 7 8 done [0.01s].
writing ... [744 rule(s)] done [0.00s].
creating S4 object  ... done [0.00s].
```

Un cop eliminades les regles redundants esdevé un conjunt final d'11 regles.

```
set of 11 rules

rule length distribution (lhs + rhs):sizes
2 3
5 6

   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  2.000   2.000   3.000   2.545   3.000   3.000

summary of quality measures:
    support         confidence          lift
 Min.   :0.1217   Min.   :0.9985   Min.   :2.567
 1st Qu.:0.1373   1st Qu.:1.0000   1st Qu.:2.571
 Median :0.2155   Median :1.0000   Median :2.571
 Mean   :0.2591   Mean   :0.9999   Mean   :2.570
 3rd Qu.:0.3887   3rd Qu.:1.0000   3rd Qu.:2.571
 Max.   :0.3890   Max.   :1.0000   Max.   :2.571

mining info:
       data ntransactions support confidence
 camp_antOP          3401     0.1        0.8
```

Regles resultants

```
   lhs                                          rhs           support   confidence lift
1  {res_calcul=NA}                           => {acceptat=NA} 0.3887092 1.0000000  2.570673
2  {res_llengua=NA}                          => {acceptat=NA} 0.3887092 1.0000000  2.570673
3  {res_aptitudinal=NA}                      => {acceptat=NA} 0.3887092 1.0000000  2.570673
4  {res_assessement=NA}                      => {acceptat=NA} 0.3890032 1.0000000  2.570673
10 {pais=Espana,finalista=0}                 => {acceptat=NA} 0.1496619 1.0000000  2.570673
15 {programa_interes=MscM,finalista=0}       => {acceptat=NA} 0.1234931 1.0000000  2.570673
20 {tipus_centre_origen=Concertat,finalista=0} => {acceptat=NA} 0.1249632 1.0000000  2.570673
50 {campanya=2015-16,finalista=0}            => {acceptat=NA} 0.1217289 1.0000000  2.570673
73 {medi_contacte=Visita,finalista=0}        => {acceptat=NA} 0.1702440 1.0000000  2.570673
74 {medi_contacte=Feria,finalista=0}         => {acceptat=NA} 0.2155248 1.0000000  2.570673
5  {finalista=0}                             => {acceptat=NA} 0.3890032 0.9984906  2.566792
```

D'aquest darrer model tant sols se'n pot extreure, geogràficament parlant, que si el candidat prové d'Espanya probablement acabi abandonant. D'aquesta forma sabem que els candidats Espanyols acostumen o bé a abandonar o bé a no superar les proves.

Resultats del subconjunt amb major suport i confidència

En aquesta ocasió es seleccionen tant sols els atributs lloc_contacte, campanya, medi_contacte, tipus_centre_origen i acceptat. Es considera que el fet d'aprovar les tres proves d'accés, *l'assesement* i ser finalista és causa directa de ser acceptat i per tant no poden formar part de cap regla predictiva.

També s'han exclòs el país i la ciutat per estar definits implícitament pel lloc de contacte, reduint així possibles distorsions al resultat.

	lloc_contacte	campanya	medi_contacte	tipus_centre_origen	acceptat
1	2_108	2014-15	Visita	Privat	NA
2	1_114	2015-16	Feria	Privat	NA
3	2_96	2015-16	Visita	Privat	0.000000
4	2_06	2014-15	Visita	Privat	NA
5	2_15	2015-16	Visita	Concertat	NA
6	1_23	2013-14	Feria	Public	NA
7	2_49	2013-14	Visita	Public	NA
8	1_16	2013-14	Feria	Public	NA
9	2_43	2014-15	Visita	Public	NA
10	2_132	2014-15	Visita	Privat	0.000000
11	1_111	2014-15	Feria	Concertat	0.000000

*Captura dels 11 primers registres del subconjunt de dades

Com a *rhs* es defineix que es volen conèixer aquelles regles que esdevenen en candidat acceptat. Per tal que el model apriori pugui trobar regles vàlides, s'ha tingut que ajustar el nivell de suport a 0,001.

Resum del model generat

```
set of 18 rules

rule length distribution (lhs + rhs):sizes
3 4 5
5 9 4

   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  3.000   3.250   4.000   3.944   4.000   5.000

summary of quality measures:
    support           confidence          lift
 Min.   :0.001176   Min.   :0.8000   Min.   :6.620
 1st Qu.:0.001176   1st Qu.:0.8125   1st Qu.:6.723
 Median :0.001470   Median :0.8333   Median :6.896
 Mean   :0.002058   Mean   :0.8709   Mean   :7.207
 3rd Qu.:0.003234   3rd Qu.:0.9615   3rd Qu.:7.957
 Max.   :0.003822   Max.   :1.0000   Max.   :8.275

mining info:
  data ntransactions support confidence
 XopBo          3401   0.001        0.8
```

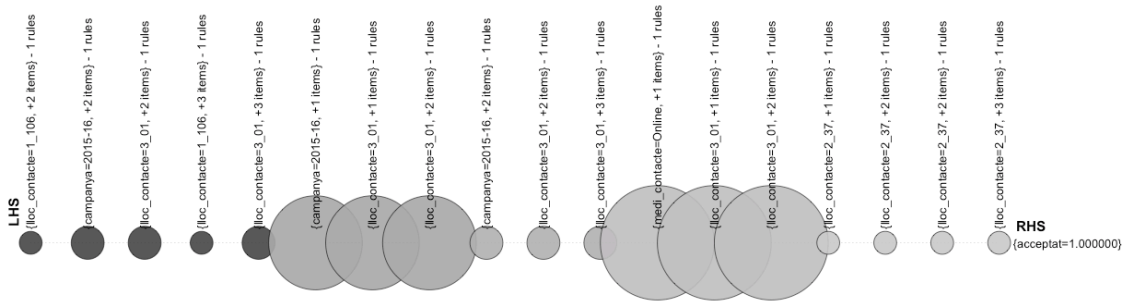
Regles resultants

	lhs		rhs	support	confidence	lift
8	{lloc_contacte=1_106,campanya=2014-15,tipus_centre_origen=Privat}	=>	{acceptat=1.000000}	0.001176125	1.0000000	8.274939
11	{campanya=2015-16,medi_contacte=Online,tipus_centre_origen=Concertat}	=>	{acceptat=1.000000}	0.001470156	1.0000000	8.274939
13	{lloc_contacte=3_01,campanya=2015-16,tipus_centre_origen=Concertat}	=>	{acceptat=1.000000}	0.001470156	1.0000000	8.274939
16	{lloc_contacte=1_106,campanya=2014-15,medi_contacte=Feria,tipus_centre_origen=Privat}	=>	{acceptat=1.000000}	0.001176125	1.0000000	8.274939
17	{lloc_contacte=3_01,campanya=2015-16,medi_contacte=Online,tipus_centre_origen=Concertat}	=>	{acceptat=1.000000}	0.001470156	1.0000000	8.274939
3	{campanya=2015-16,medi_contacte=Online}	=>	{acceptat=1.000000}	0.003234343	0.8461538	7.001872
5	{lloc_contacte=3_01,campanya=2015-16}	=>	{acceptat=1.000000}	0.003234343	0.8461538	7.001872
10	{lloc_contacte=3_01,campanya=2015-16,medi_contacte=Online}	=>	{acceptat=1.000000}	0.003234343	0.8461538	7.001872
12	{campanya=2015-16,medi_contacte=Online,tipus_centre_origen=Privat}	=>	{acceptat=1.000000}	0.001470156	0.8333333	6.895783
14	{lloc_contacte=3_01,campanya=2015-16,tipus_centre_origen=Privat}	=>	{acceptat=1.000000}	0.001470156	0.8333333	6.895783
18	{lloc_contacte=3_01,campanya=2015-16,medi_contacte=Online,tipus_centre_origen=Privat}	=>	{acceptat=1.000000}	0.001470156	0.8333333	6.895783
2	{medi_contacte=Online,tipus_centre_origen=Concertat}	=>	{acceptat=1.000000}	0.003822405	0.8125000	6.723388
4	{lloc_contacte=3_01,tipus_centre_origen=Concertat}	=>	{acceptat=1.000000}	0.003822405	0.8125000	6.723388
9	{lloc_contacte=3_01,medi_contacte=Online,tipus_centre_origen=Concertat}	=>	{acceptat=1.000000}	0.003822405	0.8125000	6.723388
1	{lloc_contacte=2_37,tipus_centre_origen=Privat}	=>	{acceptat=1.000000}	0.001176125	0.8000000	6.619951
6	{lloc_contacte=2_37,campanya=2015-16,tipus_centre_origen=Privat}	=>	{acceptat=1.000000}	0.001176125	0.8000000	6.619951
7	{lloc_contacte=2_37,medi_contacte=Visita,tipus_centre_origen=Privat}	=>	{acceptat=1.000000}	0.001176125	0.8000000	6.619951
15	{lloc_contacte=2_37,campanya=2015-16,medi_contacte=Visita,tipus_centre_origen=Privat}	=>	{acceptat=1.000000}	0.001176125	0.8000000	6.619951

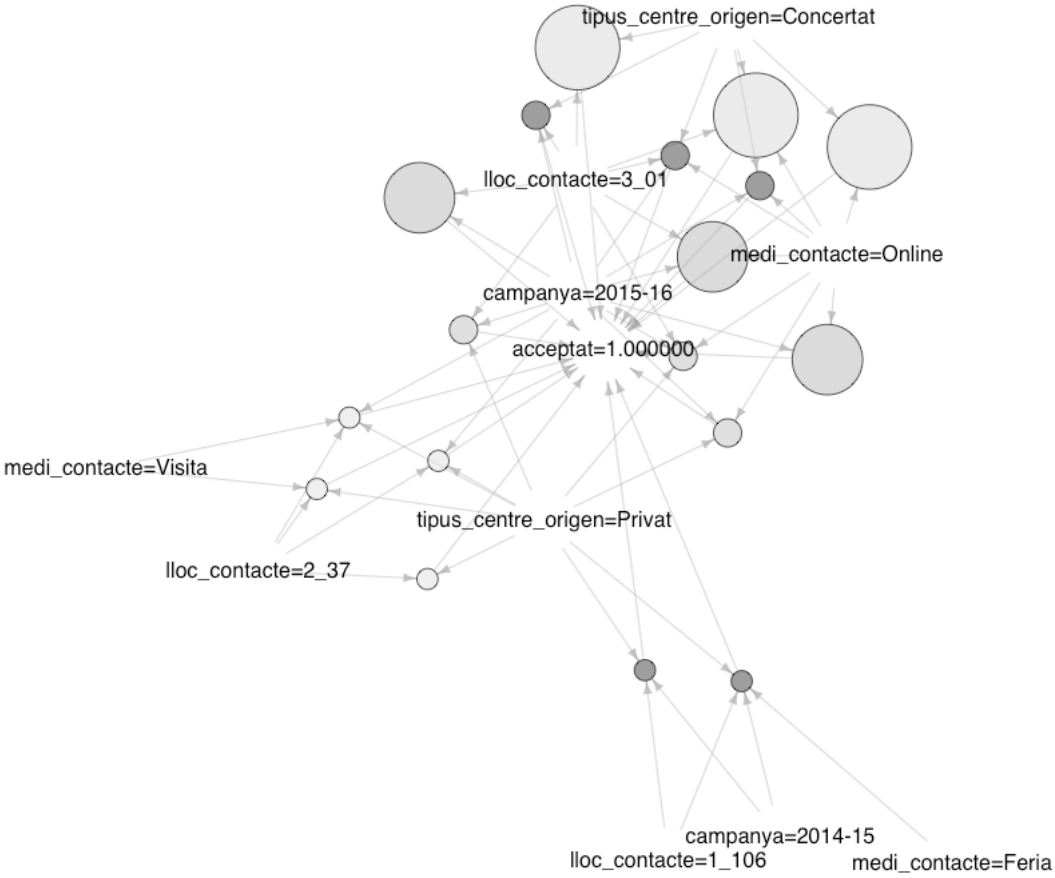
Per una banda, aquests nous resultats reforcen l'argument del model anterior sobre el fet que un candidat que hagi contactat Online, o el que és el mateix al lloc de contacte 3_01, probablement sigui acceptat.

Per altra banda, entren en joc dos nous llocs de contacte en dues campanyes diferents. Aquests són el 1_106 que correspon a una Fira de Santiago de Chile (Chile) i el 2_37 de Melbourne (Australia).

A més a més, aquest nou conjunt de regles relega a ultima posició els estudiants provinents de centres públics, sent els concertats i els privats els millors candidats.



*Matriu agrupada de les regles ordenades per confidència del model resultant.



*Graf de les regles del model resultant.

Algorisme classificatori. Naive bayes.

Es decideix procedir a l'execució d'un model classificatori degut a que no s'han pogut obtenir regions geogràfiques clares amb el model associatiu anterior.

Inicialment, sobre aquest conjunt de dades també s'ha aplicat un model classificatori mitjançant arbres decisionals, però degut a l'alta taxa d'error en tots els resultats (superior al 70%) se n'ha descartat qualsevol tipus d'interpretació vàlida.

Així doncs, s'ha optat per provar amb un altre model classificatori tot generant una xarxa bayesiana. S'han carregat els paquets d' R corresponents indicats en la preparació de l'entorn de treball i s'ha procedit a l'execució sobre el conjunt total de les dades.

D'igual forma que amb els models anteriors, s'executa l'algorisme classificatori per a cadascun dels atributs prenent com a criteri el valor de l'atribut "acceptat".

Aquest algorisme és considerat un algorisme bayesià "ingenu", on tracta cadascun dels atributs independentment fins a trobar classificacions

vàlides. La taula completa resultant de l'algorisme de Bayes mostra per a cada atribut la probabilitat de que un nou registre sigui classificat en aquell camp.

Els resultats finals tampoc són gaire clarificadors pel que fa a la definició de zones geogràfiques, però serviran per avaluar-los conjuntament amb el model associatiu en l'apartat d'interpretació dels resultats i poder així extreure'n conclusions subtils.

Dels resultats d'aquest algorisme, s'agafaran per analitzar les classificacions generades pels atributs "pais", "medi_contacte" i "seguiment_pers".

El lloc de contacte no ha permès establir cap aproximació vàlida a part de tenir un grau de fiabilitat mol reduït. A continuació es mostra una captura de la sortida pel lloc de contacte.

```
> classifier$tables
$lloc_contacte
          lloc_contacte
X2[, 6]           1_01        1_02        1_03        1_04        1_05        1_06        1_07        1_08        1_09        1_10        1_100
  0.000000 0.0017878427 0.0005959476 0.0011918951 0.0065554231 0.0017878427 0.0005959476 0.0017878427 0.0035756853 0.0148986889 0.0125148987 0.0017878427
  1.000000 0.0000000000 0.0026455026 0.0000000000 0.0079365079 0.0000000000 0.0000000000 0.0000000000 0.0026455026 0.0185185185 0.0714285714 0.0026455026
  NA       0.0000000000 0.0000000000 0.0014869888 0.0037174721 0.0022304833 0.0014869888 0.0022304833 0.0074349442 0.0230483271 0.0000000000 0.0000000000
          lloc_contacte
X2[, 6]          1_101       1_102       1_103       1_104       1_105       1_106       1_107       1_108       1_109        1_11        1_111
  0.000000 0.0208581645 0.0047675805 0.0035756853 0.0041716329 0.0023837902 0.0083432658 0.0000000000 0.0000000000 0.0053635280 0.0035756853 0.0178784267
  1.000000 0.0132275132 0.0079365079 0.0079365079 0.0052910053 0.0000000000 0.0132275132 0.0000000000 0.0052910053 0.0052910053 0.0105820106 0.0105820106
  NA       0.0141263941 0.0096654275 0.0052044610 0.0000000000 0.0000000000 0.0000000000 0.0014869888 0.0014869888 0.0000000000 0.0000000000 0.0148698885
          lloc_contacte
X2[, 6]          1_112       1_113       1_114       1_115        1_12       1_122        1_15        1_16        1_17        1_18        1_19
  0.000000 0.0143027414 0.0220500596 0.0363528010 0.0107270560 0.0005959476 0.0214541120 0.0071513707 0.0095351609 0.0041716329 0.0000000000 0.0035756853
  1.000000 0.0052910053 0.0052910053 0.0105820106 0.0026455026 0.0000000000 0.0423280423 0.0079365079 0.0132275132 0.0026455026 0.0000000000 0.0000000000
  NA       0.0208178439 0.0304832714 0.0349442379 0.0141263941 0.0007434944 0.0000000000 0.0118959108 0.0245353160 0.0029739777 0.0014869888 0.0037174721
          lloc_contacte
X2[, 6]          1_20       1_200        1_21        1_22        1_23        1_44        1_45        1_46        1_47        1_48        1_54
```

D'aquesta forma, es pot conèixer filtrant les probabilitats on és més probable que un nou candidat sigui captat.

- **pais**
  <u>Probabilitat classificatòria</u>

```
          pais
Y            Alemania      Andorra      Argelia    Australia      Belgica        Chile        China     Colombia    Dinamarca       Egipto Emiratos Arabes      Equador       Espana Estados Unidos    Filipinas
  0.000000 0.0329934013 0.0059988002 0.0215956809 0.1067786443 0.0023995201 0.0143971206 0.0257948410 0.0191961608 0.0125974805 0.0089982004    0.0197960408 0.0143971206 0.2735452909   0.1751649670 0.0041991602
  1.000000 0.0486618005 0.0121654501 0.0389294404 0.2068126521 0.0048661800 0.0316301703 0.0267639903 0.0559610706 0.0656934307 0.0316301703    0.0267639903 0.0072992701 0.1167883212   0.0705596107 0.0000000000
  NA       0.0000000000 0.0000000000 0.0000000000 0.0000000000 0.0000000000 0.0000000000 0.0000000000 0.0000000000 0.0000000000 0.0000000000    0.0015117158 0.0166288738 0.3847316704   0.1980347695 0.0060468632
          pais
Y           Finlandia      Francia       Grecia      Holanda    Hong Kong      Hungria        India      Irlanda        Japon     Jordania    Kazajstan       Kuwait   Luxemburgo      Malasia     Marruecos       Mexico
  0.000000 0.0269946011 0.0041991602 0.0011997600 0.0161967606 0.0000000000 0.0041991602 0.0017996401 0.0059988002 0.0023995201 0.0059988002 0.0041991602 0.0011997600 0.0035992801 0.0035992801 0.0203959208 0.0041991602
  1.000000 0.0194647202 0.0000000000 0.0000000000 0.0632603406 0.0000000000 0.0024330900 0.0000000000 0.0000000000 0.0000000000 0.0024330900 0.0024330900 0.0121654501 0.0024330900 0.0000000000 0.0097323601 0.0024330900
  NA       0.0340136054 0.0090702948 0.0022675737 0.0279667423 0.0015117158 0.0045351474 0.0022675737 0.0030234316 0.0030234316 0.0075585790 0.0030234316 0.0007558579 0.0075585790 0.0037792895 0.0272108844 0.0075585790
          pais
Y           Noruega      Polonia     Portugal  Reino Unido       Rusia      Suecia       Suiza      Taiwan       Tunez      Turquia    Venezuela     Viet Nam
  0.000000 0.0413917217 0.0029994001 0.0143971206 0.0071985603 0.0035992801 0.0035992801 0.0269946011 0.0005998800 0.0149970006 0.0053989202 0.0209958008 0.0137972406
  1.000000 0.0121654501 0.0072992701 0.0243309002 0.0048661800 0.0048661800 0.0000000000 0.0170316302 0.0000000000 0.0170316302 0.0000000000 0.0243309002 0.0267639903
  NA       0.0498866213 0.0075585790 0.0264550265 0.0151171580 0.0052910053 0.0037792895 0.0408163265 0.0030234316 0.0234315949 0.0052910053 0.0498866213 0.0173847317
```

<u>Nombre de casos classificats segons probabilitat. Model aleatori.</u>

```
          Alemania Andorra Argelia Australia Belgica Chile China Colombia Dinamarca Egipto Emiratos Arabes Equador Espana Estados Unidos Filipinas Finlandia
 0.000000       24       0       0         0       0     0    47        6         0      0              36      23    114           278         9        10
 1.000000       51      15      52       263       6    37     7       49        48     28              10       0      0             0         0         0
 NA              0       0       0         0       0     0     0        0         0      0               0      26    899           305         6        88

          Francia Grecia Holanda Hong Kong Hungria India Irlanda Japon Jordania Kazajstan Kuwait Luxemburgo Malasia Marruecos Mexico Noruega Polonia Portugal
 0.000000       0      0       0         0       2     0      14     0       10        12      0          0      11        0      0       2       0       0
 1.000000       0      0       0         0       0     0       0     0        0         0      0          0       0        0      0       0       2       0
 NA            19      5      90         2      12     6       0     8       11         0      8         17       0       74     18     138      16      69

          Reino Unido Rusia Suecia Suiza Taiwan Tunez Turquia Venezuela Viet Nam
 0.000000           0     6      7    49      0     0      16         0       19
 1.000000           0     0      0     0      0     0       0         0        8
 NA                34     9      4    57      5    63       0       111       30
```

- **medi_contacte**

Probabilitat classificatòria

```
          medi_contacte
Y               Feria      Online      Visita
  0.000000 0.513497301 0.003599280 0.482903419
  1.000000 0.425790754 0.111922141 0.462287105
  NA       0.554043840 0.008314437 0.437641723
```

Nombre de casos classificats segons probabilitat. Model aleatori.

```
           Feria Online Visita
  0.000000   351      0    344
  1.000000   204     13    359
  NA        1209     50    871
```

- **seguiment_pers** (0 = no, 1 = si)
  Probabilitat classificatòria

```
          seguiment_pers
Y                    0          1
  0.000000 0.47930414 0.52069586
  1.000000 0.47688564 0.52311436
  NA       0.93046107 0.06953893
```

Nombre de casos classificats segons probabilitat. Model aleatori.

```
              0    1
  0.000000  430  265
  1.000000  273  303
  NA       1527  603
```

De les lectures anteriors, es torna a interpretar que les dades font no amaguen cap patró clar pel que fa a la definició de zones geogràfiques concretes i, si les defineixen, aquestes tenen un nivell de suport molt reduït.

Com a conclusió es poden definir les següents lectures un cop aplicat el model classificatori diverses vegades

- Els països amb major probabilitat de captar alumnes i que aquests siguin acceptats són: Australia, Argèlia, Xile, Colòmbia i Dinamarca.

- Els països amb menor probabilitat de ser acceptats son: España, Estats Units, Irlanda i Suïssa. La interpretació d'aquest resultat és que aquests països tenen un gran nombre de candidats, però en canvi molts d'ells o abandonen o no passen les proves.

- Els medis de contacte Visita i Fira es troben en igualtat de condicions mentre que el medi online presenta més acceptats que no acceptats.

2.3.2. Anàlisi dels alumnes actuals. Algorismes associatiu i classificatori.

Per realitzar l'anàlisi d'aquest grup s'ha pres la decisió d'aplicar dos tipologies de models de mineria de dades, de la mateixa forma que al conjunt de dades de les campanyes anteriors.

En aquest cas s'ha utilitzat un model associatiu *apriori* i un model classificatori mitjançant arbres decisionals. Ambdós models permeten ser interpretats amb una fiabilitat acceptable, sent l'associativa la menys decisiva. Es presenten els dos models degut a que els resultats del model classificatori podran ser reafirmats pels resultats del model associatiu.

L'objectiu principal de l'anàlisi d'aquesta font de dades és la de trobar les zones geogràfiques més usuals dels estudiants actuals i definir un perfil d'aquests. En concret, s'establirà com a atribut de sortida el fet de que un estudiant sigui provinent de campanyes anteriors. D'aquesta forma, s'espera poder conèixer si l'esforç en certes regions ha sigut satisfactori així com les possibles zones potencials on seria interessant iniciar campanyes de captació.

Algorisme associatiu. Apriori.

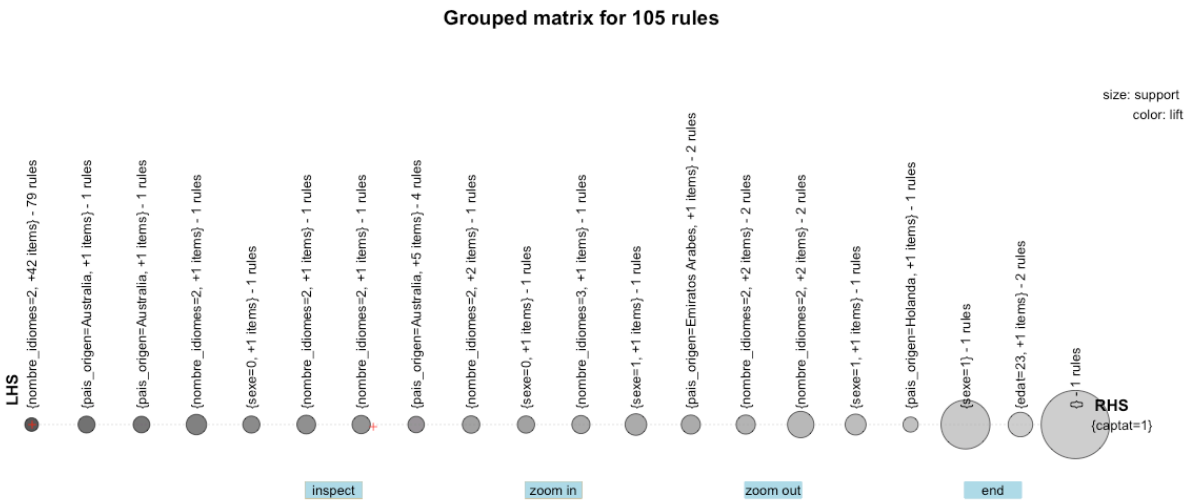
En aquest cas, ens interessa saber si els estudiants actuals provenen d'una campanya anterior (rhs = "captat = 1") i si responen a algun perfil concret (edat, país, nombre idiomes o sexe). Com el nombre d'atributs és prou reduït es selecciona el conjunt total de les dades d'entrada.

Tot seguit es mostren els resultats de les regles d'associació resultants en cadascun dels casos.

- **Captats**. De 667 regles inicials n'esdevenen 105 regles. Ordenades de major a menor confidència se'n mostren les 20 primeres. Aquestes són ordenades arbitràriament ja que fins la regla 50 totes tenen la mateixa confidència i, conseqüentment, el mateix pes. Caldrà avaluar doncs de totes aquetes quines segueixen un patró més clar en quant a regió geogràfica.

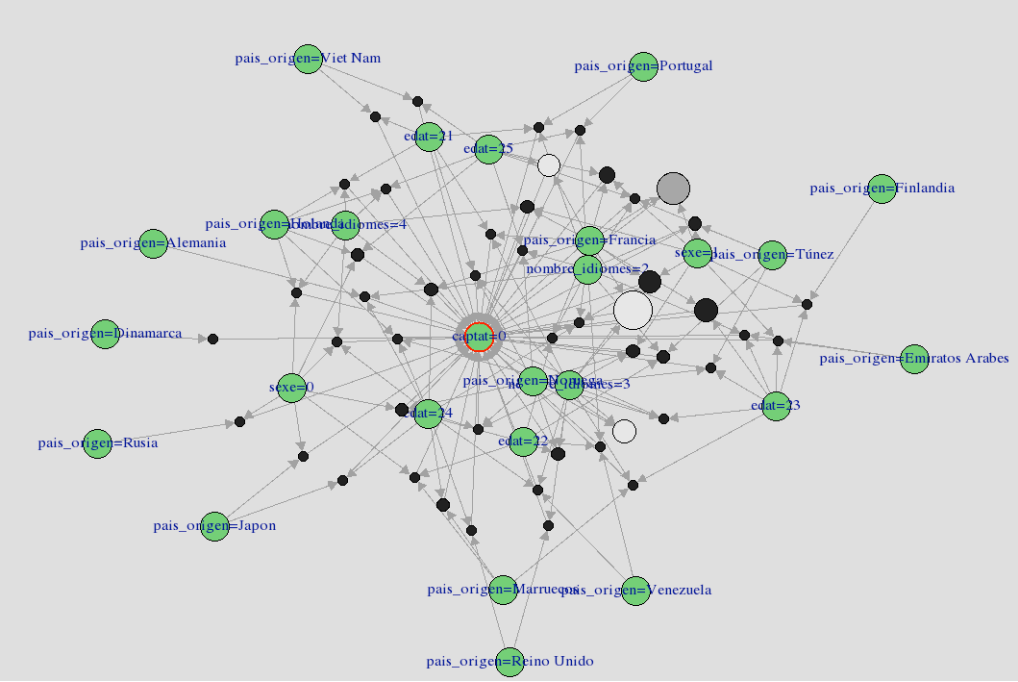
```
   lhs                                      rhs        support confidence lift
2  {pais_origen=Estats Units}           => {captat=1} 0.0025  1          1.204819
3  {pais_origen=Hungria}                => {captat=1} 0.0025  1          1.204819
4  {pais_origen=México}                 => {captat=1} 0.0025  1          1.204819
5  {pais_origen=Polonia}                => {captat=1} 0.0025  1          1.204819
6  {pais_origen=Kazajstan}              => {captat=1} 0.0025  1          1.204819
7  {pais_origen=Jordania}               => {captat=1} 0.0025  1          1.204819
8  {pais_origen=Luxemburgo}             => {captat=1} 0.0025  1          1.204819
9  {pais_origen=Grecia}                 => {captat=1} 0.0025  1          1.204819
10 {pais_origen=Argelia}                => {captat=1} 0.0050  1          1.204819
11 {pais_origen=Hong Kong}              => {captat=1} 0.0050  1          1.204819
12 {pais_origen=Equador}                => {captat=1} 0.0075  1          1.204819
13 {pais_origen=Chile}                  => {captat=1} 0.0125  1          1.204819
14 {pais_origen=Egipto}                 => {captat=1} 0.0125  1          1.204819
15 {pais_origen=Belgica}                => {captat=1} 0.0125  1          1.204819
19 {pais_origen=China}                  => {captat=1} 0.0350  1          1.204819
21 {pais_origen=Suiza}                  => {captat=1} 0.0400  1          1.204819
22 {pais_origen=Estados Unidos}         => {captat=1} 0.1075  1          1.204819
23 {pais_origen=España}                 => {captat=1} 0.1200  1          1.204819
73 {pais_origen=Rusia,edat=24}          => {captat=1} 0.0025  1          1.204819
74 {nombre_idiomes=3,pais_origen=Rusia} => {captat=1} 0.0025  1          1.204819
```

El següent gràfic permet fer zooms i inspeccionar les regles, esdevenint una eina útil en quant a la interpretació dels resultats de l'algorisme posterior.



- **No captats**. En aquest cas, totes les regles menys les 5 últimes tenen el mateix grau de confidència.

	lhs		rhs	support	confidence	lift
1	{pais_origen=Dinamarca}	=>	{captat=0}	0.0025	1	5.882353
5	{pais_origen=Reino Unido,edat=24}	=>	{captat=0}	0.0025	1	5.882353
6	{nombre_idiomes=3,pais_origen=Reino Unido}	=>	{captat=0}	0.0025	1	5.882353
7	{pais_origen=Rusia,sexe=0}	=>	{captat=0}	0.0025	1	5.882353
8	{pais_origen=Japon,edat=24}	=>	{captat=0}	0.0025	1	5.882353
9	{pais_origen=Japon,sexe=0}	=>	{captat=0}	0.0025	1	5.882353
10	{pais_origen=Marruecos,edat=24}	=>	{captat=0}	0.0050	1	5.882353
11	{pais_origen=Viet Nam,edat=21}	=>	{captat=0}	0.0025	1	5.882353
12	{pais_origen=Viet Nam,edat=25}	=>	{captat=0}	0.0025	1	5.882353
14	{pais_origen=Francia,edat=23}	=>	{captat=0}	0.0125	1	5.882353
30	{nombre_idiomes=2,pais_origen=Emiratos Arabes,edat=23}	=>	{captat=0}	0.0025	1	5.882353
31	{pais_origen=Emiratos Arabes,sexe=1,edat=23}	=>	{captat=0}	0.0025	1	5.882353
32	{pais_origen=Finlandia,sexe=1,edat=23}	=>	{captat=0}	0.0025	1	5.882353
33	{nombre_idiomes=2,pais_origen=Túnez,edat=22}	=>	{captat=0}	0.0050	1	5.882353
34	{nombre_idiomes=3,pais_origen=Túnez,edat=23}	=>	{captat=0}	0.0025	1	5.882353
35	{nombre_idiomes=2,pais_origen=Túnez,edat=25}	=>	{captat=0}	0.0025	1	5.882353
36	{nombre_idiomes=2,pais_origen=Túnez,sexe=1}	=>	{captat=0}	0.0050	1	5.882353
37	{pais_origen=Marruecos,sexe=0,edat=22}	=>	{captat=0}	0.0025	1	5.882353
40	{nombre_idiomes=3,pais_origen=Marruecos,edat=23}	=>	{captat=0}	0.0025	1	5.882353
45	{nombre_idiomes=2,pais_origen=Alemania,edat=24}	=>	{captat=0}	0.0025	1	5.882353



Caldrà analitzar amb cura aquests resultats, ja que en primera instància semblen molt difusos.

Algorisme classificatori. Arbre decisional C50.

D'igual forma que amb l'algorisme associatiu, es pren com a partida que es vol conèixer grups classificatoris a partir de si provenen de campanyes anteriors o no. Amb l'execució d'aquest canvia la perspectiva del model obtingut, on l'objectiu no és la creació de regles predictives sinó la definició de patrons d'agrupació.

Per dur a terme aquesta tasca, s'executa l'algorisme C50 i dels resultats se n'espera poder trobar aquells atributs que siguin potencialment capaços de discriminar diferents particions de les dades originals. Idòniament, l'arbre resultant més interessant serà aquell que tingui

l'atribut "captat" com a node arrel, fet que indicarà que realitzarà la partició més discriminatòria. Faltarà però que la taxa d'error resultant no sigui excessivament alta. A partir d'aquí, els diferents nivells/camins de l'arbre definiran unes característiques concretes que podran ser analitzades.

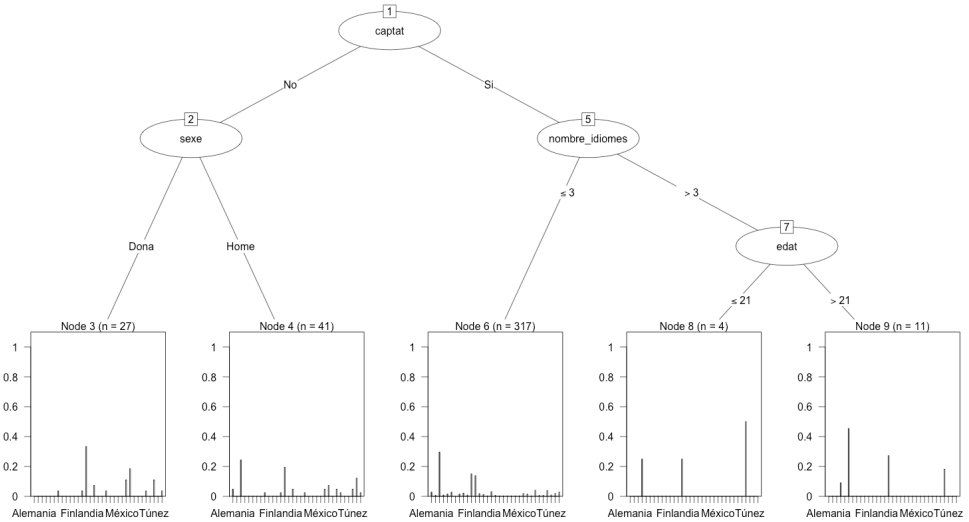
Seguidament, els resultats de l'execució de l'algorisme classificatori;.

Model resultant.

```
Read 400 cases (5 attributes) from undefined.data

Decision tree:

captat = No:
:...sexe = Dona: Francia (27/18)
:   sexe = Home: Australia (41/31)
captat = Si:
:...nombre_idiomes <= 3: Australia (317/223)
    nombre_idiomes > 3:
    :...edat <= 21: Suiza (4/2)
        edat > 21: China (11/6)

Evaluation on training data (400 cases):

          Decision Tree
        ----------------
        Size      Errors

           5   74(18.5%)   <<


        Class             Cases   False   False
                                   Pos     Neg
        -----             -----   -----   -----
        Alemania             11       0      11
        Argelia               2       0       2
        Australia           104      54       0
        Belgica               5       0       5
        Chile                 5       0       5
        China                14       0       9
        Dinamarca             1       0       1
        Egipto                5       0       5
        Emiratos Arabes       8       0       8
        Equador               3       0       3
        España               48       0      48
        Estados Unidos       44       0      44
        Finlandia             8       0       8
        Francia              22      18      13
        Grecia                1       0       1
        Holanda              17       0      17
        Hong Kong             2       0       2
        Hungria               1       0       1
        Japon                 3       0       3
        Jordania              1       0       1
        Kazajstan             1       0       1
        Luxemburgo            1       0       1
        México               1       0       1
        Marruecos            11       0      11
        Noruega              13       0      13
        Polonia               1       0       1
        Portugal             15       0      15
        Reino Unido           3       0       3
        Rusia                 3       0       3
        Suiza                16       2      14
        Túnez                8       0       8
        Venezuela            11       0      11
        Viet Nam             11       0      11
```

```
Attribute usage:

100.00% captat
 83.00% nombre_idiomes
 17.00% sexe
  3.75% edat
```

Representació gràfica de l'arbre resultant



El principal problema de l'arbre classificatori resultant és la poca uniformitat dels grups definits. No obstant, sí que se'n poden extreure algunes lectures que ajudaran a l'avaluació conjunt amb la font de dades de les campanyes anteriors.

També cal esmentar que el marge d'error és del 18,5%, fet que no és favorable alhora de prendre aquestes conclusions com a molt fiables.

La lectura del model resultant correspon a:

- La majoria de les dones amb nacionalitat francesa i dels homes australians no provenen de campanyes anteriors.
- Dels que provenen de campanyes anteriors:
  - Si parlen més de 3 idiomes, probablement vinguin de Xina o de Suïssa. A més, l'edat d'aquests serà major o igual a 21 anys.
  - Si parlen 3 idiomes o menys, els candidats probablement provindran d'Austràlia.
  - La majoria d'alumnes actuals provenen d'Austràlia, d'Espanya, dels Estats Units, d'Holanda i de Suïssa.

Aquests resultats, conjuntament amb l'anàlisi individual de cada atribut i els resultats dels models anteriors sobre les campanyes de captació, permetran definir les regions geogràfiques més adients.

2.4 Quarta fase – Avaluació i interpretació global

Les conclusions que es poden extreure dels resultats anteriorment obtinguts estan relacionades per una banda amb la fiabilitat dels resultats i per altra banda amb l'acompliment dels objectius del projecte.

Objectius, resultats i fiabilitat

El conjunt dels models resultants tant de les campanyes anteriors com dels alumnes actuals es poden avaluar i interpretar segons l'objectiu definit en l'abast del projecte.

Aquests resultats però, no disposen d'un suport molt elevat i caldria complementar-los amb alguna línia de treball futura abans de donar-los per vàlids.

Clarament, i un cop avaluades totes les fonts de dades i els models resultants, això és conseqüència directa de la poca taxa d'acceptació dels candidats. Aquesta poca eficiència, o filtre massa restrictiu d'entrada als estudis, distorsiona de forma clara els països potencialment favorables pel que fa a l'acceptació als estudis.

Caldria doncs disposar de més informació i fonts de dades diferents per a poder donar suport o bé desmentir per complet els resultats aquí exposats.

- **Patrons geogràfics descoberts**

  Aquest objectiu, i principal del projecte, és lamentablement el més difús donada la naturalesa de les dades esmentada anteriorment. No obstant, si que s'han pogut definir certs comportaments que poden ajudar a l'hora d'invertir recursos econòmics en certes zones geogràfiques.

  Existeixen països on s'han realitzat la majoria de contactes de possibles candidats però que finalment tenen una taxa d'acceptació molt menor i, per tant, no permeten predir i/o definir un patró evident. Aquests són clarament España i els Estats Units amb diferència.

  Aquesta última afirmació amaga però una lectura contradictòria, ja que precisament els països d'on provenen la majoria d'alumnes actuals corresponen a Espanya i als Estats Units, a part d'Austràlia, Holanda i Suïssa.

El model classificador i predictiu no han pogut arribar a aquesta conclusió degut a que la gran majoria d'alumnes captats durant l'històric de campanyes no han sigut acceptats. Aquesta paradoxa entre els alumnes actuals i els resultats de les campanyes tant sols podria ser corregida afegint més informació al sistema d'anàlisi.

Així doncs, es prendran com a vàlids els resultats obtinguts sobre el model classificatori dels alumnes actuals creuats amb els obtinguts de l'anàlisi bayesià. Conseqüentment, s'acaba per definir que les zones geogràfiques més interessants on invertir recursos són Espanya, els Estats Units, Austràlia, Holanda i Suïssa. També es prendrà en consideració **Xile** degut a que ha sigut l'únic país, a part d'Austràlia, que el model associatiu de les campanyes anteriors ha pogut trobar.

Acotant la zona geogràfica a ciutats en concret tant sols s'ha pogut descobrir que Melbourne (Austràlia) i Santiago de Chile (Xile) són aquelles amb major probabilitat que un estudiant sigui captat.

A més, en el cas d'Espanya el cost econòmic encara és menor a la resta de països degut a la proximitat geogràfica i el nivell de vida.

Finalment, Austràlia és la zona geogràfica per excel·lència on s'haurien de focalitzar esforços, ja que tots els models classificatoris i predictius han apuntat a ella.

Directament relacionat amb la zona destí s'ha descobert que els candidats provinents de centres privats són aquells que major rati d'acceptació han obtingut.

Resumint aquest apartat de forma esquemàtica, ja que és l'objectiu principal, obtenim que:

- Països preferibles: Espanya, Estats Units, Austràlia, Holanda, Suïssa i Xile

- Ciutats concretes: Melbourne i Santiago de Chile

- Tipus de centre destí: privats

- **Medi més favorable**

En tots els anàlisis executats, s'ha pogut observar que tant les fires com les visites tenen un índex de captació i acceptació molt similar. Aquest fet fa que no es pugui discriminar entre cap d'aquests dos mètodes alhora de desplaçar-se a un país concret.

Si que s'ha descobert, però, que els candidats provinents del medi online tenen una probabilitat molt més alta de ser acceptats. No obstant, el nombre de registres totals provinents d'aquest medi és molt inferior comparat amb els altres i, per tant, és arriscat apostar-ho tot per aquest mitjà.

Vinculat amb un dels objectius que era la de l'ajustament de la despesa, si que es podrien destinar recursos econòmics per aquesta via tot retirant-los dels països on l'acceptació final ha sigut nul·la i que tampoc es tenen estudiants actualment.

Com invertir en campanyes telemàtiques és molt més econòmic que cobrir desplaçaments en avió, allotjaments i dietes dels treballadors encarregats de la captació física, sembla raonable invertir en aquest medi donat que la seva taxa d'acceptació és més favorable.

Una altra lectura interessant, no relacionada directament amb el medi però si amb el mètode, és que el seguiment personalitzat d'un candidat interessat té una correlació directe en que l'estudiant sigui finalment acceptat. Probablement la motivació per un contacte constant faci que aquests es prenguin més seriosament el procés d'admissió.

Resumint aquest apartat obtenim que:

- Medi idoni: captació online
- Seguiment personalitzat: si

- **Avaluació dels recursos destinats actualment**

Tenint en compte que el fitxer que es disposa per despesa és el destinat a cada país, queda exclosa implícitament la lectura dels recursos destinats a la captació online.

Així doncs, comparant els resultats dels models amb la inversió destinada a cada país se n'obtenen unes lectures clares.

- La despesa i el nombre de candidats reclutats té una correlació directe, però no en acceptació.
- Els països menys eficients econòmicament parlant, segons el models associatius i classificatoris, són Espanya i els Estats Units.
- La despesa en Austràlia sembla justificada ja que en tots els models aquest país ha sigut present com a destinació favorable.

- La resta de països tenen una despesa bastant homogènia directament proporcional amb el nombre de candidats finals.

3. Conclusions

3.1. Lliçons apreses del treball

Una de les principals lliçons apreses a través de l'execució d'aquest projecte és que no per plantejar i definir uns objectius molt concrets s'han de poder extreure unes conclusions clares.

Tot i que en un projecte de mineria de dades es pot tractar i analitzar una font de dades concreta amb diversitat d'eines i metodologies capaces d'extreure'n coneixement, aquestes poden no tenir un comportament predictible prou elevat.

La principal lliçó apresa, tant amb l'execució del projecte com amb la lectura de material referent a *data mining*, és que amb gran probabilitat es podran trobar patrons i informació oculta en pràcticament qualsevol conjunt de dades, però aquesta no té perquè adaptar-se als objectius definits pel projecte.

També s'ha après que alhora de planificar el projecte cal definir amb major precisió les etapes així com afegir aquelles relacionades amb l'aprenentatge i l'adquisició de coneixements, pretenent així minimitzar al màxim les possibles desviacions de temps.

3.2. Assoliment d'objectius plantejats

Els objectius d'aquest projecte han pogut ser finalment complerts però, sent el màxim crític possible, aquests no han sigut prou contundents i clarificadors com per poder donar-los per vàlids de forma automàtica.

Finalment, com s'ha pogut veure al llarg del treball, s'ha sigut capaç de definir tant zones geogràfiques potencialment favorables com destacar un medi de captació per sobre dels altres. La destinació de recursos de forma eficient en un futur és conseqüència directa dels resultats anteriors.

En un entorn empresarial real, tot i haver trobat patrons models avaluables en les dades no es podria tenir la seguretat de que aquests resultats fossin definitius i, empresarialment parlant, els objectius no haurien estat complerts en la seva totalitat. Iniciar una línia de treball en la direcció marcada pels resultats podria suposar un fort risc i ser finalment contraproduent en els objectius establerts.

Aquest és doncs un cas clar on tot i haver executat un projecte de mineria de dades en la seva totalitat i haver-ne complert els objectius, les dades font no han permès donar els resultats amb la seguretat desitjada.

3.3. Seguiment de la planificació i metodologia

Planificació

La planificació d'aquest projecte s'ha pogut seguir sense gaires imprevistos i de forma majoritària s'han pogut complir totes les fites exceptuant la de la fase d'anàlisi. Aquesta s'ha vist endarrerida per errors en la generació de les dades i la introducció dins del sistema d'anàlisi R (problemes amb la codificació, els tipus d'atributs i les funcions a aplicar).

Per altra banda, la inexperiència en projectes de mineria de dades ha fet que s'hagi tingut que passar de la fase 2 (preparació de les dades) a la fase 3 (anàlisi d'aquestes) en diverses ocasions. Tanmateix, la fase 4 d'anàlisi també s'ha fet inevitablement paral·lelament durant l'execució de la fase 3 i, personalment, no n'estic segur que sigui la forma més correcte de dur a terme aquest tipus de projectes.

A més, penso que en la planificació s'haguessin tingut d'introduir etapes d'aprenentatge de les eines a emparar i de cerca d'informació, ja que també han suposat una alta dedicació de recursos (en aquest cas de temps).

En un projecte real, si no es disposessin dels coneixements i aptituds necessaris, caldria també contemplar-ho pel fet que podria comportar desviacions importants en la planificació.

Metodologia

Pel que fa a la metodologia, crec que ha sigut contraproduent el fet de generar un conjunt de dades fictici a partir del real. Tot i que s'han pogut extrapolar les dades en un grau bastant elevat, és probable que finalment s'hagin distorsionat els resultats.

Aquest fet també ha penalitzat en el compliment de la segona fase de preparació de les dades, tot tenint que tornar-hi posteriorment un cop començat l'anàlisi de d'aquestes. Això va ser degut a errors en la translació de valors i que no van ser descoberts fins avançada la tercera fase. Entenent que les etapes d'un projecte de *data mining* acostumen a ser cícliques, i que la preparació de les dades s'emporta gran part dels esforços, no és una sorpresa però si hagués pogut ser evitat amb una font de dades fiable i major experiència en l'execució d'aquests tipus de projectes.

No obstant, s'ha tingut en compte el cicle de vida d'un projecte de mineria de dades i, per tant, la metodologia seguida ha correspòs amb l'esperat.

Finalment, a nivell d'aprenentatge, esdevé la reflexió que hagués sigut preferible disposar de dades sòlides i testades la primera vegada que es

realitza un projecte “real” d’aquest tipus, per tal d’obtenir uns resultats més clarificadors. No obstant, aquest cas és més fidedigne a una realitat empresarial on res està preparat per obtenir un resultats satisfactoris.

3.4. Línies de treball futur

Penso que la millor línia de treball futur per tal de donar suport als resultats d’aquest projecte, o bé trobar noves regions geogràfiques favorables, hauria ser mitjançant Big Data.

Introduir en el procés d’anàlisi dades tals com densitat d’escoles de negoci per ciutat/país, nombre de mencions i *likes* de temes relacionats amb els estudis oferts a les xarxes socials, regions propenses a la generació d’empreses o amb comportaments clars de creixement, entre moltes altres, crec que pot aportar un valor afegit molt clar alhora de predir i descobrir informació que d’entrada es desconeix.

Així doncs, una bona forma de treure profit als resultats poc concisos de les dades històriques i transaccionals de les que s’han disposat en aquest projecte és la d’incorporar dades externes de tota mena i analitzar-ne els resultats de cada conjunt com un tot.

Una altra línia de treball futur, donat que un dels resultats del projecte més fiables és el que estableix que la captació online és la més eficient (inscrits/acceptats), seria la de posar en marxa campanyes de captació per via telemàtica.

D’aquesta manera, llençar campanyes publicitàries en tots els medis socials i *blogs*, oferir sessions informatives remotes amb ronda de preguntes i respostes, activar un mètode que permetés fer les provés d’accés no presencials i, en definitiva, activar el major nombre possible d’eines online que permetin arribar a candidats d’arreu del món sense tenir que desplaçar-se sembla una de les millors alternatives, tant per estalviar costos de desplaçament com per arribar a un major públic.

4. Glossari

Algorisme C50: l'algorisme C50 és un algorisme classificador evolució de l'algorisme ID3 i C4.5 per a la construcció d'arbres decisionals.

Algorisme associatiu (apriori): un algorisme associatiu és aquell que descriu un domini a partir de dependències entre conjunts de valors. Aquest permet definir comportaments o regles tot descobrint relacions entre atributs aparentment no relacionats. Es binaritzen unes dades d'entrada com a transacció i de cada transacció esdevé una regla provinent d'un conjunt d'atributs. Apriori és la llibreria d'R que permet l'execució d'aquest tipus d'algorismes.

Algorisme classificador: un algorisme classificatori és aquell que es basa en la idea de trobar atributs discriminants que generin particions de classificació prou uniformes.

Algorisme d'agregació (kmeans): l'algorisme d'agregació o de *clustering* és aquell que cerca un criteri per agrupar objectes semblants tot establint proximitat entre les dades. L'algorisme *kmeans* permet dur a terme aquesta fita a partir d'atributs numèrics.

Algorisme de Naive Bayes: l'algorisme de Naive Bayes és un algorisme classificador considerat ingenu, que permet construir una xarxa bayesiana tot analitzant cada atribut de la font de dades per separat.

Atribut (camp): és la unitat mínima d'informació que compon cada font de dades. Aquest pot ser numèric, lògic, categòric, descriptiu... El conjunt d'aquests atributs componen un domini.

Mineria de dades/*data mining*: és un procés de descobriment de coneixement desconegut o ocult sobre grans conjunts de dades.

Model: un model és una descripció articulada i abstracta d'una realitat. Un model permetrà, doncs, predir, endevinar o establir un resultat concret donat un conjunt d'atributs d'entrada.

Paràmetre rhs: en anglès *right-hand-side* és l'atribut conseqüent d'un conjunt d'atributs d'entrada (*lhs* o *left-hand-side*) en aquelles regles generades per un algorisme apriori.

5. Bibliografia

Documents

Ramón Sangüesa i Solé; Luis Carlos Molina Félix, *Apunts de l'assignatura de Mineria de Dades*, UOC, Curs 2015-16-1

T. Warren Liao; Evangelos Triantaphyllou, *Recent Advances in Data Mining of Enteprise Data: Algorithms and Applications*, 2007

Yanchang Zhao, *R and Data Mining: Examples and Case Studies*, 2013

Recursos web

Informació referent a R-Project. Lloc oficial de CRAN.
https://cran.r-project.org, finals 2015

Informació concreta sobre Mineria de Dades en R-Project
http://www.rdatamining.com/, finals 2015

Exemples concrets sobre anàlisi de dades en R
http://www.salemmarafi.com/, finals 2015

Exemples concrets sobre anàlisi de dades en R
http://connor-johnson.com/?s=c50, finals 2015

Consultes i tractament d'errors d'R
http://stackoverflow.com/, 2015