



ESCUELA SUPERIOR DE INGENIERÍA

INGENIERÍA EN INFORMÁTICA

EPA Explorer

Software de preparación, procesado y análisis de datos de la EPA
(Encuesta de Población Activa)

Curso 2016-2017

José Saúco Delgado

Cádiz, 07 de septiembre de 2017



ESCUELA SUPERIOR DE INGENIERÍA

INGENIERÍA EN INFORMÁTICA

EPA Explorer

Software de preparación, procesado y análisis de datos de la EPA
(Encuesta de Población Activa)

DEPARTAMENTO: Ingeniería Informática.

DIRECTORES DEL PROYECTO: Dña. Elisa Guerrero Vázquez y Don Andrés Yáñez Escolano.

AUTOR DEL PROYECTO: Don José Saúco Delgado

Cádiz, 07 de septiembre de 2017

José Saúco Delgado

Este documento se halla bajo la licencia FDL¹.

Según estipula la licencia, se muestra aquí el aviso de copyright. Se ha usado la versión inglesa de la licencia, al ser la única que se ha reconocido oficialmente por la FSF².

Copyright© 2017 José Saúco Delgado

Permission is granted to copy, distribute and/or modify this document under the terms of the GNU Free Documentation License, Version 1.3 or any later version published by the FSF; With no Invariant Sections, no Front-Cover Texts, and no Back-Cover Texts.

A copy of the license is included in the section entitled “GNU Free Documentation License”.

¹ Free Documentation License

² Free Software Foundation

A mi madre, al final lo hice.

*"The programmers of tomorrow are the wizards of the future.
You're going to look like you have magic powers compared to everybody else."*

Gabe Newell, fundador y presidente de Valve.

Agradecimientos

- A mi novia *Débora* por estar ahí para mí en todo momento.
- A todos y cada uno de los miembros de mi familia, en especial a mi madre. No sé cuántas veces me habrá preguntado que “como voy con el proyecto”.
- A la sección de pescadería de Mercadona, por facilitar un encuentro fortuito con Andrés. Si no llega a ser por eso no estamos hoy aquí.
- A *Elisa Guerrero Vázquez* y *Andrés Yáñez Escolano*, por ayudarme y motivarme a terminar este proyecto.

Muchas gracias a todos.

ÍNDICE GENERAL

CAPÍTULO 1. INTRODUCCIÓN.....	1
1.1. Antecedentes.....	1
1.2. Objetivos.....	3
1.3. Tecnologías empleadas.....	4
1.4. Contenido.....	5
CAPÍTULO 2. DESCRIPCIÓN GENERAL	7
2.1. Descripción	7
2.2. Framework Cliente-Servidor	7
2.3. Características principales.....	8
CAPÍTULO 3. PLANIFICACIÓN.....	9
3.1. Metodología de desarrollo	9
3.2. Diagrama de Gantt	12
3.3. Esfuerzos.....	14
3.4. Presupuesto	15
CAPÍTULO 4. DESARROLLO DEL PROYECTO	17
4.1. Especificación de requisitos del sistema	17
4.1.1. Requisitos interfaces externas.....	17
4.1.2. Requisitos funcionales	19
4.1.3. Requisitos de rendimiento	21
4.2. Análisis del sistema	21
4.2.1. Modelo de casos de uso	22
4.2.2. Modelo conceptual de datos	33
4.2.3. Modelo de comportamiento.....	34
4.3. Diseño del sistema	49
4.3.1. Arquitectura de sistema software	49
4.3.2. Diseño de base de datos	51
4.3.3. Diseño detallado del sistema.....	54
4.3.4. Interfaz con el usuario	58
4.4. Codificación.....	59
4.4.1. Implementación.....	59
4.4.2. Otros programas.....	61
4.5. Pruebas y validación	62
4.5.1. Pruebas incrementales	62
4.5.2. Entorno de pruebas.....	64
CAPÍTULO 5. CONCLUSIONES.....	67

5.1	Retrospectiva	67
5.2	Futuras ampliaciones	68
5.3	Valoraciones personales	69
BIBLIOGRAFÍA		71
Referencias		71
APÉNDICE A: GNU FREE DOCUMENTATION LICENSE V 1.3.....		75

ÍNDICE DE FIGURAS

Figura 1: Ejemplo de información obtenida de la EPA. Fuente: cadenaser.com	1
Figura 2: Contenido de los ficheros de microdatos de la EPA.....	2
Figura 3: Captura de ejemplo de PC-Axis.	3
Figura 4: Estructura cliente-servidor de Shiny. Fuente: shiny.rstudio.com	7
Figura 5: Gestión de sesiones R en Shiny. Fuente: shiny.rstudio.com	8
Figura 6: Diagrama de modelo incremental. Fuente: procesosoftware.wikispaces.com .	9
Figura 7: Diagrama de Gantt	13
Figura 8: Tiempo estimado de las iteraciones.....	14
Figura 9: Tabla días estimados y reales	15
Figura 10: Coste total del proyecto.....	16
Figura 11: Menús de Navegación de EPA Explorer	17
Figura 12: Vista de exploración de una variable	18
Figura 13: Diagrama de casos de uso.....	22
Figura 14: CU de Análisis Exploratorio de Datos de Una variable.....	23
Figura 15: CU de Análisis Exploratorio de Datos de Dos variables.....	24
Figura 16: CU de Análisis Exploratorio de Datos de Múltiples variables.....	25
Figura 17: CU de Análisis Exploratorio de Datos de Serie temporal	26
Figura 18: CU de Entrenamiento Clustering	27
Figura 19: CU de Visualizar Clustering.....	28
Figura 20: CU de Entrenamiento Reglas de Asociación	29
Figura 21: CU de Visualizar Reglas de Asociación	30
Figura 22: CU de Generación de Informes	31
Figura 23: CU de Actualización de base de datos	32
Figura 24: Diagrama conceptual de clases.....	33
Figura 25: DSS del caso de uso de Análisis Exploratorio de Datos de Una variable	34
Figura 26: DSS del caso de uso de Análisis Exploratorio de Datos de Dos variables	37
Figura 27: DSS del caso de uso de Análisis Exploratorio de Datos de Múltiples variables	38
Figura 28: DSS del caso de uso de Análisis Exploratorio de Datos de Serie temporal.....	39
Figura 29: DSS del caso de uso de Entrenamiento Clustering.....	40
Figura 30: DSS del caso de uso de Visualizar Clustering.....	42
Figura 31: DSS del caso de uso de Entrenamiento Reglas de Asociación	43
Figura 32: DSS del caso de uso de Visualizar Reglas de Asociación.....	44
Figura 33: DSS del caso de uso de Generación de Informes.....	45
Figura 34: DSS del caso de uso de Actualización de base de datos	47
Figura 35: Patrón de diseño Modelo Vista Controlador.	49
Figura 36: Estructura MVC de aplicación Shiny.....	50
Figura 37: Definición lógica de la tabla de observaciones.....	53
Figura 38: Descripción de elementos en la Interfaz.....	55

Capítulo 1

Figura 39: Descripción de elementos en el Servidor	56
Figura 40: Descripción de elementos de la clase Base de Datos.....	57
Figura 41: Menús de Navegación de EPA Explorer	58
Figura 42: Pruebas caso de uso Análisis Exploratorio de Datos de Una variable	64
Figura 43: Pruebas caso de uso Análisis Exploratorio de Datos de Dos variables	64
Figura 44: Pruebas caso de uso Análisis Exploratorio de Datos de Múltiples variable ..	65
Figura 45: Pruebas caso de uso Análisis Exploratorio de Datos de Serie Temporal.....	65
Figura 46: Pruebas caso de uso Entrenamiento Clustering.....	65
Figura 47: Pruebas caso de uso Visualizar Clustering.....	65
Figura 48: Pruebas caso de uso Entrenamiento Reglas de Asociación	66
Figura 49: Pruebas caso de uso Visualizar Reglas de Asociación.....	66
Figura 50: Pruebas caso de uso Generación de Informes.....	66
Figura 51: Pruebas caso de uso Actualización de base de datos	66

Capítulo 1. Introducción

1.1. Antecedentes

La Encuesta de Población Activa (EPA) [1], elaborada por el Instituto Nacional de Estadística (INE) [2], es un estudio estadístico destinado a capturar datos sobre el mercado de trabajo, que se utiliza para calcular la tasa de desempleo, tal y como la define la Organización Internacional del Trabajo (OIT). La figura 1 a continuación muestra un ejemplo de información extraída de la EPA.



Figura 1: Ejemplo de información obtenida de la EPA. Fuente: cadenaser.com

Los datos se recogen con periodicidad trimestral mediante entrevista personal o telefónica. Las entrevistas están distribuidas uniformemente a lo largo de las 13 semanas del trimestre. Siguiendo directrices de Eurostat, la primera semana del año es aquella que contiene el primer jueves de dicho año. El primer trimestre consta de las 13 semanas consecutivas que comienzan en la primera semana del año. Al segundo trimestre se adjudican las 13 siguientes y así sucesivamente. Puede encontrarse una descripción completa de metodología que sigue la encuesta en el sitio oficial online del INE [3].

Capítulo 1

Se realiza desde 1964, si bien ha habido diversos cambios en el procedimiento estadístico a lo largo del tiempo que afectan a la continuidad de la información. Está considerada como el mejor indicador de la evolución del empleo y desempleo en España.

La EPA se realiza en hogares distribuidos por todo el territorio nacional. Según los últimos cambios metodológicos adoptados a partir del primer trimestre de 2005, la muestra consta de 3.588 secciones y 18 viviendas por sección, excepto en las provincias de Madrid, Barcelona, Sevilla, Valencia y Zaragoza, en las que el número de entrevistas por sección es de 22. Así, el tamaño de la muestra es de unas 65.000 familias, que en la práctica quedan reducidas a unas 60.000, equivalentes a unas 200.000 personas.

El INE publica los datos obtenidos en los distintos ejercicios de la EPA, en un formato de tabla donde cada fila corresponde a una persona encuestada, y cada columna a una de las preguntas que ha contestado en dicha encuesta. La figura 2 corresponde al contenido de uno de estos ficheros de microdatos.

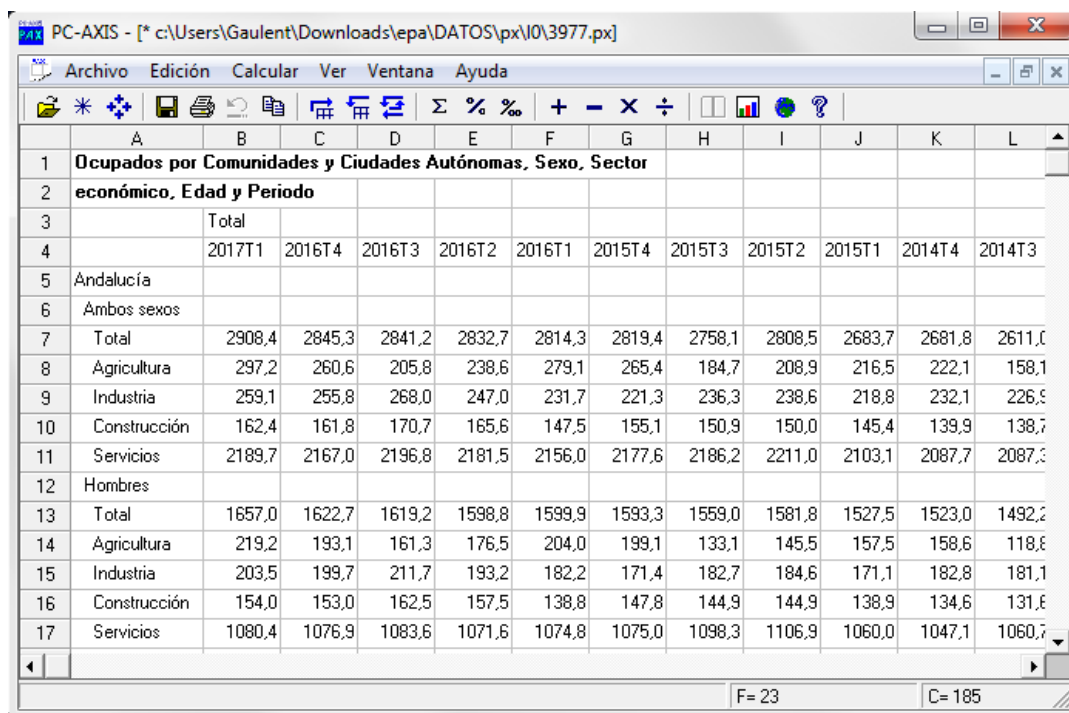
datos_2016t3: Bloc de notas

Archivo	Edición	Formato	Ver	Ayuda										
1761601000011014011000000	101	1	S1	0153	3	1	8008	11	01402647	1	4000400040006			
1761601000021014011020000	101	1	SP	0403	3	1	8508	6	0503	00103701	1	4500450045006		
1761601000021023526010000	101	1	SU	0253	3	666	1							
1761601000022030036000102	01	1												
1761601000031014011000000	101	1	SU	0213	3	1	8305		12001	1	40004000			
1761601000041014011020000	109	1	S1	0163	3	1	8308	11	16216209	1	4000400040006			
1761601000041024526010000	101	1	S1	0143	3	6610401	8308	11	18118109	1	4000400000006			
1761601000041031631000002	101	1	S1	0173	3	666	3							
1761601000051014516020000	201	1	S1	0153	3	1	8308	11	14815001	1	4000400032006			
1761601000051024526010000	201	1	SU	0213	3	666	3							
1761601000061014011000000	101	1	SG	0183	3	1	5807411		20720701	1	3500350032006			
1761601000071014511000000	101	1	SU	0203	3	666	3							
1761601000081014011020000	201	1	SU	0253	3	1	4503		13201	1	45004500			
1761601000081024026010000	201	1	SU	0243	3	666	3							
1761601000082030536000102	01	1												
1761601000082040536000102	01	1												
1761601000091013516000000	126	1	SP	0193	3	6610101	580736	0500	00601201	1	3500350000006			
1761601000101013016000000	101	1	SU	0303	3	1	5508	11	03204401	1	3900390039006			
1761601000111014016000000	401	1	SU	0343	1PE018	6610101	5508	11	16717901	1	4000400000006			
1761601000111022031000001	101	1	SP	0203	3	666	3							
1761601000121013011020000	101	1	SG	0173	3	6610101	4208	11	03809701	1	4000400000006			
1761601000121023026010000	101	1	SU	0283	3	6610101	2808	11	04604601	1	3500350000006			
1761601000131013016040000	401	1	S1	0173	3	6610401	8208	6	0536	01406201	1	4000400000006		
1761601000132020536000001	01	1												
1761601000131033571000000	128	1	SU	0183	3	666	3							
1761601000131043521010000	101	1	SU	0253	3	1	8208	6	0536	01306001	1	4000400040006		
1761601000141013516000000	101	1	S1	0163	1PE999	1	8308	6	0108	00300701	1	4000400045006		
1761601000151013016020000	101	1	SU	0283	3	1	3908	11	01505101	605200020002006				
1761601000151023021010000	148	1	SU	0233	3	1	9308	6	0609	00300501	1	4000400040006		

Figura 2: Contenido de los ficheros de microdatos de la EPA.

Además, provee de una herramienta de análisis de datos bajo Windows (PC-Axis) [4], aunque esta herramienta se limita a cálculos y gráficas estadísticas básicas, sobre

resultados que ya han sido procesados previamente. En la figura 3 a continuación, se muestra una visualización de ejemplo de la herramienta PC-Axis.



PC-AXIS - [* c:\Users\Gaulent\Downloads\epa\DATOS\px\I0\3977.px]

Archivo Edición Calcular Ver Ventana Ayuda

	A	B	C	D	E	F	G	H	I	J	K	L
1	Ocupados por Comunidades y Ciudades Autónomas, Sexo, Sector											
2	económico, Edad y Período											
3		Total										
4		2017T1	2016T4	2016T3	2016T2	2016T1	2015T4	2015T3	2015T2	2015T1	2014T4	2014T3
5	Andalucía											
6	Ambos sexos											
7	Total	2908,4	2845,3	2841,2	2832,7	2814,3	2819,4	2758,1	2808,5	2683,7	2681,8	2611,0
8	Agricultura	297,2	260,6	205,8	238,6	279,1	265,4	184,7	208,9	216,5	222,1	158,1
9	Industria	259,1	255,8	268,0	247,0	231,7	221,3	236,3	238,6	218,8	232,1	226,9
10	Construcción	162,4	161,8	170,7	165,6	147,5	155,1	150,9	150,0	145,4	139,9	138,7
11	Servicios	2189,7	2167,0	2196,8	2181,5	2156,0	2177,6	2186,2	2211,0	2103,1	2087,7	2087,3
12	Hombres											
13	Total	1657,0	1622,7	1619,2	1598,8	1599,9	1593,3	1559,0	1581,8	1527,5	1523,0	1492,2
14	Agricultura	219,2	193,1	161,3	176,5	204,0	199,1	133,1	145,5	157,5	158,6	118,8
15	Industria	203,5	199,7	211,7	193,2	182,2	171,4	182,7	184,6	171,1	182,8	181,1
16	Construcción	154,0	153,0	162,5	157,5	138,8	147,8	144,9	144,9	138,9	134,6	131,6
17	Servicios	1080,4	1076,9	1083,6	1071,6	1074,8	1075,0	1098,3	1106,9	1060,0	1047,1	1060,7

F= 23 C= 185

Figura 3: Captura de ejemplo de PC-Axis.

1.2. Objetivos

El objetivo de este Proyecto Fin de Carrera es el desarrollo de una herramienta que sirva como soporte para el análisis estadístico y minería de datos sobre los microdatos publicados por la EPA y que se denominará EPA Explorer.

La herramienta debe permitir una amplia funcionalidad sobre los datos de los distintos ejercicios de la EPA, como puede ser:

- El análisis exploratorio sobre los datos recogidos con distintos modos de visualización y representaciones.
- La aplicación de técnicas de aprendizaje computacional no supervisado, como clustering o reglas de asociación.
- La generación de informes, como la generación de notas de prensa o tablas con distintos indicadores estadísticos definidos previamente.

Capítulo 1

Además, la herramienta debe actualizar su base de datos con la información trimestral de cada ejercicio, obtenida del repositorio oficial de la EPA. Los datos se publican en forma de ficheros de texto plano en un formato no estándar, por lo que la herramienta debe ser capaz de interpretar, almacenar, procesar y normalizar dichos datos para su posterior uso.

La interfaz de usuario debe ser atractiva, visual, amigable y fácil de usar.

Finalmente, estará basada en un entorno web, por lo que será: multiplataforma, sin instalador y con cálculo centralizado en servidor dedicado.

1.3. Tecnologías empleadas

Se hará uso del lenguaje de programación R [6] para el desarrollo de la misma por su creciente popularidad en el campo de la computación estadística y sus motores de visualización gráfica. R es un paquete GNU distribuido gratuitamente bajo la Licencia Publica General de GNU (GNU GPL).

Se han utilizado varios paquetes adicionales de R para cubrir distintas necesidades dentro del proyecto. En concreto se ha hecho uso de Shiny [7] como framework para el desarrollo de una interfaz web fácil de usar y mantener.

Como entorno de desarrollo se ha elegido RStudio [8], debido a su integración con paquetes de uso muy extendido de R, como pueden ser Shiny o RMarkdown.

Aparte de los ya mencionados, el proyecto hace uso de gran cantidad de paquetes R como pueden ser las siguiente, por mencionar algunos de ellos:

- *dplyr*: Como librería genérica de transformación de datos.
- *readr*: Interpretación de datos en texto plano.
- *rcurl*: Acceso a recursos de datos a través de internet.
- *ggplot2*: Librería de generación de visualizaciones.
- *arules*: Entrenamiento y manipulación de reglas de asociación.

Como herramienta para la gestión del proyecto en materia de esfuerzo y tiempo se ha utilizado la versión 2.8.5 de la aplicación de código abierto Gantt Project [9].

Se ha utilizado **DIA** [10] en su versión 0.97.2 como herramienta de modelado para la creación de diagramas. Es distribuido bajo licencia GPL. Tiene la capacidad de trabajar con distintos tipos de diagramas como UML, entidad-relación, topologías de red o diagramas de flujo.

1.4. Contenido

Este documento está estructurado en diferentes capítulos cuyos contenidos se describen a continuación.

El capítulo 2 hace una descripción general del proyecto, contando brevemente que tipo de arquitectura se plantea utilizar, así como el alcance en funcionalidad del mismo.

El capítulo 3 trata sobre la planificación del proyecto, se hace una descripción de la metodología de desarrollo elegida y como esta afecta a nuestra planificación en términos de tiempo y esfuerzo. Además, se incluye un apartado con un posible presupuesto del proyecto.

El capítulo 4 conforma el grueso del desarrollo del proyecto, se compone de 5 apartados que describen las distintas fases principales del desarrollo del software. Estos apartados son:

- Especificación de requisitos.
- Análisis del sistema.
- Diseño del sistema.
- Codificación.
- Pruebas y validación.

El ultimo capitulo está dedicado a las conclusiones, donde se evalúa el grado de cumplimiento de los objetivos iniciales, además de presentar posibles mejoras e incorporaciones futuras.

Capítulo 2. Descripción general

2.1. Descripción

El análisis exploratorio de datos es una aproximación para resumir y visualizar las características más importantes de un conjunto de datos. Este se enfoca en explorar los datos para entender la estructura subyacente de los mismos, mostrar el origen o motivo de dichos datos o para decidir cómo dichos datos deben ser investigado por métodos estadísticos más formales.

EPA Explorer es una herramienta que servirá como punto de apoyo al estudio y análisis de tendencia de los datos recogidos por el INE en la encuesta de población activa. Esta herramienta podrá ser utilizada por personal no familiarizado con el lenguaje R o la programación.

Ha sido concebida como una herramienta donde múltiples usuarios pueden acceder a un mismo servicio centralizado de cálculo a través de un navegador web, evitando la necesidad de que los clientes dispongan de computadores muy potentes para realizar dichos cálculos.

2.2. Framework Cliente-Servidor

Para facilitar la estructura cliente servidor se ha utilizado el framework Shiny [7] del lenguaje R. Shiny nos permite definir y mantener una vista web, independiente del proceso que realiza el servidor, como queda ilustrado en la figura 4.

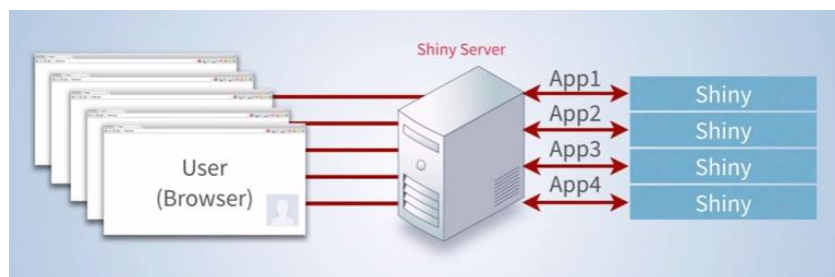


Figura 4: Estructura cliente-servidor de Shiny. Fuente: shiny.rstudio.com

El servidor mantiene en ejecución varias sesiones de R que satisfacen las distintas peticiones de cálculo realizadas por navegadores de los usuarios. Shiny gestionará las sesiones de R (o workers) necesarias para satisfacer la demanda, como puede verse en la figura 5

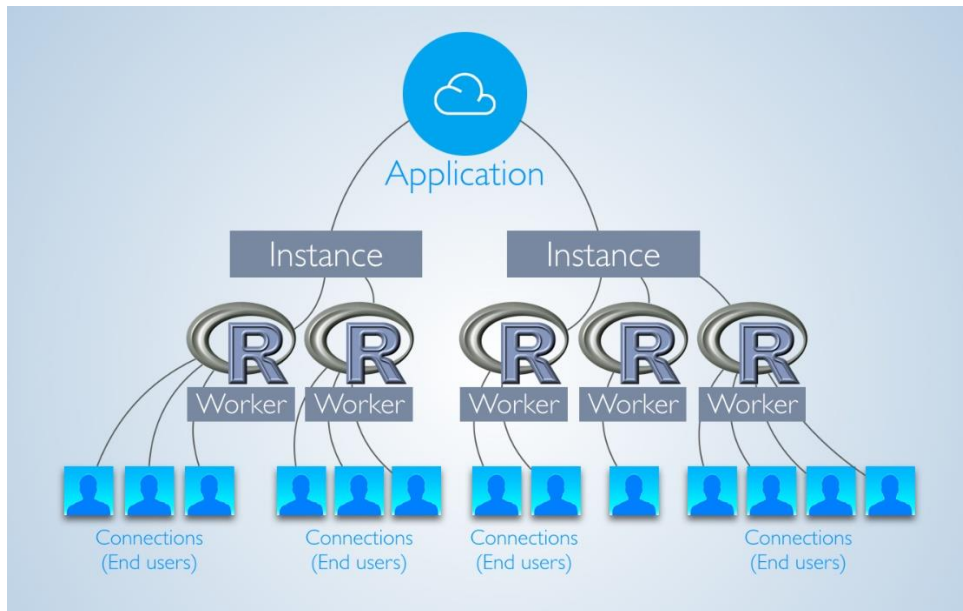


Figura 5: Gestión de sesiones R en Shiny. Fuente: shiny.rstudio.com

2.3. Características principales

A continuación, se listan los principales puntos que caracterizan nuestra herramienta:

- El usuario será capaz de obtener distintos tipos de visualizaciones y métricas de los datos contenidos de los distintos ejercicios de la EPA.
- Además, podrá lanzar en el servidor de cálculo el entrenamiento de ciertos algoritmos de aprendizaje maquina no supervisados con los datos almacenados.
- El sistema será capaz de recrear visualizaciones y distintas representaciones de los métodos de aprendizaje no supervisados anteriormente descritos.
- El usuario podrá verificar la publicación de nuevas actualizaciones del INE y actualizar su base de datos con dichos datos.
- Existirá la opción de generar informes en distintos formatos desde la propia herramienta. Estos informes serán fácilmente ampliables en el futuro.

Capítulo 3. Planificación

3.1. Metodología de desarrollo

Analizando el proyecto a desarrollar se determinó que seguir un modelo de desarrollo software de tipo incremental [11], sería la opción más apropiada para acometer el problema en cuestión. Este modelo de desarrollo se caracteriza por plantear la planificación de un proyecto en distintos bloques temporales que pasaremos a denominar iteración [12].

En cada iteración repetiremos el mismo proceso definido para el resto. De esta forma el cliente dispondría al final de cada iteración una versión del producto que funciona cumpliendo un conjunto concreto de funcionalidad acordado previamente. Como puede verse en la figura 6, en cada iteración seguirá de nuevo el proceso completo, incrementando el conjunto de funcionalidades entregadas al cliente hasta completarla al final del desarrollo.

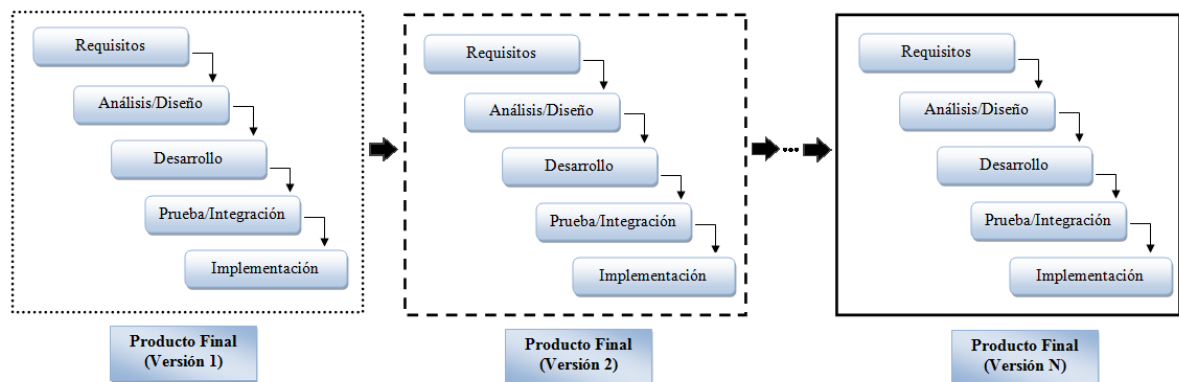


Figura 6: Diagrama de modelo incremental. Fuente: procesosoftware.wikispaces.com

Capítulo 3

La elección de esta metodología de desarrollo se llevó a cabo considerando principalmente los siguientes motivos:

- El planteamiento inicial del proyecto en cuestión, donde se plantean una serie de herramientas o aplicaciones a priori independientes a modo de toolbox, pudiendo identificar un conjunto de estas dentro de la iteración.
- La propia naturaleza del lenguaje R para trabajar de forma modular.

A continuación, se enumeran las distintas iteraciones identificadas en la elaboración de este proyecto.

Primera iteración: Conceptualización e Interpretación de los datos

En una primera fase de conceptualización de la herramienta se estudia la viabilidad del proyecto, así como las tecnologías a utilizar. Aunque parecía claro el uso del lenguaje R, se barajaban distintas alternativas de como plantear la interfaz hombre-maquina.

Después de un primer análisis de las posibles capacidades de la herramienta, se determina como prueba de viabilidad el poder hacer una interpretación de los datos a tratar de la EPA que ofrece el INE. Así pues, se genera un primer prototipo capaz de hacer lectura de estos ficheros y hacer interpretaciones básicas de los mismos.

Segunda iteración: Diseño de la base de datos

Después de esta primera fase de interpretación de los datos se determina como necesario el almacenar los mismos en una base de datos local. Esto es debido a cuestiones de eficiencia y rendimiento por el gran volumen de datos a tratar.

De esta forma en esta fase se diseñan las estrategias para hacer un uso eficiente de los datos, así como planear su captura y almacenamiento.

Tercera iteración: Incorporación de Análisis Exploratorio de Datos

En esta iteración los esfuerzos se vuelcan en estudiar las técnicas más usadas de Análisis Exploratorio de Datos o EDA (Exploratory Data Analysis), así como las posibilidades de uso de distintas visualizaciones de los datos y su encaje con la herramienta.

Una vez realizado el estudio a través de distintos cursos y fuentes de referencia se procede a sintetizar las visualizaciones o análisis más interesantes en distintas categorías.

Cuarta iteración: Incorporación del motor para exportación documental

De forma paralela a la iteración anterior se lanza la incorporación de un exportador de documentación, capaz de automatizar la obtención de ciertos informes básicos haciendo uso de los datos almacenados de la EPA.

Para esto se toman como ejemplo las notas de prensa que el propio INE genera en cada uno de sus ejercicios trimestrales, con el objetivo de automatizar lo máximo posible la generación de dichas notas de prensa.

Quinta iteración: Incorporación de Actualización de la Base de Datos

Una vez llegados a este punto se detecta como necesaria la incorporación de un mecanismo automatizado de detección de nuevas actualizaciones de los datos de la EPA por parte del INE, así como la importación y normalización de estos datos a la base de datos local.

En esta iteración se implementa este mecanismo avisando al usuario de la existencia de estos nuevos ficheros publicados y de la posibilidad de incluirlos a su repositorio.

Sexta iteración: Incorporación de Reglas de Asociación

Llegados a este punto se estudia el posible uso y explotación de los datos haciendo uso de distintas técnicas de aprendizaje máquina, donde surgen como mejores candidatos técnicas de aprendizaje no supervisado como reglas de asociación o técnicas de agrupamiento (o clustering).

Aquí se estudia que posibilidades de visualización y explotación de reglas de asociación son las más comunes entre la comunidad, y se hace una implementación de algunas de las mismas en la herramienta.

Séptima iteración: Incorporación de Técnicas de Agrupamiento

Por último, en esta iteración se continua con el esfuerzo por la implementación de técnicas de aprendizaje maquina sobre los datos de la EPA, considerando como interesante el uso de técnicas de agrupamiento (clustering).

Debido a la naturaleza categórica de los datos se considera hacer uso de algoritmos alternativos a los clásicos de clustering como k-means (MacQueen, 1967) más basados en observaciones numéricas. Se presenta como mejor candidato el algoritmo de los k-modes [13].

3.2. Diagrama de Gantt

Como herramienta para la gestión del proyecto en materia de esfuerzos y tiempo se ha utilizado la herramienta de código abierto Gantt Project.

A continuación, se muestran vistas de esta herramienta mostrando los diagramas de proyecto con las iteraciones propuestas y su estimación en el tiempo. Más abajo, en la figura 7 se ilustran en el propio diagrama estos periodos de tiempo encajados en el calendario real de desarrollo del proyecto.

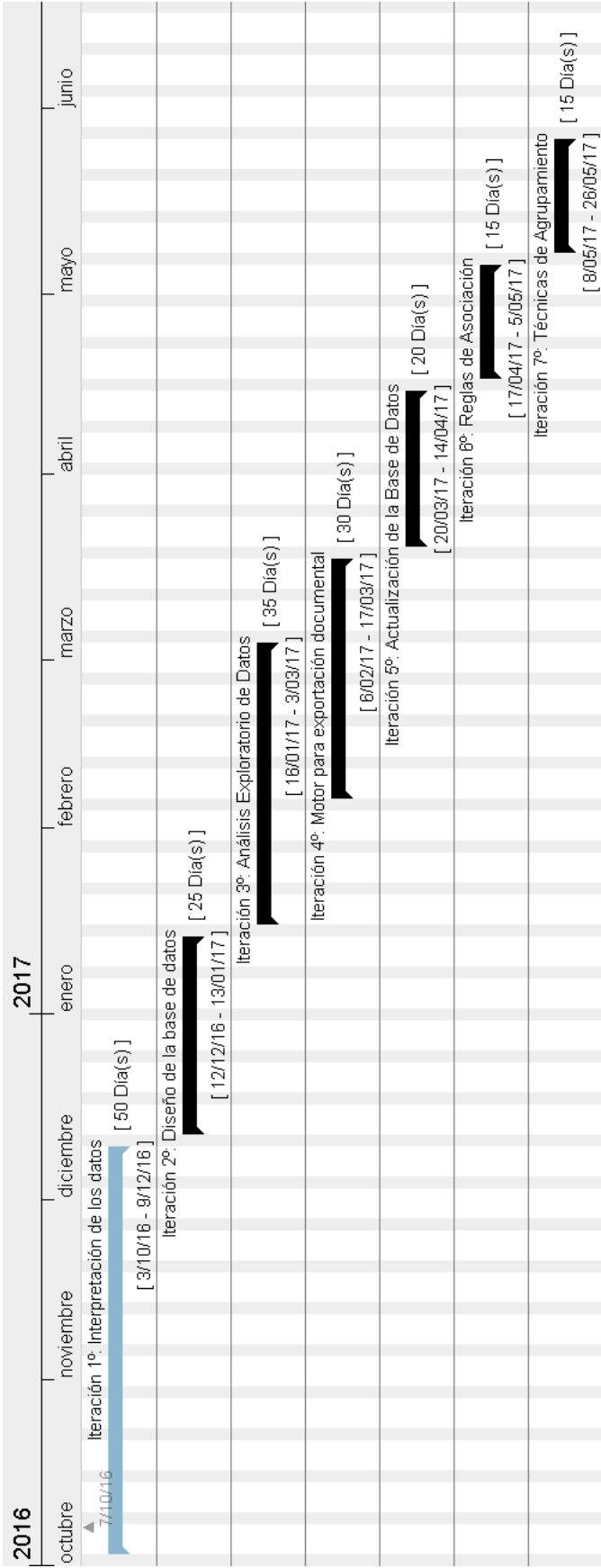


Figura 7: Diagrama de Gantt

Capítulo 3

En otra extracción de la herramienta Gantt Project, en la figura 8 se muestran las 7 iteraciones principales junto con el tiempo estimado, la primera iteración, la Interpretación de los datos ha sido la de mayor duración estimada, con 50 días de dedicación. Las ultimas iteraciones (6ª y 7ª) se estiman más cortas y de igual duración, ya que van a existir cierto solape funcional entre ellas pudiendo reutilizar gran parte del trabajo.



Nombre	Fecha de inicio	Fecha de fin	Duración
Iteración 1ª: Interpretación de los datos	3/10/16	9/12/16	50
Iteración 2ª: Diseño de la base de datos	12/12/16	13/01/17	25
Iteración 3ª: Análisis Exploratorio de Datos	16/01/17	3/03/17	35
Iteración 4ª: Motor para exportación documental	6/02/17	17/03/17	30
Iteración 5ª: Actualización de la Base de Datos	20/03/17	14/04/17	20
Iteración 6ª: Reglas de Asociación	17/04/17	5/05/17	15
Iteración 7ª: Técnicas de Agrupamiento	8/05/17	26/05/17	15

Figura 8: Tiempo estimado de las iteraciones

3.3. Esfuerzos

A continuación, se muestra una relación mostrando cada una de las iteraciones propuestas junto la estimación de esfuerzo en días a invertir por iteración justo al esfuerzo real invertido. En la figura 9 se puede observar como las estimaciones iniciales fueron demasiado optimistas, existiendo cierta demora en cada una de las fases, provocando un desajuste en el calendario de 29 días en total.

Tareas Realizadas	Estimados	Reales
Iteración 1º: Interpretación de los datos	50	60 (+10)
Iteración 2º: Diseño de la base de datos	25	30 (+5)
Iteración 3º: Análisis Exploratorio de Datos	35	38 (+3)
Iteración 4º: Motor para exportación documental	30	36 (+6)
Iteración 5º: Actualización de la Base de Datos	20	22 (+2)
Iteración 6º: Reglas de Asociación	15	17 (+2)
Iteración 7º: Técnicas de Agrupamiento	15	16 (+1)
Totales	190 días	219 días

Figura 9: Tabla días estimados y reales

La demora más importante frente al estimado inicial aparece en la iteración tercera correspondiente al motor de generación documental, ya que ha supuesto el estudio del lenguaje de marcado que no se consideró inicialmente.

3.4. Presupuesto

A continuación, se presenta un presupuesto del coste de este proyecto.

En materia del software utilizado para el desarrollo del mismo solo se ha hecho uso de software de código abierto gratuito y de libre distribución, por lo que no supondría un coste adicional en el cómputo final.

Aun así, por la naturaleza del proyecto, la implantación del mismo podría llegar a suponer la adquisición de hardware para dar soporte a la herramienta en caso de no existir en la organización. Se estima que el coste de una computadora que cumpla con los requisitos del proyecto ascendería a 1.200€ aproximadamente. Se toma como referencia una estación de trabajo con procesador i7-6700, 16GB de RAM y almacenamiento de 2TB en HDD y 120GB de SSD.

Capítulo 3

El coste humano ha supuesto la mayor parte del coste del proyecto, destinado tanto a labores de análisis como de programación. Tomando como referencia el “Convenio colectivo nacional de empresas de ingeniería y oficinas de estudios técnicos” [14], consultando la categoría de Nivel 3, vemos como el salario mínimo anual de un programador sería de 16.917,60€ brutos, y según el nivel 1 para el analista asciende a 23.618,28€ brutos.

Considerando entonces el tiempo invertido a lo largo de las distintas iteraciones del proyecto dedicado al análisis ha sido de aproximadamente 1 mes, y 7 de desarrollo en media jornada. Además, es necesario considerar otros costos indirectos asociados que se estiman como el 10% del coste total del proyecto. De esta forma podemos resumir el cómputo de costes del proyecto a continuación en la figura 10:

Tareas	Coste estimado (euros)
Análisis	984,10 €
Codificación, diseño y pruebas	4.934,3 €
Infraestructura	1.200 €
Gastos indirectos (10% total)	591,84 €
Total	7.710,24 €

Figura 10: Coste total del proyecto

Capítulo 4. Desarrollo del proyecto

4.1. Especificación de requisitos del sistema

En este apartado se describirán los requisitos que nuestra aplicación deberá cumplir, tanto de interfaces externas como funcionales.

4.1.1. Requisitos interfaces externas

A continuación, se describen los requisitos a cumplir por las interfaces de la aplicación de cara al software y al usuario.

La herramienta muestra a los usuarios una interfaz principal compuesta por una serie de vistas organizadas por un menú principal superior. Será visual, amigable y fácil de usar. La figura 11 muestra un ejemplo de dicho menú principal.

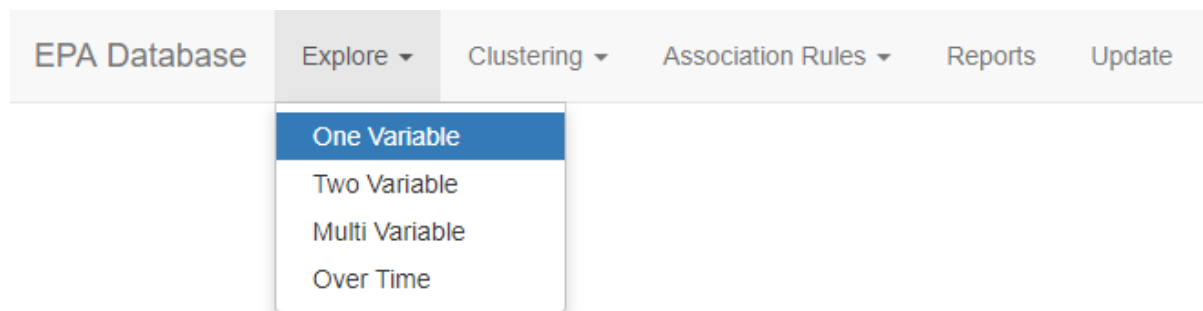


Figura 11: Menús de Navegación de EPA Explorer

El menú permitirá al usuario navegar por la aplicación pudiendo elegir cualquier funcionalidad disponible. El menú se compone de las siguientes opciones:

- Explorar
 - Una variable
 - Dos variables
 - Múltiples variables
 - Serie temporal

Capítulo 4

- Agrupación
 - Entrenamiento
 - Ver
- Reglas de Asociación
 - Entrenamiento
 - Ver
- Informes
- Actualización

El usuario interactuará con la vista concreta seleccionando con el ratón las opciones que crea convenientes en los distintos elementos visuales de la vista. La aplicación entonces reaccionara actualizando los distintos elementos en pantalla como gráficos o informes.

A continuación, en la figura 12 se muestra un ejemplo del aspecto que tendrá una vista de ejemplo de la aplicación.

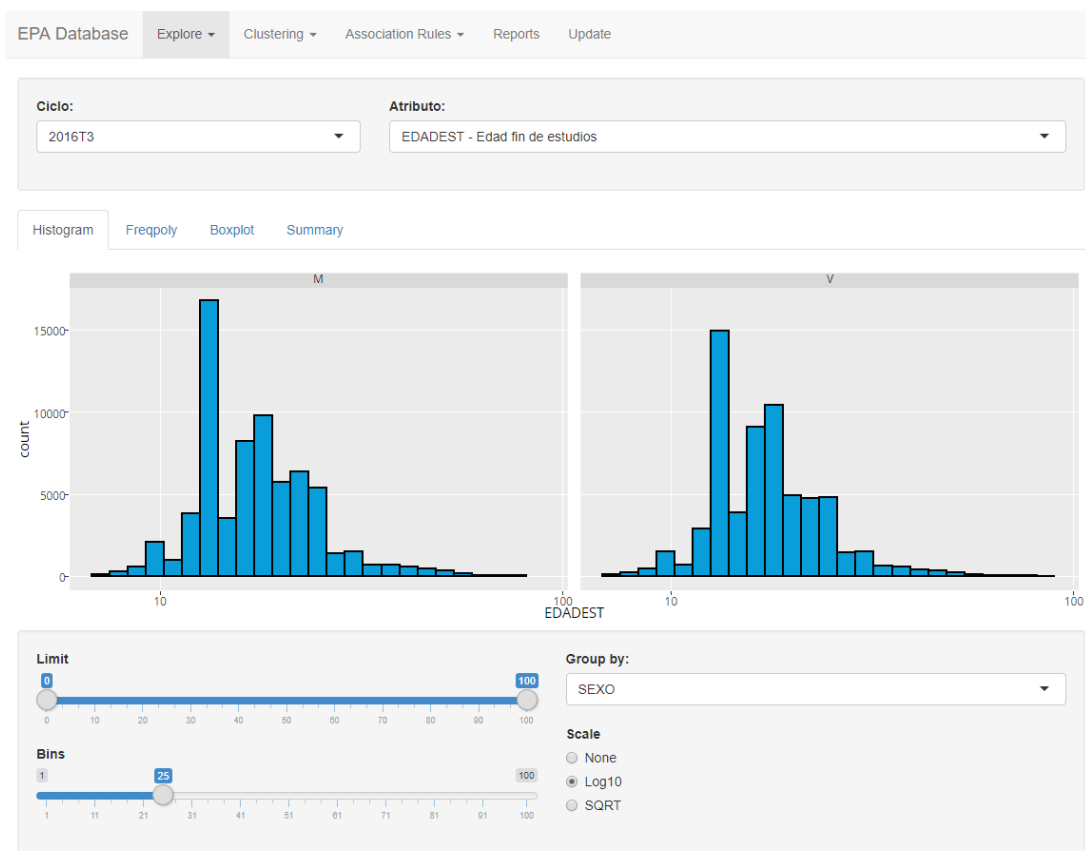


Figura 12: Vista de exploración de una variable

4.1.2. Requisitos funcionales

A continuación, se listan los requisitos que la aplicación debe cumplir. A continuación, se detallan las distintas funciones accesibles desde la interfaz principal de la aplicación, navegables en un sistema de menús en pestañas.

- **Exploración**

- Una variable:
 - Dado un ejercicio concreto y una variable, el usuario podrá explorar dicha variable haciendo uso de un conjunto de visualizaciones típicas.
 - Además, se podrán hacer ciertos ajustes a dicha visualización como agrupaciones, filtros, etc.
- Dos variables:
 - Dado un ejercicio concreto y dos variables, el usuario podrá explorar la relación entre ambas variables.
 - Como en el caso anterior, se podrán hacer ajustes a dicha visualización como agrupaciones, estudio de covarianza, etc.
- Múltiples variables:
 - Dado un ejercicio concreto y un conjunto de variables a visualizar, el usuario podrá observar con distintas representaciones como se interrelacionan dichas variables entre ellas.
- Serie temporal:
 - Seleccionando una variable, el usuario será capaz de obtener una representación de como dicha variable ha ido variando en el tiempo por los distintos ejercicios de la EPA registrados.

- **Agrupación (*clustering*):**

- Generación de los grupos (*clusters*)
 - Seleccionando un ejercicio de la EPA y ciertos valores de entrada para el algoritmo de agrupación (n° de clusters, valores iniciales de los centroides, ...), se generan los grupos o *clusters* que clasifican de forma natural los datos de entrada.
- Visualización
 - Muestra los valores de los *clusters* utilizando diferentes representaciones,

- **Reglas de Asociación:**

- Generación de las reglas de asociación
 - Seleccionando un ejercicio de la EPA y ciertos valores de entrada (el soporte, la confianza, o el tamaño de las reglas) se lanza la ejecución del algoritmo.
- Visualización
 - Muestra las reglas de asociación obtenidas haciendo uso de diferentes representaciones visuales.

- **Generación de informes:**

- Permite crear distintos informes para ser exportados en distintos formatos, como HTML, Word o PDF. Estos informes están descritos en una plantilla escrita en RMarkdown y almacenada en un fichero .Rmd en cierta ruta concreta en el servidor. La aplicación se despliega con un ejemplo de informe para la obtención de las notas de prensa de la EPA que el INE publica cada trimestre.

- **Actualización:**

- Validar los datos de trimestres disponibles actualmente
- Verificar si existen actualizaciones en los repositorios oficiales del INE.
- Incluir dichas actualizaciones en la base de datos local, para tener los datos disponibles para las distintas funciones de la herramienta.

4.1.3. Requisitos de rendimiento

La aplicación debe mantener unos mínimos tiempos de respuesta, no pudiendo interferir en ningún caso la realización de cálculos en el servidor con la libertad de operación del usuario por la interfaz.

Se intentará minimizar el uso de recursos en el servidor, aunque siempre será preferente el requisito anterior en cuanto a tiempos de respuesta, pudiendo suponer un uso de recursos adicional para cumplir dicho requisito.

4.2. Análisis del sistema

En este apartado pasaremos a definir el comportamiento de la herramienta. Para ello nos ayudaremos de diagramas de casos de uso.

Estos casos de uso están categorizados en base al menú al que pertenecen:

- Módulo de Análisis Exploratorio de Datos
 - Análisis Exploratorio de Datos de Una variable (AED_1V)
 - Análisis Exploratorio de Datos de Dos variables (AED_2V)
 - Análisis Exploratorio de Datos de Múltiples variables (AED_MV)
 - Análisis Exploratorio de Datos de Serie temporal (AED_ST)
- Agrupación
 - Entrenamiento Clustering (E_CL)
 - Visualizar Clustering (V_CL)
- Reglas de Asociación
 - Entrenamiento Reglas de Asociación (E_RA)
 - Visualizar Reglas de Asociación (V_RA)
- Informes
 - Generación de Informes (INF)
- Actualización
 - Actualización de base de datos (ACT)

4.2.1. Modelo de casos de uso

Los casos de usos ayudarán a identificar los distintos actores o roles que tendrán acceso a la herramienta y las distintas funciones que esta aporta y las distintas relaciones entre estas. En nuestro caso concreto solo se considera el rol de usuario donde todos tendrán acceso a las mismas funciones o privilegios, aunque podría variar en el futuro.

A continuación, se muestra un diagrama general mostrando todos los casos de uso de la herramienta, que posteriormente serán descritos independientemente con mayor profundidad.

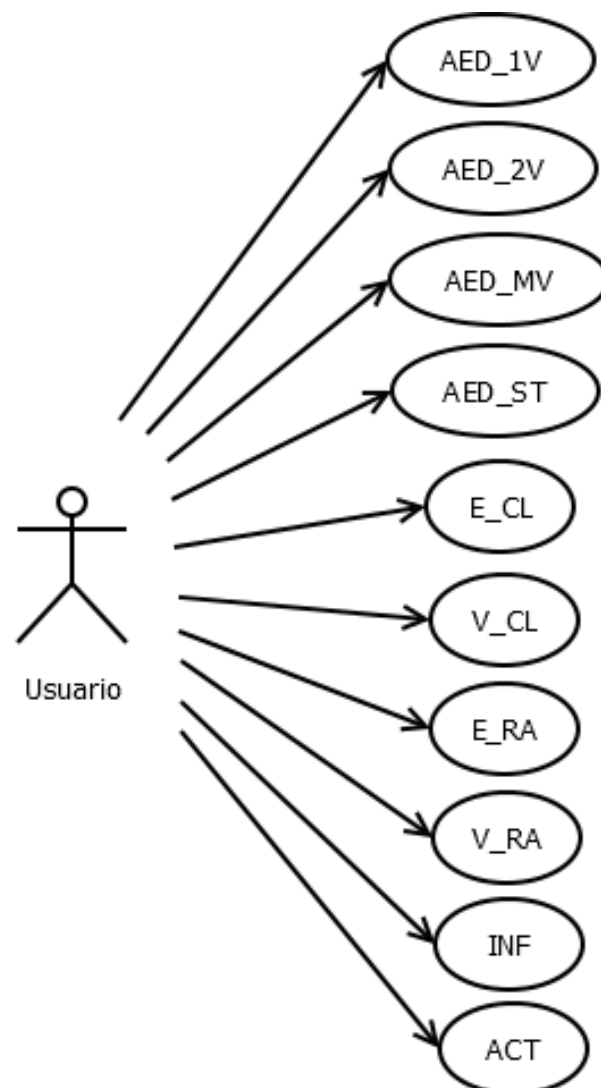


Figura 13: Diagrama de casos de uso

Descripción del caso de uso "Análisis Exploratorio de Datos de Una variable"

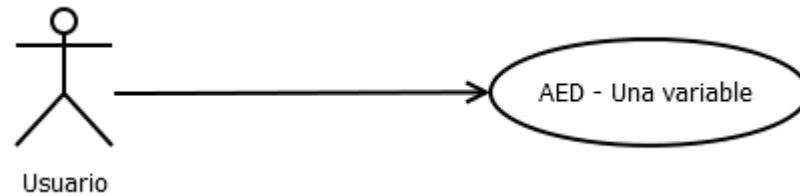


Figura 14: CU de Análisis Exploratorio de Datos de Una variable

Actor principal: Usuario.

Precondiciones: Ninguna.

Postcondiciones: Se muestran distintas visualizaciones de una variable

Escenario principal: El usuario manipula visualizaciones de una variable

Escenario principal:

1. El usuario selecciona el ejercicio y el atributo a analizar.
2. El sistema carga la información relevante a dicho conjunto de datos.
3. El usuario selecciona el tipo de visualización
4. El usuario configura la visualización ajustando distintos parámetros como:
 - Limites en la escala.
 - Agrupaciones.
 - Numero de columnas.
5. El sistema reacciona en vivo ajustando la visualización obtenida por cada selección del usuario.

Descripción del caso de uso "Análisis Exploratorio de Datos de Dos variables"

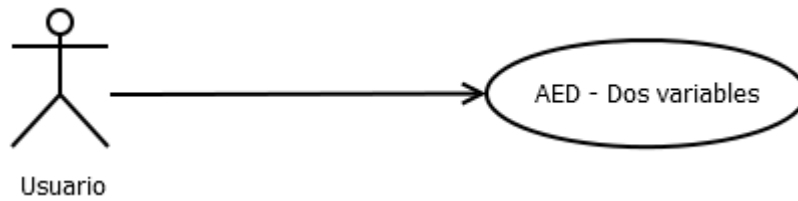


Figura 15: CU de Análisis Exploratorio de Datos de Dos variables

Actor principal: Usuario.

Precondiciones: Ninguna.

Postcondiciones: Se analiza la relación entre dos variables mediante visualizaciones.

Escenario principal: El usuario visualiza la relación entre dos variables

Escenario principal:

1. El usuario selecciona el ejercicio y los dos atributos a analizar.
2. El sistema carga la información relevante a dicho conjunto de datos.
3. El usuario selecciona el tipo de visualización
4. El usuario configura la visualización ajustando distintos parámetros como:
 - Límites en la escala.
 - Agrupaciones.
 - Jitter.
 - Estadísticos.
 - Transparencia.
5. El sistema reacciona en vivo ajustando la visualización obtenida por cada selección del usuario.

Descripción del caso de uso "Análisis Exploratorio de Datos de Múltiples variables"

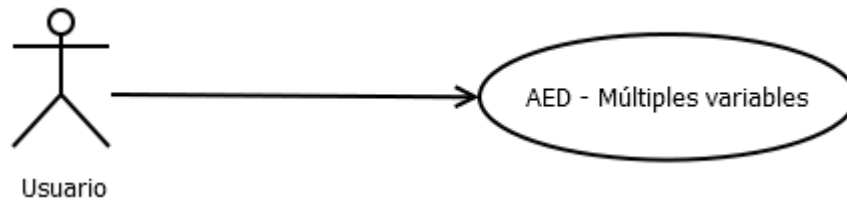


Figura 16: CU de Análisis Exploratorio de Datos de Múltiples variables

Actor principal: Usuario.

Precondiciones: Ninguna.

Postcondiciones: Se visualiza la relación entre varias variables.

Escenario principal: El usuario visualiza la relación entre múltiples variables.

Escenario alternativo 1: El usuario no selecciona los atributos a comparar.

Escenario principal:

1. El usuario selecciona el ejercicio y el conjunto de atributos a analizar.
2. El usuario lanza la generación de la visualización.
3. El sistema carga la información relevante a dicho conjunto de datos.
4. El sistema compara cada par de atributos seleccionados.
5. El sistema combina los resultados obtenidos para generar la visualización.

Escenario alternativo 1:

- 1a. El usuario no ha seleccionado los atributos a analizar.
 1. El sistema muestra el mensaje de error y pide que el usuario seleccione las variables a analizar.

Descripción del caso de uso "Análisis Exploratorio de Datos de Serie temporal"

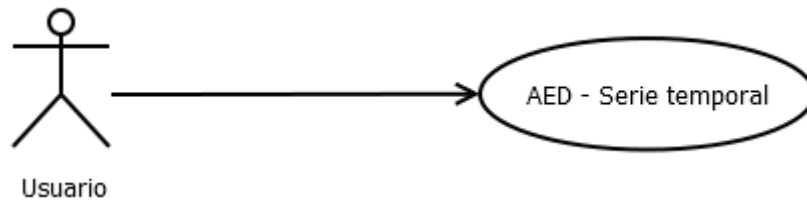


Figura 17: CU de Análisis Exploratorio de Datos de Serie temporal

Actor principal: Usuario.

Precondiciones: Ninguna.

Postcondiciones: Se visualiza la evolución de los valores de una variable en el tiempo.

Escenario principal: El usuario observa la evolución en los valores de una variable.

Escenario principal:

1. El usuario selecciona la variable a analizar en el tiempo.
2. El sistema carga la información relevante a dicho conjunto de datos.
3. El sistema muestra una visualización mostrando la evolución de dicha variable en el tiempo.

Descripción del caso de uso "Entrenamiento Clustering"

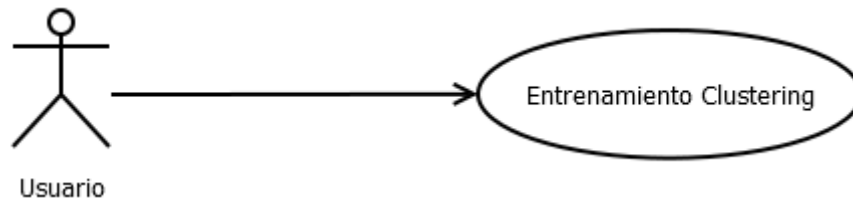


Figura 18: CU de Entrenamiento Clustering

Actor principal: Usuario.

Precondiciones: Ninguna.

Postcondiciones: Se realiza la ejecución de un algoritmo de agrupación.

Escenario principal: El usuario obtiene el resultado de la agrupación

Escenario alternativo 1: El sistema agota sus recursos en la ejecución del algoritmo.

Escenario principal:

1. El usuario selecciona el ejercicio sobre el que desea realizar la agrupación.
2. El usuario ajusta parámetros adicionales como:
 - Numero de agrupaciones.
3. El usuario ejecuta el algoritmo de agrupación.
4. El sistema almacena el resultado del algoritmo de agrupación para su posterior visualización.

Escenario alternativo 1:

- 4a. El sistema agota sus recursos en la ejecución del algoritmo.
 1. El sistema muestra el mensaje de error y pide que el usuario ajuste los parámetros adicionales

Descripción del caso de uso "Visualizar Clustering"

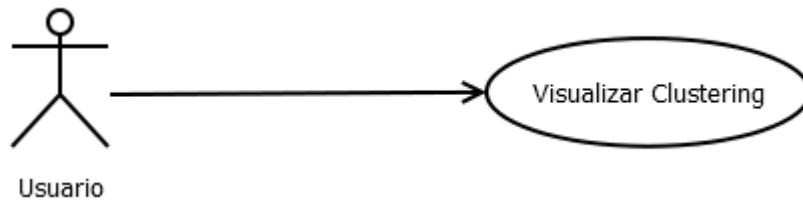


Figura 19: CU de Visualizar Clustering

Actor principal: Usuario.

Precondiciones: Debe haberse ejecutado alguna vez el algoritmo de agrupación.

Postcondiciones: Se visualiza el resultado de una agrupación.

Escenario principal: El usuario visualiza el resultado de una agrupación.

Escenario alternativo 1: No existe ejecución previa del algoritmo de agrupación.

Escenario principal:

1. El sistema obtiene el listado de las agrupaciones calculadas previamente.
2. El usuario selecciona la agrupación a analizar.
3. El usuario selecciona el tipo de visualización deseada.
4. El sistema representa la agrupación en base a la selección del usuario.

Escenario alternativo 1:

- 1a. No existe ejecución previa del algoritmo de agrupación.
 1. El sistema muestra el mensaje de error y pide que el usuario realice previamente la ejecución del algoritmo de agrupación.

Descripción del caso de uso "Entrenamiento Reglas de Asociación"

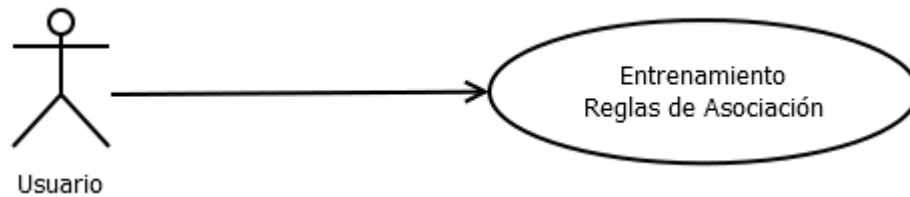


Figura 20: CU de Entrenamiento Reglas de Asociación

Actor principal: Usuario.

Precondiciones: Ninguna.

Postcondiciones: Se ejecuta el algoritmo de reglas de asociación.

Escenario principal: El usuario obtiene el resultado de las reglas de asociación.

Escenario alternativo 1: El sistema agota sus recursos en la ejecución del algoritmo.

Escenario principal:

1. El usuario selecciona el ejercicio sobre el que desea obtener las reglas de asociación.
2. El usuario ajusta parámetros adicionales como:
 - Soporte mínimo de las reglas de asociación.
 - Confianza mínima de las reglas de asociación.
 - Tamaño de las reglas de asociación.
3. El usuario ejecuta el algoritmo de obtención de reglas de asociación
4. El sistema almacena el resultado de las reglas de asociación para su posterior visualización.

Escenario alternativo 1:

- 4a. El sistema agota sus recursos en la ejecución del algoritmo.
 1. El sistema muestra el mensaje de error y pide que el usuario ajuste los parámetros adicionales

Descripción del caso de uso "Visualizar Reglas de Asociación"

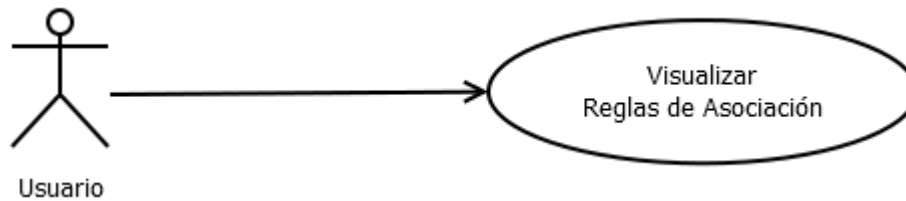


Figura 21: CU de Visualizar Reglas de Asociación

Actor principal: Usuario.

Precondiciones: Debe existir previamente algún conjunto de reglas de asociación.

Postcondiciones: Se visualiza un conjunto de reglas de asociación.

Escenario principal: El usuario visualiza un conjunto de reglas de asociación.

Escenario alternativo 1: No existe ejecución previa de reglas de asociación.

Escenario principal:

1. El sistema obtiene el listado de las reglas de asociación calculadas previamente.
2. El usuario selecciona el conjunto de reglas de asociación a analizar.
3. El usuario selecciona el tipo de visualización deseada.
4. El sistema representa el conjunto de reglas de asociación en base a la selección del usuario.

Escenario alternativo 1:

- 1a. No existe ejecución previa de reglas de asociación.
 1. El sistema muestra el mensaje de error y pide que el usuario realice previamente la ejecución de las reglas de asociación.

Descripción del caso de uso "Generación de Informes"

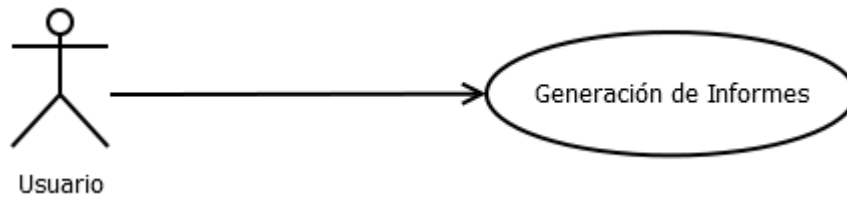


Figura 22: CU de Generación de Informes

Actor principal: Usuario.

Precondiciones: Debe existir alguna plantilla de informe.

Postcondiciones: Se realiza la generación de algún informe.

Escenario principal: El usuario obtiene un informe en el formato deseado.

Escenario alternativo 1: No existe ninguna plantilla de informe.

Escenario principal:

1. El usuario selecciona la plantilla del informe a generar.
2. El usuario selecciona información adicional a la generación del informe como:
 - Identificador del ejercicio.
 - Formato de Exportación.
3. El usuario lanza la generación del informe.
4. El sistema genera el informe acorde a la información aportada.
5. El sistema envía el fichero generado a la sesión de navegador del usuario para su descarga.

Escenario alternativo 1:

- 1a. No existe ninguna plantilla de informe.
 1. El sistema muestra el mensaje de error y avisa al usuario de la necesidad de instalar una plantilla en la herramienta.

Descripción del caso de uso "Actualización de base de datos"

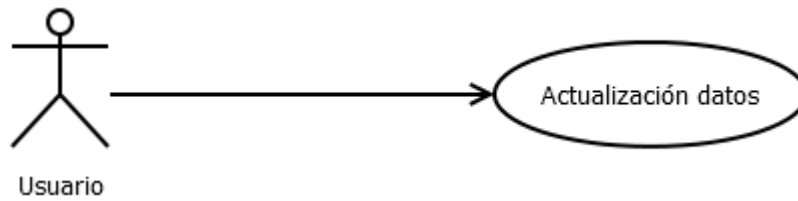


Figura 23: CU de Actualización de base de datos

Actor principal: Usuario.

Precondiciones: Debe existir conexión con el servidor ftp del INE.

Postcondiciones: Se actualizan los datos el ejercicio seleccionado.

Escenario principal: El usuario actualiza los datos disponibles para la aplicación.

Escenario alternativo 1: No existe conexión con el repositorio online del INE.

Escenario alternativo 2: No se selecciona ningún paquete de datos a actualizar.

Escenario principal:

1. El sistema accede al repositorio ftp del INE y comprueba que ficheros de datos han sido añadidos desde la última actualización.
2. El usuario selecciona el fichero de datos para actualizar la herramienta.
3. El usuario lanza el proceso de actualización.
4. El sistema descomprime y analiza el fichero de actualización.
5. El sistema actualiza los datos disponibles con los nuevos datos seleccionados.

Escenario alternativo 1:

- 1a. No existe conexión con el repositorio online del INE.
 1. El sistema muestra el mensaje de error y avisa al usuario de que no existe conexión al repositorio online del INE.

Escenario alternativo 2:

- 2a. No se selecciona ningún paquete de datos a actualizar.
 1. El sistema muestra el mensaje de error y avisa al usuario de que no se han encontrado paquetes de actualización

4.2.2. Modelo conceptual de datos

En este apartado pasaremos a definir las clases que encontraremos en nuestro sistema, así como las relaciones entre las mismas.

Concretamente en el proyecto a desarrollar se identifica un modelo simple, donde el proceso vendrá definido principalmente por los siguientes elementos:

- Interfaz: Siendo esta clase la que gestionará las operaciones introducidas por el usuario y hará las llamadas oportunas a la clase servidor para facilitar las salidas deseadas por el usuario.
- Servidor: Esta clase responderá de forma reactiva a los cambios en la clase interfaz comunicándose con la misma para entregar los resultados de los cálculos y visualizaciones.
- Datos: Existirá en el sistema una clase independiente que gestionará el acceso a la base de datos del sistema. Esta clase abstraerá al servidor del sistema concreto de base de datos utilizado, así como que tipo de lenguaje se necesita para comunicarse con el mismo.

A continuación, se muestra un diagrama de clases conceptual donde se reflejan estos elementos y su relación en el ámbito del sistema.



Figura 24: Diagrama conceptual de clases

La simpleza de este esquema radica en la naturaleza de la aplicación como herramienta de trabajo, donde el valor es aportado por los procesos de cálculo o visualización sobre los datos almacenados.

En posteriores apartados se hará un desarrollo más profundo de estos elementos en un análisis más detallado del sistema.

4.2.3. Modelo de comportamiento

En este apartado se modela como se comportará el sistema en su interacción con el usuario. Para ello se hará uso de un modelo de comportamiento por cada caso de uso anteriormente descrito.

Estos modelos se describen haciendo uso de dos componentes:

- Diagramas de secuencia: Donde se muestran de una forma visual los eventos que ocurren en dicho caso de uso en una línea temporal.
- Contrato de las operaciones del sistema: donde se describe el efecto producido por la ejecución de las operaciones mostradas en el diagrama.

Para evitar contenidos duplicados, se omitirán las operaciones que hayan sido explicadas con anterioridad.

DSS del caso de uso "Análisis Exploratorio de Datos de Una variable"

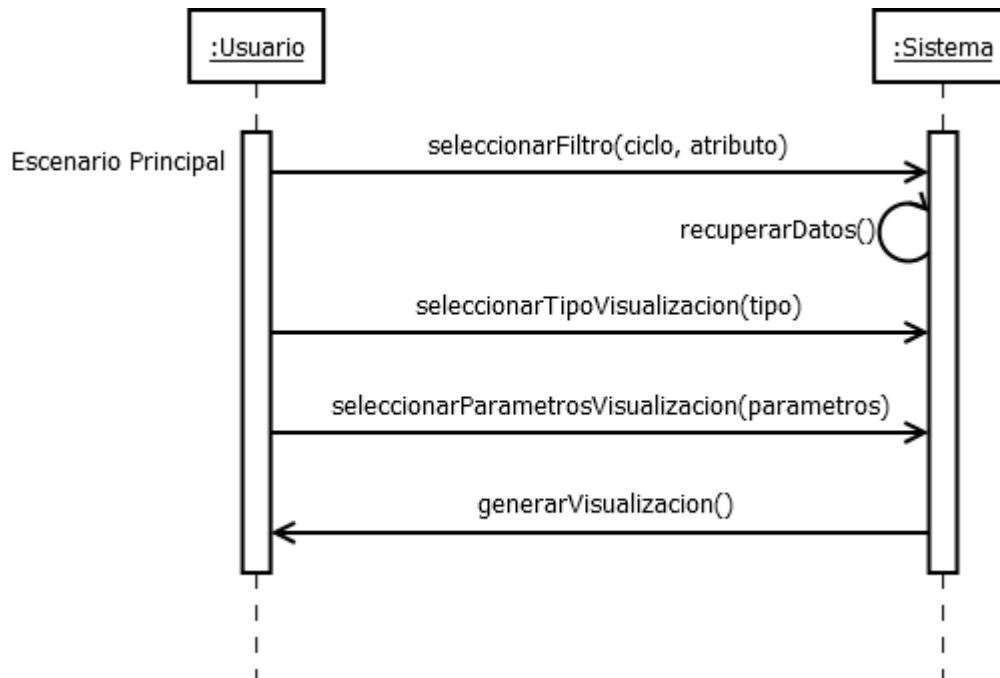


Figura 25: DSS del caso de uso de Análisis Exploratorio de Datos de Una variable

CONTRATOS

Operación: seleccionarFiltro (ciclo, atributo).

Responsabilidades: Seleccionar la fuente de datos para la visualización.

Precondiciones: Existen el atributo y el ejercicio indicados.

Postcondiciones: El sistema almacena la información del filtro aplicado para los datos.

Operación: recuperarDatos ().

Responsabilidades: Recuperar de la base de datos los indicados anteriormente.

Precondiciones: Se ha indicado previamente el filtro de datos a recuperar.

Postcondiciones: El sistema almacena el conjunto de datos siguiendo el filtro indicado.

Operación: seleccionarTipoVisualizacion (tipo).

Responsabilidades: Indicar que modo de visualización se utilizara en la representación de los datos filtrados.

Precondiciones: Los datos a representar están cargados en la memoria del sistema.

Postcondiciones: El sistema almacena el tipo de visualización seleccionado.

Operación: seleccionarParametrosVisualizacion (parámetros).

Responsabilidades: Ajustar parámetros de visualización para el tipo seleccionado.

Precondiciones: Se ha seleccionado el tipo de visualización.

Postcondiciones: El sistema almacena los parámetros de visualización seleccionado.

Operación: generarVisualizacion ().

Responsabilidades: Generar una visualización de los datos acorde a la información aportada por el usuario.

Precondiciones: Los datos a visualizar se encuentran cargados.

Postcondiciones: El sistema genera una visualización interactiva para la sesión de navegador del usuario activo.

DSS del caso de uso "Análisis Exploratorio de Datos de Dos variables"

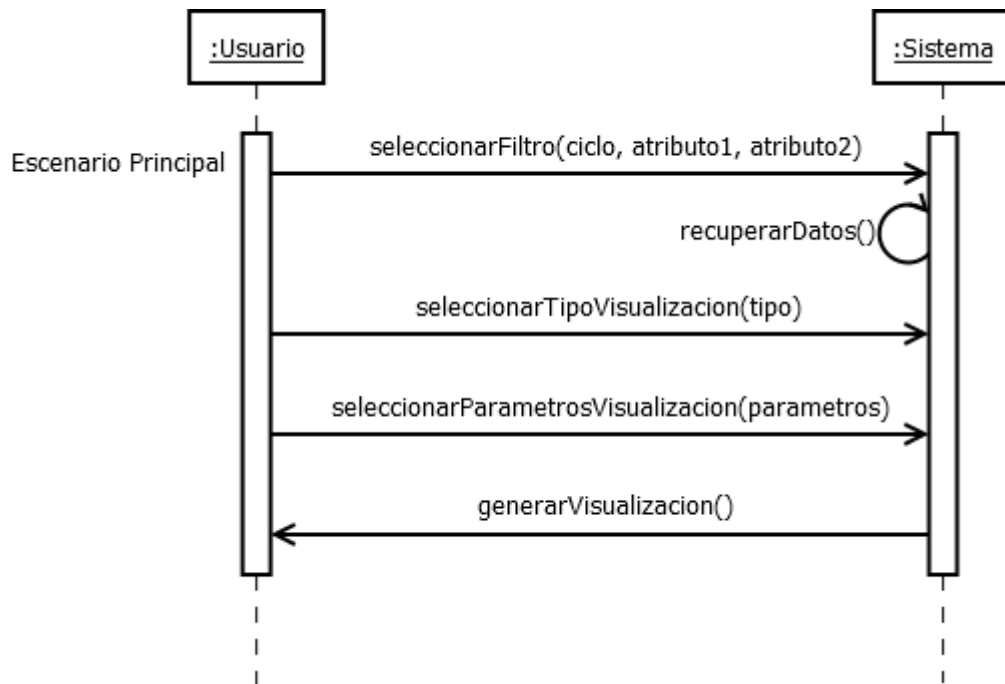


Figura 26: DSS del caso de uso de Análisis Exploratorio de Datos de Dos variables

CONTRATOS

Operación: seleccionarFiltro (ciclo, atributo1, atributo2).

Responsabilidades: Seleccionar la fuente de datos para la visualización.

Precondiciones: Existen ambos atributos seleccionados, así como el ejercicio indicado.

Postcondiciones: El sistema almacena la información del filtro aplicado para los datos.

DSS del caso de uso "Análisis Exploratorio de Datos de Múltiples variables"

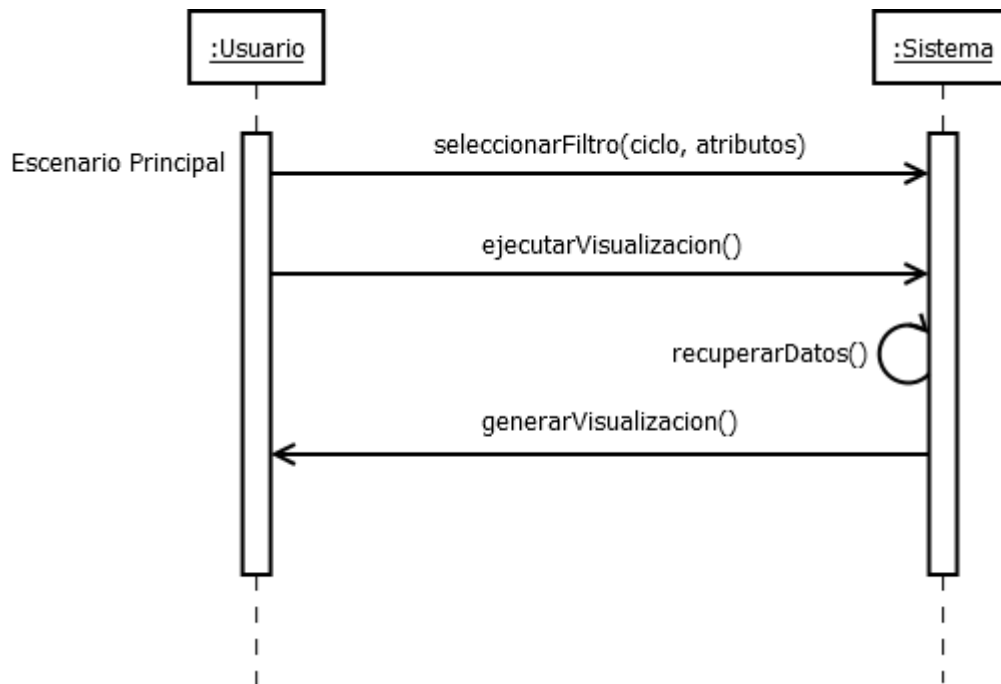


Figura 27: DSS del caso de uso de Análisis Exploratorio de Datos de Múltiples variables

CONTRATOS

Operación: seleccionarFiltro (ciclo, atributos).

Responsabilidades: Seleccionar la fuente de datos para la visualización.

Precondiciones: Existen todos los atributos seleccionados y el ejercicio indicado.

Postcondiciones: El sistema almacena la información del filtro aplicado para los datos.

Operación: ejecutarVisualizacion ().

Responsabilidades: Indicar al sistema la intención de generación de una visualización.

Precondiciones: Han sido seleccionados los atributos requeridos para la visualización.

Postcondiciones: El sistema almacena la información relevante para la generación de la visualización.

DSS del caso de uso "Análisis Exploratorio de Datos de Serie temporal"

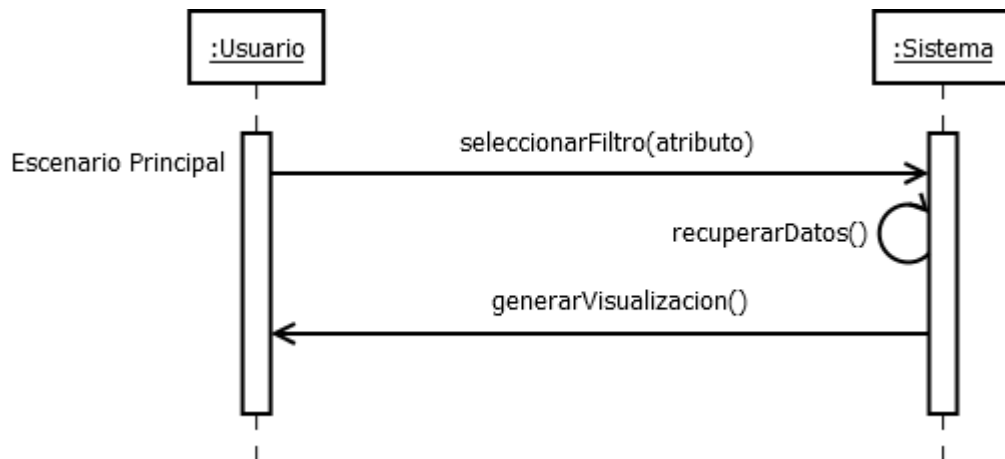


Figura 28: DSS del caso de uso de Análisis Exploratorio de Datos de Serie temporal

CONTRATOS

Operación: seleccionarFiltro (ciclo, atributo).

Responsabilidades: Seleccionar la fuente de datos para la visualización.

Precondiciones: Existe el atributo indicado por el usuario.

Postcondiciones: El sistema almacena la información del filtro aplicado para los datos.

DSS del caso de uso "Entrenamiento Clustering"

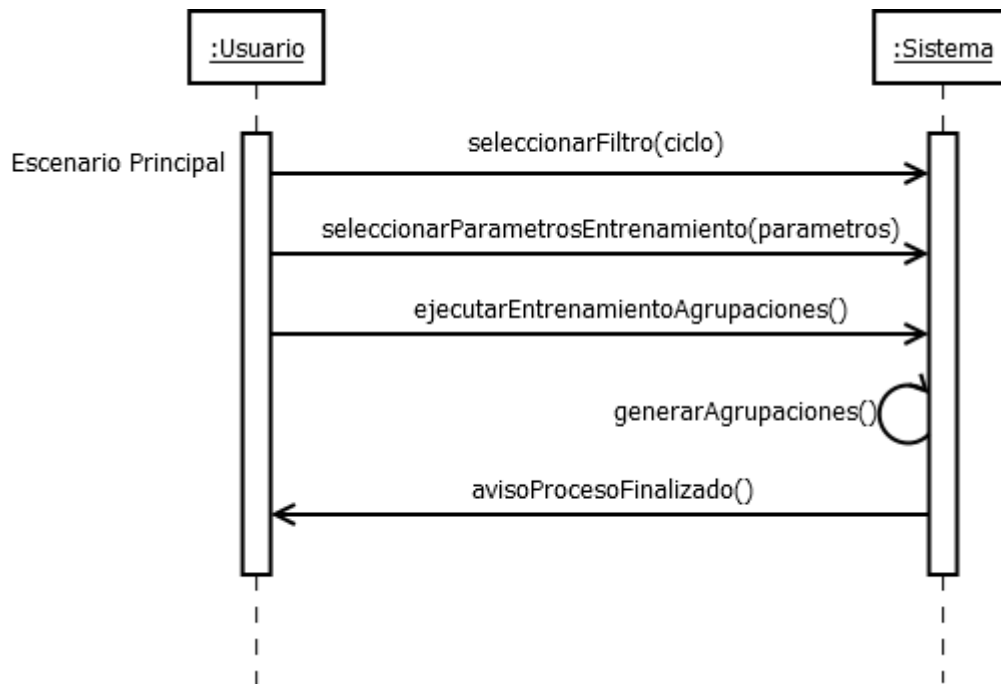


Figura 29: DSS del caso de uso de Entrenamiento Clustering

CONTRATOS

Operación: seleccionarFiltro (ciclo).

Responsabilidades: Seleccionar la fuente de datos para la visualización.

Precondiciones: Existe el ejercicio indicado por el usuario.

Postcondiciones: El sistema almacena la información del filtro aplicado para los datos.

Operación: seleccionarParametrosEntrenamiento (parámetros).

Responsabilidades: El usuario ajusta ciertos valores necesarios para el algoritmo de aprendizaje maquina seleccionado.

Precondiciones: Ninguna.

Postcondiciones: El sistema almacena los parámetros aportados por el usuario.

Operación: ejecutarEntrenamientoAgrupaciones ().

Responsabilidades: Indicar al sistema la intención de ejecutar el algoritmo de generación de agrupaciones.

Precondiciones: Han sido seleccionados el filtro de datos y los parámetros requeridos por el algoritmo.

Postcondiciones: El sistema almacena la información relevante para la ejecución del algoritmo de agrupación.

Operación: generarAgrupaciones ().

Responsabilidades: El sistema ejecuta el algoritmo de agrupaciones con los parámetros indicados por el usuario.

Precondiciones: Ninguna

Postcondiciones: El sistema almacena en disco el resultado de la ejecución del algoritmo de agrupación para su posterior revisión por el usuario.

Operación: avisoProcesoFinalizado ().

Responsabilidades: El sistema genera un aviso al usuario indicando que el proceso solicitado ha finalizado de manera satisfactoria.

Precondiciones: Ninguna

Postcondiciones: Ninguna

DSS del caso de uso "Visualizar Clustering"

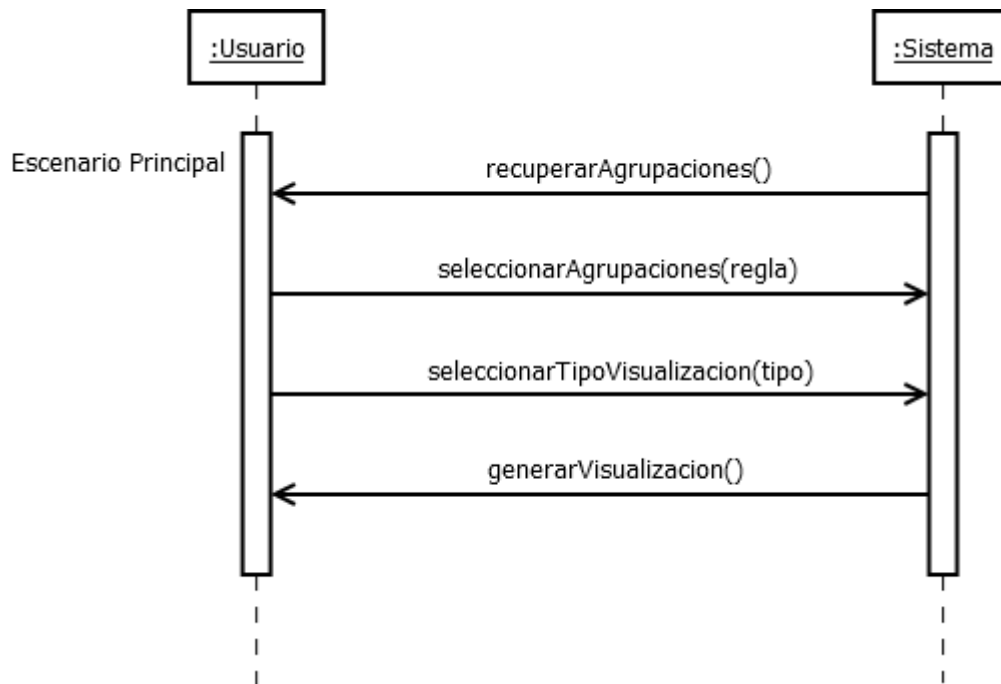


Figura 30: DSS del caso de uso de Visualizar Clustering

CONTRATOS

Operación: recuperarAgrupaciones ().

Responsabilidades: Recuperar los resultados obtenidos anteriormente por el caso de uso de “Entrenamiento Clustering”.

Precondiciones: Se ha ejecutado previamente el caso de uso “Entrenamiento Clustering”.

Postcondiciones: El sistema almacena los resultados anteriores del algoritmo de entrenamiento de agrupación.

Operación: seleccionarAgrupacion (agrupación).

Responsabilidades: El usuario indica que resultado del algoritmo de agrupación se desea explorar.

Precondiciones: Ninguna.

Postcondiciones: El sistema almacena el resultado de agrupación seleccionado.

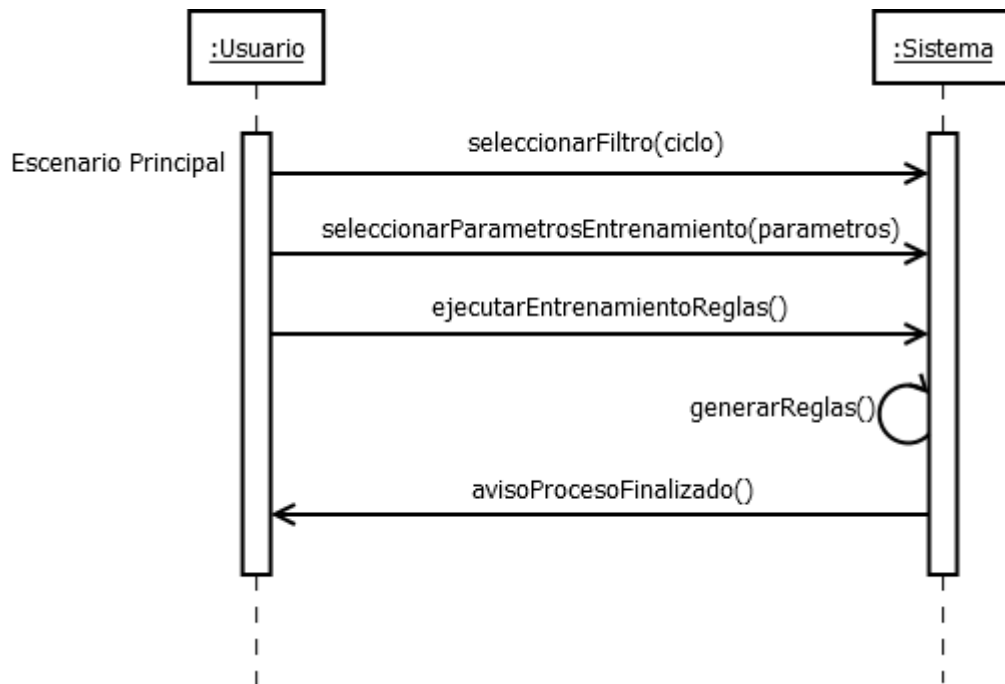
DSS del caso de uso "Entrenamiento Reglas de Asociación"

Figura 31: DSS del caso de uso de Entrenamiento Reglas de Asociación

CONTRATOS

Operación: ejecutarEntrenamientoReglas ().

Responsabilidades: Indicar al sistema la intención de ejecutar el algoritmo de reglas de asociación.

Precondiciones: Han sido seleccionados el filtro de datos y los parámetros requeridos por el algoritmo.

Postcondiciones: El sistema almacena la información relevante para la ejecución del algoritmo de reglas de asociación.

Operación: generarReglas ().

Responsabilidades: El sistema ejecuta el algoritmo de reglas de asociación con los parámetros indicados por el usuario.

Precondiciones: Ninguna

Postcondiciones: El sistema almacena en disco el resultado de la ejecución del algoritmo de reglas de asociación para su posterior revisión por el usuario.

DSS del caso de uso "Visualizar Reglas de Asociación"

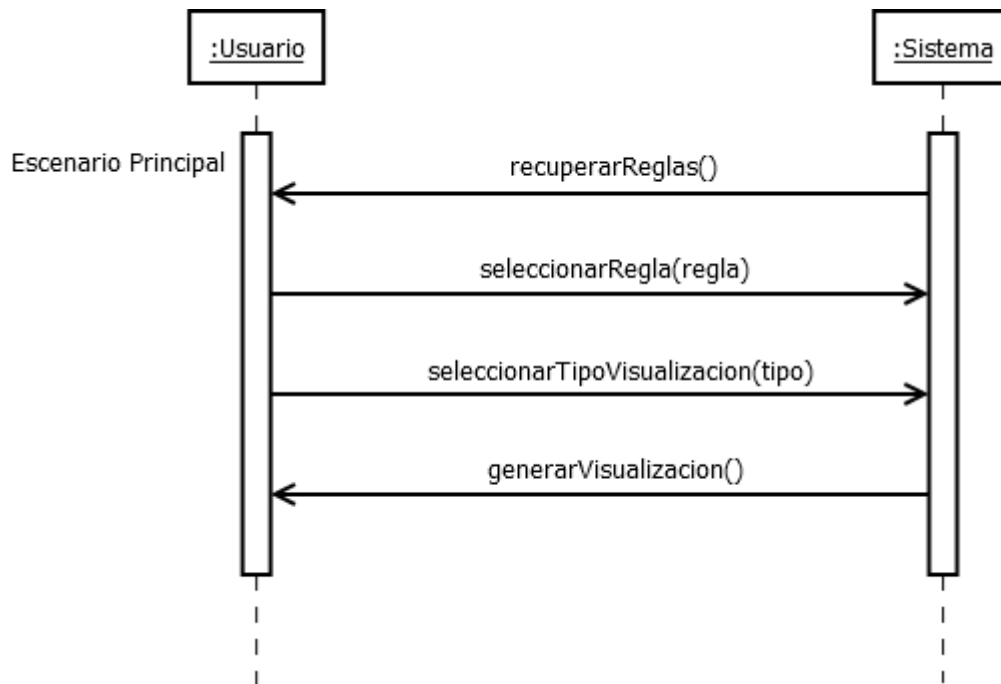


Figura 32: DSS del caso de uso de Visualizar Reglas de Asociación

CONTRATOS

Operación: `recuperarReglas ()`.

Responsabilidades: Recuperar los resultados obtenidos anteriormente por el caso de uso de “Entrenamiento Reglas de Asociación”.

Precondiciones: Se ha ejecutado previamente el caso de uso “Entrenamiento Reglas de Asociación”.

Postcondiciones: El sistema almacena los resultados anteriores del algoritmo de entrenamiento de reglas de asociación.

Operación: `seleccionarRegla (regla)`.

Responsabilidades: El usuario indica que resultado del algoritmo de reglas de asociación se desea explorar.

Precondiciones: Ninguna.

Postcondiciones: Se almacena el conjunto de reglas de asociación seleccionado.

DSS del caso de uso "Generación de Informes"

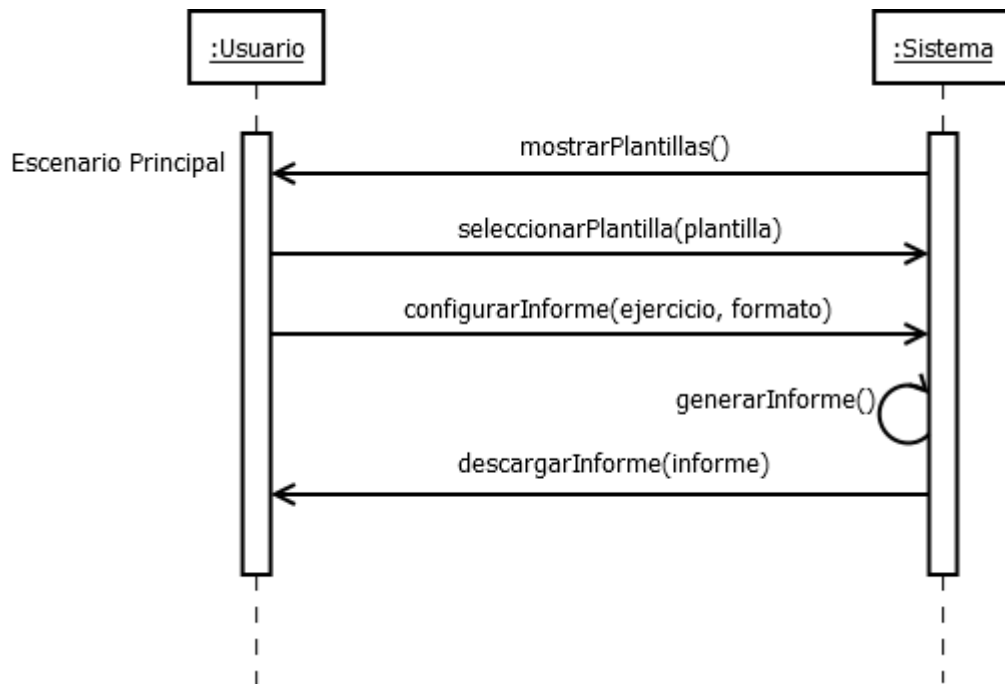


Figura 33: DSS del caso de uso de Generación de Informes

CONTRATOS

Operación: `mostrarPlantillas ()`.

Responsabilidades: Mostrar al usuario las plantillas cargadas en el sistema.

Precondiciones: Ninguna.

Postcondiciones: Se muestra al usuario la lista de plantillas para su selección.

Operación: `seleccionarPlantilla (plantilla)`.

Responsabilidades: El usuario indica la plantilla para la generación del informe.

Precondiciones: Debe existir alguna plantilla en el sistema.

Postcondiciones: El sistema almacena el identificador de plantilla seleccionada.

Capítulo 4

Operación: configurarInforme (ejercicio, formato).

Responsabilidades: El usuario ajusta el tipo de informe que desea obtener.

Precondiciones: Ninguna.

Postcondiciones: El sistema almacena la información aportada por el usuario.

Operación: generarInforme ().

Responsabilidades: El sistema hace la generación de un informe de acuerdo a la información aportada por el usuario como filtro o formatos.

Precondiciones: El usuario a aportado la información necesaria para la generación del informe.

Postcondiciones: El sistema almacena en una ruta temporal el resultado de la generación del informe.

Operación: descargarInforme (ejercicio, formato).

Responsabilidades: El sistema envía el informe al usuario.

Precondiciones: Se ha ejecutado previamente la generación de un informe.

Postcondiciones: El sistema envía el informe generado como un fichero a la sesión de navegador del usuario activo.

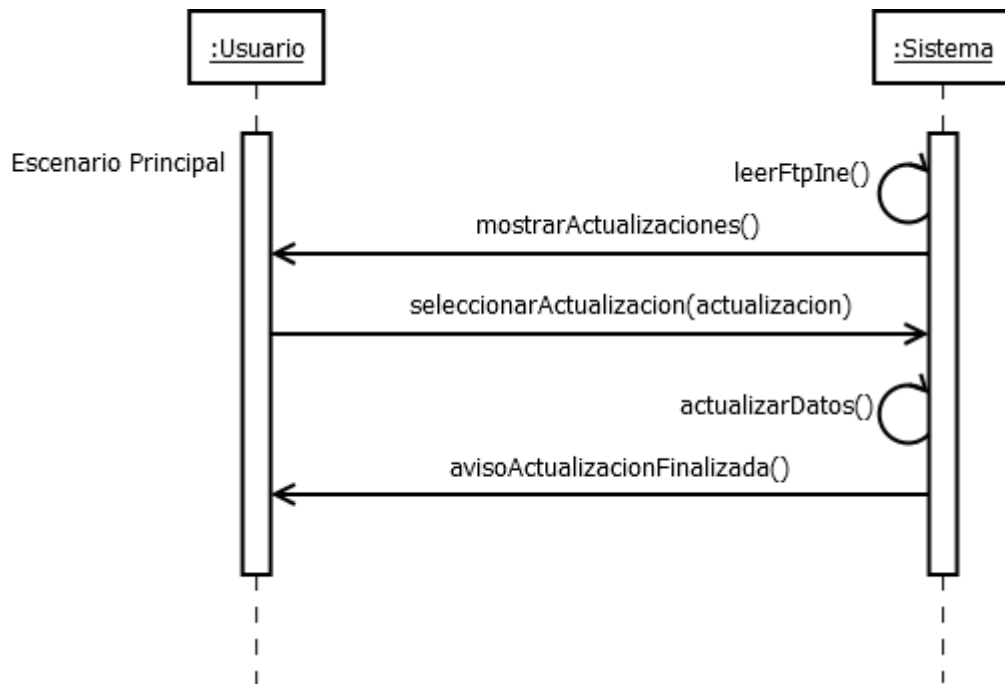
DSS del caso de uso "Actualización de base de datos"

Figura 34: DSS del caso de uso de Actualización de base de datos

CONTRATOS

Operación: leerFtpIne ().

Responsabilidades: El sistema obtiene una lista de paquetes de actualización del EPA publicados por el INE en su repositorio FTP.

Precondiciones: Ninguna.

Postcondiciones: El sistema almacena las actualizaciones disponibles en el FTP.

Operación: mostrarActualizaciones ().

Responsabilidades: El sistema muestra al usuario una selección de los paquetes seleccionables para actualizar su repositorio local de datos.

Precondiciones: El sistema debe haber obtenido la lista de paquetes disponibles.

Postcondiciones: Ninguna.

Operación: seleccionarActualizacion (actualización).

Responsabilidades: El usuario indica cuál de los paquetes de actualización disponible desea integrar en la base de datos.

Precondiciones: El paquete indicado existe en el repositorio FTP del INE.

Postcondiciones: El sistema almacena la información referente al paquete de actualización.

Operación: actualizarDatos ().

Responsabilidades: El sistema actualiza su base de datos con el paquete indicado por el usuario.

Precondiciones: El usuario ha indicado previamente el identificador de los datos a actualizar.

Postcondiciones: El sistema descarga, interpreta y almacena los datos contenidos en el paquete de actualización indicado.

Operación: avisoActualizacionFinalizada ().

Responsabilidades: El sistema genera un aviso al usuario indicando que el proceso de actualización solicitada ha finalizado de manera satisfactoria.

Precondiciones: Ninguna

Postcondiciones: Ninguna

4.3. Diseño del sistema

4.3.1. Arquitectura de sistema software

El diseño de la herramienta se ha conceptualizado haciendo uso del principio de diseño de separación de intereses.

Para ello se ha usado el framework de R, Shiny, mencionado anteriormente, que nos permite realizar un diseño muy en la línea de patrones del estilo de modelo vista controlador (o MVC) [15]. La figura 35 a continuación, ilustra un diagrama que describe dicho patrón de diseño.

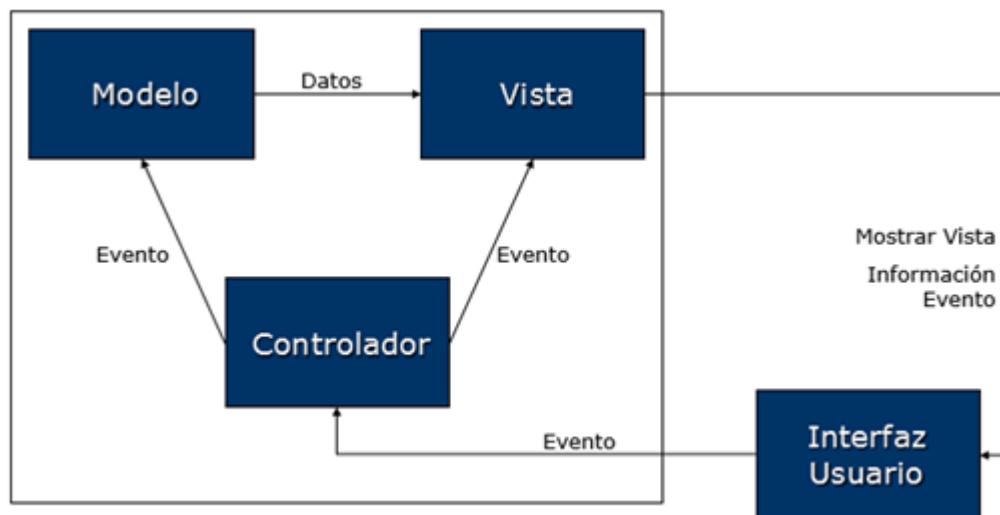


Figura 35: Patrón de diseño Modelo Vista Controlador.

Un modelo aplicación que haga uso de Shiny se compone de varios ficheros de código fuente (scripts), donde destacan principalmente dos de ellos:

- ui.R: Script que define la interfaz de usuario. Encajando con la vista en nuestro patrón MVC, este fichero contendrá la información necesaria para construir la interfaz, la definición de los formularios de entradas para las distintas operaciones, así como las salidas devueltas por el modelo.

- **server.R:** Script de servidor. Correspondiendo a la parte Modelo del patrón MVC, el fichero server.R contiene los distintos scripts que serán ejecutados de forma reactiva a las acciones del usuario sobre la interfaz. Shiny ejecutará estos scripts devolviendo sus resultados a la interfaz que se encargará de mostrarlas al usuario.

En el MVC tradicional, el controlador es necesario y explícito. Este define qué hacer cuando se reciben las solicitudes de los usuarios y qué recursos se van a movilizar para llevar a cabo las tareas necesarias descritas en el modelo. En este entorno reactivo, el controlador se convierte en una caja negra controlado por el framework.

A continuación, en la figura 36 mostramos un diagrama típico de interacción entre los distintos componentes de una aplicación Shiny.

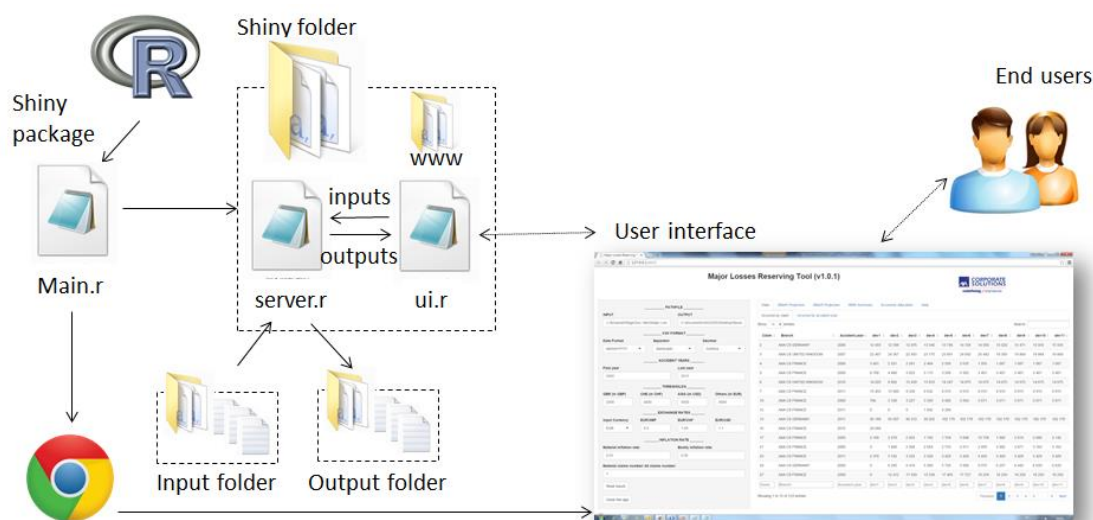


Figura 36: Estructura MVC de aplicación Shiny.

4.3.2. Diseño de base de datos

La herramienta EPA Explorer utiliza como entrada para sus distintos algoritmos y visualizaciones datos extraídos de la Encuesta de Población activa.

Estos datos consisten en una entrada u observación por cada persona encuestada en un ciclo dado. Además, cada observación esta descrita por un listado de atributos que corresponden a cada respuesta otorgada en la encuesta.

Para almacenar estos datos se detectó la necesidad de hacer uso de un pequeño sistema de gestión de base de datos. Esta decisión se vio motivada principalmente por los siguientes puntos:

- Evitar interpretar voluminosos ficheros de texto plano cada vez que queremos realizar cualquier operación sobre estos.
- Rapidez en el acceso a los datos, pudiendo crear índices de los mismos para poder hacer filtros rápidos seleccionando, por ejemplo, todas las entradas de un único ejercicio o las de un tipo de observaciones concreto.

En concreto, el sistema de gestión de base de datos seleccionado ha sido SQLite [16]. Los principales motivos para su elección han sido, ante todo:

- Funcionamiento sin ningún tipo de configuración ni instalación de ningún servidor dedicado.
- Base de datos autocontenida en un único fichero, siendo este trasladable y compatible entre distintas plataformas.
- Código bajo dominio público y libre de copyright.

A continuación, en la tabla de la figura 37, se detallan los nombres de las variables, así como el tipo que las define o una breve descripción de su contenido.

Atributo	Tipo	Descripción
CICLO	NUMERIC	Período de referencia
CCAA	FACTOR	Comunidad autónoma
PROV	FACTOR	Provincia
NIVEL	FACTOR	Nivel del registro
EDAD5	NUMERIC	Edad

Atributo	Tipo	Descripción
SEXO	FACTOR	Sexo
ECIV	FACTOR	Estado civil
REGNA	FACTOR	Provincia/Región de Nacimiento
NAC	FACTOR	Nacionalidad
EXREGNA	FACTOR	Región de nacionalidad extranjera
ANORE	NUMERIC	Años de residencia en España
NFORMA	FACTOR	Nivel de estudios
EDADEST	NUMERIC	Edad fin de estudios
CURSR	FACTOR	Cursa estudios reglados
NCURSR	FACTOR	Nivel de estudios reglados
CURSNR	FACTOR	Cursa formación no reglada
NCURNR	FACTOR	Nivel de estudios no reglados
HCURNR	NUMERIC	Horas de estudios no reglados
TRAREM	FACTOR	Trabajo remunerado
AYUDFA	FACTOR	Ayuda familiar
AUSENT	FACTOR	Se ausento al trabajo en la semana de referencia
RZNOTB	FACTOR	Razones por las que no trabajo
VINCUL	FACTOR	Vinculación con el empleo ausentes
NUEVEM	FACTOR	Ha encontrado empleo
OCUP	FACTOR	Ocupación principal
ACT	FACTOR	Actividad principal
SITU	FACTOR	Situación profesional
SP	FACTOR	Tipo de administración
DUCON1	FACTOR	Contrato indefinido o temporal
DUCON2	FACTOR	Relación laboral permanente o discontinuo
DUCON3	FACTOR	Tipo de contrato de carácter temporal
TCONT	NUMERIC	Duración del contrato o relación laboral
DREN	NUMERIC	Meses desde la renovación del contrato
DCOM	NUMERIC	Meses en la empresa
REGEST	FACTOR	Provincia/Región donde está ubicado
PARCO1	FACTOR	Jornada completa o parcial
PARCO2	FACTOR	Motivo de tener jornada parcial
HORASP	NUMERIC	Horas pactadas en contrato
HORASH	NUMERIC	Horas semanales habituales
HORASE	NUMERIC	Horas dedicadas la semana pasada
EXTRA	FACTOR	Realizó horas extraordinarias
EXTPAG	NUMERIC	Horas extraordinarias pagadas
EXTNPG	NUMERIC	Horas extraordinarias no pagadas
RZDIFH	FACTOR	Razón de la diferencia de horas
TRAPLU	FACTOR	Tiene otro u otros empleos
OCUPLU	FACTOR	Ocupación en el segundo empleo

Atributo	Tipo	Descripción
ACTPLU	FACTOR	Actividad del segundo empleo
SITPLU	FACTOR	Situación profesional segundo empleo
HORPLU	NUMERIC	Horas efectivas en el segundo empleo
MASHOR	FACTOR	Desearía trabajar más horas
DISMAS	FACTOR	Disponibilidad para trabajar más horas
RZNDISH	FACTOR	Razón para no trabajar más horas
HORDES	NUMERIC	Número de horas que desearía trabajar
BUSOTR	FACTOR	Busca otro empleo
BUSCA	FACTOR	Buscado empleo las últimas 4 semanas
DESEA	FACTOR	Desearía tener un empleo
FOBACT	FACTOR	Métodos de encontrar empleo
NBUSCA	FACTOR	Razones por las que no busca empleo
ASALA	FACTOR	El empleo que busca es asalariado
EMBUS	FACTOR	Tipo de jornada en el empleo buscado
ITBU	FACTOR	Tiempo que lleva buscando empleo
DISP	FACTOR	Disponible para trabajar en 15 días
RZNDIS	FACTOR	Razones para no trabajar en 15 días
EMPANT	FACTOR	Si ha realizado antes algún trabajo
DTANT	NUMERIC	Meses desde que dejó su último empleo
OCUPA	FACTOR	Ocupación en su último empleo
ACTA	FACTOR	Actividad donde trabajaba
SITUA	FACTOR	Situación en su anterior trabajo
OFEMP	FACTOR	Inscrito en oficina de Empleo de la administración
SIDI1	FACTOR	Estudiante
SIDI2	FACTOR	Jubilado
SIDI3	FACTOR	Labores del hogar
SIDI4	FACTOR	Incapacitado permanente
SIDI5	FACTOR	Pensión distinta a la de jubilación
SIDI6	FACTOR	Actividad no remunerada
SIDI7	FACTOR	Otras situaciones
SIDAC1	FACTOR	Trabajando
SIDAC2	FACTOR	Buscando empleo
MUN	FACTOR	Lugar de residencia hace un año
REPAIRE	FACTOR	Provincia/Región de residencia anterior
AOI	FACTOR	Actividad económica OIT
CSE	FACTOR	Condición socioeconómica
FACTOREL	NUMERIC	Factor de elevación

Figura 37: Definición lógica de la tabla de observaciones.

4.3.3. Diseño detallado del sistema

En este apartado se realiza una descripción detallada de los elementos que encontramos en el sistema.

Recordando nuestro modelo conceptual definido anteriormente, donde encontramos las definiciones de los elementos Interfaz, Servidor y Base de datos, a continuación, se describen estos elementos en profundidad.

Para una aplicación reactiva de este estilo, la definición de la interfaz vendrá dada por los elementos de entrada manipulables por el usuario que provocan una reacción inmediata en la aplicación. De esta forma, en la tabla a continuación en la se describen dichos elementos.

Interfaz	
Atributo	Descripción
arules_train_btn	Botón que lanza la ejecución de las reglas de asociación.
arules_train_ciclo	Ciclo seleccionado para el algoritmo de reglas de asociación.
arules_train_confidence	Confianza mínima para el algoritmo de reglas de asociación.
arules_train_minlen	Longitud mínima de reglas a obtener.
arules_train_support	Soporte mínimo para las reglas de asociación.
arules_view_file	Fichero de reglas de asociación a explorar.
cfg_file	Fichero de actualización para incluir a la base de datos.
cfg_update_btn	Botón que lanza la ejecución de la actualización de los datos.
cluster_train_btn	Botón que lanza la ejecución de clustering.
cluster_train_ciclo	Ciclo seleccionado para el clustering.
cluster_train_groups	Numero de agrupaciones a obtener por el clustering.
cluster_view_file	Fichero de clustering a explorar.
multi_atributo	Atributos para realizar la comparación múltiple.
multi_btn	Botón que lanza la comparación de múltiples atributos.
multi_ciclo	Ciclo seleccionado para la comparación múltiple
pair_add_10perc	Inclusión de visualización de percentil 10.

pair_add_50perc	Inclusión de visualización de percentil 50.
pair_add_90perc	Inclusión de visualización de percentil 90.
pair_add_cov	Inclusión de covarianza en comparación por pares.
pair_add_jitter	Inclusión de vibración en comparación de pares.
pair_add_mean	Inclusión de la media en comparación de pares.
pair_alpha	Selección de nivel de transparencia en comparación de pares.
pair_atributo1	Atributo 1 seleccionado para la comparación de pares.
pair_atributo2	Atributo 2 seleccionado para la comparación de pares
pair_ciclo	Ciclo seleccionado para la comparación de pares
pair_group	Criterio de agrupación en la comparación de pares.
pair_limit_x	Filtrado en el eje x para la comparación de pares.
pair_limit_y	Filtrado en el eje y para la comparación de pares.
pair_scale	Selección de Escala de los ejes en la comparación por pares.
rpt_ciclo	Selección de ciclo para el generador de informes.
rpt_file	Fichero plantilla para la generación de informes.
single_atributo	Atributo seleccionado para la exploración de un atributo.
single_bins	Numero de bloques en la visualización de un atributo.
single_ciclo	Ciclo seleccionado para la exploración de un atributo.
single_group	Criterio de agrupación en la exploración de un atributo.
single_limit	Filtrado en el eje x de los elementos visualizados.
single_scale	Selección de escala de eje en la visualización de un atributo.
time_atributo	Atributo seleccionado en la comparación en el tiempo.

Figura 38: Descripción de elementos en la Interfaz

Capítulo 4

De la misma forma que con las entradas, el servidor estará definido por las distintas salidas que variaran de forma reactiva con las entradas del usuario. Aunque estas salidas tienen un reflejo de definición en la interfaz, es en el servidor donde se les da un comportamiento a través de scripts R.

Servidor	
Atributo	Descripción
arules_train_text	Resultado de entrenamiento de reglas de asociación.
arules_view_explore	Vista tabular de las reglas de asociación.
arules_view_graph	Visualización de grafo enlazado de reglas de asociación.
arules_view_plot	Gráfico de calor de reglas de asociación.
arules_view_text	Resumen de contenido de fichero de reglas de asociación.
cfg_db_summary	Listado de paquetes de actualización de la base de datos.
cluster_train_text	Resultado del entrenamiento del algoritmo de clustering.
cluster_view_text	Resumen textual del contenido de un fichero de clustering.
multi_plot	Salida grafica de comparación entre múltiples atributos.
pair_plot	Salida grafica de comparación de dos atributos.
rpt_generate	Descarga de informe generado por la herramienta.
single_box_plot	Gráfico de cajas en la exploración de una variable.
single_freq_plot	Polígono de frecuencias en la comparación de una variable.
single_hist_plot	Histograma en la pestaña de exploración de una variable.
single_text	Resumen textual de análisis de una variable.
time_plot	Visualización de evolución de una variable en el tiempo.

Figura 39: Descripción de elementos en el Servidor

Además, definimos otro elemento adicional en el sistema de acceso a la base de datos. Este elemento nos hace de interfaz entre la base de datos y la propia herramienta, simplificándonos el proceso de extracción de datos y abstrayéndolo de la implementación del sistema de gestión de base de datos. A continuación, se muestran descritos los métodos que provee dicho elemento.

Acceso a Base de Datos	
Atributo	Descripción
check_for_updates()	Comprueba si existen actualizaciones de base de datos.
getAttrDef()	Devuelve listado de atributos definidos en las observaciones.
getData(select, where)	Extrae de la base de datos en una table de tipo data.frame el resultado de la ejecución de una búsqueda.
getMapValues(attr_name)	Devuelve la definición de un atributo dado.
getSQL(sql_query)	Ejecuta y devuelve el resultado de una query SQL.
import_file_to_db(file)	Importación de un fichero ya preparado a la base de datos.
mapToString(dframe)	Traducción de datos a valores legibles.
time_to_hours(x)	Método auxiliar de conversión de unidades de tiempo.
update_database(file)	Prepara un fichero para incorporar a la base de datos.

Figura 40: Descripción de elementos de la clase Base de Datos

4.3.4. Interfaz con el usuario

La interfaz de la herramienta ha sido diseñada con los siguientes puntos clave:

- Minimizar la carga visual al usuario
- Funcionalidad intuitiva
- Facilidad de uso

Cada una de las vistas de la herramienta se compone principalmente de tres tipos de elementos:

- Menú de Navegación
- Elementos de parámetros de entrada
- Elementos de salida de datos o visualizaciones.

La siguiente figura 41 identifica cada uno de los elementos en una vista típica de la herramienta.

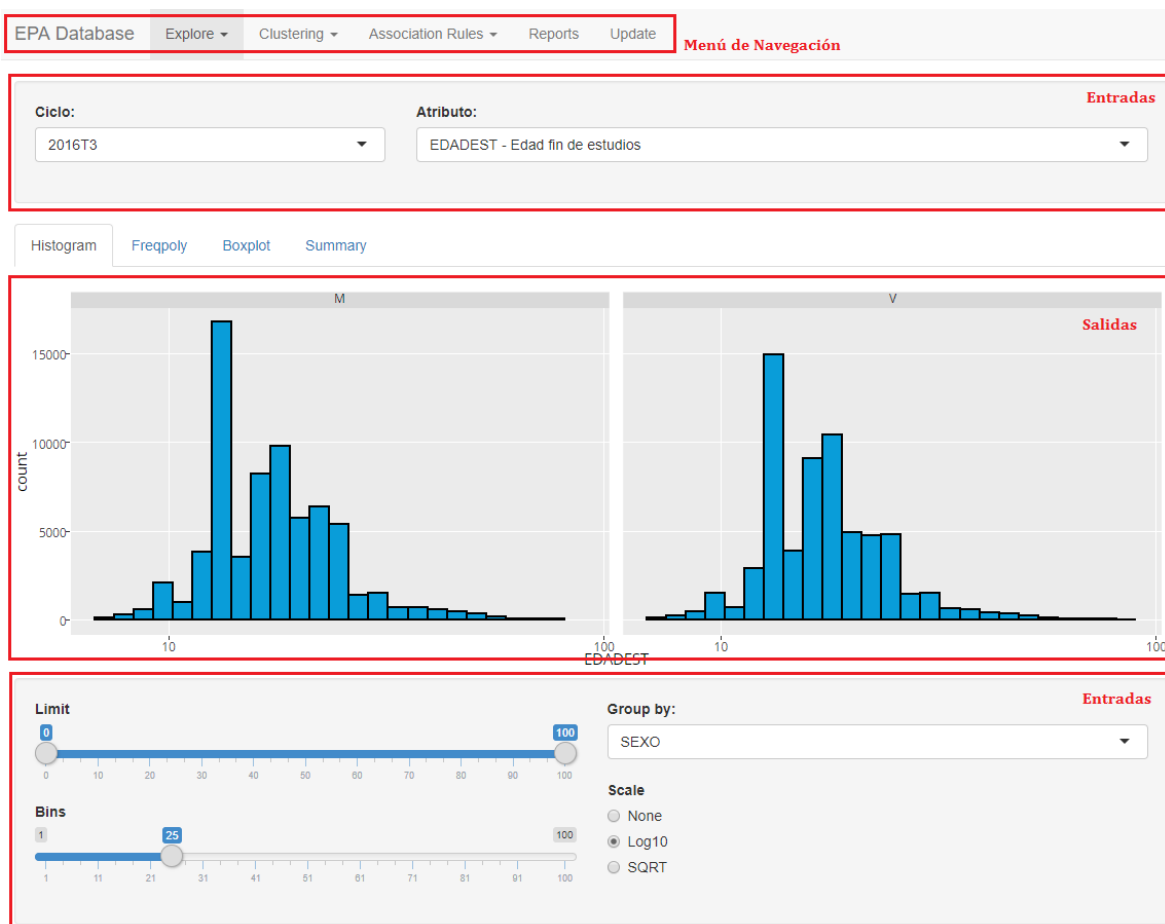


Figura 41: Menús de Navegación de EPA Explorer

4.4. Codificación

En este apartado pasamos a describir distintos elementos software que han sido necesarios para el desarrollo del proyecto.

4.4.1. Implementación

Lenguaje R

R es un entorno y lenguaje de programación con un enfoque al análisis estadístico. Es uno de los lenguajes más utilizados hoy día en el campo de la investigación por parte de la comunidad estadística, además de ser muy popular en otros campos como

- Minería de datos
- Investigación biomédica
- Bioinformática
- Matemáticas financieras.

Paquetes en R

Se han utilizado varios paquetes adicionales de R para cubrir distintas necesidades dentro del proyecto. El proyecto hace uso de gran cantidad de paquetes R como pueden ser las siguiente, por mencionar algunos de ellos:

- *rmarkdown*: Librería de renderizado de Markdown. Markdown es una sintaxis que puede ser convertida a XHTML o a otros muchos formatos.
- *RSQLite*: Incluye en R el motor de base de datos SQL y provee una interfaz compatible con DBI, la interfaz estándar para acceso a base de datos en R.
- *dplyr*: Librería genérica para manipulación de datos.
- *readr*: Provee una forma rápida y amigable de interpretación de datos tabulares encontrados en ficheros de texto plano como csv, tsv y fwf.

Capítulo 4

- *RCurl*: Librería de comunicación con servidores web haciendo uso de distintos protocolos como HTTP, FTP, FTPS o TFTP.
- *ggplot2*: Sistema de pintado de gráficos en R basado en la “gramática de los gráficos”. Produce complejos gráficos multicapa.
- *GGally*: Extensión de *ggplot2* que combina geometría de objetos con transformaciones de datos. Concretamente es usado para generación de comparación de variables por pares.
- *plotly*: Traduce gráficos generados por *ggplot2* a versiones interactivas para ser utilizadas en un navegador web, permitiendo descargas los gráficos, hacer zoom o marcar elementos con pulsaciones de ratón.
- *arules*: Provee la infraestructura para representar, manipular y analizar datos transaccionales y patrones. Provee interfaces a implementaciones C de algoritmos como Apriori y Eclat.
- *arulesViz*: Extensión del paquete *arules* con varias técnicas de visualización para reglas de asociación.
- *klaR*: Conjunto de métodos de algoritmos de clasificación y visualización.

Shiny:

Uno de los principales paquetes de los que se ha hecho uso en el proyecto es Shiny. Como se ha adelantado anteriormente, Shiny es un framework de desarrollo en R que facilita la generación de entornos web basados en la reactividad. Su enfoque está en la facilidad de uso y de mantenimiento.

SQLite

SQLite es un sistema de gestión de bases de datos relacional compatible con ACID. Funciona como una simple librería que se incorpora a la aplicación en cuestión, permitiendo que la propia herramienta gestione el acceso a la base de datos.

Esto permite ahorrar el mantenimiento de una topología habitual de sistema de sistema de gestión de bases de datos cliente-servidor.

Como consecuencia de esto, SQLite aporta ciertas características deseables para el proyecto en cuestión, como una base de datos formada por un único fichero fácil de trasladar e independiente de la plataforma, o una instalación donde no es necesaria configuración alguna de la base de datos.

4.4.2. Otros programas

RStudio

RStudio es un entorno de desarrollo (IDE) para su uso en el desarrollo de aplicaciones en lenguaje R. Incluye una consola interprete de R, un editor de sintaxis que apoya la ejecución de código, así como distintas herramientas para el trazado, la depuración y la gestión del espacio de trabajo.

Git

Git [17] es uno de los sistemas de control de configuración más utilizados hoy día. Ha sido seleccionado entre otros candidatos principalmente por su integración con RStudio y por estar familiarizado con el uso del mismo. Git se distribuye como software libre bajo licencia GNU GPL versión 2.

Otras Aplicaciones

Entre otras herramientas utilizadas en la elaboración del proyecto se ha hecho uso de la aplicación Gantt Project como herramienta para la gestión del proyecto en materia de esfuerzo y tiempo.

Además, se ha hecho uso de la herramienta **DIA** para el modelado y la generación de distintos diagramas que ilustran esta memoria de proyecto.

4.5. Pruebas y validación

En este apartado pasamos a describir detalles en la validación y verificación de nuestros requisitos de sistema, así como la estrategia general seguida para la elaboración de las pruebas.

4.5.1. Pruebas incrementales

Siguiendo la metodología de tipo incremental de desarrollo determinada para el desarrollo del proyecto, se ha ejecutado una fase de pruebas sobre cada una de las distintas iteraciones en el desarrollo del mismo.

Estas pruebas consistían principalmente en comprobar por casa caso de uso o vista que la ejecución de distintas entradas o configuraciones provocaban los efectos esperados como salidas en el navegador para el usuario o como transacciones en la base de datos.

A continuación, se describe el proceso de pruebas realizado en cada una de las iteraciones del proyecto.

- **Interpretación de los datos:** Durante esta fase inicial del proyecto, la actividad principal fue la de la interpretación de los datos extraídos del INE. Debido a esto, las pruebas consistieron en validar la consistencia de los datos en sus estructuras en memoria, comprobando si los datos extraídos correspondían a los datos reales. Parte de dichas pruebas consistieron en calcular ciertas métricas o indicadores que conociamos como válidas por las notas de prensa publicadas de la EPA.
- **Diseño de la base de datos:** En esta fase se detecta la necesidad de hacer uso de una base de datos local para los datos. Las pruebas realizadas en esta fase consisten en repetir las pruebas de la primera fase, pero tomando como entrada esta vez los datos almacenados en disco. Además, se realizaron pruebas de rendimiento para validar la necesidad del almacenamiento local, así como comprobación tipos asignados para cada atributo además de su lectura.

- **Análisis Exploratorio de Datos:** Esta ha sido una de las fases más extensas y principales del proyecto, por lo que las pruebas realizadas fueron más estrictas que en fases anteriores. En esta fase las pruebas en su gran mayoría se han basado en comprobar que las visualizaciones obtenidas correspondían a resultados esperados conocidos.
- **Exportador Documental:** De forma similar a la fase anterior, para comprobar el funcionamiento del exportador documental se han realizado pruebas en distintos formatos que han sido contrastados con información ya conocida supuesta como válida. Esta información ha sido extraída de las notas de prensa emitidas por el INE.
- **Actualización de Base de Datos:** En esta fase del desarrollo donde se implementa la captación de nuevos datos remotos en la web de la INE se comprueba cómo reacciona el sistema ante una falta de respuesta o conexión. Además, se realizan pruebas de consistencia de los datos, así como su correcta importación en la base de datos.
- **Reglas de Asociación y Clustering:** Las pruebas realizadas en esta iteración del proyecto se han basado principalmente en comprobaciones de cara al rendimiento del sistema. Siendo crítico en una herramienta de estas características el que el usuario no tenga que esperar un tiempo no determinado previamente, se prueba como afectan la alteración de ciertos parámetros en el rendimiento de los algoritmos seleccionados.

4.5.2. Entorno de pruebas

Una vez finalizado el desarrollo de las distintas iteraciones y sus pruebas unitarias, se somete el sistema al completo a unas pruebas de integración de la herramienta. Aquí se comprueba que todos los elementos unitarios que componen el software, funcionan juntos correctamente probándolos en grupo.

A continuación, se muestran en una serie de tablas el protocolo de pruebas simple utilizado para la validación de cada caso de uso individual.

Caso de uso Análisis Exploratorio de Datos de Una variable	
Acción a Realizar	Criterio de Verificación
Seleccionar ciclo y atributo para filtrar.	La visualización se ajusta al filtro seleccionado.
Seleccionar el tipo de visualización	La visualización cambia para mostrar el tipo de visualización seleccionado.
Alterar parámetros de la visualización	La visualización se altera para mostrar los parámetros de visualización seleccionados
Manipular la visualización con pulsaciones del ratón	La visualización se ajusta acorde con la selección realizada de forma interactiva.
Pulsar el botón “Download Plot”	Se genera una imagen para descargar por el navegador en formato PNG

Figura 42: Pruebas caso de uso Análisis Exploratorio de Datos de Una variable

Caso de uso Análisis Exploratorio de Datos de Dos variables	
Acción a Realizar	Criterio de Verificación
Seleccionar ciclo y los atributos para filtrar.	La visualización se ajusta al filtro seleccionado.
Seleccionar el tipo de visualización	La visualización cambia para mostrar el tipo de visualización seleccionado.
Alterar parámetros de la visualización	La visualización se altera para mostrar los parámetros de visualización seleccionados
Manipular la visualización con pulsaciones del ratón	La visualización se ajusta acorde con la selección realizada de forma interactiva.
Pulsar el botón “Download Plot”	Se genera una imagen para descargar por el navegador en formato PNG

Figura 43: Pruebas caso de uso Análisis Exploratorio de Datos de Dos variables

Caso de uso Análisis Exploratorio de Datos de Múltiples variables	
Acción a Realizar	Criterio de Verificación
No se selecciona el conjunto de atributos a mostrar.	Error: "No seleccionado atributos para comparar"
Seleccionar ciclo y el conjunto de atributos para filtrar.	La herramienta muestra el conjunto de atributos seleccionado para comparar.
Ejecutar generación de visualización.	Se presenta ante el usuario la comparación entre los atributos seleccionados previamente.
Manipular la visualización con pulsaciones del ratón	La visualización se ajusta acorde con la selección realizada de forma interactiva.
Pulsar el botón "Download Plot"	Se genera una imagen para descargar por el navegador en formato PNG

Figura 44: Pruebas caso de uso Análisis Exploratorio de Datos de Múltiples variable

Caso de uso Análisis Exploratorio de Datos de Serie Temporal	
Acción a Realizar	Criterio de Verificación
Seleccionar atributo para mostrar.	La visualización se ajusta al atributo seleccionado.
Manipular la visualización con pulsaciones del ratón	La visualización se ajusta acorde con la selección realizada de forma interactiva.
Pulsar el botón "Download Plot"	Se genera una imagen para descargar por el navegador en formato PNG

Figura 45: Pruebas caso de uso Análisis Exploratorio de Datos de Serie Temporal

Caso de uso Entrenamiento Clustering	
Acción a Realizar	Criterio de Verificación
Seleccionar el ciclo a analizar	La herramienta muestra el ciclo seleccionado para analizar.
Configurar parámetros adicionales para el algoritmo	La herramienta muestra los valores seleccionados para los parámetros adicionales para el algoritmo.
Lanzar la ejecución del algoritmo con los parámetros seleccionados.	La herramienta muestra información básica de la ejecución del algoritmo. Se genera un fichero bajo la carpeta 'clúster' con una marca de tiempo del momento de la ejecución.

Figura 46: Pruebas caso de uso Entrenamiento Clustering

Caso de uso Visualizar Clustering	
Acción a Realizar	Criterio de Verificación
No existen ejecuciones anteriores de algoritmo	Error: "No existen ejecuciones previas del algoritmo"
Seleccionar un fichero con resultados de clustering previos	La herramienta muestra información básica de la ejecución del algoritmo.
Selecciona el tipo de visualización deseado.	La herramienta muestra una visualización del fichero seleccionado siguiente el tipo de visualización dado por el usuario.

Figura 47: Pruebas caso de uso Visualizar Clustering

Caso de uso Entrenamiento Reglas de Asociación	
Acción a Realizar	Criterio de Verificación
Seleccionar el ciclo a analizar	La herramienta muestra el ciclo seleccionado para analizar.
Configurar parámetros adicionales para el algoritmo	La herramienta muestra los valores seleccionados para los parámetros adicionales para el algoritmo.
Lanzar la ejecución del algoritmo con los parámetros seleccionados.	La herramienta muestra información básica de la ejecución del algoritmo. Se genera un fichero bajo la carpeta 'arules' con una marca de tiempo del momento de la ejecución.

Figura 48: Pruebas caso de uso Entrenamiento Reglas de Asociación

Caso de uso Visualizar Reglas de Asociación	
Acción a Realizar	Criterio de Verificación
No existen ejecuciones anteriores de algoritmo	Error: "No existen ejecuciones previas del algoritmo"
Seleccionar un fichero con resultados de reglas de asociación previos	La herramienta muestra información básica de la ejecución del algoritmo.
Selecciona el tipo de visualización deseado.	La herramienta muestra una visualización del fichero seleccionado siguiendo el tipo de visualización dado por el usuario.

Figura 49: Pruebas caso de uso Visualizar Reglas de Asociación

Caso de uso Generación de Informes	
Acción a Realizar	Criterio de Verificación
No existen plantillas	Error: "No se han cargado plantillas"
Seleccionar la plantilla	La herramienta muestra la plantilla seleccionada.
Configurar parámetros del informe	La herramienta muestra los valores seleccionados para la generación del informe deseado.
Ejecutar la generación del informe.	Se genera un informe de acuerdo a la plantilla y parámetros seleccionados y en el formato adecuado.

Figura 50: Pruebas caso de uso Generación de Informes

Caso de uso Actualización de base de datos	
Acción a Realizar	Criterio de Verificación
No hay conexión al FTP del INE.	Error: "No existe conexión al repositorio de actualizaciones remoto"
No existe o no se selecciona paquete para actualizar.	Error: "No se ha seleccionado ningún paquete de actualización de datos"
Se lanza la ejecución de la actualización de datos	La herramienta tiene ahora un nuevo ciclo seleccionable desde todas las vistas.

Figura 51: Pruebas caso de uso Actualización de base de datos

Capítulo 5. Conclusiones

5.1 Retrospectiva

Una vez finalizado el proyecto, volvemos a visitar los puntos que anteriormente nos marcamos como objetivos del mismo. En este apartado intentaremos valorar cual ha sido nuestra experiencia en conseguir los objetivos marcados, así como una valoración del grado de cumplimiento de dichos objetivos.

Inicialmente nos marcamos el conseguir obtener una aplicación o herramienta capaz de servir como soporte al análisis de datos sobre los datos extraídos de la EPA e incluir ciertas capacidades de minería de datos. Creo entonces que podemos afirmar que el objetivo global ha sido conseguido.

De forma adicional, hemos podido dotar a la herramienta de ciertas capacidades interesantes, aunque no identificadas anteriormente como objetivos como pueden ser:

- Generación de informes, dado a la capacidad y explotación del paquete Rmarkdown y la disponibilidad de los datos con los que trabajamos, se presentó como una opción viable el incluir esta funcionalidad en nuestro proyecto.
- Actualización automática de los datos, permitiendo al usuario que con un par de clics mantenga su conjunto de datos actualizado con los últimos publicados por el INE.

Podemos decir también que gracias al uso y la explotación de Shiny hemos cumplido con creces los requisitos de usabilidad de la herramienta, donde tenemos una interfaz realmente simple de usar y manipular. Una interfaz reactiva, muy resistente a la introducción de posibles errores.

5.2 Futuras ampliaciones

Debido a la naturaleza modular del proyecto, el modelo de desarrollo seguido, además de la estructura de implementación que se ha seguido en el mismo, sería muy simple implementar nuevas funcionalidades a la herramienta.

Algunas de estas posibles ampliaciones a la herramienta podrían ser:

- Generalizar a otras fuentes de datos. Aunque el proyecto desarrollado se centre en los datos obtenidos de la EPA, sería posible ampliar la idea de este proyecto para que pudiera ser aplicable a datos provenientes de otras fuentes.
- Inclusión de otros algoritmos de aprendizaje máquina. Teniendo un conocimiento experto de los datos tratados, sería posible incluir paquetes adicionales a nuestra herramienta para lanzar otros tipos de minería de datos que pudiéramos creer conveniente.
- Inclusión de más tipos de informes. Por la implementación que se ha realizado en el sistema de generación de documentos sería muy simple incluir nuevas plantillas de generación para sacar todo tipo de informes sobre los datos almacenados.
- Sistema de gestión de usuarios. Sería interesante en un futuro añadir la capacidad de que un usuario se registrase en la herramienta, con la idea de poder mantener filtros de los datos bajo su perfil personal, así como una gestión de roles para por ejemplo limitar el acceso a la pestaña de actualización de datos.
- Gestión de proceso en paralelo. Existe trabajo en R dedicado a la gestión de herramientas multiproceso. Esto permitiría el entrenamiento de varios algoritmos de forma simultánea en procesadores multihilo. Desgraciadamente dicha funcionalidad no está disponible en Windows, donde una instancia de R es un proceso mono hilo, pero sería muy simple hacer el traslado a Linux para conseguir esta funcionalidad.

5.3 Valoraciones personales

Debido a mi trayectoria laboral, consistente principalmente en administración de herramientas de ciclo de vida en particular (e informático para todo en general), estoy habituado a desarrollar pequeñas herramientas o utilidades para uso de otros usuarios, no necesariamente compañeros de profesión. También me he enfrentado en muchas ocasiones con problemas derivados de tratamiento e interpretación de ficheros de datos.

Aun con esta experiencia previa en desarrollos de alguna forma similares, no siempre se han utilizado las herramientas más adecuadas para ello, o no se han explotado lo suficiente.

Este proyecto me ha ayudado a encontrar herramientas impresionantes como R o Shiny, creadas por la comunidad y de libre acceso, que explotadas adecuadamente nos proporcionan una potencia y versatilidad realmente impresionantes. Adicionalmente me ha ayudado a recordar y a valorar el valor de la información, refrescando conocimientos que cada vez más están en la boca de todos, formando parte crucial en campos como el Internet de las Cosas, el análisis de datos.

Por si no fuera poco, ya estoy viendo clara aplicabilidad de todo lo aprendido en mi trabajo, desarrollo de sistemas de combate en Navantia, donde pienso proponer el uso de este conjunto de herramientas y técnicas para satisfacer la necesidad de análisis pos-misión de la próxima fragata para la Armada Española.

En conclusión, estoy contento con el trabajo realizado y valoro de forma positiva el desarrollo del proyecto. Creo que hemos hecho un buen trabajo.

Bibliografía

Referencias

[1] Sitio oficial de la Encuesta de Población Activa

http://www.ine.es/prensa/epa_prensa.htm

[2] Sitio oficial del Instituto Nacional de Estadística

<http://www.ine.es>

[3] Descripción de Metodología vigente de la EPA

<http://www.ine.es/daco/daco43/resumetepa.pdf>

[4] Pagina de descarga de aplicación PC-Axis

<http://www.ine.es/ss/Satellite?c=Page&pagename=ProductosYServicios%2FPYSLayo ut&cid=1254735116596>

[5] Sitio de descarga de ficheros de Microdatos de la EPA

http://www.ine.es/dyngs/INEbase/es/operacion.htm?c=Estadistica_C&cid=1254736176918&menu=resultados&secc=1254736030639&idp=1254735976595

[6] Sitio oficial del Proyecto de R

<https://cran.r-project.org/>

[7] Sitio oficial del framework Shiny

<https://shiny.rstudio.com/>

[8] Sitio oficial del entorno de desarrollo RStudio

<https://www.rstudio.com/>

[9] Pagina de descarga de la aplicación Gantt Project

<http://www.ganttproject.biz/>

[10] Pagina de descarga de la aplicación DIA

<http://dia-installer.de/>

[11] Procesos Software – Modelo Iterativo

<https://procesossoftware.wikispaces.com/Modelo+Iterativo>

Bibliografía y referencias electrónicas

[12] Recurso de documentación sobre metodologías de desarrollo incremental
<https://proyectosagiles.org/desarrollo-iterativo-incremental/>

[13] Huang, Zhexue
Extensions to the k-Means Algorithm for Clustering Data Sets with Categorical Values
Kluwer Academic Publishers, 1998.

[14] Convenio colectivo nacional de empresas de ingeniería (18 de enero de 2017)
<https://www.boe.es/boe/dias/2017/01/18/pdfs/BOE-A-2017-542.pdf>

[15] Sitio de consulta sobre el patrón de diseño de modelo vista controlador.
<https://desarrolloweb.com/articulos/que-es-mvc.html>

[16] Sitio oficial de SQLite
<https://www.sqlite.org/>

[17] Sitio oficial de sistema de control de versiones Git
<https://git-scm.com/>

[18] Lantz, Brett
Machine Learning with R
Packt Publishing, Ed. 1, 2013.

[19] Rodríguez Pacheco, Erik
Unsupervised Learning with R
Packt Publishing, Ed. 1, 2015.

[20] Curso Online, Data Analysis with R
<https://in.udacity.com/course/data-analysis-with-r--ud651/>
Udacity

.

Apéndice A: GNU Free Documentation License v 1.3



Version 1.3, 3 November 2008 Copyright © 2000, 2001, 2002, 2007, 2008 Free Software Foundation, Inc.³

Everyone is permitted to copy and distribute verbatim copies of this license document, but changing it is not allowed. [B]

PREAMBLE

The purpose of this License is to make a manual, textbook, or other functional and useful document "free" in the sense of freedom: to assure everyone the effective freedom to copy and redistribute it, with or without modifying it, either commercially or noncommercially. Secondly, this License preserves for the author and publisher a way to get credit for their work, while not being considered responsible for modifications made by others.

This License is a kind of "copyleft", which means that derivative works of the document must themselves be free in the same sense. It complements the GNU General Public License, which is a copyleft license designed for free software.

We have designed this License in order to use it for manuals for free software, because free software needs free documentation: a free program should come with manuals providing the same freedoms that the software does. But this License is not limited to software manuals; it can be used for any textual work, regardless of subject matter or whether it is published as a printed book. We recommend this License principally for works whose purpose is instruction or reference.

³ <http://fsf.org/>

APPLICABILITY AND DEFINITIONS

This License applies to any manual or other work, in any medium, that contains a notice placed by the copyright holder saying it can be distributed under the terms of this License. Such a notice grants a world-wide, royalty-free license, unlimited in duration, to use that work under the conditions stated herein. The "Document", below, refers to any such manual or work. Any member of the public is a licensee, and is addressed as "you". You accept the license if you copy, modify or distribute the work in a way requiring permission under copyright law.

A "Modified Version" of the Document means any work containing the Document or a portion of it, either copied verbatim, or with modifications and/or translated into another language.

A "Secondary Section" is a named appendix or a front-matter section of the Document that deals exclusively with the relationship of the publishers or authors of the Document to the Document's overall subject (or to related matters) and contains nothing that could fall directly within that overall subject. (Thus, if the Document is in part a textbook of mathematics, a Secondary Section may not explain any mathematics.) The relationship could be a matter of historical connection with the subject or with related matters, or of legal, commercial, philosophical, ethical or political position regarding them.

The "Invariant Sections" are certain Secondary Sections whose titles are designated, as being those of Invariant Sections, in the notice that says that the Document is released under this License. If a section does not fit the above definition of Secondary then it is not allowed to be designated as Invariant. The Document may contain zero Invariant Sections. If the Document does not identify any Invariant Sections then there are none.

The "Cover Texts" are certain short passages of text that are listed, as Front-Cover Texts or Back-Cover Texts, in the notice that says that the Document is released under this License. A Front-Cover Text may be at most 5 words, and a Back-Cover Text may be at most 25 words.

A "Transparent" copy of the Document means a machine-readable copy, represented in a format whose specification is available to the general public, that is suitable for revising the document straightforwardly with generic text editors or (for images composed of pixels) generic paint programs or (for drawings) some widely available drawing editor, and that is suitable for input to text formatters or for automatic translation to a variety of formats suitable for input to text formatters. A copy made in an otherwise Transparent

file format whose markup, or absence of markup, has been arranged to thwart or discourage subsequent modification by readers is not Transparent. An image format is not Transparent if used for any substantial amount of text. A copy that is not "Transparent" is called "Opaque".

Examples of suitable formats for Transparent copies include plain ASCII without markup, Texinfo input format, LaTeX input format, SGML or XML using a publicly available DTD, and standard-conforming simple HTML, PostScript or PDF designed for human modification. Examples of transparent image formats include PNG, XCF and JPG. Opaque formats include proprietary formats that can be read and edited only by proprietary word processors, SGML or XML for which the DTD and/or processing tools are not generally available, and the machine-generated HTML, PostScript or PDF produced by some word processors for output purposes only.

The "Title Page" means, for a printed book, the title page itself, plus such following pages as are needed to hold, legibly, the material this License requires to appear in the title page. For works in formats which do not have any title page as such, "Title Page" means the text near the most prominent appearance of the work's title, preceding the beginning of the body of the text.

The "publisher" means any person or entity that distributes copies of the Document to the public.

A section "Entitled XYZ" means a named subunit of the Document whose title either is precisely XYZ or contains XYZ in parentheses following text that translates XYZ in another language. (Here XYZ stands for a specific section name mentioned below, such as "Acknowledgements", "Dedications", "Endorsements", or "History".) To "Preserve the Title" of such a section when you modify the Document means that it remains a section "Entitled XYZ" according to this definition.

The Document may include Warranty Disclaimers next to the notice which states that this License applies to the Document. These Warranty Disclaimers are considered to be included by reference in this License, but only as regards disclaiming warranties: any other implication that these Warranty Disclaimers may have is void and has no effect on the meaning of this License.

VERBATIM COPYING

You may copy and distribute the Document in any medium, either commercially or noncommercially, provided that this License, the copyright notices, and the license notice saying this License applies to the Document are reproduced in all copies, and that you add no other conditions whatsoever to those of this License. You may not use technical measures to obstruct or control the reading or further copying of the copies you make or distribute. However, you may accept compensation in exchange for copies. If you distribute a large enough number of copies you must also follow the conditions in section 3.

You may also lend copies, under the same conditions stated above, and you may publicly display copies.

COPYING IN QUANTITY

If you publish printed copies (or copies in media that commonly have printed covers) of the Document, numbering more than 100, and the Document's license notice requires Cover Texts, you must enclose the copies in covers that carry, clearly and legibly, all these Cover Texts: Front-Cover Texts on the front cover, and Back-Cover Texts on the back cover. Both covers must also clearly and legibly identify you as the publisher of these copies. The front cover must present the full title with all words of the title equally prominent and visible. You may add other material on the covers in addition. Copying with changes limited to the covers, as long as they preserve the title of the Document and satisfy these conditions, can be treated as verbatim copying in other respects.

If the required texts for either cover are too voluminous to fit legibly, you should put the first ones listed (as many as fit reasonably) on the actual cover, and continue the rest onto adjacent pages.

If you publish or distribute Opaque copies of the Document numbering more than 100, you must either include a machine-readable Transparent copy along with each Opaque copy, or state in or with each Opaque copy a computer-network location from which the general network-using public has access to download using public-standard network protocols a complete Transparent copy of the Document, free of added material. If you use the latter option, you must take reasonably prudent steps, when you begin distribution of Opaque copies in quantity, to ensure that this Transparent copy will remain thus accessible at the stated location until at least one year after the last time you

distribute an Opaque copy (directly or through your agents or retailers) of that edition to the public.

It is requested, but not required, that you contact the authors of the Document well before redistributing any large number of copies, to give them a chance to provide you with an updated version of the Document.

MODIFICATIONS

You may copy and distribute a Modified Version of the Document under the conditions of sections 2 and 3 above, provided that you release the Modified Version under precisely this License, with the Modified Version filling the role of the Document, thus licensing distribution and modification of the Modified Version to whoever possesses a copy of it. In addition, you must do these things in the Modified Version:

- A. Use in the Title Page (and on the covers, if any) a title distinct from that of the Document, and from those of previous versions (which should, if there were any, be listed in the History section of the Document). You may use the same title as a previous version if the original publisher of that version gives permission.
- B. List on the Title Page, as authors, one or more persons or entities responsible for authorship of the modifications in the Modified Version, together with at least five of the principal authors of the Document (all of its principal authors, if it has fewer than five), unless they release you from this requirement.
- C. State on the Title page the name of the publisher of the Modified Version, as the publisher.
- D. Preserve all the copyright notices of the Document.
- E. Add an appropriate copyright notice for your modifications adjacent to the other copyright notices.
- F. Include, immediately after the copyright notices, a license notice giving the public permission to use the Modified Version under the terms of this License, in the form shown in the Addendum below.
- G. Preserve in that license notice the full lists of Invariant Sections and required Cover Texts given in the Document's license notice.
- H. Include an unaltered copy of this License.

Apéndice A

I. Preserve the section Entitled "History", Preserve its Title, and add to it an item stating at least the title, year, new authors, and publisher of the Modified Version as given on the Title Page. If there is no section Entitled "History" in the Document, create one stating the title, year, authors, and publisher of the Document as given on its Title Page, then add an item describing the Modified Version as stated in the previous sentence.

J. Preserve the network location, if any, given in the Document for public access to a Transparent copy of the Document, and likewise the network locations given in the Document for previous versions it was based on. These may be placed in the "History" section. You may omit a network location for a work that was published at least four years before the Document itself, or if the original publisher of the version it refers to gives permission.

K. For any section Entitled "Acknowledgements" or "Dedications", Preserve the Title of the section, and preserve in the section all the substance and tone of each of the contributor acknowledgements and/or dedications given therein.

L. Preserve all the Invariant Sections of the Document, unaltered in their text and in their titles. Section numbers or the equivalent are not considered part of the section titles.

M. Delete any section Entitled "Endorsements". Such a section may not be included in the Modified Version.

N. Do not retitle any existing section to be Entitled "Endorsements" or to conflict in title with any Invariant Section.

O. Preserve any Warranty Disclaimers.

If the Modified Version includes new front-matter sections or appendices that qualify as Secondary Sections and contain no material copied from the Document, you may at your option designate some or all of these sections as invariant. To do this, add their titles to the list of Invariant Sections in the Modified Version's license notice. These titles must be distinct from any other section titles.

You may add a section Entitled "Endorsements", provided it contains nothing but endorsements of your Modified Version by various parties—for example, statements of peer review or that the text has been approved by an organization as the authoritative definition of a standard.

You may add a passage of up to five words as a Front-Cover Text, and a passage of up to 25 words as a Back-Cover Text, to the end of the list of Cover Texts in the Modified Version. Only one passage of Front-Cover Text and one of Back-Cover Text may be added

by (or through arrangements made by) any one entity. If the Document already includes a cover text for the same cover, previously added by you or by arrangement made by the same entity you are acting on behalf of, you may not add another; but you may replace the old one, on explicit permission from the previous publisher that added the old one. The author(s) and publisher(s) of the Document do not by this License give permission to use their names for publicity for or to assert or imply endorsement of any Modified Version.

COMBINING DOCUMENTS

You may combine the Document with other documents released under this License, under the terms defined in section 4 above for modified versions, provided that you include in the combination all of the Invariant Sections of all of the original documents, unmodified, and list them all as Invariant Sections of your combined work in its license notice, and that you preserve all their Warranty Disclaimers.

The combined work need only contain one copy of this License, and multiple identical Invariant Sections may be replaced with a single copy. If there are multiple Invariant Sections with the same name but different contents, make the title of each such section unique by adding at the end of it, in parentheses, the name of the original author or publisher of that section if known, or else a unique number. Make the same adjustment to the section titles in the list of Invariant Sections in the license notice of the combined work.

In the combination, you must combine any sections Entitled "History" in the various original documents, forming one section Entitled "History"; likewise combine any sections Entitled "Acknowledgements", and any sections Entitled "Dedications". You must delete all sections Entitled "Endorsements".

COLLECTIONS OF DOCUMENTS

You may make a collection consisting of the Document and other documents released under this License, and replace the individual copies of this License in the various documents with a single copy that is included in the collection, provided that you follow

the rules of this License for verbatim copying of each of the documents in all other respects.

You may extract a single document from such a collection, and distribute it individually under this License, provided you insert a copy of this License into the extracted document, and follow this License in all other respects regarding verbatim copying of that document.

AGGREGATION WITH INDEPENDENT WORKS

A compilation of the Document or its derivatives with other separate and independent documents or works, in or on a volume of a storage or distribution medium, is called an "aggregate" if the copyright resulting from the compilation is not used to limit the legal rights of the compilation's users beyond what the individual works permit. When the Document is included in an aggregate, this License does not apply to the other works in the aggregate which are not themselves derivative works of the Document.

If the Cover Text requirement of section 3 is applicable to these copies of the Document, then if the Document is less than one half of the entire aggregate, the Document's Cover Texts may be placed on covers that bracket the Document within the aggregate, or the electronic equivalent of covers if the Document is in electronic form. Otherwise they must appear on printed covers that bracket the whole aggregate.

TRANSLATION

Translation is considered a kind of modification, so you may distribute translations of the Document under the terms of section 4. Replacing Invariant Sections with translations requires special permission from their copyright holders, but you may include translations of some or all Invariant Sections in addition to the original versions of these Invariant Sections. You may include a translation of this License, and all the license notices in the Document, and any Warranty Disclaimers, provided that you also include the original English version of this License and the original versions of those notices and disclaimers. In case of a disagreement between the translation and the original version of this License or a notice or disclaimer, the original version will prevail.

If a section in the Document is Entitled "Acknowledgements", "Dedications", or "History", the requirement (section 4) to Preserve its Title (section 1) will typically require changing the actual title.

TERMINATION

You may not copy, modify, sublicense, or distribute the Document except as expressly provided under this License. Any attempt otherwise to copy, modify, sublicense, or distribute it is void, and will automatically terminate your rights under this License.

However, if you cease all violation of this License, then your license from a particular copyright holder is reinstated (a) provisionally, unless and until the copyright holder explicitly and finally terminates your license, and (b) permanently, if the copyright holder fails to notify you of the violation by some reasonable means prior to 60 days after the cessation.

Moreover, your license from a particular copyright holder is reinstated permanently if the copyright holder notifies you of the violation by some reasonable means, this is the first time you have received notice of violation of this License (for any work) from that copyright holder, and you cure the violation prior to 30 days after your receipt of the notice.

Termination of your rights under this section does not terminate the licenses of parties who have received copies or rights from you under this License. If your rights have been terminated and not permanently reinstated, receipt of a copy of some or all of the same material does not give you any rights to use it.

FUTURE REVISIONS OF THIS LICENSE

The Free Software Foundation may publish new, revised versions of the GNU Free Documentation License from time to time. Such new versions will be similar in spirit to the present version, but may differ in detail to address new problems or concerns. See <http://www.gnu.org/copyleft/>.

Each version of the License is given a distinguishing version number. If the Document specifies that a particular numbered version of this License "or any later version" applies to it, you have the option of following the terms and conditions either of that specified

version or of any later version that has been published (not as a draft) by the Free Software Foundation. If the Document does not specify a version number of this License, you may choose any version ever published (not as a draft) by the Free Software Foundation. If the Document specifies that a proxy can decide which future versions of this License can be used, that proxy's public statement of acceptance of a version permanently authorizes you to choose that version for the Document.

RELICENSING

"Massive Multiauthor Collaboration Site" (or "MMC Site") means any World Wide Web server that publishes copyrightable works and also provides prominent facilities for anybody to edit those works. A public wiki that anybody can edit is an example of such a server. A "Massive Multiauthor Collaboration" (or "MMC") contained in the site means any set of copyrightable works thus published on the MMC site.

"CC-BY-SA" means the Creative Commons Attribution-Share Alike 3.0 license published by Creative Commons Corporation, a not-for-profit corporation with a principal place of business in San Francisco, California, as well as future copyleft versions of that license published by that same organization.

"Incorporate" means to publish or republish a Document, in whole or in part, as part of another Document.

An MMC is "eligible for relicensing" if it is licensed under this License, and if all works that were first published under this License somewhere other than this MMC, and subsequently incorporated in whole or in part into the MMC, (1) had no cover texts or invariant sections, and (2) were thus incorporated prior to November 1, 2008.

The operator of an MMC Site may republish an MMC contained in the site under CC-BY-SA on the same site at any time before August 1, 2009, provided the MMC is eligible for relicensing.

ADDENDUM: How to use this License for your documents

To use this License in a document you have written, include a copy of the License in the document and put the following copyright and license notices just after the title page:

Copyright (C) YEAR YOUR NAME.

Permission is granted to copy, distribute and/or modify this document

under the terms of the GNU Free Documentation License, Version 1.3
or any later version published by the Free Software Foundation;
with no Invariant Sections, no Front-Cover Texts, and no Back-Cover
Texts.

A copy of the license is included in the section entitled "GNU
Free Documentation License".

If you have Invariant Sections, Front-Cover Texts and Back-Cover Texts, replace the
"with ... Texts." line with this:

with the Invariant Sections being LIST THEIR TITLES, with the
Front-Cover Texts being LIST, and with the Back-Cover Texts being LIST.

If you have Invariant Sections without Cover Texts, or some other combination of the
three, merge those two alternatives to suit the situation.

If your document contains nontrivial examples of program code, we recommend
releasing these examples in parallel under your choice of free software license, such as
the GNU General Public License, to permit their use in free software.

