Escuela Superior de Ingeniería
Dirección

C/ Chile, 1. 11002 Cádiz
Tlf. 956015100
Fax 956015101

Escuela Superior de Ingeniería

PROYECTO FIN DE CARRERA

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

Resumen del Proyecto Fin de Carrera
Software de preparación, procesado y análisis de datos de la EPA (Encuesta de Población Activa)

José Saúco Delgado, [1]Elisa Guerrero Vázquez, Andrés Yáñez Escolano

José López Fernández, 57 – Esc 3 3º C, Puerto Real (Cádiz) CP: 11510, 856212742/639328543 jose.saucodelgado@alum.uca.es

[1]Edificio Policlínico
Doctor Marañón, 3
11002 Cádiz

Extracto
El objetivo del 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. La herramienta permite una amplia funcionalidad sobre los datos de los distintos ejercicios de la EPA, como pueden ser análisis exploratorio de datos, técnicas de aprendizaje maquina no supervisado o generación de informes.

Palabras Clave: encuesta de población activa, lenguaje r, análisis de datos.

1. Introducción

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).

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.

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 [3]. 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.

1.1. 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 [5] 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.

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.

2. Descripción General

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.

3. Metodología de Desarrollo

Analizando el proyecto a desarrollar se determinó que seguir un modelo de desarrollo software de tipo incremental, 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 [6].

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. En cada iteración seguirá de nuevo el proceso completo, incrementando el conjunto de funcionalidades entregadas al cliente hasta completarla al final del desarrollo.

3.1. Planificación

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 ***Tabla 1*** 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.

Tabla 1. *Tabla días estimados y reales*

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

4.1. Interfaz

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 1*** muestra un ejemplo de dicho menú principal.

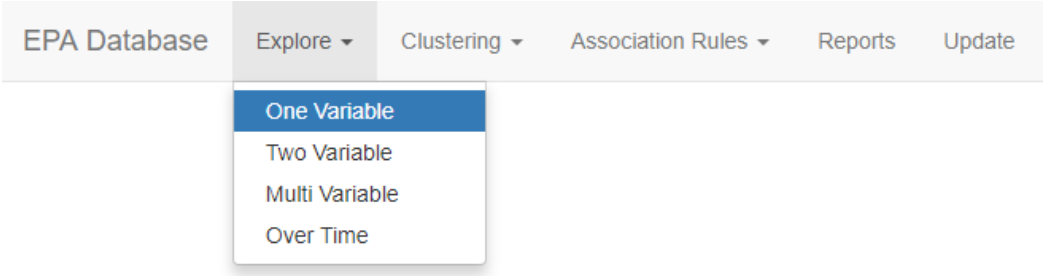

Figura 1. *Menús de Navegación de EPA Explorer*

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

1. Explorar
    - Una variable
    - Dos variables
    - Múltiples variables
    - Serie Temporal
2. Agrupación
    - Entrenamiento
    - Ver
3. Reglas de Asociación
    - Entrenamiento
    - Ver
4. Informes
5. Actualización

El usuario interactuara 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 2*** se muestra un ejemplo del aspecto que tendrá una vista de ejemplo de la aplicación.

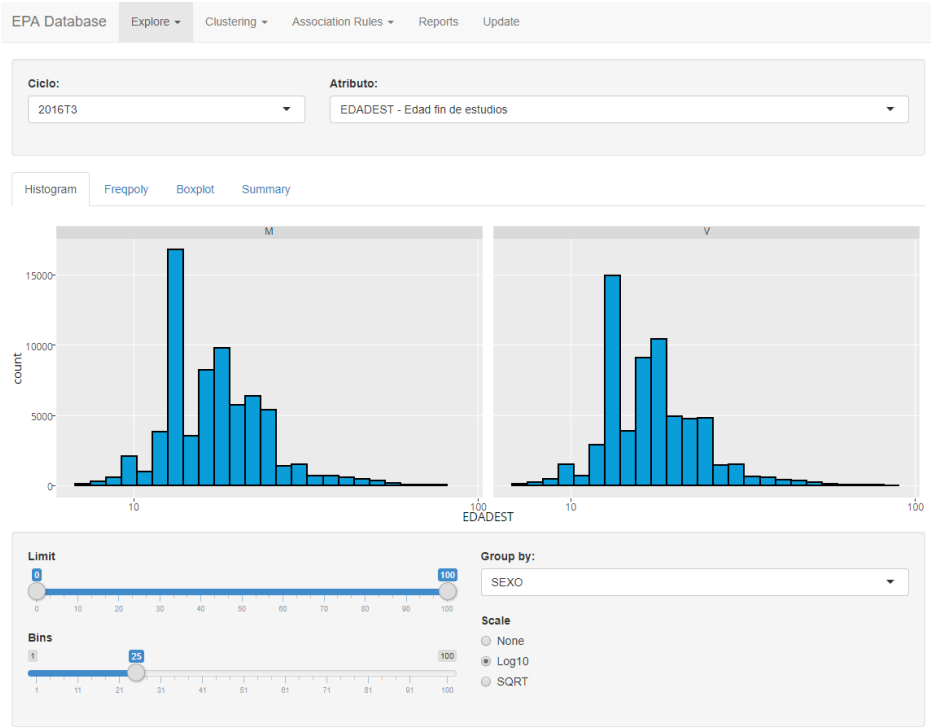

Figura 2. *Vista de exploración de una variable*

4.2. Requisitos funcionales

A continuación, se listan los requisitos que la aplicación debe cumplir:

- 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.

4.3. 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.

4.4. Tecnologías empleadas

Para el desarrollo de este proyecto se han utilizado las siguientes herramientas:

1. R, 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. Se han utilizado varios paquetes adicionales de R para cubrir distintas necesidades dentro del proyecto.
2. Shiny, uno de los principales paquetes de R de los que se ha hecho uso en el proyecto. 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.
3. SQLite, sistema de gestión de bases de datos relacional compatible con ACID. Funciona como una simple librería que ese 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.
4. RStudio, 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.
5. Git, 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.
6. 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.
7. Se ha utilizado DIA 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.

5. Conclusiones

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.

6. Bibliografía y referencias electrónicas

1. Sitio oficial de la Encuesta de Población Activa
http://www.ine.es/prensa/epa_prensa.htm

2. Sitio oficial de la Organización Internacional del Trabajo
http://www.ilo.org

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%2FPYSLayout&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. Recurso de documentación sobre metodologías de desarrollo incremental
https://proyectosagiles.org/desarrollo-iterativo-incremental

7. Curso Online, Data Analysis with R de Udacity
https://in.udacity.com/course/data-analysis-with-r--ud651/

8. B. Lantz, Machine Learning with R, Packt Publishing, (2013).

9. E. Rodríguez. Unsupervised Learning with R, Packt Publishing, (2015).

10. Z. Huang. Extensions to the k-Means Algorithm for Clustering Data Sets with Categorical Values, Kluwer Academic Publishers, (1998).

7. 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.

.