Diseño de una herramienta *business inteligence* para el análisis de encuestas de salud

Autora: Estefanía Piñeiro Crespo

Consultores: Dra. Laia Subirats Maté y Dr. Jordi Conesa Caralt

Profesoras responsables de la asignatura: Dra. Maria Pujol Jover y Dra. Teresa Sancho Vinuesa

Febrero/2018

"Tuve un gesto de desaliento: es absurdo buscar un pozo, al azar, en la inmensidad del desierto. Sin embargo, nos pusimos en marcha. Después de haber caminado durante horas en silencio, cayó la noche y las estrellas comenzaron a iluminarse. Yo las entreveía como en sueños, al tener un poco de fiebre a causa de mi sed. Las palabras del principito bailaban en mi memoria:
- Entonces tú también tienes sed ? – le pregunté.
Pero no respondió a mi pregunta. Simplemente me dijo:
- El agua puede ser buena también para el corazón…
No comprendí su respuesta pero me callé… Ya sabía que no había que interrogarlo.
Estaba cansado y se sentó. Yo me senté a su lado. Y, después de un silencio, agregó:
- Las estrellas son bellas, a causa de una flor que no se ve…
Respondí "desde luego" y miré, sin hablar, las ondulaciones de la arena bajo la luna.
- El desierto es bello… – agregó.
Y era verdad. A mí siempre me gustó el desierto. Uno se sienta sobre una duna de arena. No se ve nada. No se escucha nada. Y sin embargo hay algo que irradia en silencio…
- Lo que hace al desierto tan bello – dijo el principito – es que esconde un pozo en algún lado…"

(El PRINCIPITO, capitulo 24)

A mi hijo David, a mi marido Luis y a esa gran estrella que tanto brilla en nuestros corazones y nos ha enseñado lo más importante de la vida, saber vivirla.

Al Dr. Manuel Armayones Ruiz por ser fuente de inspiración e ideólogo.

A mis profesores la Dra. Laia Subirats Maté y el Dr. Jordi Conesa Caralt, por haber sido mi guía en este TFM, por su energía y apoyo durante este tiempo.

Al Sr. Josep Maria Suelves Responsable de la Unitat de Coordinació a la Direcció General de Salut Pública del Departament de Salut de la Generalitat de Catalunya.

A las Sras. Anna Schiaffino Rubinat, Antonia Medina Bustos y Anna Mompart Penina, del Servei de Gestió i Anàlisi de la Informació per a la Planificació Estratègica (SGAIPE). Subdirecció General de Planificació Sanitària i Professional. Direcció General de Planificació en Salut del Departament de Salut de la Generalitat de Catalunya.

Esta obra está sujeta a una licencia de Reconocimiento - No Comercial - Sin Obra Derivada 3.0 España de Creative Commons

Ficha del trabajo final

Título del trabajo:	Diseño de una herramienta business inteligence interactiva para el análisis de las encuestas de salud
Nombre del autor:	Estefanía Piñeiro Crespo
Nombre del consultor/a:	Dra. Laia Subirats Maté Dr. Jordi Conesa Caralt
Nombre del PRA:	Dra. Maria Pujol Jover Dra. Teresa Sancho Vinuesa
Fecha de entrega:	02/2018
Titulación:	*Master Inteligencia de negocios y Big Data*
Área del Trabajo Final:	*Trabajo Final de máster MIB*
Idioma del trabajo:	*Castellano*
Palabras clave	*Salud autopercibida, Business Intelligence, Cuadro de mandos, Árboles de decisiones*

Resumen del Trabajo: ***contexto de aplicación, metodología, resultados y conclusiones del trabajo.***

Ante los crecientes desafíos de la globalización económica cada vez más gobiernos y agentes públicos aumentan su interés por conocer, comprender y observar las estructuras dinámicas de un territorio para la toma de decisiones, así como fomentar la trasparencia y buen gobierno. Estas decisiones resultan altamente complejas debido a la gran cantidad de información generada. Por este motivo es necesario disponer de métodos para extraer y transformar los datos en información, y herramientas para distribuir el conocimiento generado, en el momento adecuado y a los usuarios interesados.

Los cuadros de mando son la herramienta adecuada, ayudando a la conversión de la información en indicadores que generan conocimiento.

El trabajo se basa en la creación de un prototipo de una herramienta Business Intelligence Interactiva para explorar y explicar datos de las encuestas, sobre indicadores demográficos, económicos y de bienestar. Esta herramienta se crea con el software *Tableau*.

En el proceso de descubrimiento de conocimiento de grandes volúmenes de datos es vital escoger las variables y características más adecuadas, por ello aplicamos la extracción de características de los datos y realizamos una clasificación a través de los árboles de decisión con el entorno R, con el objetivo de encontrar el mejor modelo de clasificación en función de la variable salud autopercibida.

La salud autopercibida es una medida muy importante que influye sobre la mortalidad y proporciona información sobre el estado de salud.

Abstract:

In the face of the growing economic globalization challenges, an increasing number of Governments and public agents are showing more interest in knowing, understanding and observing the territorial dynamic structures to carry out actions. These decisions end up being highly complex due to the amount of information produced. Therefore, it is necessary to have methods to extract and transform data to turn it into useful information and tools to distribute the generated knowledge at the right time and to the interested users.

Dashboards are suitable for this task, helping when it comes to converting information into knowledge generating indicators.

This work is based on the creation of a Business Intelligence prototype tool built for exploring and explaining Catalonia's survey data, about welfare indicators. This tool is created by the Tableau software.

In this process, that is carried upon a large amount of data, it is essential to choose appropriate variables and characteristics. Therefore, data characteristics are extracted and classified by decision trees in the R context, the aim of this is to classify them according to the level of healthiness each individual perceives on themselves, since it is a very important factor that influences mortality and provides information on the actual healthiness.

The self-rated health is an important measure which influences over mortality and provides information on health condition.

Tabla de contenido

1-INTRODUCCIÓN 9

1.1-CONTEXTO Y JUSTIFICACIÓN DEL TRABAJO 9
1.2-OBJETIVOS DEL TRABAJO 12
1.3- ENFOQUE Y MÉTODO A SEGUIR 12
1.4- PLANIFICACIÓN DEL TRABAJO 14
1.5-BREVE RESUMEN DE LOS PRODUCTOS OBTENIDOS 17
1.6-BREVE DESCRIPCIÓN DE LOS CAPÍTULOS DE LA MEMORIA 17

2-ESTADO DE ARTE 20

2.1- SALUD AUTOPERCIBIDA 20
2.2-INTELIGENCIA DE NEGOCIO 21
2.2.1 CONCEPTOS Y ELEMENTOS 21
2.3-VISUALIZACIÓN DE DATOS 27
2.3.1 - ¿QUÉ ES VISUALIZACIÓN DE LA INFORMACIÓN? 27
2.3.2 - BENEFICIOS DE LA VISUALIZACIÓN 27
2.3.3 - PRINCIPIOS DE DISEÑO 28
2.4- CUADROS DE MANDO 28
2.4.1 - CUADRO DE MANDOS, SU IMPORTANCIA Y REQUISITOS 28
2.4.2 - HERRAMIENTAS DISPONIBLES EN EL MERCADO 29
2.5-TABLEAU 30
2.5.1 - ¿QUÉ ES TABLEAU? 30
2.6-CASO DE ESTUDIO: ENCUESTA DE SALUD DE CATALUNYA 31
2.6.1- ¿QUÉ ES LA ESCA?, OBJETIVOS Y FINALIDADES 31
2.6.2-CONTENIDO DE LAS ENCUESTAS DE SALUD DE CATALUNYA 31
2.6.3-RESULTADOS 31

3-EVALUACIÓN DE LA INFRAESTRUCTURA 32

3.1- EVALUACIÓN DE LA ESTRUCTURA EXISTENTE 32
3.2-SELECCIÓN DE PRODUCTOS. 32

4- ANÁLISIS DE DATOS A TRAVÉS DE ALGORITMOS "ESTRUCTURALMENTE SENCILLOS" COMO LOS ÁRBOLES DE DECISIÓN. 33

4.1 OBJETIVO DEL ANÁLISIS 33
4.2 FUENTE DE DATOS 33
4.3 DEFINICIÓN DEL PROBLEMA 33
4.3.1 ELABORACIÓN DEL PRIMER MODELO DE CLASIFICACIÓN 34
4.3.2 PODA DEL ÁRBOL DE DECISIÓN 37
4.3.3 *BOOSTING* 39
4.4 EXTRACCIÓN DE CARACTERÍSTICAS 41
4.4.1 TÉCNICAS *FILTER* 41
4.4.2 TÉCNICAS WRAPPER 43

5. DISEÑO, CONSTRUCCIÓN DEL ALMACEN DE DATOS PARA EL ANÁLISIS DE ENCUESTAS DE LA ENCUESTA DE SALUD DE CATALUNYA. (ESCA) 49

5.1-REQUERIMIENTOS 49
5.1-2- NECESIDADES DE *REPORTING* 49
5.1.3- TIPOS DE USUARIOS 50

5.1.4- CARACTERÍSTICAS QUE HA DE REUNIR EL SISTEMA DE *REPORTING* 50
5.2- ANÁLISIS DE LOS DATOS. 51
5.2.1- ANÁLISIS DE LOS DATOS CEDIDOS ESCA_2016 52
5.3-DISEÑO CONCEPTUAL 59
5.3.1- TABLA DE HECHOS 59
5.3.2- DIMENSIONES 59
5.4-DISEÑO FÍSICO 60

6. CREACIÓN DE UN CUADRO DE MANDOS PARA EL ANÁLISIS DE ENCUESTAS DE LA ENCUESTA DE SALUD DE CATALUNYA. (ESCA) 65

6.1-CONECTAR LA BASE DE DATOS CON TABLEAU 65
6.2 CREAR VISUALIZACIONES 66
6.2.1-GRÁFICO DE BARRAS SALUD AUTOPERCIBIDA 66
6.2.2-GRÁFICO SEXO Y EDAD 67
6.2.3-GRÁFICO ESTUDIOS 67
6.2.4-GRÁFICA CLASE SOCIAL 68
6.2.5-MAPA 68
6.2- ELABORACIÓN Y PUBLICACIÓN EN INTERNET DEL CUADRO DE MANDO O *DASHBOARD* EN *TABLEAU* 69

7-CONCLUSIONES 70

7.1-ANÁLISIS DE DATOS DE LA BASE DE DATOS DE LA ENCUESTA DE LA COMUNIDAD DE MADRID. 70
7.2-BASE DE DATOS PREPARADOS PARA CONECTAR CON TABLEAU. 71
7.3-HERRAMIENTA DE *REPORTING* INTERACTIVA. 71
7.4-UNA VISTA GLOBAL. 72

8- TRABAJOS FUTUROS 72

9- ANEXOS 74

9.1 - ANEXO: CÓDIGO ÁRBOL DE DECISIÓN 74
9.1.1-ELABORACIÓN PRIMER MODELO DE CLASIFICACIÓN. 74
9.1.2- PODA DEL ÁRBOL DE DECISIÓN 75
9.1.3- BOOSTING 75
9.2.4-MEJORA MODELO, ELIMINAR CARACTERÍSTICAS REDUNDANTES A TRAVÉS DEL ANÁLISIS DE CORRELACIONES 76
9.1.5-MEJORA MODELO, PUNTUAR CARACTERÍSTICAS POR IMPORTANCIA: RANDOM FOREST 76
9.1.6-MEJORAR EL MODELO GENERADO APLICANDO SVD 77
9.1.7- MEJORA DEL MODELO CON EL ANÁLISIS COMPONENTES PRINCIPALES 78
9.2 - ANEXO: PREPARACIÓN DE DATOS PARA TABLEAU 80
9.2.1- CREAR LA TABLA RESPUESTAS 80

10-BIBLIOGRAFÍA 81

Ilustración 1 Jerarquía DIKW (data- information-Knoledge-Wisdom) 11
Ilustración 2-Tareas de la metodología de Kimball, denominada Business. Fuente: Kimball 14
Ilustración 3-Tareas en función de las PECs. Fuente: propia 15
Ilustración 4-Otros entregables. Fuente: propia 15
Ilustración 5-Diagrama de Gantt. Fuente: propia 15
Ilustración 6-Planificación detallada TFM 16
Ilustración 7- Ejemplo de arquitectura de inteligencia de negocio 22
Ilustración 8-Enterprise Bus Arquitecture. Fuente: Josep Curto 23
Ilustración 9-Corporate Information Factory. Fuente: Josep Curto 23
Ilustración 10-Enterprise Data Warehouse 2.0. Fuente: Bill Inmon. 24
Ilustración 11-Ejemplo cubo OLAP 26
Ilustración 12-Cuadrante mágico de febrero 2017 para herramientas de BI 30
Ilustración 13- Evaluación datos entrenamiento 534 casos, hojas 35
Ilustración 14- Evaluación datos entrenamiento 534 casos, reglas 36
Ilustración 15-Capacidad predictiva del modelo 37
Ilustración 16-Evaluación datos de entrenamiento 534 poda 38
Ilustración 17- Capacidad predictiva modelo con poda 38
Ilustración 18- Evaluación datos de entrenamiento 534 casos poda, reglas 39
Ilustración 19- Evaluación de datos de entrenamiento, Boosting 40
Ilustración 20- Capacidad predictiva Boosting 40
Ilustración 21- Gáfica correlaciones todas las variables Vs sin las variables altamente correlacionadas 42
Ilustración 22- Importancia de las variables método randomForest y la evolución del error según el número de árboles 42
Ilustración 23- Diferencias de las variables de influencia en el modelo 43
Ilustración 24-Porcentaje de varianza explicada por el modelo 44
Ilustración 25- Capacidad predictiva del árbol de decisión una vez reducida la dimensionalidad con SVD 44
Ilustración 26- Varianza de cada componente 45
Ilustración 27- Mapa de los dos primeros componentes en función de la variable salud autopercibida 45
Ilustración 28- Mapeo de las dos primeras componentes en función del SAP y los demás predictores 46
Ilustración 29- Árbol de decisión, parte a 47
Ilustración 30- Árbol de decisión, parte b 48
Ilustración 31- Tabla características sociodemográficas ESCA 2016 53
Ilustración 32- Tabla Estilo de vida ESCA 2016 54
Ilustración 33- Tabla Cobertura sanitaria ESCA 2016 54
Ilustración 34- Tabla Salud mental ESCA 2016 54
Ilustración 35- Tabla soporte social 55
Ilustración 36- Tabla limitaciones, discapacidades y autonomía ESCA 2016 55
Ilustración 37- Tabla prácticas preventivas ESCA 2016 55
Ilustración 38- Visitas Médicas, atención primaria, atención de urgencias 56
Ilustración 39- Tabla bienestar psicológico ESCA 2016 57
Ilustración 40- Tabla consumo de medicamentos ESCA 2016 57
Ilustración 41- Tabla variables generales ESCA 2016 58
Ilustración 42- Tabla morbilidad crónica y accidentes ESCA 2016 58
Ilustración 43- Tabla de hecho de las encuestas de Salud de Catalunya 59
Ilustración 44- Tabla de dimensiones de la Encuesta de salud de Catalunya 60
Ilustración 45- Variables comunes en todas las dimensiones 60
Ilustración 46- Tabla.etiquetas.csv 62
Ilustración 47. Tablas. numéricas.csv 63
Ilustración 48- tabla variables.generales.csv 63
Ilustración 49-Tabla preguntas.csv 63
Ilustración 50-Tabla Respuestas.csv 64
Ilustración 51- Conexión y relación de las tablas 65
Ilustración 52- Visualización gráfico de barras para Dashboard ESCA 66
Ilustración 53- Visualización gráfica sexo y edad para el dashboard ESCA 67
Ilustración 54- Visualización de los estudios para el dashboard de la ESCA 67
Ilustración 55- Cuadro de mandos interactivo ESCA 69

Ilustración 56- Perfil Tableau _____ *70*

1-INTRODUCCIÓN

1.1-CONTEXTO Y JUSTIFICACIÓN DEL TRABAJO

«La Era de la Información» es nuestra era:
Es un período histórico caracterizado por una revolución tecnológica centrada en las tecnologías digitales de información y comunicación, concomitante, pero no causante, con la emergencia de una estructura social en red, en todos los ámbitos de la actividad humana, y con la interdependencia global de dicha actividad. (Castells, 2005)

Una **sociedad de la información** es aquella en la cual las tecnologías que facilitan la creación, distribución y manipulación de la información juegan un papel esencial en las actividades sociales, culturales y económicas.

Cada día recibimos más información de la que necesitamos y, por tanto, mucha más de la que realmente nos pueda aportar valor. Por ello debemos ser cada vez más selectivos a la hora de satisfacer nuestra necesidad de estar informados.

En este contexto se ha propiciado la necesidad de tener mejores, más rápidos y más eficientes métodos para extraer y transformar los datos de una organización en información.

Para que la información sea útil para la toma de decisiones, son necesarias herramientas para distribuir el conocimiento generado, en el momento adecuado y a los usuarios que puedan sacar provecho de él. La inteligencia de negocio responde a dichas necesidades.

Como dice (Curto Díaz and Conesa Caralt, 2010), se entiende por **Inteligencia de negocio** al conjunto de metodologías, aplicaciones, prácticas y capacidades enfocadas a la creación y administración de información que permite tomar mejores decisiones a los usuarios de una organización.

En esta nueva Era de la información, la nueva fuente de ventaja competitiva se basa en la adecuada gestión de información, el conocimiento y el momento oportuno, la estrategia más adecuada reside en la capacidad de gestionar, compartir y cooperar, lo que requiere apertura, transparencia y coordinación, agilidad y flexibilidad para permitir una “producción distribuida”. Por todo ello, se apuesta por el modelo de innovación abierta.

La competitividad actual se basa en el conocimiento y en el paradigma de lo abierto.

Esta nueva fuente de ventaja competitiva no discrimina a los organismos sin ánimo de lucro, gubernamentales o agentes públicos, los crecientes desafíos que plantea el proceso de globalización económica están propiciando, en gran medida por el creciente interés de los gobiernos y los agentes públicos respecto al concepto y aplicaciones de la inteligencia territorial como herramientas para fomentar la confianza, el desarrollo económico y social de las regiones, así como la transparencia y buen gobierno.

Como se puede ver en la lectura, "¿Hacia una sociedad del "buen conocer"?" (Miedes-Ugarte, López and Moreno, 2014), la **inteligencia territorial** persigue conocer, comprender y observar las estructuras y dinámicas de un territorio para llevar a cabo las acciones necesarias que faciliten las sinergias entre todos los actores involucrados en el desarrollo sostenible del mismo.

Algunas normativas legales nacionales y autonómicas, establecen que la Administración tiene que dar cuenta a la ciudadanía de su actividad y de la gestión de los recursos públicos que se han puesto a su alcance y por ello debe ser transparente, se debe regir por el principio de buen gobierno y ha de ser accesible.

Los gobiernos se han comprometido a facilitar datos en abierto, con el objetivo de hacerlas transparentes a los ciudadanos, fomentar el uso y la utilización de la información generada por la administración. Con ello se persigue la creación de valor con la reutilización de la información pública, facilitar la ordenación interna de los sistemas de información dentro de la administración y fomentar la interoperabilidad entre servicios del sector público.

Los datos abiertos en los que se basa el trabajo son los generados en la página web encuesta de salud de Catalunya (ESCA) de (Departament de Salut and Generalitat de Catalunya, no date), así como los Open Data de las encuestas de salud de la Comunidad de Madrid (Portal de datos abiertos del Ayuntamiento de Madrid, no date) Estos *Open Data* recogen información de la población residente en las diferentes comunidades indicadas sobre el estado de salud, los comportamientos relacionados con la salud y el uso de los servicios sanitarios. También se dispone de un fichero ZIP, cedido por la ESCA el día 18 de diciembre de 2017 con un archivo *sav* en su interior.

Estos datos abiertos y cedidos proporcionan demasiada información al usuario que puede devenir en una sobrecarga informativa, intoxicación o infobesidad.

Como dice (Cornella, 2004), la **infoxicación** es la sobrecarga de información o exceso de información provocada por la profusión de contenidos en Internet. La causa de esta es el volumen inmanejable de datos, informaciones, contenidos y conocimientos que podemos recibir y obtener de Internet. Los usuarios tienen la dificultad de encontrar la información que les es necesaria entre tanto volumen de datos.

Por tanto, para que los datos aporten valor, estos han de ser analizados, interpretados y explicados. Por ello han de pasar por la jerarquía DIKW(data-information-knowledge-wisdom).

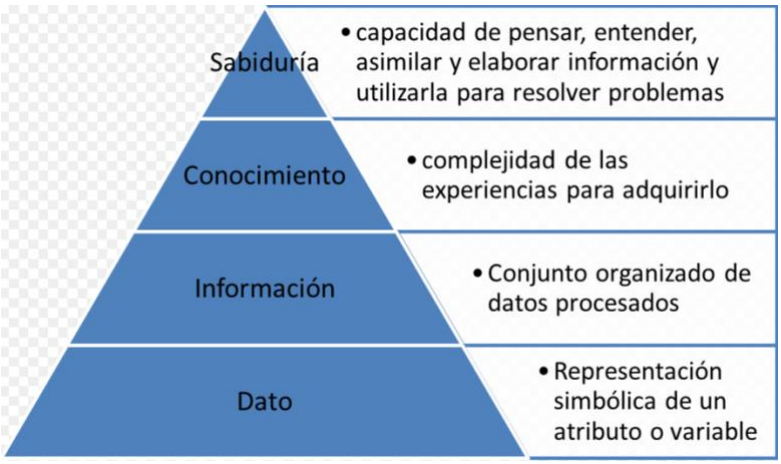


Ilustración 1 Jerarquía DIKW (data- information-Knoledge-Wisdom)

El objetivo de este trabajo es ofrecer un servicio para que esos datos abiertos aporten valor a los usuarios.

Se analizarán los datos open data de la Comunidad de Madrid y se creará un sistema de *reporting* para el análisis de las Encuestas de Salud de Catalunya.

¿Por qué un sistema de *reporting*?

- Para prevenir la potencial pérdida de conocimiento resultante del gran volumen de información Open Data publicado en la web ESCA.

- Permitir a los diferentes tipos de usuarios obtener gran cantidad de información relevante para analizar y establecer relaciones, comprender tendencias y soportar decisiones sobre el estado de salud, los comportamientos relacionados con la salud y el uso de los servicios sanitarios y todo ello orientado a la salud autopercibida de las personas.

-Para relacionar los datos de la ESCA y convertirlos en información valiosa.

- Nos permitirá analizar y entender los datos de la ESCA mediante un análisis detallado.

- Tendremos elementos visuales y gráficos para cada situación.

-Para analizar la evolución y variación de las variables sociodemográficas, como la estructura familiar, situación laboral, nivel de estudios, incertidumbre laboral condiciones de la vivienda, estado de salud, practicas preventivas, estilos de vida, utilización de servicios sanitarios y todo ello de forma dinámica.

- Permitir el análisis desde diferentes orientaciones y niveles la información.

- Automatizar la elaboración y distribución de informes para distintos usuarios.

- Mejorar el acceso a los datos siendo este fácil y rápido, permitiendo a los usuarios, de forma autónoma, hacer sus propios informes.

- A través de este tipo de información se puede lograr una visión más completa e integral de la sociedad catalana en términos del modelo social de la salud de (Dahlgren and Whitehead, 1991), que explica las causas de la desigualdad, permite entender la relación entre individuo, su entorno y la enfermedad, además de aportar valor a múltiples usuarios pudiendo redefinir sus objetivos estratégicos.

-Permitir identificar fácilmente problemas y oportunidades. Permitiendo dar respuesta a preguntas del tipo:

¿ La población que pertenece a una clase social más des favorecedora o que tiene un nivel de estudios más bajo tiene una mala salud autopercibida ?¿Cuáles son las características o atributos de las personas que tienen una buena salud autopercibida?¿Las personas que tienen una mala salud autopercibida son las más hospitalizadas? ¿….

1.2-Objetivos del trabajo

El siguiente proyecto tiene como objetivo final generar una página web dinámica que ofrezca una solución para mostrar elementos emergentes con el contenido de una base de datos. Utilizará varios tipos de recursos gráficos para representar la información necesaria contenida en la base de datos.

La aplicación consta de una única página web dinámica que muestra una serie de menús variables en función del usuario activo. Su objetivo es presentar en una pantalla el máximo de información posible para analizar la información de la población residente en Cataluña sobre el estado de salud, los comportamientos relacionados con la salud y el uso de los servicios sanitarios.

A través de las opciones disponibles en dichos menús, se despliegan una serie de cuadros de texto con los datos de entrada necesarios en cada caso. Y a partir de estos, se muestren los gráficos con el resultado de la acción seleccionada.

1.3- Enfoque y método a seguir

El trabajo sigue las metodologías de **proyectos Agíl**, en concreto el *método de desarrollo de sistemas dinámicos*, el trabajo tendrá como principios, involucrar a los consultores para tomar decisiones con precisión, entrega frecuente de productos para ser revisada y verificada por los consultores, el desarrollo del trabajo es iterativo e incremental, entregas a tiempo, evaluación de riesgos.

Esta última fase es la que da origen a trabajar con dos dataset ya que el riesgo de obtener los datos cedidos de la ESCA a tiempo era muy improbable, de hecho, estos no se nos cedieron hasta el 18 de diciembre. Por ello la primera fase del trabajo, en la que se realiza un análisis de datos, se realiza con los Open Data de la Comunidad de Madrid.

Para las visualizaciones de la herramienta de *reporting* interactiva se han tenido en cuenta las buenas prácticas y recomendaciones de (Few, 2006).

El trabajo se ha realizado en base a la aproximación al desarrollo de almacenes de datos abogado por (Kimball and Ross, 2013). La metodología se basa en lo que Kimball denomina “Ciclo de Vida Dimensional del Negocio”.

De la siguiente figura, Ilustración 2, se pueden extraer dos características clave. En primer lugar, el rol central de la tarea de definición de requerimientos, que se basa en la identificación las necesidades que tiene la organización particular con respecto al análisis de la información.

En segundo lugar, se aprecia tres rutas o caminos que se enfocan en tres diferentes áreas:

1.- Tecnología (Camino Superior). Implica tareas relacionadas con software específico.

2.- Datos (Camino del medio). En la misma diseñaremos e implementaremos el modelo dimensional, y desarrollaremos el subsistema de Extracción, Transformación y Carga (Extract, Transformation, and Load - ETL) para cargar el DW.

3.- Aplicaciones de Inteligencia de Negocios (Camino Inferior). En esta ruta se encuentran tareas en las que diseñaremos y desarrollaremos las aplicaciones de negocios para los usuarios finales.

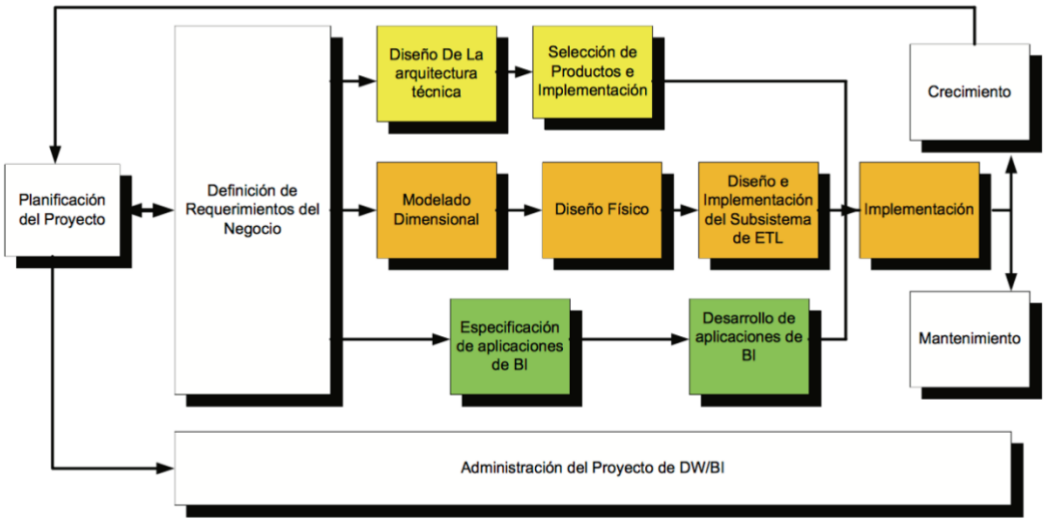


Ilustración 2-Tareas de la metodología de Kimball, denominada Business. Fuente: Kimball

1.4- Planificación del trabajo

Este proyecto está compuesto por pequeñas partes que tienen que completarse y entregarse en pocas semanas. De esta manera si hay que realizar cualquier modificación del proyecto, a petición de los consultores, sólo cambiara la parte implicada y en muy poco tiempo.

Calendario laboral y stakeholders

El proyecto tiene una fecha de inicio prevista el 18 de octubre de 2017 y la entrega final el día 23 de enero de 2018 con un solo recurso. Entre dicho período hay una serie de pequeñas entregas siguiendo la metodología Agíl, las cuáles se dividen en función del cronograma PECs del TFM y otros entregables que forman parte de tareas críticas del TFM.

Los stakeholders o interesados en el proyecto son: los consultores del TFM, la Dra. Laia Subirats y el Dr. Jordi Conesa Caralt, que desempeñarán el rol comercial.

Cronograma PECs del TFM

Las fases en las que se ha dividido el TFM corresponden a las PECs planteadas en el Plan Docente de la asignatura y se resumen en la figura 3. Existen otros entregables que son factores críticos del proyecto, que se detallan en la figura 4. Estos se entregarán a los consultores para su análisis y aprobación, incorporándose al proyecto TFM una vez aprobados.

TFM-Construcción y explotación de un almacén de datos para el análisis de las encuestas de salud de Cataluña (ESCA).				
Fecha	**Dedicación**	**Actividad**	**Inicio**	**Entrega**
18 oct. 2017	2 (horas)	Selección del trabajo	18 oct. 2017	22 oct. 2017
23 oct. 2017	20 (horas)	PEC 1 Propuesta inicial	23 oct. 2017	6 nov. 2017
7 nov. 2017	42 (horas)	PEC 2 Estructura de la memoria	7 nov. 2017	4 dic. 2017
5 dic. 2017	50 (horas)	PEC 3 Primera versión de la memoria	5 dic. 2017	8 ene. 2018
9 ene. 2018	**22 (horas)**	**Entrega final**	**9 ene. 2018**	**23 ene. 2018**

Ilustración 3-Tareas en función de las PECs. Fuente: propia

TFM-Otros entregables.				
Fecha	**Dedicación**	**Actividad**	**Inicio**	**Entrega**
5 nov. 2017	22 (horas)	Recogida, preparación y análisis de los datos	5 nov. 2017	20 nov. 2017
11 dic. 2017	10 (horas)	Diseño funcional del prototipo	dic. 2017	15 dic. 2017

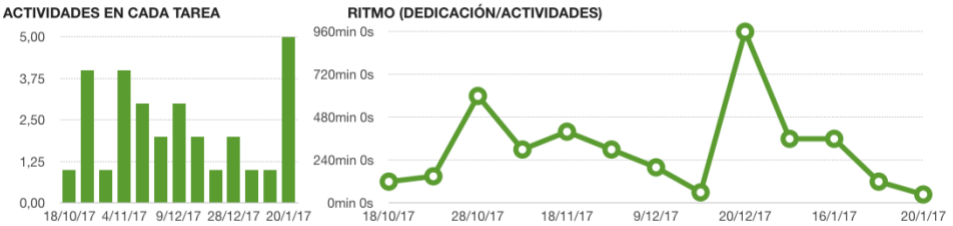
Ilustración 4-Otros entregables. Fuente: propia

Nº	TAREAS	MES DEDICACIÓN (Minutos)		octubre	noviembre	diciembre	enero	OBSERVACIONES
0	**SELECCIÓN DEL TRABAJO**		120					
	Selección del trabajo							
1	**INTRODUCCIÓN**		600					
	1-Contexto y justificación del trabajo 2-Objetivos del trabajo 3-Enfoque y método a seguir 4-Planificación del trabajo 5-Breve resumen de los productos obtenidos 6- Breve descripción de los otros capítulos de la memoria							
2	**ESTADO DE ARTE**		2280					
	1-Inteligencia de negocio 2-Visualización de datos 3-Cuadro de mandos 4-Herramienta de análisis visual. Tableau							
3	**DISEÑO, CONSTRUCCIÓN Y EXPLOTACIÓN DEL ALMACÉN DE DATOS**		4080					
	1-Análisis 2-Diseño 3-Carga de datos 4-Informes 5- Elaboración y publicación en internet del cuadro de mando o dashboard en Tableau							
4	**RESULTADOS**		80					
	Prototipo							
5	**CONCLUSIONES**		80					
6	**GLOSARIO**		80					
7	**BIBLIOGRAFIA**		80					

Ilustración 5-Diagrama de Gantt. Fuente: propia

Construcción y explotación de un almacén de datos para el análisis de las encuestas de salud de Cataluña (ESCA).

18 de octubre del 2017 - 23 de enero de 2018



ACTIVIDAD MEDIA (TAREAS)	DEDICACIÓN DE EJECUCIÓN MEDIO	RITMO MEDIO/ACTIVIDAD	DEDICACIÓN TOTAL
2,31	**572min 18s**	**306min 0s**	**7440min 0s**
ACTIVIDAD TOTAL	ACTIVIDAD MÁX. (TAREAS)	RITMO MÁX./ACTIVIDAD	
30,00	**5,00**	**48min 0s**	

FECHA INICIO	FECHA FIN	DEDICACIÓN	ACTIVIDADES EN CADA TAREA	RITMO (DEDICACIÓN/ ACTIVIDADES)	TAREAS	CONTENIDO
18/10/17	20/10/17	120min 0s	1,00	120min 0s	TAREA 1- SELECCIÓN TRABAJO	SELECCIÓN TRABAJO
21/10/17	27/10/17	600min 0s	4,00	150min 0s	TAREA 2. INTRODUCCIÓN	1- Contexto y justificación del trabajo, objetivos del trabajo, enfoque y método a seguir, planificación del trabajo
28/10/17	3/11/17	600min 0s	1,00	600min 0s	TAREA 3. ESTADO DE ARTE	2- INTELIGENCIA DE NEGOCIO
4/11/17	17/11/17	1200min 0s	4,00	300min 0s	TAREA 4- DISEÑO, CONSTRUCCIÓN Y EXPLOTACIÓN DEL ALMACÉN DE DATOS	3-EVALUACIÓN DE LA ESTRUCTURA Y ANÁLISIS, Análisis de requerimientos, análisis de fuentes de datos, análisis funcional
18/11/17	1/12/17	1200min 0s	3,00	400min 0s	TAREA 5- DISEÑO, CONSTRUCCIÓN Y EXPLOTACIÓN DEL ALMACÉN DE DATOS	3- DISEÑO, Diseño conceptual, Diseño lógico, Diseño físico
2/12/17	8/12/17	600min 0s	2,00	300min 0s	TAREA 6- DISEÑO, CONSTRUCCIÓN Y EXPLOTACIÓN DEL ALMACÉN DE DATOS	3-CARGA DE DATOS, Identificación de los procesos ETL necesarios, descripción de las acciones en cada proceso ETL
9/12/17	15/12/17	600min 0s	3,00	200min 0s	TAREA 7- ESTADO DE ARTE	2- VISUALIZACIÓN DE DATOS, definición, principios clave, principio de diseño
16/12/17	19/12/17	120min 0s	2,00	60min 0s	TAREA 8- ESTADO DE ARTE	2-CUADROS DE MANDO, definición e importancia
20/12/17	27/12/17	960min 0s	1,00	960min 0s	TAREA 9- ESTADO DE ARTE	2- HERRAMIENTA DE ANÁLISI VISUAL. TABLEAU. Breve introducción a tableau
28/12/17	15/1/17	720min 0s	2,00	360min 0s	TAREA 10- DISEÑO, CONSTRUCCIÓN Y EXPLOTACIÓN DEL ALMACÉN DE DATOS	3- INFORMES, Elaboración de informes requeridos, creación de un dashboard
16/1/17	18/1/17	360min 0s	1,00	360min 0s	TAREA 11- DISEÑO, CONSTRUCCIÓN Y EXPLOTACIÓN DEL ALMACÉN DE DATOS	3- PUBLICACIÓN EN INTERNET DEL CUADRO DE MANDO O DASHBOARD EN TABLEAU
19/1/17	19/1/17	120min 0s	1,00	120min 0s	TAREA 12- RESULTADOS	4-PROTOTIPO
20/1/17	23/1/17	240min 0s	5,00	48min 0s	TAREA 13- CONCLUSIONES, GLOSARIO, BIBLIOGRAFIA, INTRODUCCIÓN	Breve resumen productos obtenidos, breve descripción otros capítulos memoria,5-CONCLUSIONES, 6, GLOSARIO, 7-BIBLIOGRAFIA

Ilustración 6-Planificación detallada TFM

La figura 6 muestra la planificación detallada de las actividades de cada una de las tareas en que se ha dividido el proyecto. Además, cuenta con el desglose de actividades totales, actividad media por tarea, dedicación de ejecución media por tarea, el número máximo de actividades por tarea, el ritmo máximo de dedicación por actividad, y el ritmo máximo por actividad y la dedicación total del TFM. Estas pueden estar sujetas a modificaciones en caso de que se produzcan cambios y se lleven a cabo gestiones de cambio.

1.5-Breve resumen de los productos obtenidos

"Es cierto que el mundo es lo que vemos y, sin embargo, tenemos que aprender a verlo." (Merleau-Ponty)

Se pretende hacer un recorrido por el mundo de la inteligencia de negocio a través de diferentes sistemas, con el fin último, de crear un prototipo de sistema de *reporting* interactivo para la ESCA que nos permita medir, comparar, reportar, analizar, predecir, avisar, colaborar, y saber sobre la salud autopercibida de las personas de Catalunya.

Para crear este sistema, se parte de diferentes fases:

-Análisis de datos a través de algoritmos estructuralmente sencillos como los árboles de decisión para conocer bien los datos, ya que antes de crear un sistema de *reporting* debemos saber qué es lo que cuentan los datos para poder explicar historias y construir entendimiento. Esta fase se analizará en detalle en el punto 4, donde se realiza la extracción de características y clasificación a través de los árboles de decisión. (Se debe indicar que este análisis se realizó con otra base de datos diferente a ESCA con similar estructura e información (la Encuesta de Salud de la Comunidad de Madrid), ya que hasta el 18 de diciembre no se dispuso de los datos. Para ello utilizaremos el entorno R.

-Una vez realizado el análisis de datos, toca construir un **sistema de almacenamiento y gestión de datos,** (*datawarehouse*). Para ello utilizaremos las herramientas R y EXCEL. Este sistema de almacenamiento nos permitirá incorporar los datos a la herramienta de *reporting interactiva* una vez realizado el proceso de reducción de la dimensionalidad. El resultado de esta fase es un archivo EXCEL en el que constan 5 pestañas para poder conectar los datos con *Tableau*.

-Y, por último, crearemos nuestro ***sistema de reporting*** a partir de la herramienta *TABLEAU*, que nos permita contar historias y construir entendimiento con los datos a partir de las visualizaciones.

1.6-Breve descripción de los capítulos de la memoria

En el **Capítulo 2** se muestra el conocimiento indispensable para poder entender el contenido del trabajo. En primer lugar, para que el lector tenga una referencia más concreta sobre el caso de desarrollo, se explica y contextualiza la salud autopercibida a través de diferentes estudios realizados por expertos.
La segunda parte de este capítulo se adentra en la inteligencia de negocio, sus conceptos y elementos.
La tercera parte se centra en las visualizaciones, que son, su evolución con los años, los beneficios que reporta tener un sistema de visualización y por último los principios de diseño.
La cuarta parte se centra en la idea de cuadro de mando integral de Norton y Kaplan, su significado, su importancia, los requisitos que ha de tener un cuadro de mando y por último se explican las herramientas disponibles en el mercado mediante el cuadrante Gartner Inc. La última parte del capítulo 2 hace referencia

al caso de estudio ESCA, se define la ESCA, el contenido de las Encuestas de Salud de Catalunya y algunos trabajos realizados o resultados obtenidos.

En el **Capítulo 3** se realiza una evaluación de la infraestructura existente, se indican las características de la máquina con la que se trabaja y se realiza la selección de los productos para poder llevar a cabo el trabajo.

En el **Capítulo 4** se analiza la base de datos de la Encuesta de la Comunidad de Madrid a partir del algoritmo C5.0 de generación de árboles de decisión. En primer lugar, se indica el objetivo de análisis, la fuente de datos y se define el problema de análisis. A continuación, se realiza un análisis visual y de correlaciones (estas dos partes se podrán obtener en el archivo original Rmarkdown) se pasa a construir el modelo de clasificación y se realiza la optimización o mejora a través de la poda y *boosting*. Por último, se aplican las técnicas de extracción de características *Filter* y *Wrapper* para mejorar los modelos de clasificación anteriores.

En el **Capítulo 5** se diseña y construye un almacén de datos para el análisis de las encuestas de la ESCA en TABLEAU. En primer lugar, se identifican los requerimientos funcionales, donde se profundiza en las necesidades de *reporting,* los tipos de usuarios que utilizarán la herramienta y las características que debería reunir el sistema de r*eporting*. El siguiente paso es analizar los datos; en una primera instancia, se analizan los datos open data de la ESCA y la evaluación de la recogida de información, que fue infructuosa debido a que los datos Open Data son secundarios, han sido extraídos de fuentes primarias, manipulados y transformados en tablas. Con esos datos era imposible extraer más información por ello se realiza la solicitud de cesión de los datos a ESCA, fase de elevado riesgo en el proyecto y que produjo un retraso considerable en los demás entregables del TFM. Los datos son cedidos el 18 de diciembre y comenzamos el análisis de los datos del archivo. *sav*. Se realiza un diseño conceptual de la base de datos, en función de la tabla de hechos y las dimensiones. Este capítulo finaliza con el diseño físico de la base de datos.

En el último **Capítulo 6,** se realiza la creación del cuadro de mandos para la ESCA. En primer lugar, se conecta la base de datos a TABLEAU, después se construyen las visualizaciones o vistas del cuadro de mandos y se crea el cuadro de mandos, se guarda el proyecto, (por falta de permisos, y debido a que la información con la que se trabaja es muy sensible, no se realiza la publicación en TABLEAU public).

En el **Capítulo 7** se extraen las conclusiones de cada una de estas fases por separado y se realiza una conclusión en conjunto.

En el **Capítulo 8** se explica, brevemente, la continuación del TFM, cuáles deberían ser los siguientes pasos.

En el **Capítulo 9** figuran los anexos, con el código R de los análisis de datos y la parte de creación de la base de datos.

Por último, en el **Capítulo 10**, se recoge la bibliografía que hemos utilizado. Para realizar esta sección hemos contado con la herramienta Mendeley, por tanto, se puede enviar en archivo.ris para los interesados con la bibliografía utilizada.

2-Estado de arte

2.1- Salud autopercibida

"La salud autopercibida por el individuo forma parte de un grupo de indicadores socio- sanitarios complementarios a la mortalidad en fases avanzadas de la transición epidemiológica en las que el escenario de salud no sólo se define por resultados fatales (mortalidad y sus causas) sino también por la mayor o menor prevalencia de enfermedades asociadas a los estilos de vida en regímenes de baja mortalidad" (Omran, 1998).

(Idler and Benyamini, 1997) en una revisión de veintisiete trabajos que utilizaban este indicador mostraron que la salud autopercibida es una medida importante, ya que tiene influencia sobre la mortalidad y proporciona información muy importante sobre el estado de salud.

Hay varias medidas para la salud autopercibida, quizás la más popular en Europa es EuroQoL que es la que se utiliza en la encuesta de salud de Cataluña.

En España, diversos trabajos, la mayoría basados en la Encuesta Nacional de Salud (ENSE), han utilizado la salud autopercibida como indicador global de salud, así como para la medición de las desigualdades en salud en función de determinantes socioeconómicos (Gonzalo and Pasarín, 2004).

Hay varios análisis longitudinales de la salud autopercibida en España. (Tamayo-Fonseca *et al.*, 2013), realizó un estudio de diseño longitudinal, de seguimiento durante 4 años de la cohorte representativa de adultos de la Encuesta de Salud de la Comunidad Valenciana del año 2005 (ESCV 2005). En su estudio encontró que la autoevaluación de la salud es una medida predictiva de la mortalidad tanto en hombres como en mujeres, y actúa como un mediador entre las condiciones socioeconómicas, demográficas y de salud y la mortalidad. Además (Tamayo-Fonseca *et al.*, 2015) encontró que la autoevaluación de la salud es un predictor de los servicios hospitalarios.

Otro tema importante a tener en cuenta es que los factores asociados con la Salud autopercibida difieren según los grupos de edad. (Giron, 2012) en uno de sus estudios con los datos de Encuestas Nacionales de Salud realizadas en 2001, 2003, 2006 y 2011-12 establece que las tendencias de la buena salud autopercibida difieren según el género, según los factores sociodemográficos y la prevalencia de la discapacidad.

(Morcillo Cebolla *et al.*, 2014) en uno de sus estudios en la revisión de más de 20 documentos concordantes sobre la valoración de las desigualdades en la salud autopercibida de los mayores en relación con la edad, el género, los factores socioeconómicos y el nivel de dependencia, concluye que entre las personas mayores, las desigualdades en la salud autopercibida debido al estatus socioeconómico y al género han persistido en el tiempo.

Por lo tanto, es importante que los gobiernos desarrollen políticas de salud en este sector de la población.

Además, también hay estudios de salud autoinformada en España sobre discapacidades neurológicas. (Subirats *et al.*, 2015) en un estudio en el que se utilizaron datos de 730 personas con lesión cerebral adquirida realizando evaluaciones exhaustivas periódicas en los años 2006-2013, se concluyó que un sistema de monitoreo y pronóstico tiene el potencial de: (1) ahorrar costos y tiempo, (2) proporcionar más información para tomar decisiones, (3) promover la interoperabilidad, (4) facilitar la toma conjunta de decisiones y (5) mejorar las políticas de evaluación socioeconómica de la carga de la enfermedad.

Existen otros estudios de origen cardíaco como el de (Calvo *et al.*, 2013), en este estudio el objetivo es proporcionar un método para evaluar automáticamente el impacto en España de la rehabilitación cardíaca. En su estudio se concluye que los indicadores de los resultados de salud pueden proporcionar información sobre la calidad de vida obtenida después de una intervención, como la rehabilitación cardíaca.

En estos dos casos se usa el estándar de la Organización Mundial de la Salud (OMS) de la Clasificación Internacional de Funcionamiento, Discapacidad y Salud.

En Europa, también se han realizado varios estudios sobre la autoevaluación de la salud. (Molarius *et al.*, 2007) estableció que las condiciones psicosociales, las condiciones materiales y los factores del estilo de vida están relacionados de forma independiente con la mala autoevaluación de la salud.

(Abu-Omar, Rütten and Robine, 2004) encontró que también existe una relación positiva entre la actividad física y la salud autopercibida.

También hay estudios como el de (Bardage *et al.*, 2005) que incluyen la autoevaluación de la salud autopercibida comparada internacionalmente. Este concluyó que existen diferencias en la autoevaluación de la salud entre los países, pero que no se pueden explicar con diferencias en el estado, el estado de salud o la capacidad funcional.

2.2-Inteligencia de negocio

2.2.1 Conceptos y elementos

Business Intelligence (Inteligencia de Negocio), también conocida por sus siglas BI, se puede definir como:

"La combinación de tecnología, herramientas y procesos que me permiten transformar los datos almacenados en información, esta información en conocimiento y este conocimiento dirigido a un plan o una estrategia comercial. La inteligencia de negocios debe ser parte de la estrategia empresarial, esta le permite optimizar la utilización de recursos, monitorear el cumplimiento de los objetivos de la empresa y la capacidad de tomar buenas decisiones para así obtener mejores resultados" -Data Warehouse Institute.

Los principales elementos que componen una arquitectura de BI "podrían ser":

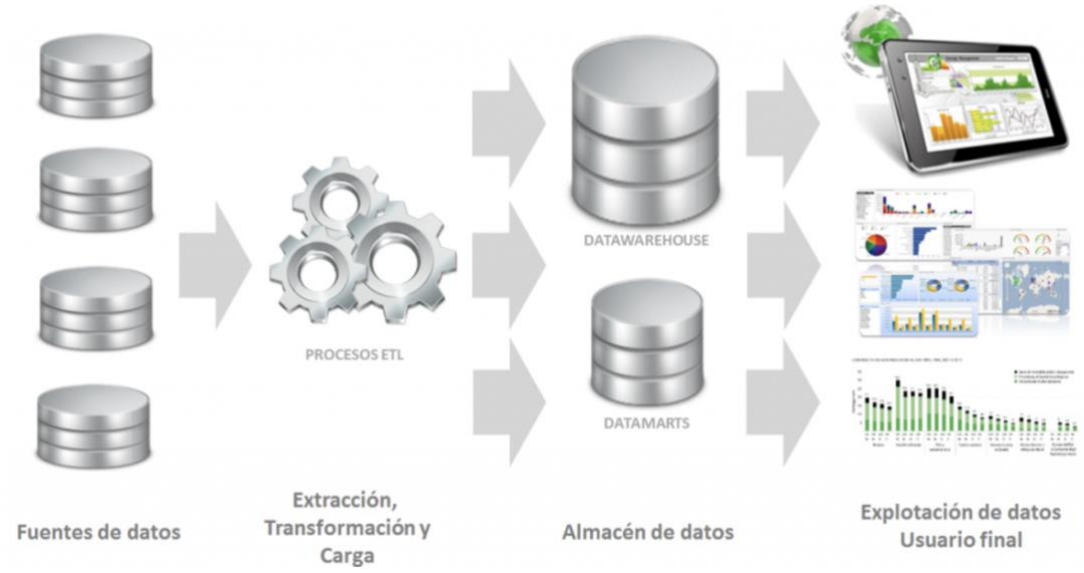


Ilustración 7- Ejemplo de arquitectura de inteligencia de negocio

2.2.1.1-Data Warehouse (DW)

El DW es un repositorio de datos global de la organización que es independiente de su utilización posterior, y ha de reunir las propiedades de estabilidad, coherencia, fiabilidad y con información histórica.

"Un DW es un conjunto de datos orientados por temas, integrados, variantes en el tiempo y no volátiles, que tienen por objetivo dar soporte a la toma de decisiones"-W.H. Inmon

"Un DW es una copia de los datos transaccionales específicamente estructurada para la consulta y el análisis" - Kimball.

Los elementos de un DW:

- ***Hechos:*** Permiten representar los procesos de negocio de la organización.

- ***Dimensiones:*** Permiten representar las distintas vistas para un cierto proceso de negocio.

- ***Métricas:*** Son los indicadores de negocio de un proceso de negocio, son aquellos conceptos cuantificables que permiten medir un proceso de negocio. Están asociados a las tablas de hecho.

Para estructurar los datos en un almacén de datos existen principalmente dos tipos de esquemas:

-***Esquema en estrella:*** Consiste en una tabla de hecho en el centro para el hecho objeto de estudio y una o varias tablas de dimensión por cada punto de vista de análisis que permita la descripción de ese hecho. En la

tabla de hecho encontramos los atributos destinados a mediar. La tabla de hechos sólo presenta uniones con dimensiones.

-***Esquema en copo de nieve:*** Es un esquema de representación derivado del esquema en estrella, en el que las tablas de dimensiones se normalizan en múltiples tablas. Así, la tabla de hechos deja de ser la única tabla del esquema que se relaciona con otras tablas, y aparecen nuevas uniones.

Existen tres enfoques en la arquitectura corporativa de un DW.

- ***Enterprise Bus Arquitecture (MB):*** Arquitectura basada en *data marts* independientes federados que pueden hacer uso de una *staging* área en caso de ser necesario.

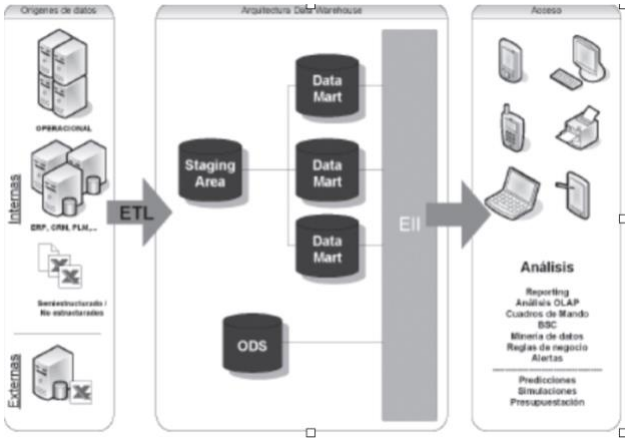


Ilustración 8-Enterprise Bus Arquitecture. Fuente: Josep Curto

- *Corporate Information Factory*: Consiste en una arquitectura en la que existe una data *warehouse* corporativo y distintas *data mart* dependientes del mismo. Nunca se consulta una data warehouse directamente.

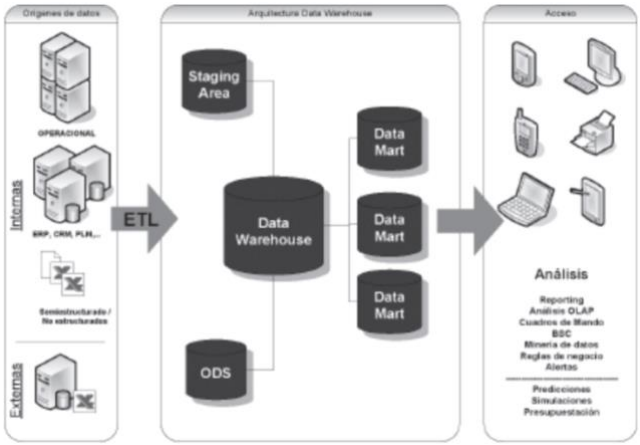


Ilustración 9-Corporate Information Factory. Fuente: Josep Curto

-***Enterprise Data Warehouse 2.0:*** Se caracteriza por incluir tanto información estructurada como no estructurada y por focalizarse en responder a todas las necesidades actuales de negocio.

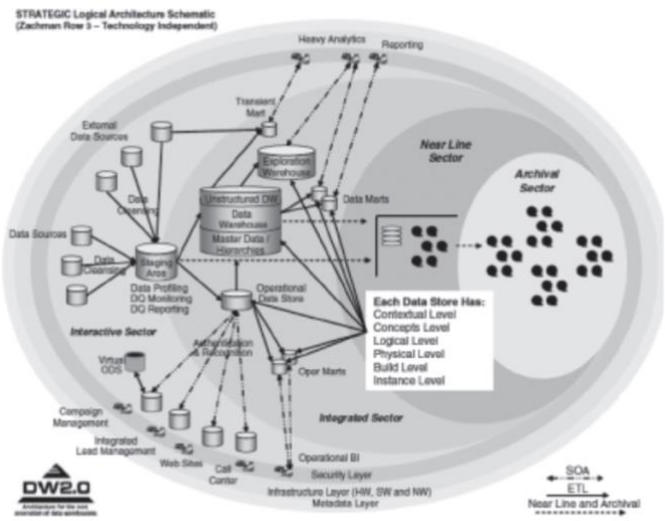


Ilustración 10-Enterprise Data Warehouse 2.0. Fuente: Bill Inmon.

2.2.1.2-ETL (extract transform load)

Los procesos *ETL* son las tecnologías de integración de datos en el *Data Warehouse* basados en la definición de fuentes de datos, procesos de extracción, transformación y adecuación a la estructura del repositorio de datos destino.

Las técnicas de integración de datos son:

-**Técnica *push***: consiste en la actualización continua y en línea del entorno destino mediante aplicaciones de integración.

-**Técnica *pull***: consiste en la provisión de datos mediante procesos *batch* en intervalos prefijados.

- **Propagación de datos**: consiste en copiar datos de un lugar de origen a un entorno destino local o remoto. Los datos pueden extraerse del origen mediante programas que generen un fichero que debe ser transportado al destino, donde se utilizará como fichero de entrada para cargar en la base de datos destino.

-**CDC (*Change Data Capture*)**: se utilizan para capturar los cambios producidos por las aplicaciones operacionales en las bases de datos origen, de tal manera que pueden ser almacenados y/o propagados a los entornos destino para que estos mantengan la consistencia con los entornos origen.

-**Consolidación de datos**: consiste en capturar cambios realizados en múltiples entornos origen y propagarlos a un único entorno destino, se almacena una copia de todos estos datos.

-**Federación de datos:** proporciona a las aplicaciones una visión lógica virtual común de una o más bases de datos. Permite acceder a diferentes

entornos origen de datos, que pueden estar en diferentes gestores de bases de datos y máquinas, y crear una visión integrada de los datos como se encontrarán en la práctica en una base de datos única e integrada.

-**Técnicas híbridas:** incluye una combinación de técnicas de integración anteriores.

Existen diferentes tecnologías de integración de datos basadas en las anteriores técnicas de integración de datos:

ETL: permite extraer datos del entorno origen, transformarlos según nuestras necesidades de negocio y cargar estos datos en entornos de destino.
Estas tecnologías pueden ser de diferentes tipos:

-*ETL de generación de código:* Constan de un entorno gráfico donde se diseñan y especifican los datos de origen, sus transformaciones y los entornos destino.

-*ETL basados en motor:* permiten crear flujos de trabajo en tiempo de ejecución definidos mediante herramientas gráficas. El entorno gráfico especifica el *mapping* de los entornos de datos de origen y destino, las transformaciones de datos necesarias, el flujo de procesos y los procesos por lotes necesarios.

-*ETL integrado en base de datos:* se incluyen capacidades ETL dentro del motor de bases de datos.

-**EII:** permite a las aplicaciones de acceso a datos dispersos de forma federada.

-**EDR:** detecta los cambios que suceden las fuentes de origen. Está soportada por técnicas de integración de datos CDC y de propagación de datos.

2.2.1.3-OLAP (Procesamiento analítico en línea)

OLAP es una solución cuyo objetivo es agilizar la consulta de grandes cantidades de datos. Para ello utiliza estructuras de datos diversas, normalmente multidimensionales o cubos OLAP, que permiten analizar bases de datos relacionales de gran volumen y variedad con una gran agilidad y rapidez, reduciendo enormemente el tiempo y los recursos empleados en el análisis.

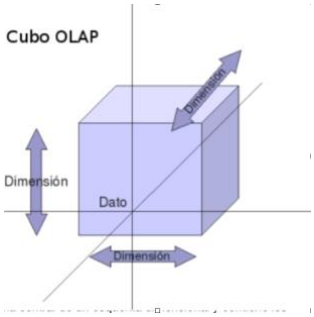


Ilustración 11-Ejemplo cubo OLAP

Los elementos del sistema analítico OLAP son:

-**Esquema**: es una colección de cubos, dimensiones, tablas de hecho y roles.

-**Cubo**: es una colección de dimensiones asociadas a una tabla de hecho.

-**Tabla de hecho**: es la tabla central de un esquema dimensional y contiene los valores de las medidas de negocio o dicho de otra forma los indicadores de negocio.

-**Tabla de dimensiones**: las tablas de dimensiones son elementos que contienen atributos (o campos) que se utilizan para restringir y agrupar los datos almacenados en una tabla de hechos cuando se realizan consultas sobre dichos datos en un entorno de almacén de datos o data mart.

-**Métrica**: son los indicadores de negocio de un proceso de negocio. Aquellos conceptos cuantificables que permiten medir nuestro proceso de negocio. Por ejemplo, en una venta tenemos el importe de la misma.

-**Jerarquía**: es un conjunto de miembros organizados en niveles.

-**Nivel**: es un grupo de miembros en una jerarquía que tienen los mismos atributos y nivel de profundidad en la jerarquía.

-**Miembro**: es un punto en la dimensión de un cubo que pertenece a un determinado nivel de jerarquía.

-**Roles**: permisos asociados a un grupo de usuarios.

-**MDX**: es el lenguaje de consulta de estructuras OLAP.

Existen diferentes tipos de OLAP, en función de la forma en que se almacenan los datos:

-**MOLAP** (Multidimensional OLAP): utiliza bases de datos multidimensionales. En definitiva, se crea un fichero que contiene todas las posibles consultas pre-calculadas.

-**ROLAP** (*Relational* OLAP): Utiliza el modelo relacional, almacenando las dimensiones como tablas relacionales y utilizando tablas adicionales para guardar la información agregada.

-**HOLAP** (*Hybrid* OLAP): Es una base de datos en la que los datos se dividen en almacenaje relacionan y multidimensional. Utilizan las tablas relacionales para guardar las cantidades más grandes de datos detallados y las multidimensionales para los datos menos detallados o agregados.

-**DOLAP** (*Desktop* OLAP): Consiste en obtener la información necesaria desde la base de datos relacional y guardarla en el escritorio. Las consultas y análisis son realizadas contra los datos guardados en el escritorio.

-***In-memory* OLAP**: Consiste en una estructura dimensional que se genera sólo a nivel de memoria y se guarda el dato original en algún formato que potencia su despliegue de esta forma (comprimido o a través de una base de datos lógica asociativa).

2.3-Visualización de datos

2.3.1 - ¿Qué es visualización de la información?

"Visualización es aquella disciplina que consiste en transformar datos en información semántica, o en crear herramientas para que un usuario complete por sí mismo dicho proceso. Por medio de una sintaxis de fronteras imprecisas y en constante evolución, basada en la conjunción de signos de naturaleza icónica (figurativos) con otros de naturaleza arbitraria y abstracta (no-figurativos). La visualización de datos y la infografía son formas, no solo de mostrar unos datos, sino de explorarlos, analizarlos y contextualizarlos. De razonar sobre ellos", (Cairo, 2011)

La **visualización de datos** es un primer paso para el análisis y proyección de los datos disponibles, utilizando los recursos del sistema visual humano como procesador para detectar patrones, tendencias y/o anomalías.
En este sentido, una visualización permite, de forma eficiente, medir y comparar datos, entre otras operaciones. Incluso, un simple resumen de un conjunto de datos puede ser mejor transmitido y comprendido mediante una visualización que mediante el uso de texto y tablas numéricas.

La disponibilidad masiva de datos y la aparición de tecnologías que permiten la creación de visualizaciones de datos dinámicas, con la utilización de la web como medio, ha propiciado el *boom* que ha tenido la visualización recientemente.

2.3.2 - Beneficios de la visualización

- La visualización de la información nos ayuda en el reto de filtrar la información relevante y centrarnos en el interés pertinente. **Evita el exceso de información.**

- Interpretamos la información visual de manera más rápida que la escrita. Los humanos somos seres visuales. **Facilita la comprensión**.

- Si creamos visualizaciones atractivas, fáciles de digerir y que aporten conocimiento nuevo, información novedosa o relevante, el éxito de que acabe siendo **compartida** está asegurado. Los seres humanos somos seres sociales.

2.3.3 - Principios de diseño

"El objetivo principal de la visualización de datos es transmitir información clara y eficaz a través de medios gráficos. Esto no significa que la visualización de datos tenga que ser aburrida para ser funcional, o extremadamente sofisticada para ser bella. Para expresar ideas de manera efectiva, tanto la dimensión estética como la funcional deben ir de la mano, y deben proporcionar determinados puntos de entrada a un conjunto de datos complejos, para comunicar sus aspectos clave de una forma más intuitiva", (Friedman, 2008).

"La visualización ideal no solo debe comunicar con claridad, sino que también debe estimular la participación y la atención del espectador", (Viegas and Wattenberg, 2006).

"El trabajo de un diseñador de información es interpretar, aclarar, sintetizar, explicar, transmitir contenidos y ponerlos en contexto para que sean relevantes", (Alcalde, 2015)

Según(Few, 2006), las ideas que deberían regir cualquier visualización de datos son:

- Una visualización debería indicar claramente cómo los valores mostrados se relacionan (comparan) los unos con los otros, o cada uno con la totalidad.

- Representar las cantidades de forma precisa.

- Facilitar la comparación entre cantidades.

- Facilitar el establecimiento de un orden de los elementos según la cantidad que representan.

- Dejar claro cómo debería usarse la visualización y cuáles son sus objetivos y para qué debería usarse.

2.4- Cuadros de mando

2.4.1 - Cuadro de mandos, su importancia y requisitos

"La medición es el primer paso para el control y la mejora. Si no se puede medir, no se puede entender. Si no se entiende, no se puede controlar. Si no se puede

controlar, no se puede mejorar" (Harrington and Tiznado Santana, 1997)

(Kaplan, 2016) propuso el cuadro de mandos en 1990 como una forma de gestionar una organización recogiendo todo tipo de aspectos que supusieran una ventaja competitiva. Además de ayudar a la toma de decisiones, también ayuda a medir el nivel de cumplimiento y desempeño en cuanto a las estrategias y objetivos marcados por parte de las organizaciones.

En los últimos años se han producido dos revoluciones en la gestión de las empresas: la primera, la revolución de la medición con la incorporación generalizada de medidas no financieras a los sistemas de información y control convencionales y, la segunda, la revolución del "ValueReporting", la transparencia absoluta mediante la cual los directivos de las empresas informan al mercado con todos los indicadores internos que utilizan en su gestión.

Los requisitos que ha conseguir alcanzar el cuadro de mandos son:

- El contenido de la información ha de ser comprensible, pertinente y confiable.
- Deberá estar orientado a los usuarios.
- Debería poder extraer, filtrar, consolidar y visualizar los datos más relevantes.
- Ha de permitir su acceso en tiempo real.
- Ha de poder visualizar tendencias, patrones, anomalías…
- Ha de servir de mecanismo de alerta, para atraer la atención del usuario, ante desviaciones importantes.
- Debe actuar como una herramienta de diagnóstico.
- Debe facilitar la toma de decisiones.
- Ha de poseer una interface fácil de usar y útil para el usuario

2.4.2 - Herramientas disponibles en el mercado

Gartner, Inc. es una consultora líder mundial en investigación y asesoramiento de tecnologías de información. Analiza las tendencias del mercado y facilita la selección de plataformas, elabora sus reconocidos cuadrantes mágicos de publicación anual sobre soluciones tecnológicas.

Ilustración 12-Cuadrante mágico de febrero 2017 para herramientas de BI

Las herramientas líderes, según Gartner, están en el cuadrante de líderes (LEADERS). Estas dan respuesta a todas o casi todas las necesidades que se pueden plantear en una empresa, tienen una buena ejecución en el momento actual y están bien posicionadas para el futuro.

Tableau está posicionada como una de las mejores en este sector, y estos dos últimos años se ha situado entre las tres líderes, junto con **Qlik** y con los servicios ofrecidos por Microsoft Cloud. Tableau ha sido reconocido por centrarse en la experiencia del cliente, con mejoras en los puntajes de los beneficios del negocio y sigue siendo percibido como el líder moderno en el mercado de BI.

Tableau es el fruto de la investigación de Jock Mackinlay, quien sentó las bases de la teoría que permite a una máquina decidir cómo representar unos datos de la mejor forma posible.

Algunos de sus trabajos se pueden ver a través del enlace web -> https://research.tableau.com/user/jock-mackinlay.

2.5-Tableau

2.5.1 - ¿Qué es Tableau?

Tableau es una herramienta que ofrece una experiencia de exploración visual de datos y la de comunicar o narrar historias a través de los mismos, de forma interactiva e Intuitiva para que los usuarios puedan acceder, preparar y analizar fácilmente sus datos sin necesidad de codificación.

Tableau permite, por medio de simples interacciones, tener distintas perspectivas sobre un conjunto de datos. Además, nos ofrece numerosas técnicas de análisis y elementos estadísticos que podemos incorporar a nuestras visualizaciones para entender mejor los datos y extraer conocimiento de los mismos a partir del uso de representaciones gráficas.

2.6-Caso de estudio: Encuesta de salud de Catalunya

2.6.1- ¿Qué es la ESCA?, objetivos y finalidades

La Encuesta de salud de Catalunya recoge información de la población residente en Catalunya, sin límite de edad, sobre el estado de salud, los comportamientos relacionados con la salud y el uso de los servicios sanitarios.

El objetivo general de la Encuesta de salud de Catalunya es identificar a través de la información recogida en ella las necesidades de salud y los servicios, establecer perfiles de población diferentes, así como la evaluación de la eficacia de las intervenciones sanitarias, los objetivos de salud y de disminución de riesgo.

La finalidad de esta encuesta se centra en la planificación, evaluación y orientación de los servicios sanitarios y políticas de salud del Gobierno de Catalunya.

2.6.2-Contenido de las encuestas de salud de Catalunya

Los cuestionarios tratan información relacionada con:

- Estado de salud: Percepción del estado de salud, problemas de salud, estado de ánimo, etc.

-Comportamiento y estilo de vida: Consumo de tabaco, alcohol, nivel de actividad física habitual, hábitos alimenticios, actividades preventivas, etc.

-Uso de los servicios sanitarios: Consumo de medicamentos, uso de servicios de atención primaria o especializada, nivel de satisfacción con los servicios sanitarios utilizados, etc.

-Otras características relevantes: Sexo, edad, nacionalidad, situación familiar, situación laboral, nivel educativo, etc.

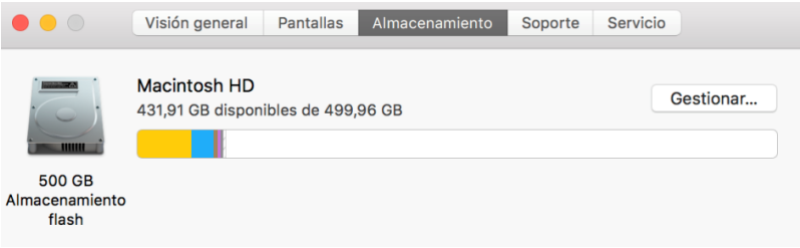
2.6.3-Resultados

Existen diferentes trabajos realizados con los datos de estas encuestas. Destacan, en particular, la comparación de indicadores de salud por grupos socioeconómicos ESCA 2006 y 2011 coordinación editorial y redacción por (Medina Antonia *et al.*, 2013) la población ocupada y trabajadora: comparación de indicadores de salud ESCA 2006 y 2011 (Brugulat, Medina, Mompart, *et al.*, 2013), y las regiones sanitarias en la ESCA 2010-2012, (Brugulat, Medina, Mompart Anna, *et al.*, 2013).

3-Evaluación de la infraestructura

3.1- Evaluación de la estructura existente

La infraestructura con la que se desarrollará el trabajo es a través de un equipo Mac con sistema operativo MacOS Sierra y las siguientes características:



Nombre del modelo:	MacBook Pro
Identificador del modelo:	MacBookPro14,3
Nombre del procesador:	Intel Core i7
Velocidad del procesador:	2,9 GHz
Cantidad de procesadores:	1
Cantidad total de núcleos:	4
Caché de nivel 2 (por núcleo):	256 KB
Caché de nivel 3:	8 MB
Memoria:	16 GB

3.2-Selección de productos.

Las aplicaciones que se utilizan para el desarrollo del trabajo son:

Para el análisis y preparación de datos se usarán **R** y **RStudio**, R es un entorno y lenguaje de programación con un enfoque al análisis estadístico. RStudio es un entorno de desarrollo integrado (IDE) para R (lenguaje de programación).

Para el diseño y modelado de la base de datos se usarán: R para crear las tablas para las relaciones y Excel al cual se exportarán las tablas para crear un único fichero Excel con las pestañas de las tablas creadas en R. Excel permite trabajar con hojas de cálculo.

Por último, TABLEAU como software de cuadro de mandos o *dashboards*. Tableau permite combinar en un mismo *dashboard* múltiples fuentes de datos para mantenerse al tanto de los datos de sensores, las entradas de operadores y los sistemas en tiempo real.

4- Análisis de datos a través de algoritmos "estructuralmente sencillos" como los árboles de decisión.

4.1 Objetivo del análisis

Con el objetivo de mejorar la salud autopercibida de las personas, se crea un modelo que determine la pertenencia de una observación a la clase buena, regular o mala salud autopercibida.

Con este modelo predictivo funcionando, las administraciones públicas contarían con una herramienta que les advertiría de cuándo se producen ciertos patrones que podrían llevar a una persona a tener una buena/regular/mala salud autopercibida.

4.2 Fuente de datos

Los datos con los que realizaremos los experimentos han sido escogidos y descargados de los open Data de la comunidad de Madrid, accesible en el siguiente enlace:
http://datos.madrid.es/portal/site/egob/menuitem.c05c1f754a33a9fbe4b2e4b284f1a5a0/?vgnextoid=77e22cbf3ee07510VgnVCM1000001d4a900aRCRD&vgnextchannel=374512b9ace9f310VgnVCM100000171f5a0aRCRD&vgnextfmt=default.
La encuesta se ha elaborado para obtener información sobre temas relacionados con la salud de los ciudadanos de Madrid; tanto sobre hábitos y estilos de vida como sobre los principales problemas de salud de los ciudadanos: enfermedades crónicas, dolencias, accidentes, limitaciones funcionales, etc.

La muestra, de 802 encuestas válidas representa a la población, de 16 y más años, residente en la ciudad de Madrid. Como criterio de estratificación se ha considerado una agrupación de los distritos de la ciudad en cuatro grupos según el nivel de desarrollo de su residencia. Por lo que sólo permite la explotación general a nivel de ciudad o de agrupación de distritos, pero no su explotación a nivel de distrito.

4.3 Definición del problema

En el proceso de descubrimiento de conocimiento en grandes volúmenes de datos es vital escoger las variables y características más adecuadas al presentar el algoritmo. Este problema puede tener diferentes enfoques: saber escoger las mejores características de los datos a partir de su análisis preliminar, eliminando las redundantes o reducir la dimensionalidad de los datos generando nuevas características. Las buenas prácticas requieren buenos atributos (*predictors).*

En segundo lugar, vamos a aplicar un algoritmo a los datos para descubrir el conocimiento desconocido previamente. Para ello utilizaremos algoritmos cuyo objetivo es poder predecir cuándo se producirá el hecho contenido en una variable llamada clasificadora. Estos modelos se llaman clasificadores o

predictivos. Y en concreto, utilizaremos los árboles de decisión. A los árboles de decisión que clasifican los datos del conjunto de entrada en función de una variable clasificadora categórica (es decir, que toma un conjunto finito de valores) se les llama árboles de clasificación. Estos árboles de decisión dan el nombre común CART al trabajo descrito por Breiman y Friedman en 1984.

En primer lugar, se cargan los datos, se realiza un preprocesado de los datos y se trabaja la calidad de los mismos. En esta primera fase se detecta que el conjunto de datos sufre la maldición de la dimensionalidad (*The curse of dimensionality*), con el conjunto de datos con 115 variables y 802 observaciones. Es imposible construir un buen modelo de clasificación, ya que se reducen las posibilidades de encontrar en el dataset todos los casos posibles que ilustren la combinación de las distintas dimensiones. Por tanto, faltarán ejemplos disponibles que permitan entrenar nuestro algoritmo de clasificación.

Para poder construir un buen modelo de clasificación para extraer conocimiento válido y comprensible de los predictores de la salud autopercibida, deberemos de ajustar nuestro conjunto de datos transformando los datos brutos para que nos aporte una mayor información significativa y utilizable que disminuya el desorden del ambiente (disminución de la entropía).

Para reducir la entropía, recodificaremos la variable clasificadora salud autopercibida en tres niveles en vez de cinco: malo o muy malo, regular, bueno o muy bueno, los extremos muy malo y muy bueno, podríamos considerarlos valores extremos, y en gráficas tener en cuenta estos 5 niveles.

4.3.1 Elaboración del primer modelo de clasificación

Para la elaboración del primer modelo de clasificación hemos utilizado el algoritmo generador de árboles de decisión C5.0.

Antes de abordar este tema se debe indicar que el dataset se divide en dos conjuntos (training y test set) según la regla del 2/3 partes. Y, por tanto, el conjunto de entrenamiento tiene 534 observaciones y el de test 268. Se confirma la existencia de una representación parecida de la variable clasificadora en los dos conjuntos.

Librería(C50)

Utilizaremos el algoritmo creado por Ross Quinlan C5.0. El árbol de decisión se construye mediante el control de la evolución de la entropía que podemos esperar de cada decisión (test) formando así las ramas del árbol. El crecimiento del árbol se controla mediante la noción de pureza de los nodos, y este termina cuando las ramas son puras o el número de instancias que corresponde a cada hoja es pequeño.

```
Call:
C5.0.default(x = xtrain[, -11], y = as.factor(xtrain$B1))

Classification Tree
Number of samples: 534
Number of predictors: 114

Tree size: 21

Non-standard options: attempt to group attributes


Call:
C5.0.default(x = xtrain[, -11], y = as.factor(xtrain$B1))
```

```
Evaluation on training data (534 cases):

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

            21   35( 6.6%)   <<


           (a)   (b)   (c)    <-classified as
          ----  ----  ----
           396     1     3    (a): class -0.56614405291637
             4    28     1    (b): class 0.681239272465303
            25     1    75    (c): class 1.92862259784698


        Attribute usage:

        100.00% B2_6
         80.71% C3
         21.35% C2_4
         19.29% EDAD_C
         16.85% F3
         13.86% J5
         10.49% A2
          9.55% B2_4
          6.55% B2_7
          4.68% C2_5
          3.93% H0
          3.37% E1_12_Est
          3.00% C2_9
          2.25% A3
          2.25% J3_AN
          1.87% H1
          0.75% A1
```

Ilustración 13- Evaluación datos entrenamiento 534 casos, hojas

El árbol se divide en 21 hojas, 534 casos y 114 variables, *Size* es el número de hojas no vacías en el árbol y Errors muestra el número y porcentaje de casos mal clasificados. El árbol, con 21 hojas con una tasa de error del 6,6%.

Clasifica correctamente los datos de entrenamiento: 499 casos y se equivoca en 56.

Las variables que más influyen en el modelo son:

1- __B.2.6__ Estado de salud: Durante las dos últimas semanas, ¿cómo calificaría su salud general?

2- __C.3__ Durante los últimos 12 MESES, ¿ha tenido que limitar o reducir la realización de sus actividades habituales a causa de este o estos problemas crónicos que me ha mencionado?

3- __C2_4__ Diabetes

4- __EDAD_C__ EDAD

5- __F3__ ¿Podría indicarme, aproximadamente, cuántas horas duerme

habitualmente al día? Incluya las horas de siesta.

```
Evaluation on training data (534 cases):

                Rules
           ----------------
             No      Errors

             12   43( 8.1%)   <<


            (a)   (b)   (c)    <-classified as
           ----  ----  ----
            391     1     8    (a): class
-0.56614405291637
              4    28     1    (b): class
0.681239272465303
             28     1    72    (c): class
1.92862259784698


        Attribute usage:

         94.76% B2_6
         86.33% C3
         74.91% C2_4
         71.16% B2_4
         68.54% B2_7
         61.24% C2_5
         36.70% EDAD_C
         15.17% F3
         11.80% A2
         11.80% J5
          7.30% C2_9
          7.30% E1_12_Est
          3.56% J3_AN
          3.18% H0
          2.43% A3
          0.37% A1
```

Ilustración 14- Evaluación datos entrenamiento 534 casos, reglas

El modelo tiene 12 reglas, 43 errores con una tasa de error de 8,1%.

El modelo clasifica los datos de entrenamiento según las reglas 491 correctamente y se equivoca en 64.

Ahora pasamos a evaluar el modelo, a través del estudio de los aciertos y errores en el dataset test.

```
   Cell Contents
|-------------------------|
|                       N |
|         N / Table Total |
|-------------------------|

Total Observations in Table:  268

                 | predicted B1
       actual B1 | -0.56614405291637 | 0.681239272465303 |  1.92862259784698 |         Row Total |
-------------------|-------------------|-------------------|-------------------|-------------------
-|
-0.56614405291637 |               187 |                 4 |                 5 |               196
                 |             0.698 |             0.015 |             0.019 |
-------------------|-------------------|-------------------|-------------------|-------------------
-|
0.681239272465303 |                 5 |                 4 |                 6 |                15
                 |             0.019 |             0.015 |             0.022 |
-------------------|-------------------|-------------------|-------------------|-------------------
-|
 1.92862259784698 |                19 |                 1 |                37 |                57
                 |             0.071 |             0.004 |             0.138 |
-------------------|-------------------|-------------------|-------------------|-------------------
-|
     Column Total |               211 |                 9 |                48 |               268
-------------------|-------------------|-------------------|-------------------|-------------------
-|

[1] 85.07463
```

Ilustración 15-Capacidad predictiva del modelo

Nuestro modelo tiene un porcentaje del 85,07% correctamente clasificados de los datos test. Ha clasificado erróneamente 40 casos de 268, la tasa de error es del 14,92% y ha acertado en la predicción en los 228 casos restantes.

4.3.2 Poda del árbol de decisión

Podemos mejorar el modelo realizando una poda. El algoritmo C5.0 construye el árbol en dos fases. En la primera fase se construye un árbol completo, intentando clasificar correctamente todos los elementos del conjunto de entrenamiento, la segunda fase se revisa, si alguna de las ramas generadas tiene una elevada tasa de error y no aporta nada se seguirán realizando particiones en dicha rama. Si ese fuese el caso, se colapsan los registros como una sola hoja en ese nivel y se recalcula la tasa de error.

```
Evaluation on training data (534
cases):

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

            24   32( 6.0%)   <<


           (a)   (b)   (c)    <-
classified as
          ----  ----  ----
           395     1     4    (a):
class -0.56614405291637
             4    28     1    (b):
class 0.681239272465303
            21     1    79    (c):
class 1.92862259784698


        Attribute usage:

        100.00% B2_6
         83.15% C3
         21.35% C2_4
         19.29% EDAD_C
         16.85% F3
         13.86% J5
         10.49% A2
          9.55% B2_4
          6.55% B2_7
          4.68% C2_5
          3.93% H0
          3.56% C2_10
          3.37% E1_12_Est
          3.00% C2_9
          2.25% A3
          2.25% J3_AN
          1.87% H1
          1.12% F8
          0.75% A1
```

Ilustración 16-Evaluación datos de entrenamiento 534 poda

El modelo de poda ha incrementado el número de hojas a 24, y ha reducido la tasa de error a 6% con 32 errores. El modelo clasifica el dataset de entreno correctamente 502 casos y falla en 53, se reducen los errores en el conjunto de entrenamiento.

Para el caso de las reglas, el número de reglas se ha reducido a 13 y la tasa de error se reduce a 7,3%, con 39 errores. El modelo ha clasificado el conjunto de entrenamiento 495 casos correctamente, cometiendo un error en 60 casos.

Para el conjunto test, el modelo una vez efectuada la poda ha aumentado un poco su capacidad predictiva a 85.44%. El modelo es capaz de clasificar el conjunto test correctamente en 229 y cometer un error en 27 casos.

Pred.PODA	-0.56614405291637	0.681239272465303	1.92862259784698
-0.56614405291637	186	4	17
0.681239272465303	4	4	1
1.92862259784698	6	7	39

Ilustración 17- Capacidad predictiva modelo con poda

```
Evaluation on training data (534
cases):

                Rules
           ----------------
             No      Errors

             13   39( 7.3%)   <<


            (a)   (b)   (c)    <-
classified as
           ----  ----  ----
            391     1     8    (a):
class -0.56614405291637
              4    28     1    (b):
class 0.681239272465303
             24     1    76    (c):
class 1.92862259784698


         Attribute usage:

          95.32% B2_6
          88.20% C3
          74.91% C2_4
          71.16% B2_4
          68.54% B2_7
          61.24% C2_5
          36.70% EDAD_C
          15.17% F3
          14.04% J5
          11.80% A2
           9.74% J3_AN
           7.30% C2_9
           7.30% E1_12_Est
           6.18% H1
           3.18% H0
           2.43% A3
           0.37% A1
```

Ilustración 18- Evaluación datos de entrenamiento 534 casos poda, reglas

4.3.3 *Boosting*

Otra técnica interesante para mejorar la calidad del árbol obtenido son las técnicas de *boosting*. La idea es que podemos crear un proceso que itere la creación de modelos, utilizando la información extraída de un modelo en el siguiente. Así, cuando el segundo modelo se genera, ya se dispone de la información de los errores que ha cometido el primero (de hecho, todos los anteriores) y esto permite que el primer modelo (y anteriores) voten (en el sentido de aportar su decisión tomada) en el momento de asignación del valor de la observación, de acuerdo al error cometido.

```
Call:
C5.0.default(x = xtrain[, -11], y =
as.factor(xtrain$B1), trials = 100)

Classification Tree
Number of samples: 534
Number of predictors: 114

Number of boosting iterations: 100
Average tree size: 24.5

Non-standard options: attempt to group
attributes


Call:
C5.0.default(x = xtrain[, -11], y =
as.factor(xtrain$B1), trials = 100)
```

```
     (a)   (b)   (c)    <-classified as
    ----  ----  ----
     400                (a): class -0.56614405291637
            33          (b): class 0.681239272465303
                 101    (c): class 1.92862259784698


Attribute usage:

100.00% A3
100.00% B2_1
100.00% B2_4
100.00% B2_6
100.00% C2_4
100.00% C2_16
100.00% C3
100.00% D1
100.00% D2
100.00% E1_3
100.00% E1_12
100.00% E1_12_Est
100.00% F3
100.00% F9S1
100.00% G5B_3
```

Ilustración 19- Evaluación de datos de entrenamiento, Boosting

En este caso el algoritmo boosting calcula el error cometido sobre 100 de estos mismos subconjuntos, poda toda una rama de cálculos potenciales, y el error en este caso para los datos de entrenamiento son 0, ha acertado en clasificar en todos los casos. Y lo mismo pasa con las reglas.

Veamos la capacidad predictiva del modelo para el conjunto test:

```
   Cell Contents
|-------------------------|
|                       N |
|         N / Table Total |
|-------------------------|


Total Observations in Table:  268


             | predicted B1
      actual B1 | -0.56614405291637 | 0.681239272465303 |  1.92862259784698 |         Row Total |
-------------|-------------------|-------------------|-------------------|-------------------|
-0.56614405291637 |               191 |                 0 |                 5 |               196 |
                 |             0.713 |             0.000 |             0.019 |                   |
-------------|-------------------|-------------------|-------------------|-------------------|
0.681239272465303 |                 4 |                 5 |                 6 |                15 |
                 |             0.015 |             0.019 |             0.022 |                   |
-------------|-------------------|-------------------|-------------------|-------------------|
 1.92862259784698 |                16 |                 2 |                39 |                57 |
                 |             0.060 |             0.007 |             0.146 |                   |
-------------|-------------------|-------------------|-------------------|-------------------|
    Column Total |               211 |                 7 |                50 |               268 |
-------------|-------------------|-------------------|-------------------|-------------------|


[1] 87.68657
```

Ilustración 20- Capacidad predictiva Boosting

El modelo ha aumentado su capacidad predictiva a 87,68%, esta técnica ha mejorado el modelo.

4.4 Extracción de características

El principio divide y vencerás, para grandes cantidades de variables o dominio a estudiar, no nos deja analizar exhaustivamente las relaciones entre todos los datos, sus dependencias, correlaciones…

La eficiencia del algoritmo de clasificación de la población según su salud autopercibida, depende de la naturaleza y calidad de los datos que se utilizan. Un número demasiado grande de variables genera un sobreajuste, si el ajuste es demasiado fuerte, el modelo ya no refleja una ley general al fenómeno estudiado, sino solamente las características del conjunto de entrenamiento, perdiendo sus capacidades de clasificación cuando se aplique sobre nuevas variables explicativas.

Para seleccionar del dataset de 115 variables las más relevantes para el problema a resolver, utilizaremos dos técnicas:

- **Filter** que excluye las variables irrelevantes. Para ello utilizaremos el algoritmo *Random Forest* y el análisis de correlaciones.

- **Wrapper** que genera nuevas características, en este caso utilizaremos el análisis de Componentes Principales (PCA) y descomposión en valores singulares (SVD)

4.4.1 Técnicas *filter*

4.4.1.1- Eliminar características redundantes a través del análisis de correlaciones

Podemos comprobar a través del análisis de correlaciones la influencia de las variables en el conjunto del dataset y eliminar las características redundantes.

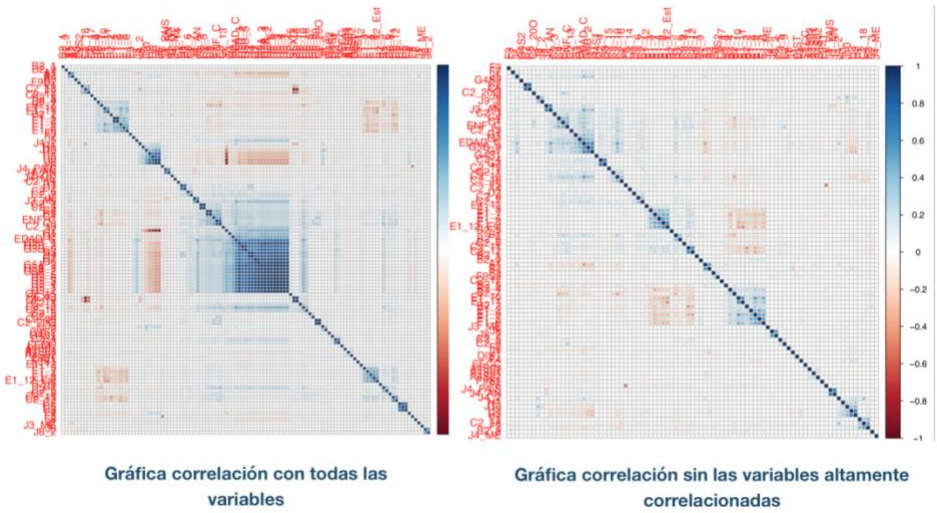


Ilustración 21- Gáfica correlaciones todas las variables Vs sin las variables altamente correlacionadas

Este método tiene muchas limitaciones, ya que lo que busca eliminar son las variables de buenas a primeras en función de su coeficiente de correlación alto, considerando que las variables funcionan independientemente en lugar de combinadamente.

4.4.1.2- Puntuar características por importancia: Random Forest

Cuando hemos creado el primer árbol de decisión con C5.0 este nos daba la relación de las variables más influyentes en el modelo clasificador, podemos utilizar otro algoritmo, *Random Forest*, que proporcionará una lista muy interesante de variables clasificadas según su nivel de importancia. Otra de las ventajas de la aplicación de este algoritmo es que se ejecuta eficientemente en grandes conjuntos de datos, permite trabajar con cientos de variables independientes sin excluir ninguna, recupera eficazmente los valores perdidos de un dataset (función = rfImpute).

Para este caso vamos a indicar al algoritmo *randomForest* que realice el estudio con 100 árboles y con muestras de reemplazo. La proporción de datos de entrenamiento y test es como en el caso anterior 2/3.

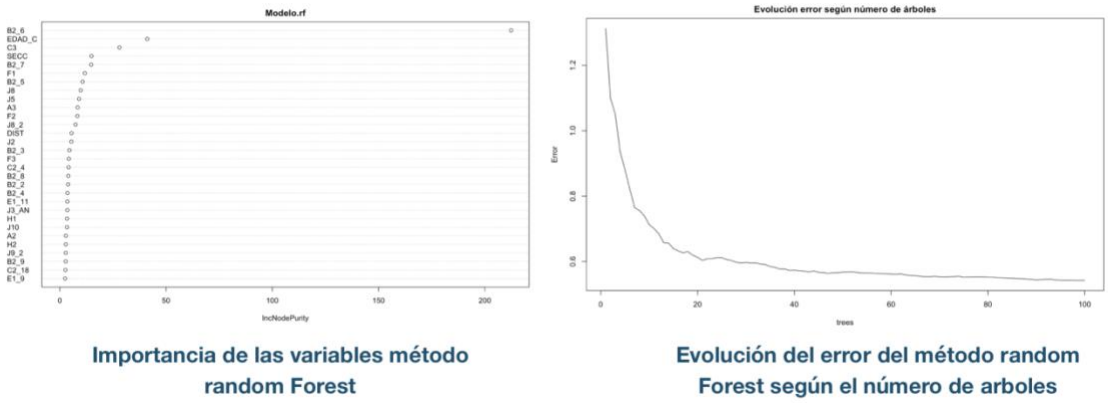


Ilustración 22- Importancia de las variables método randomForest y la evolución del error según el número de árboles

El problema de este algoritmo es que favorece las variables categóricas que tienen un mayor número de niveles por encima de las que tienen un número más reducido y favorece los grupos más pequeños cuando las variables están correlacionadas, además tienen el inconveniente de que es un método de difícil representación.
Veamos las diferencias de la importancia de las variables en función del algoritmo.

Ilustración 23- Diferencias de las variables de influencia en el modelo

Los métodos anteriores tienen por objeto determinar las variables con mayor eficacia clasificadora en el conjunto de datos de la encuesta de salud de CM, aunque se ha de considerar que ciertos atributos no funcionan solos, sino de manera combinada con otras variables.

Para solucionar los problemas planteados anteriores se opta por el análisis de componentes principales (PCA) y descomposición en valores singulares (SVD), en lugar de eliminar variables, se construyen nuevas, tales que su ranking resulte más sencillo, y a continuación se realizará la selección.

4.4.2 Técnicas Wrapper

4.4.2.1-Mejorar el modelo generado aplicando SVD

Podemos utilizar este método de descomposición factorial para la extracción de características. Aplicamos la descomposición factorial llamada "Singular Value Descomposition"(SVD) para encontrar una nueva forma de representar las variables de la encuesta de salud de la Comunidad de Madrid con mayor correlación, perdiendo la mínima información.

Los 42 factores principales representan el 80% de la información del conjunto de datos.

Veamos el porcentaje de varianza explicada.

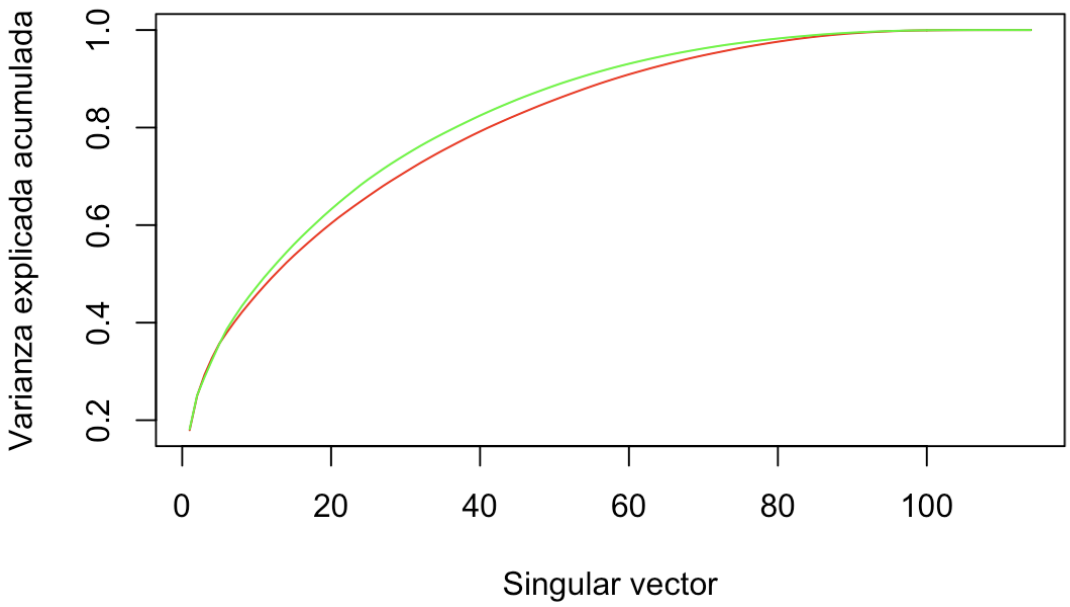


Ilustración 24-Porcentaje de varianza explicada por el modelo

Creamos otro árbol de decisión con la incorporación a los datos de la variable C1 con 42 variables que acumula el 80% de varianza.

Nuestro modelo ha empeorado con la incorporación de la nueva variable, ya que añade más información redundante al conjunto datos. Esta nueva variable aporta una menor información significativa y utilizable y hace que aumente la entropía.

```
   Cell Contents
|-------------------------|
|                       N |
|         N / Table Total |
|-------------------------|


Total Observations in Table:  268


                 | predicted B1
        actual B1 | -0.56614405291637 | 0.681239272465303 |  1.92862259784698 |         Row Total
|
------------------|-------------------|-------------------|-------------------|------------------
-|
-0.56614405291637 |               187 |                 4 |                 5 |               196
|
                  |             0.698 |             0.015 |             0.019 |
|
------------------|-------------------|-------------------|-------------------|------------------
-|
0.681239272465303 |                 5 |                 2 |                 8 |                15
|
                  |             0.019 |             0.007 |             0.030 |
|
------------------|-------------------|-------------------|-------------------|------------------
-|
 1.92862259784698 |                19 |                 0 |                38 |                57
|
                  |             0.071 |             0.000 |             0.142 |
|
------------------|-------------------|-------------------|-------------------|------------------
-|
     Column Total |               211 |                 6 |                51 |               268
|
------------------|-------------------|-------------------|-------------------|------------------
-|


[1] 84.70149
```

Ilustración 25- Capacidad predictiva del árbol de decisión una vez reducida la dimensionalidad con SVD

La capacidad predictiva del modelo ha descendido a 84.70%, el modelo en un conjunto de test es capaz de acertar 227 casos y equivocarse en 26 casos.

Nuestro modelo ha empeorado con la incorporación de la nueva variable, ya que añade más información redundante al conjunto datos. Esta nueva variable aporta una menor información significativa y utilizable y hace que aumente la entropía.

4.4.2.2-Análisis de componentes principales

El algoritmo PCA reduce la dimensionalidad a través de una transformación lineal que escoge un nuevo sistema de coordenadas para el conjunto original de datos, en el cual la varianza de mayor tamaño del conjunto de datos se recoge en el primer eje (llamado el primer componente principal), la segunda varianza más grande en el segundo eje, y así sucesivamente.

El PCA permite pasar a un nuevo conjunto de variables: los componentes o factores, que se caracterizan por estar relacionados entre sí y que, además, pueden ordenarse de acuerdo con la información que llevan incorporada.

En resumen, nos aparecen los valores propios de cada componente principal, la proporción de varianza explicada por cada componente y la varianza explicada acumulativa. La proporción que explican los 44 primeros componentes es del 80 %. Podemos conocer el coeficiente para cada variable del dataset dentro de cada componente (*loadings*), en el componente 1 con una proporción del 17,95% de explicación del componente uno en el data set, tenemos la variable EDAD_C con un coeficiente del -17,7 % y así para todas las variables de los componentes.

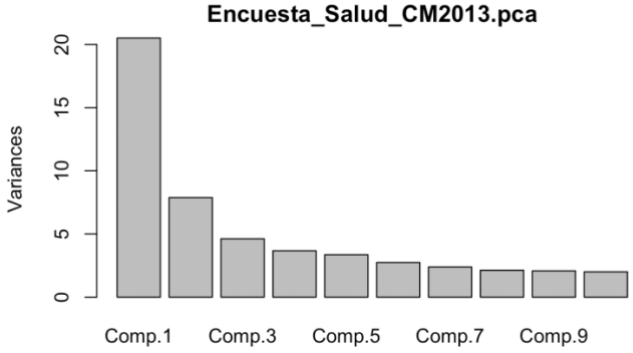


Ilustración 26- Varianza de cada componente

Dibujemos las observaciones en función de la salud autopercibida. Para ello dibujamos los scores de los 44 componentes principales y asignamos color en función de la variable clasificadora salud autopercibida.

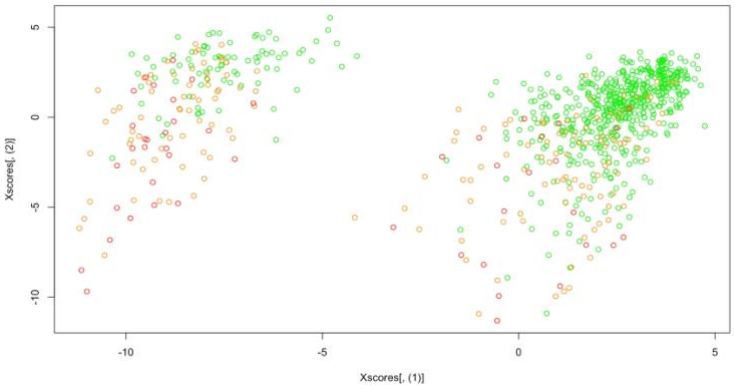


Ilustración 27- Mapa de los dos primeros componentes en función de la variable salud autopercibida

Se puede ver que es bastante complicado diferenciar entre los individuos con muy buena, buena, regular, salud autopercibida, existiendo mucho solapamiento. Por ello es que solo estos dos componentes representan juntos tan solo el 24,82% de la explicación.

Veamos otro gráfico,

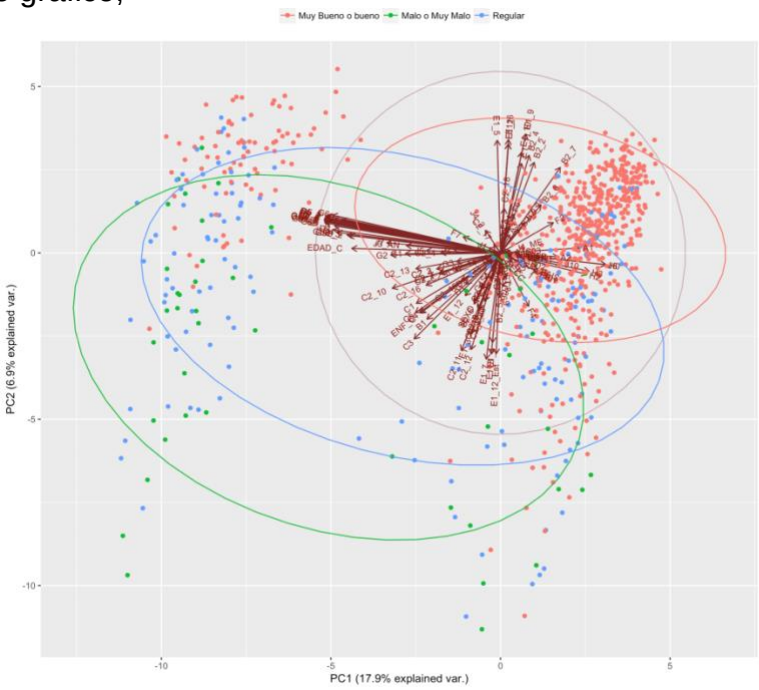


Ilustración 28- Mapeo de las dos primeras componentes en función del SAP y los demás predictores

Azul	Salud autopercibida Regular. Estos individuos se caracterizan por estar la mayoría de ellos en el tercer cuadrante; los individuos del tercer cuadrante son los que tienen más edad, (ver flecha EDAD_C), padecen algún tipo de enfermedad o trastorno crónico tanto autopercibido como reconocido por el médico, y mala salud mental (variables E1_1:E1_12). Se caracterizan mayoritariamente por mala condición física.
Verde	Salud autopercibida mala o muy mala; estos individuos son muy parecidos a los del grupo azul, aunque la diferencia es que hay más individuos que pertenecen al cuarto cuadrante, se caracterizan por haber tenido algún accidente, o tener reducidas horas de sueño.
Rojo	Salud autopercibida buena y muy buena; estos individuos están la mayoría en el cuadrante 1, caracterizado por las personas que viven en el hogar, y los menores que viven en el hogar, el tiempo que tardan en desplazarse al trabajo, el hecho de trabajar, el tipo de contrato.

Para nuestro conjunto de datos ninguno de estos dos métodos anteriores para mejorar el árbol de decisión es aplicable, ya que la lógica de utilizar estos algoritmos SVD y PCA es para *datasets* con altas correlaciones, donde si es posible reducir la dimensionalidad sin pérdida de información.

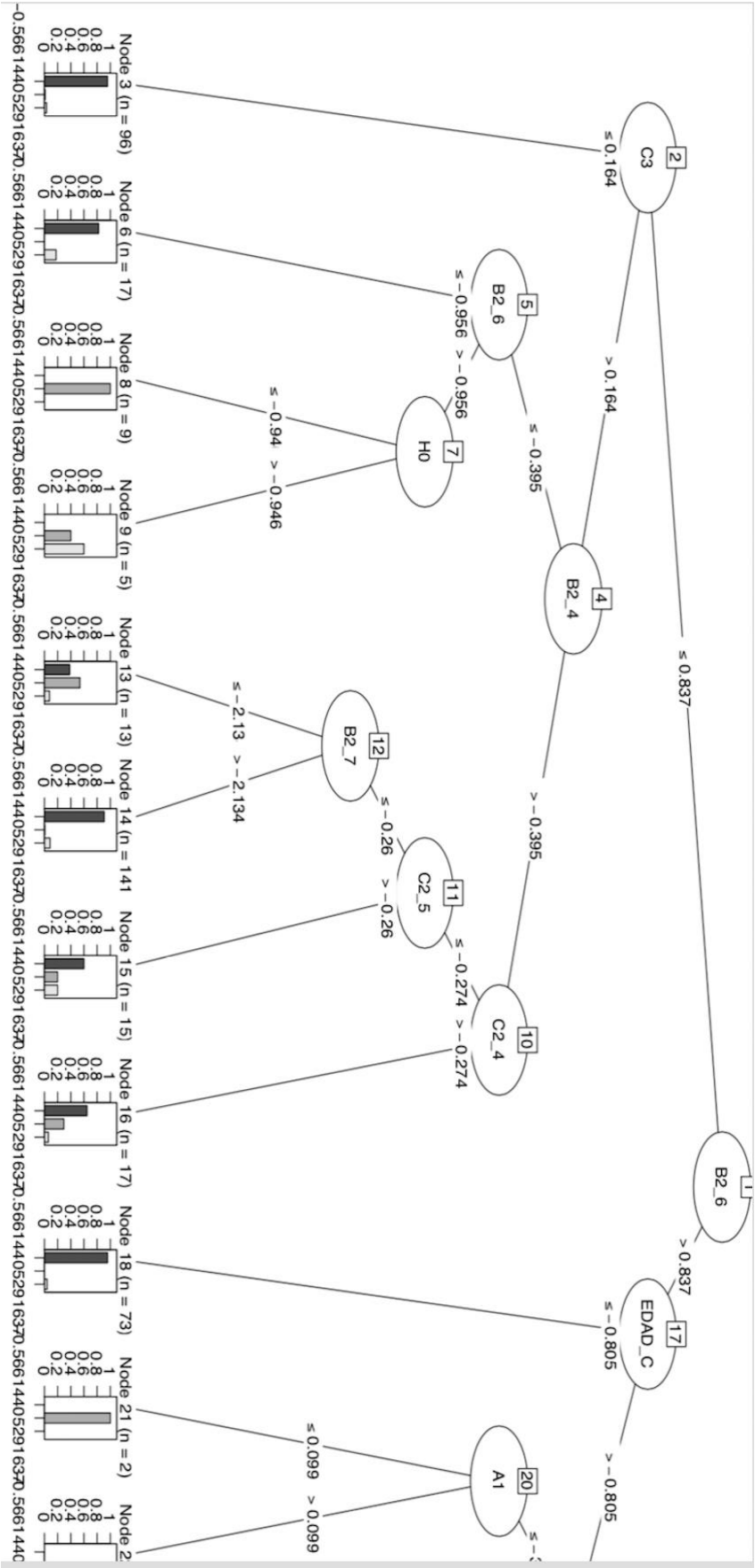


Ilustración 29- Árbol de decisión, parte a

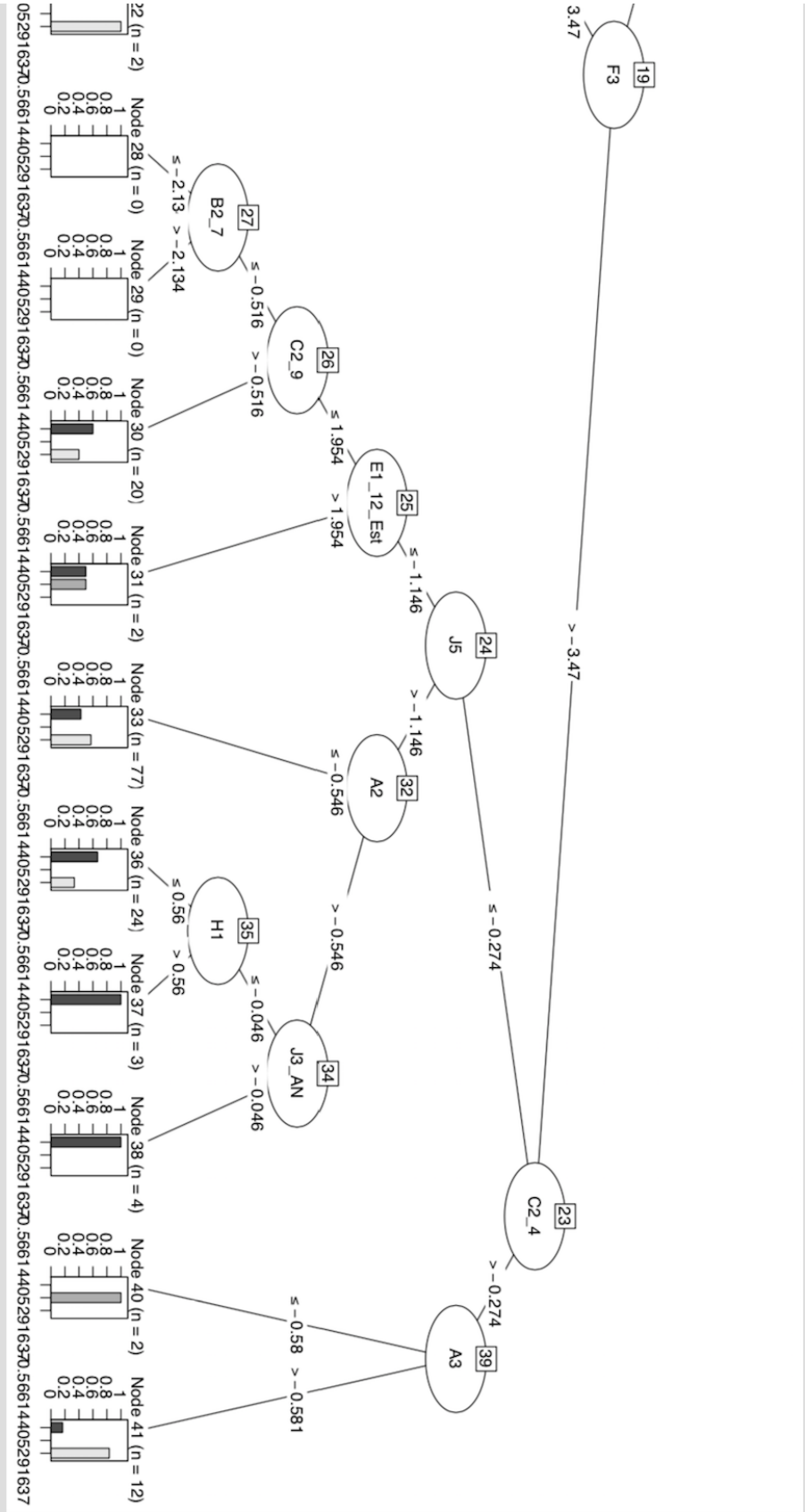


Ilustración 30- Árbol de decisión, parte b

5. DISEÑO, CONSTRUCCIÓN DEL ALMACEN DE DATOS para el análisis de encuestas de la Encuesta de Salud de Catalunya. (ESCA)

5.1-Requerimientos

Los requerimientos funcionales definen las operaciones que el sistema será capaz de realizar y describen las características que deberá contener la información alojada.

Las necesidades de los usuarios, en el caso de estudio, residen en la ausencia de una herramienta de *reporting* interactiva que les permita dar respuesta a través de visualizaciones sencillas a preguntas relacionadas con información de la población residente en Cataluña sobre el estado de salud, los comportamientos relacionados con la salud y el uso de los servicios sanitarios.

Es por ello que se les ofrece un prototipo de cuadro de mandos que permita la presentación o explicación de la historia o conocimiento a través de los datos de la ESCA, así como explorar, filtrar e interactuar con los datos de la ESCA para construir entendimiento.

Entender los factores relevantes de la encuesta de salud de Catalunya es crítico para obtener una comprensión sólida de sus requerimientos, y crucial para la traslación de estos requerimientos en el diseño de nuestra base de datos y sistema de *reporting*. Lo lógico es que el paso anterior de análisis de datos con algoritmos de clasificación lo hubiéramos realizado con los datos de la ESCA, pero estos no fueron cedidos hasta el día 18 de diciembre de 2017.

5.1-2- Necesidades de *reporting*

Se analizan las necesidades de *reporting* para determinar el nivel de agregación de datos, el tipo de interrogaciones y la estructura de los universos de datos.

El sistema de *reporting* ha de permitir:

- Conocer las variables de la encuesta de salud de Catalunya, características demográficas; estructura familiar, situación laboral, nivel de estudios, incertidumbre laboral, condiciones de la vivienda, estado de salud, prácticas preventivas, estilos de vida y utilización de servicios sanitarios.

- Este conocimiento debe poderse analizar desde diferentes puntos de vista. proporcionando diferentes perspectivas para el análisis. En este caso serían:

    - Por región sanitaria geográfica

- Por sexo
- Por edad
- Por clase social
- Por nivel de estudios
- Por periodo anual (el fichero cedido solo cuenta con un periodo anual, en un proyecto real se debería contar con todos los periodos anuales para estudiar la evolución del problema)

5.1.3- Tipos de usuarios

El sistema de *reporting* dinámico está orientado a dos tipos diferentes de usuarios:

- Los usuarios avanzados especializados en datos (ESCA)los cuales valoran, sobre todo, poder explorar, filtrar e interactuar con los datos que les presentemos; y

- Los usuarios no especializados (consumidores de información de la ESCA), cuyo valor percibido es la visualización en sí. Solo buscan una presentación o explicación de la historia o conocimiento que estemos transmitiendo.

5.1.4- Características que ha de reunir el sistema de *reporting*

- Debe ser fácil, utilizable e intuitivo para el usuario. Debe proporcionar un resumen entre 5 y 9 conceptos, que incluyan el sexo, edad, nivel de estudios, clase social, nivel económico, comunidad autónoma y año.
- En este resumen se deben poder resumir las áreas principales. El resumen debería utilizar un código de colores donde se use una escala del 1-peor al 3 -mejor, y colores en forma de semáforo ,1rojo, 2naranja y 3verde.
- Los gráficos han de ser interactivos y permitir a través de un menú variable seleccionar varios valores de todas las dimensiones.
- El sistema de *reporting* también debería poder permitir ver un detalle de los resultados si se quiere para obtener más información.
- La visualización debería, al menos, poder comparar entre al menos 2 dimensiones, por ejemplo, una gráfica de barras horizontal que muestre en el eje de las Y los diferentes sexos de los individuos (mujer y hombre), en el eje de las X la cantidad de individuos que consume alcohol, diferenciándose por color en función de su Salud autopercibida.
- La visualización de la región geográfica debería, al menos, mostrar las diferentes regiones y poder seleccionarlas, mostrar por colores las regiones donde mayor proporción de habitantes por habitante tienen esa variable seleccionada. Se podría optar por un mayor detalle, y seleccionar la dimensión sexo, nivel educativo ...
- Las cantidades que se muestren han de estar representadas de forma precisa.
- Ha de permitir la comparación entre cantidades de las diferentes dimensiones.

- Permitir establecer un orden a través de la cantidad de población con esa variable, y poder detectar máximos y mínimos.
- Deberá añadir una leyenda que explique la historia de los datos y el funcionamiento de la visualización (ha de servir de guía al usuario).

5.2- Análisis de los datos.

Del análisis de las fuentes de información puede desprenderse información clave y la evolución del proyecto TFM, así como la identificación de los riesgos vinculados.

Como punto de partida, se cuenta con la información blanca contenida en las fuentes de datos abiertos de la Encuesta de Salud de Catalunya.
Hay un total de 11 encuestas en la página web de la ESCA, realizadas a una muestra representativa de la sociedad catalana seleccionada aleatoriamente.

La primera encuesta data de 1994. La siguiente se realizó ocho años más tarde en 2002. Las dos siguientes se realizan transcurridos cuatro años, concretamente en 2006 y 2010, y las consecutivas con periodicidad anual hasta 2017, año en curso. (Actualmente, a fecha de 15 de enero se ha añadido otra con datos en progreso de 2018).

De los resultados de las 11 Encuestas de Salud de Catalunya, solo disponemos de los años 2011 hasta el 2016 de fuentes de datos open data con formato descargable Excel .xlsx. En los otros años, el formato es pdf y los datos al no ser estructurados no pueden ser incorporados a una base de datos relacional.

En resumen, contamos con dos niveles de datos abiertos:

a) 5 ficheros (resultados ESCA 1994, 2002, 2006, 2010 y 2017 encuesta en curso) publicados en el sitio web de ESCA de nivel 1.

b) 11 ficheros (resultados ESCA 2011, 2012, 2013, 2014, 2015, 2016) publicados en el sitio web ESCA de nivel 2.

La evolución del plan de recogida de datos ha resultado negativa. Con los datos open data de la ESCA disponibles en la página web ESCA, es difícil cumplir con los requerimientos de información. Los datos de los que disponemos ya han sido extraídos de fuentes primarias, depurados y publicada la información agregada en forma de tablas.

El problema reside en que los datos son secundarios. La ESCA los había recopilado para propósitos distintos al nuestro.

Es imposible reconfigurar los datos a fin de incrementar su utilidad. Los datos no son válidos, ni aplicables al problema que se pretende resolver.

Para dar solución a estos problemas hemos seguido el protocolo para solicitar la cesión de los datos a la ESCA, un proceso burocrático entre la ESCA, la UOC, los consultores y yo misma, que ha tenido una duración de varias semanas hasta concluir el 18 de diciembre de 2017, momento en el que obtenemos el fichero ZIP con un archivo .sav en su interior con los datos de la ESCA para el periodo 2016.

5.2.1- Análisis de los datos cedidos ESCA_2016

Los datos de la ESCA tienen el mismo problema que los de la Encuesta de Salud de la comunidad de Madrid, se trata de unos datos con un elevado número de dimensiones: 1377 y con un reducido número de observaciones: 4818 que no permiten entrenar nuestro algoritmo clasificador.

Los datos de la ESCA contienen metadatos con etiquetas de variables, valores y valores perdidos.

Con el objetivo de crear un prototipo de herramienta de *reporting* interactiva, no se incorporan de manera exhaustiva, el total de dimensiones debido a los reducidos recursos, limitación temporal y a la incidencia surgida con el retraso en la obtención de los datos cedidos Por ello, utilizaremos solo unas variables, las cuáles han sido seleccionadas por la importancia de los modelos de clasificación, a partir de los árboles de decisión del análisis de las Encuestas de Madrid y una selección arbitraria, considerando la información Open data de la ESCA.

Estas variables se alojarán en diferentes archivos .csv creados en R a partir del archivo.sav cedido por la ESCA, en función de los siguientes términos:

- Características sociodemográficas
- Cobertura sanitaria
- Estado de salud y calidad de vida
- Morbilidad crónica y accidentes
- Consumo de medicamentos y restricciones actividad física
- Salud mental
- Limitaciones, discapacidades y autonomía personal
- Prácticas preventivas
- Soporte social
- Bienestar psicológico
- Estilo de vida
- Actividad física
- Visitas medicas
- Atención primaria, atención de urgencias y hospitalizaciones

Cada uno de estos archivos ha de contener un mínimo de 5 variables y un máximo de 9 que sean los más representativos o influyentes del concepto al que pertenece el grupo clasificador de preguntas.

La estructuración rígida y potente de las arquitecturas BI permite obtener excelentes rendimientos en términos de *reporting* o de navegación en los datos.

Nuestro sistema de BI trabaja sobre datos estructurados. La arquitectura BI proporcionara recorridos interactivos implementados en los cubos OLAP.

Veamos esas fuentes de datos creadas.

5.2.1.1 Tablas de los grupos de preguntas de la ESCA

	Nombre atributo fichero	Descripción	Tipo de dato	Nivel de información
Características sociodemograficas	V15	Nacionalidad	Cadena	4 niveles(1-Española, 2-Española y extranjera , 3-Extranjera, 9-NC)
	AV20	Probabilidad de perder el trabajo	Cadena	7 niveles (0- NP(Menores), 1-Muy probable, 2-Probable, 3-Improbable, 4-Muy improbable, 8-NP(UM no trabaja), 9-NC) Existencia de NA
	AV115	Satisfacción con las condiciones laborales	Cadena	7 niveles(o-NP(Menores), 1-Muy satisfaceho, 2- Satisfecho, 3- Insatisfecho, 4-Muy instatisfecho, 8-NP (no trabaja), 9-NS/NC) Existencia de NA.
	V7_DISC1	Número de personas discapacitadas en la casa	Cadena	7 niveles(0-Ninguno, 1-Una ,2-Dos, 3- Tres, 4- Cuatro o más, 8- NP (Vive solo), 9- NS/NC
	V116	Como llega a fin de mes	Cadena	8 niveles(1-Mucha dificultad, 2-Dificultad, 3-Cierta dificultad, 4-Cierta facilidad, 5-Facilidad, 6-Mucha facilidad, 8_NP (la persona informada no es progenitor), 9-NS/NC.
	AV19R	Situación laboral de la persona entrevistada	Cadena	7 niveles (0-NP (menores), 1- TRabaja, 2-Parado, 3- Trabaja en casa, 4- Estudia, 5-Otros, 99-NC)

Ilustración 31- Tabla características sociodemográficas ESCA 2016

	Nombre atributo fichero	Descripción	Tipo de dato	Nivel de información
Estilo de vida	PREDI_7	Bebidas azucaradas	Número entero	Métrica enteros , valores 77- NP(cuestionario indirecto y menores), 99-NS/NC) Existencia de NA
	PREDI_8	Copas de vino a la semana	Número entero	Métrica enteros , valores 77- NP(cuestionario indirecto y menores), 99-NS/NC) Existencia de NA
	PREDI_13	Consume carne blanca en vez de roja	Cadena	4 niveles(0-NP, 1-Sí, 2-No, 3-No consume carne) Existencia de NA
	ACT_FIS-R	Actividad física saludable des de 2016	Cadena	4 niveles(1- saludable, 2- No saludable, 7-NP (cuestionario indirecto y menores), 8-NP(70 años y más). Existencia de NA
	MOB1	Medio de transporte utilizado habitualmente	Cadena	7 niveles(0-NP(cuestionario indirecto y menores), 1-En coche o moto, 2-Trasporte público , 3-A pie, 4- Bicicleta, 5-Otra posibilidad, 6-No se desplaza) Existencia de NA
	CONS_TABAC_R	Consumo de tabaco	Cadena	5 niveles(0-NP, 1-Fumador, 2-Exfumador, 3-No fumador, 9-NS/NC) Existencia de NA
	INDEX_ALCOH	Índice de consumo de alcohol	Cadena	NC

Ilustración 32- Tabla Estilo de vida ESCA 2016

	Nombre atributo fichero	Descripción	Tipo de dato	Nivel de información
Cobertura sanitaria	V34_1	Cobertura sanitaria	Cadena	5 niveles (1- Solo cobertura pública, 2- Cobertura pública y privada, 3- Solo cobertura privada, 4- No tiene ninguna cobertura, 9- NC)
	V35	Modalidad de servicios sanitarios más utilizados el último año	Cadena	7 niveles (1- Ninguno, 2- Servicios sanitarios públicos , 3- Seguro mutua obligatoria , 4- Seguro individual o de empresa, 5- Médicos particulares, 6- Otros servicios, 9-NS/NC)
	V36	Grado de satisfacción con los servicios utilizados el ultimo año	Cadena	7-Niveles (0-NP, 1-Muy satisfecho, 2-Satisfecho, 3-Insatisfecho, 4-Muy insatisfecho, 8- NP (No ha utilizado ninguno), 9- NS/NC) Existen valores NA

Ilustración 33- Tabla Cobertura sanitaria ESCA 2016

	Nombre atributo fichero	Descripción	Tipo de dato	Nivel de información
Salud mental	INDEX_GOLD	Índice de salud mental (GHQ)	Cadena	3 niveles (0-NP(cuestionario indirecto y menores), 1-Riesgo de mala salud mental, 2- Salud mental normal) Existencia de NA

Ilustración 34- Tabla Salud mental ESCA 2016

	Nombre atributo fichero	Descripción	Nivel de información
Soporte social	D16_5:DD16 012_11	Soporte social	Variables con todas las observaciones NA, esta tabla no la podemos crear, ya que no tiene valores en los registros.

Ilustración 35- Tabla soporte social

	Nombre atributo fichero	Descripción	Tipo de dato	Nivel de información
Limitaciones, discapacidades y autonomia	V48	Limitado para realizar las actividades cotidianas	Cadena	4 niveles(1- Sí, gravemente limitado/a, 2-Sí, limitado/a no gravemente, 3-No limitado/a, 9- NS/NC)
	V49	Antigüedad de la limitación	Cadena	8 niveles(1-Menos de 6 meses, 2- Entre 6 meses y año, 3-Entre 1 y 5 años, 4- Entre 5 y 10 años, 5- Mas de 10 años, 6- Siempre, 8-NP (No tiene limitaciones), 9-NS/NC
	DISCAP	Tiene alguna discapacidad	Cadena	3 niveles(1-Sí, 2-No tiene, 9-NS/NC)

Ilustración 36- Tabla limitaciones, discapacidades y autonomía ESCA 2016

	Nombre atributo fichero	Descripción	Tipo de dato	Nivel de información
Prácticas preventivas	AV60	Vacuna de la gripe	Cadena	5 niveles (0-NP (cuestionario indirecto y menores), 1- Sí, 2-No, 8-NP(Menores de 60 años), 9- NS/NC) Existe valores NA
	AV61	Medir la tensión arterial periodicamente	Cadena	4 niveles (0-NP (cuestionario indirecto y menores), 1- Sí, 2-No, 9- NS/NC) Existe valores NA
	AV62	Revisión nivel colesterol periodica	Cadena	4 niveles (0-NP (cuestionario indirecto y menores), 1- Sí, 2-No, 9- NS/NC) Existe valores NA
	AV63	Mamografia periodica	Cadena	4 niveles (0-NP (cuestionario indirecto y menores), 1- Sí, 2-No, 9- NS/NC) Existe valores NA
	AV66	Citologia periodica	Cadena	6 niveles (0-NP (cuestionario indirecto y menores), 1- Sí, 2-No, 3- No se ha realizado ninguna, 8- NP (hombre), 9-NS/NC) Existe valores NA
	CC1	Prueba de sangre en heces	Cadena	5 niveles (0-NP (cuestionario indirecto y menores), 1- Sí, 2-No, 8-NP(Menores de 40. años), 9- NS/NC) Existe valores NA
	CC4	Prueba colonoscopia	Cadena	5 niveles (0-NP (cuestionario indirecto y menores), 1- Sí, 2-No, 8-NP(Menores de 40. años), 9- NS/NC) Existe valores NA

Ilustración 37- Tabla prácticas preventivas ESCA 2016

	Nombre atributo fichero	Descripción	Tipo de dato	Nivel de información
Visitas médicas, atención primaria, atención de urgencias	VISIT_MED	Visitas médicas últimos 12 meses	Cadena	3 niveles (1-Sí, 2-No, 9-NS/NC)
	VISIT_MED_ESP	Visitas médicas especialista 12 meses	Cadena	3 niveles (0-No, 1-Sí, 9-NS/NC)
	VISIT_MED_MEG_PD	Visitas medicina general 12 meses	Cadena	3 niveles (0-No, 1-Sí, 9-NS/NC)
	V106_R	Ha sido hospitalizada el otro año	Cadena	3 niveles(1-Sí, 2-No, 9 NS/NC)
	NC1	Atención sanitaria no atendida	Cadena	3 niveles(1-Sí, 2-No, 9-NSNC)
	V104	Ha utilizado el servicio de urgencias el último años	Número entero	Metrica entero, valores 0- Ninguno y 99-NS/NC
	NC2	Motivo de la atención sanitaria no atendida	Cadena	14 niveles(1-Perdida de horas en el trabajo, 2-Miedo a perder el trabajo, 3-Incapacidad física por desplazarse, 4-Problemas en los transportes, 5-No resuelven el problema de salud, 6-Tener que cuidar de un enfermo de casa, 7-Le han dado hora para más tarde, 8-No era muy grave, 9-Lo resolvió en la farmacia, 10-Tiene otros remedios curativos, 11-Problemas económicos , 88-NP(No ha tenido problema de salud), 98-Otros motivos, 99-NS/NC
	V105	Lugar de la visita a urgencias	Cadena	8 niveles (1- Hospital público , 2-Hospital privado , 3-Llamó a un servicio de urgencias médicas, 4-Centro de atención primaria o continuada, 5- Consulta de un medico o centro privado, 6-Otro lugar, 8-NP(no fue a ningún servicio de urgencias), 9-NS/NC)

Ilustración 38- Visitas Médicas, atención primaria, atención de urgencias

	Nombre atributo fichero	Descripción	Tipo de dato	Nivel de información
Bienestar psicólogos	AV40a1-5	Esta ansioso o deprimido	Cadena	7 niveles(0-NP(Menores), 1-No está, 2-Esta levemente, 3-Esta moderadamente, 4-Esta muy ansioso o deprimido, 5-Esta extremadamente ansioso o deprimido, 9-NC) Existencia de valores NA.
	BP1_1	Optimista respecto al futuro	Cadena	6 niveles (0-NP(cuestionario indirecto y menor), 1-Nunca, 2-Muy pocas veces, 3-Algunas veces, 4-Frecuentemente, 5-Siempre) Existencia valores NA
	BP1_3	Se ha sentido relajada	Cadena	6 niveles (0-NP(cuestionario indirecto y menor), 1-Nunca, 2-Muy pocas veces, 3-Algunas veces, 4-Frecuentemente, 5-Siempre) Existencia valores NA
	BP1_8	Se ha sentido bien con sigo misma	Cadena	6 niveles (0-NP(cuestionario indirecto y menor), 1-Nunca, 2-Muy pocas veces, 3-Algunas veces, 4-Frecuentemente, 5-Siempre) Existencia valores NA
	BP1_11	Ha sido capaz de tomar sus propias decisiones	Cadena	6 niveles (0-NP(cuestionario indirecto y menor), 1-Nunca, 2-Muy pocas veces, 3-Algunas veces, 4-Frecuentemente, 5-Siempre) Existencia valores NA
	BP1_12	Se ha sentido querida	Cadena	6 niveles (0-NP(cuestionario indirecto y menor), 1-Nunca, 2-Muy pocas veces, 3-Algunas veces, 4-Frecuentemente, 5-Siempre) Existencia valores NA
	BP1_14	Se ha sentido alegre	Cadena	6 niveles (0-NP(cuestionario indirecto y menor), 1-Nunca, 2-Muy pocas veces, 3-Algunas veces, 4-Frecuentemente, 5-Siempre) Existencia valores NA

Ilustración 39- Tabla bienestar psicológico ESCA 2016

	Nombre atributo fichero	Descripción	Tipo de dato	Nivel de información
Consumo de medicamentos	MEDICAM_AD	Han tomado algún medicamento en los dos últimos días adultos	Cadena	4 niveles (0-NP(Menores), 1-Sí, 2-No, 9-NS/NC) Existencia de NA
	NMEDICAM_AD	Número de medicamentos tomados en los dos últimos días	Número entero	Metrica entero, valores 98-NP(Menores), 99-NS/NC, existencia de NA
	MEDICAM_NENS	Han tomado algún medicamento en los dos últimos días	Cadena	4 niveles (0-NP(Adultos), 1-Sí, 2-No, 9-NS/NC) Existencia de NA
	NMEDICAM_NENS	Número de medicamentos tomados en los dos últimos días	Cadena	Metrica entero, valores 98-NP(Adultos), 99-NS/NC, existencia de NA

Ilustración 40- Tabla consumo de medicamentos ESCA 2016

	Nombre atributo fichero	Descripción	Tipo de dato	Nivel de información
Variables generales	NUM	Número de la encuesta	Número entero	Métrica
	ANY	Sexo de la persona entrevistada	Cadena	2 niveles (1-Hombre, 2-Mujer)
	EDAT_2	Edad agrupada de la persona entrevistada	Cadena	5 niveles(1-0 a 14 años, 2-15 a 44 años, 3-45 a 64 los, 4-65 a 74 años, 5- 75 años y más)
	AV18_R	Nivel máximo de estudios finalizados	Cadena	6 niveles(0-NP (Menores), 1- Primarios o sin estudios, 2- Secundarios, 3- Universitarios, 4-Otra posibilidad, 9-NC
	CLASS_PR_2011_3CAT	Clase social de la persona entrevistada en 3 categorías	Cadena	5 niveles (1- Directores, gerentes y profesionales universitarios, 2- Ocupaciones intermedias y trabajadores por cuenta propia, 3- Trabajadores manuales, 88-Np (PR no ha trabajado nunca), 99-NC)
	V1	Sexo de la persona entrevistada	Cadena	2 niveles (1-Hombre, 2-Mujer)
	RS_N	Región sanitaria	Cadena	7 niveles(61-LLeida, 62- Camp de Tarragona, 63-Terres del ebro, 64-Girona, 67-Catalunya Central, 71- Alt Pirineu y Aran, 78-Barcelona)
	V37_R	Estado de salud autopercibida	Cadena	2 niveles (1-Buena, 2-Mala, 9-NC)

Ilustración 41- Tabla variables generales ESCA 2016

	Nombre atributo fichero	Descripción	Tipo de dato	Nivel de información
Morbilidad crónica y accidentes ⊞	V41	Trastorno crónico autopercibida	Cadena	3 niveles (1- Sí, 2-No, 9-NS/NC)
	TRAST_ADU	Padecer algún trastorno crónico adultos	Cadena	4 niveles (0-NP (Menores), 1- Sí, 2- No, 9-NS/NC) Existencia de NA
	TRAST_MENORS	Padecer algún trastorno crónico menores	Cadena	4 niveles (0-NP (Adultos), 1- Sí, 2- No, 9-NS/NC) Existencia de NA
	ACCID	Tipo de accidente	Cadena	3 niveles (1- Sí, 2- No, 9-NS/NC) Existencia de NA
	V45	Localización último accidente	Cadena	11 niveles (1- Dentro de casa, 2- Trabajo, 3-Desplazándose al trabajo, 4- Escuela, 5- Via pública (yéndose al trabajo o escuela), 6-Vía Pública (fuera del trabajo o estudio), 7- Centro deportivo, 8 Otro lugar en tiempo libre, 9- Otro lugar, 88 NP(na ha tenido ningún accidente el último año), 99- NS/NC) Existencia de NA

Ilustración 42- Tabla morbilidad crónica y accidentes ESCA 2016

5.3-Diseño conceptual

5.3.1- Tabla de hechos

Llamamos evento o **Hecho** a una operación que se realiza en el negocio en un tiempo determinado. Los hechos son objeto de análisis para la toma de decisiones.

En nuestro proyecto, la tabla de hechos corresponde a las respuestas de la Encuesta de Salud de Catalunya por las muestras encuestadas.

Por lo tanto, el propio registro es el hecho y la medida es el conteo de registros.

La tabla de hecho tiene los siguientes atributos y métricas:

- Atributos:
- NUM: Número de encuesta

- Métricas:
  - count_Reg: Conteo de registros en función del sexo, edad, estudios, SAP, clase social, provincia y año.

Tabla de hecho	Descripción
h_Encuesta	Recoge las respuestas de la Encuesta de Salud de Catalunya

Ilustración 43- Tabla de hecho de las encuestas de Salud de Catalunya

5.3.2- Dimensiones

Una **Dimensión** es una característica de un hecho que permite su análisis posterior, en el proceso de toma de decisiones. El atributo es la información asociada a una dimensión y que tiene como objetivo enriquecerla.

Por lo tanto, las preguntas se dividen en diferentes grupos según su pertenencia, cobertura sanitaria, estilo de vida… Las preguntas dependerán del tipo de encuestado que se diferenciará por un número ID y estarán relacionadas con las preguntas y el tipo de preguntas, y las métricas o conteo de registros se realizará en función de los puntos de vista de las variables generales que estas, a su vez, estarán relacionadas con las respuestas del encuestado.

Dimensiones	Atributos
d_f_carac_sociode	Nacionalidad, probabilidad de perder el trabajo, satisfacción condiciones de trabajo, personas discapacitadas en casa, como llega a fin de mes, situación laboral.
d_f_cobert_sani	Cobertura sanitaria, modalidad servicios sanitarios, grado de satisfacción.
d_f_esta.salud	SAP, IMC adultos, IMC menores, sueño, problemas caminar, problemas bañarse y vestirse, problemas actividades cotidianos
d_f_morb_acci	Trastorno crónico autopercibida, padecer algún trastorno adultos, padecer algún trastorno menores, tipo de accidentes, localización accidente.
d_f_cons_med_rest	Toma de medicamentos adultos, número de medicamentos tomados adultos, Toma de medicamentos niños, número de medicamentos tomados niños.
d_f_salud_mental	Índice de salud mental
d_f_limit_discp	Limitación para hacer actividades cotidianas, antigüedad limitación, discapacidad
d_f_pract_prev	Vacuna gripe, medir tensión arterial, revisión colesterol, mamografía, citología, prueba de sangre heces, prueba colonoscopia
d_f_factores_psico	Ansioso/a o deprimido/a, optimista respecto el futuro, sentirse relajado, sentirse bien consigo mismo, capaz de tomar sus propias decisiones, sentirse querido, sentirse alegre
d_f_estilo_vida	Bebidas azucaradas, copas de vino, consumir carne blanca o roja, actividad física , medio de trasporte habitual, consumo tabaco, indice consumo alcohol.
d_f_Vist_Med_urg	Visitas médicas, visitas medicas especialistas, visitas medicina general, atención sanitaria no atendida, motivo a la atención sanitaria no atendida, servicio de urgencias, lugar de visita a urgencias
d_var_general	Sexo, región sanitaria, estudios, clase, estado de salud autopercibida, año.

Ilustración 44- Tabla de dimensiones de la Encuesta de salud de Catalunya

5.4-Diseño físico

En el modelo físico podemos ver las relaciones entre las tablas. Estas se relacionan a través del identificador único ID = NUM, que es común para cada una de ellas. Existen una serie de variables comunes para cada una de las tablas: el año, la edad, la clase social, la región sanitaria, los estudios y la clasificación de la salud autopercibida, ya que estas son las variables generales, aquellas que describen las características de los encuestados.

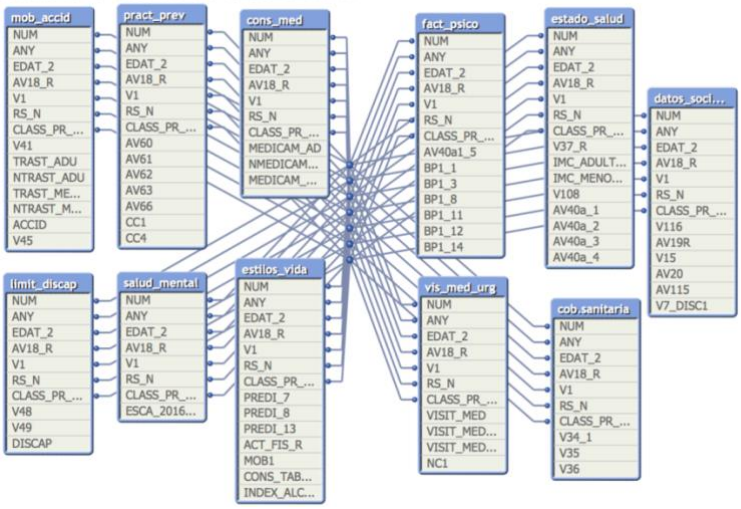


Ilustración 45- Variables comunes en todas las dimensiones

Nuestro modelo trata de dar respuesta a las preguntas de la Encuesta de Salud de Catalunya, podemos dividir las preguntas en 4 tipos diferentes.

- Preguntas sociodemográficas - Aquellas que describen características de los encuestados, a esta categoría se uniría la tabla datos sociodemográficos y la tabla de variables generales.

- Preguntas si/no/quizás - A esta categoría se unirían todas las variables que siguen ese tipo de estructura, se incluirían las siguientes variables:

    - Cobertura sanitària
    - Carn blanca en lloc de vermella
    - Limit activ.quotidianes
    - Discapacitat
    - Trastorn autopercebit
    - Trastorns persones adultes
    - Trastorn menors
    - Accident
    - Vacuna grip
    - Mesurar tensió arterial
    - Rev. colesterol
    - Mamografia
    - Citologia
    - Prova caques
    - Colonoscòpia
    - Visita metge
    - Visita especialista
    - Visita pediatra
    - No atenció
    - Servei urgències
    - Hospitalitzat
    - Discapacitats a casa
    - Medicaments persones adultes
    - Medicaments per nens

- Preguntas con varias niveles u opciones diferentes - En esta categoría se incluirían:

    - Modalitat de servei
    - Tabac
    - Transport utilitzat
    - Antiguitat límit
    - Localització accident
    - Motiu no atenció
    - LLoc visites d'urgències

- Preguntas de escala Likert- En esta categoría se incluirían:
    - Grau de satisfacció
    - Activitat física
    - Index_alcohol

- Deprimit/ansiós
- Optimista
- Relaxat
- Bé amb si mateix
- Capaç de prendre decisions
- Estimat
- Alegre
- INDEX_GOLD
- Probabilitat de perdre la feina
- Satisfacció condicions laborals

- Preguntas con respuesta numérica- En esta categoría se incluirían:
  - N_begudes alcohòliques
  - N_copes vi
  - N_trastorns persones adultes
  - N_trastorns menors
  - N_medicaments adults
  - N_medicaments nens

En primer lugar, debemos realizar un preprocesado de los datos para aumentar la calidad de estos y una transformación en las variables a fin de adecuar nuestro dataset, el preprocesado incluirá:

- Unir las anteriores tablas a partir de merge y la variable NUM que es el identificador clave.
- Cambiar los niveles de las variables, por ejemplo la variable limitación de actividades cotidianas de 4 niveles(1- Sí, gravemente limitado/a, 2-Sí, limitado/a no gravemente, 3-No limitado/a, 9- NS/NC) cambiaremos por 3 niveles (Sí, No, NS/NC).
- Modificar el título de la variable.
- Cambiar el orden de las variables a fin de agrupar las variables en función de las cinco categorías de pregunta a la que pertenezcan.
- Crear otro dataset con las variables numéricas.

A continuación, se muestra el dataset tabla.etiquetas.csv que incorpora la unión de las diferentes tablas indicadas unidas a partir de la variable NUM que hacía referencia al identificador de la persona encuestada. Este dataset tendrá 61 variables y 4818 observaciones.

ESCA.ord

Filter

	ID_ECSA	Any	SAP	Classe.social	Provincia	Sexe	Estudis	Edat	Cobertura.sànitaria	Carn.blanca.en.lloc.de.vermella	Limit.activ..quotidianes	Discapacitat
1	29330	2016-01-01	Mala salut	I. Directors i gerents i professionals universitaris	Lleida	Home	Universitaris	15 a 44 anys	Sí	No	Sí	No
2	29331	2016-01-01	Bona salut	III. Treballadors manuals	Lleida	Home	Universitaris	15 a 44 anys	Sí	Sí	No	No
3	29332	2016-01-01	Bona salut	III. Treballadors manuals	Lleida	Dona	NP (menors)	0 a 14 anys	Sí	NA	No	No
4	29333	2016-01-01	Bona salut	III. Treballadors manuals	Lleida	Dona	Secundaris	15 a 44 anys	Sí	Sí	No	No
5	29334	2016-01-01	Bona salut	III. Treballadors manuals	Lleida	Dona	Universitaris	45 a 64 anys	Sí	Sí	No	No
6	29335	2016-01-01	Bona salut	III. Treballadors manuals	Lleida	Home	NP (menors)	0 a 14 anys	Sí	NA	No	No
7	29336	2016-01-01	Mala salut	III. Treballadors manuals	Lleida	Home	Secundaris	15 a 44 anys	Sí	Sí	No	No
8	29337	2016-01-01	Bona salut	I. Directors i gerents i professionals universitaris	Lleida	Home	Secundaris	45 a 64 anys	Sí	No	No	No
9	29338	2016-01-01	Bona salut	II. Ocupacions intermèdies i treballadors per compte ...	Lleida	Home	Secundaris	45 a 64 anys	Sí	No	No	No
10	29339	2016-01-01	Mala salut	III. Treballadors manuals	Lleida	Home	Primaris o sense estudis	75 anys i més	Sí	Sí	Sí	Sí

Showing 1 to 11 of 4,818 entries

Ilustración 46- Tabla.etiquetas.csv

El segundo paso es crear otro dataset a partir del anterior con los valores de las variables en clase numérica. A este dataset le llamaremos tablas.numerica.csv y tendrá la misma dimensión que el anterior.

	ID_ECSA	Any	SAP	Classe.social	Província	Sexe	Estudis	Edat	Cobertura.sànitaria
1	29330	2016	Mala salut	I. Directors i gerents i professionals universitaris	Lleida	Home	Universitaris	15 a 44 anys	3
2	29331	2016	Bona salut	III. Treballadors manuals	Lleida	Home	Universitaris	15 a 44 anys	3
3	29332	2016	Bona salut	III. Treballadors manuals	Lleida	Dona	NP (menors)	0 a 14 anys	3
4	29333	2016	Bona salut	III. Treballadors manuals	Lleida	Dona	Secundaris	15 a 44 anys	3
5	29334	2016	Bona salut	III. Treballadors manuals	Lleida	Dona	Universitaris	45 a 64 anys	3
6	29335	2016	Bona salut	III. Treballadors manuals	Lleida	Home	NP (menors)	0 a 14 anys	3
7	29336	2016	Mala salut	III. Treballadors manuals	Lleida	Home	Secundaris	15 a 44 anys	3
8	29337	2016	Bona salut	I. Directors i gerents i professionals universitaris	Lleida	Home	Secundaris	45 a 64 anys	3
9	29338	2016	Bona salut	II. Ocupacions intermèdies i treballadors per compte	Lleida	Home	Secundaris	45 a 64 anys	3

Ilustración 47. Tablas. numéricas.csv

Luego creamos otro dataset con las variables generales, a este dataset le llamaremos variables.generales.csv, y tendrá una dimensión de 8 variables y 4818 observaciones.

	id_Ecsa	Any	SAP	Classe.social	Província	Sexe	Estudis	Edat
1	29330	2016	Mala salut	I. Directors i gerents i professionals universitaris	Lérida	Home	Universitaris	15 a 44 anys
2	29331	2016	Bona salut	III. Treballadors manuals	Lérida	Home	Universitaris	15 a 44 anys
3	29332	2016	Bona salut	III. Treballadors manuals	Lérida	Dona	NP (menors)	0 a 14 anys
4	29333	2016	Bona salut	III. Treballadors manuals	Lérida	Dona	Secundaris	15 a 44 anys
5	29334	2016	Bona salut	III. Treballadors manuals	Lérida	Dona	Universitaris	45 a 64 anys
6	29335	2016	Bona salut	III. Treballadors manuals	Lérida	Home	NP (menors)	0 a 14 anys
7	29336	2016	Mala salut	III. Treballadors manuals	Lérida	Home	Secundaris	15 a 44 anys
8	29337	2016	Bona salut	I. Directors i gerents i professionals universitaris	Lérida	Home	Secundaris	45 a 64 anys
9	29338	2016	Bona salut	II. Ocupacions intermèdies i treballadors per compte propi	Lérida	Home	Secundaris	45 a 64 anys
10	29339	2016	Mala salut	III. Treballadors manuals	Lérida	Home	Primaris o sense estudis	75 anys i més

Ilustración 48- tabla variables.generales.csv

El cuarto dataset, se denomina preguntas.csv. En este se crean 3 columnas que hacen referencia a las preguntas de la encuesta, este estará compuesto de 3 variables (identificador de las preguntas, las etiquetas de las preguntas, y la categoría a la que pertenecen las preguntas). Este dataset como ya se ha indicado tiene 3 variables y 53 filas.

	Preguntas.etiquetas	id.preguntas	Preguntas.grupo
1	Cobertura.sànitaria	1	Cobert.sànit
2	Carn.blanca.en.lloc.de.vermella	2	Estil.vida
3	Limit.activ..quotidianes	3	Limit.discap.autonomia
4	Discapacitat	4	Limit.discap.autonomia
5	Transtorn.autopercibit	5	morb.cronica.accidents
6	Trastorn.persones.adultes	6	morb.cronica.accidents
7	trastorn.persones.menors	7	morb.cronica.accidents
8	Accident	8	morb.cronica.accidents
9	Vacuna.grip	9	prevenció
10	Mesurar.tensió.arterial	10	prevenció

Ilustración 49-Tabla preguntas.csv

Y por último se crea el dataset respuestas. Este dataset se denominará respuestas.csv y consta de cuatro variables y 255.354 observaciones. Este dataset registra el identificador del usuario, las etiquetas de las preguntas, las respuestas por cada una de esas preguntas y los valores son las respuestas a las preguntas con los datos trasformados a numéricos.

	ID_ECSA	Preguntas	Respuestas	valores
1	27367	Accident	No	2
2	27367	Activitat.física	NA	NA
3	27367	Alegre	NA	NA
4	27367	Antiguitat.limit	NP (no es troba limitat)	6
5	27367	Bé.amb.si.mateix	NA	NA
6	27367	Capaç.prendre.decisions	NA	NA
7	27367	Carn.blanca.en.lloc.de.vermella	NA	NA
8	27367	Citologia	NP (qüestionari indirecte + menors)	2
9	27367	Cobertura.sànitaria	Sí	3
10	27367	Colonoscòpia	NA	NA

Ilustración 50-Tabla Respuestas.csv

El código R para la preparación de los datos para *tableau* está en el anexo 8.2.

Ahora pasaremos esos datos a formato excel, ya hemos creado nuestros cinco conjuntos de datos adaptados a las necesidades de visualización para el sistema de visualización TABLEAU, estos son:

- tabla.etiquetas.xlsx
- tabla.numérica.xlsx
- variables.generales.xlsx
- preguntas.xlsx
- respuestas.xlsx

Creamos una única base de datos, y unimos las hojas anteriores en un único archivo, para poder manejarlo fácilmente. Este se llamará **datos_esca_tableau.**

El siguiente paso será conectarnos a *Tableau*.

6. Creación de un cuadro de mandos para el análisis de encuestas de la Encuesta de Salud de Catalunya. (ESCA)

6.1-Conectar la base de datos con TABLEAU

Hemos añadido, en primer lugar, la conexión al dataset variables.generales, y hemos unido a este la base de datos respuestas a través del identificador común número de persona encuestada (ide_esca), y a este segundo dataset lo hemos unido con el dataset preguntas a través de las etiquetas de las preguntas.

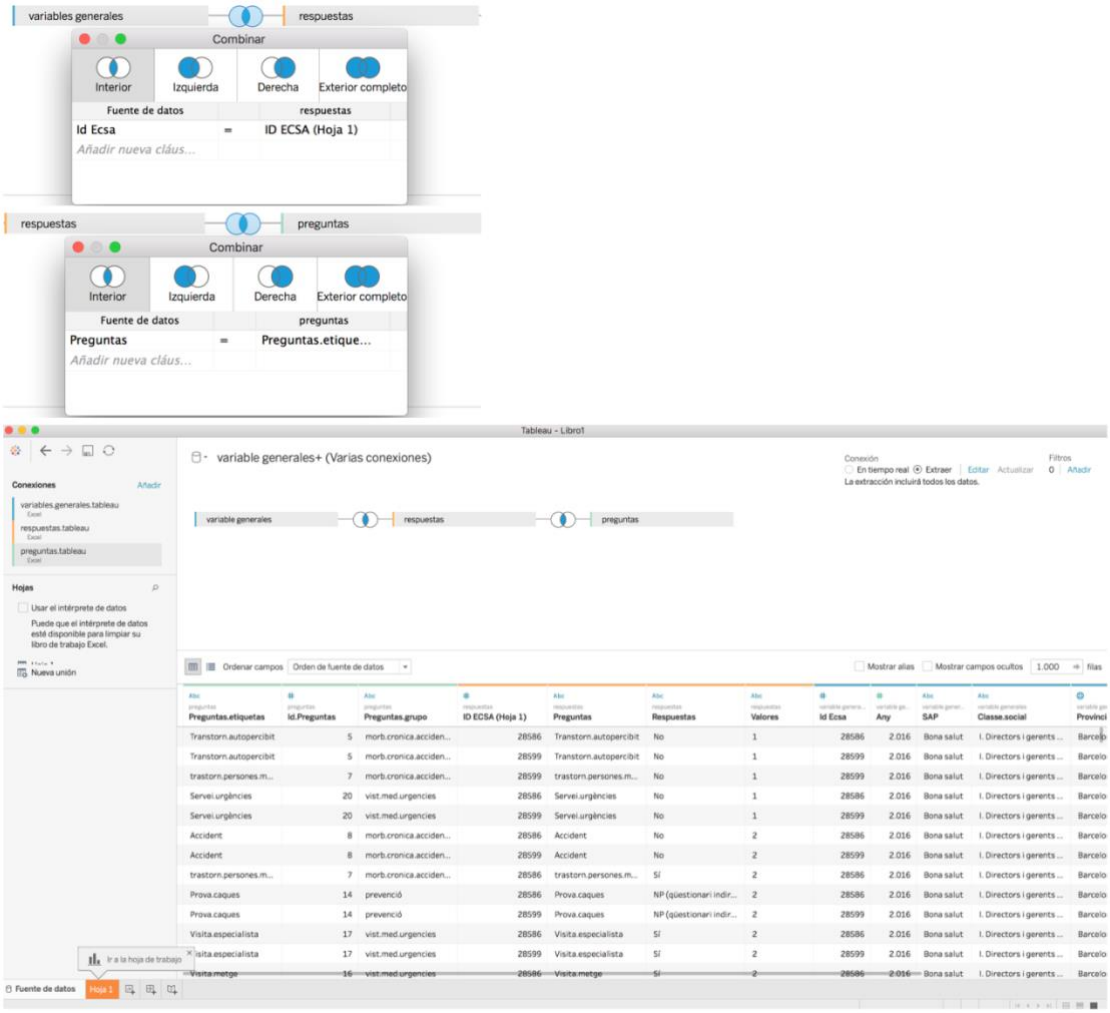


Ilustración 51- Conexión y relación de las tablas

Presionamos el botón extraer. La extracción incluirá todos los datos de las fuentes de datos conectadas.

Guardaremos esta fuente de datos en el repositorio de *Tableau,* dentro de la carpeta fuentes de datos y le denominaremos esca.tfm.tde

Ahora hay que renombrar las variables y especificar la clase de datos que son, por ejemplo en cuanto a las provincias, cabría indicar el rol geográfico de

provincia/región. Hay variables a las que tendremos que indicar que son datos numéricos enteros.

6.2 Crear visualizaciones

En primer lugar, se crearán las visualizaciones de forma independiente para cada una de las vistas. Las vistas de nuestra herramienta de *reporting* interactiva son 5:

6.2.1-Gráfico de barras salud autopercibida

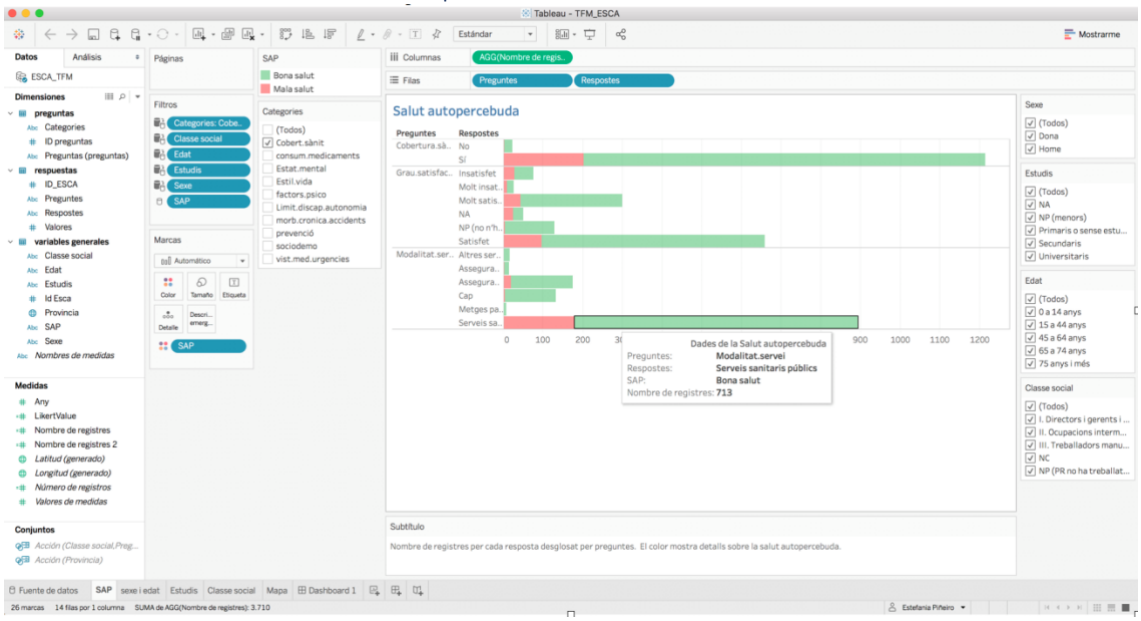


Ilustración 52- Visualización gráfico de barras para Dashboard ESCA

El gráfico muestra el número de registros por cada respuesta de las preguntas de la Encuesta de Salud de Catalunya. Esta gráfica de barras permite filtrar por las categorías de las preguntas en función de si pertenecen al grupo de cobertura sanitaria, consumo de medicamentos, estilo de vida, estado mental…

También tenemos la opción de filtrar según las variables sociodemográficas de las personas encuestadas. Para el color de las gráficas se opta por colores semáforo que muestren la salud autopercibida de los registros en función de las preguntas vs respuestas y de la salud autopercibida.

6.2.2-Gráfico sexo y edad

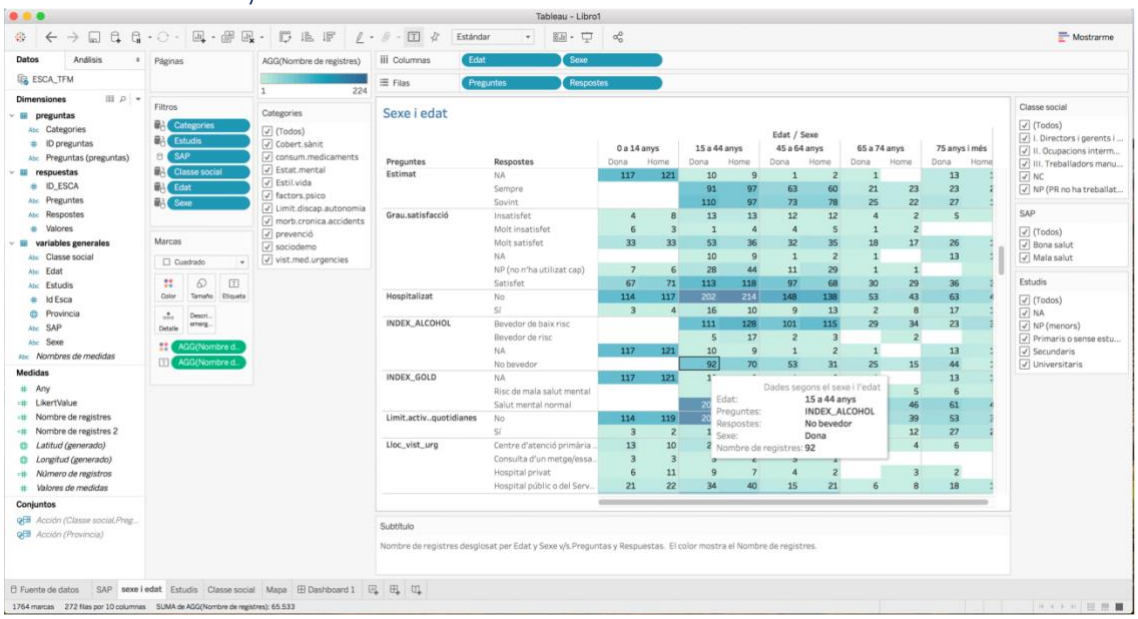


Ilustración 53- Visualización gráfica sexo y edad para el dashboard ESCA

Este gráfico muestra el número de registros desglosados por edad y sexo v/s. Preguntas y respuestas. El color muestra el número de registros.
Como en el caso anterior podemos filtrar por categorías de las preguntas y por las características sociodemográficas de las personas.

6.2.3-Gráfico estudios

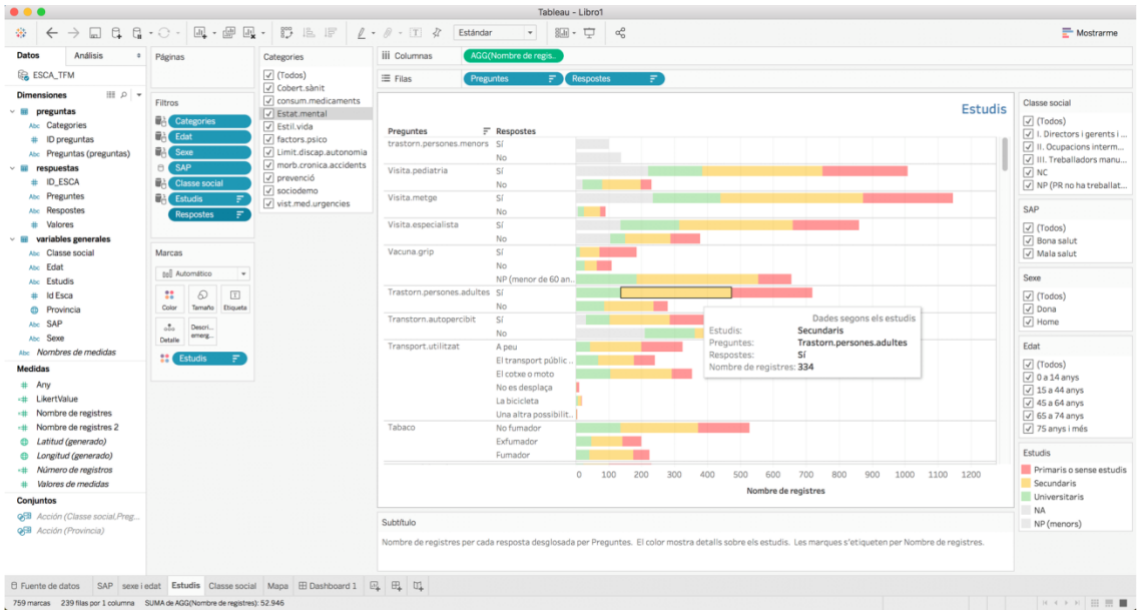
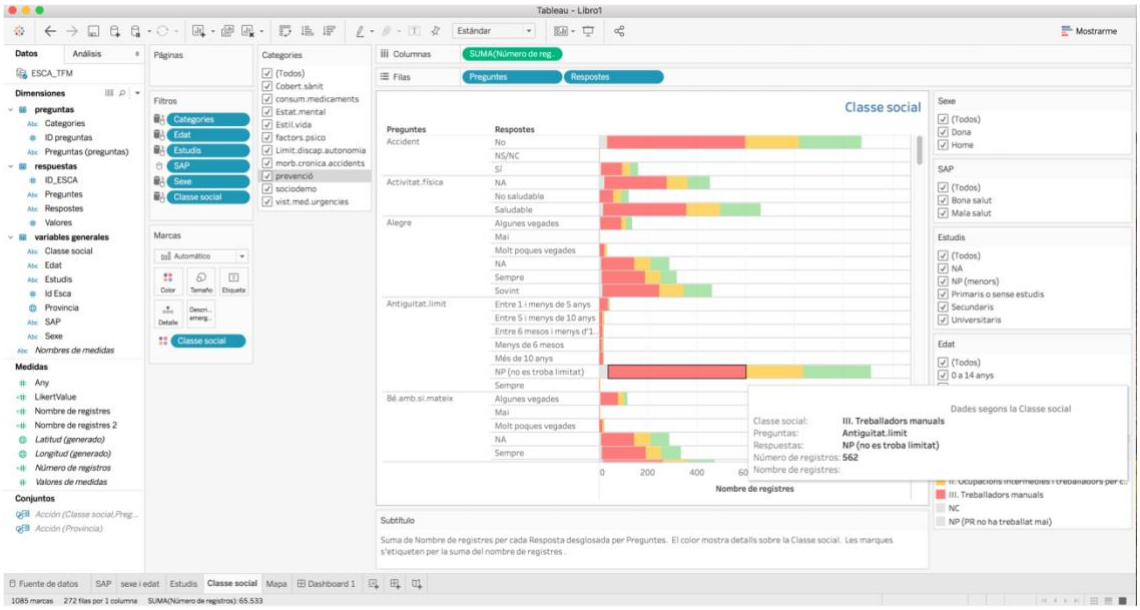


Ilustración 54- Visualización de los estudios para el dashboard de la ESCA

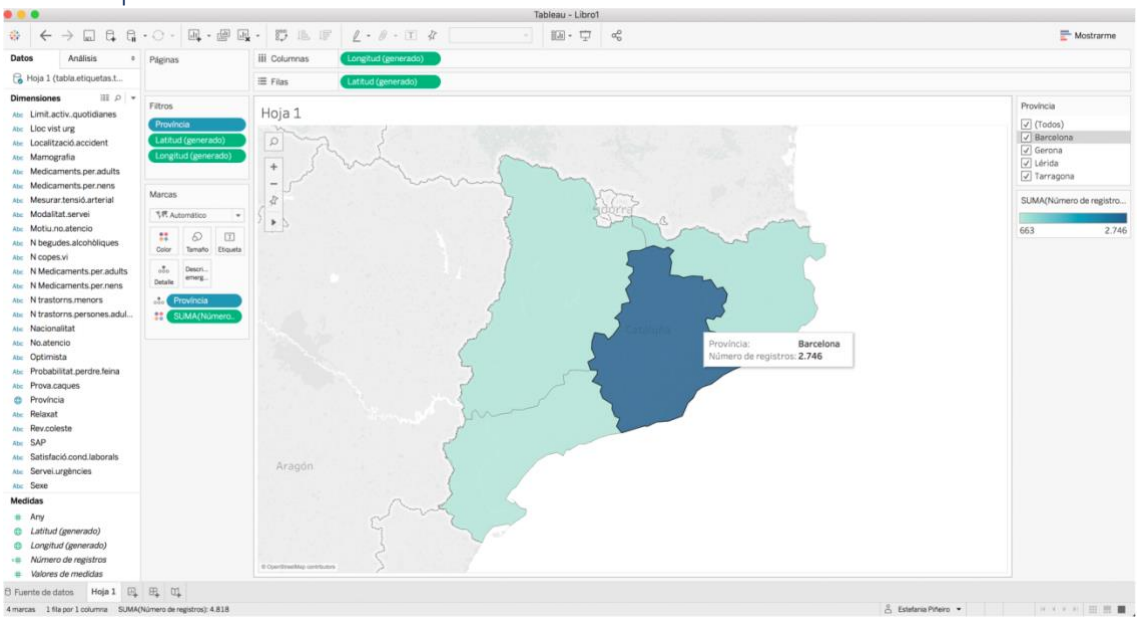
Este gráfico muestra el número de registros por cada respuesta desglosada por las preguntas. El color muestra detalles sobre los estudios. Las marcas se etiquetan por el número de registros.
Como en los casos anteriores podemos filtrar por categorías de las preguntas y por las características sociodemográficas de los encuestados.

6.2.4-Gráfica clase social



Este gráfico muestra el número de registros por cada respuesta desglosada por las preguntas. El color muestra detalles sobre la clase social. Las marcas se etiquetan por el número de registros.
Como en los casos anteriores, podemos filtrar por categorías de las preguntas y por las características sociodemográficas de los encuestados.

6.2.5-Mapa



Mapa basado en longitud y latitud generadas. El color muestra el número de registros, a mayor tonalidad mayor es el número de registros. Los detalles se muestran por Provincia.

Una vez terminadas las visualizaciones según las vistas, ya podemos crear nuestro *Dashboard* a partir de estos.

6.2- Elaboración y publicación en internet del cuadro de mando o *dashboard* en *Tableau*

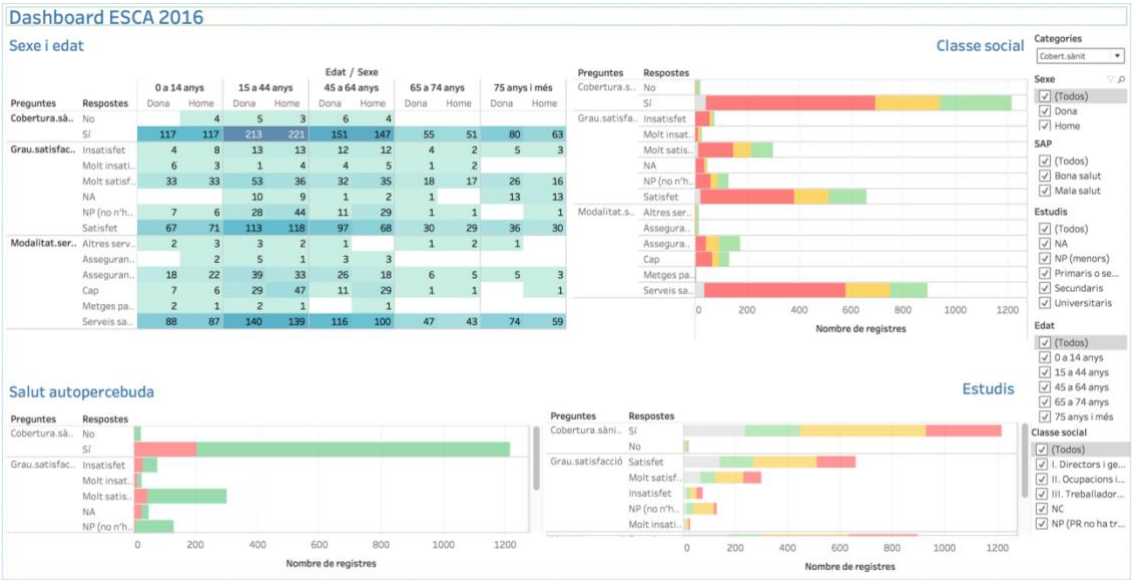


Ilustración 55- Cuadro de mandos interactivo ESCA

El *Dashboard* final consta de 4 vistas, una con cada una de las visualizaciones anteriores.
El *dashboard* permite filtrar por categorías de preguntas y por las características sociodemográficas de las personas.

Tableau permite, una vez creado el libro o *dashboard*, subirlo a internet a la página de *Tableau Public*, a través del usuario y *password,* que previamente hemos dado para crear la cuenta y así poder compartirlo de forma fácil. (Debido a la sensibilidad de los datos y a la ausencia de autorización de la ESCA este paso no se ha concluido).

También disponemos de otras opciones que nos permiten incorporar una descripción, realizar actualizaciones, ver nuestras bases de datos, las vistas del libro de trabajo e incluso, conocer cuántas personas han visto el *dashboard* o las visualizaciones.

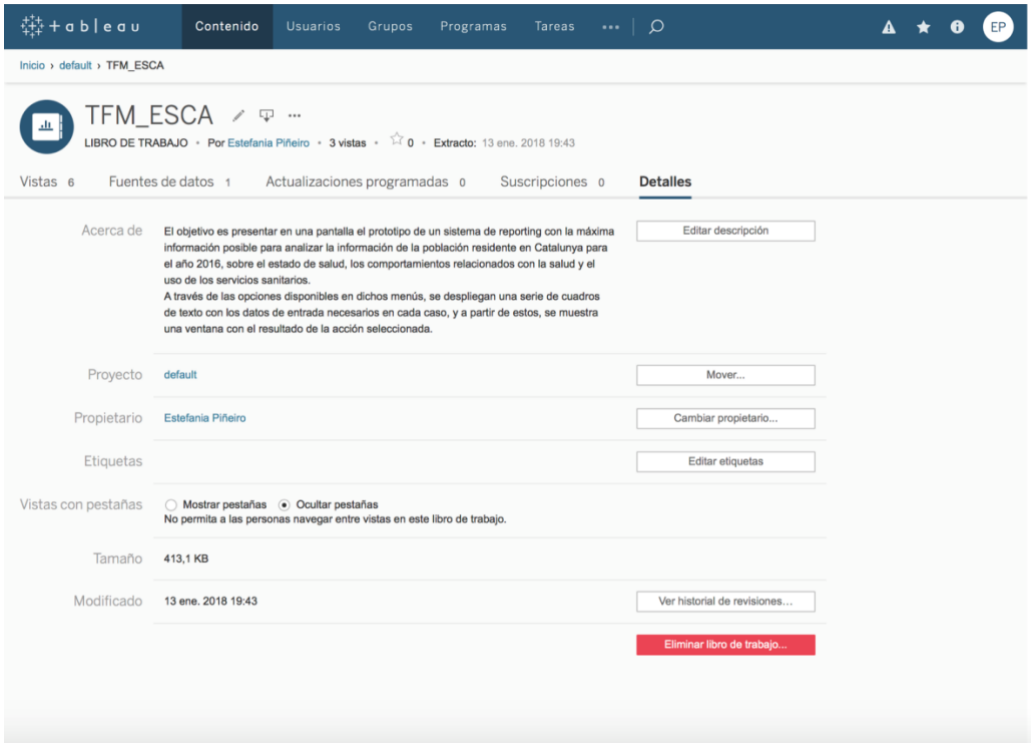


Ilustración 56- Perfil Tableau

7-CONCLUSIONES

En las fases desarrolladas anteriormente, se han obtenido tres productos diferentes. En primer lugar, se han obtenido unos resultados de la tercera parte del trabajo en el que se realiza un análisis de datos de la comunidad de Madrid, a través de algoritmos de clasificación. En segundo lugar, hemos obtenido una base de datos con los datos relacionados para conectar con TABLEAU. Y en tercer lugar, se ha creado una herramienta de *reporting* interactiva que analiza las respuestas de la Encuesta de Salud de Catalunya.

A continuación, se desarrollan las conclusiones extraídas de cada producto:

7.1-Análisis de datos de la base de datos de la Encuesta de la comunidad de Madrid.

Nuestro dataset tiene la maldición de la dimensionalidad, con 802 observaciones y 115 variables es imposible construir un buen modelo de clasificación, debido a que se reducen las posibilidades de encontrar en el dataset todos los casos posibles que ilustren la combinación de las distintas dimensiones. Por tanto, no dispondremos de ejemplos que permitan entrenar nuestro algoritmo de clasificación.

Para poder construir un buen modelo de clasificación para extraer conocimiento

válido y comprensible de los predictores de la salud autopercibida. Se ha ajustado nuestro conjunto de datos, transformando los datos brutos para que nos aporten una mayor información significativa y utilizable que disminuya el desorden del ambiente (disminución de la entropía).

La técnica que ha optimizado nuestro algoritmo de clasificación ha sido la técnica poda con un 0.37 puntos y en mayor medida la técnica boosting incrementando la capacidad predictiva en 2,61 puntos.

En el caso de extracción de características por el método *filter*, la técnica del análisis de las correlaciones interpreta que las variables funcionan por si solas cuando en algunos casos lo hacen de forma combinada. La técnica *randomforest,* para determina la importancia de las variables en el árbol de decisión, da mayor relevancia a aquellas en las que tienen mayor número de niveles y favorece los grupos más pequeños cuando las variables están correlacionadas, añadiendo que este modelo hace compleja la interpretación del modelo convirtiéndose en una caja negra.

En el caso de extracción de características por técnica *Wrapper*, el análisis de componentes principales no tiene sentido, ya que no existen altas correlaciones entre las variables. Lo que indica que no existe información redundante y, por lo tanto, un número menor de variables no será capaz de explicar gran parte de la variabilidad total. Lo mismo ocurre para la técnica SVD, nuestro modelo ha empeorado con la incorporación de la nueva variable ya que añade más información redundante al conjunto datos. Esta nueva variable aporta una menor información significativa y utilizable y hace que aumente la entropía.

7.2-Base de datos preparados para conectar con TABLEAU.

En esta fase hemos obtenido como resultado una base de datos. Esta base de datos consta de un solo fichero .xlsx con 5 pestañas, estas pestañas nos permiten relacionar los datos para conectar con TABLEAU. Es una base de datos sencilla y para su construcción no se ha necesitado nada más que R y Excel. A través de la carga de este fichero y la unión de los datos de las variables generales con el de las respuestas por el ID, que es el número de encuestado, y este último, por las preguntas, relacionándolo a partir de la pregunta, conseguiremos relacionar la base de datos.

El resultado nos permite obtener las respuestas a cualquiera de las preguntas, en función de las variables generales de los encuestados.

7.3-Herramienta de *reporting* interactiva.

Al conectar la base de datos que hemos creado con TABLEAU esta nos permite realizar visualizaciones para dar respuesta a las preguntas de la ESCA. El cuadro de mandos se divide en 4 pestañas, aunque podríamos añadir una quinta con las regiones (Esta posibilidad puede ser objeto de desarrollo, con más detalle, en trabajos futuros). Esas visualizaciones se dividen en sexo y edad, estudios, clase social y SAP.

El usuario, de forma interactiva, puede seleccionar la categoría de las preguntas que estaría interesado en conocer, y filtrar, en función de las variables sociodemográficas que considere. El resultado final es un cuadro de mandos con 4 vistas, en función de los filtros seleccionados. (En la presentación se añade un vídeo del testeo de la herramienta de *reporting* interactiva. No ha sido posible su publicación del dashboard en internet debido a la sensibilidad de los datos y a la falta de autorización de la ESCA).

7.4-Una vista global.

Una herramienta de *reporting interactiva* ha de adaptarse a las necesidades y requerimientos del usuario. Cumplir con estas exigencias no convierte el sistema de *reporting* en un instrumento perfecto o idóneo. Antes que nada, es necesario realizar un preprocesado de los datos, analizar su calidad y conocerlo a través de algoritmos de clasificación o predicción. Antes de elaborar un sistema de *reporting* debemos conocer qué es lo que cuentan los datos para poder saber qué explicar y cómo hacerlo.

Para el caso de la ESCA, si se hubiera realizado el mismo análisis que con los datos de la comunidad de Madrid, la dimensionalidad de 1377 variables se habría reducido a unas menos, con lo que se reduciría la entropía. Hubiéramos sido más eficientes al seleccionar las variables o preguntas que forman parte de los diferentes grupos de preguntas, y hubiéramos tenido una herramienta de *reporting* interactiva más potente.

8- Trabajos futuros

Ahora tocaría perfeccionar, en primer lugar, como se expuso anteriormente; se debería realizar un análisis de datos para el fichero de la ESCA_2016.sav a través del algoritmo generador de árboles de decisión, pero en este caso se incluiría el 10-fold cross validation en el árbol de decisión, esta técnica produce modelos correctamente válidos, poco sujetos al overfitting y robustos frente a pequeñas variaciones en las variables explicativas.

También se añadiría un nuevo método, las *máquinas de soporte vectorial*, que requieren un pequeño número de observaciones para su entrenamiento y capaces de resistir a un gran número de dimensiones.

Una vez depurados los datos, reducida su dimensionalidad y obtenido los predictores más influyentes, podría crearse una base de datos con todas las variables y todas las observaciones más relevantes. Para este caso, para optimizar y mejorar la eficiencia se usaría una base de datos relacional con MySQL que nos proporcionaría un entorno de herramientas integrado para el diseño y modelado de bases de datos, desarrollo de SQL, y administración de bases de datos.

Y por último y muy importante, la base de datos ha de tener más observaciones, es por ello que ha de incluir más períodos temporales para poder analizar la evolución.

El resultado de las visualizaciones de nuestro cuadro de mando interactivo habría sido diferente si hubiera contado con más series temporales.

Nuestra apuesta sería por representaciones visuales en series temporales.

9- Anexos

9.1 - Anexo: Código árbol de decisión

9.1.1-Elaboración primer modelo de clasificación.

```
library(tree)
# Selección muestra entrenamiento
ECM.tree = tree(xtrain$B1~.,xtrain[,-11],subset=train)
summary(ECM.tree)

plot(ECM.tree);text(ECM.tree,pretty=0)

ECM.tree

library(gmodels)
tree_pred <- predict(ECM.tree, xtest)
CrossTable(xtest$B1,tree_pred,prop.chisq=FALSE,prop.c=FALSE,prop.r=FALSE,dnn =
c('actual B1','predicted B1'))

#El mismo ejercicio realizado con la función C5.0 del package C50.
library(C50)
model1B <- C50::C5.0(xtrain[,-11],as.factor(xtrain$B1))
model1B
summary(model1B)

model1B.rules <- C50::C5.0(xtrain[,-11],as.factor(xtrain$B1), rules= TRUE)
summary(model1B.rules)

#Importancia de las variables
importancia.v<- varImp(model1B, scale =FALSE)
importancia.v
plot(importancia.v, top=10)

#Dibuja árbol de decisión
options(repr.plot.width=10, repr.plot.height=5)
plot(model1B)

#Evaluación primer modelo de clasificación
library(gmodels)
B1_pred <- predict(model1B, xtest)
CrossTable(xtest$B1,B1_pred,prop.chisq=FALSE,prop.c=FALSE,prop.r=FALSE,dnn =
c('actual B1','predicted B1'))

# Porcentaje correctamente clasificados
tabla.pred.B1<- table(B1_pred, xtest$B1)
```

```
100 * sum(diag(tabla.pred.B1)) / sum(tabla.pred.B1)
```

9.1.2- Poda del árbol de decisión

```
#creamos la poda
set.seed(123)
model1B.PODA <- C50::C5.0(xtrain[,-11],as.factor(xtrain$B1), control = C5.0Control(noGlobalPruning = TRUE))
summary(model1B.PODA)

model1B.PODA.RULES <- C50::C5.0(xtrain[,-11],as.factor(xtrain$B1),rules= TRUE, control = C5.0Control(noGlobalPruning = TRUE))
summary(model1B.PODA.RULES)

options(repr.plot.width=10, repr.plot.height=5)
plot(model1B.PODA)


#Capacidad predictiva del modelo de poda
Pred.PODA<- predict(model1B.PODA, xtest, type = "class")

# Porcentaje correctamente clasificados
tabla.pred.poda<- table(Pred.PODA, xtest$B1)
100 * sum(diag(tabla.pred.poda)) / sum(tabla.pred.poda)
```

9.1.3- Boosting

```
#Creamos el modelo con 100 repeticiones
set.seed(123)
model1B.Boosting <- C50::C5.0(xtrain[,-11],as.factor(xtrain$B1), trials=100)
model1B.Boosting
summary(model1B.Boosting)

model1B.boosting.rules <- C50::C5.0(xtrain[,-11],as.factor(xtrain$B1), trials=100, rules= TRUE)
summary(model1B.boosting.rules)

#Importancia de las variables
importancia.v.boo<- varImp(model1B.Boosting, scale =FALSE)
importancia.v.boo
plot(importancia.v.boo, top=10)

#Dibuja árbol de decisión
options(repr.plot.width=10, repr.plot.height=5)
plot(model1B.Boosting)

#Evaluación de la mejora con la técnica boosting
```

```
B1_pred.boo <- predict(model1B.Boosting, xtest)
CrossTable(xtest$B1,B1_pred.boo,prop.chisq=FALSE,prop.c=FALSE,prop.r=FALSE,dnn  =
c('actual B1','predicted B1'))

# Porcentaje correctamente clasificados
tabla.pred.B1.boo<- table(B1_pred.boo, xtest$B1)
100 * sum(diag(tabla.pred.B1.boo)) / sum(tabla.pred.B1.boo)
```

9.2.4-Mejora modelo, eliminar características redundantes a través del análisis de correlaciones

```
# Variables con alta correlación que se pueden eliminar
library(mlbench)
library(corrplot)
library(caret)


correlacionMatrix.2 <- cor(Encuesta_Salud_norm)

#Gráfica correlación
corrplot(correlacionMatrix.2 ,order = "hclust")

#Gráfica diferenciando en función de la salud autopercibida
corrplot(correlacionMatrix.2[,],                    pch='*',col=c("green",                    "red",
"orange")[unclass(Encuesta_Salud_CM2013[,11])])

## Variables con alta correlación
MaxCorrelated.2 <- findCorrelation (correlacionMatrix.2 , cutoff = 0.80)
print(MaxCorrelated.2 )


#Procedemos a aplicar el filtro y repetimos el gráfico, observando cómo desaparecen
las características y cómo quedan nuestros datos ahora:
Filter <- Encuesta_Salud_CM2013.N2[,-MaxCorrelated.2 ]
Encuesta_Salud_CM2013.N2.new <- cor(Filter)
corrplot(Encuesta_Salud_CM2013.N2.new, order = "hclust")
```

9.1.5-Mejora modelo, puntuar características por importancia: Random Forest

```
library(randomForest)
#Aplicamos el algoritmo
Modelo.rf<-randomForest(B1 ~ .,
                data=xtrain, # datos para entrenar
                ntree=100,          # cantidad de arboles
                mtry=115,           # cantidad de variables
                replace=T)         # muestras con reemplazo

#Obtenemos la importancia de cada variable en el proceso de clasificación
```

```
importancia.rf<-importance(Modelo.rf)

#Importancia de las variables en formato gráfico
varImpPlot(Modelo.rf)

#evolución del error según el número de árboles
plot(Modelo.rf, main = "Evolución error según número de árboles")
```

9.1.6-Mejorar el modelo generado aplicando SVD

```
#Utilizamos SVD para extraer características
SAP <- xtrain[,11]
VarSAP <- xtrain[,-11]

svdVarSAP <- svd(VarSAP)
svdVarSAPT <- svd(xtest[,-11])

#Matrices generadas
svdVarSAP$d
head(svdVarSAP$u)
svdVarSAP$v

#La varianza retenida por cada componente principal puede obtenerse de la siguiente manera:
valor.propio <- (svdVarSAP$d)^2
varianza <- valor.propio*100/sum(valor.propio)  # Porcentaje varianzas
acumulativa.var <- cumsum(varianza) # Acumulativas varianza
valor.propio.encuestaCM.active <- data.frame(valor.propio, varianza,acumulativa.var)
head(valor.propio.encuestaCM.active, n=50)

#Potrcentaje de la varianza acumulada explicada
plot(cumsum(svdVarSAP$d^2/sum(svdVarSAP$d^2)),type="l",col="red",xlab="Singular vector",ylab="Varianza explicada acumulada")
lines(cumsum(svdVarSAPT$d^2/sum(svdVarSAPT$d^2)),col="green")

Y <- xtrain$B1
Y<-as.factor(Y)

# Asignamos rojo muy malo y muy malo, naranga regular, verde bueno y azul muy bueno.
plot(svdVarSAP$u[,(1)],svdVarSAP$u[,(2)])
points(svdVarSAP$u[Y==-0.56614405291637   ,1],svdVarSAP$u[Y==-0.56614405291637 ,2],col="green")
points(svdVarSAP$u[Y==0.681239272465303,1],svdVarSAP$u[Y==0.681239272465303, 2],col="red")
points(svdVarSAP$u[Y==1.92862259784698,1],svdVarSAP$u[Y==1.92862259784698,2] ,col="orange")
```

```
#Añadir una nueva variable C1 a los datos usando el primer SVD que acumula el 80% de varianza
C1a42<-svdVarSAP$u[,c(1:42)]
C1a42t<-svdVarSAPT$u[,(1:42)]

Xtrainc1<- cbind(xtrain, C1a42)
xtestc1<- cbind(xtest, C1a42t)

# Construimos un nuevo árbol con las nuevas caracteríasticas
set.seed(123)
modelB2 <- C50::C5.0(Xtrainc1[,c(1:10,12:116)],as.factor(Xtrainc1$B1))
modelB2
summary(modelB2)

#Gráfica del árbol de decisión
options(repr.plot.width=10, repr.plot.height=5)
plot(modelB2)

#Evaluamos el segundo modelo
B1_pred.2 <- predict(modelB2, xtestc1)
CrossTable(xtestc1$B1,B1_pred.2,prop.chisq=FALSE,prop.c=FALSE,prop.r=FALSE,dnn  =
c('actual B1','predicted B1'))

# Porcentaje correctamente clasificados
tabla.pred.2.B1<- table(B1_pred.2, xtestc1$B1)
100 * sum(diag(tabla.pred.2.B1)) / sum(tabla.pred.2.B1)
```

9.1.7- Mejora del modelo con el Análisis componentes principales

```
# Aplicamos el algoritmo de componentes principales
Encuesta_Salud_CM2013.pca <- princomp(Encuesta_Salud_norm)

#Observamos el resultado
Encuesta_Salud_CM2013.pca

#Podemos sacar un resumen del análisis para observar algunos detalles más.
summary(Encuesta_Salud_CM2013.pca)

#El resumen nos aparecen los valores propios de cada componente principal, la proporción de varianza explicada por cada componente y la varianza explicada acumulativa.

#Esta opción nos muestra más información
summary(Encuesta_Salud_CM2013.pca,loadings=T)
#Esta opción nos da un coeficiente  para cada variable del dataset dentro de cada componente (loadings),
```

```
#Grafico de la varianza de cada componente
screeplot(Encuesta_Salud_CM2013.pca)
```

#El siguiente paso es analizar el componente obtenido para poder extraer información del mismo. Utilizaremos scores que es una matriz generada por la función princomp() y que contiene la combinación lineal del total de componentes.

```
# Revisamos scores
head(Encuesta_Salud_CM2013.pca$scores)

#Dibujamos las observaciones en función de la salud autopercibida
# Asignamos a Y si supera o no el curso
Y <- Encuesta_Salud_CM2013[,11]

# Dibujamos primero todos los scores de los componentes
Xscores=(Encuesta_Salud_CM2013.pca$scores)

# Asignamos color a los dos primeros componentes debido al elevado tiempo de poner color a las 44 componentes
plot(Xscores[,(1)],Xscores[,(2)])

points(Xscores[Y=="Muy Bueno o bueno",1],Xscores[Y=="Muy Bueno o bueno",2],col="green")
points(Xscores[Y=="Regular",1],Xscores[Y=="Regular",2],col="yellow")
points(Xscores[Y=="Malo o Muy Malo",1],Xscores[Y=="Malo o Muy Malo",2],col="red")
```

#Crearemos un biplot que muestre la relación entre las dos primeras componentes principales y la salud autopercibida, se colorea cada punto de acuerdo con la la salud autopercibida y se dibuja una linea circular para cada grupo.

```
# Previo
library(devtools)
library(ggbiplot)

salud.autopercibida <-Encuesta_Salud_CM2013[,11]

g <- ggbiplot(Encuesta_Salud_CM2013.pca, obs.scale = 1, var.scale = 1,
          groups = salud.autopercibida, ellipse = TRUE,
          circle = TRUE)
g <- g + scale_color_discrete(name = '')
g <- g + theme(legend.direction = 'horizontal',
           legend.position = 'top')
print(g)
```

9.2 - Anexo: Preparación de datos para tableau

9.2.1- Crear la tabla respuestas

```
#Creamos un nuevo dataset con las respuestas de la ESCA
library(tidyr)
library(dplyr)
ESCA.ord %>%
  gather(Preguntas, Respuestas, -ID_ECSA, -Any, -Sexe,- Classe.social,-SAP,
        -Província, -Estudis, -Edat , convert = TRUE) %>%
  arrange(ID_ECSA, Preguntas)-> ESCA.ord.2
ESCA.n %>%
  gather(Preguntas, Respuestas, -ID_ECSA, -Any, -Sexe,- Classe.social,-SAP,
        -Província, -Estudis, -Edat , convert = TRUE) %>%
  arrange(ID_ECSA, Preguntas)-> ESCA.ord.3

valores<-ESCA.ord.3$Respuestas

RESPUESTAS<- data.frame(ESCA.ord.2[,c(1, 9,10)], valores)
```

10-Bibliografía

Abu-Omar, K., Rütten, A. and Robine, J.-M. (2004) 'Self-rated health and physical activity in the European Union.', *Sozial- und Praventivmedizin*, 49(4), pp. 235–42. Available at: http://www.ncbi.nlm.nih.gov/pubmed/15357524 (Accessed: 15 January 2018).

Alcalde, I. (2015) *Visualización de la información : de los datos al conocimiento*. Edited by UOC. Barcelona. Available at: http://cataleg.uoc.edu/record=b1056640~S1%2Acat (Accessed: 15 January 2018).

Bardage, C. *et al.* (2005) 'Self-rated health among older adults: a cross-national comparison.', *European journal of ageing*, 2(2), pp. 149–158. doi: 10.1007/s10433-005-0032-7.

Brugulat, P., Medina, A., Mompart Anna, *et al.* (2013) 'Les regions sanitàries a l'ESCA Servei del Pla de Salut Direcció General de Planificació i Recerca en Salut', p. 14. Available at: http://salutweb.gencat.cat/web/.content/home/el_departament/estadistiques_sanitaries/enquestes/02_enquesta_catalunya_continua/documents/arxius/plasalut_regions_sanitaries_esca_2010-2012.pdf (Accessed: 15 January 2018).

Brugulat, P., Medina, A., Mompart, A., *et al.* (2013) 'Població ocupada i aturada: comparació d'indicadors de salut ESCA Servei del Pla de Salut Direcció General de Planificació i Recerca en Salut'. Available at: http://salutweb.gencat.cat/web/.content/home/el_departament/estadistiques_sanitaries/enquestes/01_enquesta_catalunya/documents/informe_ocupats.pdf (Accessed: 15 January 2018).

Cairo, A. (2011) *El arte funcional : infografía y visualización de información*. Alamut.

Calvo, M. *et al.* (2013) 'Automatic assessment of socioeconomic impact on cardiac rehabilitation.', *International journal of environmental research and public health*. Multidisciplinary Digital Publishing Institute (MDPI), 10(11), pp. 5266–83. doi: 10.3390/ijerph10115266.

Castells, M. (2005) *La era de la información: Economía, sociedad y cultura. | Alianza Editorial*. tercera edición. Madrid. Available at: https://www.alianzaeditorial.es/libro.php?id=1107620&id_col=100508 (Accessed: 14 January 2018).

Cornella, A. (2004) *Infoxicación: buscando un orden en la información*. 1ªEdición. Edited by librosinfonimia. Madrid. Available at: http://www.infonomia.com/wp-content/uploads/2014/05/948_infoxicacion.pdf (Accessed: 14 January 2018).

Curto Díaz, J. and Conesa Caralt, J. (2010) *Introducción al Business Intelligence*. Editorial UOC.

Dahlgren, G. and Whitehead, M. (1991) *Policies and strategies to promote social equity in health. Background document to WHO - Strategy paper for Europe, Institute for Futures Studies, Arbetsrapport*.

Departament de Salut and Generalitat de Catalunya (no date) *Enquesta de salut de Catalunya (ESCA). Departament de Salut*. Available at: http://salutweb.gencat.cat/ca/el_departament/estadistiques_sanitaries/enquestes/esca// (Accessed: 15 January 2018).

Few, S. (2006) *Information dashboard design : the effective visual communication of data*. O'Reilly. Available at: https://archive.org/details/pdfy--fQ3cC8TeDUArgti (Accessed: 15 January 2018).

Friedman, V. (2008) *Data Visualization and Infographics — Smashing Magazine*. Available at: https://www.smashingmagazine.com/2008/01/monday-inspiration-data-visualization-and-infographics/ (Accessed: 15 January 2018).

Giron, P. (2012) 'Determinants of self-rated health in Spain: differences by age groups for adults', *The European Journal of Public Health*, 22(1), pp. 36–40. doi: 10.1093/eurpub/ckq133.
Gonzalo, E. and Pasarín, M. I. (2004) '[Health among the elderly].', *Gaceta sanitaria*, 18 Suppl 1, pp. 69–80. Available at: http://www.ncbi.nlm.nih.gov/pubmed/15171860 (Accessed: 15 January 2018).
Harrington, H. J. and Tiznado Santana, M. A. (1997) *Mejoramiento de los procesos de la empresa*. Colombia: McGraw-Hill. Available at: https://www.casadellibro.com/libro-mejoramiento-de-los-procesos-de-la-empresa/9789586001687/403983 (Accessed: 15 January 2018).
Idler, E. L. and Benyamini, Y. (1997) 'Self-rated health and mortality: a review of twenty-seven community studies.', *Journal of health and social behavior*, 38(1), pp. 21–37. Available at: http://www.ncbi.nlm.nih.gov/pubmed/9097506 (Accessed: 15 January 2018).
Kaplan, R. S. (2016) *El cuadro de mando integral = The balanced scorecard*. 3ª. Gestión 2000.
Kimball, R. and Ross, M. (2013) *Data Warehouse Toolkit: The Definitive Guide To Dimensional Modeling: Amazon.es: Ralph Kimball And Margy Ross: Libros*. 3ªEdition. John Wiley & Sons, Inc. Available at: https://www.amazon.es/Data-Warehouse-Toolkit-Definitive-Dimensional/dp/8126544279/ref=asap_bc?ie=UTF8 (Accessed: 14 January 2018).
Medina Antonia *et al.* (2013) *Comparació d'indicadors de salut per grups socioeconòmics ESCA Servei del Pla de Salut Direcció General de Planificació i Recerca en Salut, Generalitat de Catalunya*. Available at: http://salutweb.gencat.cat/web/.content/home/el_departament/estadistiques_sanitaries/enquestes/01_enquesta_catalunya/documents/informe_classe_social.pdf (Accessed: 15 January 2018).
Miedes-Ugarte, B., López, S. and Moreno, A. (2014) *Escenarios alternativos para las TICs y el conocimiento territorial en el marco de las transiciones socioecológicas: ¿Haceia una sociedad del 'buen conocer'?*
Molarius, A. *et al.* (2007) 'Socioeconomic conditions, lifestyle factors, and self-rated health among men and women in Sweden', *The European Journal of Public Health*, 17(2), pp. 125–133. doi: 10.1093/eurpub/ckl070.
Morcillo Cebolla, V. *et al.* (2014) 'Desigualdades en la salud autopercibida de la población española mayor de 65 años', *Gaceta Sanitaria*. Elsevier Doyma, 28(6), pp. 511–521. doi: 10.1016/J.GACETA.2014.05.008.
Omran, A. (1998) 'The epidemiologic transition theory revisited thirty years later.', *World Health Statistics Quarterly*, 51(2–4), pp. 99–119. Available at: https://www.popline.org/node/527048 (Accessed: 15 January 2018).
Portal de datos abiertos del Ayuntamiento de Madrid (no date) *Encuesta de Salud de la Ciudad de Madrid 2013 - Portal de datos abiertos del Ayuntamiento de Madrid*. Available at: http://datos.madrid.es/portal/site/egob/menuitem.c05c1f754a33a9fbe4b2e4b284f1a5a0/?vgnextoid=77e22cbf3ee07510VgnVCM1000001d4a900aRCRD&vgnextchannel=374512b9ace9f310VgnVCM100000171f5a0aRCRD&vgnextfmt=default. (Accessed: 15 January 2018).
Subirats, L. *et al.* (2015) 'Monitoring and Prognosis System Based on the ICF for People with Traumatic Brain Injury.', *International journal of environmental research and public health*. Multidisciplinary Digital Publishing Institute (MDPI), 12(8), pp. 9832–47. doi: 10.3390/ijerph120809832.

Tamayo-Fonseca, N. *et al.* (2013) 'Self-rated health and mortality: a follow-up study of a Spanish population', *Public Health*, 127(12), pp. 1097–1104. doi: 10.1016/j.puhe.2013.09.003.
Tamayo-Fonseca, N. *et al.* (2015) 'Self-rated health and hospital services use in the Spanish National Health System: a longitudinal study.', *BMC health services research*. BioMed Central, 15, p. 492. doi: 10.1186/s12913-015-1158-8.
Viegas, F. B. and Wattenberg, M. (2006) 'Communication-minded visualization: A call to action [Technical forum]', *IBM Systems Journal*, 45(4), pp. 801–812. doi: 10.1147/sj.454.0801.