Selecció i Implantació d'una eina BI en assaig clínic sobre tabaquisme

Alumne:
Vanesa Granado Font
Grau Enginyeria Informàtica

Consultor:
Humberto Andrés Sanz

14 Gener 2017


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

FITXA DEL TREBALL FINAL

Títol del treball:	*Selecció i implantació d'una eina BI en assaig clínic sobre tabaquisme.*
Nom de l'autor:	*Vanesa Granado Font*
Nom del consultor:	*Humberto Andrés Sanz*
Data de lliurament (mm/aaaa):	*01/2017*
Àrea del Treball Final:	*Business Intelligence*
Titulació:	*Grau Enginyeria informàtica*

Resum del Treball (màxim 250 paraules):

Fruit del problema expressat pel personal sanitari, que treballa en un assaig clínic sobre tabaquisme, en obtenir coneixement sobre les dades que tenen emmagatzemades, versem el nostre treball en oferir una possible solució.

Per aquest motiu, el nostre projecte tracta sobre la tria d'una eina de Business Intelligence i la seva implementació.

El projecte mostra com, després de reunir la informació necessària de l'estudi i de les necessitats de l'equip sanitari, es fa una anàlisi dels requeriments i es decideix utilitzar *RapidMiner Studio.*

Per últim, s'implementa el software i es dóna resposta a cinc qüestions demanades per l'equip, finalitzant així el esmentat treball.

Abstract (in English, 250 words or less):

Due to the problem expressed by the sanitary staff who works in a clinical trial about smoking, and thanks to the data provided by them, we develop our project to give them a possible solution.

For this reason, our project is about selecting a Business Intelligence tool and its implementation.

After collecting the necessary information about the trial and the sanitary staff needs, an analysis of the requirements is carried out and it is decided to use RapidMiner Studio.

Finally, the software is implemented to solve five important demands required by the sanitary staff.

Paraules clau (entre 4 i 8):

Business Intelligence, ETL, RapidMiner Studio, Mineria de dades, dashboards, informació, coneixement

Índex

Llista de figures

1. Introducció

1.1 Context i justificació del Treball

Amb l' explosió, dels últims anys, de les xarxes socials i el canvi en la manera d'interactuar els usuaris amb la tecnologia ha fet que es generi una gran quantitat de dades on la majoria aporta informació però no coneixement.

D'aquí la necessitat d' incorporar les solucions **Business Intelligence (BI)** que ens permetin gestionar aquestes dades.

Els darrers últims anys, els sistemes Business Intelligence van integrant-se a les organitzacions com una eina indispensable per optimitzar els seus processos i reduir costos. Però aquests han esdevingut una eina fonamental en àrees d' investigació on es genera gran quantitat de dades en poc temps, com és el món sanitari.

Aquest últim punt és el que ha motivat aquest TFG. En una primera reunió l'equip d'investigació, encarregat de dur a terme un assaig clínic sobre el tabaquisme, expressa la seva preocupació per no poder extreure informació que aporti algun coneixement de les dades que tenen registrades. Aquestes dades les tenen guardades en un excel molt extens però que no li veuen la utilitat. Aquest fet fa que l'equip estigui en una constant preocupació i angoixa fent que baixi el rendiment alhora d' obtenir les dades dels pacients.

Així doncs, no és difícil veure la preocupació del responsable de l'assaig, ja que és possible que acabi fracassant i de la utilitat que els proporcionaria un sistema de Business Intelligence.

Es proposa triar una de les eines i fer la implantació en local per així començar a treballar la informació i poder veure resultats a partir de les dades que tenen emmagatzemades al full de càlcul.

L' equip d'investigadors veuen amb bons ulls la proposta, entreguen el fitxer excel amb les dades.

1.2 Objectius del Treball

L'objectiu general del treball versa sobre l'obtenció de coneixements específics sobre Business Intelligence. L' oportunitat que ofereix treballar sobre un projecte des de les fases inicials fins a l'entrega final, proporciona una experiència que no s' obté amb la teoria.

Com a objectius principals específics:

- Atenen als requisits i a les necessitats escollir una eina del mercat adequada que compleixi les expectatives del grup d'assaig.
- Fer-ne la implantació local i desenvolupar els processos que aportin informació sobre les següents qüestions plantejades per l'equip:
    - Obtenir visió general ràpida de dades bàsiques com sexe, edat iniciació, cigarretes que fuma al dia, nivell d'estudis, status en l'estudi o intents previs per deixar de fumar.
    - Verificar l'impacte de la gamificació.
    - Proporcionar coneixement que permeti determinar als centres l'època adient per realitzar les campanyes.
    - Fer una predicció aproximada al pes i nivell de monòxid de carboni a l'any de deixar de fumar.
    - Relació entre marques de tabac, dependència a la nicotina i nivell de motivació.
- Fer una formació a l'equip per a que ells mateixos puguin fer alguns processos, almenys els més senzills.

1.3 Enfocament i mètode seguit

El següent TFG té un enfoc principalment acadèmic, sense deixar de banda la possibilitat de poder adaptar-se en un escenari real.

Les dades de l'assaig clínic del qual disposem en alguns casos són incompletes per tema de temporalitat, així que farem una simulació del que podria ser un escenari hipotètic i les omplirem a l' atzar. Per aquesta raó, la següent memòria quedarà merament en l'àmbit acadèmic.

Aquest TFG té un component elevat sobre l'àrea sanitària i investigació, així doncs l'autora de la mateixa estarà en contacte amb un membre de l'assaig clínic per tal de resoldre els dubtes.

Per tal de poder triar una de les eines BI, es buscarà informació relativa a diferents fonts com: llibres tècnics sobre BI, articles online i professionals especialitzats en Business Intelligence. Tot això anirà degudament especificat a la bibliografia.

1.4 Planificació del Treball

1.4.1 Anàlisi de requeriments

El recursos necessaris per la realització del projecte són els següents:

- **Recursos Tècnics**:

HARDWARE	• PC o MAC OS x amb tecnología Intel i5, 8GB RAM, 500GB HDD
SOFTWARE	• SO Windows o MAC OS X • Microsoft Officce 2013 • Acrobat Reader PDF • RapidMiner Studio v7.3.0

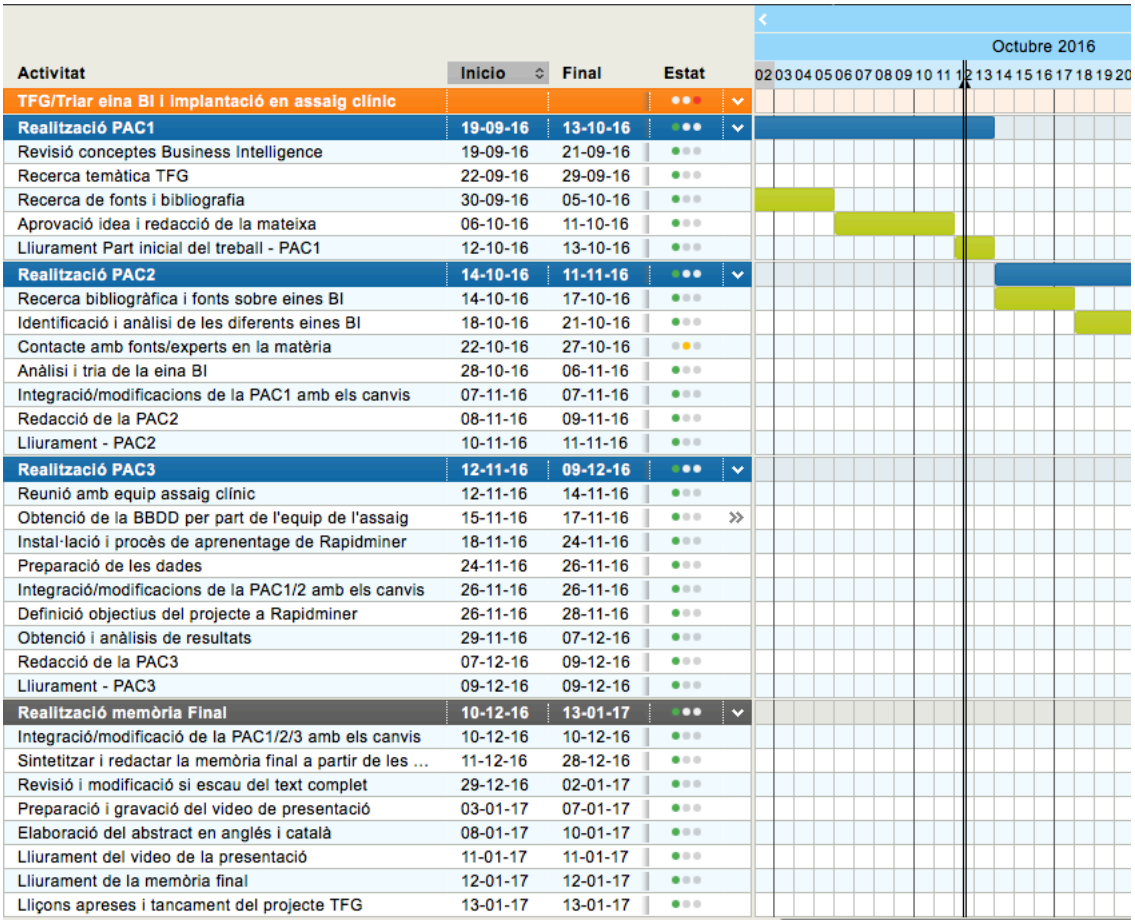
- **Recursos Humans**:

ROL	RESPONSABILITATS
Cap de projecte	• Planificació i estimació del projecte • Gestió de l'abast i seguiment del projecte • Tancament del projecte
Consultor UOC	• Seguiment del projecte • Resoldre dubtes durant el projecte • Determinar la qualitat dels lliurables temporals
Analista	• Anàlisi de requeriments
Consultor	• Validar requeriments i comprovar qualitat dels lliurables.
Enginyer especialista en BI	• Realitzar el procés ETL i processos a RapidMiner. • Obtenció de resultats per posterior anàlisi.

1.4.2 Diagrama de Gantt

El següent **diagrama de Gantt** mostra les diferents tasques i subtasques necessàries per realitzar el TFG. Aquestes han estat pensades i programades perquè coincideixin amb les entregues programades definides al pla docent.

Així doncs s'estructura en quatre blocs diferenciats entre ells i que es mostra a continuació:



Activitat	Inicio	Final	Estat
TFG/Triar eina BI i implantació en assaig clínic			
Realització PAC1	**19-09-16**	**13-10-16**	
Revisió conceptes Business Intelligence	19-09-16	21-09-16	
Recerca temàtica TFG	22-09-16	29-09-16	
Recerca de fonts i bibliografia	30-09-16	05-10-16	
Aprovació idea i redacció de la mateixa	06-10-16	11-10-16	
Lliurament Part inicial del treball - PAC1	12-10-16	13-10-16	
Realització PAC2	**14-10-16**	**11-11-16**	
Recerca bibliogràfica i fonts sobre eines BI	14-10-16	17-10-16	
Identificació i anàlisi de les diferents eines BI	18-10-16	21-10-16	
Contacte amb fonts/experts en la matèria	22-10-16	27-10-16	
Anàlisi i tria de la eina BI	28-10-16	06-11-16	
Integració/modificacions de la PAC1 amb els canvis	07-11-16	07-11-16	
Redacció de la PAC2	08-11-16	09-11-16	
Lliurament - PAC2	10-11-16	11-11-16	
Realització PAC3	**12-11-16**	**09-12-16**	
Reunió amb equip assaig clínic	12-11-16	14-11-16	
Obtenció de la BBDD per part de l'equip de l'assaig	15-11-16	17-11-16	
Instal·lació i procès de aprenentage de Rapidminer	18-11-16	24-11-16	
Preparació de les dades	24-11-16	26-11-16	
Integració/modificacions de la PAC1/2 amb els canvis	26-11-16	26-11-16	
Definició objectius del projecte a Rapidminer	26-11-16	28-11-16	
Obtenció i anàlisis de resultats	29-11-16	07-12-16	
Redacció de la PAC3	07-12-16	09-12-16	
Lliurament - PAC3	09-12-16	09-12-16	
Realització memòria Final	**10-12-16**	**13-01-17**	
Integració/modificació de la PAC1/2/3 amb els canvis	10-12-16	10-12-16	
Sintetitzar i redactar la memòria final a partir de les ...	11-12-16	28-12-16	
Revisió i modificació si escau del text complet	29-12-16	02-01-17	
Preparació i gravació del video de presentació	03-01-17	07-01-17	
Elaboració del abstract en anglés i català	08-01-17	10-01-17	
Lliurament del video de la presentació	11-01-17	11-01-17	
Lliurament de la memòria final	12-01-17	12-01-17	
Lliçons apreses i tancament del projecte TFG	13-01-17	13-01-17	

Figura 1. Diagrama de Gantt - Diagrama complet online aquí

1.5 Breu sumari de productes obtinguts

El productes obtinguts al llarg del cicle de vida del projecte han estat els següents:

- Fitxers dels processos realitzats a RapidMiner Studio amb els models predictius, estructura ETL i resultats obtinguts
- Memòria del projecte
- Presentació virtual
- Arxiu de suport a la presentació
- Informe d'autoavaluació

1.6 Breu descripció dels altres capítols de la memòria

En els següents apartats d'aquesta memòria s'aniran descrivint les tasques realitzades per a l' obtenció dels objectius plantejats en l' apartat 1.2.

L' apartat 2 conté tot el desenvolupament del treball a partir de l'anàlisi i la planificació inicial del projecte.

Així doncs, aquest inicia amb el procés de tria de l'eina BI. Aquest recull una introducció als conceptes bàsics sobre Business Intelligence, una parada obligatòria al tractament de dades i el procés ETL i una descripció dels tipus d'eines principals. Acabant amb l' explicació del perquè s'arriba a escollir RapidMiner Studio prioritzant requisits com; software multiplataforma, llicència gratuïta, fugir de serveis al núvol per tal de no interferir amb la LOPD actual aplicada al centre i que tingués la possibilitat de treure gràfics, informes, dashboards i explotar la mineria de dades.

A continuació en l'apartat 2.2 s'entra de ple a la implantació de l'eina, explicant primer el producte, els requisits tècnics i la preparació del fitxer de dades inicial, facilitat per l'equip, per a importar des de RapidMiner Studio.

Un cop l' entorn és l'apropiat per l' explotació de les dades facilitades, es mostra com s'han treballat les qüestions concretes que preocupaven l'equip investigador. D'aquesta manera, l'apartat 2.3 recull aquest últim punt i en mostra el resultat.

L' apartat 3 mostrarà les conclusions un cop finalitzat i avaluat el resultat del projecte. En ell es parlarà de les lliçons apreses, reflexió sobre l' obtenció dels objectius inicials, anàlisi sobre el seguiment de la planificació inicial i possibles línies de treball pendent.

2. Desenvolupament del projecte

2.1 Procés de tria de l'eina BI

2.1.1 Introducció al Business Intelligence

Segons Gartner, la consultora internacional especialitzada en tecnologies de la informació, defineix el ***Business Intelligence*** (BI) com a "*terme general que inclou les aplicacions, la infraestructura i les eines i millors pràctiques que permeten l'accés i anàlisi de la informació, per millorar i optimitzar les decisions i el rendiment*" [2].

L'objectiu del BI versa sobre l'anàlisi de les dades que es tenen dels diferents inputs/fonts i així poder proporcionar coneixement específic que ajudi a la presa de decisions.

[Dades + anàlisi = Coneixement]

Les fases d'un projecte complet de BI s'inicien amb la definició d'un model de negoci, on es fa un anàlisi exhaustiu de la situació actual. Això permet definir les metodologies i tenir coneixement sobre el resultat de l'aplicació d' aquestes.

Posteriorment es recull la informació necessària i que serà la que s'utilitzarà pel posterior anàlisis. Normalment aquestes dades es presenten en gestors de base de dades, sota un model "entitat-relació" i seguint el llenguatge de consulta conegut com **SQL** (Structured Query Language).

No obstant, la informació pot estar emmagatzemada en diferents sistemes i amb múltiples formes. Com per exemple, el cas que ens ocupa. On la informació es facilitada per part de l'equip de l'assaig clínic en un full de càlcul.

Un cop es té localitzada la informació, o model de dades, comença el procés anomenat en BI com **ETL** (Extract, Transform and Load), explicat amb més detall en el següent apartat. Aquest prepara les dades per ser analitzades, utilitzant les eines i tècniques adequades. El resultat final ha de proporcionar el

coneixement suficient per determinar si el resultat obtingut requereix més requeriments o pel contrari aporta el que s'estava buscant.

2.1.2 ETL

El concepte ETL introduït en l'apartat anterior fa referència a un dels processos més importants dintre d'un projecte de BI, ja que serà el que "alimentarà" el **Data Warehouse** o el magatzem de dades utilitzat per al posterior anàlisi.

En el procés es realitzen les diferents tasques:

- **Extracció:** Es recupera tota la informació en brut de les diferents fonts on poden estar emmagatzemades, ja siguin **fonts internes** (fulls de càlcul, documents, etc.) **com externes** (CRM, ERP, Xarxes Socials, etc.). És important intentar causar el menor impacte possible a la font origen, ja que aquesta segurament estarà realitzant altres tasques i es podria veure afectat el rendiment.

- **Transformació:** En aquest pas es fa neteja de la informació redundant. Es corregeixen valors erronis i es completen valors donant qualitat a les dades i com a conseqüència reduint els errors. També es filtren les dades, s'homogeneïtzen i s'agrupen.

- **Càrrega:** Finalment es valida i es carrega la nova informació al Data Warehouse del qual s'analitzarà la informació.

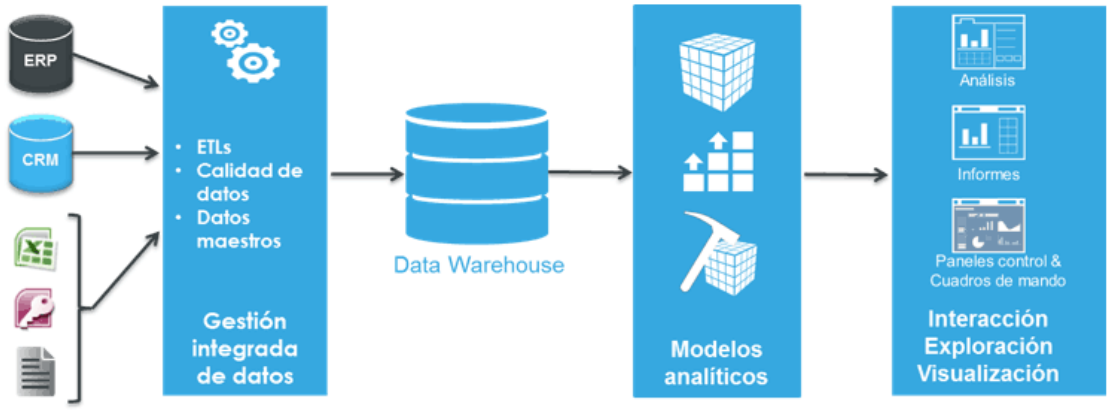


Figura 2. Esquema procés BI - Origen imatge

2.1.3 Tipus d'eines BI

Les eines BI són les encarregades d'analitzar i explotar les dades del repositori d'informació obtingut després del procés ETL.

A continuació es desglossen les principals eines d' Intel·ligència de Negoci:

- **Informes i consultes:** Aquestes són eines que, mitjançant la recollida i presentació de les dades permeten mostrar molta informació d'una manera organitzada i estructurada, facilitant el posterior anàlisi per part de l'usuari final. Aquests poden ser desenvolupats a mida de les necessitats que es requereixin en cada cas.

- **Quadres de comandament:** Aquests panells són útils per mostrar informació als usuaris de manera molt gràfica i concisa. Així d'un cop d'ull és fàcil tindre una visió general del que està reflectint. A diferència dels informes, no ofereixen tanta informació. Els quadres de comandament els podem dividir en dos grups:

    - ***Analítics o Dashboards*:** Aquests estan destinats a elaborar informes dels objectius estratègics de l' empresa i mostrar els indicadors claus de rendiment (**KPI**). Al mostrar un conjunt de

KPI's proporcionen una instantània de la situació actual que es vol advertir quan es defineixen.

- ***Integrals o Scoreboards*:** A diferencia dels anteriors aquests es defineixen a nivell estratègic i inclouen a tota l' organització. Així doncs, proporciona una visió més estratègica que permet definir objectius i indicadors d'evolució. Facilita la presa de decisions i millora la seva certesa ja que permeten la correcció a mida que es van generant els successos.

- **OLAP (On-Line Analytical Processing):** El processament analític en línia dona multidimensionalitat a les dades, indexació especialitzada i capacitats intensives de càlcul. Això fa que aquestes eines permetin la realització de consultes complexes a les bases de dades. En resum, un sistema OLAP es capaç de suportar requeriments complexes d'anàlisi, és capaç d'analitzar les dades des de diferents perspectives i permet treballar amb gran quantitat de dades.

  Hi ha diferents tipus de sistemes OLAP, els quals la principal diferència resideix en la forma d'accedir a les dades:

  - ***Relacional OLAP (ROLAP)*:** El motor d'aquesta tecnologia transforma dinàmicament les consultes dels usuaris en SQL on els resultats es relacionen mitjançant taules creuades i conjunts multidimensionals per tornar el resultat als usuaris.
    OLAP permet accedir directament a les base de dades relacionals, accedint en forma d'estrella en la majoria dels casos. La principal avantatja es que no hi ha limitació en el volum de les dades.
  - ***Multidimensional OLAP (MOLAP)*:** S'accedeix mitjançant base de dades multidimensional. La seva avantatja principal és la rapidesa en els temps de resposta, per contra no rendeix tan bé quan requereix carregar de nou el cub al realitzar algun canvi en les seves dimensions.

- ***Híbrida OLAP (HOLAP)*:** És una tecnologia que busca aprofitar els avantatges de les dues anteriors accedint a les dades d'alt nivell mitjançant una base de dades multidimensional i als atòmics per mitjà d'una relacional.

- **Mineria de dades (Data Mining):** Aquestes eines són utilitzades per descobrir certs patrons ocults, tendències i correlacions presentant la informació de manera senzilla i accessible als usuaris per poder preveure, solucionar o simular problemes. El data mining incorpora la utilització de tecnologies basades en arbres de decisió, xarxes neuronals, regles d'inducció, anàlisi de sèries temporals i visualització de dades que permet una extracció de coneixement útil a partir de les dades d'origen. [3]

Figura 3. Gràfic tipus d'eines BI

2.1.4 Identificació d'algunes eines actuals del mercat

Per tal d'iniciar aquest començar aquest apartat sense parlar de Gartner Inc., consultora nord-americana que mitjançant la investigació posa a l'abast informes sobre estudis relacionats amb les tecnologies de la informació.

Així doncs, el passat febrer 2016 va publicar un estudi/informe sobre les diferents solucions actuals del mercat en qüestió de BI. El quadrant màgic és un gràfic on, atenent als diferents criteris de la companyia, classifica als diferents productes analitzats en 1 dels quatre quadrants:

- **Niche players**: En fase encara inicial, que encara que tenen una bona base, els falta encara desenvolupar gran part de funcionalitat.
- **Visionaires:** Molt innovadors però amb mancança de fiabilitat i maduresa.
- **Challengers:** Productes força complets però que els falta fiabilitat.
- **Leaders:** Solucions molt completes que cobreixen la majoria de necessitats que es poden donar en l'organització. [4]

Figura 4. Magic Quadrant for Business Intelligence and Analytics Platforms

Segons la figura 3 veiem com Tableau, Qlik i Microsoft ocupen l' espai reservat per les solucions líder del mercat. Destacant en funcionalitat, fiabilitat i maduresa.

Fins aquest punt es té una visió general i es decideix cercar productes que compleixin els següents requisits:

- Que sigui multiplataforma (almenys, Windows i Mac OS)
- Llicència gratuïta (falta de pressupost per opcions amb llicenciament)

- Versió escriptori (local) com a conseqüència de la LOPD

Així doncs, es descarten ràpidament Microsoft i Tableau per no disposar de versions gratuïtes.

Es decideix investigar més profundament la versió gratuïta de **Qlikview Desktop** accedint a la seva pàgina web i contactant amb l' empresa per tal d'extreure tota la informació possible sobre el software.

La seva interfície gràfica és intuïtiva, clara i entenedora la qual cosa fa que sigui una eina fàcil en la seva utilització.

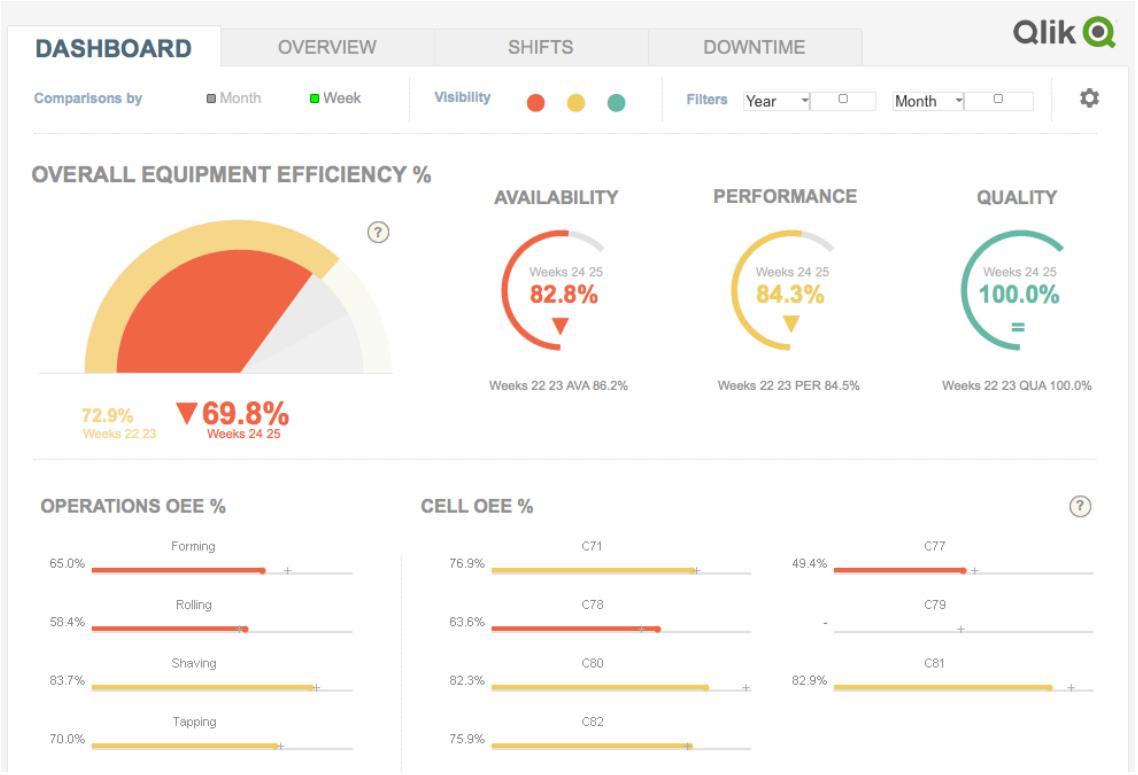


Figura 5. Interface de QlikView Desktop

La gran quantitat de funcions fa de Qlikview una eina molt completa però es desestima la seva utilització per algunes limitacions bàsiques com són:

- No està disponible per diferents Sistemes Operatius

- Només funciona amb arxius locals. [5]

Una altra possibilitat es la utilització de **JasperSoft** que disposa d' una versió "Community" gratuïta Open Source. En aquest cas, la solució és descartada al veure que dita versió no disposa dels mòduls de dashboards i la impossibilitat per emmagatzemar els canvis en els informes interactius [6]

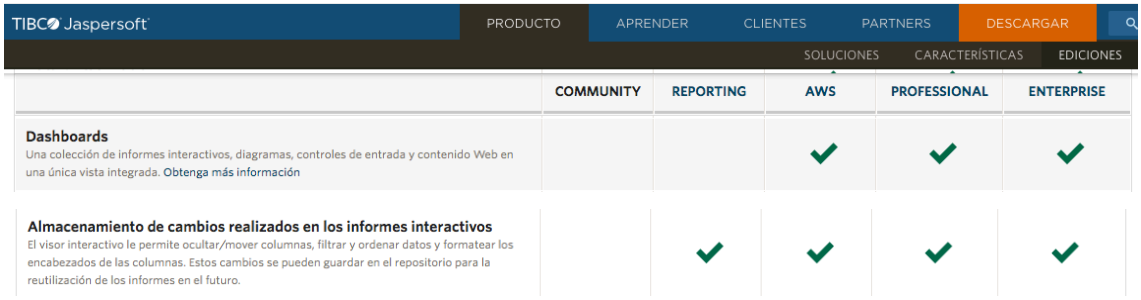


Figura 6. Opcions de Jaspersoft

Finalment, s'analitza l'opció **RapidMiner Studio** el qual proporciona una gran funcionalitat, ampliable amb diferents extensions, amb llicència gratuïta i multiplataforma.

Un altre punt a tenir en compte és el suport que rep de la comunitat i la fiabilitat d' aquest, ja que RapidMiner és la versió millorada de **Yale**. També permet la utilització d'algorismes inclosos en **Weka.**

Així doncs, analitzant els pros i contres dels paràgrafs anteriors s'arriba a la conclusió d'utilitzar RapidMiner Studio. Aquest ens ofereix funcionalitat, fiabilitat, baix cost al no haver llicenciament i permetre la instal·lació local, suport de la comunitat i una interfície gràfica força intuïtiva.

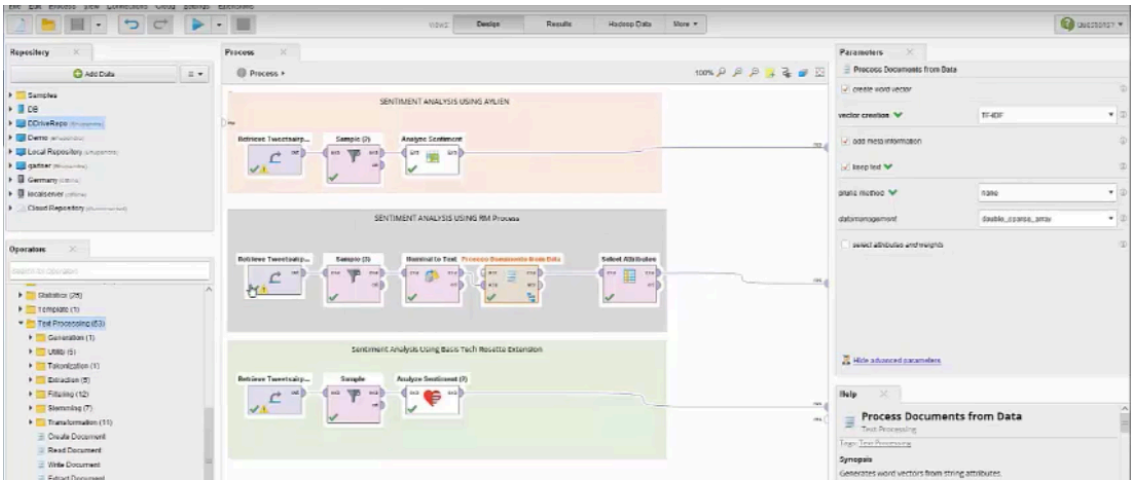


Figura 7. Mostra del panell de RapidMiner Studio

2.2 Implementació RapidMiner Studio

Rapidminer és una plataforma **Open Source**, líder del mercat en el seu sector, per a l'anàlisi de dades i mineria de dades.

La potència de Rapidminer permet la creació d' estructures senzilles de dades per a obtenció de coneixement més general, fins a la modelització de processos per avançats per fer prediccions.

La plataforma "tot en un", com ells l'anomenen en la seva pàgina web, permet accelerar la construcció de fluxos de treball d'anàlisi complets, la preparació de dades per a la modelització de la implementació de negocis en un únic entorn, millorant notablement l'eficiència i l' operativitat per l' obtenció de coneixement.

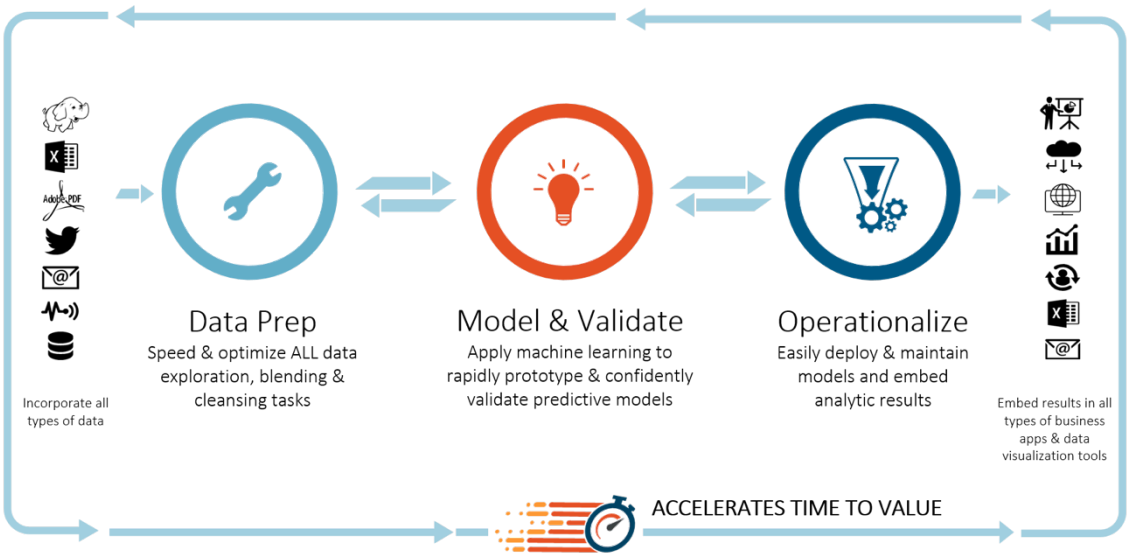


Figura 8. Esquema funcionalitat – Origen Imatge

La plataforma permet aplicar el sistema ETL, explicat en capítols anteriors, sense necessitat de recórrer a altres eines per tal finalitat.

Un cop les dades passen aquest procés ja es pot aplicar el model adequat segons objectiu i fer-ne la validació.

Finalment, els resultats poden ser lliurats en infinitat de gràfics, informes i/o arxius totalment personalitzables a les necessitats de cada usuari. Aquest últim

punt facilita en gran mesura el posterior anàlisi de resultats al mostrar les dades de forma ordenada i clara.

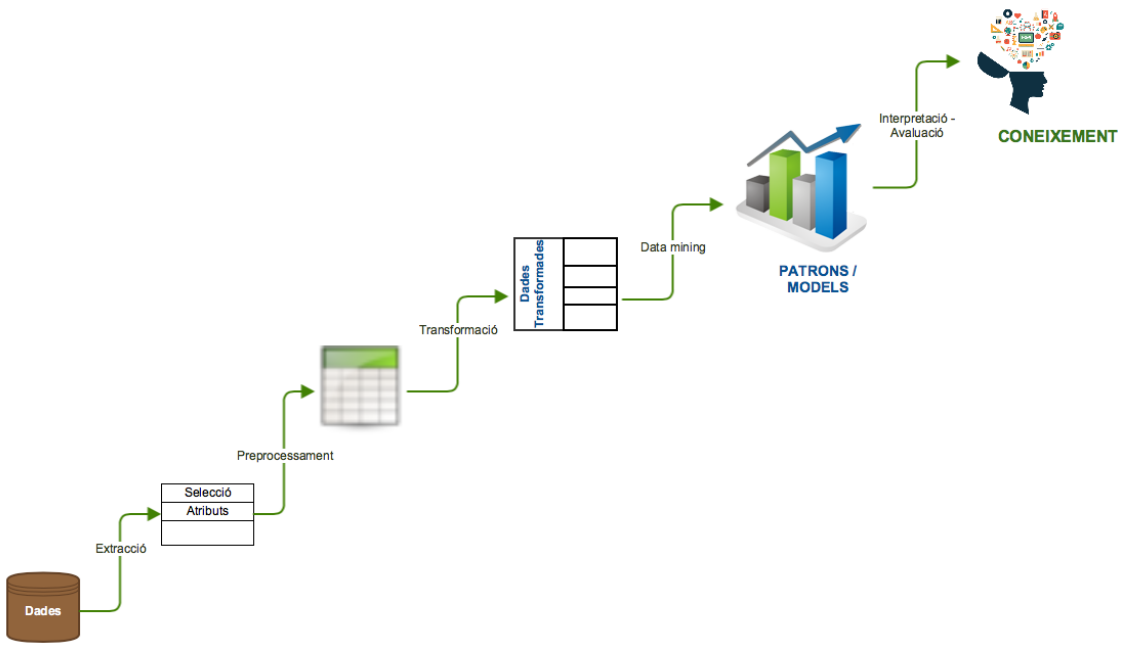


Figura 9. Gràfic propi del procés per a l'obtenció de coneixement

2.2.1 Entorn de treball

L'entorn de treball utilitzat per fer la implantació ha estat el següent:

- **Hardware**: Mac Mini Late 2012, i7 2,3 Ghz, 10 GB RAM i 1TB SSD
- **Sistema Operatiu**: Mac OS Sierra
- **Eina BI**: RapidMiner Studio v7.3.0 amb llicència gratuïta
- **Software**: Adobe Acrobat Reader DC i Microsoft Office 2011.
- **Xarxa**: Connexió Ethernet 100MBps LAN i sortida a Internet

2.2.2 Instal·lació i configuració

La instal·lació de Rapidminer Studio en l'equip de treball compta de les següents fases:

- Registre a la pàgina web de l'empresa seguint el formulari.
- Un cop el sistema valida el registre, es tria el sistema operatiu on anirà instal·lada l'aplicació.

Figura 10. Captura Web en tria de SO

- En aquest punt es descarrega i es guarda el fitxer executable Rapidminer-studio-osx-7.3.0.dmg a l'equip, s'executa mitjançant doble clic i començarà el procés d'extracció del programa amb extensió .app Un cop finalitzi sol·licitarà moure'l a la carpeta "*Aplicacions*".
- S'executa l'arxiu Rapidminer Studio.app i arranca l'aplicació amb la pantalla principal.

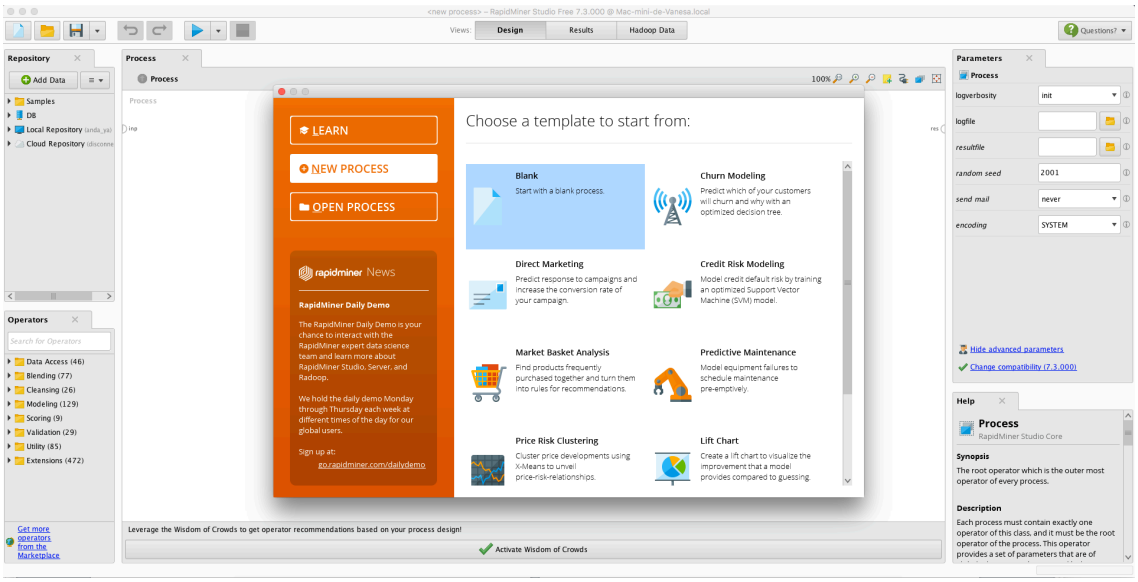


Figura 11. Captura pantalla inicial RapidMiner Studio

- Rapidminer permet la instal·lació d' extensions per ampliar les seves funcionalitats des del Marketplace. Aquesta opció no serà aplicable al nostre treball donat que les funcions que porta per defecte són suficients.

La configuració de panells que ve per defecte en RapidMIner es prou completa per no haver de canviar-la. No obstant, ofereix possibilitats de personalització.

La pestanya "Repository" contempla les opcions d'emmagatzematge de les dades. En ell es pot configurar el Data Warehouse.

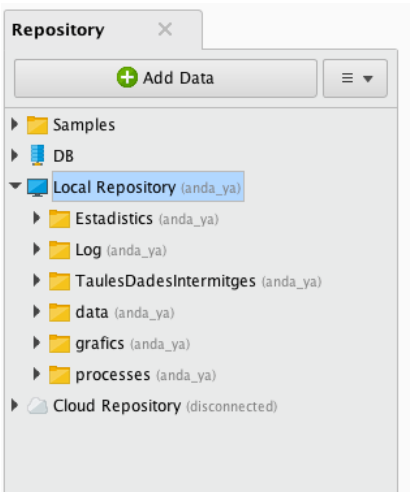


Figura 12. Captura de la pestanya de configuració Repository

En l'apartat 2.3 es mostrarà la càrrega del fitxer excel original al repositori del programa per tal de començar a tractar les dades segons procés ETL.

2.2.3 Preparació de les dades

Les dades origen disponibles són lliurades per l'equip mèdic de l'assaig clínic en un excel. El fitxer està ple de paraules i codificacions mèdiques que necessiten ser descodificats, així que hi ha una reunió per tal de fer-ne una revisió.

De la reunió se'n desprèn que les dades han estat introduïdes de manera manual a l' excel a partir de les diferents entrevistes i seguiments als grups de persones avaluades. Per aquesta raó, es detecten errors humans en la seva introducció.

El primer pas a realitzar es eliminar tots els camps que no es necessitaran pel treball i/o tenen alguna connotació, per petita que sigui, que pugi entrar amb conflicte amb la llei de protecció de dades (LOPD). Aquest pas ha reduït significativament el nombre de camps i s'ha passat d' uns 65 a uns 40 camps útils.

Un segons pas ha estat la eliminació de registres sense consistència o als quals mancaven dades claus per l'estudi. Aquest procés a deixat 520 registres.

Per últim, s´ha emmagatzemat l' arxiu resultant per a poder ser llegit des de RapidMiner.

A continuació, es detalla els camps que s'han utilitzat:

- **ID_Pacient**: *Nº identificatiu del pacient (autonumèric)*
- **Sexe**
- **Edat**
- **Nivell_estudis:** *Camp acotat a estudis primaris, estudis secundaris, estudis universitaris, Post-graus universitaris o sense estudis.*

- **Status_estudi:** *Actiu (Individu ha finalitzat l'estudi, que dura 1 any satisfactòriament i se'l considera ex fumador) o Inactiu (No ha finalitzat l'estudi perquè a recaigut o perquè ha abandonat l'assaig clínic).*
- **App_activada:** *Si (si disposen de la app al mòbil) o No.*
- **Puntuació_Richmond:** *Test que indica la motivació. Rang del 1 al 10 (essent 1 baixa i 10 màxima motivació).*
- **Puntuació_Fagerstrom:** *Test que indica la dependència a la nicotina. Rang del 0 al 10 (essent 0 sense dependència i 10 dependència molt elevada)*
- **Edat_inici_tabaquisme:** *Edat amb que van començar a fumar.*
- **Cigarretes_que_fuma_dia:** *Nombre de cigarretes o similars que fuma al dia.*
- **Marca_tabac:** *Marca que fuma de manera habitual.*
- **Intents_previs_deixar_fumar:** *Nombre d'intents que ha fet l'individu, abans d'iniciar l'estudi, per deixar de fumar.*
- **InspectionBeforeD_pes:** *Pes de l'individu en Kg el dia abans de començar amb l' estudi.*
- **InspectionBeforeD_Nivell_Co:** Carboximetria → *Nivell de CO expirat de l'individu el dia abans de començar amb l'estudi.*
- **15D_FollowInspection_pes:** *Pes de l'individu en Kg als 15 dies de començar amb l'estudi.*
- **15D_FollowInspection_Nivell_Co:** *Carboximetria als 15 dies de començar amb l'estudi.*
- **30D_FollowInspection_pes:** *Pes de l'individu en Kg als 30 dies de començar amb l'estudi.*
- **30D_FollowInspection_Nivell_Co:** *Carboximetria als 30 dies de començar amb l'estudi.*
- **3M_FollowInspection_pes:** *Pes de l'individu en Kg als 3 mesos de començar amb l'estudi.*
- **3M_FollowInspection_Nivell_Co:** *Carboximetria als 3 mesos de començar amb l'estudi.*
- **6M_FollowInspection_pes:** *Pes de l'individu en Kg als 6 mesos de començar amb l'estudi.*

- **6M_FollowInspection_Nivell_Co:** *Carboximetria als 6 mesos de començar amb l'estudi.*
- **12M_FollowInspection_pes:** *Pes de l'individu en Kg als 12 mesos de començar amb l'estudi i dia de finalització.*
- **12M_FollowInspection_Nivell_Co:** *Carboximetria als 12 mesos de començar amb l'estudi i dia de finalització.*

2.3 Definició de processos per a l'obtenció de resultats

En les diferents converses mantingudes amb l'equip d'assaig mostren la seva preocupació de la següent manera, citem textualment:

" *No disposem d' informació clara. Ara només tenim un excel amb moltes dades que no ens diuen res.* "

Per aquesta raó es proposa a l' equip que defineixin entre 4 i 5 objectius importants per a ells, i que serveixin de punt de partida quan a l'obtenció de coneixement sobre les dades emmagatzemades de l'estudi que estan realitzant.

A continuació, es defineixen els 5 objectius/qüestions sobre les quals es treballarà:

- 1er. → Donar una visió general ràpida de dades bàsiques com sexe, edat d'iniciació, cigarretes que fuma al dia, Nivell d'estudis, Status en l'estudi o intents previs per deixar de fumar.
- 2n. → Un dels objectius de l'assaig era verificar l'impacte de la gamificació per deixar de fumar. Així que a la meitat se'ls va instal·lar una app al mòbil per veure si aquesta ajudava en el procés.
- 3r. → Els centres d 'atenció primària van una mica perduts amb el llançament de campanyes antitabac i els aniria molt bé tindre dades per poder planificar-ho amb més efectivitat.
- 4rt. → Poder predir els pes i els nivells de monòxid de carboni en aire expirat i veure si hi ha canvis significatius a cap de l'any.
- 5è. → Extreure informació sobre les marques de tabac, la dependència a la nicotina i el nivell de motivació.

2.3.1 1er. Objectiu – Visió general de dades bàsiques

Per tal de poder donar informació bàsica a l'equip i que es puguin fer una idea general d' una manera ràpida es decideix utilitzar els dashboards del propi programa.

A continuació es mostra el procés per a l' obtenció de resultats:

El primer que es necessita, tal com s' havia comentat en aparats anteriors, és la configuració d' un repositori a Rapidminer on es mantindran les dades, processos i resultats.

Es prem el botó "*Add data*" i es selecciona la ubicació del fitxer origen, al tractar-se d'un excel s' obre una previsualització de les dades, on ja es pot iniciar el procés ETL de transformació, seleccionant el tipus de dada o si es vol carregar o no.

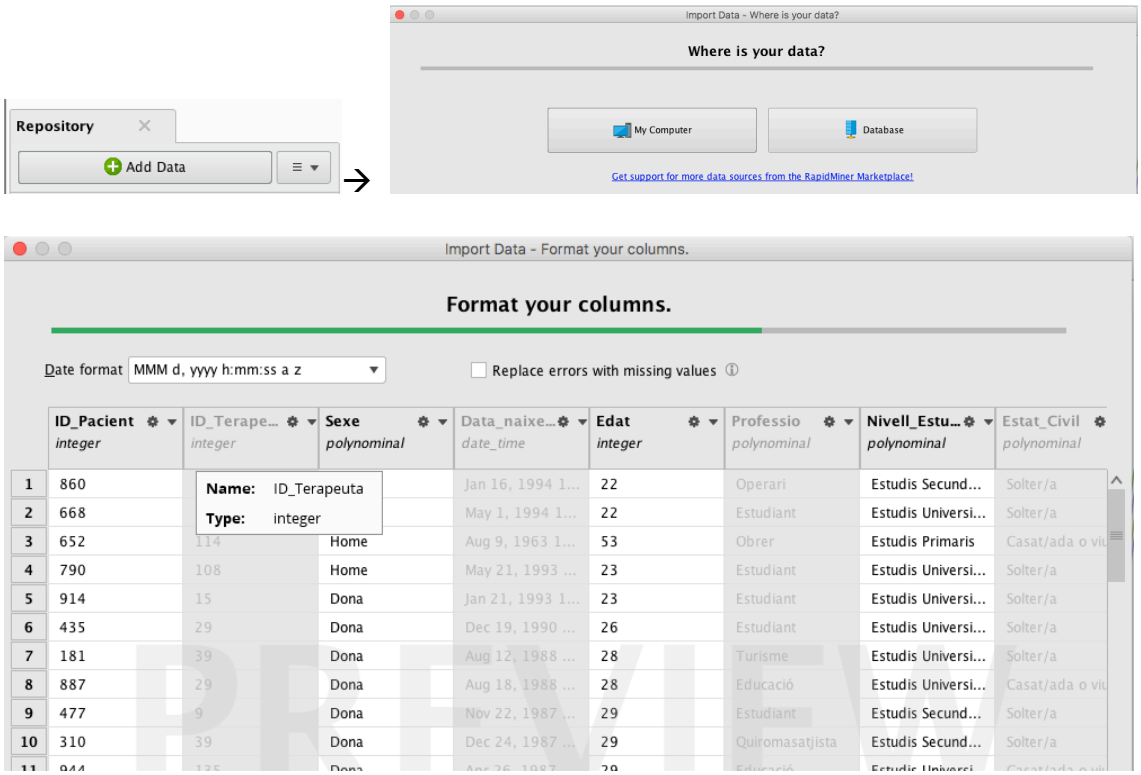


Figura 13. Configuració del repositori i extracció de dades

Al finalitzar demanarà emmagatzemar-lo al repositori per a la seva utilització.

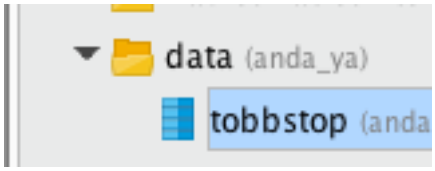


Figura 14. Captura de les dades emmagatzemades al repositori

En els següents passos s'haurà d'indicar a RapidMiner el que ha de processar. A la finestra de processos s'apliquen les instruccions que ha de seguir a mode de diagrama guiat per tal de poder executar-lo i mostrar els resultats.

Primer de tot s'ha d'afegir el camp de lectura de les dades, on s'ha carregat l'arxiu del repositori (aquest pas és necessari i comú a tots els processos següents).

Un cop feta l' extracció de les dades, en aquest primer objectiu, se seleccionen els camps necessaris per visualitzar amb l' opció "*Select Attribute*" i es connecten a la línia de sortida de resultats perquè RapidMiner pugui interpretar el procés i llençar els resultats.

La figura 15 ens mostra com queda el procés ETL complet.

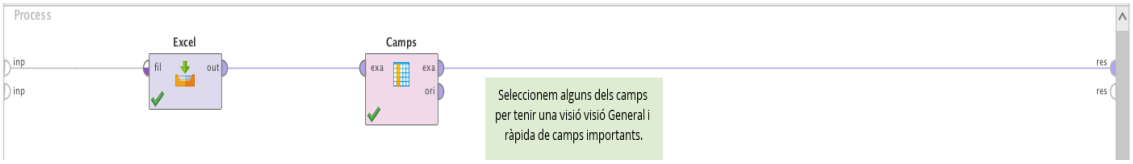


Figura 15. Procés objectiu 1

L' execució mostra els resultats a la pestanya "*Results*" i dels quals ofereix diferents visualitzacions: En forma de taula, de manera estadística segons camps o com a gràfic.
En aquest cas la visualització estadística és la que genera els dashboards que es necessiten pel primer objectiu. Les següents figures mostren el potencial que ofereixen.

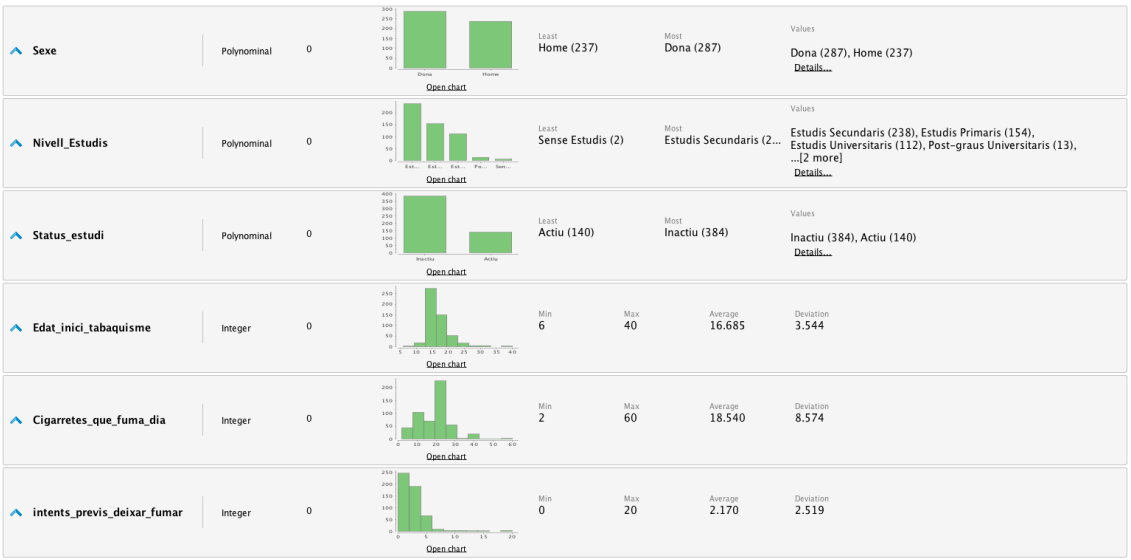


Figura 16. Dashboard objectiu 1

Des del dashboards es pot navegar directament als gràfics i visualitzar les dades d'un camp concret per facilitar la lectura. El següent gràfic, mostra informació sobre el nivell d'estudis.

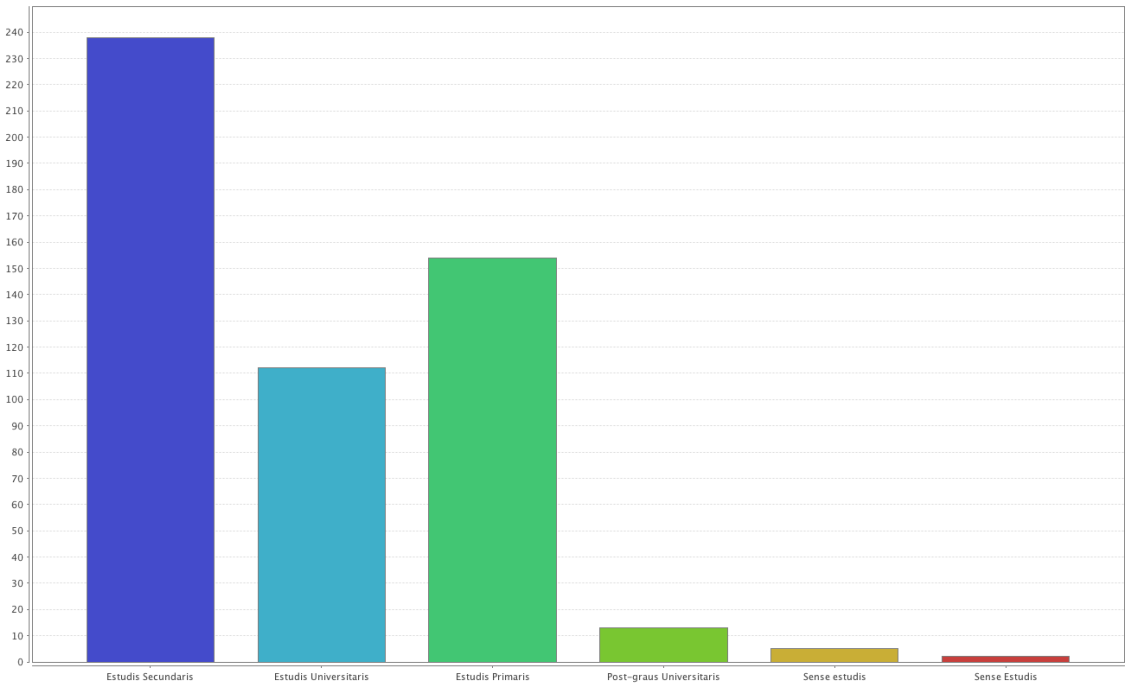


Figura 17. Gràfic del nivell d'estudis

Les dades anteriors mostren un perfil potencialment fumador: Dona amb estudis secundaris, amb edat d'iniciació des dels 13 als 16 anys i que fumen entre 20 i 25 cigarretes al dia.

2.3.2 2n. Objectiu – Impacte de la gamificació

L' assaig clínic va desenvolupar una app per dispositius mòbils i es va facilitar a la meitat de les mostres de l'estudi. Es necessita saber l'impacte que aquesta eina ha tingut en persones que han finalitzat l'estudi positivament.

Seguint amb la base del procés anterior el procés de transformació ha estat el següent:

- Discretitzar les dades corresponents al camp edat i agrupar-les per rangs d' edat:
  - [18 - 30]
  - [31 - 45]
  - [46 - 60]
  - [61 - 75]
  - [76 - 99]

- Aplicar un filtre per carregar només els que han acabat l'estudi correctament.
- Selecció d'atributs a mostrar *(Edat, app_activada)*

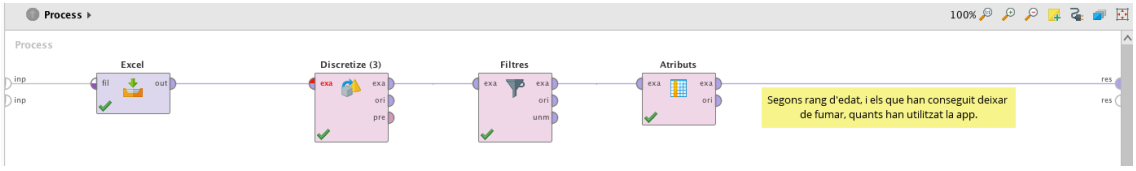


Figura 18. Procés objectiu 2

El resultat obtingut es el següent:

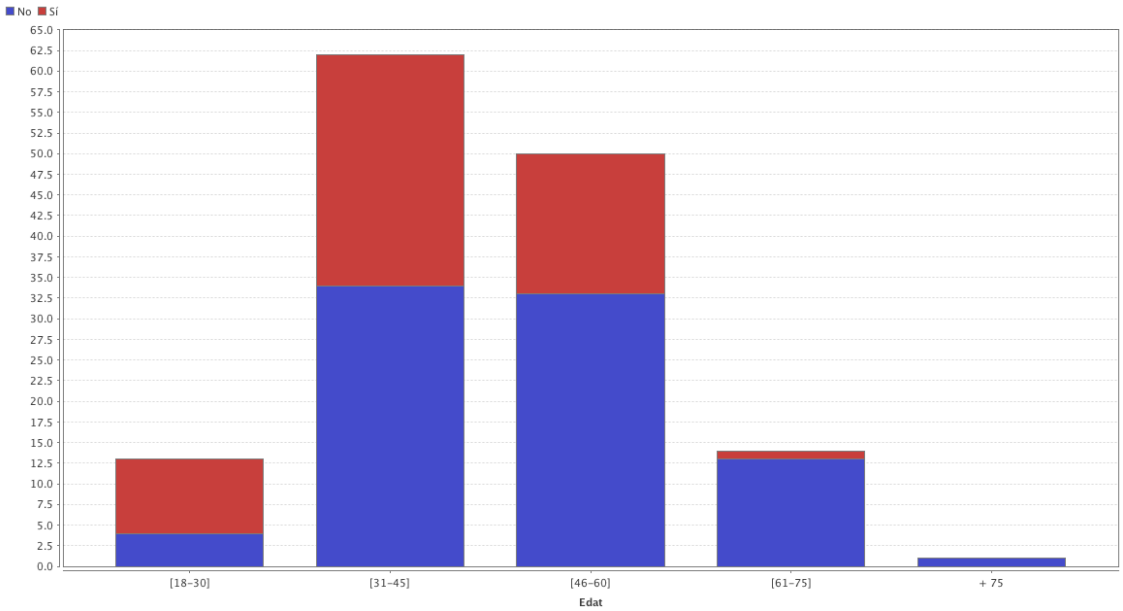


Figura 19. Gràfic impacte de la gamificació

El gràfic mostra un impacte favorable en rangs d'edat de [18-30] i en menys mesura en el rang [31-45]. L' impacte cau quan l'edat creix. És un resultat previsible si pensem que les noves tecnologies encara són força recents.

És previsible que aquestes dades canviïn en un futur.

2.3.3 3r. Objectiu – Època més favorable per realitzar la campanya antitabac des dels centre d'atenció primària (CAP)

Des dels centres d'atenció primària van una mica perduts quant a la planificació de les campanyes antitabac. Segons el INE (Institut Nacional d'estadística) la millor època per deixar de fumar són les vacances però ells noten més predisposició en altres diferents.

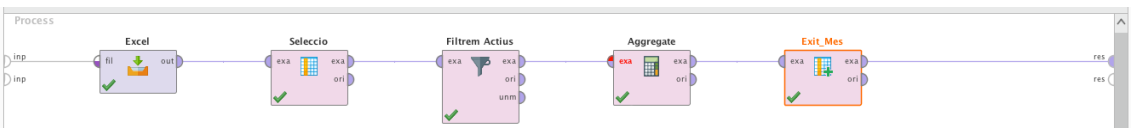


Figura 20. Procés Objectiu 3

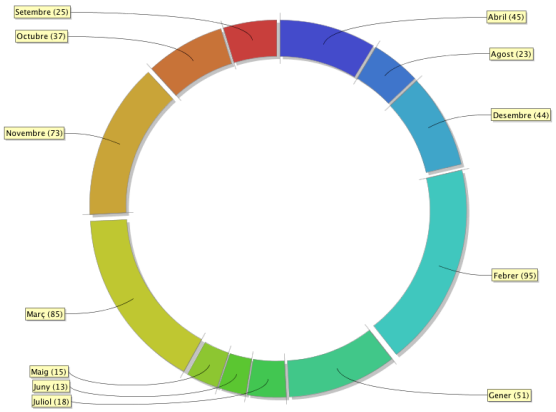
En aquest cas, s'ha portat a terme un procés per determinar quina època de l'any hi ha més gent que inicia el procés per deixar de fumar i així planificar-se les campanyes millor i que puguin ser més efectives.

Aquest procés requereix d'un *aggregate* per tal de poder agrupar el camp "*Època*" i generat un atribut nou que guardarà el resultat de contar els registres per cada època, Count(Epoca).

De manera comparativa, s'han mostrat els resultats dels que han deixat de fumar i els totals que van començar amb l'estudi i trobem que durant el primer trimestre de l'any és quan la gent inicia el deshabituallament tabàquic.

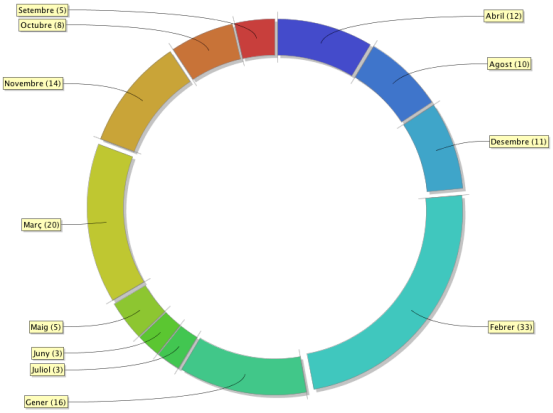
Tanmateix, si apliquem una simple regla de 3 entre el que van començar, amb els que ho van aconseguir, ens dóna un percentatge d'èxit més elevat al mes d'agost (43,48%), seguit de febrer (34,74%) i maig (33,33%)

Total individus de l'estudi



Row No.	Total	Mes
1	45	Abril
2	23	Agost
3	44	Desembre
4	95	Febrer
5	51	Gener
6	18	Juliol
7	13	Juny
8	15	Maig
9	85	Març
10	73	Novembre
11	37	Octubre
12	25	Setembre

Individus que van acabar l'estudi amb èxit



Row No.	Total	Mes
1	12	Abril
2	10	Agost
3	11	Desembre
4	33	Febrer
5	16	Gener
6	3	Juliol
7	3	Juny
8	5	Maig
9	20	Març
10	14	Novembre
11	8	Octubre
12	5	Setembre

Figura 21. Informe comparatiu objectiu 3

Si ben és cert que les dades mostren que el percentatge d'èxit és més elevat en la gent que comença a deixar de fumar en vacances, coincidint amb l'estadística de l'INE, no podem obviar que el primer trimestre de l'any és l'escollit per iniciar el procés de deixar l'hàbit tabàquic. Així doncs, recomanem una planificació de les campanyes tenint en compte aquestes dades.

2.3.4 4rt. Objectiu – Estimació de pes i Nivell CO al cap de l'any

Hi ha la creença que deixar de fumar engreixa molt a les persones i en alguns cops aquesta dada és determinant en alguns grups d'edat que estan força preocupats. Aprofitant que en les visites es fa un registre del pes i dels nivells de CO es vol obtenir una predicció al cap de l'any de deixar de fumar.

En aquest cas es comença amb el pes i es crea un procés estudiant el valor recollit en les diferents visites. Per fer una estimació del pes d'una persona en finalitzar l'any es fa servir un model predictiu estadístic, un model de regressió lineal amb validació.

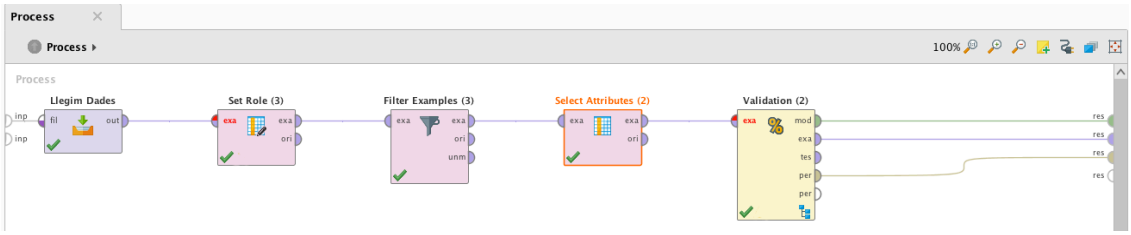


Figura 22. Procés Objectiu 4

Una vegada s'ha extret i transformat les dades de manera similar als casos anteriors s' afegeix un model de regressió lineal. Aquest constarà d' una part d'entrenament del model predictiu, del qual agafarà 10 registres a l'atzar, crearà el model i posteriorment el testejarà amb tots els altres registres. D'aquesta manera obtenim un model amb validació.

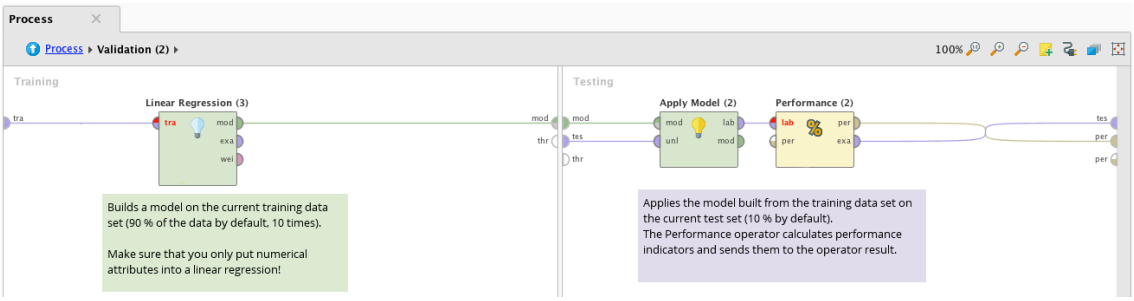


Figura 23. Model predictiu regressió lineal

A continuació ens mostrarà les dades resultant i tindrem la fórmula per poder predir el pes a partir del pes inicial (InspectionBeforeD_pes) amb una desviació d'uns 5 kg. aproximadament.

A continuació es mostren les captures amb totes les dades del model predictiu.

Attribute	Coefficient	Std. Error	Std. Coefficient	Tolerance	t-Stat	p-Value
InspectionBeforeD_pes	0.988	0.028	0.950	1	35.826	0
(Intercept)	7.128	2.051	?	?	3.475	0.001

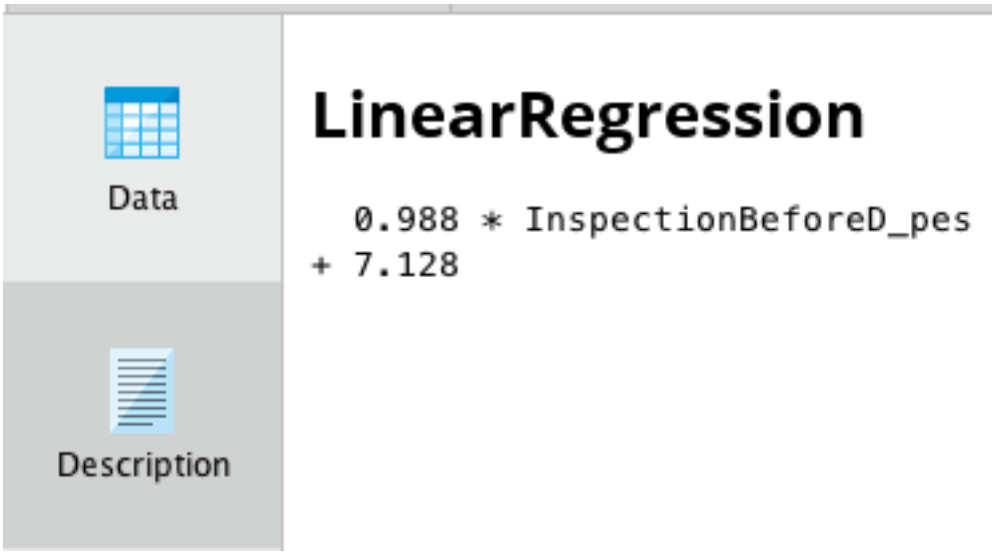


Figura 24. Captures del model predictiu pes

També s'ha realitzat un gràfic on es mostren els registres del pes al llarg de l'any i on es veu un petit augment de pes en la majoria, no gaire significatiu.

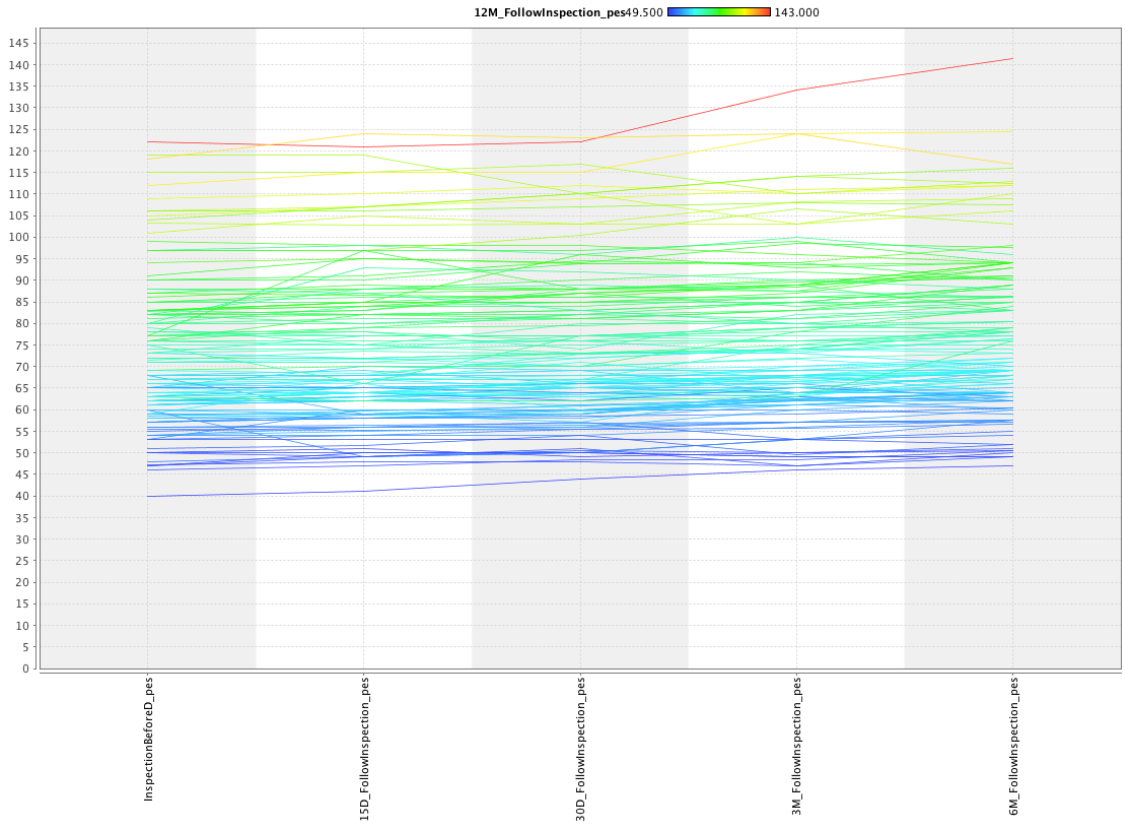


Figura 25. Gràfic de tendència de pes anual

Per validar el model i aprofitant que es disposa de les dades reals al cap de dotze mesos de deixar de fumar, s'ha realitzat la figura 26 on mostra la comparativa del model predictiu amb les dades reals.

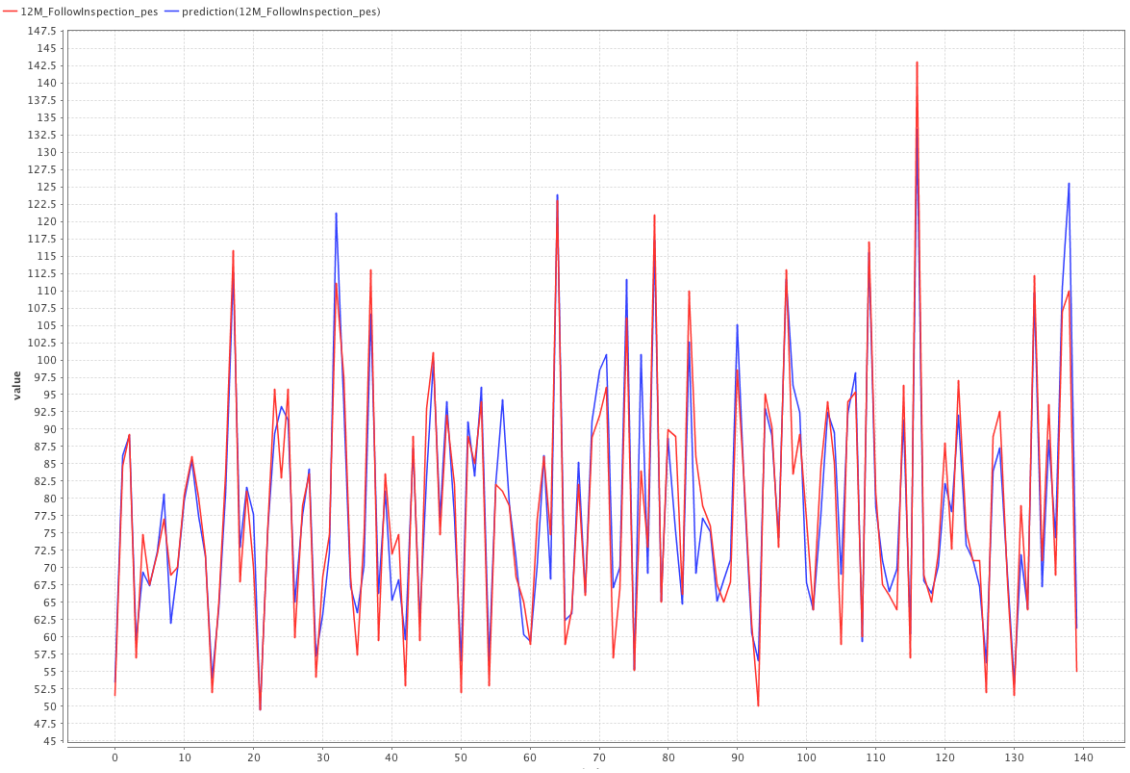


Figura 26. Gràfic comparatiu model predictiu amb dades reals de pes

A continuació, es farà gairebé el mateix per veure que passa amb el nivell de CO en l' expiració. El procés utilitzat és el mateix que per al pes, però en aquest cas s'ha fet servir una model de regressió lineal múltiple tenint en compte tots els registres de les visites. A continuació es mostren els resultats:

Attribute	Coefficient	Std. Error	Std. Coefficient	Tolerance	t-Stat	p-Value
InspectionBeforeD_Nivel...	0.008	0.004	0.132	0.972	2.154	0.033
15D_FollowInspection_...	-0.008	0.012	-0.046	0.971	-0.667	0.506
30D_FollowInspection_...	-0.004	0.018	-0.015	0.879	-0.200	0.842
3M_FollowInspection_Ni...	0.032	0.030	0.087	0.558	1.055	0.293
6M_FollowInspection_Ni...	0.380	0.049	0.641	0.542	7.755	0.000
(Intercept)	-0.098	0.064	?	?	-1.539	0.126

data in a table

PerformanceVector

```
PerformanceVector:
root_mean_squared_error: 0.385 +/- 0.181 (mikro: 0.425 +/- 0.000)
squared_error: 0.181 +/- 0.132 (mikro: 0.181 +/- 0.544)
```

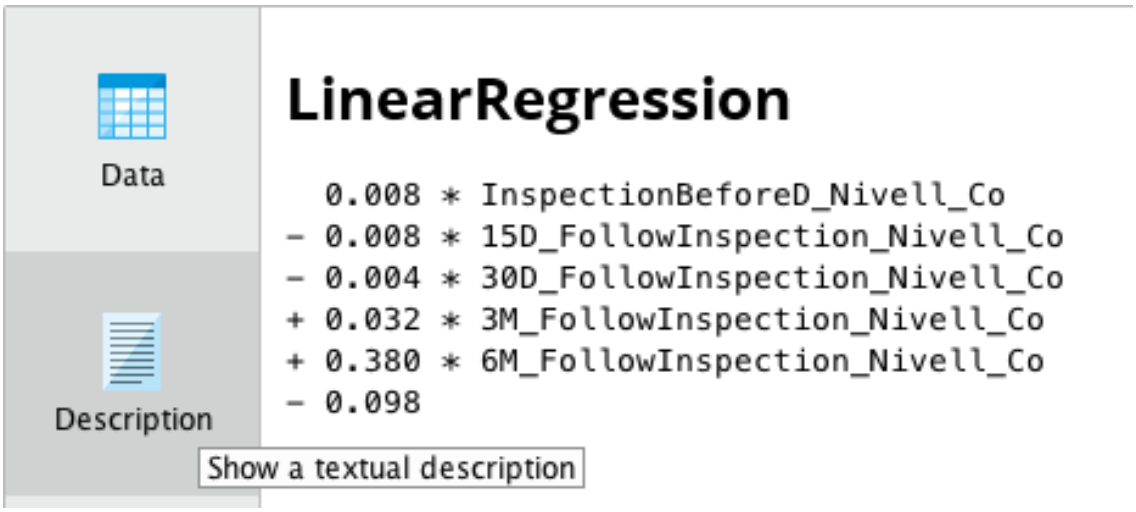


Figura 27. Captures del model predictiu nivell CO

Tal i com s'havia fet amb el pes, es mostra gràfic per veure la tendència .

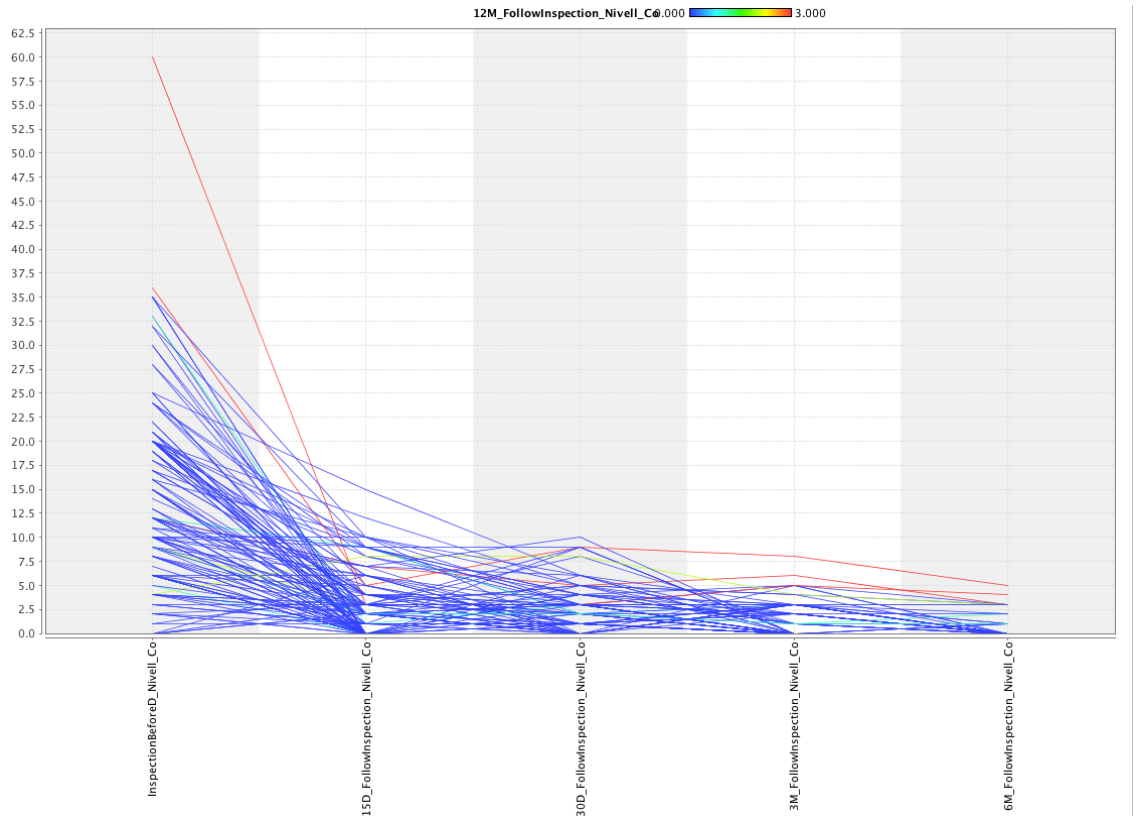


Figura 28. Gràfic de tendència del nivell CO

En resum, els models realitzats mostren com al cap de l'any hi ha una tendència a guanyar una mica de pes, no gaire rellevant. En canvi en el nivell de CO es veu una tendència baixista molt més marcada, sobretot els 15 primers dies.

2.3.5 5è. Objectiu – Relació entre les maques de tabac, dependència i nivell de motivació per deixar de fumar

El darrer objectiu planteja una relació entre les maques de tabac, amb l'índex de dependència a la nicotina (Test de Fagerstrom) i l'índex de motivació (Test de Richmond).

En aquest cas s'ha decidit crear un model basat en un arbre de decisió, on es relacionarà les tres variables que ha demanat l'equip, juntament amb el sexe.

El procés plantejat és el següent:

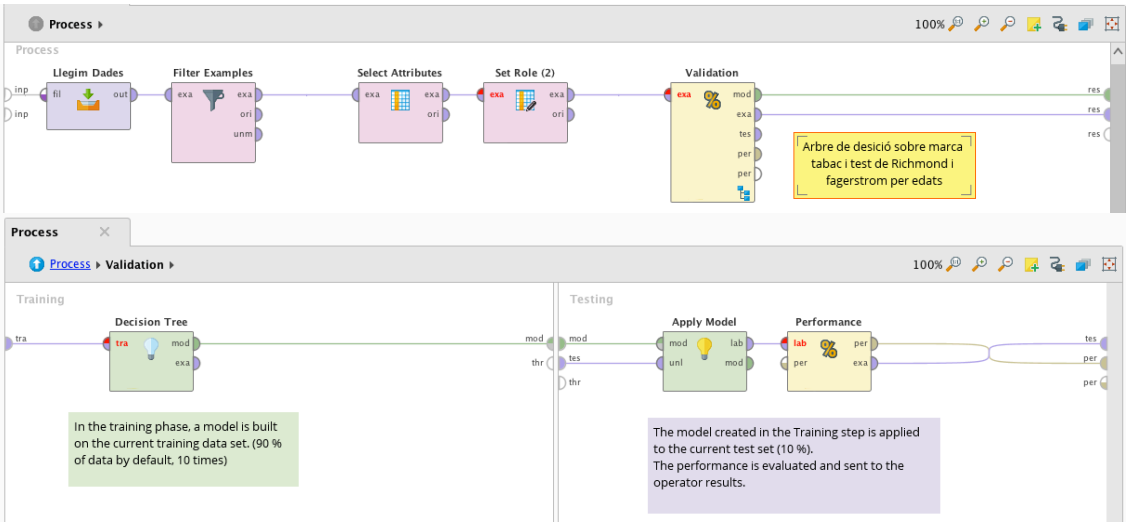


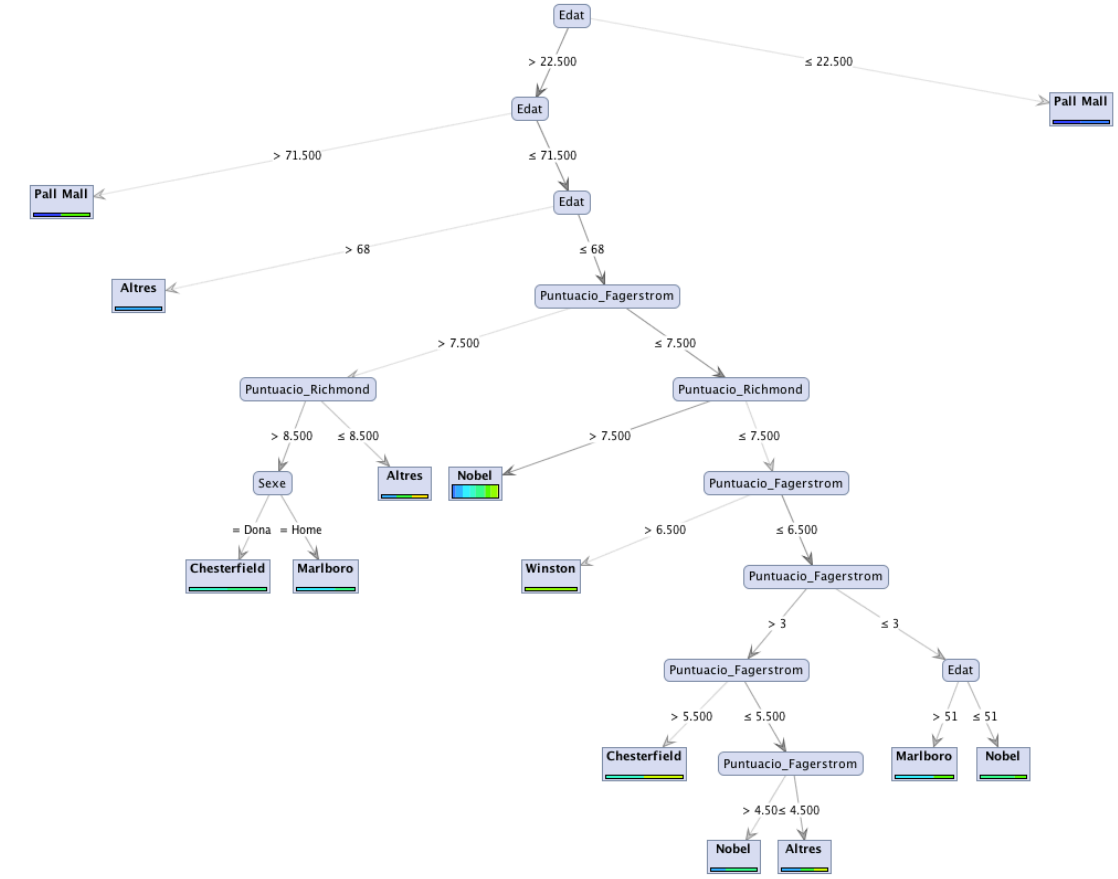
Figura 29. Procés Objectiu 5

Es tracta d'un model validat després de fer l' extracció i transformació de les dades com en processos anteriors.

El filtratge i la selecció dels atributs que s'ha utilitzat és: marca tabac, sexe, edat, puntuació Richmond i puntuació Fagerstrom.

Aquest model també incorpora el procés d'entrenament, el qual agafa 10 registres aleatòriament i després fa el testeig, com al model anterior.

Els resultats són els següents:



Tree

```
Edat > 22.500
|   Edat > 71.500: Pall Mall {Pall Mall=1, Tabac d'embolicar (50g)=0, Altres=0, Marlboro=0, Chesterfield=0, Nobel=0, Camel=0, Ducados=1, Winston=0, Lucky Strike=0, Fortuna=0, Celtas=0, Gold Coast=0, Davidoff=0}
|   Edat ≤ 71.500
|   |   Edat > 68: Altres {Pall Mall=0, Tabac d'embolicar (50g)=0, Altres=2, Marlboro=0, Chesterfield=0, Nobel=0, Camel=0, Ducados=0, Winston=0, Lucky Strike=0, Fortuna=0, Celtas=0, Gold Coast=0, Davidoff=0}
|   |   Edat ≤ 68
|   |   |   Puntuacio_Fagerstrom > 7.500
|   |   |   |   Puntuacio_Richmond > 8.500
|   |   |   |   |   Sexe = Dona: Chesterfield {Pall Mall=0, Tabac d'embolicar (50g)=0, Altres=0, Marlboro=0, Chesterfield=1, Nobel=1, Camel=0, Ducados=0, Winston=0, Lucky Strike=0, Fortuna=0, Celtas=0, Gold Coast=0, Da
|   |   |   |   |   Sexe = Home: Marlboro {Pall Mall=0, Tabac d'embolicar (50g)=0, Altres=0, Marlboro=2, Chesterfield=0, Nobel=1, Camel=0, Ducados=0, Winston=0, Lucky Strike=0, Fortuna=0, Celtas=0, Gold Coast=0, Davido
|   |   |   |   Puntuacio_Richmond ≤ 8.500: Altres {Pall Mall=0, Tabac d'embolicar (50g)=0, Altres=1, Marlboro=0, Chesterfield=0, Nobel=0, Camel=1, Ducados=0, Winston=0, Lucky Strike=0, Fortuna=1, Celtas=0, Gold Coast=
|   |   |   Puntuacio_Fagerstrom ≤ 7.500
|   |   |   |   Puntuacio_Richmond > 7.500: Nobel {Pall Mall=0, Tabac d'embolicar (50g)=9, Altres=16, Marlboro=17, Chesterfield=12, Nobel=20, Camel=2, Ducados=12, Winston=15, Lucky Strike=2, Fortuna=0, Celtas=0, Gold (
|   |   |   |   Puntuacio_Richmond ≤ 7.500
|   |   |   |   |   Puntuacio_Fagerstrom > 6.500: Winston {Pall Mall=0, Tabac d'embolicar (50g)=0, Altres=0, Marlboro=0, Chesterfield=0, Nobel=0, Camel=0, Ducados=0, Winston=2, Lucky Strike=0, Fortuna=0, Celtas=0, Gold
|   |   |   |   |   Puntuacio_Fagerstrom ≤ 6.500
|   |   |   |   |   |   Puntuacio_Fagerstrom > 3
|   |   |   |   |   |   |   Puntuacio_Fagerstrom > 5.500: Chesterfield {Pall Mall=0, Tabac d'embolicar (50g)=0, Altres=0, Marlboro=0, Chesterfield=1, Nobel=0, Camel=0, Ducados=0, Winston=0, Lucky Strike=1, Fortuna=0, (
|   |   |   |   |   |   |   Puntuacio_Fagerstrom ≤ 5.500
|   |   |   |   |   |   |   |   Puntuacio_Fagerstrom > 4.500: Nobel {Pall Mall=0, Tabac d'embolicar (50g)=0, Altres=1, Marlboro=0, Chesterfield=0, Nobel=2, Camel=0, Ducados=0, Winston=0, Lucky Strike=0, Fortuna=0, Celt
|   |   |   |   |   |   |   |   |   Puntuacio_Fagerstrom ≤ 4.500: Altres {Pall Mall=0, Tabac d'embolicar (50g)=0, Altres=3, Marlboro=0, Chesterfield=0, Nobel=0, Camel=2, Ducados=0, Winston=0, Lucky Strike=2, Fortuna=0, Cel
|   |   |   |   |   |   Puntuacio_Fagerstrom ≤ 3
|   |   |   |   |   |   |   Edat > 51: Marlboro {Pall Mall=0, Tabac d'embolicar (50g)=0, Altres=0, Marlboro=2, Chesterfield=0, Nobel=0, Camel=0, Ducados=1, Winston=0, Lucky Strike=0, Fortuna=0, Celtas=0, Gold Coast=0,
|   |   |   |   |   |   |   Edat ≤ 51: Nobel {Pall Mall=0, Tabac d'embolicar (50g)=0, Altres=0, Marlboro=0, Chesterfield=0, Nobel=3, Camel=0, Ducados=1, Winston=0, Lucky Strike=0, Fortuna=0, Celtas=0, Gold Coast=0, Dav
Edat ≤ 22.500: Pall Mall {Pall Mall=1, Tabac d'embolicar (50g)=1, Altres=0, Marlboro=0, Chesterfield=0, Nobel=0, Camel=0, Ducados=0, Winston=0, Lucky Strike=0, Fortuna=0, Celtas=0, Gold Coast=0, Davidoff=0}
```

Figura 30. Arbre de decisió i esquema.

De manera addicional es fa un gràfic de barres per veure la relació entre les marques de tabac i l' addicció que presentaven els fumadors.

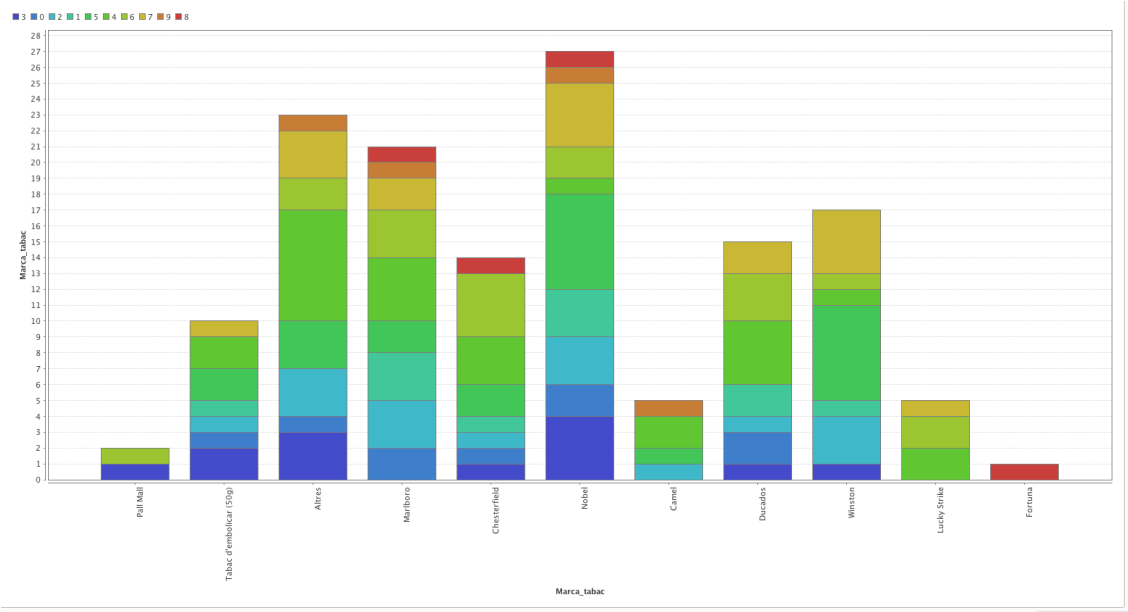


Figura 31. Gràfic addició segons marca de tabac

Amb l'arbre de decisió ja es pot veure quines marques creen menys addicció per edats i sexe, com la Pall Mall. Això pot portar a pensar que algunes marques com Winston, Marlboro o Nobel fiquen a les seves cigarretes certs productes o químics que les fan més addictives al consumidor.

Aleshores, es pot relacionar que segons la marca de tabac que estigui consumint una persona que vol deixar de fumar, tindrà més o menys addicció i com a conseqüència la dificultat en el deshabituallament tabàquic variarà.

***Nota addicional**: En cap cas es vol donar a entendre que una marca és millor que una altra. Totes creen addicció i són perjudicials per la salut.*

3. Conclusions

Aquest treball ha fet que tinguem una visió molt més acurada de tot el que implica un projecte de Business Intelligence. Encara que som conscients que aquest és només una introducció del que seria un projecte real, ens ha permès descobrir, les fases i la terminologia pròpia i fer un petit salt de la part teòrica, apressa en assignatures de sistemes de la informació.

Hem après que el Business Intelligence com a tal no arregla res o no suposa una millora en si mateixa, sinó que és un medi per tal d'arribar a un fi concret.

Donat que era el primer projecte sobre BI que realitzàvem aquest no té una dificultat en si mateix molt elevat però sí que dóna resposta a un cas real com és el problema de l'equip d'assaig clínic.

Els objectius marcats s'han assolit en termes generals i s'ha pogut proporcionar coneixement que en un principi no es tenia. L' únic objectiu pendent ha estat la formació a personal de l'assaig perquè pugui fer alguns processos simples a RapidMiner Studio. La falta de temps per ambdues parts no ho ha fet possible, almenys fins a la data.

D'altra banda, s'ha anat seguit la planificació de manera força fidel, introduint els canvis que el consultor va proposar donada la seva experiència per garantir l'èxit del TFG. I que podem afirmar ara com ara, de manera encertada, ja que en un principi es va valorar l'opció de fer els dashboards amb el propi excel.

Com a línies futures d'exploració seria l'aprenentatge més precís del programa RapidMiner i de la gran quantitat d'opcions que té. Així com la investigació de sistemes dedicats a la ETL.

4. Glossari

Bussiness Intelligence (BI)
Conjunt de processos, tecnologia i estratègies enfocades a aportar coneixement i així ajudar a la presa de decisions.

SQL
De l'anglès Structured Query Language

Open source
Software desenvolupat i distribuït lliurement, on el codi es obert a tothom.

ETL
De l'anglès Extact, Transform and Load. Procés d' extracció, transformació i càrrega de dades utilitzat en entorns BI seguint les pautes i models de dades que marca el Data Warehouse

Data Warehouse
Base de dades o repositori destinat a emmagatzemar la informació de diversos orígens i sistemes dissenyat específicament per mantenir històrics, mineria de dades, anàlisis multidimensionals, etc.

Dashboards
Es pot denominar Quadre de comandament i informa de l'evolució de paràmetres fonamentals.

KPI
De l'anglès Key Performance Indicator, són claus de rendiment.

OLAP
De l'anglès On-Line Analytical Processing. Eines que permeten la realització de consultes complexes a les base de dades.

ROLAP
De l'anglès Relational On-Line Analytical Processing. Amb capacitats OLAP accedeixen a la base de dades relacional

MOLAP
De l'anglès Multidimensional On-Line Analytical Processing. Amb capacitats OLAP accedeixen a bases de dades multidimensionals.

HOLAP
De l'anglès Hybrid On-Line Analytical Processing. Aprofita les avantatges de les ROLAP i les MOLAP

LOPD
Llei orgànica de Protecció de Dades

ERP
De l'anglès Enterprise Resource Planning, son sistemes que integren els diferents àmbits que es poden donar en una organització com producció, distribució, vendes, etc.

CRM
De l'anglès Costumer Relationship Management, model de gestió basada en la satisfacció amb els clients.

YALE
Programa Open Source per a projectes de BI

WEKA
Plataforma de software per a la mineria de dades i l'aprenentatge automàtic de la universitat de Waikato

5. Bibliografia

[1] **Web:** Sinnexus, Octubre 2016
URL: http://www.sinnexus.com/business_intelligence/

[2] **Web:** Gartner, Novembre 2016
URL: http://www.gartner.com/it-glossary/business-intelligence-bi

[3] **Web:** AMJTelecom, Novembre 2016
URL: http://www.amjtelecom.com/bi_componentes.php

[4] **Web:** Gartner, Octubre 2016
URL: https://www.gartner.com/doc/reprints?id=12XXET8P&ct=160204

[5] **Web:** Qlik Community, Novembre 2016
URL: https://community.qlik.com/docs/DOC-1799

[6] **Web:** JasperSoft, Novembre 2016
URL: https://www.jaspersoft.com/es/node/77688

Altra bibliografia consultada:

- Introducción al Business Intelligence - Jordi Conesa / Josep Curto - Editorial UOC

- **Web:** Rapidminer, Novembre 2016
  URL: https://rapidminer.com

- **Web**: Business Intelligence: Competir con información, Cano, J. L. Banesto, Fundación Cultural. Octubre 2016
  URL: http://itemsweb.esade.edu/biblioteca/archivo/Business_Intelligence_competir_con_informacion.pdf

- **Web:** Wikipedia, Conceptes: *Almacén de datos.* Octubre 2016
  URL: https://es.wikipedia.org/wiki/Almacén_de_datos

- **Web:** IBM. Creación de árboles de decisión. Novembre 2016
  URL:http://www.ibm.com/support/knowledgecenter/es/SSLVMB_22.0.0/com.ibm.spss.statistics.help/spss/tree/idh_idd_treegui_main.htm

6. Annexos

Aquest TFG no ha generat annexos rellevant per ser adjunts.