Estat de l'Art de l'Anàlisi Predictiu

Arnau Sargatal Prat
Grau en Enginyeria Informàtica

Xavier Martinez Fontes

14/01/2019

Llicències alternatives (triar alguna de les següents i substituir la de la pàgina anterior)

A) Creative Commons:

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

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

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

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

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

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

B) GNU Free Documentation License (GNU FDL)

C) Copyright

FITXA DEL TREBALL FINAL

Títol del treball:	*Estat de l'art de l'anàlisi predictiu*
Nom de l'autor:	*Arnau Sargatal Prat*
Nom del consultor:	*Xavier Martinez Fontes*
Data de lliurament (mm/aaaa):	*01/2019*
Àrea del Treball Final:	*Business Intelligence*
Titulació:	*Grau en Enginyeria informàtica*

Resum del Treball (màxim 250 paraules):

Actualment les tecnologies de la informació es troben en un punt d' inflexió degut a la digitalització de organitzacions i els grans volums de dades que genera. L'anàlisi de dades ha esdevingut clau per a les estratègies corporatives, i més si aquest anàlisi te la capacitat de determinar el futur probable de un esdeveniment o la probabilitat que es produeixi un situació, aquest concepte s'anomena l'Anàlisi Predictiu.

En aquest treball es realitza un estudi de l'estat de l'art de l'Anàlisi Predictiu, una branca de coneixement de l'anàlisi de dades, que generalment prediu ocurrències i les seves probabilitats a partir dels seus històrics de dades. Utilitza tècniques com l'aprenentatge computacional o la mineria de dades, entre d'altres per crear processos de predicció.

El treball comença amb una introducció sobre el concepte d'anàlisi predictiu, destacant la diferenciació de dades i informació, i la suposició de la predicció. També es toca un tema important com els aspectes ètics i legals en la manipulació de dades, ja que actualment es una qüestió molt oberta i que genera debat de manera continuada.

A partir d'aquestes definicions, s'evoluciona el projecte cap a la descripció de l'anàlisi predictiu i les seves ramificacions, on es desenvolupen les tipologies d'anàlisi, problemàtiques i tècniques de modelatge. Posteriorment s'analitzen diferents eines de modelatge tant a nivell de codi obert com software propietari, i s'exemplifiquen diferents models amb l'ús d'eines de codi obert.

Finalment es conclou amb l'exploració de les tendències actuals i casos d'ús real de l'aplicació d l'Anàlisi Predictiu.

Abstract (in English, 250 words or less):

Information technology is currently at a turning point due to the digitization of organizations and the large volumes of data it generates. Data analysis has become a key to corporate strategies, and more if this analysis has the ability to determine the probable future of an event or the likelihood of a situation, this concept is Predictive Analysis.

In this work, a study of the state of the art of Predictive Analysis is done, it is a branch of knowledge of data analysis, which generally predicts occurrences and their probabilities based on their historical data. It uses techniques such as machine learning or data mining, among others to create prediction processes.

The work begins with an introduction on the concept of predictive analysis, highlighting the differentiation of data and information, and the assumption of prediction. An important issue such as the ethical and legal aspects of data manipulation, as it is actually a very open question that generates debate on an on-going way.

Based on these definitions, the project is developed towards the description of predictive Analysis and its branches, where the typologies of analysis, problematic and modelling techniques are developed. Subsequently, different modelling tools are analysed, at open source level and proprietary software, and different models are exemplified with the use of open source tools.

Finally, it concludes with the exploration of the current tendencies and cases of real use of the Predictive Analysis application.

Paraules clau (entre 4 i 8):

Anàlisi, Predictiu, Algoritme, Computacional, Modelatge, Aprenentatge.

Índex

1. Introducció 1
1.1 Context i justificació del Treball 1
1.2 Objectius del Treball 2
1.3 Enfocament i mètode seguit 3
1.4 Planificació del Treball 3
1.5 Breu sumari de productes obtinguts 5
1.6 Breu descripció dels altres capítols de la memòria 5
2. Anàlisi Predictiu 6
2.1. Dades i informació 6
2.2. Suposició vs. Predicció 7
2.3. Aspectes ètics i legals sobre l'anàlisi predictiu 8
2.4. Descripció de l'anàlisi predictiu 9
2.4.1. Que és l'anàlisi predictiu? 9
2.4.2. Perquè és important l'anàlisi predictiu? 9
2.4.3. Com funciona l'anàlisi predictiu? 10
2.4.4. Objectius de l'Anàlisi Predictiu 10
2.5. Tipologies d'anàlisi i validació 11
2.5.1. Models Predictius 11
2.5.2. Models Descriptius 12
2.5.3. Models de decisió 12
2.5.4. Models Prescriptius 13
2.5.5. Models Combinats 14
2.5.6. Models Incremental 14
2.5.7. Validació de models 15
2.6. Problemàtiques de l'anàlisi predictiu 16
2.7. Tècniques de l'anàlisi predictiu i modelatge estadístic 18
2.7.1. Reducció Dimensional 18
2.7.2. Regressió 20
2.7.3. Arbres de decisió 22
2.7.4. Estadística Bayesiana 24
2.7.5. Xarxes Neuronals 26

2.7.6. Deep Learning 28
2.7.7. Modelatge Combinat 28
2.7.8. Modelatge Basat en Instancies 30
2.7.9. Clustering 31
2.7.10. Selecció de Models i Avaluació 32
2.7.11. Regularització 33
2.7.12. Sistema de regles (Rule Based System) 33
2.8. Aprenentatge Computacional (Machine Learning) 34
2.8.1. Tipus d'aprenentatge 34
2.8.2. Comparativa entre tipus d'aprenentatges 35
2.9. Mineria de Dades (Data Mining) 37
2.9.1. Procés de mineria de dades 37
3. Eines 38
3.1. Codi Obert 39
3.1.1. R i RStudio 40
3.1.2. Scikit-learn (Python) 48
3.2. Software Propietari / Comercial 56
3.2.1. IBM SPSS 56
3.2.2. MATLAB 56
3.2.3. SAP HANA 57
3.2.4. Altres 57
3.4. R vs. Python 60
3.4.1. Anàlisi de R 61
3.4.2. Anàlisi de Python 61
4. Tendències Actuals i Aplicabilitat 62
4.1. Tendències actuals i casos d'ús 62
4.1.1. Sector Energètic 62
4.1.2. Manteniment predictiu 64
4.1.3. Sector Esports 65
4.1.4. Sector Salut 66
4.1.5. Sector Banca 66
5. Conclusions 68
5.1. Conclusions Finals 68
5.2. Línies de Futur 69

6. Glossari 70
7. Bibliografia 72
7.1. Webgrafía 72
7.2. Referències d'imatges 76
8. Annexos 77

Llista de Figures

Figura 1. Diagrama de Gantt. 4
Figura 2. Regressió lineal. [54] 20
Figura 3. Regressió Logística. [55] 22
Figura 4. Perceptró Multicapa.[56] 27
Figura 5. Interfície RStudio. 40
Figura 6. Iris Setosa Kmeans 1. 42
Figura 7. Iris Setosa Kmeans 2. 43
Figura 8. Arbres de decisió 1. 45
Figura 9. Arbres de decisió 2. 46
Figura 10. Arbres de decisió 3. 46
Figura 11. Arbres de decisió 4. 47
Figura 12. Millors Programari d' Anàlisi Predictiu 58
Figura 13. Comparativa R vs. Python. 61

1. Introducció

1.1 Context i justificació del Treball

Information is the oil of the 21st century and analytics the engine. (Peter Sondergaard)

Aquesta frase de Peter Sondergaard, reflecteix el canvi desenvolupat en l'àmbit de la gestió de dades en les últimes dècades, l'analítica de dades ha deixat de ser qualsevol acció empresarial aïllada per convertir-se en una eina corporativa a nivell global. Cada vegada més empreses mitjançant la transformació digital, integren dins la seva estratègia empresarial aquesta tipologia de processos degut a l'ampli ventall de possibilitats que apareixen al seu davant.

La gestió i estudi de les transformacions de dades, en qualsevol de les seves branques, des de una visió corporativa dona una conclusió evident. Qualsevol empresa que adopti una estratègia de negoci basada en l'analítica de dades produeix avantatges en segments de l'organització com benefici financer, canvi cultural en la dinàmica de l'empresa i impacte en l'estratègia de la companyia. Tres estaments fonamentals dins una organització per a l'adaptació al canvi digital.

A mesura que l'anàlisi de dades avança, engloba més i més contextos. Es podria posar el cas dels fraus cap a les asseguradores, que es comptabilitzen en bilions d'euros anualment. Aquesta tipologia d'empresa intentar provar que les reclamacions de certs clients són fraudulentes, i que això pot suposar un cost més elevat que el cost original de la reclamació. Per aquest motiu moltes empreses han anat adherint-se a l'aprenentatge automàtic i als models predictius per detectar el frau.

Aquest procés ajuda a identificar les reclamacions que han de ser investigades pels auditors humans. No només redueix els costos temporals del treballadors de les asseguradores, sinó que també incrementa l'oportunitat de reclamar els diners robats per reclamacions fraudulentes.

En un altre àmbit, com el de la salut, l'anàlisi predictiu pot detectar quan un pacient necessita atenció mèdica o ajudar a optimitzar els recursos dels centres de salut. En certs hospitals de Estats Units s'ofereixen solucions d'anàlisi predictiu per ajudar a reduir els índexs de readmissió de l'hospital, preveure la probabilitat que un pacient contragui malalties terminals i predir la probabilitat que un pacient perdi la seva cita.

El sistema de salut hospitalari treballar juntament amb professionals de la ciència de dades que revisen les dades històriques de readmissió i determinen un pla de cura adequat a la llar de l'afectat en el moment de l'alta per tal de

reduir les readmissions. En certs centres es va obtenir una reducció relativa del 39% en la taxa de readmissió de 30 dies per a totes les causes, incloent una reducció relativa del 52% en la taxa de readmissió de 30 dies de pacients amb un diagnòstic principal d'insuficiència cardíaca.

S'observa doncs com l'anàlisi de dades i mes concretament l'anàlisi predictiu, es un concepte que engloba i englobarà empreses del sector privat, públic i altres organitzacions, i mitjançant el qual es pretén substituir les intuïcions per decisions fonamentades, o reduir el desconeixement per a la presa de decisions alineades amb objectius concrets.

Aprofitant el nínxol de mercat que ofereix l'anàlisi predictiu, grans empreses del mon dels sistemes d'informació com Microsoft, Oracle, IBM, SAP han enfocat part dels seus esforços a desenvolupar eines per a la realització de tasques i processos d'anàlisi predictiu. No obstant també existeixen tecnologies de codi obert de gran qualitat i amb les mateixes capacitats de computació.

1.2 Objectius del Treball

L'objectiu d'aquest treball es establir una estructura de coneixement de l'estat de l'art de l'Anàlisi predictiu i els diferents punts del que comporta.

Aquest TFG es desenvoluparà en el marc de l'àmbit BI (Business Intelligence) i es farà èmfasi en els fonaments de l'anàlisi predictiu, fent un recorregut per les diferents tècniques que engloba, i a partir d'aquestes es centrarà el projecte en els models predictius, dels quals se'n farà una descripció i representació, per tal de poder visualitzar les diferents tècniques que utilitzen els models.

Un cop descrits els models predictius, es centrarà el treball en les eines que s'utilitzen per a l'anàlisi predictiu, on es descriuran diferents plataformes de modelatge estadístic i predictiu, com puguin ser R, Oracle BI, SAP, etc.

Finalment i per veure l'aplicació en la realitat de l'anàlisi predictiu, es representaran diferents casos reals i/o tendències en diferents àmbits com puguin ser urbanisme, sector de la moda, sector de l'automoció, etc.

Com a objectius principals es proposen:

- Anàlisi i desenvolupament dels fonaments de l'Anàlisi Predictiu
- Anàlisi i desenvolupament dels principals models i tècniques de l'Anàlisi Predictiu
- Anàlisi i desenvolupament de les eines de modelatge de l'Anàlisi Predictiu
- Anàlisi i desenvolupament de tendències i casos d´ús de l'Anàlisi Predictiu

1.3 Enfocament i mètode seguit

Per a fer un control de les fites temporals durant el desenvolupament del treball, es desenvolupa un diagrama de Gantt on es marcaran els temps de cada PAC i els seus punts específics a desenvolupar. L'enfocament i la metodologia per al correcte desenvolupament del TFG, destinarà els recursos temporals a fer recer-ca i investigació dels següents punts:

- Recerca i investigació sobre els fonaments de l'anàlisi predictiu
- Recerca i investigació sobre les principals tècniques i models predictius
- Recerca i investigació sobre el tipus d'eines utilitzades
    - Anàlisi de les plataformes
- Recerca i investigació sobre tendències actuals i casos d'ús.
- Conclusions
- Presentació

1.4 Planificació del Treball

Per a fer un control de les fites temporals durant el desenvolupament del treball, es facilita un diagrama de Gantt on es marcaran els temps de cada PAC i els seus punts específics a desenvolupar. El seguiment del contingut de les tasques i el desenvolupament dels objectius de contingut s'ha realitzat paral·lelament amb el professor col·laborador de la UOC per a aquesta assignatura.

S'adjunta diagrama de Gantt per a representar la planificació temporal de l'estructura del treball i el desenvolupament dels diferents punts establerts.

Id	Modo de tarea	Nombre de tarea	Duración	Comienzo	Fin
1		TFG_BI_SoA_MP	16 días	lun 24/09/18	lun 15/10/18
2		Elaboració PAC1	16 días	lun 24/09/18	lun 15/10/18
3		Definició conceptual	7 días	lun 24/09/18	mar 02/10/18
4		Recerca de informació	5 días	mié 03/10/18	mar 09/10/18
5		Estructurar TFG	3 días	mié 10/10/18	vie 12/10/18
6		Repàs i entrega PAC1	1 día	lun 15/10/18	lun 15/10/18
7		Elaboració PAC2	25 días	mar 16/10/18	lun 19/11/18
8		Revisió conceptes PAC1	2 días	mar 16/10/18	mié 17/10/18
9		Recerca fonaments AP	6 días	jue 18/10/18	jue 25/10/18
10		Recerca Principals MP	6 días	vie 26/10/18	vie 02/11/18
11		Recerca Principals Plataformes MP	5 días	lun 05/11/18	vie 09/11/18
12		Redacció PAC2 + Reestructuració informació	5 días	lun 12/11/18	vie 16/11/18
13		Repàs i entrega PAC2	1 día	lun 19/11/18	lun 19/11/18
14		Elaboració PAC3	16 días	mar 20/11/18	mar 11/12/18
15		Revisió conceptes PAC2	1 día	mar 20/11/18	mar 20/11/18
16		Reestructuració informació i criteris utilitzats	1 día	mar 20/11/18	mar 20/11/18
17		Revisar i finalitzar Capitol 2	4 días	mié 21/11/18	lun 26/11/18
18		Redactar i finalitzar Capitol 3	4 días	mié 21/11/18	lun 26/11/18
19		Recerca tendencies actuals i Casos d'ús (Capitol 4)	2 días	mar 27/11/18	mié 28/11/18
20		Redacció conceptes PAC3	6 días	jue 29/11/18	jue 06/12/18
21		Repàs i entrega PAC3	3 días	vie 07/12/18	mar 11/12/18
22		Lliurament memòria / producte final	34 días	mié 12/12/18	lun 28/01/19
23		Revisió PAC3 i modificacions	7 días	mié 12/12/18	jue 20/12/18
24		Redactar apartat 3.3.	5 días	mié 12/12/18	mar 18/12/18
25		Redacció de conclusions i bibliografia	5 días	mié 12/12/18	mar 18/12/18
26		Integració PAC1+PAC2+PAC3	6 días	vie 21/12/18	vie 28/12/18
27		Revisió TFG	5 días	lun 31/12/18	vie 04/01/19
28		Preparació vídeo	2 días	lun 07/01/19	mar 08/01/19
29		Retocs finals	2 días	mié 09/01/19	jue 10/01/19
30		Lliurament TFG + Memòria Final	1 día	vie 11/01/19	vie 11/01/19
31		Qualificació	1 día	lun 28/01/19	lun 28/01/19

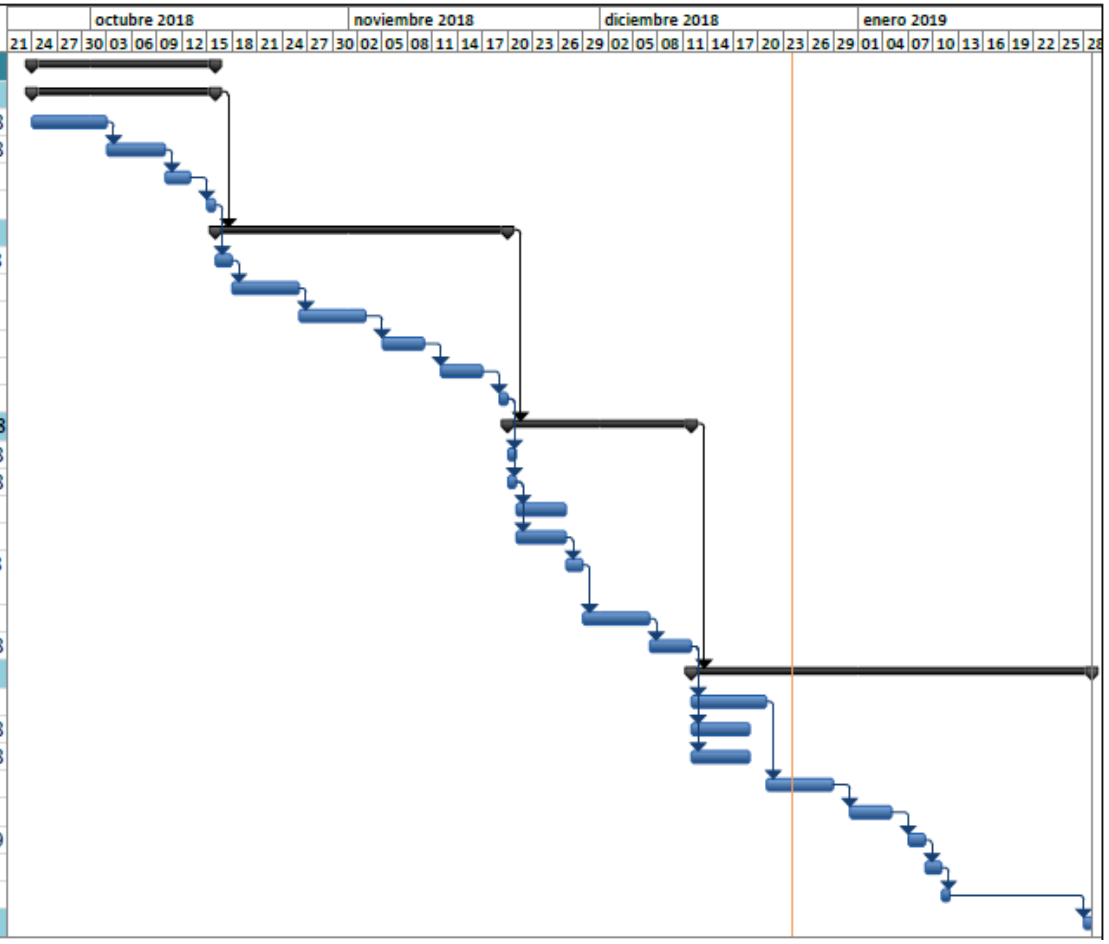


Figura 1. Diagrama de Gantt.

1.5 Breu sumari de productes obtinguts

L'estructura principal del treball consta de 5 capítols:

- El primer capítol fa una petita introducció justificant l'àmbit en que es desenvolupa el treball, es defineixen els objectius principals, es descriu la planificació temporal i l'estructura del treball.
- En el segon capítol es fa èmfasi en el concepte de anàlisi predictiu, es recorren les tècniques estadístiques que utilitza i les diferents branques de modelatge predictiu.
- En els tercer capítol es fa una descripció de diferents eines que s'utilitzen dins de l'anàlisi predictiu.
- En el quart capítol es descriuen les tendències i casos reals d'ús de l'aplicació de l'anàlisi predictiu.
- El cinquè i últim capítol es realitzen les conclusions del treball.

1.6 Breu descripció dels altres capítols de la memòria

- En el sisè capítol hi ha el glossari, on es defineixen els acrònims i conceptes més importants per a desenvolupar el treball.
- En el setè capítol es descriu la bibliografia utilitzada i consultada per a desenvolupar el contingut del projecte.
- En el vuitè capítol s'hi adjunten els annexos del treball.

2. Anàlisi Predictiu

Torture the data, and it will confess to anything. (Ronald Coase)

2.1. Dades i informació

Actualment es viu en l'era de la informació. De la mateixa manera que en èpoques històriques anteriors, com la industria va generar la revolució industrial, el desenvolupament dels sistemes d'informació ha desencadenat l'era de la informació. En aquests últims anys, la informació ha esdevingut poder per a les empreses que la saben gestionar, ja que els hi genera coneixement dels seus propis serveis i de la competència. Conceptualment la informació és la assignació de significat a les dades. Per entrar en més detall sobre dades i informació a continuació es defineixen els termes. [1]

Dades → Informació → Coneixement

El concepte de dades inicials en l'anàlisi predictiu normalment s'anomena "raw data" o dades en brut, que sol ser una conjunt de caràcters com text, números i símbols sense significat que normalment s'han de processar dins un context determinat. Pel que fa al concepte de informació, aquesta és el resultat de processar dades a través de maquinari. Si en aquestes dades processades posteriorment se'ls hi assigna context i significat, esdevenen informació. Per tant la informació són dades amb significat.

En les tecnologies de la informació, els símbols, caràcters, imatges o números són dades. Aquests són inputs que les tecnologies de la informació necessiten per processar i donar interpretació amb significat. La informació pot ser de fets, coses, conceptes o qualsevol tòpic rellevant, i sol anar enllaçada amb preguntes com qui, com, quan, perquè, etc.

En referencia al coneixement, aquest concepte esdevé real quan es produeix com a resultat de entendre informació proporcionada i s'utilitza per relacionar conceptes o solucionar problemes. Per tant a la informació amb alguna aplicació o ús esdevindrà coneixement. [2]

S'observa doncs com actualment les tecnologies de la informació proporcionen la capacitat en que qualsevol objecte (persones, mobiliari, etc.) generi dades i per tant informació derivada en coneixement. Qualsevol tipus d'acció quotidiana, actualment pot ser registrada mitjançant dispositius. Com a exemples clars es podria posar, moviments de GPS, pagaments bancaris, cerques en buscadors, etc.

La tendència actual i en un futur, és que el desenvolupament continuat de la tecnologia obre portes a nous camps d'estudi de la informació i presa de decisions, i més si branques com la internet de les coses continuen creixent a aquest ritme frenètic.

2.2. Suposició vs. Predicció

Sovint en el dia a dia, les persones a través de la visualització de situacions, lectura, o altres fenòmens elaboren imatges mentals que desencadenen en judicis. Però es important entendre aquests successos i la informació que es proporciona per a formular correctament prediccions, que a diferencia de les suposicions estan fonamentades. [3]

Les suposicions solen ser conclusions que no necessàriament se sustenten en evidència o que se sustenten en molt poca. En aquest sentit, les suposicions poden interferir els aspectes subjectius dels individus i també alguns falsos coneixements que certes persones podrien considerar com a veritables i vàlids. Les suposicions poden ser veritat o no, però en molts casos la precisió de les mateixes és només producte de la sort. Alguns exemples de suposicions podrien ser els següents:

- Afirmar que el contingut de un llibre es dolent perquè la imatge de la portada no es agradable.

- Dir que sortirà cara al jugar a cara o creu.

- Últimament ha disminuït el volum de pesca del mediterrani, i afirmar que el 2020 no hi haurà peixos al mar.

Pel que fa al concepte de predicció, consisteix a treure conclusions sobre alguna cosa a partir de les observacions i evidències observades en els successos o fenòmens. Aquestes evidències serveixen per a elaborar conclusions de manera raonable de manera que pot ser explicada i sustentada. Alguns exemples de prediccions podrien ser els següents:

- Si s'escalfa el gel, aquest es fondrà

- Si es veu fum a l'horitzó, afirmar que hi ha foc.

Com s'observa, les prediccions consisteixen en seguir un procés lògic en el qual s'uneixen elements prèviament coneguts. Serà important doncs diferenciar be entre suposició i predicció en el context de l'anàlisi predictiu degut a les connotacions que implica.

2.3. Aspectes ètics i legals sobre l'anàlisi predictiu

L'ús de l'anàlisi de dades i anàlisi predictiu ha plantejat inquietuds ètiques relacionades amb qüestions com la confidencialitat, la comunicació, els registres i la supervisió, i molt més. Sense cap tipus de dubte la gestió de grans volums de dades i l'anàlisi de dades ha donat a formular ètiques relacionades amb qüestions com la confidencialitat, la comunicació, els registres i la supervisió, així com qüestions relacionades amb la privadesa i la seguretat.

Fer un ús ètic de grans conjunts de dades, s'està convertint cada vegada més en una tasca amb més consideració per a grans empreses i industries que utilitzen solucions d'anàlisi predictiu. La privacitat del client, l'ús adequat de les dades dins el límits jurídics, les praxis durant el procés d'anàlisi de cada empresa poden suposar seriosos riscos per les empreses en un futur.

L'ús ètic de les dades, hauria de convertir-se en quelcom primordial per les empreses i indústries, a causa dels constants canvis en les barreres legals dins de l'àmbit de la manipulació de dades. L'abús de analítiques predictives en els hàbits dels clients a vegades pot provocar una reacció contraria a l'esperada, en lloc de veure-ho com una millora del servei. [4]

La clau en aquest tipus de context, resideix en la formació d'hàbits. El resultat de la gestió de la informació, pot ser en la majoria de vegades que es recopili més informació del clients, que la que desitjarien aquests mateixos. Per la millora de aspectes ètics fonamentals, les empreses s'hauran de centrar en com s'utilitzaran les dades recopilades, creació de oportunitats i auditar els processos d'anàlisi involucrats.

Des de un punt de vista jurídic els grans volums de dades, han de tenir en compte qüestions ètiques comentades anteriorment com la com la privacitat, la seguretat, l'emmagatzematge i conservació de dades, així com els drets dels consumidors, la propietat intel·lectual o la reutilització pel sector privat de continguts elaborats per les administracions públiques, entre d'altres.

S'ha de tenir en compte també els efectes que poden tenir el tractament de les dades i les prediccions generades sobre les persones i les seves llibertats.

La protecció de la informació ha estat sempre una tasca essencial per a les empreses que realitzen enquestes, màrqueting o publicitat, en l'era de la informació, i actualment encara més. Mes enllà del compliment de les obligacions derivades del compliment normatiu, les empreses han de tenir en compte altres qüestions, com la transparència i l'ètica com a factors clau a l'hora de establir relacions amb els consumidors. Els usuaris cada vegada més volen informació de quin ús es faran amb les dades, quant de temps i amb quina finalitat. La privacitat i la seguretat de la informació guanyen protagonisme a l'hora de decidir els serveis que utilitza. [5] [6]

2.4. Descripció de l'anàlisi predictiu

2.4.1. Que és l'anàlisi predictiu?

L'anàlisi predictiu es un concepte que engloba un conjunt de tècniques estadístiques que utilitzen històrics de dades per predir esdeveniments futurs. Les dades històriques s'utilitzen per a crear models matemàtics que reprodueixi les tendències més importants. El model predictiu generat s'utilitza doncs per a predir el que succeirà a continuació o accions a dur a terme per tal de obtenir els resultats òptims segons el cas. Les tècniques estadístiques més utilitzades són la mineria de dades, el modelatge predictiu i l'aprenentatge computacional.

El nucli de l'anàlisi predictiu resideix en enllaçar les relacions entre variables actuals i les variables històriques de esdeveniments passats, i explotar-les per predir tendències futures. Cal remarcar que la precisió i la usabilitat dels resultats depèn en la major part del nivell i la qualitat dels filtres en les variables d'entrada. L'anàlisi predictiu es sovint definit com a predicció però a un nivell més detallat de granularitat. La diferencia entre una predicció i una previsió resideix en que com s'ha comentat anteriorment l'anàlisi predictiu és una tecnologia que aprèn de l'experiència (històric de dades) per a fer millor decisions o predir resultats amb una major precisió, mentre que la previsió es basa en suposicions generades de manera aleatòria.

En les últims anys, l'anàlisi predictiu, degut als avenços de la tecnologia ha avançat en molts aspectes i àrees com l'anàlisi de grans volums de dades i l'aprenentatge automàtic. Es una branca de la ciència de dades que esta agafant importància i reconeixement. Dins el context dels grans volums de dades, l'anàlisi predictiu, obté les dades de sensors, instruments i sistemes connectats al món real. Aquest tipus de dada dona tal informació en conjunt, que els teixits empresarials les poden incloure per a la presa de decisions.

2.4.2. Perquè és important l'anàlisi predictiu?

Actualment la informació per a les empreses te un valor incalculable, amb l'augment de la competitivitat, cada negoci busca generar avantatges competitius a l'hora de oferir productes o serveis en el mercat. Els models predictius donen un gir a les empreses amb estratègies convencionals i les ajuden a resoldre problemes quotidians de manera diferent, o renovar estratègies corporatives. Amb l'ús de l'anàlisi predictiu les empreses poden ser més precises a l'hora de efectuar prediccions, de manera que a l'hora de planificar els recursos, poden esser molt més efectives.

Com a cas d'ús per mostrar la importància d'aquest concepte estadístic de manera resumida es posa l'exemple següent. Un fabricant de maquinaria pot estar desenvolupant un producte que li està costant entrar al mercat per la

complexitat de la demanda. Llavors els desenvolupadors del producte afegeixen capacitats predictives a la solució desenvolupada per augmentar el valor afegit de cares al client. La utilització de recursos predictius en el producte pot anticipar fallades, anticipar-se a trencaments, reduir costos de manteniment i operatius. Un cas encara més concret podria ser el de uns sensors amb capacitat de mesurar les vibracions dels components de un ascensor, per tal de indicar la necessitat de manteniment de l'aparell, abans de que falli la màquina.

2.4.3. Com funciona l'anàlisi predictiu?

L'anàlisi predictiu és el procés de fer un ús de les dades per a realitzar prediccions de futur basades en aquestes. Aquest procés on es tracten les dades engloba les dades, diferents tècniques de modelatge estadístic i aprenentatge automàtic per tal de crear models predictius. Tot aquest procés el que pretén es crear una predicció de caire quantitatiu sobre el futur. La majoria de vegades s'utilitzen tècniques d'aprenentatge automàtic supervisat per predir valors futurs (Quan tardarà la televisió en espatllar-se?) o calcular probabilitats (Quina probabilitat hi ha que un client pugui comprar un Iphone X?).

L'objectiu de l'anàlisi predictiu a nivell comercial esta clar, utilitzar les dades per estalviar recursos (temps i diners, en la majoria de casos). Els processos de modelatge sovint utilitza masses de dades heterogènies amb capacitat de generar resultats clars i precisos amb l'objectiu definit o especificacions prèvies.

2.4.4. Objectius de l'Anàlisi Predictiu

Com s'ha comentat en el punt 2.4.2., aquesta tipologia d'anàlisi està essent de vital importància en estratègies empresarials, les organitzacions estan recorrent a aquestes metodologies per ajudar a resoldre problemes complexes i descobrir noves oportunitats. Els objectius poden ser molt variats, els processos i tècniques, són adaptables i aplicables en qualsevol àmbit. Per tant dependrà del camp que s'apliquin i el resultat que se'n vulgui obtenir, no obstant alguns dels objectius estratègics més importants serien els següents.

- Competitivitat, per tal de competir de manera eficaç i competitiva.
- Creixement, en vendes, clients, i exploració de nous nínxols de mercat.
- Millora de les polítiques de seguretat, fent èmfasi a la gestió, detecció i prevenció de frau.
- Millora Operacional, cercant avantatges competitius i millorant les capacitats de les organitzacions.
- Millora en la satisfacció dels clients, avançant els moviments corporatius a les expectatives dels clients.
- Millora en la presa de decisions

Avançar segons els objectius de l'anàlisi predictiu, proporciona gran impacte en les àrees on es desenvolupa l'activitat, doncs es important tenir clar la tipologia de problema que es vol enfocar i disposar de les eines adequades.[7]

2.5. Tipologies d’anàlisi i validació

2.5.1. Models Predictius

Els models predictius utilitzen metodologies estadístiques i de càlcul per a predir les tendències sobre un esdeveniment, que succeirà en un futur. Funciona analitzant dades històriques i actuals i generant un model que permeti predir els resultats futurs En el modelatge predictiu, es recopilen dades, es formula un model estadístic, es fan prediccions i el model es valida (o es revisa) a mesura que hi ha dades addicionals disponibles. Els models predictius analitzen el rendiment passat per avaluar la probabilitat que un client exhibeixi un comportament específic en el futur. Aquesta categoria també inclou models que busquen patrons de dades subtils per respondre a preguntes sobre el rendiment del client, com ara models de detecció de fraus.

El modelatge predictiu actualment està entrat amb molta força a les aplicacions de negoci. Un dels usos més habituals de la d’aquest tipus de models predictiu és la publicitat i el màrqueting en línia. Els modeladors utilitzen les dades històriques dels navegadors web, utilitzant-los mitjançant algoritmes per determinar quins tipus de productes els usuaris poden estar interessats i el que probablement facin clic. Un altre aplicació dels models de predicció seria com a filtre de Spam en el correu, on es fan servir models de predicció per identificar la probabilitat que un missatge donat sigui correu brossa.

En la detecció de fraus, la modelització predictiu s'utilitza per identificar els forats en un conjunt de dades que apunten cap a l'activitat fraudulenta. I en la gestió de relacions amb els clients (CRM), el modelatge predictiu s'utilitza per orientar la missatgeria als clients que tenen més possibilitats de realitzar una compra.

En aquest tipus de models un dels punts més rellevants és la recollida de les dades adequades per utilitzar durant el desenvolupament dels algoritmes. Segons algunes estimacions, els científics de dades gasten aproximadament el 80% del seu temps en aquest pas.

A diferencia dels models descriptius que ajuden a entendre el que va passar, o altres models que ajuden a comprendre les relacions entre successos i determinen per què va passar alguna cosa. Els models predictius es basen en esdeveniments futurs.

De manera genèrica es podria dir que hi ha dos tipus de models predictius. Els models de classificació que preveuen a quina agrupació pertanyerà una classe. Per exemple, es podria classificar si algun client te risc de abandonar l’entitat o empresa, quin nivell de morositat pot tenir, etc. O també hi ha els models de

regressió, que preveuen successos de manera numèrica, per exemple, el nombre d'ingressos que generarà un client durant el proper any o el nombre de mesos que un fabricant fallarà pot fallar en proveir d'un servei.

Les tècniques de modelatge predictiu es desenvolupen més endavant en l'apartat 2.7. Algunes de les tècniques de modelització predictiva més utilitzades són els arbres de decisió, la regressió i les xarxes neuronals.

2.5.2. Models Descriptius

L'anàlisi descriptiva és l'exploració de dades o continguts, generalment realitzats manualment, per respondre a la pregunta "Què va passar?" (O Què passa?). Són analítiques que descriuen el passat. Caracteritzada per la intel·ligència empresarial tradicional (BI) i visualitzacions com ara gràfics, gràfics de barres, línia gràfics, taules o narracions generades. S'utilitzen aquest tipus de models quan es necessita comprendre a nivell global el que està passant en certes circumstancies i quan es vulgui resumir i descriure aspectes diferents a aquestes.

L'anàlisi descriptiu pretén proporcionar una imatge precisa del que ha succeït en un context específic i com això es diferencia d'altres períodes comparables. Aquestes mètriques de rendiment es poden utilitzar per marcar àrees de força i debilitat per tal d'informar l'estratègia de gestió dins de l'entorn analitzat. [8]

Les tècniques estadístiques utilitzades en aquesta categoria solen ser de aritmètica bàsica (Sumes, mitjanes, percentatges de canvis, etc.). Normalment, les dades bàsiques són un recompte o un agregat d'una columna de dades filtrada a la qual s'aplica la processos de càlcul simples. Exemples habituals d'anàlisi descriptiu són informes que proporcionen informació històrica sobre la producció, finances, operacions, vendes, finances, inventaris i clients de la companyia.

2.5.3. Models de decisió

Els models de decisió utilitzen per al modelatge de una o un conjunt de decisions, les quals es poden reproduir de manera semblant al llarg del temps. La modelització de decisions agrupa els elements necessaris per a determinar el procés per arribar al millor resultat, que millorarà o optimitzarà el rendiment d'un sistema. [8] Exemples de metodologies de decisió serien:

- Arbres de decisió
- Diagrama SWOT (DAFO): avalua tots els punts forts, febles, oportunitats i amenaces de la decisió.

- Multi votació: Aquest mètode s'utilitza quan molta gent està involucrada en una situació.
- Anàlisi conjunt: Aquesta tècnica ajuda a identificar les preferències dels clients mentre pren decisions. La investigació determina el que la gent realment agrada i valora els vostres serveis.

- Anàlisi de Pareto. Si s'ha de prendre moltes decisions, aquest mètode ajudarà a establir prioritats. Es veuen quines decisions s'han de fer primer afegint una puntuació a totes elles.

- Matriu de decisions. Amb aquest mètode, s'avaluen totes les opcions de la decisió que es vol prendre. En primer lloc, es crea una taula. A la primera columna es posa les opcions de la decisió i en la primera fila es posa tots els factors que poden influir en la seva decisió.

2.5.4. Models Prescriptius

Es una branca relativament nova, permet als usuaris "prescriure" una sèrie de diferents accions possibles i guiar-les cap a una solució. El que pretén aquest tipus de models es assessorar dins de un context determinat. L'anàlisi prescriptiva intenta quantificar l'efecte de les decisions futures per tal d'assessorar sobre possibles resultats abans de prendre decisions. En el millor dels casos, les anàlisis prescriptives prediuen no només què passarà, sinó també per què passarà proporcionant recomanacions sobre accions que aprofitaran les prediccions.

Aquesta tipologia d'anàlisis van més enllà dels models descriptius i predictius recomanant una o més accions possibles. Fonamentalment es preveu múltiples escenaris futurs i permeten als agents que intervenen avaluar una sèrie de resultats possibles basats en les seves accions. Les anàlisis prescriptives utilitzen una combinació de tècniques i eines, com ara normes empresarials, algorismes, aprenentatge automàtic i procediments de modelatge computacional. Aquestes tècniques s'apliquen a l'entrada de molts conjunts de dades diferents, incloent dades històriques i transaccionals, aquests processos s'alimenten de grans volums de dades en temps real.

Aquest models són complexes d'administrar, i la majoria de les empreses encara no les utilitzen en el seu negoci diari. La seva implantació de manera correcta, pot tenir un gran impacte en la manera com les es prenen les decisions en diferents contextos com l'empresarial, públic, etc. A nivell d'exemple les empreses més grans utilitzen amb èxit l'anàlisi prescriptiva per optimitzar la producció, la programació i l'inventari a la cadena de subministrament per assegurar-se que ofereixen els productes adequats en el moment adequat i que optimitzen l'experiència del client.

2.5.5. Models Combinats

Els modelatge combinat és una tècnica predictiva molt per augmentar la precisió dels processos de l'aprenentatge automàtic. Els models "ensemble" o combinats, són processos que utilitzen diversos algorismes d'aprenentatge per obtenir un millor rendiment predictiu del que es podria obtenir de qualsevol dels algorismes d'aprenentatge per si sol, disminuir la variància (bagging), el biaix (boosting) o millorar les prediccions (stacking).. A diferència dels models tradicionals, els models de conjunt d'aprenentatge automàtic consisteixen únicament en un conjunt finit concret de models alternatius, que permeten una estructura molt més flexible a l'hora de implementar models alternatius. [10]

Els mètodes de grup es poden dividir en dos grups:

- *Mètodes seqüencials* on els aprenentatges de base es generen de forma seqüencial (p. ex. AdaBoost).L'objectiu principal dels mètodes seqüencials és explotar la dependència entre els aprenentatges bàsics.
- *Mètodes paral·lels* on els aprenentatges de base es generen en paral·lel (p. ex., Random forest).L'objectiu principal dels mètodes paral·lels consisteix a explotar la independència entre els aprenentatges bàsics, ja que es pot reduir dràsticament l'error utilitzant la mitjana.

La majoria dels mètodes combinats utilitzen un algoritme d'aprenentatge únic base per produir aprenentatges homogenis, és a dir, aprenents del mateix tipus, que condueixen a conjunts homogenis.

També hi ha alguns mètodes que utilitzen aprenents heterogenis, és a dir, aprenents de diferents tipus, que condueixen a conjunts heterogenis. Perquè els mètodes de conjunt siguin més precisos que qualsevol dels seus membres individuals, els aprenents bàsics han de ser tan precisos com sigui possible i tan diversos com sigui possible. [11]

2.5.6. Models Incremental

És una tècnica de modelatge predictiu que modifica directament l'impacte incremental d'un tractament (com ara una acció de màrqueting) sobre el comportament d'un individu o grup. L'objectiu de modelització incremental és trobar clients persuasibles. El modelatge incremental es pot aplicar a qualsevol resultat modelat, com l'efecte del fertilitzant en els rendiments de cultius o l'enviament de missatges de correu electrònic en campanyes polítiques. On el model predictiu tradicional es centra en el resultat, el model incremental es centra en l'eficàcia del tractament. [12]

El model Incremental en altres paraules, pot ajudar a identificar persones que només compraran productes com a conseqüència de rebre un cupó de descompte o un anunci personalitzat. La utilització d'aquests models pot ajudar a les empreses a maximitzar els beneficis mantenint un cost publicitari al mínim. Aquest anàlisi pot ajudar també a detectar possibles pèrdues de clients als que no els agrada que els enviïn cupons o anuncis.

Per exemple, es posa el cas de un negoci que te un apartat de productes per a dones embarassades. És possible que la botiga vulgui predir quins clients probablement estiguin embarassats per enviar cupons promocionals per a productes per a l'embaràs o nadons, el model incremental capacitarà al negoci per elaborar quest tipus de estudi. [13]

Tot i que els models d'elevació poden ser realment beneficiosos, de vegades poden ser difícils d'implementar. El repte més gran és trobar el mètode òptim per modelitzar l'impacte incremental del tractament sobre la resposta de l'individu. [14]

2.5.7. Validació de models

Amb els models creat, es fa necessari comprovar l'actuació del procés complet. Per provar el model d'anàlisi predictiu elaborat, es divideix el conjunt de dades en dos conjunts: els conjunts de dades d'entrenament i el conjunt de dades de test. Aquests conjunts de dades s'han de seleccionar a l'atzar i han de ser una bona representació de la població real. Les dades utilitzades hauran de complir certes característiques per a la correcta validació:

- Utilitzar dades similars per ambdós conjunts de dades.
- Normalment el conjunt de dades d'entrenament és significativament més gran que el conjunt de dades de prova.
- L'ús del conjunt de dades de prova (test) ajuda a evitar errors com ara l'overfiting.
- El model d'entrenament s'executa amb les dades de del conjunt test per veure el rendiment del model.

Alguns científics de dades prefereixen tenir un tercer conjunt de dades que tingui característiques similars a les dels dos primers: un conjunt de dades de validació. La idea és que si utilitzeu activament les dades de prova per perfeccionar el vostre model, haureu d'utilitzar un conjunt independent (tercer) per comprovar la precisió del model. Tenir un conjunt de dades de validació que no s'ha utilitzat com a part del procés de desenvolupament del model, garanteix una estimació neutral de la precisió i l'eficàcia del model. [15]

2.6. Problemàtiques de l'anàlisi predictiu

Malauradament en la pràctica de l'anàlisi predictiu no solament hi ha problemàtiques de caire numèric o de procediments matemàtics. Molts factors intervenen en el desenvolupament d'aquest tipus d'activitats predictives, la planificació, organització, vies d'accés a dades, són per exemple barreres que es poden trobar en les metodologies predictives. Les infraestructures en el sistemes de informació també són un punt rellevant en les problemàtiques, accessos a magatzems de dades, centralització dels magatzems, poca capacitat d'emmagatzematge, etc.

La relació de l'anàlisi de dades amb el món empresarial, porta també obstacles de caràcter corporatiu, ja que les empreses utilitzen aquests mètodes per a objectius econòmics i les correlacions entre variables utilitzades sol ser complexa. A continuació es detalles algunes de les problemàtiques més freqüents. [16]

Disponibilitat de dades
Sovint es complex i costós el fet de trobar dades valides per a ser tractades, o construir models a partir de dades amb poca relació entre elles, el procés de cerca i obtenció de dades a vegades pot ser bastant problemàtic.

Dades brutes
En la recollida de dades, aquestes no venen perfectament estructurades i amb valors clars i concisos, el grans volums de dades solen ser imperfectes, amb atributs on hi falten valors, o registres amb dades redundants. El filtratge i tractament previ de la neteja es és un punt molt important en el procés inicial de l'anàlisi de dades.

Problema de representació
Un cop realitzades les prediccions, la representació gràfica dels resultats pot resultar complicat. La comunicació dels resultats es un pilar fonamental per tal de que les persones que han de llegir o entendre el que ha sortit del procés de predicció puguin treuen les seves pròpies conclusions i siguin fàcilment comprensibles.

Problema de diversificació
La gran capacitat d'emmagatzematge que tenen actualment les empreses fa que tinguin també gran diversitat de informació i dades (tipus, formats, ubicacions). Per tal de que es puguin agrupar i utilitzar aquestes s'han d'alinear en el context en que se'n farà ús, d'aquesta manera s'obtenen agrupacions de dades de qualitat.

Dades perdudes
Com s'ha comentat, les empreses tenen grans repositoris de dades que augmenten dia a dia. El problema de la diversificació, representa un handicap a l'hora d'analitzar les dades i fer un ús de les eines disponibles, provocant alguna vegada transformacions de dades, o que s'hagin de realitzar processos on es perd informació. Aquest efecte te un impacte en la predicció, ja que no s'han utilitzat les dades extretes en la seva totalitat en els conjunts

d'entrenament. No obstant, cada vegada menys es produeix aquest problema degut a l'avanç de les aplicacions d'analítica predictiva.

Integració de dades

Lligat amb les aplicacions que permeten fer una gestió de la informació durant el procés d'anàlisi, hi ha el problema d'integració. Moltes de les aplicacions són de proveïdors independents entre ells pel que a vegades es fa complexa la integració amb aplicacions empresarials i el flux de dades que generen les empreses. En el casos mes complexes es pot arribar a tenir pèrdues de importants volums de dades, cosa que faria ineficient el procés de integració. Aquest es un problema important a tenir en compte ja que la gestió per integrar diferents tecnologies com CRMs, ERPs, o d'altres influeix a nivell corporatiu en el fet de ser mes o menys eficient, controlar processos empresarials, etc.

Velocitat i escalabilitat

Amb que les empreses generin grans volums de dades i que els processos d'anàlisi estiguin controlats no n'hi ha prou. Aquests processos predictius requereixen actualment de grans velocitats de transmissió i processament. Es important doncs optimitzar processos, i escollir be les eines per tal de treballar amb una velocitat i transferència de dades adequada. S'observa com una de les qüestions més importants a tenir en compte en aquest punt és tenir la capacitat de enviar les dades a una velocitat optima, independentment del volum.

Evolució de la predicció

En la formulació de resultats de caràcter predictiu, es possible que les variables variïn d'un dia per l'altre, que com a conseqüència modificaria les prediccions elaborades. Les empreses amb ús d'aquests mètodes de predicció han de preveure aquests "girs de volant" i tenir plans d'accions per a possibles desviacions, amb les mesures corresponents a aplicar.

2.7. Tècniques de l’anàlisi predictiu i modelatge estadístic

Mitjançant el tractament de volums de dades es poden explicar histories complexes. Per extreure informació i explicar el relat corresponent es requereix de diverses metodologies, que en qualsevol cas no sempre són les mateixes.

Aquests mètodes matemàtics i de càlcul, mitjançant un procés iteratiu, desenvolupen el model mitjançant un conjunt de dades d'entrenament, i després aquest es provat i es validat per determinar la precisió a l’hora de realitzar prediccions. Sovint es solent utilitzar varis mètodes per a un mateix conjunt de dades, d’aquesta manera s’enfoca el problema des de diferents angles i comparant-los entre ells, es troba la solució més eficient. A continuació s’exploren diferents tècniques predictives.

2.7.1. Reducció Dimensional

Analitzar totes les variables de grans conjunts de dades en detall es una tasca complicada, per a realitzar aquest tipus de processos podria suposar una gran inversió de recursos, i sobretot cost computacional. Per tant es necessita una millor manera de tractar els grans volums de dades, de manera que podrem extreure ràpidament patrons i coneixements. Alguns avantatges de la reducció de la dimensionalitat serien:

- Com menys dades, menys espai d’emmagatzematge es necessita.
- Menys dimensions condueixen a menys temps de
- Alguns algorismes no funcionen bé quan tenim dimensions grans. Per tant, reduir aquestes dimensions ha de passar perquè l'algoritme sigui útil
- Elimina les funcions redundants. Es a dir funcions que utilitzaran variables amb alta correlació
- Millor visualització de les dades. Reduir el l’espai a 2D o 3D ens permetrà extreure gràfics i observar patrons amb més claredat

La reducció de la dimensionalitat en conjunts de dades es un factor a tenir en compte a l’hora reduir el nombre de variables aleatòries a considerar. Aquests algoritmes de reducció dimensional obtenen un conjunt de variables principals. La reducció de la dimensionalitat es pot fer de dues maneres diferents:

- Selecció de funcions, es a dir conservant les variables més rellevants del conjunt de dades original.
- Reducció de la dimensionalitat o extracció de funcions, es a dir trobar un conjunt més petit de variables noves, cada una de les quals és una combinació de les variables d'entrada, que conté bàsicament la mateixa informació que les variables d'entrada.

A continuació es mostren diverses tècniques de reducció de la dimensionalitat, i se n'expliquen les més importants a nivell d'ús d'anàlisi predictiu. [17]

Principal Component Analysis (PCA)

El PCA és una tècnica que ajuda a extreure un nou conjunt de variables a partir d'un gran conjunt de variables existents. Aquestes variables acabades d'extreure es diuen Components Principals. Un component principal és una combinació lineal de les variables originals.

Els components principals s'extreuen de manera que el primer component principal explica la variància màxima del conjunt de dades, el segon component principal intenta explicar la variància restant del conjunt de dades i no està correlacionada amb el primer component principal, el tercer component principal intenta explicar la variància que no s'explica pels dos primers components principals, etc. [18]

Independent Component Analysis (ICA)

L'anàlisi de components independents (ICA) es basa en la teoria de la informació i és també una de les tècniques de reducció de dimensionalitat més utilitzades. La gran diferència entre PCA i ICA és que la PCA busca factors no correlacionats mentre que l'ICA busca factors independents. Si dues variables no estan correlacionades, significa que no hi ha cap relació lineal entre elles. Si són independents, vol dir que no depenen d'altres variables. Per exemple, l'edat d'una persona és independent del que aquesta persona menja, o la quantitat de televisió que mira.

Aquest algorisme suposa que les variables donades són barreges lineals d'algunes variables latents desconegudes. També assumeix que aquestes variables latents són mútuament independents, és a dir, que no depenen d'altres variables i, per tant, es denominen components independents de les dades observades.

Altres tècniques no tan comunes de reducció de dimensionalitat serien:

- Principal Component Regression (PCR)
- Linear Discriminant Analysis (LDA)
- Flexible Discriminant Analysis (FDA)
- Quadratic Discriminant Analysis (QDA)
- Non-negative matrix factorization (NMF or NNMF)
- Sammon Mapping

2.7.2. Regressió

Les tècniques de regressió són dels pilars fonamentals de l'anàlisi predictiu. Aquest tipus de tècniques es centren en construir una equació matemàtica com a model de representació de les interaccions entre variables. Les tècniques de regressió lineals i logístiques solen ser els algoritmes amb més ús del modelatge predictiu. Hi ha un ventall molt ampli de formes de regressió, que es poden efectuar. Cada model es centra en les seves pròpies condicions específiques, i les pròpies relacions entre variables.

A nivell més tècnic, l'anàlisi de regressió és una tècnica de modelatge predictiu que investiga la relació entre una variable (variable) i variable dependent (predeterminada). Aquesta tècnica s'utilitza per a la predicció, el modelatge de sèries temporals i la relació de l'efecte causal entre les variables. Per exemple, la relació entre la conducció ràpida i la quantitat d'accidents de trànsit per part d'un conductor la millor manera de ser estudiada és a través de la regressió. A continuació és mostra un exemple gràfic de regressió lineal, on s'exemplifica un model de regressió que relaciona l'alçada de un infant respecte el temps, en mesos. D'aquesta manera al gràfic s'observa la relació de l'edat vers l'alçada de l'infant. [19]

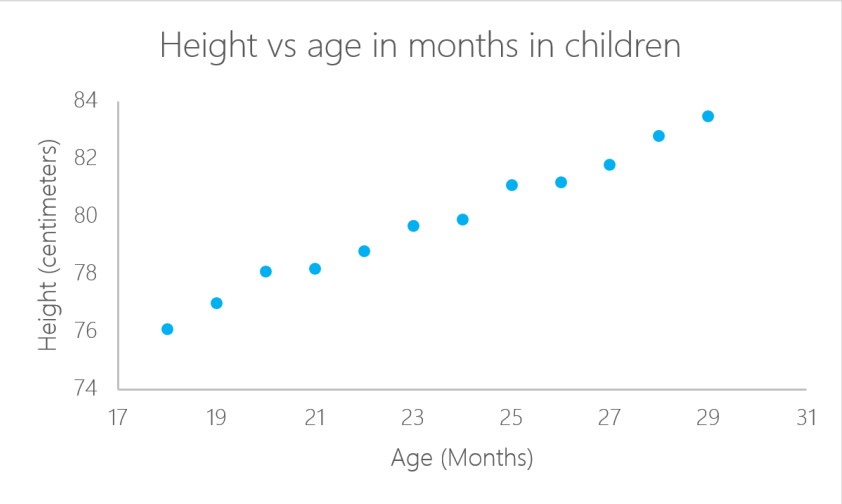


Figura 2. Regressió lineal. [54]

Hi ha diversos tipus de tècniques de regressió disponibles per fer prediccions. Aquestes tècniques es basen principalment en tres mètriques (nombre de variables independents, tipus de variables dependents i forma de línia de regressió). A continuació s'expliquen les dues tècniques de regressió més utilitzades:

Regressió Lineal

La regressió lineal és una de les tècniques de modelització més utilitzades, degut a la seva baixa complexitat sol ser d'entre les primeres tècniques que utilitza la gent per iniciar-se en el modelatge predictiu. El funcionament d'aquesta tècnica es basa en que la variable dependent és contínua, la variable o variables independent pot ser contínua o discreta, i la línia de regressió és lineal. [20]

La regressió lineal estableix una relació entre la variable dependent (Y) i una o més variables independents (X), utilitzant la línia entre els punts que s'ajusti millor (anomenada també línia de regressió). Es representa mitjançant una equació del tipus Y = a + b * X + e, on a és la ordenada d'origen, b és pendent de la línia i e és terme d'error. Aquesta equació es pot utilitzar per predir el valor de la variable objectiu en funció de la variable predictora especificada. [21]

La diferència entre la regressió lineal simple i la regressió lineal múltiple és que, la regressió lineal múltiple té variables independents (> 1), mentre que la simple regressió lineal té només 1 variable independent.

Quan es te un gran volum de mostres i es complica l'ajust de la línia de regressió, s'obté la millor línia d'ajust mitjançant el mètode Least Square. Aquest calcula la línia més adequada per a les dades observades minimitzant la suma dels quadrats de les desviacions verticals de cada punt de dades a la línia. Com que les desviacions són el primer quadrat, quan s'hi afegeix, no hi ha cancel·lació entre valors positius i negatius. El rendiment del model es pot avaluar mitjançant la mètrica R-square. Punts importants de la regressió lineal:

- Hi ha d'haver una relació lineal entre variables independents i dependents
- La regressió múltiple pateix multicolinearialitat, autocorrelació, heteroskedasticitat.
- La regressió lineal és molt sensible als "Outliers". Pot afectar terriblement la línia de regressió i eventualment els valors previstos.
- La multicolinearialitat pot augmentar la variació de les estimacions del coeficient i fer que les estimacions siguin molt sensibles als canvis menors del model, provocant que les estimacions dels coeficient siguin inestables.

Regressió Logística (Logistic Regression)

La regressió logística és l'anàlisi de regressió adequada per dur a terme quan la variable dependent és binària. Com tots els anàlisis de regressió, la regressió logística és una tècnica d'anàlisi predictiu. La regressió logística s'utilitza per descriure les dades i explicar la relació entre una variable binaria dependent i una o més variables nominals, ordinaris, intervals o de nivell de relació.

Matemàticament, un model logístic binari té una variable dependent amb dos valors possibles, com passar / fracassar, guanyar / perdre, viure / morta o sana / malalt; Aquests són representats per una variable indicadora, on els dos valors estan etiquetats com "0" i "1". [22]

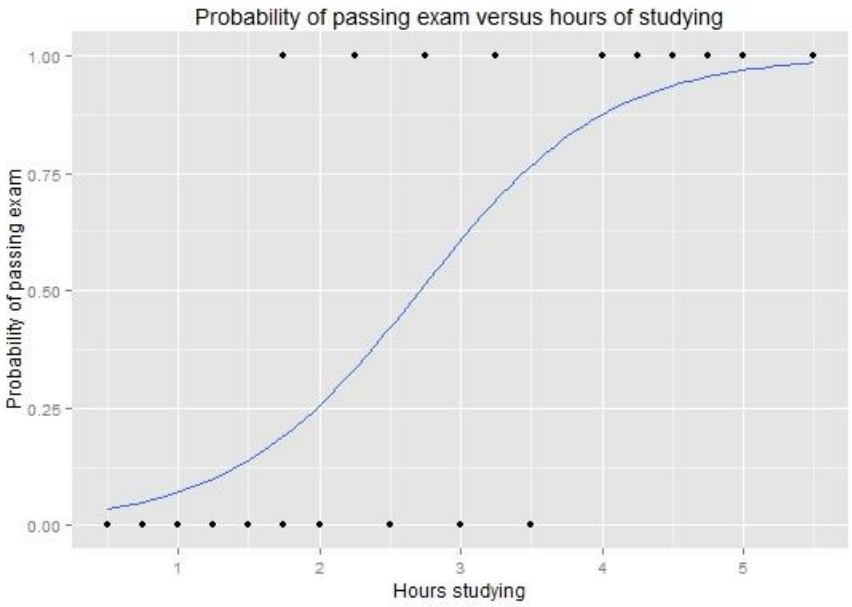


Figura 3. Regressió Logística. [55]

Altres tècniques de regressió no tant comunes podrien ser:

- Regressió Mínima Quadrangular Ordinària (OLSR)
- Regressió Esglaonada (Stepwise Regression)
- Multivariate Adaptive Regression Splines (MARS)
- Locally Estimated Scatterplot Smoothing (LOESS)

2.7.3. Arbres de decisió

Els algoritmes d'aprenentatge basats en l'arbre de decisió es consideren un dels millors i més utilitzats mètodes d'aprenentatge supervisats. Els mètodes basats en arbre permeten models predictius amb gran precisió, estabilitat i facilitat d'interpretació. Són adaptables a la resolució de qualsevol tipus de problema (classificació o regressió). [23]

L'arbre de decisió és un tipus d'algorisme d'aprenentatge supervisat utilitzat principalment en problemes de classificació. Funciona tant per a variables d'entrada i sortida categòriques com contínues. En aquest tipus de tècnica es divideix la població o mostra en dos o més conjunts homogenis que es basin en el diferenciador més significatiu de les variables d'entrada. Els tipus d'arbres de decisió es basen segons la variable objectiu proposada. [24]
Poden ser de dos tipus:

- Arbre de decisió variable categòrica, que té una variable de destinació categòrica
- Arbre de decisió de variable contínua, on l'arbre de decisió té una variable de destinació contínua.

Avantatges:

- Facilitat de comprensió, gràficament són intuïtius i fins i tot persones alienes a l'anàlisi de dades poden comprendre'ls, els usuaris poden relacionar fàcilment la seva hipòtesi.
- Es de gran ajuda en l'exploració de dades. Identificar les variables i les relacions entre aquestes.
- En comparació a altres tècniques de modelatge, no es requereix un filtre tant exhaustiu de neteja de dades.
- El tipus de dades no genera restriccions sobre el model, es pot gestionar variables numèriques i categòriques.
- Es un mètode no paramètric, això significa que els arbres de decisió no tenen cap hipòtesi sobre la distribució espacial i l'estructura del classificador.

Inconvenients:

- Sobre-ajust(Over-fitting), el sobre-ajust de les dades es un dels grans problemes d'aquest model. Es pot solucionar resolent restriccions sobre els paràmetres del model i la poda.
- Si es treballa amb variables contínues es perd informació quan es categoritza les variables ne diferents categories. [25]

**S'exemplifica l'algoritme C5.0 en els apartats 3.1.1 i 3.1.2 amb les eines R i Python.*

Algunes tècniques que utilitzen els arbres de decisió podrien ser:

- Arbre de classificació i regressió (CART)
- Iterative Dichotomiser 3 (ID3)
- C4.5
- C5.0
- M5
- Chi-squared Automatic Interaction Detection (CHAID)
- Decision Stump
- Arbres de decisió condicionals

2.7.4. Estadística Bayesiana

Les estadístiques bayesianes són una tècnica estadística centrada en aplicar la probabilitat en problemes estadístics. Proporciona eines matemàtiques modificar les nostres creences sobre esdeveniments aleatoris per tal de veure noves dades sobre aquests esdeveniments. La inferència bayesiana interpreta la probabilitat com una mesura de confiança que un individu pot tenir sobre el succés d'un esdeveniment en particular.

Es a dir, es pot tenir una creença prèvia sobre un esdeveniment en concret, però si apareixen noves variables, es possible que les nostres creences variïn en funció d'aquestes. Les estadístiques bayesianes, proporcionen una metodologia que incorpora les creences de un individu, juntament amb proves anteriors, per tal de produir creences posteriors, i així actualitzar de manera racional les creences subjectives en funció de les dades aportades. Es proporciona un exemple perquè quedi el concepte més clar. [26]

Es suposa que de 4 carreres del campionat del món de motociclisme, en Marc Marquez n'ha guanyat 3 mentre que en Valentino Rossi només una. Per tant, si s'hagués de apostar per qui guanyaria la pròxima carrera, segurament la majoria apostaríem per en Marc. Però i si et diguessin que va ploure la vegada que va guanyar en Valentino i també una vegada que va guanyar en Marc, per qui apostaries?

Ara veiem com la probabilitat de que guanyi en Valentino ha augmentat de manera sorprenent. Però quant ha augmentat? De manera genèrica, l'estudi d'aquest tipus de situacions implicaria l'estadística bayesiana per tal de poder explicar els possibles successos futurs.

Naive Bayes

Es tracta d'una tècnica de classificació basada en el teorema de Bayes amb una hipòtesi d'independència entre els predictors. Simplificant l'explicació, un classificador de Naive Bayes suposa que la presència d'una característica particular en una classe no està relacionada amb la presència d'una altra característica. Per exemple, es pot considerar que una fruita és una poma si és de color vermell, rodona i d'uns 6 centímetres de diàmetre.

Fins i tot si aquestes característiques depenen de l'altra o de l'existència de les altres característiques, totes aquestes propietats contribueixen independentment a la probabilitat que aquesta fruita sigui una poma i per això es coneix com a "Naive". [27]

El model Naive Bayes és fàcil de construir i és especialment útil per a conjunts de dades molt grans. Juntament amb la simplicitat, Naive Bayes és conegut per

superar mètodes de classificació fins i tot molt sofisticats. Els avantatges i inconvenients d'utilitzar aquest tipus de tècnica serien els següents:

Avantatges:

- És fàcil i ràpid predint la classe de conjunt de dades de prova. També funciona bé en la predicció de classes múltiples.
- Funciona millor en els casos en que les variables d'entrada siguin categòriques en comparació amb les variables numèriques.

Inconvenients:

- Si la variable categòrica té una categoria (en el conjunt de dades de la prova), no observada prèviament en el conjunt de dades de formació, el model assignarà una probabilitat zero i no podrà fer una predicció. Sovint es coneix com a "freqüència zero".
- És un mal estimador. Al classificar un punt de dades determinat, la tècnica calcula primer la probabilitat en que creu que el punt de dades pertanyerà a cada etiqueta de la classe possible. Es produeix una classificació de l'etiqueta segons la selecció de la classe associada on la probabilitat és gran. Llavors la correlació entre la probabilitat i l'etiqueta de la classe, no es correlacionen amb la confiança de classificació
- Mala assumpció de característiques independents. El model Naive Bayes s'anomena "Naive" precisament perquè assumeix que les característiques són independent. Tanmateix, el model pot aconseguir una precisió forta de classificació, fins i tot en els casos en què la seva hipòtesi de la independència condicional és significativament inexacta.

**S'exemplifica el mètode Naive Bayes en l'apartat 3.1.2 amb l'eina Python.*

Altres algoritmes relacionats amb la tècnica de l'estadística Bayesiana serien:

- Gaussian Naive Bayes
- Multinomial Naive Bayes
- Bayesian Network (BN)
- Bayesian Belief Network (BBN)
- Averaged One-Dependence Estimators (AODE)

2.7.5. Xarxes Neuronals

Les Xarxes neuronals o també anomenades Xarxes Neuronals Artificials, es podrien considerar com una representació artificial del sistema nerviós humà. Fent un símil amb aquest, cada neurona està connectada amb moltes altres neurones i fa d'entrada d'informació per aquestes a través de les dendrites. Quan s'excita a l'entrada, la neurona processa la informació a través dels axiomes a les altres neurones connectades. [28]

En les xarxes neuronals artificials, el funcionament es semblant. Una neurona recull informació de totes les entrades i realitza una operació sobre elles, i posteriorment transmet la sortida a altres neurones que estan connectades amb ella. La xarxa neuronal es divideix en tres tipus de capes:

- Capa d'entrada: les observacions s'alimenten a través d'aquestes neurones
- Capes amagades: són les capes intermèdies entre entrada i sortida que ajuden a la Xarxa Neural a conèixer les complicades relacions que hi intervenen.
- Capa de sortida: la sortida final s'extreu de les dues capes anteriors.

Les xarxes neuronals ajuden a agrupar i classificar. Es poden considerar com una capa de agrupació i classificació sobre de les dades que s'emmagatzemen i es gestionen. Aquestes ajuden a agrupar dades no etiquetades d'acord amb les similituds entre les entrades d'exemple i classifiquen les dades quan tenen un conjunt de dades etiquetat per entrenar.

Les xarxes neuronals també poden extreure característiques administrades a altres algoritmes per agrupament i classificació, per la qual cosa es pot pensar en xarxes neuronals profundes com a components d'aplicacions de màquina-aprenentatge més grans que impliquen algorismes d'aprenentatge, classificació i reforç.

Les xarxes neuronals funcionen molt bé per:

- Cercar associacions o descobrir regularitats en un conjunt de patrons;
- Analitzar grans volums de dades on el nombre de variables o la diversitat de dades són molt grans;
- Les relacions entre variables estan poc definides o són difícils de descriure adequadament amb enfocaments convencionals.

En alguns casos, les xarxes neuronals es consideren com a "força bruta", perquè són efectius, però en alguns aspectes són ineficients en el seu enfocament de modelització, ja que no poden fer suposicions sobre dependències funcionals entre la 'entrada i la sortida. [29] [30]

Perceptró

En l'aprenentatge automàtic, el perceptró és un algorisme per a l'aprenentatge supervisat dels classificadors binaris. Un classificador binari és una funció que pot decidir si una entrada, representada per un vector de nombres, pertany a alguna classe específica. Es tracta d'un tipus de classificador lineal, és a dir, un algoritme de classificació que fa les seves prediccions basades en una funció predictora lineal que combina un conjunt de pesos amb el vector de característiques.

El valor de f (x) (0 o 1) s'utilitza per classificar x com una instància positiva o negativa, en el cas d'un problema de classificació binària. Si b és negatiu, la combinació ponderada d'entrades ha de tenir un valor positiu superior a |b| per empènyer la neurona classificadora sobre el llindar 0.

L'algoritme perceptró també es denomina perceptró d'una capa, per distingir-lo d'un perceptró multicapa, que és un nom erroni per a una xarxa neuronal més complexa. Com a classificador lineal, el perceptró d'una sola capa és la xarxa neuronal més senzilla d'alimentació.

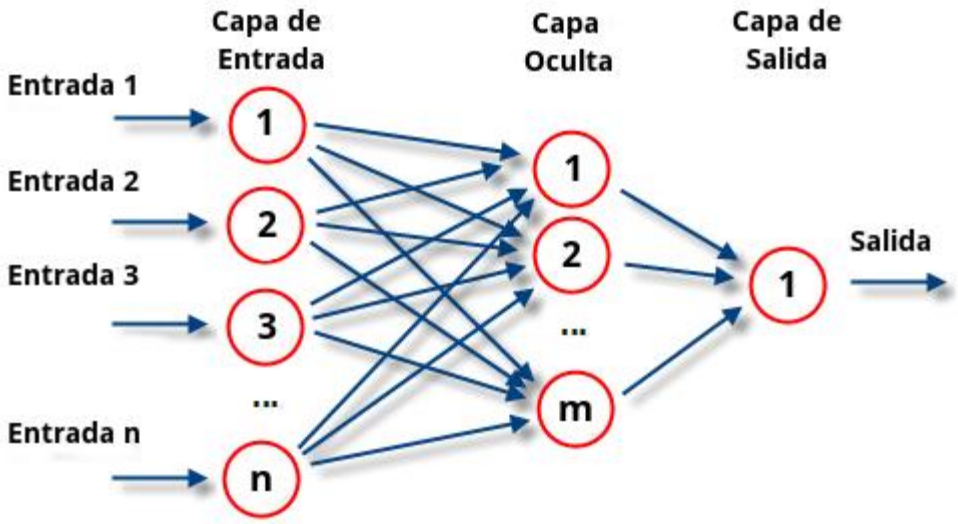


Figura 4. Perceptró Multicapa.[56]

Un exemple del que es podria extreure amb el perceptró seria el preu de l'habitatge segons una zona concreta. En el model es farien entrar les variables d'entrada i segons certes classes s'anirien classificant les mostres.

Altres algoritmes relacionats amb la tècnica de Xarxes Neuronals serien:

- Radial Basis Function Network (RBFN)
- Back-Propagation
- Hopfield Network

2.7.6. Deep Learning

La majoria dels mètodes de Deep Learning utilitzen arquitectures de xarxes neuronals, per la qual cosa els models de deep learning es denominen sovint xarxes neuronals profundes. El terme "deep" sol referir-se al nombre de capes ocultes a la xarxa neuronal. Les xarxes neuronals tradicionals només contenen 2-3 capes ocultes, mentre que les xarxes d'aprenentatge profundes poden tenir fins a 150 capes. Els models de deep learning es formen mitjançant l'ús de grans conjunts de dades etiquetades i arquitectures de xarxes neuronals que aprenen característiques directament de les dades sense necessitat d'extracció de funcions manuals. [31]

Alguna de les aplicacions dels models de Deep Learning seria ajudar a millorar la traducció automàtica del text utilitzant xarxes apilades de xarxes neuronals i permetent traduccions d'imatges.

Alguns algoritmes de la tècnica Deep Learning serien:
- Stacked Auto-Encoders
- Deep Boltzman Machine (DBM)
- Deep Belief Networks (DBM)
- Convolutional Neural Network (CNN)
- Stacked Auto-Encoders

2.7.7. Modelatge Combinat

Els mètodes combinats són algoritmes d'aprenentatge computacional que combinen diversos models base per produir un model predictiu òptim. Aquesta tècnica que combina diferents algoritmes és la millor manera de millorar el rendiment del model predictiu, ja que s'agafen les millor característiques dels algoritmes implicats. [32]

Boosting

L'algoritme "Boosting" utilitza un seguit de processos per tal de enfortir els aprenents febles i convertir-los en aprenents forts. Es fa l'explicació de l'algoritme mitjançant un exemple de identificació de correu brossa. Per a classificar un correu com a brossa primer de tot s'ha de identificar si aquest el considerem brossa o no. Considerarem el següent per a classificar:

- El correu electrònic només té un fitxer d'imatge (imatge promocional), és SPAM
- El correu electrònic només té enllaços, és SPAM
- El cos de correu electrònic consisteix en frases com "Has estat seleccionat per un concurs xxx", és SPAM
- Correu electrònic des de un domini oficial "uoc.edu", no és SPAM

- Correu electrònic des d'una font coneguda, no és SPAM

Amb aquestes indicacions s'han definit diverses regles per classificar un correu electrònic com a "spam" o "no correu brossa". Però individualment aquestes regles no tenen prou força per a classificar els correus electrònics, doncs s'anomenaran aprenents dèbils. Per convertir l'aprenent feble en un aprenentatge fort, combinarem la predicció de cada alumne feble utilitzant mètodes com l'ús de la mitjana o ponderació. A dalt, s'ha definit 5 aprenents febles.

D'aquests 5, 3 es vota com a 'SPAM' i 2 són votats com 'No és un SPAM'. En aquest cas, de manera predeterminada, es considerarà un correu electrònic que contingui les característiques esmentades com SPAM perquè es te un vot superior (3) per a 'SPAM'.

Per tant després de posar l'exemple de boosting, s'observa que per trobar una regla feble, s'apliquen algorismes d'aprenentatge bàsic amb una distribució diferent. Cada algoritme d'aprenentatge a base de temps s'aplica, genera una nova regla de predicció feble. Es tracta d'un procés iteratiu. Després de moltes iteracions, l'algorisme d'enfortiment combina aquestes regles febles en una única regla de predicció sòlida.

Finalment, combina els resultats de l'alumne feble i crea un gran aprenent que millora el poder de predicció del model. La promoció paga un enfocament més alt en exemples que no estan classificats o tenen errors més grans precedint regles febles. El boosting en general disminueix l'error de biaix i crea models predictius sòlids.

Alguns algoritmes de metodologia boosting serien:

- AdaBoost (Adaptive Boosting)
- Gradient Boosting Machines (GBM)
- Gradient Boosting Regression Trees (GBRT)

Bagging

La tècnica del Bagging s'utilitza quan l'objectiu és reduir la variància d'un arbre de decisió. La idea és crear diversos subconjunts de dades a partir de la mostra de d'entrenament seleccionada a l'atzar amb substitució. Cada grup de dades dels subconjunts s'utilitza per entrenar els arbres de decisió. Com a resultat, s'acaba amb un conjunt de models diferents. S'utilitzen les mitjanes de les prediccions de tots els arbres de decisió, això fa que sigui més robust que un arbre de decisió.

Random Forest
La tècnica dels Random Forest és un algoritme d'aprenentatge supervisat i és una extensió dels Arbres de decisió. Aquest algoritme construeix múltiples arbres de decisió i els combina per obtenir una predicció més precisa i estable, de manera que redueix la correlació entre els classificadors individuals. [33]

Altres algoritmes relacionats amb la metodologia combinada serien:
- Bootstraped Aggregation (Bagging)
- Stacked Generalization (Blending)

2.7.8. Modelatge Basat en Instancies

El modelatge basat en instancies és un és tracta de un grup d'algoritmes que realitza una comparació de instancies noves o problemàtiques amb instancies observades en el conjunt d'entrenament, guardades en la memòria. S'anomenen basats en instancies ja que generen les hipòtesis des de les mateixes dades del conjunt d'entrenament, per tant si el volum de dades augmenta la complexitat de computació de la hipòtesis augmentarà també.

Un avantatge que l'aprenentatge basat en instància té sobre altres mètodes d'aprenentatge automàtic és la seva capacitat d'adaptar el seu model a dades no vistes prèviament. A continuació es mostra el funcionament de l'algoritme kNN. [34]

K-Nearest Neighbours (kNN)
L'algoritme kNN utilitza la 'similitud de característiques' per predir valors de qualsevol nou punt de dades. El nou punt s'assigna un valor basat en el grau d'aproximació als punts del conjunt d'entrenament. A continuació es realitza una explicació simplificada pas a pas de l'algoritme.

1. En primer lloc, es calcula la distància entre el punt nou i cada punt d'entrenament. Els medotes de càlcul més comuns són la distancia Euclidiana i Manhattan per variables continues o Hamming per variables categòriques.
2. Es seleccionen els punts de dades k més propers (segons la distància).
3. La mitjana d'aquests punts de dades és la predicció final del nou punt.

Altres algoritmes relacionats amb la metodologia basada en instancies serien:
- Learning Vector Quantization (LVQ)
- Locally Weighted Learning (LWL)
- Self-Organizing Map (SOM)

2.7.9. Clustering

La tècnica del Clustering és la tasca de dividir la població o els punts de dades en diversos grups, de manera que els punts de dades dels mateixos grups són més semblants a altres punts de dades del mateix grup que els d'altres grups. En paraules simples, l'objectiu és segregar grups amb trets similars i assignar-los a clústers.

El model clústering es pot dividir en dos tipologies de grups:

- Hard Clúster, on cada punt de dades pertany a un clúster completament o no.
- Soft Clúster, on en lloc de posar cada punt de dades en un clúster diferent, s'assigna una probabilitat o probabilitat que aquestes dades s'assignin en aquells clústers. [35]

A continuació s'explica el funcionament de l'algoritme K-Means.

K-Means
El K-means és un algoritme d'agrupació en busca en cada iteració els màxims grups de dades agrupats. L'execució de l'algoritme es realitza seguint els següents passos: [36]

1. Especificar el nombre desitjat de clústers K: Escollim k = 2 per a "x" punts de dades en l'espai en 2-D.
2. S'assigna de manera aleatòria cada punt de dades a un clúster.
3. Es calculen els centroides del clúster.
4. Es reasignen els punts al centroide del clúster més proper.
5. Es torna a calcular els centroides dels clústers amb les reassignacions.
6. Es repeteixen els passos 4 i 5 fins que no hi hagi millors opcions de clustering.

**S'exemplifica l'algoritme kNN i K-means els apartats 3.1.1 i 3.1.2 amb les eines R i Python.*

Altres algoritmes relacionats amb la metodologia Clustering serien:

- K-Means
- Expectation Maximization
- Hierarchical Clustering

2.7.10. Selecció de Models i Avaluació

La selecció i de models és la tasca de seleccionar un model estadístic a partir d'un conjunt de models candidats, donat un conjunt de dades. La selecció de models també implica però casos de integració de models, en que dos o més algoritmes intervenen per a resoldre un mateix objectiu. Hi ha la teoria de la navalla de Ockham, en desenvolupa el fet que el criteri de selecció de el millor model d'aprenentatge sol ser el model més simplificat.

El procés de selecció de models es una tasca fonamental per optimitzar el rendiment de la solució aportada. Les tècniques de selecció de models poden donar també un feedback de probabilitat del model en si mateix. El biaix i la variància són mesures importants també a l'hora de construir estimadors qualitatius, l'eficiència també ha d'entrar en consideració. A continuació s'explora l'algoritme de validació creuada. [37]

Cross-validation: evaluating estimator performance
El mètode Cross Validation és una tècnica que consisteix a reservar una mostra particular d'un conjunt de dades sobre el qual no s'entrena en el model i posteriorment, provar-la en el model abans de donar-lo com a finalitzat per computar la seva actuació. Aquests són els passos implicats en la validació creuada:

1. Reservar un conjunt de dades de la mostra total
2. Realitzar l'entrenament amb el conjunt de dades restant i el procés algorítmic escollit
3. Utilitzar la mostra reservada del conjunt inicial per realitzar la validació.

Amb el tercer pas, utilitzant el conjunt de dades reservat ajuda a mesurar l'eficàcia i rendiment del model seleccionat. El resultat positiu d'aquest procés de validació implica que el model es vàlid i es pot seguir endavant amb l'anàlisi predictiu

Les mètriques d'avaluació del model s'utilitzen per avaluar la bondat d'ajust entre model i dades, comparar diferents models, en el context de la selecció de models i predir com es preveu que les prediccions (associades a un model i conjunt de dades específics) siguin precises. [38] [39]

Intervals de confiança
Els intervals de confiança s'utilitzen per avaluar la fiabilitat d'una estimació estadística. Els intervals de confiança amples fan que un model sigui pobre o que les dades siguin molt sorolloses si els intervals de confiança no milloren canviant el model.

Altres models de selecció i avaluació podrien ser:

- Tuning the hyper-parameters of an estimator
- Model evaluation: quantifying the quality of predictions
- Model persistence
- Validation curves: plotting scores to evaluate models
- Matriu de confusió

2.7.11. Regularització

La regularització és una tècnica que modifica lleugerament l'algoritme d'aprenentatge de tal manera que el model generalitza millor. Això a la vegada millora el rendiment del model en les dades que no es veuen. La regularització ajuda a reduir el problema de l'Overfitting. Alguns algoritmes que utilitza la tècnica de regularització són:

- Ridge Regression
- Least Absolute Shrinkage and Selection Operator (LASSO)
- Elastic Net
- Least Angle Regression (LARS) [40]

2.7.12. Sistema de regles (Rule Based System)

Els sistemes basats en regles són un tipus simple d'intel·ligència artificial, que utilitzen una sèrie d'afirmacions IF-THEN que guien a una computadora per arribar a una conclusió o recomanació. Existeixen dos components principals clau, un conjunt de dades sobre una situació i un conjunt de dades sobre com tractar els fets.

El conjunt de fets, es coneix també com a base del coneixement. Aquests fets són una combinació de dades i condicions. Els conjunts de regles o motor de regles, descriuen la relació entre afirmacions IF i THEN. Amb aquests dos conceptes bàsics, és possible construir un sistema bàsic d'intel·ligència artificial com una eina per recomanar l'elecció de sabates en un dia concret. Els fets utilitzats en aquest cas podrien incloure "És un dia de la setmana", "És fred", "Està plovent". Es poden construir unes regles perquè una d'intel·ligència artificial conclogui que l'elecció de les sabates, poden ser un botes d'aigua i un paraigües. [41]

Algunes tècniques que utilitza el Sistema basat en regles són:

- One Rule (OneR)
- Zero Rule (ZeroR)
- Repeated Incremental Pruning to Produce Error Production (RIPPER)

2.8. Aprenentatge Computacional (Machine Learning)

L'aprenentatge computacional o automàtic, es una branca de la intel·ligència artificial, que utilitza tècniques estadístiques com xarxes neuronals, xarxes bayesianes, SVM, K-neighbours, etc. Perquè els sistemes informàtics tinguin la capacitat "d'aprendre" a partir de conjunts de dades proporcionats, i millorar l'actuació de la tasca al qual se li aplica de manera progressiva.

Aquesta disciplina explora l'estudi i la construcció d'algorismes que poden aprendre i realitzar prediccions sobre les dades. Aquests algoritmes que conformen l'aprenentatge automàtic, tenen la capacitat de superar-se a si mateixos mitjançant l'execució de certes instruccions especifiques i prenent decisions a partir d'aquests processos de transformació executats. [42]

2.8.1. Tipus d'aprenentatge

A nivell general es pot sub-dividir en 4 tipus d'algoritmes l'àmbit de l'aprenentatge automàtic.

Aprenentatge supervisat
Aquest algoritme consisteix en una variable de resultat que s'ha de predir a partir d'un determinat conjunt de de dades d'entrades. Mitjançant aquest conjunt de variables, es genera una funció que assigna entrades a les sortides desitjades. Es un procés de formació continua fins que el model aconsegueix un nivell de precisió desitjat en les dades d'entrenament. Per tant de manera més simplificada es podria dir que es te coneixement de la resposta que hauria de donar el procés algorítmic amb les variables d'entrada. Exemples de tècniques d'aprenentatge supervisat són: Regressió, Arbres de decisió, Random Forests, kNN, Regressió logística, etc.

Aprenentatge no supervisat
En aquest algoritme, no es te cap variable objectiu o cap resultat per predir. S'utilitza per agrupar la població en diferents grups, també s'utilitza àmpliament per segmentar els clients en diferents grups per a una intervenció específica. Els algoritmes que intervenen en aquest tipus d'aprenentatge són més de categorització. Algun exemple d'aprenentatge no supervisat: K-means.

Aprenentatges semi-supervisats
Els algorismes d'aprenentatge automàtic semi-supervisats es troben entre l'aprenentatge supervisat i no supervisat, ja que utilitzen dades etiquetades i no etiquetades en l'entrenament, normalment es fa servir una petita quantitat de dades etiquetades i una gran quantitat de dades no etiquetades en els processos d'aprenentatge. Els sistemes que utilitzen aquest mètode són capaços de millorar considerablement la precisió de l'aprenentatge. En general,

l'aprenentatge semi-supervisat tria quan les dades etiquetades han adquirit recursos qualificats i rellevants per entrenar-lo.

Aprenentatge de reforç
El seu funcionament és a partir de màquines capacitades per prendre decisions especifiques. El procés és el següent, la màquina està exposada a un entorn on s'entrena contínuament amb prova i error. Aquesta màquina aprèn de l'experiència passada i intenta captar el millor coneixement possible per prendre decisions empresarials precises. Com els humans, que aprenen per interacció, l'aprenentatge per reforç és una aproximació computacional. Un possible exemple seria el següent.

Es considera el cas de aprendre a caminar, on el nen és un agent que tracta de manipular l'entorn (que és la superfície on camina) prenent accions (caminant) i ell / ella tracta d'anar d'un estat (a cada pas que ell / ella pren) a una altra. El nen obté una recompensa (diguem xocolata) quan realitza un èxit en la tasca (fent un parell de passos) i no rebrà cap xocolata (una recompensa negativa) quan no pugui caminar. Es tracta d'una descripció simplificada d'un problema d'aprenentatge de reforç.

Aquest bucle d'aprenentatge per reforç produeix una seqüència d'estat, acció i recompensa. La idea central de l'aprenentatge de reforç es basa en la idea de la hipòtesi de recompensa. Tots els objectius es poden descriure mitjançant la maximització de la recompensa acumulada esperada. És per això que en l'Aprenentatge de reforç, per tenir el millor comportament, s'ha de maximitzar la recompensa acumulada esperada.

2.8.2. Comparativa entre tipus d'aprenentatges

Aprenentatge Supervisat vs. Aprenentatge no supervisat
La diferència principal entre l'aprenentatge supervisat i l'aprenentatge no supervisat són les dades utilitzades en qualsevol dels mètodes d'aprenentatge automàtic. Ambdós mètodes d'aprenentatge requereixen de dades. No obstant, les dades d'entrada utilitzades en l'aprenentatge supervisat són conegudes i estan etiquetades. Això vol dir que la màquina només s'encarrega del rol de determinar els patrons ocults de les dades ja etiquetats. Les dades utilitzades en l'aprenentatge no supervisat no es coneixen ni s'identifiquen. És el treball de la màquina classificar i etiquetar les dades en brut abans de determinar els patrons o funcions ocultes de les dades d'entrada.

Una altra de les diferències destacades entre l'aprenentatge supervisat i l'aprenentatge no supervisat és la complexitat computacional. L'aprenentatge supervisat és un mètode complex mentre que el mètode d'aprenentatge no supervisat és menys complex. Una de les raons que fa que l'aprenentatge

supervisat sigui més complex és el fet que cal comprendre i etiquetar les entrades mentre es troba en l'aprenentatge no supervisat, no es requereix comprendre i etiquetar les entrades.

L'altra diferència predominant aquests dos aprenentatges és la precisió dels resultats produïts després de cada cicle d'anàlisi. Tots els resultats generats a partir del mètode supervisat són més precisos i fiables en comparació amb els resultats generats a partir del mètode no supervisat. Un dels factors que explica aquesta diferencia, és perquè les dades d'entrada són ben conegudes i etiquetades, la qual cosa significa que la màquina només analitzarà els patrons ocults. Això no coincideix amb el mètode d'aprenentatge no supervisat on la màquina ha de definir i etiquetar les dades d'entrada abans de determinar els patrons o funcions ocultes. [43]

Aprenentatge Supervisat vs. Aprenentatge per reforç
En l'aprenentatge supervisat, hi ha un "supervisor" amb coneixement de l'entorn. Alguns problemes deriven de que hi ha tantes combinacions de subtasques que l'agent pot realitzar per aconseguir l'objectiu. De manera que crear un "supervisor" és pràcticament poc pràctic. En aquest tipus de problemes, és més factible aprendre de les pròpies experiències i obtenir-ne coneixement. En tant l'aprenentatge supervisat i el reforç, hi ha un mapatge d' entrada i sortida. Però en l'aprenentatge de reforç, existeix la funció de recompensa que actua com a feedback a l'agent en comparació amb l'aprenentatge supervisat.

Aprenentatge no supervisat vs. Aprenentatge per reforç
En l'aprenentatge per reforç, hi ha un mapatge d'entrada a la sortida que no està present en l'aprenentatge no supervisat. En l'aprenentatge no supervisat, la tasca principal és trobar els patrons subjacents més que no pas el mapatge. Per exemple, si la tasca consisteix a suggerir un article de notícies a un usuari, un algoritme d'aprenentatge no supervisat analitzarà articles similars que la persona hagi llegit prèviament i suggereixi a qualsevol d'ells. Mentre que un algorisme d'aprenentatge de reforç obtindrà una resposta constant de l'usuari suggerint alguns articles de notícies i, a continuació, construeixi un "gràfic de coneixement" del qual els articles els agradarà.

Aprenentatge Semi-Supervisat vs. Aprenentatge per reforç
L'aprenentatge semi-supervisat al ser una combinació d'aprenentatge supervisat i no supervisat, es diferencia de l'aprenentatge per reforç de manera similar a l'aprenentatge supervisat i semi-supervisat. Es mapeja de manera directe, mentre que l'aprenentatge per reforç no ho fa.

2.9. Mineria de Dades (Data Mining)

La mineria de dades es defineix com el procés que consisteix en extreure coneixement útil i comprensible, prèviament desconegut, de grans volums de dades emmagatzemades en diferents formats des de diferents fonts (SQL Server, Oracle, Excel, etc.). Per tant la tasca fonamental de la mineria de dades és trobar models intel·ligible a partir de les dades recollides. Això també implica descobrir les relacions, patrons i tendències observades de grans conjunts de dades, per posteriorment identificar informació útil i transcendent per prendre alguna decisió significativa. [44]

2.9.1. Procés de mineria de dades

Les etapes per la realització de la minera de dades sempre són les mateixes, independentment de la tècnica especifica a utilitzar. El conjunt de les parts del procés de mineria de dades també és conegut com a KDD (Knowledge Discovery in Databases). El procés KDD és el següent:

1. Aprenentatge del domini de l'aplicació, informació prèvia i plantejament dels objectius de l'aplicació.
2. Creació del conjunt de dades de destí, escollir el conjunt de dades o subconjunt de variables sobre les que es realitzarà el tractament.
3. Neteja de les dades i processament, en aquest punt es realitzen diverses operacions de caràcter important, com eliminar el soroll, recollida de informació necessària per modelar, determinar les estratègies per aconseguir informació dels camps que falten, etc.
4. Reducció de dades i projecció. En aquest punt s'exploren les característiques clau per representar les dades, depenent de l'objectiu i tasca a realitzar.
5. Elecció de la funció de mineria de dades, s'escull el model per el qual es processaran les dades.
6. Elecció de l'algoritme de mineria de dades, s'escullen els models algorítmics per la cerca de patrons.
7. Execució de la mineria de dades, es fa l'estudi i interpretació dels patrons detectats o es retrocedeix en algun punt anterior.
8. Interpretació, es fa una lectura de la informació extreta i s'esculla la informació útil.
9. Utilització del coneixement descobert.

3. Eines

Les organitzacions necessiten analitzar grans volums de dades, per extreure idees que les ajudin a informar i generar prediccions sobre el futur. Els softwares per a l'anàlisi predictiu són eines amb capacitats avançades, i dominis tant extensos que poden contenir processos que inclouen anàlisis estadístics, aprenentatge computacional, mineria de dades, etc.

Grans corporacions com IBM, SAP, SAS, etc. Des de principis dels 70 que estan en continuat desenvolupament d'eines centrades en la anàlisis de dades per a donar servei a empreses, centre educatius i entitats del sector públic. L'evolució constant dels sistemes de informació i les infraestructures ha permès que empreses mitjanes i petites tinguin accés a eines d'anàlisi. La globalització ha reduït preus de recursos fent que l'adquisició d'aquest tipus de sistemes per a tercers sigui molt més assequible. [45]

Les eines utilitzades per a l'anàlisi permeten, un ús senzill i eficaç dels mètodes matemàtics i un control dels processos d'anàlisi a nivell més profund. D'aquesta manera es brinden noves oportunitats en aquest camps permetent segmentar dades i informació, per convertir-ho posteriorment en paquets de dades útils per l'anàlisi d'estratègies empresarials.

Actualment existeix un gran ventall de possibilitats pel que fa eines d'anàlisis, algunes de codi obert i d'altres amb software propietari. En la gran majoria hi ha diferents nivells de coneixement per part de l'usuari, algunes per a realitzar consultes a un nivell més simple i superficial per part de l'usuari, i d'altres a nivells mes profunds i complexos.

A continuació es farà un anàlisi d'eines de codi obert com R i Python mentre que s'esmentarà les de codi propietari sense entrar-hi en profunditat, ja que no es te accés a llicencies de tipus educatiu. La selecció de les eines a analitzar s'ha fet en funció de la senzillesa d'accés a l'eina, instal·lació fàcil i qualitat de l'entorn.

En primer lloc es descriurà els entorns de les eines d'anàlisi escollides i s'exemplificarà l'ús d'aquestes mitjançant l'aplicació de casos de tècniques de modelatge predictiu com el Clustering, Arbres de decisió i Naive Bayes, en segon lloc es farà una comparativa entre les eines analitzades de codi obert.

3.1. Codi Obert

Les solucions d'anàlisi predictiu de codi obert són desenvolupades per comunitats globals de programadors, analistes i estadístics que us desenvolupen programari per oferir solucions d'anàlisi predictiu de cost zero o baix cost a les organitzacions.

El programari, desenvolupat de col·laborativament i distribuït de forma gratuïta, Fa que moltes solucions de codi obert siguin utilitzades per empreses de totes les indústries. Moltes de les eines, tenen grans capacitats i ofereixen amplis ventalls d'opcions. Permeten des de tècniques de agrupació fins a validacions dels models algorítmics. L'ús d'eines de codi obert permet destinar menys recursos en persones i projectes que en llicències d'eines de caràcter privat.

Les solucions de codi obert no són sistemes tancats, nous mòduls estan en constant desenvolupament. Si hi ha un paquet base o llibreria que no fa el que es necessita, hi haurà un complement gratuït disponible que ho faci. En molts programaris, hi ha una comunitat activa que genera i desenvolupa solucions de codi obert. Això vol dir que hi ha un grup de gent disposat a compartir coneixements i donar suport. Les solucions basades en codi obert ofereixen també molta flexibilitat a l'usuari final en termes d'integració amb altres productes i enfocaments per resoldre un problema.

Algunes de les opcions que ofereix el mercat actual són,

1. **Scikit-learn**
   Scikit-learn és una biblioteca d'aprenentatge de màquina de codi obert per al llenguatge de programació Python.
2. **R i RStudio**
   R és un llenguatge de programació de programari lliure i un entorn de programari per a la informàtica estadística i els gràfics.
3. **RapidMiner**
   RapidMiner és una plataforma de programari que proporciona un entorn integrat per a l'aprenentatge automàtic, la mineria de dades, la mineria de text, l'anàlisi predictiu i l'anàlisi de negocis.
4. **Weka**
   Weka és un popular programa de màquina-aprenentatge escrit en Java, desenvolupat a la Universitat de Waikato, Nova Zelanda. Conté una col·lecció d'eines de visualització i algorismes d'anàlisi i modelització predictiu.
5. **Orange:**
   Orange és una suite de programari per a la mineria de dades i l'aprenentatge automàtic basada en components. Inclou un conjunt de components per al processament previ de dades, la modelització, l'avaluació de models i les tècniques d'exploració.
6. **KNIME**
   KNIME és una plataforma d'anàlisi, reporting i integració d'informació de codi obert que integra diversos components per a l'aprenentatge automàtic i la mineria de dades.

3.1.1. R i RStudio

Que es R?

R és un llenguatge i entorn que proporciona mitjans per a desenvolupar informàtica estadística a nivell teòric i gràfic. Es un projecte GNU de codi obert, que proporciona un ampli ventall de possibilitats de estadístiques (modelització lineal i no lineal, proves estadístiques clàssiques, anàlisi de sèries de temporals, classificació, agrupació, etc.) i tècniques gràfiques, i és altament extensible.

Una de les fortaleses de R és la facilitat amb què es poden produir processos de qualitat amb una estructura ben dissenyada. R està disponible com a programari lliure sota els termes de la llicència pública general GNU de la Free Software Foundation en forma de codi font. Compila i executa una gran varietat de plataformes UNIX i sistemes similars (incloent FreeBSD i Linux), Windows i MacOS.

Entorn R i RStudio

R és un conjunt integrat d'instal·lacions de programari per a la manipulació, càlcul i visualització gràfica de dades. Inclou una instal·lació de paquets de manipulació i emmagatzematge de dades, un conjunt d'operadors per a càlculs matricials, una gran col·lecció integrada i coherent d'eines intermediàries per a l'anàlisi de dades, instal·lacions gràfiques per a l'anàlisi de dades i visualització en pantalla, i un llenguatge de programació desenvolupat per ser senzill i eficaç que inclou condicionals, bucles, funcions recursives definides per l'usuari i instal·lacions d'entrada i sortida. A continuació es mostra la interfície gràfica de RStudio

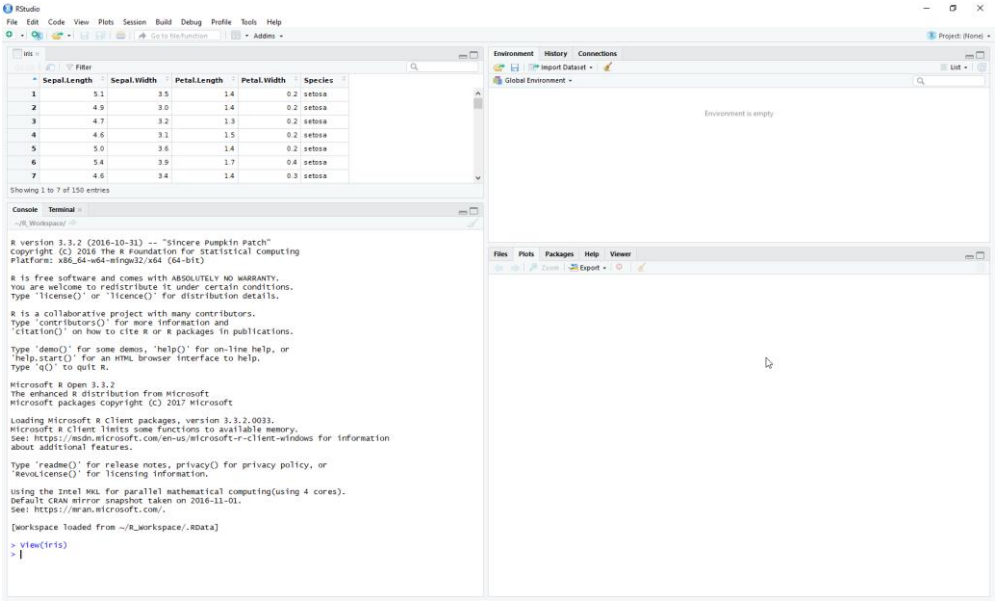


Figura 5. Interfície RStudio.

S'observa com el programari proporciona una estructura simple on hi ha uneditor de text, consola, visualitzador de variables i visualitzador de gràfics. A continuació es mostren el procés de execució de 2 models diferents com el K-means i els Arbres de decisió.

Clustering k-means

En aquest exemple es mostra un exemple de Clustering utilitzant l'algoritme K-Means, on es farà una anàlisi de les dades basat en la similitud entre les dades. El que es pretén amb l'exploració d'aquest conjunt de dades és crear diferents agrupacions de dades (clusters) per tal de que si es te una mostra es pugui classificar segons les variables d'entrada amb una agrupació en concret. [46]

La k significa clúster, s'ha d'especificar el nombre de clústers en que es vol que s'agrupin les dades. L'algorisme assigna aleatòriament cada observació a un clúster, i troba el centroide de cada clúster. Seguidament fa 2 passos més, es re-asignen els punts de dades al centroide més proper, i es torna a calcular els centroides. Es repeteixen aquests passos, fins que no hi ha més variacions en els centroides.

El següent conjunt de dades explorat, conté dades de l'espècie de flor d'iris on es registren dades sobre la longitud del sèpal, l'amplada del sèpal, la longitud del pètal i l'amplada del pètal de flors de diferents espècies.

```
> head(iris)
```

	Sepal.Length	Sepal.Width	Petal.Length	Petal.Width	Species
1	5.1	3.5	1.4	0.2	setosa
2	4.9	3.0	1.4	0.2	setosa
3	4.7	3.2	1.3	0.2	setosa
4	4.6	3.1	1.5	0.2	setosa
5	5.0	3.6	1.4	0.2	setosa
6	5.4	3.9	1.7	0.4	setosa

```
>
```

Després d'explorar les dades, s'observa que Petal.Length i Petal.Width són similars entre la mateixa espècie, però que varien considerablement entre diferents espècies, com es mostra a continuació:

```
> library(ggplot2)
> ggplot(iris, aes(Petal.Length, Petal.Width, color = Species)) +
geom_point()
```

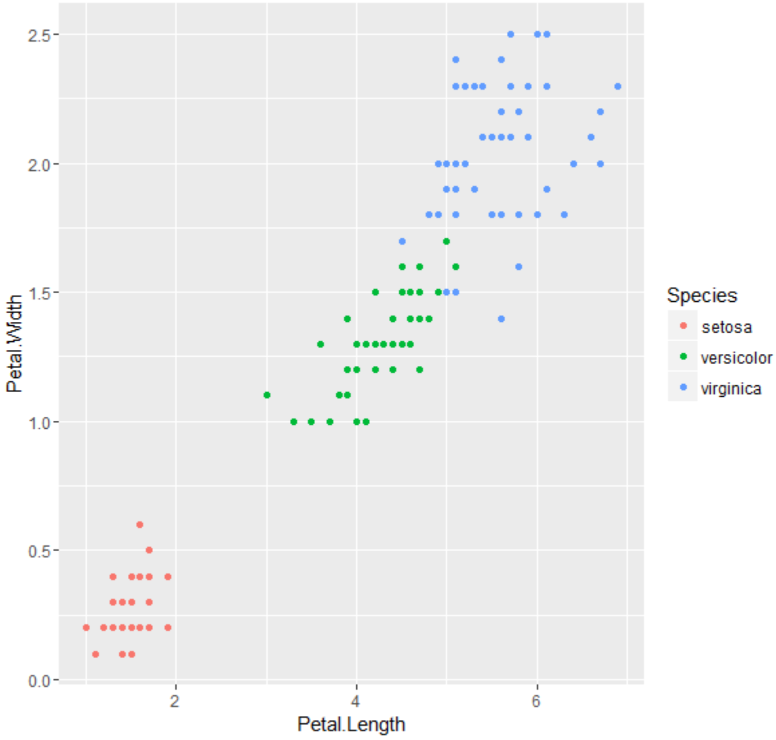


Figura 6. Iris Setosa Kmeans 1.

Ara que s'ha mostrat les dades, s'agrupen segons en funció de les variables. Atès que les assignacions de clúster inicials són aleatòries, es defineix la llavor per tal de que es pugui reproduir l'exemple.

```
> set.seed(20)
> irisCluster <- kmeans(iris[, 3:4], 3, nstart = 20)
> irisCluster
```

```
K-means clustering with 3 clusters of sizes 50, 52, 48

Cluster means:
  Petal.Length Petal.Width
1     1.462000    0.246000
2     4.269231    1.342308
3     5.595833    2.037500

Clustering vector:
  [1] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 2 2 2 2 2 2 2 2 2 2 2 2 2 2
 [65] 2 2 2 2 2 2 2 2 2 2 2 2 2 3 2 2 2 2 2 3 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 3 3 3 3 3 3 2 3 3 3 3 3 3 3 3 3 3 3 3 2 3 3 3 3 3 3 2 3
[129] 3 3 3 3 3 3 3 3 3 3 2 3 3 3 3 3 3 3 3 3 3 3

Within cluster sum of squares by cluster:
[1]  2.02200 13.05769 16.29167
 (between_SS / total_SS =  94.3 %)

Available components:

[1] "cluster"      "centers"      "totss"        "withinss"     "tot.withinss" "betweenss"    "size"         "iter"
[9] "ifault"
> |
```

Al saber que hi ha 3 espècies implicades, mitjançant l'algoritme, s'agrupa les dades en 3 clústers, i com que les assignacions inicials són aleatòries, s'especifica nstart = 20. Això vol dir que R provarà 20 assignacions aleatòries diferents i, a continuació, seleccionarà el que té el valor més baix de variació dins del clúster. S'observen els clústers centroides, els clústers als quals s' assigna cada punt de dades i la variació del clúster. Es compara els clústers amb les especies.

```
> table(irisCluster$cluster, iris$Species)

  setosa versicolor virginica
1     50          0         0
2      0         48         4
3      0          2        46
>
```

Com s'observa, les dades que pertanyen a les espècies setosa s'agrupen al clúster 3, versicolor al clúster 2 i virginica al clúster 1. L'algorisme classifica erròniament dos punts de dades pertanyents a versicolor i sis punts de dades pertanyents a virginica. Es treu el gràfic per veure les agrupacions.

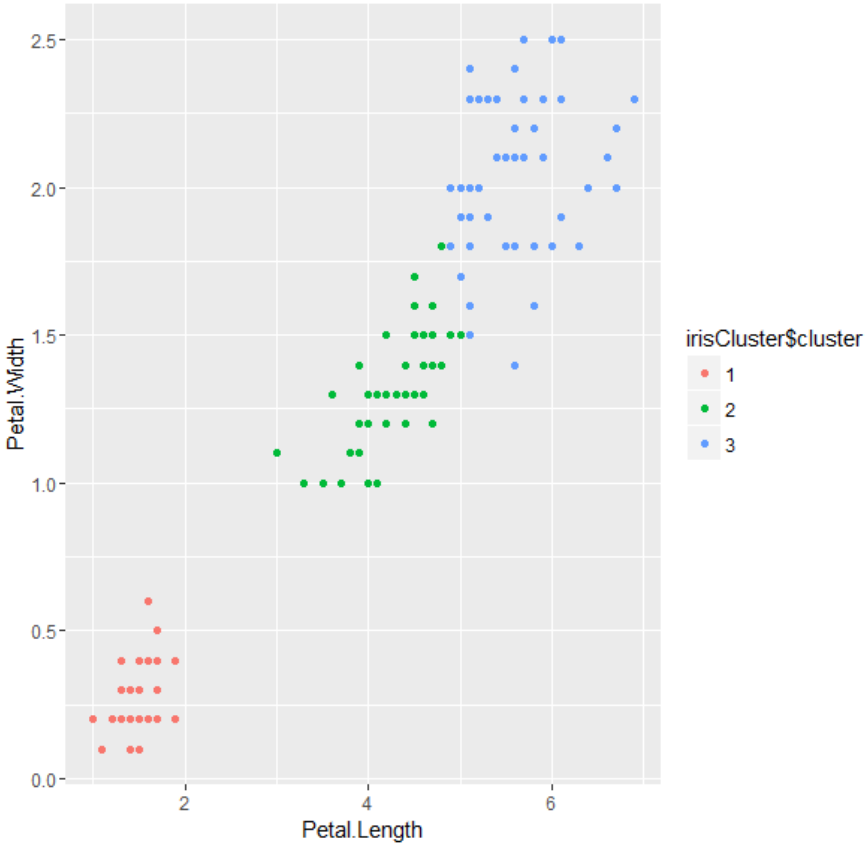


Figura 7. Iris Setosa Kmeans 2.

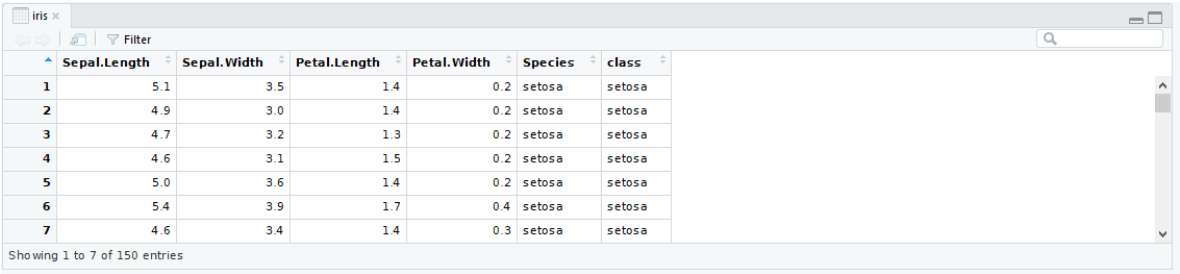
En la figura 7, s'observa com s'han agrupat finalment les diferents mostres segons l'amplada i la llargada del pètal, diferenciant les agrupacions de classes per colors. Com es comenta en l'apartat anterior, hi ha alguna mostra mal classificades, degut a que poden haver-hi excepcions dins de la mateixa classe i que per tant entrarien dins una agrupació errònia com és el cas.

Arbres de decisió

En aquest sub-apartat s'explorarà el cas dels arbres de decisió amb l'entorn R. El casos que es plantejaran tenen la finalitat de mostrar com varia el resultat final del model en funció del nombre de variables d'entrada, i més en els casos dels arbres de decisió, que poden tenir multitud de variables d'entrada de caràcter classificatori. Els exemples següents mostraran com es desenvolupa l'arbre en funció del nombre de variables i la classe, ja que depenent de les entrades la classificació de la mostra esdevindrà una o altra. El que es interessant d'aquest sub-apartat es veure com es desenvolupen les ramificacions en funció dels paràmetres d'entrada.

A continuació és realitzen els scripts per crear els diferents arbres de decisió, s'utilitzarà la funció ctree () del paquet "party". Per a realitzar l'exemple d'aquest tipus d'algoritmes s'utilitzarà el conjunt de dades iris com en cas anterior. [47]

```
> View(iris)
> iris$class<-as.factor(iris$Species)
> summary(iris)
```



iris

Filter

	Sepal.Length	Sepal.Width	Petal.Length	Petal.Width	Species	class
1	5.1	3.5	1.4	0.2	setosa	setosa
2	4.9	3.0	1.4	0.2	setosa	setosa
3	4.7	3.2	1.3	0.2	setosa	setosa
4	4.6	3.1	1.5	0.2	setosa	setosa
5	5.0	3.6	1.4	0.2	setosa	setosa
6	5.4	3.9	1.7	0.4	setosa	setosa
7	4.6	3.4	1.4	0.3	setosa	setosa

Showing 1 to 7 of 150 entries

```
  Sepal.Length    Sepal.Width     Petal.Length    Petal.Width          Species          class
 Min.   :4.300   Min.   :2.000   Min.   :1.000   Min.   :0.100   setosa     :50   setosa     :50
 1st Qu.:5.100   1st Qu.:2.800   1st Qu.:1.600   1st Qu.:0.300   versicolor:50   versicolor:50
 Median :5.800   Median :3.000   Median :4.350   Median :1.300   virginica :50   virginica :50
 Mean   :5.843   Mean   :3.057   Mean   :3.758   Mean   :1.199
 3rd Qu.:6.400   3rd Qu.:3.300   3rd Qu.:5.100   3rd Qu.:1.800
 Max.   :7.900   Max.   :4.400   Max.   :6.900   Max.   :2.500
```

Com s'observa, en el resum del conjunt de dades, hi ha 150 observacions totals: cada classe d'iris té 50 observacions cadascuna. També es detallen algunes estadístiques descriptives bàsiques de cadascuna de les quatre dimensions de les diferents flors.

Per a exemplificar, primer es crearà un exemple d'un arbre del qual dependrà una variable. Primer es fa una predicció de la classe d'iris en funció de la longitud del sèpal, l'arbre de decisió és el següent.

```
> tree <- ctree(class~Sepal.Length, data=iris)
> plot(tree)
```

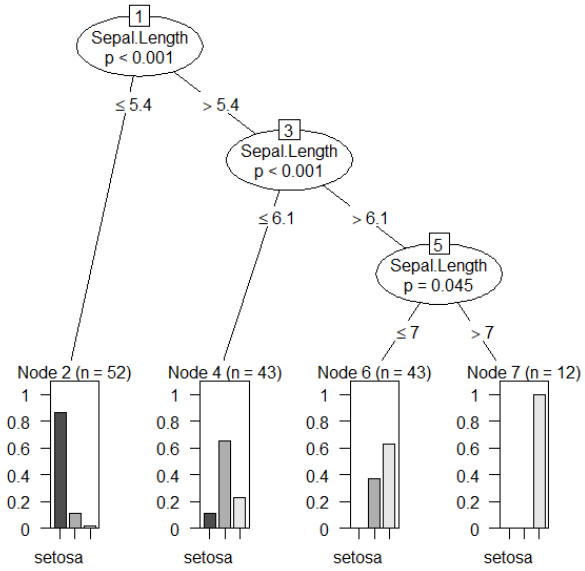


Figura 8. Arbres de decisió 1.

La interpretació de l'arbre de decisió és la següent, si es comença a fer la lectura des de l'arrel de l'arbre, el primer node de decisió determina si la iris te la longitud del sepal.length és inferior o igual a 5.4, cau cap al grup de flors d'iris. Aquesta lectura indica que hi ha 52 mostres de flors que tenen el sèpal inferior o igual a 5,4. D'aquestes 52 flors, aproximadament el 80% d'ells cauen en la primera classe d'iris que és setosa. aproximadament un 15% cau en la segona classe que és versicolor i la resta cau en la classe virginica. L'eix Y representa la proporció de les flors totals d'aquest grup que corresponen a cadascuna de les classes de flor d'iris (setosa, versicolor, virginica).

En el següent node, la variable sepal.length torna a determinar la decisió de l'arbre. Indica si les dues fulles són inferiors o iguals a 6.1 o superior a 6.1. en aquest cas si la flor te un sèpal amb longitud inferior o igual a 6.1, cau en el segon grup. Observant el segon gràfic, es pot indicar que la majoria de les 43 flors d'aquest grup són de color iris versicolor.

Si s'explora el tercer node. La longitud del sèpal es inferior o igual a 7 la classificació cau al tercer grup, sinó al quart. En mirar el tercer gràfic, hi ha 43 flors en el tercer grup i la majoria d'aquestes flors són de classe virginica, però, es pot veure que encara hi ha una bona quantitat de flors de iris versicolor. En el quart grup, hi ha les 12 flors restants. Totes aquestes flors pertanyen a la classe triple d'iris virginica.

Doncs s'observa que a nivell genèric l'arbre de decisió indica que les flors tipus iris setosa tendeixen a tenir una longitud de sèpal inferior, les flors de iris versicolor tenen sèpals de mitja longitud i les flors de iris virginica tendeixen a tenir la longitud de sèpal més llarga.

En el següent exemple es farà la predicció de la classe en funció de l'ample del sèpal de cada iris.

```
> tree <- ctree(class~Sepal.Width, data=iris)
> plot(tree2)
```

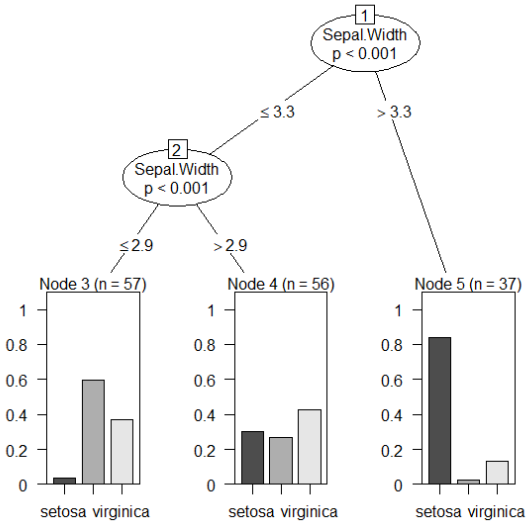


Figura 9. Arbres de decisió 2.

Al observar l'arbre de decisió, es veu com s'han generat 3 grup en comptes de 4 com en l'exemple anterior. En els resultats obtinguts, hi ha més uniformitat que abans. Les conclusions principals serien que l'iris setosa tendeix a tenir sèpals més amplis, els versicolors solen tenir sèpals més estrets, i la virginica té més varietat en l'amplada sèpal.

En el següent exemple, les ramificacions de l'arbre de decisió aniran funció de dos variables, les quals seran les dimensions dels sèpals.

```
> tree<-ctree(class~Sepal.Length+Sepal.Width, data=iris)
> plot(tree)
```

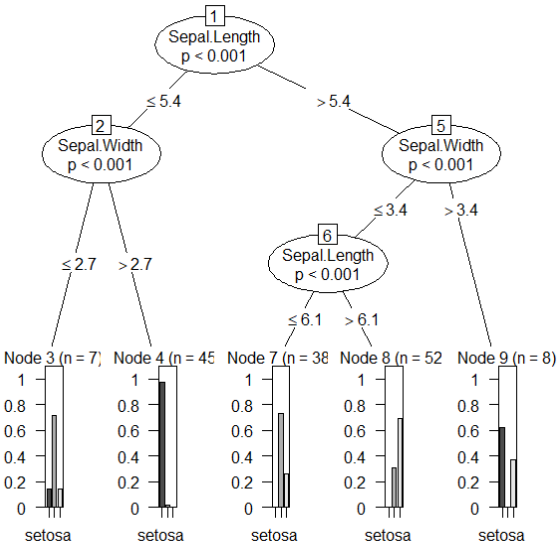


Figura 10. Arbres de decisió 3.

En aquest cas la classificació final, indica que si la longitud del sepal es menor o igual que 5.4 i posteriorment l'amplada del sepal és inferior o igual que 2.7 tendeix a ser versicolor, mentre que si es superior a 2.7 es virginica. Per l'altre costat, si la longitud del sepal és superior a 5.4, i posteriorment l'amplada del sepal es superior a 3.4, aquesta te opcions de ser setosa o versicolor, mentre que si es va per l'altre branca amb la longitud de sepal inferior o giual a 3.4, es reclassifica altre vegada amb la variable longitud de sepal, tot-hi que ambós resultats obtinguts la els resultats són bastant uniformes, però sense cap setosa.

En l'últim cas, s'agafen totes les variables per a crear l'arbre de decisió.

```
> tree <-ctree(class~Sepal.Length + Sepal.Width + Petal.Length +
Petal.Width, data=iris)
> plot(tree)
```

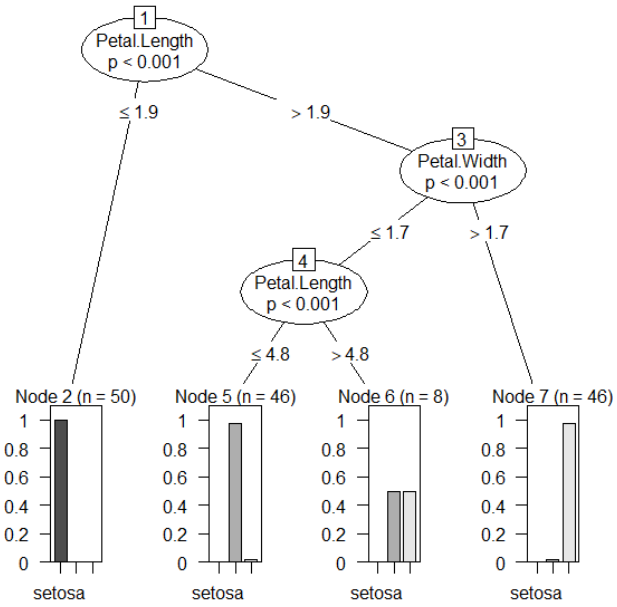


Figura 11. Arbres de decisió 4.

S'observa l'arbre de decisió que inclou les quatre variables (longitud sèpal, amplada sèpal, longitud del pètal i amplada del pètal) en el model de predicció. En aquest cas, només hi ha dos factors que s'utilitzen en els nodes per a generar decisions la longitud del pètal i l'amplada del pètal. Això ens vol dir que aquests dos factors són més importants a l'hora de diferenciar quin tipus de classe iris pertany a cada flor. Els factors longitud del sèpal i l'amplada del sèpal no són necessaris per predir quina classe pertanyen les flors.

3.1.2. Scikit-learn (Python)

Scikit-learn (anteriorment scikits.learn) és una biblioteca per a l'aprenentatge computacional de programari lliure per a la programació de llenguatge Python. Inclou diversos algorismes de classificació, regressió i anàlisi de grups entre els quals es mostren vectors, arbres aleatoris, tècniques de boosting, K-means i DBSCAN. Està dissenyada per interoperar amb les biblioteques numèriques i científiques NumPy i SciPy.

A continuació es mostren exemples de diverses metodologies de com desenvolupar algoritmes mitjançant la programació en Python.

Clustering K-nn

En el cas del l'script de clustering en python, el que es vol demostrar és com mitjançant la metodologia del kNN (veïns més propers), es classifica una mostra en funció de les mostres d'entrenament que més properes. El codi desenvolupat processa les dades proporcionades des de un fitxer extern per tal de classificar la mostra segons les K instancies del conjunt d'entrenament.

En aquest cas, es farà ús del mateix conjunt de dades que anteriorment,sobre l'espècie de flor iris. El conjunt de dades, està format per 150 observacions de flors d'iris de tres espècies diferents. Hi ha 4 mesures de flors donades: longitud del sèpal, amplada del sèpal, longitud del pètal i amplada del pètal, totes a la mateixa unitat de centímetres. L'atribut predit és l'espècie, que pot ser setosa, versicolor o virginica.

És un conjunt de dades estàndard on es coneix l'espècie per a totes les instàncies. Com a tal, podem dividir les dades en conjunts de dades d'entrenament i prova i utilitzar els resultats per avaluar la implementació del nostre algorisme. La precisió de la classificació sobre aquest problema sol ser superior al 90% correcte, normalment un 96% o superior.

El primer que fa és carregar el fitxer de dades. Les dades estan en format CSV sense cap línia de capçalera o sense cometes. Es pot obrir el fitxer amb la funció i llegir les línies de dades utilitzant la funció del lector en el mòdul CSV.

A continuació, es divideix les dades en un conjunt de dades de entrenament que el kNN utilitzarà per fer prediccions i un conjunt de dades de prova que podem utilitzar per avaluar la precisió del model. A continuació es converteixen les mesures de les flors carregades com a cadenes en nombres, posteriorment es divideix el conjunt de dades aleatòriament en entrenament i prova. Es selecciona un ràtio de 67/33 per a entrenament / prova, és una ràtio estàndard utilitzat.

Es defineix una funció anomenada loadDataset que carrega un CSV amb el nom del fitxer proporcionat i el divideix aleatòriament en conjunts de dades de trens i proves mitjançant la relació de divisió proporcionada.

Per fer les prediccions cal calcular la similitud entre dues instàncies de dades donades. Això és necessari perquè es pugui localitzar les K instàncies de dades més semblants en el conjunt de dades d'entrenament a una mostra en concret.

Tenint en compte que les quatre mesures de flors són numèriques i tenen les mateixes unitats, podem utilitzar directament la mesura de distància euclidiana. A més, volem controlar quins camps cal incloure en el càlcul de distància. Concretament, només es vol incloure els primers 4 atributs.

Ara que tenim una mesura de semblança, es vol utilitzar per recopilar les instàncies més semblants de k per a una instància donada. Aquest és un procés senzill de calcular, es calcula la distància per a totes les instàncies i es selecciona un subconjunt amb els valors de distància més petits. A sota hi ha la funció getNeighbors que retorna k els veïns més semblants del conjunt d'entrenament per a una instància de prova determinada.

Una vegada que localitzats els veïns més semblants per a una instància de prova, la propera tasca és idear una resposta que prediqui resultats basats en aquests veïns. A continuació es proporciona una funció per obtenir la resposta més gran de la votació d'un nombre de veïns.

Una manera fàcil d'avaluar la precisió del model és calcular una proporció de les prediccions correctes totals de totes les prediccions fetes, anomenades precisió de classificació. A sota hi ha la funció getAccuracy que suma les prediccions correctes totals i retorna la precisió com un percentatge de classificacions correctes.

A continuació es mostra el codi utilitzat en Pyhton. [48]

```
import csv
import random
import math
import operator

def loadDataset(filename, split, trainingSet=[] , testSet=[]):
        with open(filename, 'rb') as csvfile:
            lines = csv.reader(csvfile)
            dataset = list(lines)
            for x in range(len(dataset)-1):
                for y in range(4):
                    dataset[x][y] = float(dataset[x][y])
                if random.random() < split:
                    trainingSet.append(dataset[x])
                else:
                    testSet.append(dataset[x])


def euclideanDistance(instance1, instance2, length):
        distance = 0
```

```
        for x in range(length):
                distance += pow((instance1[x] - instance2[x]), 2)
        return math.sqrt(distance)

def getNeighbors(trainingSet, testInstance, k):
        distances = []
        length = len(testInstance)-1
        for x in range(len(trainingSet)):
                dist = euclideanDistance(testInstance, trainingSet[x], length)
                distances.append((trainingSet[x], dist))
        distances.sort(key=operator.itemgetter(1))
        neighbors = []
        for x in range(k):
                neighbors.append(distances[x][0])
        return neighbors

def getResponse(neighbors):
        classVotes = {}
        for x in range(len(neighbors)):
                response = neighbors[x][-1]
                if response in classVotes:
                        classVotes[response] += 1
                else:
                        classVotes[response] = 1
        sortedVotes = sorted(classVotes.items(), key=operator.itemgetter(1),
reverse=True)
        return sortedVotes[0][0]

def getAccuracy(testSet, predictions):
        correct = 0
        for x in range(len(testSet)):
                if testSet[x][-1] == predictions[x]:
                        correct += 1
        return (correct/float(len(testSet))) * 100.0

def main():
        # prepare data
        trainingSet=[]
        testSet=[]
        split = 0.67
        loadDataset('iris.data', split, trainingSet, testSet)
        print ('Train set:'' + repr(len(trainingSet)))
        print ('Test set:''+ repr(len(testSet)))
        # generate predictions
        predictions=[]
        k = 3
        for x in range(len(testSet)):
                neighbors = getNeighbors(trainingSet, testSet[x], k)
                result = getResponse(neighbors)
                predictions.append(result)
```

```
            print('> predicted=' + repr(result) + ', actual=' + repr(testSet[x][-1]))
    accuracy = getAccuracy(testSet, predictions)
    print('Accuracy: ' + repr(accuracy) + '%')

main()
```

En executar l'exemple, s'observen els resultats obtinguts per a cada predicció en comparació amb el valor de classe real del conjunt de proves. Al final de l'execució, s'extreu la precisió del model, per sobre el 98%.

```
...
> predicted='Iris-virginica', actual='Iris-virginica'
> predicted='Iris-virginica', actual='Iris-virginica'
> predicted='Iris-virginica', actual='Iris-virginica'
> predicted='Iris-virginica', actual='Iris-virginica'
> predicted='Iris-virginica', actual='Iris-virginica'
Accuracy: 98.0392156862745%
```

El resultat de l'execució del codi, el que pretén mostrar és la capacitat de classificació amb el model kNN. Es treu per la línia de comandes el tipus iris de la mostra i el tipus predit. Del conjunt de resultats en calcula la precisió, que permet saber si el model implementat està realment fent un anàlisi de prediccions fiable o no.

Naive Bayes

En l'exemple del metode Naïve Bayes, s'utilitzarà un conjunt de dades que fa referencia a Pima Indians Diabetis. En aquest cas la finalitat del model algorítmic implementat és la de classificar una instància en funció dels seus atributs mitjançant la probabilitat de que els atributs pertanyin a una classe o altra.

Aquest fitxer està format per 768 observacions de dades mèdiques per a patents de Pima indians. Els registres descriuen mesures instantànies preses del pacient com la seva edat, el nombre de vegades que han estat embarassades i el treball de la sang. Tots els pacients són dones de 21 anys o més. Tots els atributs són numèrics, i les seves unitats varien d'atribut a atribut.

Cada registre té un valor de classe que indica si el pacient va patir una aparició de diabetis dins dels 5 anys posteriors a la presa de mesures (1) o no (0). Una bona precisió de predicció seria del 70% -76%.

El primer que es carrega el fitxer de dades. Les dades estan en format CSV sense cap línia de capçalera o cap cometes. Es pot obrir el fitxer amb la funció open i llegir les línies de dades utilitzant la funció del lector en el mòdul CSV.

A continuació, es divideixen les dades en un conjunt de dades de entrenament que Naive Bayes pot utilitzar per fer prediccions i un conjunt de dades de test que es poden utilitzar per avaluar la precisió del model. Es divideix el conjunt de dades aleatòriament conjunt d'entrenament i conjunts de dades de test amb

una proporció del 67% de entrenament i 33% de prova. La funció splitDataset () dividirà un conjunt de dades determinat en entrenament i test.

El model naive bayes està format per un resum de les dades del conjunt de dades d'entrenament. Aquest resum s'utilitza quan es fan prediccions.

El resum de les dades de formació recollides implica la mitjana i la desviació estàndard per a cada atribut, per valor de classe. Per exemple, si hi ha dos valors de classe i 7 atributs numèrics, es necessita una desviació mitjana i estàndard per a cada atribut (7) i la combinació de valor de classe (2), que són 14 resums d'atributs. Aquests són necessaris quan es fan prediccions per calcular la probabilitat de valors d'atributs específics que pertanyen a cada valor de classe.

La primera tasca consisteix a separar les instàncies de conjunt de dades de formació per valor de classe, de manera que podem calcular les estadístiques de cada classe. Es pot fer creant un mapa de cada valor de classe a una llista de les instàncies que pertanyen a aquesta classe i ordeni el conjunt de dades d'instàncies en les llistes adequades.

Posteriorment s'ha de calcular la mitjana de cada atribut per a un valor de classe. La mitjana és la tendència central de les dades, i la s'utilitzarà com a mitjà de la nostra distribució gaussiana en calcular les probabilitats. També cal calcular la desviació estàndard de cada atribut per un valor de classe. La desviació estàndard descriu la variació de la difusió de les dades i la es farà servir per caracteritzar la propagació esperada de cada atribut en la distribució gaussiana al calcular les probabilitats.

Ara es pot fer un resum del conjunt de dades. Per a una llista determinada d'instàncies (per a un valor de classe) es calcula la mitjana i la desviació estàndard per a cada atribut. La funció zip agrupa els valors de cada atribut a través de les instàncies de dades a les seves pròpies llistes perquè es pugui calcular els valors de desviació mitjana i estàndard de l'atribut.

Es pot utilitzar una funció Gaussiana per estimar la probabilitat d'un valor d'atribut donat, donat la mitjana i desviació estàndard coneguda per l'atribut estimat a partir de les dades de formació. Tenint en compte que els resums d'atributs on es preparen per a cada atribut i valor de classe, el resultat és la probabilitat condicional d'un valor d'atribut donat un valor de classe. A la funció CalculateProbability () on es calcula primer l'exponent i després la divisió principal.

Ara que es pot calcular la probabilitat d'un atribut pertanyent a una classe, es pot combinar les probabilitats de tots els valors d'atribut d'una instància de dades i trobar una probabilitat de la instància de dades completa pertanyent a la classe. Es combinen les probabilitats multiplicant-les.

En el calculateClassProbabilities () es calcula la probabilitat d'una instància de dades determinada, es calcula multiplicant junts les probabilitats d'atribut per a cada classe.

Finalment, es pot estimar la precisió del model fent prediccions per a cada instància de dades en el nostre conjunt de dades test. El metode getPredictions () farà això i retornarà una llista de prediccions per a cada instància de prova.

A continuació es mostra el codi utilitzat. [49]

```
# Exemple de Naive Bayes implementat en Python
import csv
import random
import math

def loadCsv(filename):
	lines = csv.reader(open(filename, "rb"))
	dataset = list(lines)
	for i in range(len(dataset)):
		dataset[i] = [float(x) for x in dataset[i]]
	return dataset

def splitDataset(dataset, splitRatio):
	trainSize = int(len(dataset) * splitRatio)
	trainSet = []
	copy = list(dataset)
	while len(trainSet) < trainSize:
		index = random.randrange(len(copy))
		trainSet.append(copy.pop(index))
	return [trainSet, copy]

def separateByClass(dataset):
	separated = {}
	for i in range(len(dataset)):
		vector = dataset[i]
		if (vector[-1] not in separated):
			separated[vector[-1]] = []
		separated[vector[-1]].append(vector)
	return separated

def mean(numbers):
	return sum(numbers)/float(len(numbers))

def stdev(numbers):
	avg = mean(numbers)
	variance = sum([pow(x-avg,2) for x in numbers])/float(len(numbers)-1)
	return math.sqrt(variance)

def summarize(dataset):
	summaries = [(mean(attribute), stdev(attribute)) for attribute in
zip(*dataset)]
	del summaries[-1]
```

```
        return summaries

def summarizeByClass(dataset):
        separated = separateByClass(dataset)
        summaries = {}
        for classValue, instances in separated.iteritems():
                summaries[classValue] = summarize(instances)
        return summaries

def calculateProbability(x, mean, stdev):
        exponent = math.exp(-(math.pow(x-mean,2)/(2*math.pow(stdev,2))))
        return (1 / (math.sqrt(2*math.pi) * stdev)) * exponent

def calculateClassProbabilities(summaries, inputVector):
        probabilities = {}
        for classValue, classSummaries in summaries.iteritems():
                probabilities[classValue] = 1
                for i in range(len(classSummaries)):
                        mean, stdev = classSummaries[i]
                        x = inputVector[i]
                        probabilities[classValue] *= calculateProbability(x, mean,
stdev)
        return probabilities

def predict(summaries, inputVector):
        probabilities = calculateClassProbabilities(summaries, inputVector)
        bestLabel, bestProb = None, -1
        for classValue, probability in probabilities.iteritems():
                if bestLabel is None or probability > bestProb:
                        bestProb = probability
                        bestLabel = classValue
        return bestLabel

def getPredictions(summaries, testSet):
        predictions = []
        for i in range(len(testSet)):
                result = predict(summaries, testSet[i])
                predictions.append(result)
        return predictions

def getAccuracy(testSet, predictions):
        correct = 0
        for i in range(len(testSet)):
                if testSet[i][-1] == predictions[i]:
                        correct += 1
        return (correct/float(len(testSet))) * 100.0

def main():
        filename = 'pima-indians-diabetes.data.csv'
        splitRatio = 0.67
```

```
        dataset = loadCsv(filename)
        trainingSet, testSet = splitDataset(dataset, splitRatio)
        print('Split {0} rows into train={1} and test={2} rows').format(len(dataset),
len(trainingSet), len(testSet))
        # prepare model
        summaries = summarizeByClass(trainingSet)
        # test model
        predictions = getPredictions(summaries, testSet)
        accuracy = getAccuracy(testSet, predictions)
        print('Accuracy: {0}%').format(accuracy)

main()
```

Executar l'exemple següent proporciona una sortida com la següent:

```
Split 768 rows into train=514 and test=254 rows
Accuracy: 76.3779527559%
```

El resultat de l'execució del codi, el que pretén mostrar és la capacitat de classificació amb el model Naive Bayes. Es treu per la línia de comandes la divisió realitzada per al conjunt test i entrenament, i posteriorment la precisió del model implementat, que permetrà saber si el model és fiable, que com s'observa en aquest cas te més de un 70% de precisió per tant es podria considerar fiable a l'hora de realitzar prediccions.

3.2. Software Propietari / Comercial

Actualment moltes empreses aposten per el software propietari degut a l'increment d'interès per l'anàlisi de dades i, molta gent està interessada a entrar en aquest àmbit, però moltes vegades els hi manca una part important de formació com és la programació, o entendre els llenguatges de programació. Per aquest motiu moltes organitzacions que desenvolupen plataformes de analítiques de dades estan enfocant i centrant els seus esforços en crear plataformes amigables per a gent no experta en l'art de la ciència de dades i analítica predictiva.

A continuació es mostra un seguit de plataformes que requereixen o no de coneixements d'anàlisi de dades, però que les seves llicències són de pagament.

3.2.1. IBM SPSS

Una de les marques mes reconegudes a nivell mundial com IBM, proveeix amb una eina molt potent per a l'anàlisi de dades. IBM SPSS, inclou un panell de control on es monitoritza i gestiona el flux de treball de manera gràfica. Aquesta eina pot treballar amb dades estructurades i no estructurades.

La plataforma IBM SPSS proporciona mòduls per a l'anàlisi estadística avançada, que inclou una àmplia biblioteca d'algorismes d'aprenentatge automàtic, anàlisi de text, integració amb eines de codi obert, integració amb grans dades i implementació transversal en aplicacions.

La usabilitat i flexibilitat, fan que tingui gran acollida dins projectes amb gran variabilitat de mida. Un dels punts forts que te aquesta eina, és la integració amb plataformes de codi obert, com per exemple R i Python, compta amb més de 130 extensions per integrar eines externes. També proporciona una interfície senzilla que funciona de manera "Drag & Drop" amb múltiples funcions i orígens de dades.

IBM també proporciona una eina complementaria, IBM Watson, per a la construcció i implementació de models d'aprenentatge automàtic i Deep Learning. On es pot descobrir, neteja i transformar dades de manera interactiva, i utilitzar-ho amb eines de codi obert i integrant llibreries com Jupyter i RStudio.

3.2.2. MATLAB

MATLAB es una eina que combina un entorn d'escriptori per a l'anàlisi de dades i creació dels processos de disseny amb un llenguatge de programació de caire matemàtic, basat en matrius i arrays. Incorpora una gran quantitat d'eines dins la mateixa plataforma, totes elles ben documentades.

Les aplicaions de MATLAB permeten treballar immediatament de manera interactiva combinant l'accés directe a grans col·leccions d'algoritmes amb una realimentació visual immediata. Es poden fer iteracions continuades fins a

obtenir els resultats desitjats i, que després, generi automàticament un programa per reproduir o automatitzar el treball processat.

3.2.3. SAP HANA

SAP HANA combina una base de dades que compleix amb ACID amb serveis d'aplicacions, analítiques d'alta velocitat i eines flexibles d'adquisició de dades en una única plataforma. Com una BBDD SAP HANA emmagatzema i recupera dades utilitzades per altres aplicacions, com el CRM. També proporciona la integració d'altres eines i fonts per donar flexibilitat a l'hora de carregar dades i construir processos en temps real. Els serveis que proporciona, i la integració, donen suport al desenvolupament i implementació de aplicacions de anàlisis predictiu i processos de dades associats.

3.2.4. Altres

Google Cloud AutoML

Cloud AutoML forma part de les ofertes de Google Learning per a màquines d'aprenentatge computacional que permeten a les persones amb coneixements limitats sobre tècniques estadístiques relacionades amb l'aprenentatge computacional, crear models d'alta qualitat. Aquest servei fa que sigui més simple formar models de reconeixement d'imatges. Té una interfície "Drag & Drop" que permet que l'usuari carregui imatges, entreni el model i desplegui aquests models directament a Google Cloud. [50]

Cloud AutoML Vision està basat en tecnologies de cerca de transferència de Google i d'arquitectura neural (entre d'altres). Aquesta eina ja està sent utilitzada per moltes organitzacions.

Microsoft Azure ML Studio

Azure ML Studio és una plataforma senzilla però potent, basada en el navegador ML. Té un entorn visual "Drag & Drop" , no hi ha cap requisit de codificació. A la xarxa hi ha publicats tutorials per a la gent que vol iniciar-se en aquest món de l'analisi predicitu i utilitzar l'eina.

En aquest capítol, s'ha parlat de diverses eines que treballen per automatitzar diversos aspectes de la resolució d'un problema d'analítica predictiva. Alguns d'ells estan en una fase de recerca, alguns són de codi obert i altres ja s'estan utilitzant en la indústria. Com s'ha comentat en la introducció d'aquest apartat, moltes d'aquestes eines són més adequades per a persones que no estan familiaritzades amb la programació i la codificació.

Com a resum de l'apartat 3.2. s'adjunta un quadre resum dels software actuals i la seva presencia en el mercat de l'anàlisi predictiu en tots els segments de

mercat, situant-los dins la gràfica segons la presencia de mercat i la satisfacció del client. [57]

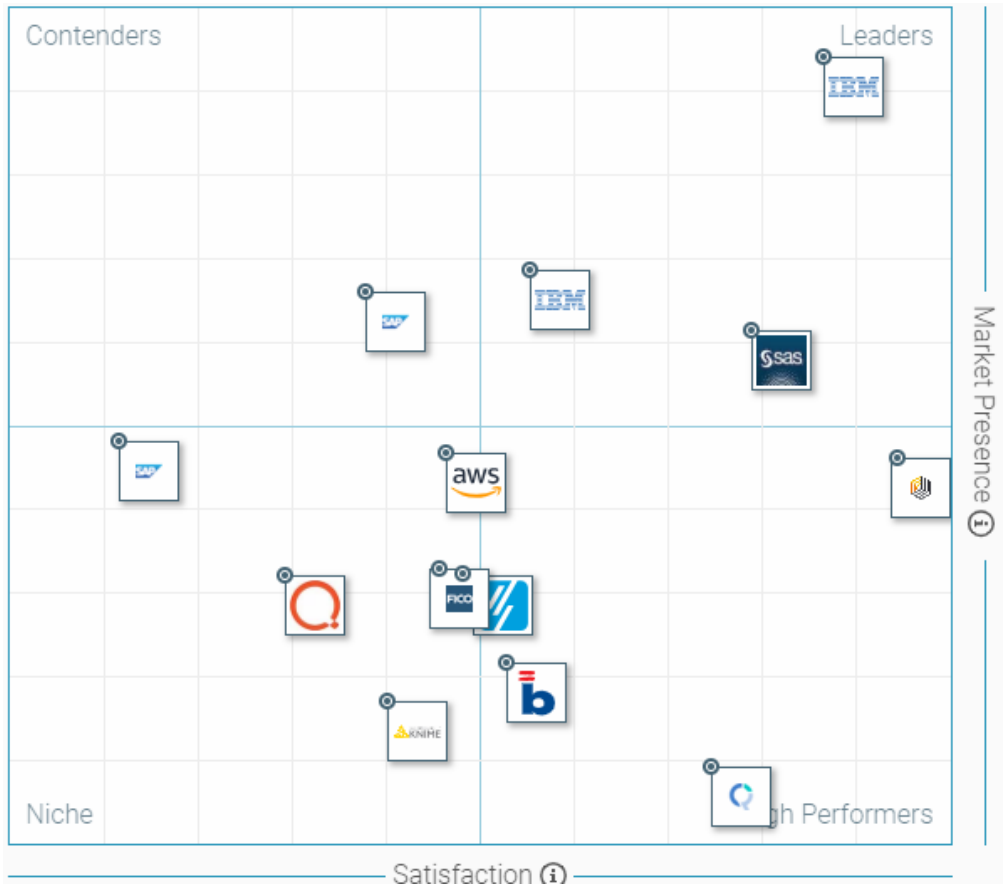


Figura 12. Millors Programari d'Anàlisi Predictiu

Per a poder incloure el producte dins la gràfica adjunta, el programari ha de contenir les següents característiques:

- Exploració i anàlisi tant de dades estructurades com no estructurades
- Crear conjunts de dades i/o visualitzacions de dades a partir de dades compilades
- Crear models predictius
- Permetre la importació i exportació de suites d'oficina o d'altres canals de recollida de dades

S'observa com el programari d'IBM SPSS Statistics és el millor situat pel que fa a quota de mercat. De manera pròxima el segueixen SAS i RapidMiner, els tres tenen facilitat d'ús de cares a l'usuari, coincideixen amb els requeriments de l'usuari i tenen una alta capacitat de integrar-se amb l'entorn existent. Altres factors que fan que estiguin al top 3 dels millors programaris són, la facilitat de instal·lació, la qualitat del suport i la facilitat de administració de l'entorn.

3.3. Comparativa software Codi Obert vs. Software Propietari.

Escollir l'eina adequada per a desenvolupar un projecte d'anàlisi predictiu és una qüestió important, actualment existeix un ventall tant ampli en el mercat que és difícil donar resposta a aquesta pregunta, Python, R, MATLAB, SAP, Oracle BI, etc. Però prèviament es podria separar en dos grans grups, el programari de codi obert i programari privat o comercial.

El programari de codi obert, està lliurement exposat a la xarxa, on tothom hi pot tenir accés i descarregar-lo, pot ser redistribuït i modificat de manera lliure i gratuïta, amb les seves limitacions. Per altra banda el programari comercial, està creat i desenvolupat per una empresa especifica.

Alguns avantatges del programari de codi obert respecte el comercial, es que la majoria del llenguatges tenen ecosistemes rics, es a dir, per exemple Python disposa de una potent llibreria per algoritmes d'anàlisi predictiu, com SciKit-Learn, que permet executar models de classificació, regressió, clustering i d'altres. Python també té paquets especialitzats per al deep learning i xarxes neuronals.

El fet de que sigui programari de codi obert, fa que atregui a una gran comunitat de gent, que entre tots s'ajuden per a resoldre les fallades i problemes del programari i també desenvolupar noves característiques. A part de que es disposen de fòrums i pagines especialitzades en resoldre els problemes més comuns.

Una de les característiques més importants però, és el baix cost que representa implementar una solució de codi obert respecte a tenir que adquirir llicències per a un programari comercial. La descàrrega de programes de codi obert i la instal·lació dels paquets necessaris són fàcils i l'adopció d'aquest procés pot accelerar el desenvolupament i reduir els costos.

D'altra banda, per a un programari amb llicència, una llicència pot agrupar les tarifes de configuració i manteniment per a la capacitat operativa d'ús diari, el suport necessari per resoldre problemes inesperats i una garantia d'implementació completa de les capacitats del programa. Les aplicacions empresarials, sovint van acompanyades d'un preu elevat, però proporcionen suport continu i especialitzat dels seus productes. El cost comparable de gestionar i prestar serveis a programes de codi obert, que sovint no tenen suport dedicat, és difícil de determinar respecte a als programes amb llicència.

Una altra característica atractiva del programari de codi obert és la seva flexibilitat. Python per exemple permet als usuaris utilitzar diferents entorns de desenvolupament integrats (IDEs) que tenen diverses característiques o funcions diferents, en comparació amb SAS, o MATLAB. Aquesta funcionalitat permet a l'usuari escollir la millor interfície per als objectius de l'organització.

Escollir doncs quin tipus de programari es una decisió complexa. Les eines de modelatge de dades de codi obert són atractives per la seva tendència natural

a estimular la innovació, l'adaptabilitat i la flexibilitat que proporcionen a les organitzacions. Però les solucions de programari propietari són també atractives ja que ofereixen suport especialitzat, moltes d'elles 24/7.

Les eines d'anàlisi de codi obert, sens dubte estan en augment, ja que cada vegada més, incrementa el nombre de persones amb habilitats tècniques i capacitats d'aprenentatge.

3.4. R vs. Python

Posteriorment a l'ús de les eines Python i R, es fa una comparativa entre ambdues, destacant els punts forts i febles de cadascuna, però també els punts en comú. Amb els exemples, s'ha verificat com són dues eines amb capacitats per a generar models d'anàlisi predictiu i que permeten realitzar tasques per al processament, transformació i visualització de dades.

R i Python són llenguatges de programació de codi obert amb un gran nombre de seguidors a darrera. Contínuament es van actualitzant la biblioteques i amplien el ventall de possibilitats que ofereixen. R s'utilitza principalment per a l'anàlisi estadística, mentre que Python proporciona un enfoc més genèric en l'àmbit de la ciència de dades.

Ambdós es podria dir que són dos dels llenguatge més capdavanters en referencia als llenguatges de programació orientats a la ciència de la informació. Python és un llenguatge de propòsit general amb una sintaxi llegible. R, però, està construït pels estadístics i engloba el seu llenguatge específic.

Python té biblioteques per a funcions matemàtiques, estadístiques i per el camp de la intel·ligència artificial, com les utilitzades en els algoritmes implementats, per tant és un llenguatge ideal per a desenvolupar algoritmes i models d'aprenentatge computacional.

R te una vessant més acadèmica i científica. Està dissenyat per respondre problemes estadístics, d'aprenentatge computacional i ciències de la informació. És una eina adequada per a la ciència de les dades a causa de les seves nombroses i potents biblioteques. Aquest llenguatge consta de molts paquets per realitzar gràfiques i mineria de dades.

A continuació és mostra una taula amb un recull dels paràmetres per a la comparació de les eines.

Paràmetres	R	Python
Objectiu	Anàlisi de dades i estadística	Desplegament i producció
Usuaris	Academic i I+D	Programadors i desenvolupadors
Usabilitat	Fàcil de utilitzar, llibreries disponibles	Facilitat de construcció de models
Curva d'aprenentatge	Dificultat en l'inici	Lineal
Popularitat del llenguatge	4.23% en el 2018	21.69% en el 2018
Integració	S'executa localment	Integrat en Apps
Tasques	Facilitat per a l'obtebció de resultats	Bo per al desplegament de models
BBDD	Grans BBDD	Grans BBDD
IDE	Rstudio	Spyder, Ipthon Notebook
Llibreries i paquets importants	tydiverse, ggplot2, caret, zoo	pandas, scipy, scikit-learn, TensorFlow, caret

Figura 13. Comparativa R vs. Python.

S'observa en la figura 13, com Python te una vessant que tendeix més cap a models d'anàlisi predictiu, com llibreries per a algoritmes d'intel·ligència artificial, aprenentatge computacional, mentre que R es dirigeix cap processos estadístics. Actualment és complex de determinar perquè és millor un o altre, i per quines àrees es millor un o altre, ja que continuadament estan sortint llibreries per a qualsevol tipus de model estadístic i matemàtic.

Per tant la tria es basaria en les objectius a assolir per a l'organització, coneixement dels llenguatges i entorns, temps de inversió per a l'aprenentatge i eina utilitzada per l'organització.

3.4.1. Anàlisi de R

R és un llenguatge desenvolupat per a estadístics. L'eina R te una bona interfície gràfica senzilla i intuïtiva, consta de multitud de paquets i una comunitat d'usuaris activa, no obstant te una corba d'aprenentatge feixuga, però que amb el temps és va simplificant. Consta d'una potent representació gràfica on la visualització de dades és una tasca senzilla, alguns paquets de visualització són ggplot2, ggvis, googleVis i rCharts. Tot i que algunes llibreries tenen dependències entre elles, és a dir depenen unes d'altres.

Un dels principals problemes és que requereix de gran quantitat de capacitat de càlcul de la CPU durant els seus processos, això pot fer que en ordinadors amb poca capacitat els processos esdevinguin lents. Existeixen eines complementaries com RStudio, RMarkdown i Shiny.

3.4.2. Anàlisi de Python

Pyhton és un llenguatge de propòsit general que és fàcil i intuïtiu, al contrari que R. Això li dóna una corba d'aprenentatge més lineal, i facilita el fet de desenvolupar eines d'anàlisi de dades. Consta de diferents IDEs que faciliten les tasques de desenvolupament, tot i això pel que fa la visualització de resultats, python te algunes biblioteques, però amb un ventall limitat i menys extens que R, això fa que sigui una tasca més complexa.

4. Tendències Actuals i Aplicabilitat

Per guanyar avantatge competitiu, les empreses, a part de passar pel procés de digitalització han de reconèixer la importància de la informació, i integrar els processos d'analítica de dades i d'anàlisi predictiu per desenvolupar estratègies basades en les importacions de les dades del seu client. En aquest apartat es recorre les tendències actuals i es llista una sèrie de casos d'ús reals d'organitzacions que han implementat processos d'anàlisi predictiu per a la millora de les seves actuacions i operacions.

4.1. Tendències actuals i casos d'ús

4.1.1. Sector Energètic

Optimització energètica
En el sector energètic, hi ha gran recerca per disminuir el consum de recursos i evitar a la contribució al canvi climàtic. En els EUA una empresa dedicada a reduir el consum energètic i convertir en eficients energèticament els recursos dels seus clients. Elabora serveis que es basen en la integració de diverses tecnologies amb la seva plataforma per a una millor gestió energètica. Ofereixen serveis com anàlisis de dades, tecnologies de modelització i optimització, aplicacions dissenyades per a usos propis i serveis administratius. Els serveis de l'empresa ofereixen valor als propietaris i agents de la construcció en qualsevol etapa del cicle de vida d'un edifici. Un dels últims casos en que ha aplicat la metodologia per a l'anàlisi predictiu és el desenvolupament de algorismes proactius per a l'optimització energètica de la climatització en edificis de gran escala. [51]

Per posar en context el mètode desenvolupat, s'ha de tenir en compte que els edificis d'oficines, hospitals i altres edificis comercials a gran escala representen prop del 30% de l'energia consumida a tot el món. Els sistemes de climatització en aquests edificis sovint són ineficients ja que no tenen en compte canvis en els patrons meteorològics, els costos d'energia variables o les propietats tèrmiques de l'edifici.

Aquesta organització ha desenvolupat una plataforma de programari basada en núvol que redueix el consum d'energia de la climatització en un 10-25% durant una jornada. Els algoritmes aplicats i mètodes de aprenentatge automàtic optimitzen de manera continuada el rendiment dels sistemes de climatització basant-se en pronòstics meteorològics a curt termini i variables relacionades amb la demanda energètica de l'entorn.

El desafiament de l'empresa residia en la necessitat de desenvolupar un procés algorítmic amb capacitat per a processar de manera continuada grans volums de dades provinents de diferents fonts (tals com els mesuradors, els termòmetres i els sensors de pressió del sistema de climatització), i altres com pronòstics del clima i diversos tipus de variables d'entrada. S'ha de tenir en compte que un edifici de grans dimensions conté milers de punts de dades, per tant les persones encarregades del projecte van tenir que realitzar un procés

previ de filtratge i processament de dades. Per a executar el procés de optimització, els enginyers i científics van haver de crear un model matemàtic precís del flux termodinàmics i energètics d'un edifici.

La solució final adoptada per l'empresa, recorre un procés on s'importen i es visualitzen de 3 a 12 mesos de dades de temperatura, pressió i energia que inclouen milers de milions de punts de dades. S'utilitzen diverses tècniques estadístiques i d'aprenentatge automàtic per detectar pics i buits, i eliminar el soroll produït per errors del sensor i altres fonts utilitzant funcions de filtratge.

Posteriorment mitjançant un model matemàtic es processen les dades filtrades i aquest que comprova diversos paràmetres, correlaciona la temperatura i la humitat ambientals amb la potència consumida pel sistema de climatització.

En el procés de modelatge s'utilitzen tècniques com la regressió, SVM, Naive Bayes i els algorismes d'aprenentatge automàtic tipus clustering com k-means per a segmentar les dades i determinar les aportacions relatives de gas, electricitat, vapor i energia solar a processos de calefacció i refrigeració.

Un cop segmentada la informació, es construeix un model que captura l'efecte del sistema de climatització i les condicions ambientals sobre temperatures internes a cada zona, així com sobre el consum total d'energia de l'edifici. S'analitzen els variables generades pel sistema de control de climatització per estimar el consum global d'energia.

Diàriament es calcula el cost d'energia que representa el que el client hauria pagat per l'energia consumida del sistema de climatització sense la plataforma de gestió de consum. L'estalvi pot arribar a percentatges del 10% fins al 25%.

Amb aquest procés de anàlisis predictiu el proveïdor genera grans volums de dades analitzades i visualitzades, la velocitat de desenvolupament de l'algorisme augmenta a mesura que s'acumula informació i es de la mateixa manera que s'optimitza el sistema, amb més volums de històrics s'obtenen resultats més precisos.

4.1.2. Manteniment predictiu

L'objectiu del manteniment predictiu és predir quan es pot produir un fracàs de l'equip i, en segon lloc, prevenir l'aparició del fracàs en realitzar el manteniment. El seguiment del fracàs futur permet planificar el manteniment abans de la fallada. L'ideal és que el manteniment predictiu permeti que la freqüència de manteniment sigui tan baixa com sigui possible per evitar el manteniment reactiu no planificat, sense incórrer en costos associats a fer massa manteniment preventiu.

El manteniment predictiu utilitza equips de control per avaluar el rendiment d'un objecte en temps real. Un element clau en aquest procés és l' internet de les coses, permet que els diferents actius i sistemes es connectin, treballin conjuntament de manera que comparteixin dades i aquestes siguin analitzades per que es pugui actuar sobre l'equip.

IoT es basa en sensors per capturar informació. Alguns exemples d'ús dels sensors del manteniment predictiu inclouen anàlisi de vibracions, anàlisi d'oli, imatges tèrmiques i observació d'equips. Alguns exemples de manteniment predictiu es s'expliquen a continuació.

Sector Oil (Gas i petroli)
L'evolució de l'analítica predictiva, ha desencadenat en moltes branques, i una d'aquestes és el manteniment predictiu. Un cas en particular és el de una empresa d'extracció de gas i petroli, que ha desenvolupat un software amb capacitats de manteniment predictiu per a equips d'extracció, utilitzant tècniques de l'aprenentatge automàtic com a peça clau.

En el cas de l'empresa petrolífera en concret, te períodes de gran demanda, on es treballen durant 24 hores els set dies de la setmana. D'un sol pou d'extracció hi poden haver fins a 20 camions funcionant de manera simultània, realitzant el procés de injecció i extracció dels fluids. Aquests camions duen unes bombes, que estan compostes de peces i components de mida reduïda, però que globalment reparar-les suposa un cost elevat (100.000 € per camió).

Doncs la tasca de supervisar el funcionament i desgast de les bombes i/o camions es converteix en un punt crític a l'hora de estalviar costos i recursos a l'empresa. Com a solució adoptada, s'ha posat sensors als camions bomba per extreure dades i aplicar metodologies d'aprenentatge automàtic per predir el seu cicle de vida.

Es posa el cas de que un camió bomba falla, degut a l'alta demanda, aquest s'ha de substituir de manera immediat o sinó es generen pèrdues milionàries per a l'empresa petrolífera. De la mateixa manera un manteniment massa freqüent consumeix molts recursos, mentre que poc manteniment no és bo pel funcionament continuat de les maquines, per tant el manteniment predictiu en aquest cas intenta buscar el manteniment òptim.

Per a desenvolupar un sistema predictiu d'aquest calibre la organització, necessitava importar conjunts de dades referents a la temperatura, pressió, vibració etc. A partir dels conjunts de dades recollits, es va analitzar la informació per tal de determinar en quines situacions les dades tenien major influencia en el desgast dels equips. D' aquesta manera es va determinar en quins casos la maquina era mes susceptible a fallades.

Es van realitzar diferents processos d'aprenentatge automàtic, conjuntament amb models de xarxes neuronals per obtenir resultats més precisos en les opcions de fallada. Posteriorment es van realitzar proves de camp per determinar si eren efectives les situacions trobades, i els resultats van confirmar la capacitat del sistema per a realitzar un anàlisi predictiu.

Amb aquesta implantació a l'empresa li suposa un estalvi anual de fins a 10 milions d'euros i dins el pressupost general un 30-40%, els diners del quals es poden destinar a altres sectors de l'empresa i no al manteniment de components dels camions bomba.

Sector Accessibilitat
En altres camps com el de l'accessibilitat també s'ha optat per desenvolupar processos d'anàlisi predictiu. Una multinacional finlandesa, una empresa de la indústria de l'ascensor i d'escales mecàniques, ha signat un acord amb el gegant IBM que li transformarà les operacions i capacitats tecnològiques de l'organització. Ha desenvolupat una eina de manteniment predictiu connectada 24 hores a la setmana, els set dies a la setmana, més concretament la plataforma Watson IoT, que gestiona aquest tipus de manteniment amb les maquines de l'organització mitjançant una plataforma interactiva.

La plataforma Watson IoT connecta els ascensors a la xarxa i permet a l'empresa aplicar metodologies d'aprenentatge automàtic i analítica de dades per millorar les estratègies de manteniment a nivell global. Amb aquesta implantació com en el cas del sector del cru, es minimitza el cost de la gestió de manteniment optimitzant-ne les reparacions, i sobretot els costos relacionats amb desplaçaments per part dels operaris. La solució de IBM ofereix dos metodologies diferenciades per crear models predictius, la primera de manera personalitzada i la segona basada en la tipologia de maquina connectada

4.1.3. Sector Esports

En aquesta branca es posa un exemple de com una entitat de futbol americà, aborda un aspecte clau com és l'experiència dels fans i optimització dels recursos de l'estadi. Degut a la necessitat que tenien de donar visibilitat a totes les operacions de l'estadi en el dia de partit i tenir una millor gestió dels recursos de l'entitat.

Com a mostra d'un dia de partit es podria dir, que durant la temporada regular en un partit estàndard hi podria haver 3.000 empleats, 80.000 seguidors, 13.000 pàrquings i 600 punts de venda (entre begudes, menjar i accessoris), entre d'altres serveis. L'entitat requeria de una solució que permetés prendre

decisions de cares a un futur a mig i llarg termini, per tal de afrontar de manera proactiva els possibles problemes de l'organització en un dia de partit.

L'entitat va desenvolupar una plataforma de gestió que permetia gestionar dades en temps real en relació a les operacions i al mateix temps, crear prediccions de cares a futures jornades de partit. Els alts càrrecs de l'organització actualment poden prendre decisions sobre les operacions que inclouen la millora de la gestió del pàrquing, estands de menjar i begudes, venda de accessoris, venda de entrades i les xarxes socials. També poden aprofundir en àrees específiques, així com dades de referència de temporades anteriors.

Amb aquestes accions el club ha augmentat els ingressos, optimitzat el rendiment de les operacions i millorat l'experiència del fan. [52]

4.1.4. Sector Salut

En l'àmbit de la salut, un cas real que s'explora, es realitza un desenvolupament d'algorismes de detecció per reduir les falses alarmes a les unitats de cures intensives desenvolupat per l'institut de les ciències de la Salut de la República Checa.

Aquesta entitat va fer un anàlisi de les falses alarmes d'electrocardiògrafs i altres dispositius de monitorització de pacients. Ja que són un problema greu en les unitats de cures intensives (UCI). Es va trobar que fins a un 86% de les alarmes de la UCI eren falses i un altre indicava que menys del 10% era important per a la gestió del pacient. Com a conseqüència el soroll de les falses alarmes pertorba el somni dels pacients, i la freqüència de les falses alarmes pot fer que el personal clínic resti importància als avisos donant lloc a temps de resposta més lents. L'objectiu doncs era reduir la incidència de falses alarmes de la UCI.

La metodologia desenvolupada per l'equip de investigació va donar coma resultat una taxa positiva del 92%i una taxa negativa 88%. Desenvolupant posteriorment, una plataforma amb capacitat de gestió de les falses alarmes.

4.1.5. Sector Banca

L'ús de les tecnologies de la informació en el sector bancari, s'ha convertit en una necessitat per tal de competir amb altres organitzacions. Els bancs s'han adonat que l'anàlisi predictiu pot ajudar-los a destinar els seus recursos de manera eficient, prendre decisions més intel·ligents i millorar el rendiment. Algunes de les possibilitats que ofereix l'anàlisi predictiu per aquest sector són: [53]

Detecció de frau
L'aprenentatge automàtic és fonamental per a la detecció i la prevenció efectiva de fraus que impliquen targetes de crèdit, comptabilitat, assegurances i molt més. La detecció de frau és essencial per proporcionar seguretat als clients i als empleats. Com més aviat un banc detecti frau, més ràpid pot restringir

l'activitat del compte per minimitzar la pèrdua. En implementar una sèrie d'esquemes de detecció de fraus, els bancs poden aconseguir la protecció necessària i evitar pèrdues significatives. La transformació dels coneixements teòrics profunds en aplicacions pràctiques exigeix experiència en tècniques de mineria de dades, com l'associació, el clúster, la predicció i la classificació.

Gestió de dades del client
Els bancs estan obligats a recollir, analitzar i emmagatzemar quantitats massives de dades. Però en lloc de considerar-ho com un simple exercici de compliment, els mètodes d'aprenentatge computacional i d'anàlisi predictiu poden transformar-la en una possibilitat d'obtenir més informació sobre els seus clients per generar noves oportunitats d'ingressos.

Comercialització personalitzada
La clau de l'èxit en màrqueting és fer una oferta personalitzada que s'adapti a les necessitats i preferències del client en particular. L'anàlisi de dades permet crear un màrqueting personalitzat que ofereixi el producte adequat a la persona adequada en el moment adequat en el dispositiu adequat. La mineria de dades s'utilitza àmpliament per a la selecció d'objectius per identificar els clients potencials d'un producte nou.

Els científics de dades utilitzen les dades de compra conductuals, demogràfiques i històriques per construir un model que prediqui la probabilitat que la resposta d'un client sigui una promoció o una oferta. Per tant, els bancs poden fer una difusió eficient i personalitzada i millorar les seves relacions amb els clients.

Segmentació del client
La segmentació del client significa separar els grups de clients segons el seu comportament o característiques específiques. Amb metodologies com clustering, arbres de decisió, regressió logística, etc. Es segmenta el volum de clients i es descobreixen segments d'alt valor i baix valor .

5. Conclusions

In God we trust. All others must bring data. (W. Edwards Deming)

5.1. Conclusions Finals

El futur de l'anàlisi de dades recau en una de les seves múltiples branques anomenada Anàlisi predictiu. Aquest àmbit de l'anàlisi de dades es centra principalment en el descobriment de oportunitats, aplicacions i tendències futures d'un domini de informació. L'anàlisi predictiu, és i s'està convertint en una àrea de interès per a la majoria de organitzacions i comunitats científiques, ja que utilitza metodologies i processos algorítmics per a la predicció de esdeveniments. Mitjançant aquests tipus de transformacions de dades les empreses, investigadors, o d'altres entitats tenen la possibilitat d'escollir i alinear les seves estratègies amb possibles successos de futur i d'aquesta manera millorar les probabilitats d'èxit i optimitzar el treball realitzat davant dels objectius proposats.

No obstant la aplicació de metodologies d'anàlisi predictiu són processos complexos en que s'ha de seguir uns estàndards i uns procediments. Per exemple, la recollida i neteja de dades es un pas de vital importància, ja que té afectació directe en el resultat final de les prediccions. De la mateixa manera els processos de validació són importants per mesurar el rendiment i la precisió del model implementat.

Per altra banda cada vegada més empreses relacionades amb programari de anàlisi de dades, estan centrant els seus esforços en produir software amb capacitats de modelatge predictiu, i estan oferint eines amb amplis ventalls de possibilitats. Tot i que també la comunitat de analistes que treballen en organitzacions i centres de recerca, s'està fent més ús d'eines de codi obert, com R o Python entre d'altres per a crear models d'anàlisi predictiu.

En relació amb els objectius plantejats inicialment al treball, s'ha recorregut una aproximació conceptual a l'anàlisi predictiu i sub-conceptes que el conformen, s'han descrit els principals mètodes d'anàlisi predictiu i el procediment de les tècniques d'anàlisi. Posteriorment s'han descrit eines de codi obert i de llicencies comercials, i s'ha fet una comparació entre elles. Finalment s'ha exemplificat l'ús de metodologies d'anàlisi predictiu a través de casos reals.

Pel que fa al seguiment de la planificació del treball, ja sigui en relació amb l'entrega o la correcció, s'ha dut a terme tal com estava planejat, tot i petites variacions degudes a temes laborals i acadèmics. En definitiva, no hi hagut grans afectacions a la planificació.

La metodologia ha estat basada en un àmbit més teòric que pràctic, atès que el present treball té com a objecte l'estat de l'art de l'anàlisi predictiu, una àrea centrada en la recerca i no en la implementació d'un cas real.

5.2. Línies de Futur

Pel que fa a possibles línies de treball d'investigació d'aquest treball, es podria haver explorat el tema de la Privadesa i Propietat de les Dades. El conflicte amb temes relacionats amb la privacitat i propietat de les dades està molt present, ja que hi ha moltes organitzacions que estan treballant perqué que hi hagi un lliure mercat de dades, i aquestes siguin de lliure accés.

Una altra línia de treball podria ser l'anàlisi de l'ús de les dades per part de un consumidor en un àmbit concret. És a dir centrar-se en l'ús que fa un usuari per determinar futures intencions. Aquest anàlisi de dades, actualment és molt popular en el camp de la publicitat en línia, al utilitzar l'anàlisi predictiu per millorar l'efectivitat dels anuncis.

Un altre línia de treball interessant és l'escalabilitat dels models o algoritmes, Quan es disposa de més dades, millora el sistema responsable de realitzar l'anàlisi predictiu, però caldria confirmar si redueix el rendiment i si segueix sent igual d'efectiu. Ambdues qüestions tenen un gran interès, per tal de determinar que passaria amb diversos models al injectar grans volums de dades.

I per últim podria resultar interessant la recerca de l'estudi dels ecosistemes de dades. Actualment s'estan generant grans quantitats de dades, de les que en deriven problemes com la propietat i la privacitat d'aquestes. Aquestes grans quantitats de dades comporten l'augment d'empreses relacionades amb extraccions de dades i la seva venta posterior a organitzacions privades. Per tant qüestions com la existència d'ecosistemes de dades o quines retroalimentacions entre empreses del sector privat creen aquets ecosistemes poden resultar de gran interès per l'àmbit de la intel·ligència artificial com també de l'àmbit legal.

6. Glossari

Algorisme Procediment de càlcul que consisteix a acomplir un seguit ordenat i finit d'instruccions que condueix, un cop especificades les dades, a la solució que el problema genèric en qüestió té per a les dades considerades.

Anàlisi Estudi d'un problema des del punt de vista de la informació, descomponent-lo en unitats més petites, esbrinant-ne l'estructura, aïllant els tractaments bàsics de la informació i dissenyant els algorismes que els realitzin.

Aprenentatge Procés pel qual un individu, un grup o una col·lectivitat adquireixen trets o complexos culturals, tals com el llenguatge, els prejudicis, les normes, les creences, les regles de conducta.

Computació Avaluar indirectament (una quantitat, especialment el temps) pel càlcul de certes dades.

Clúster Conjunt de coses.

CRM Customer Relationship Managment. Gestor de relacions amb el client.

Dada Cadascun dels grups d'operadors o factors que consisteixen en un conjunt de xifres, caràcters alfabètics o símbols que no denoten cap condició, valor o estat.

Eina Objecte fet per a una acció determinada i utilitzat directament per la mà per a actuar sobre la matèria.

ERP Enterprise Resource Managment. Gestor de recursos d'Empresa.

Estadística Ciència, mètode, tècniques, operació d'anàlisi matemàtica, que permeten d'estudiar numèricament amb el màxim de precisió els fenòmens col·lectius incompletament coneguts.

Entorn de Desenvolupament Integrat (IDE)

Informació Contingut d'una o més dades, fent abstracció de la representació concreta que adopta.

Intel·ligència de negoci (BI) Habilitat per transformar les dades en informació, i la informació en coneixement, de manera que es pugui optimitzar el procés de presa de decisions en els negocis.

Mètode Camí que se segueix, manera ordenada, sistemàtica, de procedir, per a arribar a un fi.

Operació Acció que implica l'aplicació d'un principi, d'una regla, especialment formant part d'una sèrie d'accions, d'un pla.

Procés Manera de descabdellar-se una acció progressiva.

Regressió Estudi de la millor aproximació d'una variable estadística *y* a partir d'una família donada de variables estadístiques $x_1,..., x_n$, mitjançant combinacions lineals del tipus $a_1x_1,..., a_nx_n + b$.

7. Bibliografia

7.1. Webgrafía

[1] Cambridgeinternational.org. (2018). *Cambridge international*. [online] Available at: https://www.cambridgeinternational.org/Images/285017-data-information-and-knowledge.pdf [Accessed 17 Oct. 2018].

[2] Ecomputernotes.com. (2018). *What do you mean by Data and Information ?*. [online] Ecomputernotes.com. Available at: http://ecomputernotes.com/fundamental/information-technology/what-do-you-mean-by-data-and-information [Accessed 17 Oct. 2018].

[3] Color, A. (2018). *Método científico: Inferencias y predicciones - Articulos - ABC Color*. [online] Abc.com.py. Available at: http://www.abc.com.py/articulos/metodo-cientifico-inferencias-y-predicciones-903221.html [Accessed 17 Oct. 2018].

[4] New America. (2018). *Hand-in-Hand: Ethics and Predictive Analytics*. [online] Available at: https://www.newamerica.org/education-policy/edcentral/hand-hand-ethics-and-predictive-analytics/ [Accessed 17 Oct. 2018].

[5] Jul 03, 2., Case, N., Documents, R. and Raise, T. (2018). *Legal Issues in Big Data: 2017 - Royse Law Firm*. [online] Royse Law Firm. Available at: https://rroyselaw.com/technology-transactions/agtech/legal-issues-big-data-2017/ [Accessed 17 Oct. 2018].

[6] Merino, P. (2018). *Aspectos legales que hay que tener en cuenta en el tratamiento del big data - Ecommerce News*. [online] Ecommerce News. Available at: https://ecommerce-news.es/aspectos-legales-que-hay-que-tener-en-cuenta-en-el-tratamiento-del-big-data-4313 [Accessed 18 Oct. 2018].

[7] IBM Big Data & Analytics Hub. (2019). *The strategic impact of predictive analytics*. [online] Available at: https://www.ibmbigdatahub.com/blog/strategic-impact-predictive-analytics [Accessed 7 Jan. 2019].

[8] Halo. (2018). *Descriptive, Predictive, and Prescriptive Analytics Explained*. [online] Available at: https://halobi.com/blog/descriptive-predictive-and-prescriptive-analytics-explained/ [Accessed 20 Oct. 2018].

[9] PDFfiller. (2018). *Decision matrix, Pareto analysis and other effective tools for smart business decisions*. [online] Available at: https://blog.pdffiller.com/decision-matrix-pareto-analysis-and-other-effective-tools-for-making-smart-business-decisions/ [Accessed 20 Oct. 2018].

[10] Mlwave.com. (2018). *Kaggle Ensembling Guide | MLWave*. [online] Available at: https://mlwave.com/kaggle-ensembling-guide/ [Accessed 20 Oct. 2018].

[11] Stats and Bots. (2018). *Ensemble Learning to Improve Machine Learning Results*. [online] Available at: https://blog.statsbot.co/ensemble-learning-d1dcd548e936 [Accessed 20 Oct. 2018].

[12]En.wikipedia.org. (2018). *Uplift modelling*. [online] Available at: https://en.wikipedia.org/wiki/Uplift_modelling#Measuring_uplift [Accessed 20 Oct. 2018].

[13] Dummies. (2018). *Basics of Uplift Predictive Analytics Models - dummies*. [online] Available at: https://www.dummies.com/programming/big-data/data-science/basics-of-uplift-predictive-analytics-models/ [Accessed 21 Oct. 2018].

[14] Predictive Analytics Times. (2018). *Uplift Modeling: Making Predictive Models Actionable - Predictive Analytics Times - machine learning & data science news*. [online] Available at:

https://www.predictiveanalyticsworld.com/patimes/uplift-modeling-making-predictive-models-actionable/8578/ [Accessed 21 Oct. 2018].

[15] Dummies. (2018). *How to Test the Predictive Analysis Model - dummies*. [online] Available at: https://www.dummies.com/programming/big-data/data-science/how-to-test-the-predictive-analysis-model/ [Accessed 21 Oct. 2018].

[16] Gay, E. (2018). *Top 7 Struggles with Prediction Analytics & How to Solve Them - 2018 Edition*. [online] Emcien. Available at: https://emcien.com/predictive-analytics-problems-2018/ [Accessed 22 Oct. 2018].

[17] En.wikipedia.org. (2018). *Dimensionality reduction*. [online] Available at: https://en.wikipedia.org/wiki/Dimensionality_reduction#Feature_selection [Accessed 26 Oct. 2018].

[18] En.wikipedia.org. (2018). *Principal component analysis*. [online] Available at: https://en.wikipedia.org/wiki/Principal_component_analysis [Accessed 26 Oct. 2018].

[19] dummies. (2018). *How to Utilize Linear Regressions in Predictive Analytics - dummies*. [online] Available at: https://www.dummies.com/programming/big-data/data-science/how-to-utilize-linear-regressions-in-predictive-analytics/ [Accessed 26 Oct. 2018].

[20] En.wikipedia.org. (2018). *Linear regression*. [online] Available at: https://en.wikipedia.org/wiki/Linear_regression [Accessed 26 Oct. 2018].

[21] En.wikipedia.org. (2018). *Linear regression*. [online] Available at: https://en.wikipedia.org/wiki/Linear_regression [Accessed 26 Oct. 2018].

[22] En.wikipedia.org. (2018). *Logistic regression*. [online] Available at: https://en.wikipedia.org/wiki/Logistic_regression [Accessed 26 Oct. 2018].

[23] En.wikipedia.org. (2018). *Decision tree*. [online] Available at: https://en.wikipedia.org/wiki/Decision_tree [Accessed 26 Oct. 2018].

[24] Towards Data Science. (2018). *Decision Trees in Machine Learning – Towards Data Science*. [online] Available at: https://towardsdatascience.com/decision-trees-in-machine-learning-641b9c4e8052 [Accessed 26 Oct. 2018].

[25] Analytics, B., Simplified!, D. and Ray, S. (2018). *Decision Tree | Predictive Analytics*. [online] Analytics Vidhya. Available at: https://www.analyticsvidhya.com/blog/2015/01/decision-tree-simplified/2/ [Accessed 27 Oct. 2018].

[26] Quantstart.com. (2018). Bayesian Statistics: A Beginner's Guide | QuantStart. [online] Available at: https://www.quantstart.com/articles/Bayesian-Statistics-A-Beginners-Guide [Accessed 27 Oct. 2018].

[27] Learning, M., R), 6. and Ray, S. (2018). *6 Easy Steps to Learn Naive Bayes Algorithm (with code in Python)*. [online] Analytics Vidhya. Available at: https://www.analyticsvidhya.com/blog/2017/09/naive-bayes-explained/ [Accessed 04 Nov. 2018].

[28] En.wikipedia.org. (2018). *Artificial neural network*. [online] Available at: https://en.wikipedia.org/wiki/Artificial_neural_network [Accessed 04 Nov. 2018].

[29] Skymind. (2018). *A Beginner's Guide to Neural Networks and Deep Learning*. [online] Available at: https://skymind.ai/wiki/neural-network#concrete [Accessed 05 Nov. 2018].

[30] Nielsen, M. (2018). *Neural Networks and Deep Learning*. [online] Neuralnetworksanddeeplearning.com. Available at: http://neuralnetworksanddeeplearning.com/ [Accessed 05 Nov. 2018].

[31] En.wikipedia.org. (2018). *Deep learning*. [online] Available at: https://en.wikipedia.org/wiki/Deep_learning [Accessed 05 Nov. 2018].

[32] Towards Data Science. (2018). *Ensemble Methods in Machine Learning: What are They and Why Use Them?*. [online] Available at: https://towardsdatascience.com/ensemble-methods-in-machine-learning-what-are-they-and-why-use-them-68ec3f9fef5f [Accessed 06 Nov. 2018].

[33] Towards Data Science. (2018). *The Random Forest Algorithm – Towards Data Science*. [online] Available at: https://towardsdatascience.com/the-random-forest-algorithm-d457d499ffcd [Accessed 09 Nov. 2018].

[34] Ke.tu-darmstadt.de. (2018). [online] Available at: http://www.ke.tu-darmstadt.de/lehre/archiv/ws0910/mldm/ibl.pdf [Accessed 09 Nov. 2018].

[35] Towards Data Science. (2018). *The 5 Clustering Algorithms Data Scientists Need to Know*. [online] Available at: https://towardsdatascience.com/the-5-clustering-algorithms-data-scientists-need-to-know-a36d136ef68 [Accessed 10 Nov. 2018].

[36] Scikit-learn.org. (2018). *2.3. Clustering — scikit-learn 0.20.2 documentation*. [online] Available at: https://scikit-learn.org/stable/modules/clustering.html#k-means [Accessed 10 Nov. 2018].

[37] Analytics, B. and R), I. (2018). *Improve Your Model Performance using Cross Validation (in Python / R)*. [online] Analytics Vidhya. Available at: https://www.analyticsvidhya.com/blog/2018/05/improve-model-performance-cross-validation-in-python-r/ [Accessed 10 Nov. 2018].

[38] Dr. Sebastian Raschka. (2018). *Model evaluation, model selection, and algorithm selection in machine learning*. [online] Available at: https://sebastianraschka.com/blog/2016/model-evaluation-selection-part1.html [Accessed 14 Nov. 2018].

[39] Datasciencecentral.com. (2018). *11 Important Model Evaluation Techniques Everyone Should Know*. [online] Available at: https://www.datasciencecentral.com/profiles/blogs/7-important-model-evaluation-error-metrics-everyone-should-know [Accessed 14 Nov. 2018].

[40] codeburst. (2018). *What is Regularization in Machine Learning? – codeburst*. [online] Available at: https://codeburst.io/what-is-regularization-in-machine-learning-aed5a1c36590 [Accessed 14 Nov. 2018].

[41] En.wikipedia.org. (2018). *Rule-based machine learning*. [online] Available at: https://en.wikipedia.org/wiki/Rule-based_machine_learning [Accessed 18 Nov. 2018].

[42] Expertsystem.com. (2018). *What is Machine Learning? A definition - Expert System*. [online] Available at: https://www.expertsystem.com/machine-learning-definition/ [Accessed 18 Nov. 2018].

[43] Heath, N. (2018). *What is machine learning? Everything you need to know | ZDNet*. [online] ZDNet. Available at: https://www.zdnet.com/article/what-is-machine-learning-everything-you-need-to-know/ [Accessed 18 Nov. 2018].

[44] Techopedia.com. (2018). *What is Data Mining? - Definition from Techopedia*. [online] Available at: https://www.techopedia.com/definition/1181/data-mining [Accessed 18 Nov. 2018].

[45] PAT RESEARCH: B2B Reviews, Buying Guides & Best Practices. (2018). *Predictive Analytics Tools - Compare Reviews, Features, Pricing in 2019 - PAT RESEARCH: B2B Reviews, Buying Guides & Best Practices*. [online] Available at: https://www.predictiveanalyticstoday.com/predictive-analytics-tools/ [Accessed 18 Nov. 2018].

[46] Kodali, T. (2018). *K Means Clustering in R*. [online] DataScience+. Available at: https://datascienceplus.com/k-means-clustering-in-r/ [Accessed 23 Nov. 2018].

[47] Wong, K. (2018). *Chapter 24: Decision Trees*. [online] Ademos.people.uic.edu. Available at: https://ademos.people.uic.edu/Chapter24.html [Accessed 23 Nov. 2018].

[48] Brownlee, J. (2018). *Tutorial To Implement k-Nearest Neighbors in Python From Scratch*. [online] Machine Learning Mastery. Available at: https://machinelearningmastery.com/tutorial-to-implement-k-nearest-neighbors-in-python-from-scratch/ [Accessed 06 Dec. 2018].

[49] Brownlee, J. (2018). *Naive Bayes Classifier From Scratch in Python*. [online] Machine Learning Mastery. Available at: https://machinelearningmastery.com/naive-bayes-classifier-scratch-python/ [Accessed 06 Dec. 2018].

[50] BBVAOpen4U. (2018). *El ranking de las mejores soluciones de análisis predictivo para empresas*. [online] Available at: https://bbvaopen4u.com/es/actualidad/el-ranking-de-las-mejores-soluciones-de-analisis-predictivo-para-empresas [Accessed 06 Dec. 2018].

[51] Es.mathworks.com. (2018). *BuildingIQ Develops Proactive Algorithms for HVAC Energy Optimization in Large-Scale Buildings*. [online] Available at: https://es.mathworks.com/company/user_stories/buildingiq-develops-proactive-algorithms-for-hvac-energy-optimization-in-large-scale-buildings.html [Accessed 09 Dec. 2018].

[52] McKendrick, J. and →, V. (2018). *How Real-Time Sensors Can Reduce Sports Injuries - RTInsights*. [online] RTInsights. Available at: https://www.rtinsights.com/how-real-time-sensors-can-reduce-sports-injuries/ [Accessed 09 Dec. 2018].

[53] ActiveWizards: data science and engineering lab. (2018). *Top 9 Data Science Use Cases in Banking*. [online] Available at: https://activewizards.com/blog/top-9-data-science-use-cases-in-banking/ [Accessed 09 Dec. 2018].

7.2. Referències d'imatges

[54] DataCamp Community. (2018). *Linear Regression R*. [online] Available at: https://www.datacamp.com/community/tutorials/linear-regression-R [Accessed 26 Oct. 2018].

[55] En.wikipedia.org. (2018). *Logistic regression*. [online] Available at: https://en.wikipedia.org/wiki/Logistic_regression [Accessed 26 Oct. 2018].

[56] Es.wikipedia.org. (2018). *Perceptrón multicapa*. [online] Available at: https://es.wikipedia.org/wiki/Perceptr%C3%B3n_multicapa#/media/File:RedNeuronalArtificial.png [Accessed 24 Dec. 2018].

[57] G2 Crowd. (2019). *Best Predictive Analytics Software in 2019 | G2 Crowd*. [online] Available at: https://www.g2crowd.com/categories/predictive-analytics [Accessed 7 Jan. 2019].

8. Annexos