Hidrocarburos: Análisis de Datos de Estaciones de Servicio

Pablo Burgos Casado (Jefe de Área Desarrollo (SGTIC - MITYC))

María Teresa Simino Rueda

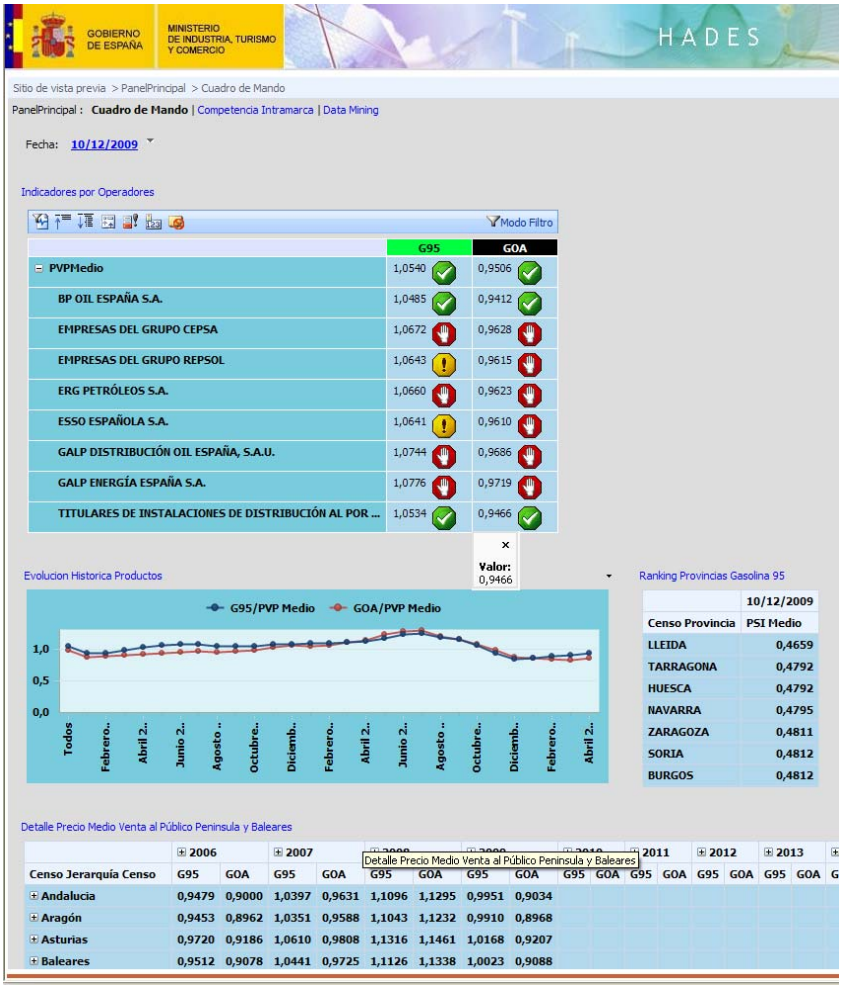
Rubén Pérez Gómez

Israel Santos Montero

María Ángeles Rodelgo Sanchez

1. INTRODUCCIÓN

HADES es un cuadro de mando formado por un sistema de indicadores que facilita la toma de decisiones y muestra un análisis completo de los precios de los combustibles



El cuadro de mando recoge los principales indicadores relacionados con la información remitida por las estaciones de servicio y por los operadores del sector petrolífero y los presenta de un modo claro y útil para la gestión del sistema. Es un sistema que nos

informa de la evolución de los parámetros fundamentales del sector de los Hidrocarburos en España.

HADES realiza **un análisis de los datos** enviados por las estaciones de servicio terrestres para los dos productos más significativos: **Gasolina 95** y **Gasoleo A**.

Se compone de dos partes, un cuadro de mandos que explota un **Datamart** con la información almacenada de los precios por operador, por provincia y por día y otra para la minería de datos o **datamining** permitiendo observar comportamientos contrarios a la libre competencia.

2. DESCRIPCIÓN DEL CUADRO DE MANDOS

El cuadro de Mando se compone:

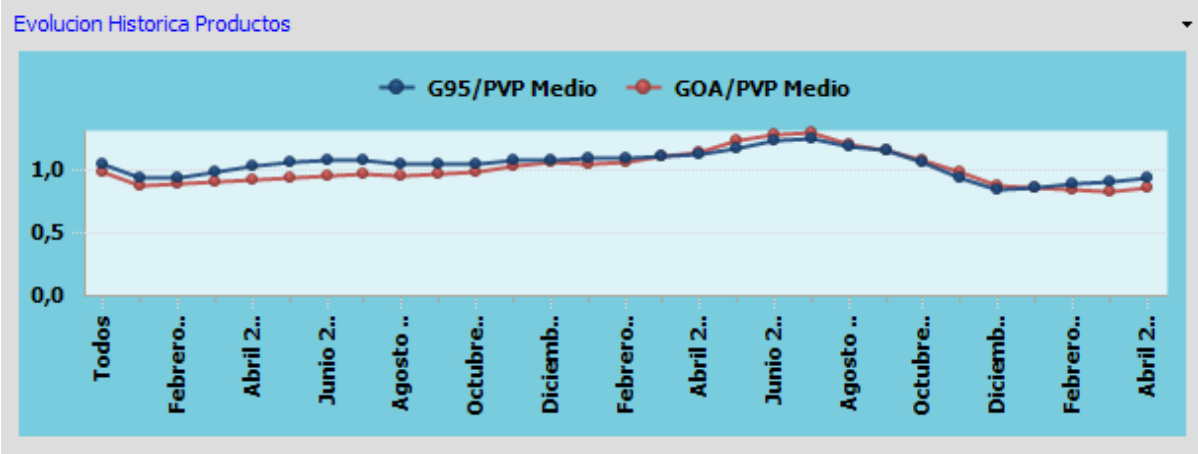
Una primera pestaña con las siguientes regiones:

1. Indicadores por Producto (Gasolina 95 ó Gasóleo A) y operador a una fecha concreta (por defecto, la fecha del día anterior). Se han definido KPI´s (Key Performance Indicators) para cada producto y para los operadores más relevantes, para estudiar las desviaciones de precios medios para cada cada tipo de combustible.

	G95	G0A
PVPMedio	1,0491	0,9465
BP OIL ESPAÑA S.A.	1,0432	0,9366
EMPRESAS DEL GRUPO CEPSA	1,0601	0,9550
EMPRESAS DEL GRUPO REPSOL	1,0561	0,9557
ERG PETRÓLEOS S.A.	1,0660	0,9623
ESSO ESPAÑOLA S.A.	1,0635	0,9598
GALP DISTRIBUCIÓN OIL ESPAÑA, S.A.U.	1,0768	0,9676
GALP ENERGÍA ESPAÑA S.A.	1,0771	0,9714
TITULARES DE INSTALACIONES DE DISTRIBUCIÓN AL POR ...	1,0525	0,9458

- En los indicadores se muestra el precio medio del PVP (precio de venta al público con impuestos incluidos) agrupado para los operadores más significativos en Península y Baleares
- Se pondría el indicador en
    - Verde si es menor o igual al PVP medio de Península y Baleares
    - Amarillo si es menor o igual al (PVP medio) * 1,01, es decir, si no superan en el 1% a la media
    - Rojo si superan a (PVP medio) * 1,01

2. Gráfica con la evolución mensual de los precios medios de ambos productos (G95 y gasoleo A) en Península y Baleares



3. Matriz con el precio medio del PVP de cada producto en Península y Baleares que permitiera navegación por las dimensiones.

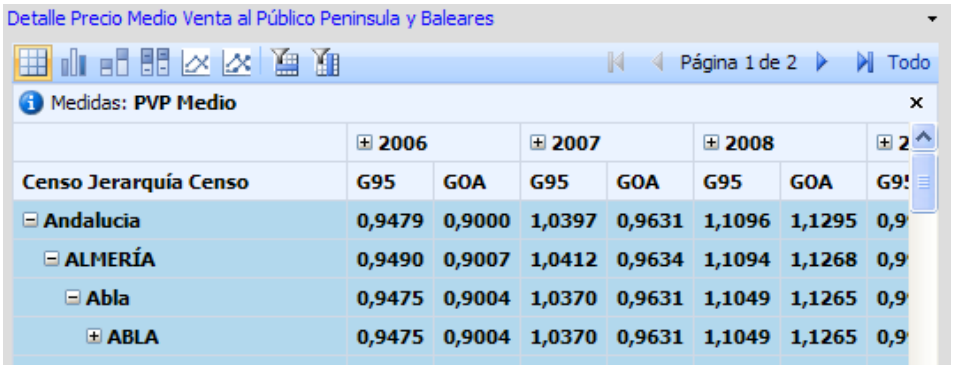
Detalle Precio Medio Venta al Público Peninsula y Baleares

Medidas: **PVP Medio**

	2006		2007		2008		2009	
Censo Jerarquía Censo	G95	GOA	G95	GOA	G95	GOA	G95	G
Andalucia	0,9479	0,9000	1,0397	0,9631	1,1096	1,1295	0,9954	0,
Aragón	0,9453	0,8962	1,0351	0,9588	1,1043	1,1232	0,9913	0,
Asturias	0,9720	0,9186	1,0610	0,9808	1,1316	1,1461	1,0172	0,
Baleares	0,9512	0,9078	1,0441	0,9725	1,1126	1,1338	1,0027	0,
Cantabria	0,9466	0,8976	1,0367	0,9608	1,1055	1,1277	0,9902	0,
Castilla la Mancha	0,9746	0,9262	1,0648	0,9893	1,1351	1,1559	1,0202	0,
Castilla y León	0,9459	0,8990	1,0356	0,9617	1,1059	1,1287	0,9923	0,
Cataluña	0,9748	0,9253	1,0624	0,9862	1,1316	1,1520	1,0181	0,
Comunidad Valenciana	0,9766	0,9149	1,0675	0,9785	1,1351	1,1430	1,0219	0,
Extremadura	0,9497	0,9019	1,0385	0,9633	1,1103	1,1328	0,9965	0,
Galicia	0,9743	0,9138	1,0636	0,9758	1,1347	1,1440	1,0216	0,
Madrid	0,9724	0,9252	1,0644	0,9890	1,1311	1,1526	1,0179	0

Desde el informe, podemos ir navegando hasta ver el máximo detalle.

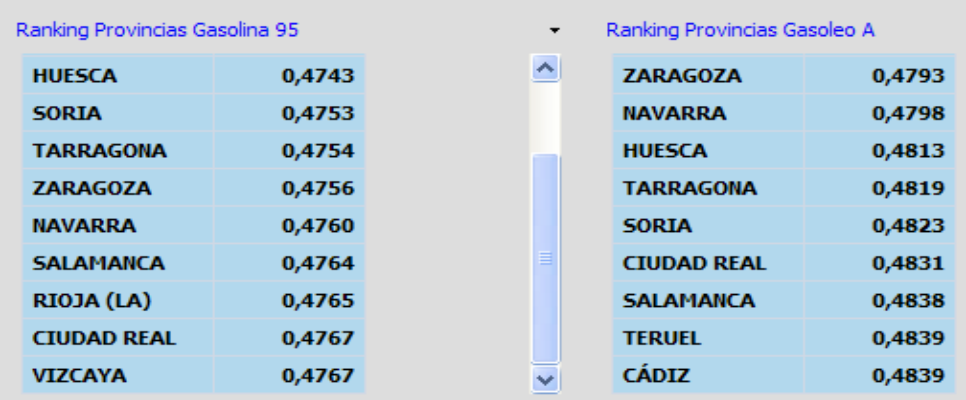
Por ejemplo **haciendo Drill Down hasta ver el PVPMedio por EESS**



Censo Jerarquía Censo	2006 G95	2006 G0A	2007 G95	2007 G0A	2008 G95	2008 G0A	2 G9
Andalucia	0,9479	0,9000	1,0397	0,9631	1,1096	1,1295	0,9
ALMERÍA	0,9490	0,9007	1,0412	0,9634	1,1094	1,1268	0,9
Abla	0,9475	0,9004	1,0370	0,9631	1,1049	1,1265	0,9
ABLA	0,9475	0,9004	1,0370	0,9631	1,1049	1,1265	0,9

También podríamos verlo para un año en concreto por operador o por operador para diversos años

4. Ranking de las 10 provincias más baratas para el precio medio del PSI (precio sin impuestos) de cada producto en Península y Baleares a una fecha concreta (por defecto, la fecha del día anterior).



Ranking Provincias Gasolina 95		Ranking Provincias Gasoleo A	
HUESCA	0,4743	ZARAGOZA	0,4793
SORIA	0,4753	NAVARRA	0,4798
TARRAGONA	0,4754	HUESCA	0,4813
ZARAGOZA	0,4756	TARRAGONA	0,4819
NAVARRA	0,4760	SORIA	0,4823
SALAMANCA	0,4764	CIUDAD REAL	0,4831
RIOJA (LA)	0,4765	SALAMANCA	0,4838
CIUDAD REAL	0,4767	TERUEL	0,4839
VIZCAYA	0,4767	CÁDIZ	0,4839

3. TECNICAS DE MINERIA DE DATOS (DATAMINING) EN HADES

La minería de datos es el proceso de extraer conocimiento útil y comprensible, previamente desconocido, desde grandes cantidades de datos almacenados. Es un proceso en el que partiendo de unos datos de entrada se generan unos modelos de salida, aplicando técnicas y algoritmos para extraer patrones de los datos. Estos modelos serán los que permitirán tomar decisiones estratégicas basándose en la información extraída de los datos.

Las tareas de minería de datos pueden ser predictivas o descriptivas.

- Modelos predictivos: Estiman valores de variables de interés a partir de valores de otras variables (predictoras). Un ejemplo de este tipo de sistemas es el que utilizan los bancos a la hora de conceder una hipoteca.

- Modelos descriptivos identifican patrones que explican o resumen los datos. Pueden ser:

- Reglas de asociación. El ejemplo más clásico es el de los distintos productos en el supermercado, se agrupan mediante reglas de asociación para incentivar la compra de productos relacionados (ej huevos-aceite)

- Agrupamiento o Clustering: agrupación de casos homogéneos.

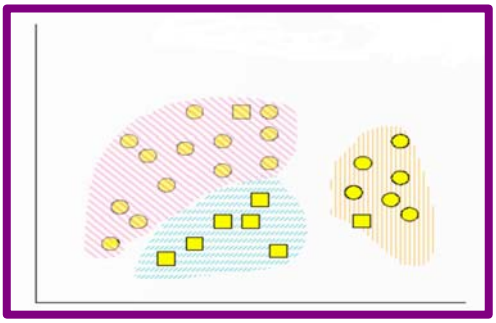
En nuestro caso, se ha orientado el estudio a los modelos de clustering porque se estudian los posibles agrupamientos de los precios dentro de cada provincia para gasolina 95 y gasóleo A. Se trata de estudiar la competencia de precios entre los distintos operadores.

idMunici...	FormaR...	Horario	PP_Precio	Cluster
5	1	2	0.911	cluster_0
6	2	2	0.891	cluster_2
	2	1	0.891	cluster_2
9	2	1	0.911	cluster_0
6	1	2	0.897	cluster_2
8	2	2	0.879	cluster_1
2	2	1	0.892	cluster_2
6	2	2	0.899	cluster_2
4	1	2	0.911	cluster_0
6	1	2	0.911	cluster_0
6	1	2	0.911	cluster_0
	2	2	0.894	cluster_2
6	2	2	0.891	cluster_2
6	2	2	0.885	cluster_2
	2	2	0.891	cluster_2
6	1	2	0.911	cluster_0
6	1	2	0.911	cluster_0

CLUSTERING EN HADES

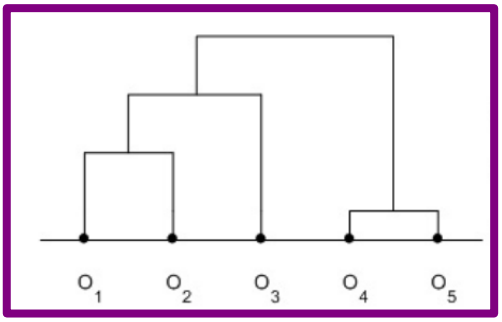
- Clustering particional

Se estudia para cada provincia las estaciones de servicio que formarían parte de los grupos o clusters identificados en función del precio. Todas las estaciones de servicio pertenecen a alguno de los k clusters, los cuales son disjuntos.



- Clustering ascendente jerárquico.

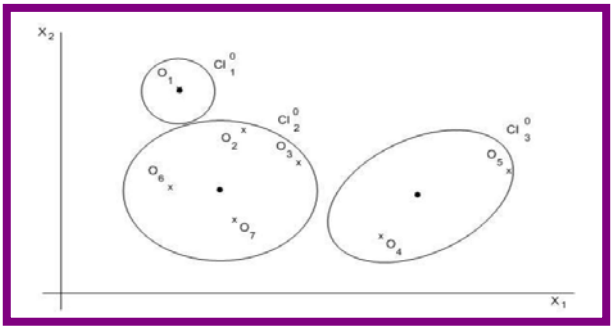
Se estudia el dendograma para crear un conjunto de agrupaciones anidadas hasta construir un árbol jerárquico. Permite obtener el valor de k para cada provincia cada día.



K-MEANS EN HADES

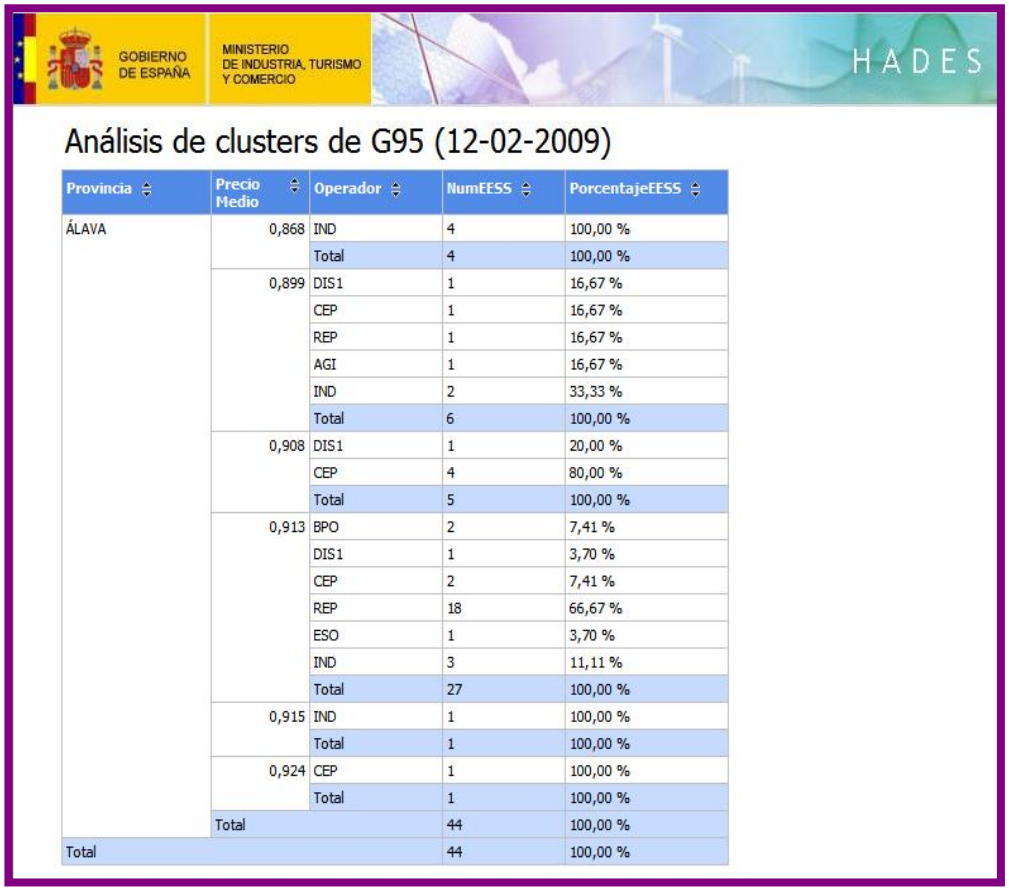
Se emplea el algoritmo K-MEANS (McQueen año 1967) que es el método de clustering particional más utilizado.

El principal problema radica en la elección del parámetro k, que identifica el número de grupos disjuntos. Para cada provincia se determina el valor del número de clusters (en base a las distancias medias y al dendograma (clustering jerárquico)) y se realizan los agrupamientos con respecto al centroide más cercano.



INFORMES DE CLUSTERING EN HADES

Se pueden obtener distintos informes por combustible, fechas y/o provincias.


GOBIERNO DE ESPAÑA
MINISTERIO DE INDUSTRIA, TURISMO Y COMERCIO
HADES

Análisis de clusters de G95 (12-02-2009)

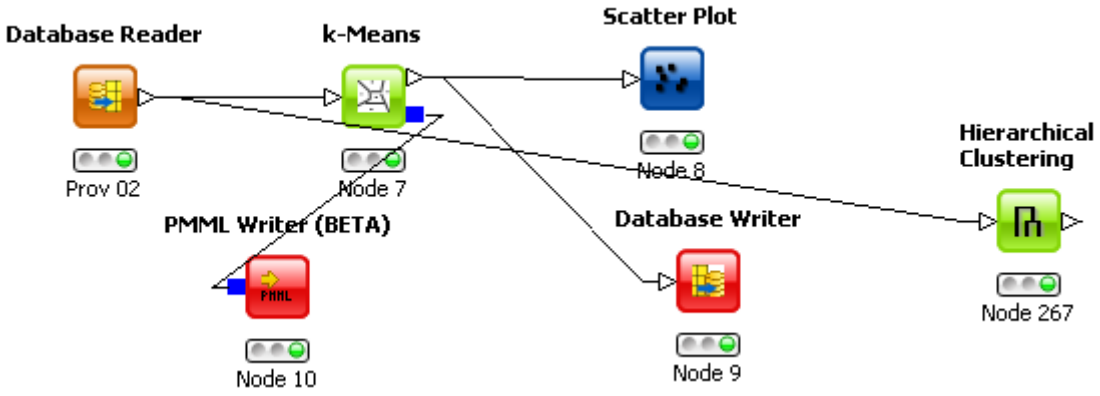
Provincia	Precio Medio	Operador	NumEESS	PorcentajeEESS
ÁLAVA	0,868	IND	4	100,00 %
		Total	4	100,00 %
	0,899	DIS1	1	16,67 %
		CEP	1	16,67 %
		REP	1	16,67 %
		AGI	1	16,67 %
		IND	2	33,33 %
		Total	6	100,00 %
	0,908	DIS1	1	20,00 %
		CEP	4	80,00 %
		Total	5	100,00 %
	0,913	BPO	2	7,41 %
		DIS1	1	3,70 %
		CEP	2	7,41 %
		REP	18	66,67 %
		ESO	1	3,70 %
		IND	3	11,11 %
		Total	27	100,00 %
	0,915	IND	1	100,00 %
		Total	1	100,00 %
	0,924	CEP	1	100,00 %
		Total	1	100,00 %
	Total		44	100,00 %
Total			44	100,00 %

En la imagen del informe se observa para una provincia (Álava) la distribución de las estaciones de servicio en estos grupos, identificando que operadores establecen precios más caros. El estudio se puede realizar dentro de cada provincia bien por precio medio (el informe reflejado), bien por municipio. Con estos informes se pueden extraer conclusiones respecto a seguimiento de precios por las estaciones de servicio del mismo operador, e identificar posibles indicios de políticas contrarias a la libre competencia dentro de un grupo de estaciones de servicio o de un operador.

Gracias a la aplicación Knime se crear los clusters de datos (creando un fichero XML y emigrándolo a Microsoft SQL Server 2005) de manera que luego se pueden plasmar en el cubo que alimenta los datos de PerfomancePoint.

Knime

Es una herramienta de software libre que permite realizar tareas de minería de datos según los diferentes algoritmos de los modelos. Mediante la herramienta para un algoritmo determinado se pueden arrastrar los nodos que realizan las distintas operaciones (lectura de base de datos, ejecución del algoritmo concreto, representación, etc) para ejecutarlos automáticamente una vez definido el proyecto.



En el caso concreto del proyecto HADES se emplean algoritmos de clustering. Los datos de cada bloque hacen referencia a una provincia y tipo de combustible de las distintas estaciones de servicio.

4. ALCANCE DEL PROYECTO

Tiene como objetivo agilizar la consulta de grandes cantidades de datos. La información se estructura en cubos multidimensionales o hipercubos y se compone de hechos numéricos llamados medidas que se clasifican por dimensiones.

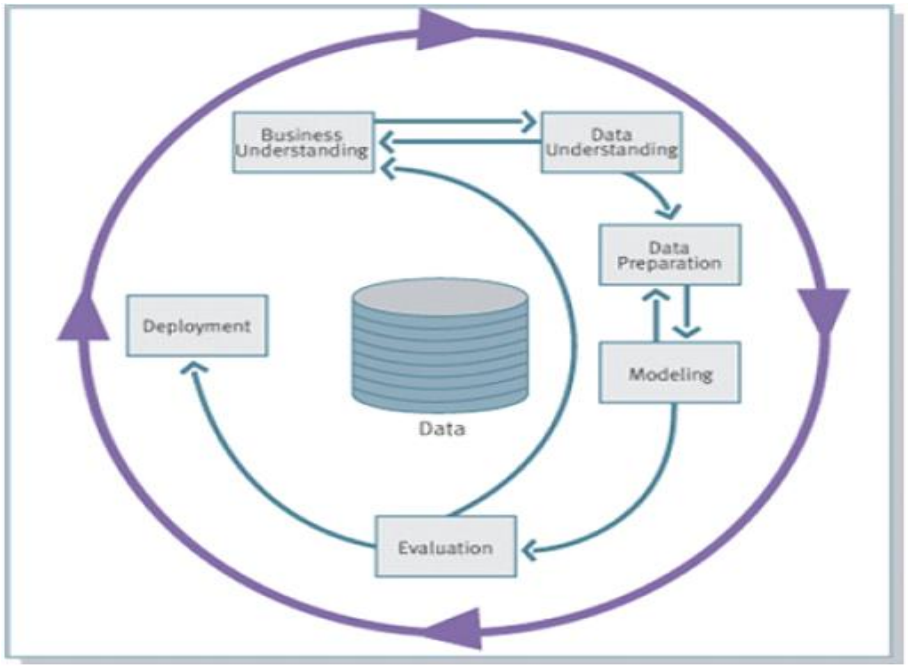
La aplicación está dirigida a aquellos usuarios que necesiten una consulta rápida, teniendo en cuenta el gran volumen de datos que existe debido a la recepción de la información enviada por los operadores y los propietarios de las distintas estaciones de servicio**. Actualmente se dispone de más de 16 millones de registros en el sistema operacional que recoge la información de precios de las distintas estaciones de servicio.** Las consultas sobre esta base de datos del sistema operacional para devolver gran cantidad de información remitida por distintas fuentes (estaciones de servicios, operadores) repercutía en elevados tiempos de respuesta en los informes. Para resolver esta problemática **se decidió crear una sistema de Datamart con la información consolidada de precios de estaciones de servicio por día y estación de servicio, permitiendo de este modo agilizar las consultar y la explotación de la información del sistema.** Para ello se establecieron procedimientos almacenados para realizar las **tareas de extracción transformación y carga, mediante DTS, de modo que en la base de datos multidimensional se consolidara el precio de cada día de la estación de servicio.** Mediante procesos batch se obtenía la información actualizada en el cuadro de mando de manera automática y se obtenían tiempos de respuesta inmediatos.

La definición del sistema cubrió la definición del **modelo multidimensional, para poder luego navegar por las distintas dimensiones del modelo (geográfica (ccaa, provincia, municipio, localidad), producto (gasolina95, gasóleo A), temporal (año, mes, día), operador).** También posteriormente se evaluaron los distintos algoritmos mediante la herramienta de minería de datos anime de software libre para identificar la información a explotar y la técnica más adecuada.

5. METODOLOGIA

Para el desarrollo se utilizo la **metodología CRISP-DM (CRoss Industry Standard Process for DataMining) que es un estándar para la realización de proyectos de minería de datos que reduce el tiempo de recuperación de la inversión realizada.**

El modelo de proceso proporciona una descripción del ciclo de vida de un proyecto, conteniendo sus correspondientes fases, tareas y las relaciones entre estas tareas.



El ciclo de vida de un proceso de minería de datos consta de 6 fases cuya secuencia no es estricta sino dependiente del resultado de la última fase llevada a cabo. El círculo externo simboliza la naturaleza cíclica de la minería de datos y las flechas pequeñas indican las dependencias más importantes y frecuentes entre fases.

- Comprensión del negocio: esta fase inicial se centra en la comprensión de los objetivos y requisitos del proyecto desde una perspectiva del cliente final para, posteriormente, diseñar un plan preliminar que defina el problema y permita la consecución de los objetivos. Aquí se identificaron con los gestores aquella información más relevante para incluir en el cuadro de mandos.
- Comprensión de los datos: esta fase comienza con una colección inicial de datos y lleva a cabo acciones para familiarizarse con ellos, medir su calidad, descubrir primeras ideas e, incluso, plantear hipótesis o subconjuntos de información que permitan descubrir tendencias ocultas. Se plantearon distintos tipos de indicadores y de informes basados en tablas y gráficos que pudieran dar una idea inmediata de los datos del sistema.
- Preparación de datos: se realizará una serie de transformaciones para obtener el conjunto final de datos que alimentará los algoritmos usados para la generación de modelos. Se prepararon los procedimientos almacenados y los DTS para cargar las nuevas tablas del modelo multidimensional.

- Modelado: en esta fase se aplican diversos algoritmos a los datos, calibrando sus parámetros con valores óptimos. Es muy habitual volver a la fase anterior por tener algunas de estas técnicas requisitos específicos sobre los datos o por cambios en el plan inicial debido a la extracción de nueva información durante esta fase. Con la herramienta libre anime se definió el proyecto que permitiera explotar la información contenida en la base de datos multidimensional medinte técnicas de clustering.
- Evaluación: Al llegar a esta fase ya se tiene al menos un modelo válido, desde el punto de vista del análisis de datos. Antes de llegar a la fase final, se evaluará en profundidad cada uno de los modelos revisando los pasos seguidos en su construcción para verificar la consecución de los objetivos y evitar que cualquier omisión impida alcanzar la mejor decisión posible.
- Implantación: Fase de aplicación de los modelos generados en un entorno de producción normal, de tal forma que los resultados sean organizados y presentados de forma útil para el cliente. Esta suele considerarse la última fase del proyecto pero no es así, puesto que los datos obtenidos en esta fase pueden realimentar nuevamente los modelos para mejorar las estimaciones realizadas.

La minería de datos se encuadra dentro de un proceso mucho mayor conocido como KDD (Knowledge Discovery from Databases).

6. BENEFICIARIOS

Todos aquellos usuarios que necesiten obtener de manera sencilla una visión general de los datos referentes a los precios de las estaciones de servicio, recoge los principales indicadores y los presenta de un modo claro y útil. En concreto los gestores de hidrocarburos que en base a la información puedan realizar estudios e informes sobre competencia en el sector de suministro de productos petrolíferos.

7. ARQUITECTURA

Se ha utilizado **Microsoft Office PerformancePoint Server 2007** para crear el cuadro de mandos, el servicio se sirve a través de Internet Information Server (IIS).

La información se estructura en cubos multidimensionales o hipercubos, apoyado en la base de datos **Microsoft SQL Server 2008**. El sistema operacional esta desarrollado con Visual Stio 2005 y ataca una base de datos Microsoft SQL Server 2005**.** Los DTS de carga y procedimientos almacenados permiten alimentar la base de datos multidimensional en Sql Server 2008

Los gestores acceden a través de su navegador a la aplicación sin necesidad de descargarse ningún pluging.

La parte de minería de datos se implementó con la herramienta de software libre Knime. Knime es una plataforma modular de exploración de datos que realiza tareas de minería de datos a través de los distintos algoritmos implementados con técnicas de minería de datos.

El **modelo con los datos de los distintos clusters se genera en Knime y se exporta como un XML**, **que es cargado en base de datos en la correspondiente tabla de cada provincia** con los datos de precios por estación para cada ejecución del proyecto de minería de datos.

Los usuarios acceden a través de su navegador a los informes predefinidos con Reporting Services 2005. Para otros informes que puedan necesitar en base a la

información almacenada disponen del Report Builder 2005, una herramienta que permite a los usuarios definirse sus propios reports e informes atacando vistas y tablas definidas del modelo.

8. RELEVANCIA

Hades es una herramienta destinada a facilitar el análisis de los datos de los combustible más vendidos Gasolina 95 (G95) y Gasoleo A (GOA) y por lo tanto los más importantes, gracias al cuadro de mando es fácil ver cual ha sido la evolución de los precios desde a nivel nacional hasta llegar a local. También se puede ver esta evolución desde un día concreto del año hasta las medias mensuales o anuales.

Esta información es considerada como relevante para los ciudadanos de acuerdo a lo dispuesto en el artículo 20 de la orden ITC 2308/2008 que regula la remisión de información por parte de los suministradores de productos petrolíferos al MITYC.

En base a la información recogida, la CNE y la Comisión de Defensa de la Competencia realizan estudios e incoan expedientes sancionadores por actividades contrarias a la libre competencia en el sector de los productos petrolíferos. El volumen económico manejado en este sector y la importancia del mismo, así como la repercusión en la economía del país justifican las actuaciones realizadas y el control sobre el citado sector.