UNIVERSITAT POLITÈCNICA DE VALÈNCIA

DSIC
DEPARTAMENTO DE SISTEMAS INFORMÁTICOS Y COMPUTACIÓN

REENMARCADO DE MODELOS MULTIDIMENSIONALES A DIFERENTES GRANULARIDADES

Autor:
Jorge Rafael Soro Doménech

Dirigido por:
José Hernández Orallo

Titulación:
Máster en Ingeniería del Software, Métodos Formales y Sistemas de Información

Valencia, Julio de 2014

Dedicado a mi madre, que ya no está con nosotros, pero que gracias a ella he podido llegar a todo lo que me he propuesto.

Jorge Soro Doménech
Valencia, 2014

RESUMEN

Dado un conjunto de datos multidimensional, podemos generar cubos a cualquier nivel de granularidad usando las jerarquías de las dimensiones. Los cubos son estructuras de datos multidimensionales, en las cuales el almacenamiento de los datos se realiza en una matriz multidimensional.

Sobre cada cubo, podemos extraer modelos de minería de datos y luego podemos aplicarlo sobre otro conjunto al **mismo nivel** de agregación. El problema es que, en unos datos multidimensionales, el número posible de cubos crece geométricamente con el número de niveles por dimensiones. Se hace, por tanto, inviable entrenar modelos a todos los posibles niveles.

Como alternativa, la idea del **reenmarcado de modelos multidimensionales a diferentes granularidades** se basa en aprender un modelo a **nivel inferior** (de mayor granularidad) y agregar las predicciones para poderlo usar a cualquier nivel superior. Otras opciones serían aprender unos pocos modelos a varios niveles bien elegidos y combinarlos mediante agregación y desagregación para cualquier nivel. Estas técnicas entran dentro del ámbito de "reenmarcado" (reframing) de modelos de minería de datos.

En este trabajo, evaluamos las aproximaciones anteriores (mismo nivel y nivel inferior) sobre varias bases de datos multidimensionales usando diferentes medidas de evaluación.

Palabras clave: Minería de datos, cubos de datos, dimensión, base datos multidimensional, reenmarcado de modelos minería de datos, granularidad de base de datos, modelos de predicción de datos, proyecto R.

ABSTRACT

Given a set of multidimensional data, we can generate cubes at any level of granularity using dimension hierarchies. Cubes are multidimensional data structures in which data storage is made in a multidimensional matrix.

On each cube, we can extract data mining models and then apply them over another set at the ***same level*** of aggregation. The problem is that, in multidimensional data, the possible number of cubes grows geometrically with the number of levels for dimensions. It becomes therefore unfeasible to train models for all possible levels.

The idea of **Reframing of multidimensional models at different granularities** consists in learning a model at the **lower level** (high granularity) and aggregating the predictions so that it can be used at any higher level. Other options would be to learn a few models at various levels and combine them through aggregation and disaggregation for any level. These techniques fall within the scope of *reframing* of data mining models.

In this paper, we evaluate the above approaches (*same level* and *lowest level*) over several multi-dimensional databases, using different assessment measures.

Keywords: Data Mining, data cubes, dimension, multidimensional database, data mining models, reframing, database granularity, data prediction models, R project.

AGRADECIMIENTOS

Este trabajo se ha realizado con la ayuda, apoyo, comprensión y paciencia de muchas personas y sería injusto no mencionar a todos quienes, de diferentes formas, han compartido conmigo la experiencia de este trabajo.

Quiero agradecer a mi familia el apoyo que me ha dado en todo momento y recordar a mi madre que ya no está, que seguro que se alegraría de ver lo que he conseguido desde que entré en la universidad.

A José Hernández Orallo, por conseguirme un trabajo a nivel de mi disponibilidad, por ayudarme infinidad de veces en todos los aspectos, por guiarme y ayudarme con todo y gracias a todo lo que ha hecho y me ha insistido, he llegado a finalizar el trabajo.

A Adolfo Martínez por su ayuda en el final del trayecto del trabajo y compartir su experiencia conmigo.

A todos los amigos, compañeros de universidad y de trabajo que me han estado preguntando todo el tiempo sobre el trabajo.

A mi pareja que ha estado siempre a mi lado para todo y me ha animado a seguir adelante con el máster.

A todo el mundo, por la comprensión de que no he podido dedicar el 100% de mí por falta de tiempo.

TABLA DE CONTENIDO

RESUMEN 5
ABSTRACT 6
AGRADECIMIENTOS 7
TABLA DE CONTENIDO 9
LISTA DE TABLAS 11
LISTA DE ILUSTRACIONES 13
1 INTRODUCCIÓN 15
1.1 MOTIVACIÓN 15
1.2 OBJETIVOS 16
1.3 ORGANIZACIÓN 17
2 ANTECEDENTES 19
2.1 ALMACENES DE DATOS 19
2.2 MODELO MULTIDIMENSIONAL 20
2.3 MINERÍA DE DATOS 22
2.4 HERRAMIENTA “R PROJECT” 24
3 REENMARCADO DE MODELOS MULTIDIMENSIONALES A DIFERENTES GRANULARIDADES 25
3.1 APROXIMACIONES 25
3.2 MODELOS UTILIZADOS 26
3.3 CONJUNTO DE DATOS DE EJEMPLO *TOY* 27
4 EXPERIMENTOS 35
4.1 METODOLOGÍA EXPERIMENTAL 35
4.2 RELLENADO DE CEROS 36
4.3 CONJUNTO DE DATOS *CAR FUEL EMISSIONS* 38
4.4 CONJUNTO DE DATOS *AROMADB* 45
4.4.1 AROMADB INDICADOR *DOLLARS* 46
4.4.2 AROMADB INDICADOR *QUANTITY* 51
4.5 CONJUNTO DE DATOS *GEN-VARIATIONS* 55
5 CONCLUSIONES Y TRABAJO FUTURO 61
5.1 DISCUSIÓN 63
5.2 TRABAJO FUTURO 64
BIBLIOGRAFÍA Y APÉNDICES 67
A BIBLIOGRAFÍA 67

B DETALLES CONJUNTOS DE DATOS 68
B.1 CONJUNTO DE DATOS *CAR FUEL EMISSIONS* 68
B.1.1 DISEÑO 70
B.1.2 PREPARACIÓN 70
B.1.3 DETALLE 73
B.1.4 JERARQUÍAS 75
B.1.5 CUBOS DE DATOS 76
B.2 CONJUNTO DE *DATOS AROMADB* 79
B.2.1 DISEÑO 80
B.2.2 DETALLE 81
B.2.3 JERARQUÍAS 86
B.2.4 CUBOS DE DATOS 86
B.3 CONJUNTO DE DATOS *GEN VARATIONS* 88
B.3.1 DISEÑO 88
B.3.2 DETALLE 89
B.3.3 JERARQUÍAS 91
B.3.4 CUBOS DE DATOS 92
B.4 ENLACES WEB DE CONJUNTOS DE DATOS 94
C SCRIPT MDHM 99
C.1 FUNCIONES MDHM 100
C.2 SCRIPT MDHM 101
D R-PROJECT 103
D.1 ¿POR QUÉ R? 104
D.2 OBTENCIÓN E INSTALACIÓN DE R 104
D.3 PAQUETES 105

LISTA DE TABLAS

TABLA 1 : BASES DE DATOS OLTP FRENTE OLAP 20
TABLA 2: DETALLE ATRIBUTOS DEL CONJUNTO DATOS TOY 28
TABLA 3: JERARQUÍAS PARA EL CONJUNTO TOY 28
TABLA 4: CUBO LOWEST LEVEL (TRAIN) DEL CONJUNTO TOY 29
TABLA 5: CUBO LOWEST LEVEL (TEST) DEL CONJUNTO TOY 30
TABLA 6: CUBO SAME LEVEL (TRAIN) CONJUNTO TOY 31
TABLA 7: CUBO SAME LEVEL (TRAIN) CONJUNTO TOY 31
TABLA 8: RESULTADO CONJUNTO DE DATOS TOY 32
TABLA 9: RESULTADO TÉCNICA MEAN PARA EL CONJUNTO CAR FUEL EMISSION 45
TABLA 10: RESULTADOS TÉCNICA MEAN CONJUNTO AROMADB (DOLLARS) 50
TABLA 11: RESULTADOS TÉCNICA MEAN CONJUNTO AROMADB (QUANTITY) 54
TABLA 12 : RESULTADOS TÉCNICA MEAN DEL CONJUNTO GEN VARIATIONS 59
TABLA 13: MANUFACTURERS DESCARTADOS (CAR FUEL EMISSION) 72
TABLA 14: MANUFACTURER CONSIDERADOS PARA EXPERIMENTO (CAR FUEL EMISSION) 73
TABLA 15: DIMENSIÓN YEAR (CAR FUEL EMISSION) 73
TABLA 16: DIMENSIÓN FUEL (CAR FUEL EMISSION) 73
TABLA 17: DIMENSIÓN ENGINE (CAR FUEL EMISSION) 73
TABLA 18: DIMENSIÓN TRANSMISSION (CAR FUEL EMISSION) 74
TABLA 19: DIMENSIÓN EURO (CAR FUEL EMISSION) 74
TABLA 20: DIMENSIÓN CAR (CAR FUEL EMISSION) 74
TABLA 21: HECHO CAR FUEL EMISSIONS (CAR FUEL EMISSION) 75
TABLA 22: JERARQUÍA CAR (CAR FUEL EMISSION) 75
TABLA 23: JERARQUÍA EURO (CAR FUEL EMISSION) 75
TABLA 24: JERARQUÍA TRANS (CAR FUEL EMISSION) 76
TABLA 25: JERARQUÍA ENGINE (CAR FUEL EMISSION) 76
TABLA 26: JERARQUÍA FUEL (CAR FUEL EMISSION) 76
TABLA 27: JERARQUÍA PROMOTION (CAR FUEL EMISSION) 76
TABLA 28: CUBOS (CAR FUEL EMISSION) 78
TABLA 29: TODAS LAS TABLAS DEL CONJUNTO AROMADB 81
TABLA 30: DIMENSIÓN PERIOD (AROMADB) 81
TABLA 31: DIMENSIÓN PRODUCT (AROMADB) 81
TABLA 32: DIMENSIÓN CLASS (AROMADB) 82
TABLA 33: DIMENSIÓN STORE (AROMADB) 82
TABLA 34: DIMENSIÓN MARKET (AROMADB) 82
TABLA 35: DIMENSIÓN PROMOTION (AROMADB) 83
TABLA 36: HECHO SALES (AROMADB) 83
TABLA 37: SQL VISTA SALES (AROMADB) 84
TABLA 38: DETALLE VISTA SALES (AROMADB) 85
TABLA 39: JERARQUÍA PROMOTION (AROMADB) 86
TABLA 40: JERARQUÍA CLASS (AROMADB) 86
TABLA 41: JERARQUÍA PRODUCT (AROMADB) 86
TABLA 42: JERARQUÍA PERIOD (AROMADB) 86
TABLA 43: JERARQUÍA STORE (AROMADB) 86
TABLA 44: CUBOS (AROMADB) 87
TABLA 45: DIMENSIÓN SPEC GEN-VARIATIONS 89
TABLA 46: DIMENSIÓN DBANK GEN-VARIATIONS 89
TABLA 47: DIMENSIÓN PHENO GEN-VARIATIONS 90

TABLA 48: DIMENSIÓN DATE GEN-VARIATIONS 90
TABLA 49: DIMENSIÓN GENO GEN-VARIATIONS 90
TABLA 50: HECHO GEN-VARIATIONS 91
TABLA 51: JERARQUÍA SPEC (GEN VARIATIONS) 91
TABLA 52: JERARQUÍA DBANK (GEN VARIATIONS) 91
TABLA 53: JERARQUÍA PHENOTYPE (GEN VARIATIONS) 92
TABLA 54: JERARQUÍA GENOTYPE (GEN VARIATIONS) 92
TABLA 55: JERARQUÍA DATE (GEN VARIATIONS) 92
TABLA 56: CUBOS GEN-VARIATIONS 93
TABLA 57: ENLACES WEB DE CONJUNTOS DE DATOS 98

LISTA DE ILUSTRACIONES

ILUSTRACIÓN 1: DIAGRAMA MULTIDIMENSIONAL CONJUNTO DE DATOS TOY..... 27
ILUSTRACIÓN 2: RESULTADO TÉCNICA MEAN CONJUNTO TOY..... 34
ILUSTRACIÓN 3: RESULTADOS EXPERIMENTO CONJUNTO CAR FUEL EMISSION (1)..... 39
ILUSTRACIÓN 4: RESULTADOS EXPERIMENTO CONJUNTO CAR FUEL EMISSION (2)..... 40
ILUSTRACIÓN 5: RESULTADOS EXPERIMENTO CONJUNTO AROMADB - DOLLARS (1) 47
ILUSTRACIÓN 6: RESULTADOS EXPERIMENTO CONJUNTO AROMADB - DOLLARS (2) 48
ILUSTRACIÓN 7: RESULTADOS EXPERIMENTO CONJUNTO AROMADB-QUANTITY (1)..... 51
ILUSTRACIÓN 8: RESULTADOS EXPERIMENTO CONJUNTO AROMADB-QUANTITY (2)..... 52
ILUSTRACIÓN 9: RESULTADOS EXPERIMENTO CONJUNTO AROMADB - DOLLARS (1) 56
ILUSTRACIÓN 10: RESULTADOS EXPERIMENTO CONJUNTO AROMADB - DOLLARS (2) 57
ILUSTRACIÓN 11: DIAGRAMA ENTIDAD RELACIÓN DATA CAR FUEL EMISSION 70
ILUSTRACIÓN 12: DIAGRAMA MULTIDIMENSIONAL AROMADB 80
ILUSTRACIÓN 13: DIAGRAMA ENTIDAD RELACIÓN AROMADB 80
ILUSTRACIÓN 14: DIAGRAMA MULTIDIMENSIONAL GENO-VARIATIONS..... 88
ILUSTRACIÓN 15: DIAGRAMA ENTIDAD RELACIÓN GEN-VARIATIONS 89

1 INTRODUCCIÓN

En este capítulo introductorio se describen la motivación de realización de este trabajo, qué objetivos nos hemos planteado y cómo se organiza la documentación.

1.1 MOTIVACIÓN

Toda la información que representa la realidad nos permite tomar decisiones. Desde hace años, los sistemas informáticos organizan y recopilan información para ayudar en la toma de decisiones. En los últimos años, se ha automatizado el proceso de almacenamiento, organización y recuperación de la información con la ayuda de los sistemas de bases de datos.

Hoy en día, difícilmente encontraremos sistemas los cuales no dispongan de la información almacenada, ya sea de forma estructurada, en bases de datos, no estructurada, en ficheros planos, etc. Hasta hace unos años, la intención del almacenamiento de la información se ha hecho con un fin concreto y luego no se realizaba una posterior explotación de los datos. En la actualidad, la situación ha cambiado de manera que se está invirtiendo en sistemas de explotación de información, ya sea enfocado al análisis o a las predicciones, creando sistemas de almacenamiento de información que contienen históricos con lo cual se puede analizar la información en muchos aspectos.

Los almacenes de datos se crean con el objetivo de mantener de forma consolidada el histórico de la información en una tabla de hechos (registros de ventas, consumos…) y describir de forma detallada los registros con las dimensiones (tablas descriptivas de un sistema de base de datos, supermercados, ciudad…), con lo cual se obtiene un modelo multidimensional de un sistema de información.

Una manera de explotar la información de un almacén de datos, es obteniendo modelos de minería de datos, los cuales se entrenan para obtener predicciones de la información y ayudar a la toma de decisiones.

Con este trabajo final de máster se han escogido diferentes conjuntos de datos, los cuales después de pre-procesarlos para que se adapten a los requisitos para el estudio, los comparamos con dos aproximaciones (*lowest level* y *same level*) de aplicación de modelos de minería de datos para obtener predicciones a diferentes niveles de agregación.

Este trabajo se desarrolla en el contexto del proyecto europeo "Reframe" (Bristol, Strasbourg, & Valencia, 2013), que se tiene como idea fundamentalla reutilización del conocimiento adquirido en forma de modelos de minería de datos, mediante el reenmarcado de los modelos (en lugar de su sustitución por nuevos modelos que se reentrenen en un nuevo contexto de aplicación). El proyecto "Reframe" tiene como principal objetivo abordar el desafío de rediseñar el proceso de datos con el fin de tener en cuenta la noción de contexto. En este trabajo, enfocamos el término "contexto" como el nivel de agregación en una base de datos multidimensional.

Para conseguir de una manera eficiente los cálculos y predicciones realizados con cubos de datos, se ha utilizado uno de los lenguajes libres más potente que existe: "R" (detalle en el apéndice D). Se han utilizado unos scripts proporcionados por el departamento DSIC para realizar la investigación. También se ha hecho uso de un servidor para realizar alguna de las ejecuciones que requería una gran cantidad de recursos de memoria.

1.2 OBJETIVOS

El propósito general de este trabajo es realizar predicciones de cubos de datos generados a distintos conjuntos de datos a diferentes niveles de granularidad.

Primero, se entrenarán los modelos y luego los aplicaremos sobre otro conjunto al mismo nivel de aplicación. Esta aproximación la denominamos *same level*. Aquí es donde surge el problema de que en unos datos multidimensionales el número posible de cubos crece de forma geométrica con el número de niveles por dimensiones. Se hace, por tanto, inviable entrenar modelos a todos los posibles niveles.

Como alternativa, se realiza el entrenamiento al nivel más bajo de agregación, con lo cual se obtiene de una forma muy detallada y precisa toda la información sobre un hecho con la agregación del indicador correspondiente. Luego estas predicciones se usarán para predecir cubos del mismo conjunto a diferentes niveles de agregación a través de la agregación de las predicciones. Esta es la aproximación *lowest level*.

Objetivos principales:

- Realizar predicciones de cubos a diferentes granularidades, en concreto usando la aproximación al nivel más bajo (*lowest level*) y al mismo nivel de agregación (*same level*).
- Comparar el comportamiento de las dos aproximaciones en cada una de las técnicas utilizadas, sobre diferentes conjuntos de datos como origen de datos.

1.3 ORGANIZACIÓN

La memoria de este trabajo fin de máster comienza con este capítulo introductorio, en la cual se intenta motivar y describir los objetivos que se pretenden estudiar en esta investigación y en cómo está estructurado.

En el capítulo dos, para poder llegar a entender mejor el conjunto del trabajo, se describen los almacenes de datos, qué es un modelo multidimensional y en qué consiste la minería de datos. Para terminar el capítulo, se describe la herramienta que se ha utilizado para realizar el trabajo.

En el tercer capítulo, se explican de forma detallada las aproximaciones que se han estudiado en el trabajo, qué modelos se han utilizado para los experimentos y una descripción detallada de cómo funciona la ejecución para un conjunto de datos del script en R reutilizado en este trabajo (MDHM) para obtener los resultados deseados para cada uno de los experimentos.

Seguidamente en el cuarto capítulo, primero se explica cómo funciona y para qué sirve la funcionalidad de rellenado de ceros utilizada en los experimentos. Luego se detallan los experimentos realizados para cada uno de los conjuntos de datos y el resultado obtenido, representado con gráficas los resultados y realizando una comparación detallada para cada uno de los modelos aplicados a estos conjuntos.

Le sigue un último capítulo donde se discuten las conclusiones obtenidas y se proponen trabajos futuros que puedan seguir el estudio de este trabajo.

Para terminar, se adjunta una serie de apéndices, entre ellos, el detalle de cada uno de los conjuntos de datos utilizados, la descripción de forma detallada del script MDHM y, finalmente, una breve descripción de R, explicando cómo instalar tanto la herramienta como alguno de sus paquetes para poder utilizar de forma rápida el script.

2 ANTECEDENTES

En este capítulo, se describe la tecnología de almacenes de datos, diferenciando los sistemas transaccionales y analíticos. Seguidamente se explica qué es un modelo multidimensional y en qué consiste la minería de datos. Para finalizar, se realiza una breve descripción de la herramienta utilizada para la realización del trabajo.

2.1 ALMACENES DE DATOS

Generalmente, la información que se quiere investigar sobre un cierto dominio de la organización se encuentra en bases de datos y otras fuentes muy diversas, tanto internas como externas.

El incremento tan grande de uso de los sistemas de información creados en bases de datos ha ayudado a generalizar el uso de herramientas de análisis de información que permiten obtener informes complejos, resúmenes, cuadros de mandos, etc., sobre la información almacenada con el objetivo de ayudar en el proceso de toma de decisiones. En la actualidad, es necesario distinguir dos sistemas de información distintos: el procesamiento transaccional (bases de datos OLTP) y el procesamiento analítico (bases de datos OLAP).

Hablamos de un **sistema transaccional** (OLTP) cuando en tiempo real se está trabajando con el sistema. Es decir, cuando se están realizando transacciones, actualizaciones y consultas a la base de datos con el fin operacional. Son sistemas destinados para el trabajo diario.

Por otra parte, los **sistemas analíticos** (OLAP) se basan en el procesamiento analítico en tiempo real de una base de datos realizando operaciones exclusivamente de consulta, en las cuales se realizan operaciones de agregación y cruces de información. El objetivo de estas consultas es obtener informes y cuadros de mando para el apoyo a la toma de decisiones.

Se puede distinguir en la tabla siguiente algunos puntos interesantes de cada uno de los dos sistemas, ya sea OLTP u OLAP. Según para que se vaya a utilizar un sistema de información, puede interesar un sistema u otro.

BASE DE DATOS TRANSACCIONAL (OLTP)	ALMACÉN DE DATOS (OLAP)
Almacén de datos transaccional	Almacén de datos histórico
Almacén de datos con información diaria	Almacén de datos con máximo detalle y la información esta agregada a diferentes niveles
Bases de datos medianas	Bases de datos muy grandes
Los datos son dinámicos, pueden cambiar en cualquier momento	Los datos son estáticos
Los procesos son repetitivos	Los procesos no son previsibles
Tiempo de respuesta mínimo	Tiempo de respuesta variable
Soporta decisiones diarias	Soporta decisiones estratégicas

Tabla 1 : Bases de Datos OLTP frente OLAP

Ya diferenciados los dos tipos de sistemas de información anteriores (OLTP y OLAP), pasamos a describir qué es un **almacén de datos**: una base de datos en la cual se almacena un conjunto de datos históricos detallados, que están integrados y organizados de tal forma que permiten la explotación de la información de forma eficiente y organizada con las herramientas analíticas que extraen de estas bases de datos informes, resúmenes, análisis, etc., con el fin de ayudar a la toma de decisiones estratégicas.

2.2 MODELO MULTIDIMENSIONAL

El **modelo multidimensional** es el modelo conceptual más utilizado para los almacenes de datos. Los datos se organizan en hechos, que son descritos a través de los atributos y medidas. Los hechos pueden ser detallados más o menos según el nivel de dimensión al que se acceda. Lo más importante del modelo multidimensional es la posibilidad de obtener la información sobre los hechos a diferentes niveles de agregación de forma instantánea.

Todo conjunto de dimensiones y de hechos se organiza en estrellas que pueden ser estrellas simples o jerárquicas (copo de nieve).

Si se tienen diferentes ámbitos, se creará una estrella que representará cada uno de los ámbitos, llamada "datamart". Cada datamart contendrá unas medidas y dimensiones propias que serán diferentes a las demás estrellas. Pero suele compartirse alguna dimensión como puede ser la de tiempo, ya que, como se ha comentado anteriormente, un almacén de datos representa información histórica.

Cuando no se pasa de más de tres dimensiones en un modelo multidimensional, podemos representar todos los niveles de agregación en un cubo. Un **cubo** está estructurado por casillas, cada casilla contiene un valor para cada dimensión a su correspondiente nivel de agregación. Cada una de las casillas representa un hecho. Cuando tenemos más de tres dimensiones, la representación sería realmente un "hipercubo", pero el término "cubo" se sigue utilizando igualmente.

Un cubo de datos provee una visión multidimensional de la información y permite realizar pre cálculos para que luego terceras aplicaciones (normalmente denominadas aplicaciones OLAP) tengan un acceso rápido a la información ya calculada. Las operaciones OLAP más comunes son:

- **Roll-up**: se trata de realizar agregaciones de los diferentes niveles de dimensiones para obtener el mínimo detalle.
- **Drill-down**: se trata de llegar al máximo detalle que puede aportar las dimensiones mediante la navegación por los diferentes niveles de las dimensiones hasta llegar al más detallado.
- **Slice y Dice**: se selecciona la información y se visualiza.
- **Pivotaje**: las dimensiones se cambian de columna a fila y viceversa.

Según lo comentado anteriormente, una razón por la que se crea un almacén de datos separado de una base de datos transaccional es el poder realizar la explotación de la información de una forma muy eficiente. Con el objetivo de obtener la eficiencia

deseada, según (Hernández-Orallo, Ramirez-Quintana, & Ferri-Ramirez, 2004) , "los sistemas de almacenes de bases de datos pueden implementarse utilizando dos tipos de esquemas físicos:

- **ROLAP (Relational OLAP):** físicamente, el almacén de datos se construye sobre una base de datos relacional.
- **MOLAP (Multidimensional OLAP):** físicamente, el almacén de datos se construye sobre estructuras basadas en matrices multidimensionales."

2.3 MINERÍA DE DATOS

"La minería de datos", según (Hernández-Orallo, Ramirez-Quintana, & Ferri-Ramirez, 2004) "es un término relativamente moderno que integra numerosas técnicas de análisis de datos y extracción de modelos. Aunque se basa en varias disciplinas, algunas de ellas más tradicionales, se distingue de ellas en la orientación más hacia el fin que hacia el medio, hecho que permite nutrirse de todas ellas sin prejuicios. Y el fin lo merece: ser capaces de extraer patrones, de describir tendencias" […] y, de en general, de "sacar partido a la información que nos rodea hoy en día, generalmente heterogénea y en grandes cantidades, permite a los individuos y a las organizaciones comprender y modelar de una manera más eficiente y precisa el contexto en el que deben actuar y tomar decisiones."

En (Witten & Frank, 2000) se define la minería de datos como "el proceso de extraer conocimiento útil y comprensible, previamente desconocido, desde grandes cantidades de datos almacenados en distintos formatos. Es decir, la tarea fundamental de la minería de datos es encontrar modelos inteligibles a partir de los datos." […] "Y el uso de los patrones descubiertos debería ayudar a tomar decisiones más seguras que reporten, por tanto, algún beneficio a la organización".

La minería de datos es un campo que se puede relacionar con otras disciplinas, como las bases de datos, la recuperación de información, la estadística, el aprendizaje

automático, los sistemas para la toma de decisión, la visualización de datos, la computación paralela y distribuida y otras.
Se puede utilizar para aplicaciones financieras y banca, análisis de mercado, distribución y en general comercio, seguros y salud privada, educación, medicina, biología, telecomunicaciones, entre otras.

Como se ha descrito anteriormente, la minería de datos es un proceso cuyo objetivo es obtener nuevo conocimiento a partir de las bases de datos. Este proceso se compone de diferentes pasos:

1. Integración y recopilación de la información de los orígenes de datos
2. Selección, limpieza y transformación de los datos
3. Obtener modelos de minería de datos
4. Evaluar e interpretan los resultados
5. Difundir y utilizar el nuevo conocimiento obtenido

"La generación de un modelo de minería de datos", según (Hernández-Orallo, 2012) "forma parte de un proceso mayor que incluye desde la formulación de preguntas acerca de los datos y la creación de un modelo para responderlas, hasta la implementación del modelo en un entorno de trabajo."

Ya explorados los datos, puede que nos encontremos con que no hay la cantidad suficiente de datos para obtener nuevos modelos de minería de datos y que, por tanto, se debe buscar más datos o se pueden generar diferentes modelos para descubrir si responde o no a las necesidades y en el caso que no, volver a plantear el problema.

Si se han obtenido los modelos, sólo quedaría interpretarlos y evaluarlos para obtener el nuevo conocimiento y compartirlo.

2.4 HERRAMIENTA “R PROJECT”

R es un lenguaje para el cálculo estadístico (R Project, 2014), generación de gráficos y permite la definición de nuevas técnicas mediante funciones. Surge a partir del lenguaje S que fue desarrollado por John Chambers en la empresa Laboratorios Bell. Dispone de una amplia variedad de técnicas estadísticas y gráficas de forma gratuita. Actualmente, es uno de los lenguajes más utilizados en el mundo de la investigación estadística. Los rasgos más característicos son:

- Capacidad de manipular y modificar datos y funciones.
- Los gráficos de alta calidad: visualización de datos y creación de gráficos.
- Existen paquetes R que se pueden descargar con el asistente de descarga que tiene internamente R.
- El lenguaje está orientado a objetos.

R está disponible en el sistema operativo GNU que es similar al UNIX, y también compilado para distintos sistemas operativos como puede ser Windows. Su *General Public Licence* (GPL) no pone ninguna restricción al uso, sólo restringe su distribución. Utiliza una interfaz de línea de comandos; sin embargo, existen herramientas visuales que están disponibles también gratuitamente, como la que se ha usado en este trabajo: RStudio (R Studio, 2014), que es un entorno de desarrollo integrado para R. Está disponible en código abierto y en ediciones comerciales y se puede utilizar en las plataformas Windows, Mac y Linux. También dispone de *RStudio server*, que se puede utilizar desde la web.

3 REENMARCADO DE MODELOS MULTIDIMENSIONALES A DIFERENTES GRANULARIDADES

En este capítulo se describen las dos aproximaciones principales: same level y lowest level, las técnicas predictivas que se utilizarán y se ilustrará todo con un ejemplo.

3.1 APROXIMACIONES

Se han propuesto dos aproximaciones distintas para realizar la predicción de los cubos y comparar cuál de las dos es la más adecuada para realizar predicciones sobre cubos de datos. Las dos aproximaciones se basan en los niveles de agregación que un cubo pueda tener y cómo realizar las predicciones de los cubos a distintos niveles para estudiar si son mejores o no. Las aproximaciones que presentamos son:

- *Same Level*
- *Lowest Level*

Same Level realiza predicciones sobre el mismo nivel de agregación del cubo, es decir, se entrena un modelo para un cubo y se aplica el modelo sobre otros datos pero con el mismo nivel de resolución.

Por otra parte, la aproximación ***Lowest Level*** consiste en realizar el cálculo del cubo del nivel más bajo y obtener las predicciones al nivel de agregación a este nivel más detallado. A continuación, se agregan las predicciones para el cubo que se desee.

3.2 MODELOS UTILIZADOS

En este trabajo se han utilizado 5 modelos distintos, disponibles en la herramienta R, para ver cual se comporta mejor.

- **MEAN**

  Uno de los medios más sencillos que hemos utilizado para medir la tendencia de los datos es simplemente calcular la media aritmética de la variable de salida sobre el conjunto de entrenamiento.

- **M5P (Regression tree from Weka)**

  Implementa una rutina para generar modelos de árboles de decisión M5 y sus reglas.

- **LRW (Linear Regression from Weka)**

  La regresión lineal es la técnica más clásica para estudiar la existente relación entre las variables predictivas y una variable de salida.

- **KNN (Package kknn in R)**

  Se trata de una técnica de clasificación que denomina vecinos a un conjunto de prototipos (k) más cercanos al patrón que se quiere clasificar. Se realizan predicciones, cogiendo como ejemplo los casos más parecidos al que se quiere predecir. Se debe especificar una métrica para medir la proximidad. La métrica por defecto es la euclídea.

- **SVM (Function svm in package e1071 in R)**

  El modelo de las máquinas de vectores de soporte (Cristianini-N & Scholkopf-B, 2002) pertenece a la familia de los clasificadores lineales, puesto que induce separadores lineales o hiperplanos en espacios de características de muy alta dimensionalidad generados por el kernel con un sesgo inductivo muy particular

3.3 CONJUNTO DE DATOS DE EJEMPLO *TOY*

A continuación, se detalla el conjunto de datos TOY. Se trata de unos datos de ejemplo, invitados, que representan un supermercado. Empezamos con la estructura del conjunto que podemos ver en el diagrama multidimensional de la ilustración 1.

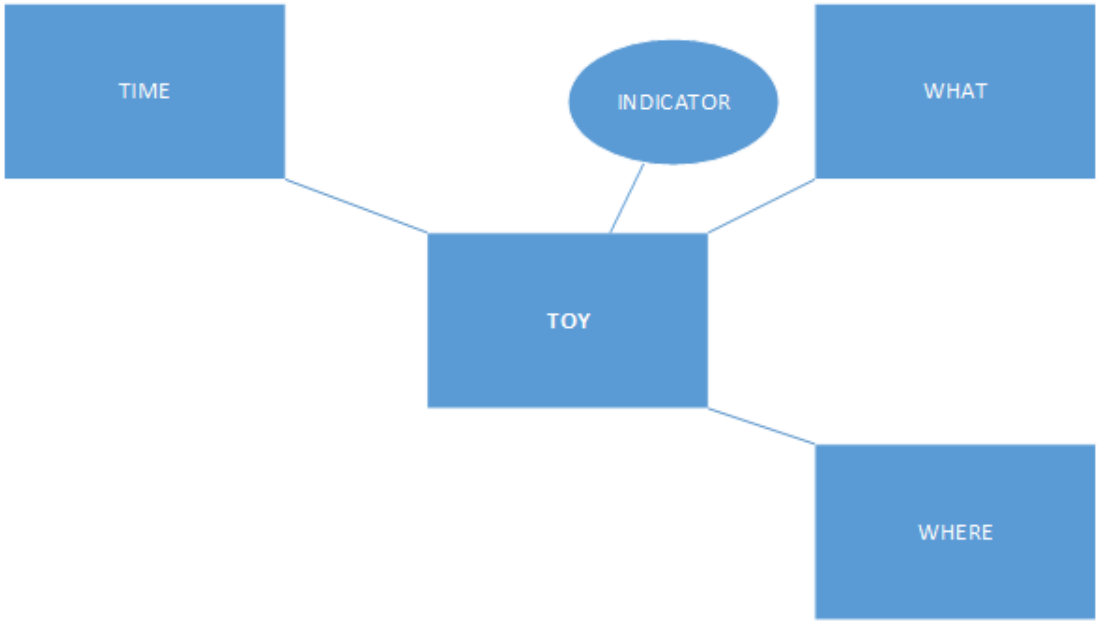

Ilustración 1: Diagrama multidimensional conjunto de datos TOY

Para que quede más claro, se detalla sus atributos y hechos en un diagrama entidad relación en la ilustración 2.

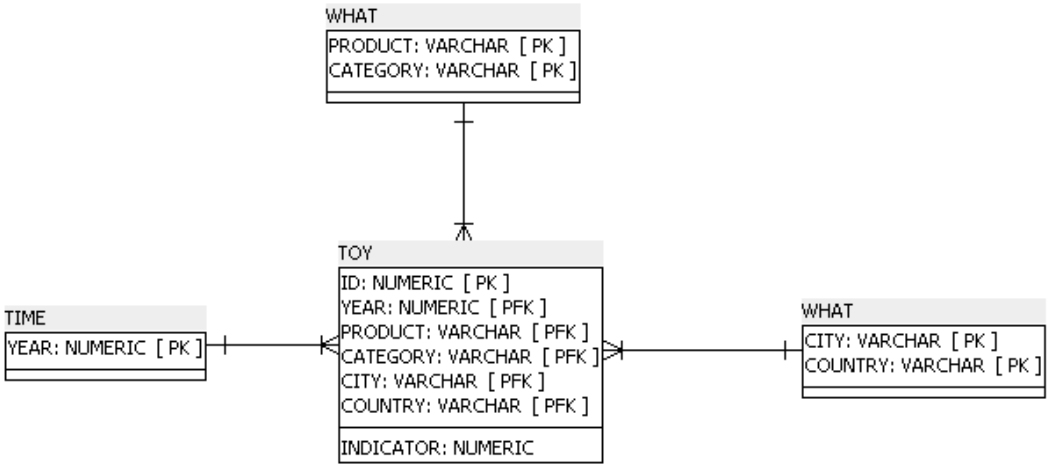

Ilustración 2: Diagrama entidad relación del conjunto de datos TOY

Ya conocida la estructura del modelo multidimensional del conjunto de datos de prueba, se representa en la tabla 2 el detalle de sus atributos.

ID	CITY	COUNTRY	PRODUCT	CATEGORY	YEAR	INDICATOR
1	VAL	ES	Tomato	Vegetables	1999	10
2	MAD	ES	Beer	Alcohol	1999	20
3	LON	UK	Whisky	Alcohol	1999	30
4	VAL	ES	Tomato	Vegetables	2000	0
5	VAL	ES	Beer	Alcohol	2000	12
6	LON	UK	Whisky	Alcohol	2000	33
7	MAD	ES	Tomato	Vegetables	2000	13
8	LON	UK	Tomato	Vegetables	2000	45
9	VAL	ES	Tomato	Vegetables	2001	6
10	MAD	ES	Beer	Alcohol	2001	3
11	LON	UK	Whisky	Alcohol	2001	0
12	VAL	ES	Beer	Alcohol	2001	5
13	VAL	ES	Beer	Alcohol	2002	6
14	LON	UK	Whisky	Alcohol	2002	183
15	LON	UK	Tomato	Vegetables	2002	25
16	VAL	ES	Tomato	Vegetables	2002	26

Tabla 2: Detalle atributos del conjunto datos TOY

Se definen en la tabla 3 las jerarquías formadas a partir de los atributos que hemos podido visualizar en el modelo multidimensional de TOY, para que el script genere de forma automática todas las posibles combinaciones de cubos, dónde ROLLED_UP_LEVEL marca el tope de la jerarquía y tiene como valor "- - - ".

JERARQUÍA	ATRIBUTOS
hierarchyWHERE	CITY, COUNTRY, ROLLED_UP_LEVEL
hierarchyWHAT	PRODUCT, CATEGORY, ROLLED_UP_LEVEL
hierarchyTIME	YEAR

Tabla 3: Jerarquías para el conjunto TOY

Luego, hay que especificar el año por el que se van a particionar los cubos para generar los conjuntos de entrenamiento y de test para realizar el experimento. En el caso del conjunto TOY, se obtiene una partición por el año 2001.

Además, se crea una columna "Count" que se le asigna el contenido de la columna INDICATOR (que es el indicador a agregar del conjunto TOY) y que representa las unidades vendidas y es el valor a predecir.

Por su sencillez, se escoge la técnica **MEAN** para describir las dos aproximaciones (*lowest level* y *same level*). El cálculo de la aproximación a nivel más bajo de agregación, *lowest level* obtiene como resultado el cubo a más bajo nivel de detalle que se particiona en un conjunto de entrenamiento. Podemos ver el resultado en la tabla 4.

LOWEST LEVEL (TRAIN)			
CITY	**PRODUCT**	**YEAR**	**COUNT**
MAD	Beer	1999	20
VAL	Tomato	1999	10
LON	Whisky	1999	30
VAL	Beer	2000	12
LON	Tomato	2000	45
MAD	Tomato	2000	13
VAL	Tomato	2000	0
LON	Whisky	2000	33

Tabla 4: Cubo Lowest Level (train) del conjunto TOY

Ya entrenada la técnica MEAN en el conjunto de datos, se obtiene un modelo el cual se aplica al cubo de datos a nivel más bajo para obtener las predicciones deseadas con el resultado de la tabla siguiente que representa el conjunto de datos de prueba con la aproximación *lowest level*. En el caso de la técnica MEAN para el conjunto de datos TOY, podemos visualizar en la tabla 5 que las predicciones (columna preds de la tabla y

siempre con un valor de 23.286, al ser la media de los valores de la columna COUNT de la Tabla 4) no son muy acertadas respecto a los valores que se obtienen después de la agregación del cubo de datos (columna actual de la tabla).

LOWEST LEVEL (TEST)				
CITY	**PRODUCT**	**YEAR**	**PREDS**	**ACTUAL**
MAD	Beer	2001	23.286	3
VAL	Beer	2001	23.286	5
VAL	Tomato	2001	23.286	6
LON	Whisky	2001	23.286	0
VAL	Beer	2002	23.286	6
LON	Tomato	2002	23.286	25
VAL	Tomato	2002	23.286	26
LON	Whisky	2002	23.286	183

Tabla 5: Cubo Lowest Level (test) del conjunto TOY

Ya calculadas las predicciones al nivel más bajo de agregación de un cubo de datos (*lowest level*), estas predicciones ya se pueden utilizar para agregarlas en cualquier otro nivel de agregación. Por ejemplo, si realizamos las predicciones al cubo que se compone de los atributos CITY y YEAR tenemos:

SAME LEVEL (TEST)			
CITY	YEAR	COUNT	PREDS
LON	2001	0	23.286
MAD	2001	3	23.286
VAL	2001	5,5	46.571
LON	2002	104	46.571
VAL	2002	16	46.571

Tabla 5b: Cubo Lowest Level (test) conjunto TOY después de la agregación

Ahora, vamos a detallar la aproximación *same level* sobre este mismo cubo, el que se compone de los atributos CITY y YEAR. El conjunto de entrenamiento de este cubo es el de la tabla 6.

SAME LEVEL (TRAIN)		
CITY	**YEAR**	**COUNT**
LON	1999	30
MAD	1999	20
VAL	1999	10
LON	2000	78
MAD	2000	13
VAL	2000	0

Tabla 6: Cubo Same Level (train) conjunto TOY

Una vez ya entrenado el conjunto de datos y obtenido el modelo MEAN para el nivel de resolución correspondiente (la media de la columna COUNT de la tabla 6, con resultado 25.667), se aplica al cubo del nivel de agregación y se obtiene el conjunto de datos de prueba con las predicciones que vemos en la tabla 7.

SAME LEVEL (TEST)			
CITY	YEAR	COUNT	PREDS
LON	2001	0	25.667
MAD	2001	3	25.667
VAL	2001	5,5	25.667
LON	2002	104	25.667
VAL	2002	16	25.667

Tabla 7: Cubo Same Level (test) conjunto TOY

Calculados todos los cubos de datos con las dos aproximaciones, sólo queda obtener los resultados finales dónde se resume el error cuadrático que es el que nos interesa

estudiar para ver qué aproximación obtiene mejores resultados para este experimento.

En la tabla 8, tenemos los resultados de aplicar la técnica MEAN al conjunto de datos TOY. La columna RESOLUTION contiene los diferentes cubos de datos. Las columnas TRAIN y TEST tienen la cantidad de filas de cada conjunto de datos de entrenamiento y de prueba, que sumados forman el cubo de datos original. La columna *SAME LEVEL* (MEAN) y *LOWEST LEVEL* (MEAN) contienen el error cuadrático producido para el cubo en concreto para las dos aproximaciones. La última columna, LOWEST LEVEL-BEST nos dice si el nivel *lowest level* obtiene menos error que el *same level* (WIN), si obtiene el mismo error (DRAW) o si obtiene un error mayor (LOSE).

RESOLUCIÓN					
RESOLUTION	**TRAIN**	**TEST**	**SAME LEVEL (MEAN)**	**LOWEST LEVEL (MEAN)**	**LOWEST LEVEL-BEST**
CITY, PRODUCT, YEAR	8	8	3483.329	3483.329	DRAW
CITY, CATEGORY, YEAR	8	8	3483.328	3483.328	DRAW
CITY, ---, YEAR	6	5	998.8507	993.8227	WIN
COUNTRY, PRODUCT, YEAR	7	7	3399.7901786	3449.560547	LOSE
COUNTRY, CATEGORY, YEAR	7	7	3399.7901786	3449.560547	LOSE
COUNTRY, ---, YEAR	4	4	933.7118055	959.2717014	LOSE
---, PRODUCT, YEAR	6	6	3437.4294	3446.2148	LOSE
---, CATEGORY, YEAR	4	4	769.56336805	755.12586805	WIN
---, ---, YEAR	2	2	232.29125	231.86328	WIN

Tabla 8: Resultado conjunto de datos TOY

Podemos destacar que, cuando existe más de un nivel de agregación, la aproximación *lowest level* se está comportando peor obteniendo errores mayores que la aproximación *same level*. Al contrario, a niveles altos de agregación, los resultados son mejores para la aproximación de *lowest level* teniendo un error cuadrático menor que la aproximación *same level*.

Y para finalizar el ejemplo, en la ilustración 2 tenemos un gráfico con el resultado de la técnica MEAN aplicada al conjunto de datos TOY. Aquí se ve que los resultados del error cuadrático son iguales para las dos aproximaciones en los casos en dónde hay menos nivel de agregación, pero en algunos casos la aproximación *lowest level* obtiene un error cuadrático mayor (cubos 4 – 6) y en otros por poca diferencia (décimas) mejores resultados (cubos 8 - 9).

El conjunto de datos TOY es meramente ilustrativo y no podemos sacar ninguna conclusión del mismo. El objetivo era entender cómo funcionan las dos aproximaciones y cómo vamos a comparar los resultados. En el capítulo siguiente ya abordamos conjuntos de datos más realistas.

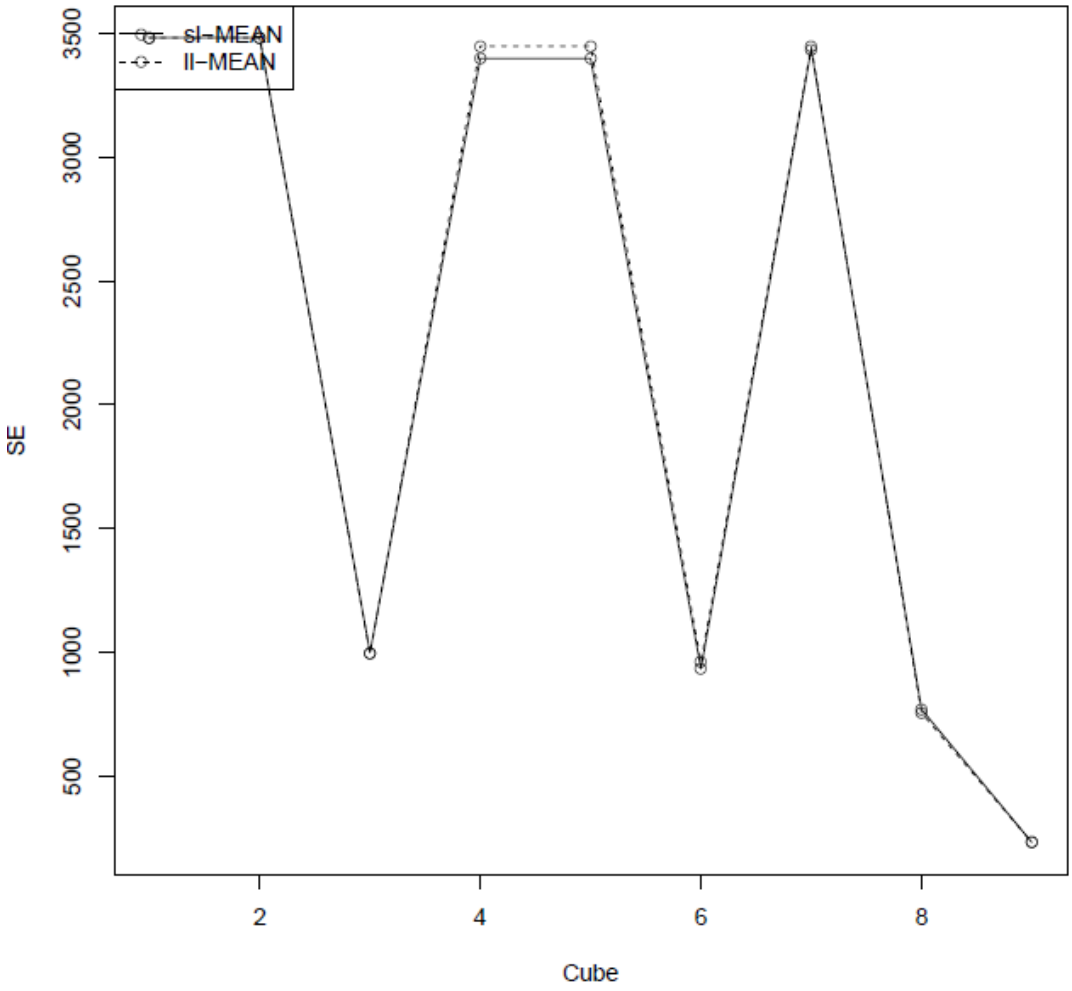

Ilustración 2: Resultado técnica MEAN conjunto TOY

4 EXPERIMENTOS

En este capítulo se presentan los resultados de las dos aproximaciones para tres conjuntos de datos multidimensionales.

4.1 METODOLOGÍA EXPERIMENTAL

Al igual que se hizo con los datos del conjunto TOY en el capítulo anterior, los experimentos consisten en la generación de todos los cubos posibles, pero para diferentes conjuntos de datos. A partir de aquí, se particiona el conjunto de entrenamiento y de prueba utilizando la dimensión de temporal. Se genera un modelo para la aproximación *lowest level* (ll-model) utilizando el conjunto de entrenamiento. Después generamos un modelo para cada uno de los niveles utilizando la aproximación *same level* (sl-model) con el conjunto de entrenamiento. se agrega el ll-model para resumir sus predicciones y comparar los resultados con el sl-model.

Las dos aproximaciones han sido aplicadas a los conjuntos de datos en los que se ha trabajado, obteniendo el error cuadrático medio y comparando cada uno de los casos, para determinar cuál de las dos aproximaciones obtiene mejores resultados.

A partir de todos los cálculos y cubos generados con los experimentos, tanto al nivel *lowest level* como al nivel *same level* se destacan los cálculos siguientes para cada uno de los cubos que se vuelcan a un fichero resultado para cada técnica:

- **mse**: representa el error cuadrático medio para la aproximación *same level*, que es la diferencia de las predicciones realizadas al conjunto de prueba y lo que se espera que tiene que dar como resultado.
- **mse_ll:** representa el error cuadrático medio para el *lowest level*, que es la diferencia de las predicciones realizadas al conjunto de prueba y lo que se espera que se tiene que obtener como resultado.
- **ll_best**: se obtiene cuál de los dos errores cuadráticos es mayor, si devuelve un LOW es que gana el error cuadrático a *same level*, si devuelve un WIN es que

gana el error cuadrático a *lowest level* y si devuelve DRAW es que los errores cuadráticos son iguales.

- **ll_reldif**: Diferencia de los errores cuadráticos a *lowest level* y *same level*.
- **resolution**: listado de cubos obtenidos a partir de las jerarquías definidas.

Cabe comentar que los cuatro conjuntos de datos utilizados para los experimentos tienen distintas características (incluido el conjunto TOY de ejemplo en el capítulo anterior). En algunos de los conjuntos ha sido necesario realizar un tratamiento especial de los datos, muy laborioso, que está descrito en el anexo al trabajo, para que cumplieran los requisitos necesarios para los estudios realizados.

Después de la sección siguiente, donde se trata la técnica de rellenado de ceros, se describen todos los experimentos realizados justificando los resultados obtenidos con gráficas representativas para cada uno de los conjuntos de datos, en las cuales se representa en el eje Y el error cuadrático y en el eje X el número de cubo del estudio.

Y para concluir cada experimento, se muestra una tabla con los resultados de la técnica MEAN (la tabla final con todos los resultados es muy grande) con todos los cubos generados, los errores de cada uno de ellos y si la aproximación *lowest level* es mejor, peor o igual que la *same level*.

4.2 RELLENADO DE CEROS

Muchos conjuntos de datos sólo contienen los datos positivos pero no los negativos (p.ej. no se vendió ningún tomate el día 12 en la tienda de Valencia, pero sabemos que su valor es 0).

En estos casos, el objetivo es generar todas las posibles combinaciones para todos los atributos importantes (están en las jerarquías especificadas para generar los cubos) de un conjunto rellenando con ceros cuando estos tienen como valor NULL o NA.

Por ejemplo, cuando realizamos las particiones por año puede que para una cierta combinación de atributos de un cubo no exista valor. Ppor tanto, la función lo que hace es generar todas las posibles combinaciones que se puedan dar rellenando con ceros, para que a la hora de realizar las funciones de agregación esa combinación se tenga en cuenta y no se descarte como pasaría si no se aplica esta funcionalidad.

La aplicación de la funcionalidad no se ha hecho para todos los conjuntos de datos, ya que en unos casos tenía sentido pero para otros no. Se ha aplicado al conjunto **AROMADB** dónde se calculan todos los cubos obteniendo el total de ventas y el total de productos vendidos, ya que una combinación no existente claramente indicaba que la venta había sido cero..

Para el conjunto de **GEN-VARIATIONS** también es preciso ya que se quiere saber si en un año determinado no se ha producido una variación concreta. Si esa fila no está podemos asumir que la variación no tiene lugar (aunque aquí la seguridad no es tan alta como en el caso anterior).

El conjunto **CARS-FUEL-EMISSIONS** es el único en el que no se ha aplicado la funcionalidad, ya que si no tenemos datos de consumo para un coche determinado, no quiere decir que ésta sea cero, sino que la desconocemos. De hecho, en este conjunto de datos, en vez de la suma, el indicador se agrega con la media, ya que lo que interesa conocer es la media de CO_2 que ha consumido un determinado coche, marca, etc, a medida que se va agregando.

En resumen, la técnica de rellenado de ceros se ha aplicado correctamente al conjunto de datos GEN VARIATIONS y AROMADB pero no al conjunto Cars Fuel Emissions que no tenía sentido.

4.3 CONJUNTO DE DATOS *CAR FUEL EMISSIONS*

El conjunto de datos **CAR FUEL EMISSIONS** original (detallado en el apéndice B.1), es una publicación de la Agencia de Certificación de Vehículos (VCA), una agencia del Reino Unido que proporciona este conjunto de datos de forma libre y gratuita.

Es una base de datos libre que contiene información sobre los consumos de los coches teniendo en cuenta muchos aspectos, como pueda ser el motor, marca, tipo de combustible, etc… Para este experimento, nos hemos centrado en el indicador de consumo de CO_2.

Contiene 45.500 filas de las cuales nos hemos quedado con 4.000 después de un proceso de tratamiento de los datos y de filtrado (detallado en el apéndice B.1.2) para hacer posible la ejecución del experimento, ya que con la cantidad de filas que tenía el conjunto original nos era imposible por falta de recursos de memoria. También tiene una gran cantidad de jerarquías la cual requiere mucha memoria.

El indicador que hemos estudiado es el consumo de CO_2, con lo que nos interesaba estudiar las predicciones del consumo medio de CO_2 para todos los cubos. Las particiones del conjunto de entrenamiento y de prueba se realizan por el año 2006. Se ha dividido en 6 jerarquías

- CAR (manufacturer.model.description, manufacturer.model,manufacturer,rolled_up_level),
- EURO (euro_standard,rolled_up_level),
- TRANS (transmission,transmission_type,rolled_up_level),
- ENGINE (engine_capacity,rolled_up_level),
- FUEL (fuel_type,rolled_up_level),
- PROMOTION (year).

Esto resulta en la generación de 4x2x3x2x2x1 cubos que hacen un total de 96 cubos (detallados en el apéndice B1.4 y apéndice B.1.5).

Como hemos dicho, en este conjunto no tiene sentido realizar un rellenado de ceros (la técnica que comentamos en la sección 4.1) ya que se va a calcular la media del consumo del indicador CO_2. Por tanto, la función de agregación utilizada es la media y no la suma, como para todos los otros conjuntos. Los resultados obtenidos se pueden ver en las siguientes gráficas.

En la ilustración 3, se representa el resultado con una visión global. Visualizamos que la técnica que mejor se comporta es la ll-KNN y ll-SVM solapandose las dos aproximaciones para prácticamente todos los cubos. Pero vemos los errores tan grandes que produce KNN a *same level* (ll-KNN) en determinados cubos. Le siguen M5P, LRW y MEAN solapándose también las dos aproximaciones. No se ve nada claro para este caso, cual de las dos aproximaciones se comporta mejor.

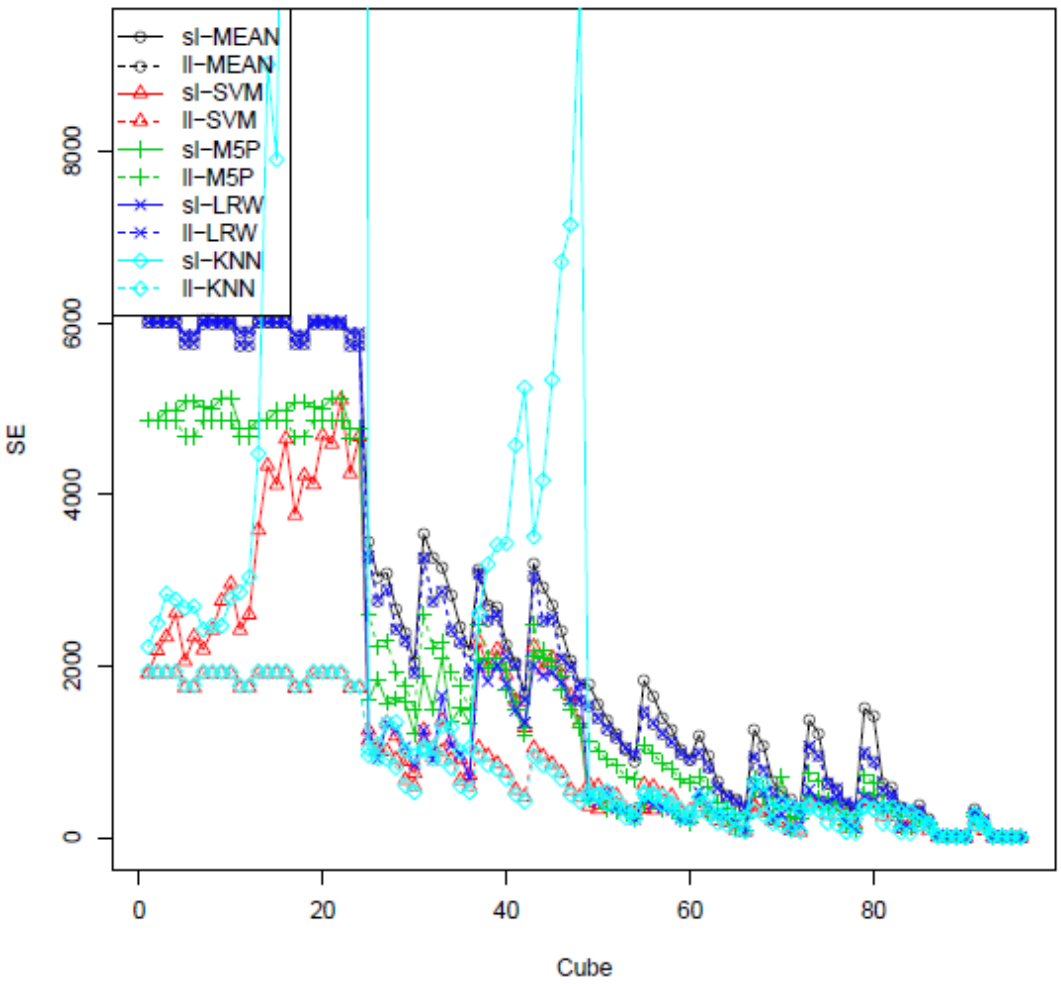

Ilustración 3: Resultados experimento conjunto CAR FUEL EMISSION (1)

Haciendo zoom, obtenemos la ilustración 4 en la cual podemos ver que del cubo 0 al 20 todas las técnicas dibujadas están fuera del rango de las gráficas porque las predicciones no son buenas para dichos cubos en cualquiera de las técnicas. Del cubo 21 al 50 ya visualizamos las técnicas a *lowest level* SVM, KNN y a *same level* LRW que son las que mejor se comportan. Del cubo 50 al 80 aparecen todas las técnicas a nivel *lowest level* menos la de LRW que aparece la de *same level* solapando la de *lowest level* y la de KNN, que se visualizan errores grandes en la línea del *same level* para pocos casos. En estos cubos se está prediciendo mejor. Para terminar, podemos ver que del cubo 80 al 96 el error es bajo para todas las técnicas, destacando la aproximación *lowest level*.

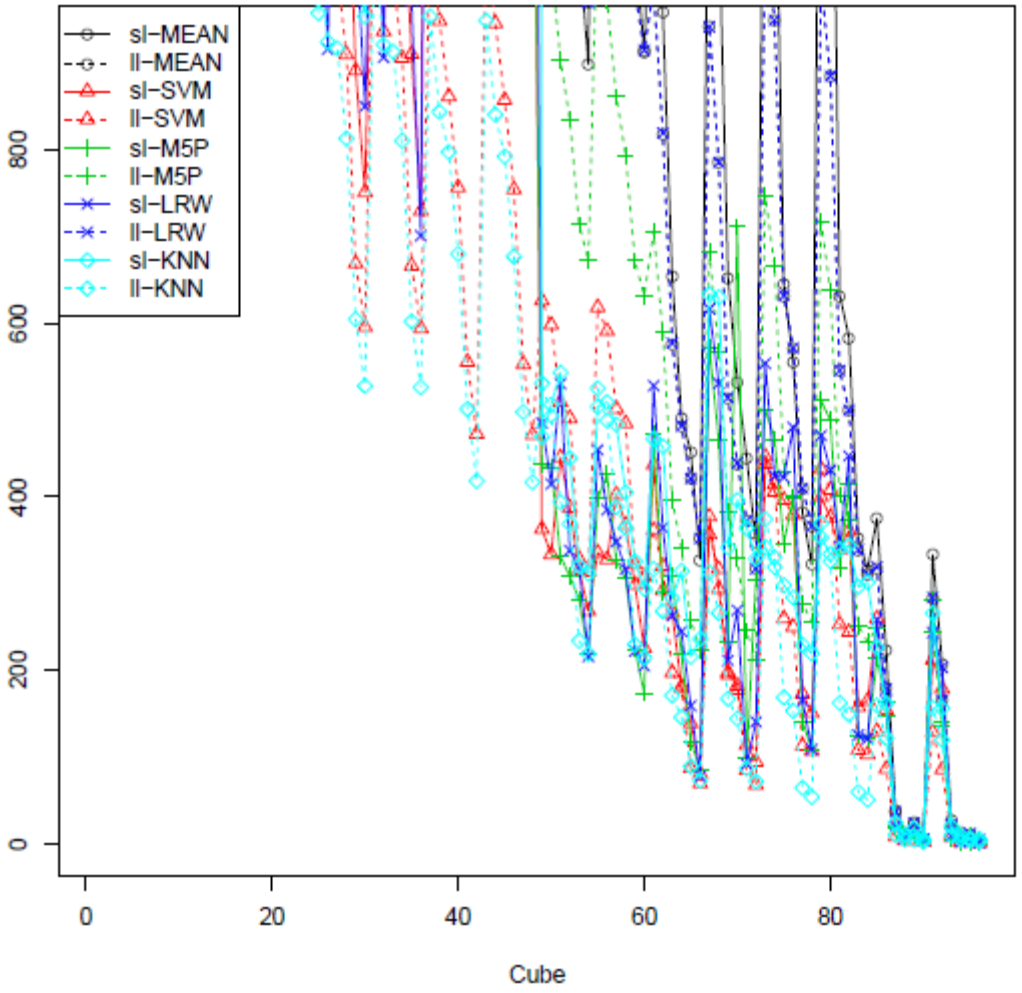

Ilustración 4: Resultados experimento conjunto CAR FUEL EMISSION (2)

Para terminar, se muestran en la tabla 9 el listado de todos los cubos generados en el experimento, con la cantidad de muestras para cada conjunto de datos (traint/test), el error producido en las dos aproximaciones y si la técnica MEAN con la aproximación a *lowest level* es mejor que la *same level*, peor o igual.

	Resolution	Train	Test	sl-MEAN	ll-MEAN	ll_best-MEAN
1	year, manufacturer,model,description, engine_capacity, euro_standard, transmission, fuel_type	1267	2597	6009,44	6009,44	draw
2	year, manufacturer,model,description, engine_capacity, euro_standard, transmission, ---	1267	2597	6009,44	6009,44	draw
3	year, manufacturer,model,description, engine_capacity, euro_standard, transmission_type, fuel_type	1265	2593	6005,35	6003,25	WIN
4	year, manufacturer,model,description, engine_capacity, euro_standard, transmission_type, ---	1265	2593	6005,35	6003,25	WIN
5	year, manufacturer,model,description, engine_capacity, euro_standard, ---, fuel_type	1063	2417	5860,20	5758,47	WIN
6	year, manufacturer,model,description, engine_capacity, euro_standard, ---, ---	1063	2417	5860,20	5758,47	WIN
7	year, manufacturer,model,description, engine_capacity, ---, transmission, fuel_type	1240	2583	6027,17	5993,73	WIN
8	year, manufacturer,model,description, engine_capacity, ---, transmission, ---	1240	2583	6027,17	5993,73	WIN
9	year, manufacturer,model,description, engine_capacity, ---, transmission_type, fuel_type	1237	2579	6021,52	5987,54	WIN
10	year, manufacturer,model,description, engine_capacity, ---, transmission_type, ---	1237	2579	6021,52	5987,54	WIN
11	year, manufacturer,model,description, engine_capacity, ---, ---, fuel_type	1039	2405	5892,25	5744,67	WIN
12	year, manufacturer,model,description, engine_capacity, ---, ---, ---	1039	2405	5892,25	5744,67	WIN
13	year, manufacturer,model,description, ---, euro_standard, transmission, fuel_type	1267	2597	6009,44	6009,44	draw
14	year, manufacturer,model,description, ---, euro_standard, transmission, ---	1267	2597	6009,44	6009,44	draw
15	year, manufacturer,model,description, ---, euro_standard, transmission_type, fuel_type	1265	2593	6005,35	6003,25	WIN

16	year, manufacturer,model,description, ---, euro_standard, transmission_type, ---	1265	2593	6005,35	6003,25	WIN
17	year, manufacturer,model,description, ---, euro_standard, ---, fuel_type	1063	2416	5853,87	5751,98	WIN
18	year, manufacturer,model,description, ---, euro_standard, ---, ---	1063	2416	5853,87	5751,98	WIN
19	year, manufacturer,model,description, ---, ---, transmission, fuel_type	1240	2583	6027,17	5993,73	WIN
20	year, manufacturer,model,description, ---, ---, transmission, ---	1240	2583	6027,17	5993,73	WIN
21	year, manufacturer,model,description, ---, ---, transmission_type, fuel_type	1237	2579	6021,52	5987,54	WIN
22	year, manufacturer,model,description, ---, ---, transmission_type, ---	1237	2579	6021,52	5987,54	WIN
23	year, manufacturer,model,description, ---, ---, ---, fuel_type	1039	2404	5885,99	5738,18	WIN
24	year, manufacturer,model,description, ---, ---, ---, ---	1039	2404	5885,99	5738,18	WIN
25	year, manufacturer,model, engine_capacity, euro_standard, transmission, fuel_type	763	1466	3450,50	3275,30	WIN
26	year, manufacturer,model, engine_capacity, euro_standard, transmission, ---	629	1165	3045,65	2765,13	WIN
27	year, manufacturer,model, engine_capacity, euro_standard, transmission_type, fuel_type	720	1240	3083,20	2888,18	WIN
28	year, manufacturer,model, engine_capacity, euro_standard, transmission_type, ---	582	937	2672,00	2429,27	WIN
29	year, manufacturer,model, engine_capacity, euro_standard, ---, fuel_type	517	842	2400,38	2298,03	WIN
30	year, manufacturer,model, engine_capacity, euro_standard, ---, ---	396	600	2040,27	1918,41	WIN
31	year, manufacturer,model, engine_capacity, ---, transmission, fuel_type	698	1441	3539,20	3256,54	WIN
32	year, manufacturer,model, engine_capacity, ---, transmission, ---	535	1142	3261,15	2752,07	WIN
33	year, manufacturer,model, engine_capacity, ---, transmission_type, fuel_type	648	1211	3159,31	2870,01	WIN
34	year, manufacturer,model, engine_capacity, ---, transmission_type, ---	487	913	2839,72	2416,59	WIN
35	year, manufacturer,model, engine_capacity, ---, ---, fuel_type	457	821	2459,20	2283,46	WIN
36	year, manufacturer,model,	319	584	2179,81	1910,26	WIN

	engine_capacity, ---, ---, ---					
37	year, manufacturer,model, ---, euro_standard, transmission, fuel_type	666	1392	3136,80	3074,96	WIN
38	year, manufacturer,model, ---, euro_standard, transmission, ---	537	1078	2711,48	2539,45	WIN
39	year, manufacturer,model, ---, euro_standard, transmission_type, fuel_type	602	1123	2683,17	2591,42	WIN
40	year, manufacturer,model, ---, euro_standard, transmission_type, ---	475	802	2259,50	2105,83	WIN
41	year, manufacturer,model, ---, euro_standard, ---, fuel_type	427	736	2039,82	2010,26	WIN
42	year, manufacturer,model, ---, euro_standard, ---, ---	313	484	1676,43	1609,57	WIN
43	year, manufacturer,model, ---, ---, transmission, fuel_type	597	1366	3189,32	3047,49	WIN
44	year, manufacturer,model, ---, ---, transmission, ---	432	1056	2910,46	2525,79	WIN
45	year, manufacturer,model, ---, ---, transmission_type, fuel_type	527	1095	2723,25	2563,21	WIN
46	year, manufacturer,model, ---, ---, transmission_type, ---	372	780	2407,41	2090,27	WIN
47	year, manufacturer,model, ---, ---, ---, fuel_type	362	716	2066,47	1987,66	WIN
48	year, manufacturer,model, ---, ---, ---, ---	230	470	1792,43	1600,39	WIN
49	year, manufacturer, engine_capacity, euro_standard, transmission, fuel_type	329	514	1778,77	1535,19	WIN
50	year, manufacturer, engine_capacity, euro_standard, transmission, ---	274	422	1567,71	1393,51	WIN
51	year, manufacturer, engine_capacity, euro_standard, transmission_type, fuel_type	274	352	1378,27	1269,97	WIN
52	year, manufacturer, engine_capacity, euro_standard, transmission_type, ---	211	282	1213,85	1171,34	WIN
53	year, manufacturer, engine_capacity, euro_standard, ---, fuel_type	182	227	1037,19	1041,43	LOSE
54	year, manufacturer, engine_capacity, euro_standard, ---, ---	141	183	898,42	969,99	LOSE
55	year, manufacturer, engine_capacity, ---, transmission, fuel_type	277	494	1826,98	1471,70	WIN
56	year, manufacturer, engine_capacity, ---, transmission, ---	218	405	1650,81	1333,29	WIN
57	year, manufacturer, engine_capacity, ---, transmission_type, fuel_type	222	333	1399,12	1210,56	WIN
58	year, manufacturer, engine_capacity, ---, transmission_type, ---	164	268	1264,46	1114,44	WIN
59	year, manufacturer, engine_capacity, ---, ---, fuel_type	147	214	1035,50	983,36	WIN

60	year, manufacturer, engine_capacity, ---, ---, ---	111	174	913,84	915,10	LOSE
61	year, manufacturer, ---, euro_standard, transmission, fuel_type	236	392	1196,77	986,95	WIN
62	year, manufacturer, ---, euro_standard, transmission, ---	187	290	958,38	819,50	WIN
63	year, manufacturer, ---, euro_standard, transmission_type, fuel_type	167	201	654,54	577,10	WIN
64	year, manufacturer, ---, euro_standard, transmission_type, ---	119	138	491,10	481,40	WIN
65	year, manufacturer, ---, euro_standard, ---, fuel_type	102	118	452,31	420,97	WIN
66	year, manufacturer, ---, euro_standard, ---, ---	72	79	327,92	351,69	LOSE
67	year, manufacturer, ---, ---, transmission, fuel_type	173	374	1255,62	942,05	WIN
68	year, manufacturer, ---, ---, transmission, ---	126	277	1072,67	785,53	WIN
69	year, manufacturer, ---, ---, transmission_type, fuel_type	112	184	653,09	514,15	WIN
70	year, manufacturer, ---, ---, transmission_type, ---	75	127	533,34	438,87	WIN
71	year, manufacturer, ---, ---, ---, fuel_type	64	108	445,55	372,19	WIN
72	year, manufacturer, ---, ---, ---, ---	44	73	347,41	316,88	WIN
73	year, ---, engine_capacity, euro_standard, transmission, fuel_type	222	357	1367,86	1059,69	WIN
74	year, ---, engine_capacity, euro_standard, transmission, ---	174	277	1220,48	949,49	WIN
75	year, ---, engine_capacity, euro_standard, transmission_type, fuel_type	152	161	644,75	631,92	WIN
76	year, ---, engine_capacity, euro_standard, transmission_type, ---	112	119	556,04	571,47	LOSE
77	year, ---, engine_capacity, euro_standard, ---, fuel_type	94	91	381,81	409,52	LOSE
78	year, ---, engine_capacity, euro_standard, ---, ---	69	65	321,55	366,31	LOSE
79	year, ---, engine_capacity, ---, transmission, fuel_type	166	339	1508,54	993,32	WIN
80	year, ---, engine_capacity, ---, transmission, ---	123	262	1413,51	885,05	WIN
81	year, ---, engine_capacity, ---, transmission_type, fuel_type	101	143	632,67	545,59	WIN
82	year, ---, engine_capacity, ---, transmission_type, ---	69	104	583,61	499,85	WIN
83	year, ---, engine_capacity, ---, ---, fuel_type	56	80	352,68	338,95	WIN
84	year, ---, engine_capacity, ---, ---, ---	38	56	315,77	310,02	WIN

85	year, ---, ---, euro_standard, transmission, fuel_type	99	196	374,81	318,92	WIN
86	year, ---, ---, euro_standard, transmission, ---	67	126	222,13	178,52	WIN
87	year, ---, ---, euro_standard, transmission_type, fuel_type	47	39	39,57	36,27	WIN
88	year, ---, ---, euro_standard, transmission_type, ---	28	20	14,09	11,07	WIN
89	year, ---, ---, euro_standard, ---, fuel_type	26	20	22,33	23,74	LOSE
90	year, ---, ---, euro_standard, ---, ---	15	10	5,88	5,39	WIN
91	year, ---, ---, ---, transmission, fuel_type	63	185	332,85	283,87	WIN
92	year, ---, ---, ---, transmission, ---	39	118	207,01	164,60	WIN
93	year, ---, ---, ---, transmission_type, fuel_type	24	32	27,65	21,88	WIN
94	year, ---, ---, ---, transmission_type, ---	12	16	12,27	8,78	WIN
95	year, ---, ---, ---, ---, fuel_type	12	16	8,09	11,22	LOSE
96	year, ---, ---, ---, ---, ---	6	8	4,68	4,42	WIN

Tabla 9: Resultado técnica MEAN para el conjunto Car Fuel Emission

4.4 CONJUNTO DE DATOS *AROMADB*

El conjunto de datos **AROMADB** (IBM, 2014) contiene información ficticia (detallado en el apéndice B.2) de un histórico de ventas de una red de supermercados, en el cual se registran las ventas realizadas de los productos disponibles, contando la cantidad de productos vendidos y el coste de la venta.

Con todas las jerarquías de las que se puede disponer del conjunto **AROMADB**, se podrían formar hasta 420 cubos con la combinación de todas ellas. Por falta de recursos de memoria, no se ha podido utilizar la totalidad de atributos de los que dispone el conjunto y se ha elimina para el estudio la jerarquía que se compone del producto, ya que aporta poco valor para el estudio.

Los cubos que se llegan a conseguir son los generados a partir de las cuatro jerarquías que se describen a continuación:

- PROMOTION (KEY_PROMO,PROMO_TYPE,ROLLED_UP_LEVEL),
- CLASS (KEY_CLASS ,ROLLED_UP_LEVEL),
- PERIOD (SAL_YEAR),

- STORE (KEY_STORE, KEY_MKT, MKT_HQ_CITY, MKT_HQ_STATE, MKT_DISTRICT, MKT_REGION, ROLLED_UP_LEVEL).

Con ellas se obtienen 3x2x1x7 cubos que hacen un total de 42 cubos (detallado en el apéndice B.2.3 y apéndice B.2.4).

Se determinó que el año de partición de entrenamiento y de test es 2006. Se hace el estudio de dos indicadores, DOLLARS y QUANTITY de los cuales se han obtenido los siguientes resultados, que son similares.

4.4.1 AROMADB INDICADOR *DOLLARS*

Para el caso del indicador **DOLLARS** representa la cantidad en dólares de las ventas de un producto en un determinado supermercado, entre otras características, podemos ver en la ilustración 5 (con un mejor detalle y resolución) y se puede ver poca información interesante), que el modelo que mejor se comporta para este caso es el KNN, le sigue el modelo LRW solapándose con el modelo M5P y podemos ver que aparece el modelo MEAN con un error cuadrático similar. Para todas estas técnicas las dos aproximaciones tienen similares resultados a excepción de la técnica SVM que su error cuadrático a *lowest level* se dispara pero a *same level* obtiene resultados similares a todas las otras técnicas.

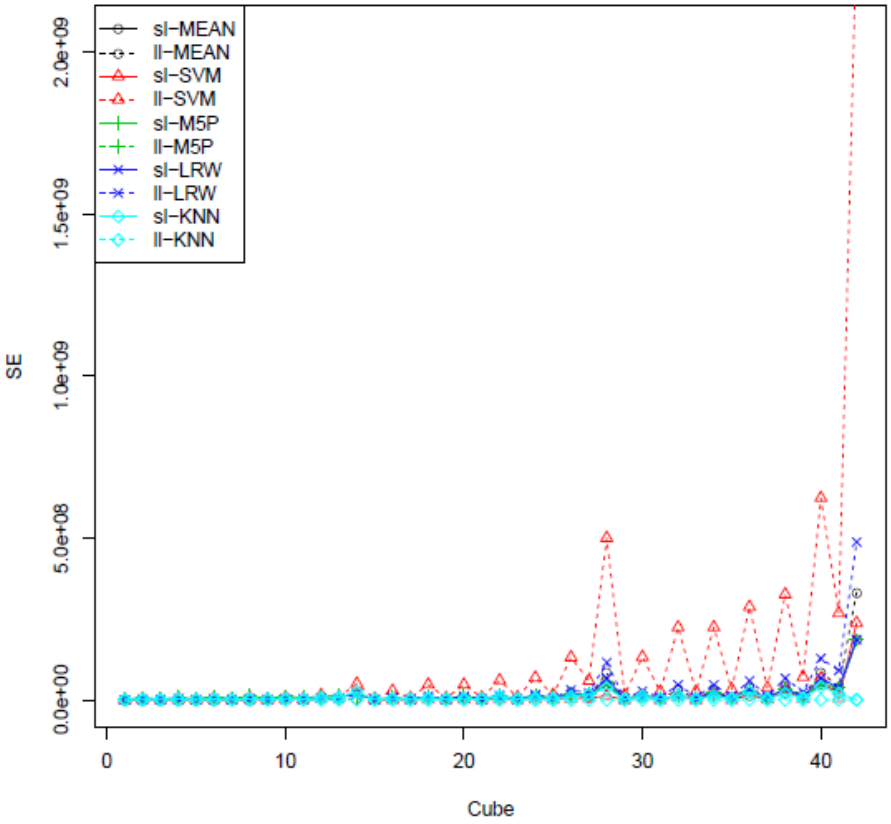

Ilustración 5: Resultados experimento conjunto AROMADB - DOLLARS (1)

Para poder distinguir los resultados mejor, a continuación la ilustración 6 con más zoom en el cual se puede distinguir algo mejor las diferencias. Se ve claramente que el mejor modelo que se está comportando es KNN con la aproximación *same level,* luego todos los otros aparecen solapados con lo cual el error cuadrático es similar para las dos aproximaciones, quitando que el error que produce el modelo SVM a *lowest level* es mayor que todos los otros en las dos aproximaciones.

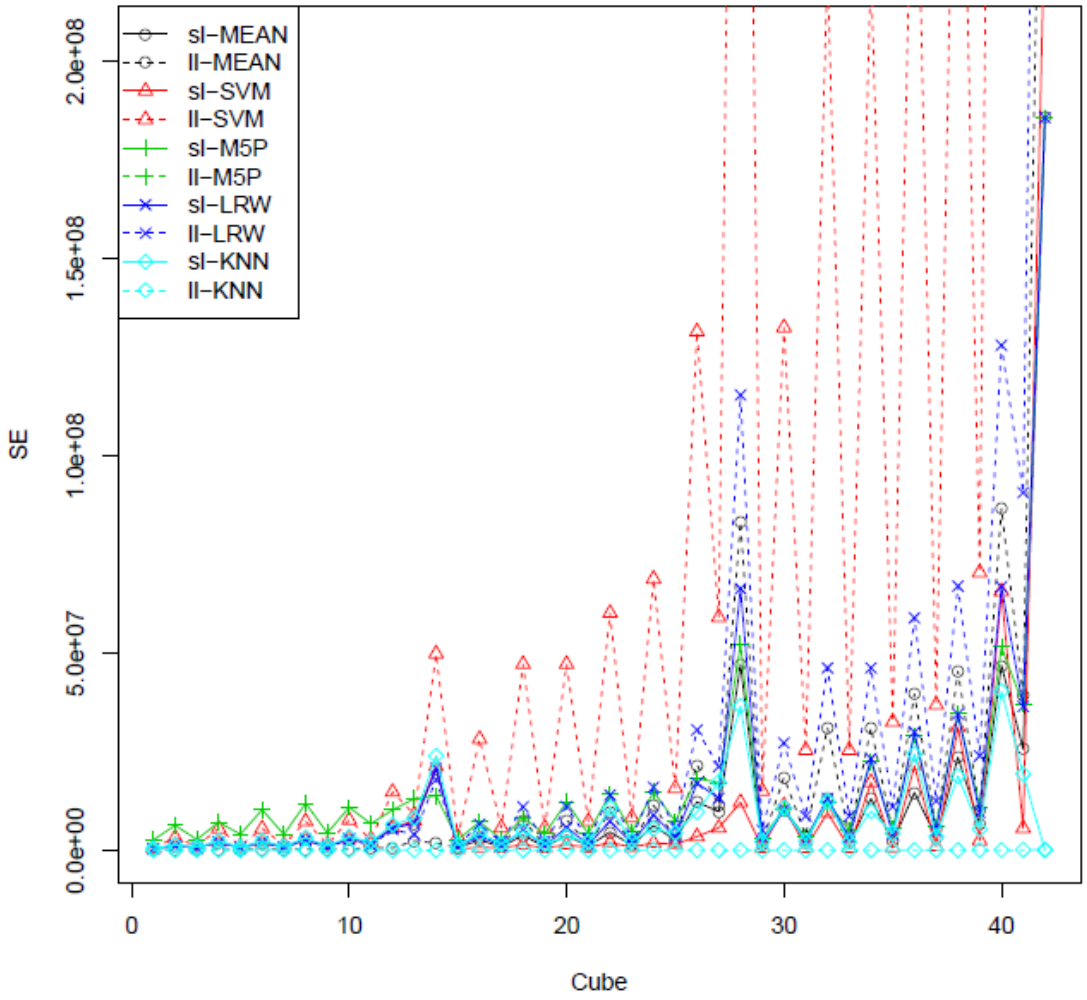

Ilustración 6: Resultados experimento conjunto AROMADB - DOLLARS (2)

El experimento concluye con los resultados de todas las técnicas aplicadas a los cubos generados. En la tabla 10, podemos ver para la técnica MEAN, el listado de cubos generados, la cantidad de muestras para cada conjunto (train/test), los errores de cada cubo y si la aproximación *lowest level* obtiene un error mejor, peor o igual que la de *same level*.

	Resolution	Train	Test	sl-MEAN	ll-MEAN	ll_best -MEAN
1	KEY_PROMO, KEY_STORE, KEY_CLASS, SAL_YEAR	59616	29808	392854,19	146152,01	WIN
2	KEY_PROMO, KEY_STORE, ---, SAL_YEAR	6624	3312	1187954,15	137929,68	WIN
3	KEY_PROMO, KEY_MKT, KEY_CLASS, SAL_YEAR	49680	24840	653372,19	241899,59	WIN
4	KEY_PROMO, KEY_MKT, ---, SAL_YEAR	5520	2760	2044762,31	205463,85	WIN
5	KEY_PROMO, MKT_HQ_CITY, KEY_CLASS, SAL_YEAR	49680	24840	653372,19	241899,59	WIN
6	KEY_PROMO, MKT_HQ_CITY, ---, SAL_YEAR	5520	2760	2044762,31	205463,85	WIN
7	KEY_PROMO, MKT_HQ_STATE, KEY_CLASS, SAL_YEAR	46368	23184	826465,14	312048,28	WIN
8	KEY_PROMO, MKT_HQ_STATE, ---, SAL_YEAR	5152	2576	2711773,04	255320,35	WIN
9	KEY_PROMO, MKT_DISTRICT, KEY_CLASS, SAL_YEAR	26496	13248	950457,35	345740,27	WIN
10	KEY_PROMO, MKT_DISTRICT, ---, SAL_YEAR	2944	1472	2934265,43	293949,46	WIN
11	KEY_PROMO, MKT_REGION, KEY_CLASS, SAL_YEAR	13248	6624	1805859,41	655042,27	WIN
12	KEY_PROMO, MKT_REGION, ---, SAL_YEAR	1472	736	5616450,32	560734,48	WIN
13	KEY_PROMO, ---, KEY_CLASS, SAL_YEAR	3312	1656	6850957,01	2362597,08	WIN
14	KEY_PROMO, ---, ---, SAL_YEAR	368	184	20741624,60	1810121,94	WIN
15	PROMO_TYPE, KEY_STORE, KEY_CLASS, SAL_YEAR	1944	972	556214,35	649528,91	LOSE
16	PROMO_TYPE, KEY_STORE, ---, SAL_YEAR	216	108	2656534,84	4667200,35	LOSE
17	PROMO_TYPE, KEY_MKT, KEY_CLASS, SAL_YEAR	1620	810	849457,05	1081257,80	LOSE
18	PROMO_TYPE, KEY_MKT, ---, SAL_YEAR	180	90	3807253,42	7758230,39	LOSE
19	PROMO_TYPE, MKT_HQ_CITY, KEY_CLASS, SAL_YEAR	1620	810	849457,05	1081257,80	LOSE
20	PROMO_TYPE, MKT_HQ_CITY, ---, SAL_YEAR	180	90	3807253,42	7758230,39	LOSE

21	PROMO_TYPE, MKT_HQ_STATE, KEY_CLASS, SAL_YEAR	1512	756	1036621,57	1375093,98	LOSE
22	PROMO_TYPE, MKT_HQ_STATE, ---, SAL_YEAR	168	84	4600360,98	9820973,39	LOSE
23	PROMO_TYPE, MKT_DISTRICT, KEY_CLASS, SAL_YEAR	864	432	1318048,02	1576516,61	LOSE
24	PROMO_TYPE, MKT_DISTRICT, ---, SAL_YEAR	96	48	6238782,13	11368219,75	LOSE
25	PROMO_TYPE, MKT_REGION, KEY_CLASS, SAL_YEAR	432	216	2540976,39	3004611,94	LOSE
26	PROMO_TYPE, MKT_REGION, ---, SAL_YEAR	48	24	12225528,79	21701762,25	LOSE
27	PROMO_TYPE, ---, KEY_CLASS, SAL_YEAR	108	54	9791329,13	11423358,22	LOSE
28	PROMO_TYPE, ---, ---, SAL_YEAR	12	6	47177694,41	83337114,06	LOSE
29	---, KEY_STORE, KEY_CLASS, SAL_YEAR	324	162	1438287,17	2200116,65	LOSE
30	---, KEY_STORE, ---, SAL_YEAR	36	18	10357182,96	18490935,35	LOSE
31	---, KEY_MKT, KEY_CLASS, SAL_YEAR	270	135	1927399,81	3696627,52	LOSE
32	---, KEY_MKT, ---, SAL_YEAR	30	15	13106384,31	31084504,91	LOSE
33	---, MKT_HQ_CITY, KEY_CLASS, SAL_YEAR	270	135	1927399,81	3696627,52	LOSE
34	---, MKT_HQ_CITY, ---, SAL_YEAR	30	15	13106384,31	31084504,91	LOSE
35	---, MKT_HQ_STATE, KEY_CLASS, SAL_YEAR	252	126	2203292,54	4712516,73	LOSE
36	---, MKT_HQ_STATE, ---, SAL_YEAR	28	14	14605753,82	39596534,70	LOSE
37	---, MKT_DISTRICT, KEY_CLASS, SAL_YEAR	144	72	3310570,60	5387630,28	LOSE
38	---, MKT_DISTRICT, ---, SAL_YEAR	16	8	23589505,52	45352300,11	LOSE
39	---, MKT_REGION, KEY_CLASS, SAL_YEAR	72	36	6514123,14	10298013,84	LOSE
40	---, MKT_REGION, ---, SAL_YEAR	8	4	46898159,87	86739358,77	LOSE
41	---, ---, KEY_CLASS, SAL_YEAR	18	9	25662910,60	39340982,95	LOSE
42	---, ---, ---, SAL_YEAR	2	1	185756506,97	332207482,04	LOSE

Tabla 10: Resultados técnica MEAN conjunto AROMADB (Dollars)

4.4.2 AROMADB INDICADOR *QUANTITY*

El indicador **QUANTITY** representa la cantidad de productos vendidos en un supermercado. Los resultados a nivel general son similares a los del indicador **DOLLARS**, pero existe alguna diferencia que no podemos ver en la ilustración 7 pero si en la siguiente la cual tiene más zoom. Sólo podemos ver que la técnica SVM a *lowest level* tiene un error mayor que todas las otras técnicas (en las dos aproximaciones) que se solapan en un error menor incluido la *same level* de SVM.

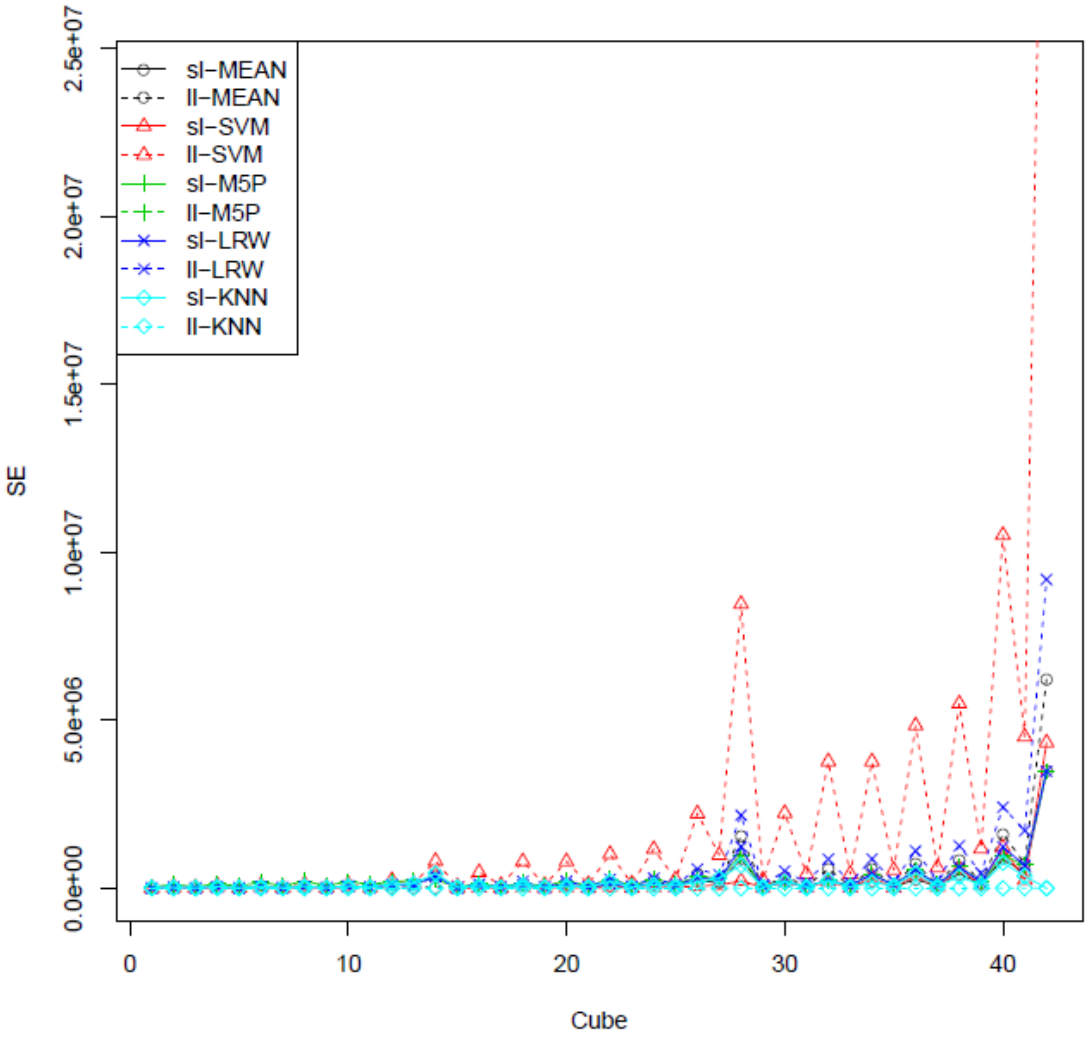

Ilustración 7: Resultados experimento conjunto AROMADB-QUANTITY (1)

A continuación se muestra con más zoom la ilustración 8 en el cual se pueden visualizar los datos mejor y distinguir mejor los resultados obtenidos. Destacar que para este indicador, los errores cuadráticos en general son más altos que el otro indicador, pero el comportamiento de las dos aproximaciones para todas las técnicas no ha variado. Vemos que la técnica SVM a *lowest level* es la que mayor error obtiene seguida de la LRW a *lowest level* también. Luego, ya no se puede distinguir todas las otras técnicas porque están solapadas entre ellas, pero se podría decir que tienen un error cuadrático similar.

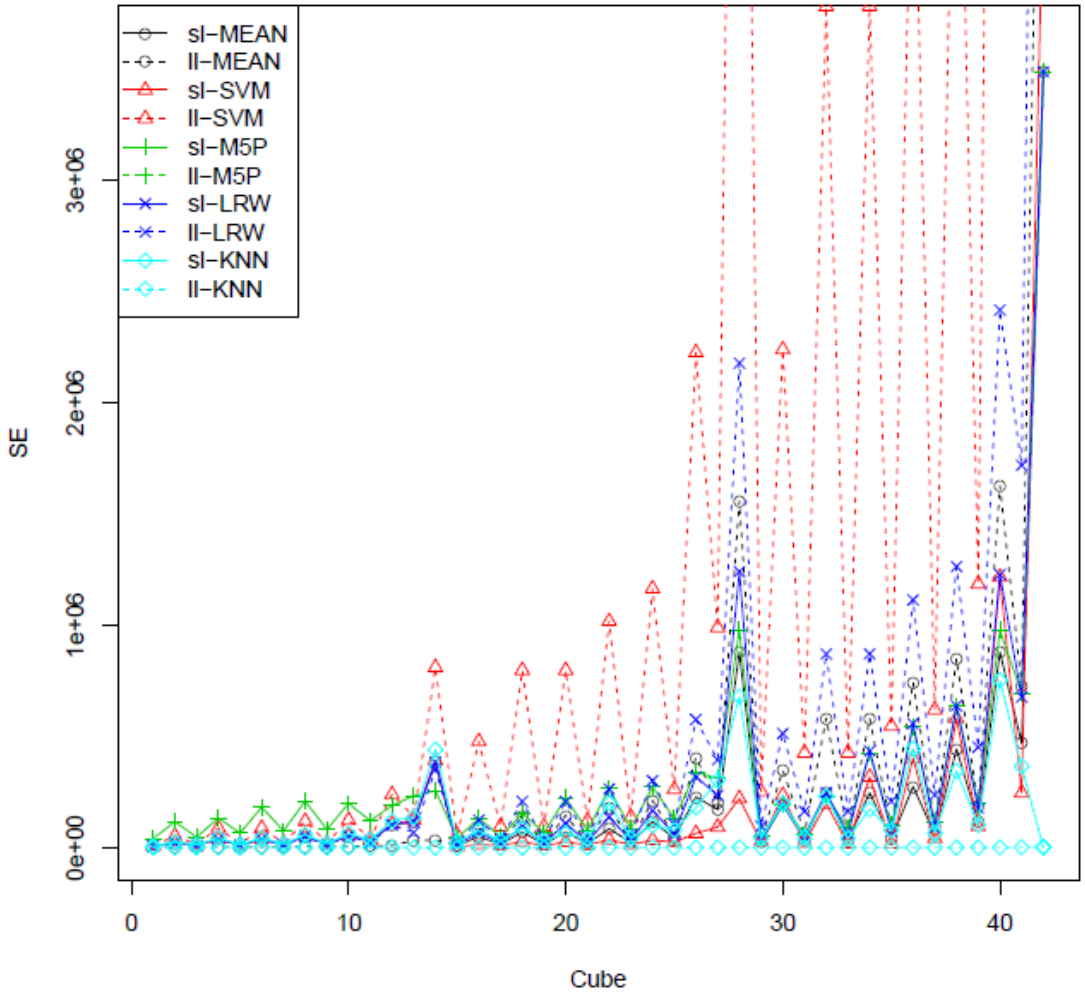

Ilustración 8: Resultados experimento conjunto AROMADB-QUANTITY (2)

El experimento concluye con los resultados de todas las técnicas aplicadas a los cubos generados.

En la tabla 11 podemos ver para la técnica MEAN, el listado de cubos generados, la cantidad de muestras para cada conjunto (train/test), los errores de cada cubo y si la aproximación *lowest level* obtiene un error mejor, peor o igual que la de *same level*.

	Resolution	Train	Test	sl-MEAN	ll-MEAN	ll_best-MEAN
1	KEY_PROMO, KEY_STORE, KEY_CLASS, SAL_YEAR	59616	29808	6693,18	1965,83	WIN
2	KEY_PROMO, KEY_STORE, ---, SAL_YEAR	6624	3312	22106,87	2468,80	WIN
3	KEY_PROMO, KEY_MKT, KEY_CLASS, SAL_YEAR	49680	24840	11204,51	3213,47	WIN
4	KEY_PROMO, KEY_MKT, ---, SAL_YEAR	5520	2760	38339,89	3682,61	WIN
5	KEY_PROMO, MKT_HQ_CITY, KEY_CLASS, SAL_YEAR	49680	24840	11204,51	3213,47	WIN
6	KEY_PROMO, MKT_HQ_CITY, ---, SAL_YEAR	5520	2760	38339,89	3682,61	WIN
7	KEY_PROMO, MKT_HQ_STATE, KEY_CLASS, SAL_YEAR	46368	23184	14186,98	4136,38	WIN
8	KEY_PROMO, MKT_HQ_STATE, ---, SAL_YEAR	5152	2576	50159,89	4665,36	WIN
9	KEY_PROMO, MKT_DISTRICT, KEY_CLASS, SAL_YEAR	26496	13248	16196,60	4640,03	WIN
10	KEY_PROMO, MKT_DISTRICT, ---, SAL_YEAR	2944	1472	54709,42	5398,50	WIN
11	KEY_PROMO, MKT_REGION, KEY_CLASS, SAL_YEAR	13248	6624	30702,05	8812,38	WIN
12	KEY_PROMO, MKT_REGION, ---, SAL_YEAR	1472	736	104465,39	10314,92	WIN
13	KEY_PROMO, ---, KEY_CLASS, SAL_YEAR	3312	1656	115232,17	32123,07	WIN
14	KEY_PROMO, ---, ---, SAL_YEAR	368	184	383715,55	34076,66	WIN
15	PROMO_TYPE, KEY_STORE, KEY_CLASS, SAL_YEAR	1944	972	9757,72	11372,94	LOSE
16	PROMO_TYPE, KEY_STORE, ---, SAL_YEAR	216	108	49650,74	87107,99	LOSE
17	PROMO_TYPE, KEY_MKT, KEY_CLASS, SAL_YEAR	1620	810	14883,50	18895,40	LOSE
18	PROMO_TYPE, KEY_MKT, ---, SAL_YEAR	180	90	71398,99	144787,23	LOSE
19	PROMO_TYPE, MKT_HQ_CITY, KEY_CLASS, SAL_YEAR	1620	810	14883,50	18895,40	LOSE

20	PROMO_TYPE, MKT_HQ_CITY, ---, SAL_YEAR	180	90	71398,99	144787,23	LOSE
21	PROMO_TYPE, MKT_HQ_STATE, KEY_CLASS, SAL_YEAR	1512	756	18130,24	23995,58	LOSE
22	PROMO_TYPE, MKT_HQ_STATE, ---, SAL_YEAR	168	84	85586,08	183358,94	LOSE
23	PROMO_TYPE, MKT_DISTRICT, KEY_CLASS, SAL_YEAR	864	432	23092,46	27637,88	LOSE
24	PROMO_TYPE, MKT_DISTRICT, ---, SAL_YEAR	96	48	116688,88	212321,65	LOSE
25	PROMO_TYPE, MKT_REGION, KEY_CLASS, SAL_YEAR	432	216	44492,54	52718,02	LOSE
26	PROMO_TYPE, MKT_REGION, ---, SAL_YEAR	48	24	228426,63	405359,46	LOSE
27	PROMO_TYPE, ---, KEY_CLASS, SAL_YEAR	108	54	170392,12	201449,74	LOSE
28	PROMO_TYPE, ---, ---, SAL_YEAR	12	6	879561,85	1557602,01	LOSE
29	---, KEY_STORE, KEY_CLASS, SAL_YEAR	324	162	26352,11	40458,02	LOSE
30	---, KEY_STORE, ---, SAL_YEAR	36	18	194458,26	346847,56	LOSE
31	---, KEY_MKT, KEY_CLASS, SAL_YEAR	270	135	35182,35	67813,12	LOSE
32	---, KEY_MKT, ---, SAL_YEAR	30	15	246493,14	581770,36	LOSE
33	---, MKT_HQ_CITY, KEY_CLASS, SAL_YEAR	270	135	35182,35	67813,12	LOSE
34	---, MKT_HQ_CITY, ---, SAL_YEAR	30	15	246493,14	581770,36	LOSE
35	---, MKT_HQ_STATE, KEY_CLASS, SAL_YEAR	252	126	40116,29	86686,77	LOSE
36	---, MKT_HQ_STATE, ---, SAL_YEAR	28	14	274118,03	743507,50	LOSE
37	---, MKT_DISTRICT, KEY_CLASS, SAL_YEAR	144	72	60561,61	99034,95	LOSE
38	---, MKT_DISTRICT, ---, SAL_YEAR	16	8	443015,31	850000,85	LOSE
39	---, MKT_REGION, KEY_CLASS, SAL_YEAR	72	36	119195,79	189454,02	LOSE
40	---, MKT_REGION, ---, SAL_YEAR	8	4	880451,83	1626562,04	LOSE
41	---, ---, KEY_CLASS, SAL_YEAR	18	9	468695,03	725238,17	LOSE
42	---, ---, ---, SAL_YEAR	2	1	3484998,50	6233887,23	LOSE

Tabla 11: Resultados técnica MEAN conjunto AROMADB (Quantity)

4.5 CONJUNTO DE DATOS *GEN-VARIATIONS*

Gen Variations (Pastor, y otros, 2012), (Celma & Martin, 2011) es un conjunto de datos proporcionado por el centro de investigación en métodos de producción de software (PROS), que contiene variaciones genéticas con todas sus características (detallado en el apéndice B.3).

Para obtener las predicciones y poder realizar el conteo, se ha creado un nuevo indicador COUNT que le asigna el valor entero 1 a cada una de las filas para luego realizar las agregaciones y predicciones de los cubos. En el conjunto de datos Gen Variations, se definen las siguientes jerarquías:

- SPEC (ESPEC.Eff, ROLLED_UP_LEVEL)
- DBANK (DBANK,ROLLED_UP_LEVEL)
- PHENOTYPE (PHENO.Name,PHENO.ICD10,PHENO.ICD10.Cat,ROLLED_UP_LEVEL)
- GENOTYPE (GENO_ID, GENO_Chrom, ROLLED_UP_LEVEL)
- DATE (DATE.Year)

Con todas ellas, hemos obtenido 2x2x4x3x1 cubos que generan un total de 48 cubos (detallado en apéndice B.3.3 y apéndice B.3.4).

En la ilustración 9, se puede distinguir que la que peor resultados obtiene es el modelo SVM a *lowest level*. Luego, que el modelo que mejor se comporta es el KNN a *lowest level* y por último aparecen todos los otros modelos con las dos aproximaciones que estan solapadas y no se distingue la diferencia entre ellas.

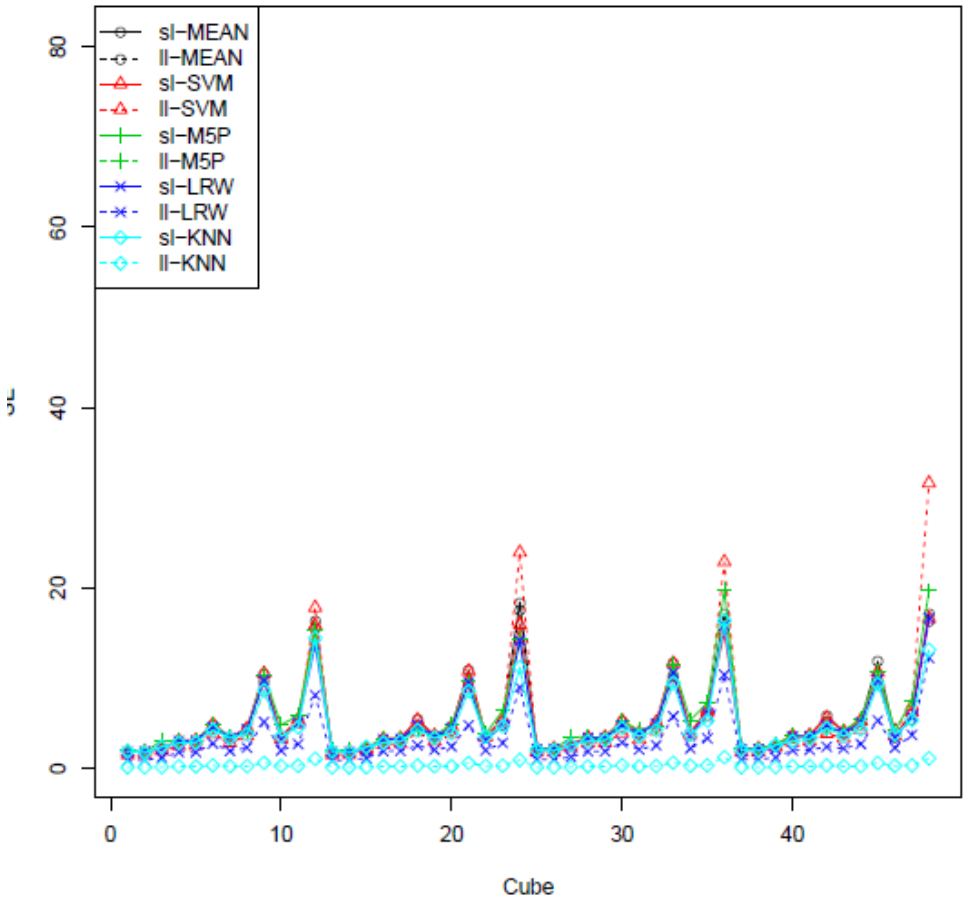

Ilustración 9: Resultados experimento conjunto AROMADB - DOLLARS (1)

En la ilustración 10, se ve claramente que el modelo que mejores resultados obtiene es el KNN a *lowest level*. Le sigue LRW a *lowest level* y luego ya podemos ver que todos los otros modelos tanto para la aproximación *lowest level* como *same level* obtienen similares resultados sin poder destacar ninguno de ellos.

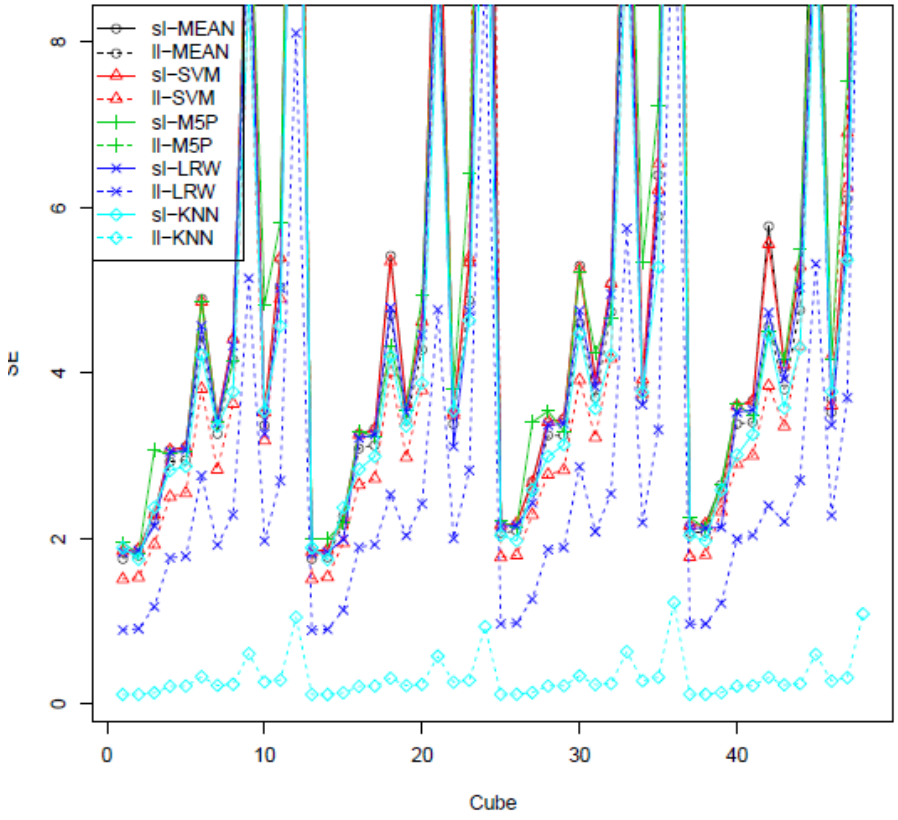

Ilustración 10: Resultados experimento conjunto AROMADB - DOLLARS (2)

Para finalizar el experimento, sólo queda ver la tabla 12 en la cual se representan los resultados de la técnica MEAN, con un listado de los cubos generados, la cantidad de muestras para los conjuntos (train/test), los errores generados para cada cubo y si la aproximación *lowest level* es mejor, igual o peor que la aproximación *same level*.

	Resolution	Train	Test	sl-MEAN	ll-MEAN	ll_best-MEAN
1	SPEC,Eff, DBANK, PHENO,Name, DATE,Year, GENO,ID	164160	47880	1,84	1,77	WIN
2	SPEC,Eff, DBANK, PHENO,Name, DATE,Year, GENO,Chrom	109440	31920	1,87	1,79	WIN
3	SPEC,Eff, DBANK, PHENO,Name, DATE,Year, ---	10944	3192	2,30	2,21	WIN
4	SPEC,Eff, DBANK, PHENO,ICD10, DATE,Year, GENO,ID	56160	16380	3,07	2,93	WIN
5	SPEC,Eff, DBANK, PHENO,ICD10, DATE,Year, GENO,Chrom	37440	10920	3,11	2,95	WIN
6	SPEC,Eff, DBANK, PHENO,ICD10, DATE,Year, ---	3744	1092	4,90	4,46	WIN
7	SPEC,Eff, DBANK, PHENO,ICD10,Cat,	14400	4200	3,45	3,28	WIN

	DATE,Year, GENO,ID					
8	SPEC,Eff, DBANK, PHENO,ICD10,Cat, DATE,Year, GENO,Chrom	9600	2800	4,41	4,16	WIN
9	SPEC,Eff, DBANK, PHENO,ICD10,Cat, DATE,Year, ---	960	280	10,53	9,64	WIN
10	SPEC,Eff, DBANK, ---, DATE,Year, GENO,ID	1440	420	3,52	3,37	WIN
11	SPEC,Eff, DBANK, ---, DATE,Year, GENO,Chrom	960	280	5,37	5,04	WIN
12	SPEC,Eff, DBANK, ---, DATE,Year, ---	96	28	16,32	15,64	WIN
13	SPEC,Eff, ---, PHENO,Name, DATE,Year, GENO,ID	82080	23940	1,84	1,77	WIN
14	SPEC,Eff, ---, PHENO,Name, DATE,Year, GENO,Chrom	54720	15960	1,87	1,79	WIN
15	SPEC,Eff, ---, PHENO,Name, DATE,Year, ---	5472	1596	2,28	2,19	WIN
16	SPEC,Eff, ---, PHENO,ICD10, DATE,Year, GENO,ID	28080	8190	3,26	3,09	WIN
17	SPEC,Eff, ---, PHENO,ICD10, DATE,Year, GENO,Chrom	18720	5460	3,32	3,13	WIN
18	SPEC,Eff, ---, PHENO,ICD10, DATE,Year, ---	1872	546	5,42	4,71	WIN
19	SPEC,Eff, ---, PHENO,ICD10,Cat, DATE,Year, GENO,ID	7200	2100	3,63	3,42	WIN
20	SPEC,Eff, ---, PHENO,ICD10,Cat, DATE,Year, GENO,Chrom	4800	1400	4,62	4,29	WIN
21	SPEC,Eff, ---, PHENO,ICD10,Cat, DATE,Year, ---	480	140	10,94	9,58	WIN
22	SPEC,Eff, ---, ---, DATE,Year, GENO,ID	720	210	3,54	3,39	WIN
23	SPEC,Eff, ---, ---, DATE,Year, GENO,Chrom	480	140	5,33	4,89	WIN
24	SPEC,Eff, ---, ---, DATE,Year, ---	48	14	18,45	17,65	WIN
25	---, DBANK, PHENO,Name, DATE,Year, GENO,ID	82080	23940	2,16	2,07	WIN
26	---, DBANK, PHENO,Name, DATE,Year, GENO,Chrom	54720	15960	2,18	2,09	WIN
27	---, DBANK, PHENO,Name, DATE,Year, ---	5472	1596	2,70	2,58	WIN
28	---, DBANK, PHENO,ICD10, DATE,Year, GENO,ID	28080	8190	3,41	3,24	WIN
29	---, DBANK, PHENO,ICD10, DATE,Year, GENO,Chrom	18720	5460	3,45	3,25	WIN
30	---, DBANK, PHENO,ICD10, DATE,Year, ---	1872	546	5,31	4,61	WIN
31	---, DBANK, PHENO,ICD10,Cat, DATE,Year, GENO,ID	7200	2100	3,92	3,70	WIN
32	---, DBANK, PHENO,ICD10,Cat, DATE,Year, GENO,Chrom	4800	1400	5,08	4,74	WIN
33	---, DBANK, PHENO,ICD10,Cat,	480	140	11,71	10,31	WIN

	DATE,Year, ---					
34	---, DBANK, ---, DATE,Year, GENO,ID	720	210	3,85	3,69	WIN
35	---, DBANK, ---, DATE,Year, GENO,Chrom	480	140	6,39	5,91	WIN
36	---, DBANK, ---, DATE,Year, ---	48	14	17,18	16,43	WIN
37	---, ---, PHENO,Name, DATE,Year, GENO,ID	41040	11970	2,16	2,07	WIN
38	---, ---, PHENO,Name, DATE,Year, GENO,Chrom	27360	7980	2,18	2,09	WIN
39	---, ---, PHENO,Name, DATE,Year, ---	2736	798	2,65	2,54	WIN
40	---, ---, PHENO,ICD10, DATE,Year, GENO,ID	14040	4095	3,61	3,39	WIN
41	---, ---, PHENO,ICD10, DATE,Year, GENO,Chrom	9360	2730	3,67	3,40	WIN
42	---, ---, PHENO,ICD10, DATE,Year, ---	936	273	5,78	4,57	WIN
43	---, ---, PHENO,ICD10,Cat, DATE,Year, GENO,ID	3600	1050	4,09	3,81	WIN
44	---, ---, PHENO,ICD10,Cat, DATE,Year, GENO,Chrom	2400	700	5,27	4,77	WIN
45	---, ---, PHENO,ICD10,Cat, DATE,Year, ---	240	70	11,87	9,55	WIN
46	---, ---, ---, DATE,Year, GENO,ID	360	105	3,70	3,54	WIN
47	---, ---, ---, DATE,Year, GENO,Chrom	240	70	6,09	5,39	WIN
48	---, ---, ---, DATE,Year, ---	24	7	17,25	16,43	WIN

Tabla 12 : Resultados técnica MEAN del conjunto Gen Variations

5 CONCLUSIONES Y TRABAJO FUTURO

Hemos utilizado los conjuntos de datos para entrenarlos y generar modelos de minería de datos a diferentes niveles de agregación (*lowest level, same level*), para obtener cubos de datos agregados con sus predicciones correspondientes, como ya se ha descrito anteriormente.

La búsqueda y exploración de los conjuntos de datos (detallado en apéndice B.4) que nos pudieran servir para los experimentos no ha sido fácil, ya que, para que un conjunto se adaptara a las necesidades de los experimentos realizados necesitaba ser un conjunto con una estructura multidimensional, con varias jerarquías con al menos dos niveles y además que mantuviera un histórico de la información a un nivel suficiente de detalle.

El formato original de los datos no importa si es en un tipo de base de datos, en ficheros planos, en Excel, etc. Se realizan los procesos de transformación adecuados, ya sea usando SQL o scripts. Al final, el conjunto de datos **Gen Variations** se nos ha proporcionado por el grupo PROS (Celma & Martin, 2011) (Pastor, y otros, 2012) que contiene variaciones genéticas estaba originalmente en una base de datos. El conjunto **Cars Fuel Emissions** (VCA, 2014) es libre, gratuito y además tiene muchos indicadores que se pueden estudiar, aunque nosotros nos hemos centrado solo en el consumo de CO_2 de los coches, y se tuvo que transformar de formatos planos. Para terminar, el último conjunto **DBAROMA** (IBM, 2014) es un conjunto con datos ficticios de ventas de una red de supermercados proporcionada gratuitamente por la página de IBM, que tenía una estructura relacional.

Los conjuntos de datos disponibles para los experimentos se han tenido que preprocesar en tamaño para adaptarlos a las necesidades requeridas. Por ejemplo, el conjunto de datos Cars Fuel Emission (detallado en apéndice B.1.2). También reducir el conjunto de Cars Fuel Emission, ya que este tiene una variabilidad de sus atributos

muy grande y a consecuencia de esto, el coste de ejecución del experimento es muy elevado.

La gestión de la agregación en R no es sencilla, ya que no es un lenguaje de base de datos, y la gestión en memoria de todos los cubos no es sencilla. La representación gráfica de resultados y su comparación también es una tarea compleja, dado la cantidad de cubos y de modelos a comparar. Como es habitual, estas son tareas poco vistosas pero que han consumido un gran porcentaje del esfuerzo realizado en el proyecto. No obstante, todo esto era necesario para tener una visión global de las dos aproximaciones, tal y como era el objetivo del proyecto y que analizamos a continuación.

5.1 DISCUSIÓN

Una vez ya ejecutados y analizados los experimentos, parece claro que sí que hay diferencias entre las dos aproximaciones. También es claro que no hay ninguna que funcione mejor que la otra de manera generalizada. Lo que no queda claro es cuál de las aproximaciones descritas se comporta mejor en un escenario u otro.

Las gráficas que se han extraído de los experimentos no hacen fácil la tarea de distinguir cual obtiene menor error cuadrático de forma clara, ya que en prácticamente todos los casos la mayoría de las técnicas se solapa con otras en las dos aproximaciones estudiadas, *lowest level* y *same level*.

Lo que sí podemos decir es que la técnica que parece ser se comporte mejor es la KNN a *lowest level* en muchos de los experimentos y en algún caso a *same level*. Todas las otras técnicas se solapan entre ellas destacando que en alguno de los casos, como puede ser SVM a *lowest level* produce un error cuadrático muy grande para muchos de los experimentos.

El aprendizaje en el nivel más bajo parece ser la forma más versátil y económica, por lo menos, para prácticamente todas las técnicas aunque tal vez no para SVM. Pero la agregación del modelo a nivel más bajo no siempre da los mejores resultados. Pero no hemos visto un patrón que aclare que el modelo a nivel más bajo es mejor que el modelo del mismo nivel o viceversa.

Por tanto, la conclusión principal del trabajo es que la aproximación *lowest level* es una aproximación tan o más válida como la de *same level*. De hecho, es la primera que debería probarse en una aplicación real, dado su coste (sólo se entrena un modelo). Si existen los recursos y capacidad de mantener muchos modelos, la aproximación *same level* también debería explorarse, ya que en algunos casos y dependiendo de las técnicas puede dar mejores resultados.

5.2 TRABAJO FUTURO

Existen algunos aspectos que seguramente son importantes explorar y profundizar en trabajos futuros y que, seguramente, permitirán mejorar los resultados obtenidos, ya sea con conjuntos de datos distintos o mejorando el script MDHM con otras funcionalidades que obtengan mejores resultados y más claros. A continuación se describen algunos aspectos:

Mejorar el rendimiento del Script: en muchos de los casos que nos hemos encontrado, hemos tenido problemas de rendimiento de memoria física por la cantidad de combinaciones que la técnica de rellenado de ceros nos creaba. Se puede estudiar y modificar el script para una mejor gestión de este caso.

Nuevos conjuntos de datos: sería interesante coger nuevos conjuntos de datos de diferentes características a los utilizados en este trabajo y obtener los resultados para compararlos con estos y ver si siguen comportándose de la misma forma o se puede distinguir algún comportamiento que destaque y/o nos aclare en algunos de los casos, qué modelo se comporta mejor entre todos los que no hemos podido visualizar de forma clara en las gráficas.

Nuevos técnicas de modelado: También afectados por el rendimiento, para este trabajo se ha intentado utilizar diferentes modelos que al final se han tenido que descartar por no disponer de máquinas que pudieran soportar la memoria necesaria para la ejecución de estos. Por lo tanto nos hemos quedado con los descritos en el (detallado en la sección 3.2), pero otras técnicas podrían explorarse.

Nuevas gráficas: Como se ha visto en todos los experimentos, no hemos encontrado un patrón claro para distinguir muchas de las técnicas utilizadas para los experimentos en las gráficas generadas con R, teniendo en el eje X el número de cubo y en el eje Y el error cuadrático medio de cada uno. Es necesario encontrar nuevas formas de

representar la información para distinguir de una forma clara y concisa las técnicas, y así poder sacar mejores conclusiones.

BIBLIOGRAFÍA Y APÉNDICES

A BIBLIOGRAFÍA

(15 de Abril de 2014). Obtenido de R Studio: http://www.rstudio.com/

Bristol, U., Strasbourg, U. d., & Valencia, U. P. (2013). Obtenido de Reframe d2k project: http://www.reframe-d2k.org/

Celma, M., & Martin, A. (2011). Integrating Human Genome Variation Data: An Information System Approach. In Database and Expert Systems Applications (DEXA). *22nd International Workshop*, 65-69.

Cristianini-N, & Scholkopf-B. (2002). Support vector machines and kernel methods: the new. *Ai Magazine*, 23(3):31.

Descarga R Project. (2014). Obtenido de R: http://cran.r-project.org/bin/windows/base/

Han, J., Kamber, M., & Pei, J. (2012). *Data Mining: Concepts and Techniques.*

Celma Giménez, M., Hernández-Orallo, J. (2012). Almacenes de Datos y Minería de Datos - Tema 1. Valencia.

Hernández-Orallo, J., Ramirez-Quintana, M. J., & Ferri-Ramirez, C. (2004). *Introducción a la Minería de Datos.* Hernández-Orallo, Ramírez-Quintana y Ferri: Pearson.

IBM. (18 de 01 de 2014). Obtenido de IBM Developer Works: http://www.ibm.com/developerworks/data/tutorials/dm0607cao/dm0607cao-pdf.pdf

Pastor, O., Casamayor, J., Celma, M., Mota, L., Pastor, M., & Levin, A. (2012). Conceptual modeling of human genome: integration challenges. In Conceptual Modelling and Its Theoretical Foundations. *Springer Berlin Heidelberg*, 231-250.

Project, R. (20 de 05 de 2014). Obtenido de R Project: http://www.r-project.org/

Quinlan, R. J. (1992). 5th Australian Joint Conference on Artificial Intelligence. (W. Scientific, Ed.) 343-348.

VCA. (16 de 03 de 2014). Obtenido de Car Fuel Data: http://carfueldata.direct.gov.uk/downloads/default.aspx

Witten, I. H., & Frank, E. (2000). Practical machine learning tools and techniques with java implementations.

B DETALLES CONJUNTOS DE DATOS

Los conjuntos de datos escogidos y detallados a continuación han sido fruto de una búsqueda extensa a través de la red. El listado completo de todos los sitios visitados para encontrar conjuntos válidos para los experimentos se detalla en el apéndice B.4.

Se han utilizado 3 conjuntos de datos con disponibilidad inmediata y libre de licencias, con características diferentes para analizar y comparar los resultados obtenidos después de aplicar las técnicas y modelos ya descritas. Estos están descritos en los siguientes apéndices: apéndice B1, apéndice B2 y apéndice B3.

B.1 CONJUNTO DE DATOS *CAR FUEL EMISSIONS*

Car Fuel Emissions (Car Fuel Data, 2014) nos proporciona las cifras del consumo de combustible, entre los años 2000 y 2013, en pruebas de condiciones específicas y por tanto, no necesariamente puede coincidir con las condiciones de la conducción en la vida real. Una gran variedad de factores influyen en el consumo real de combustible, por ejemplo, el estilo y comportamiento de la conducción, así como el medio ambiente y las condiciones bajo las cuales el vehículo está sometido. Este conjunto es una publicación de la Agencia de Certificación de Vehículos (VCA), una agencia del Reino Unido que proporciona este conjunto de datos de forma libre y gratuita.

El conjunto de datos **Car Fuel Emissions** se compone de una extensa variabilidad de todos sus atributos, lo cual, al realizar los experimentos, se nos ha hecho inviable tratar con dicha variabilidad, con lo cual se opta por pre procesar los datos para que haya menos variabilidad.

El conjunto resultante para el experimento no tiene todos los campos originales, ya que se han quitado los que no son relevantes para el estudio como se ha indicado anteriormente.

Tenemos un hecho principal DATA_CAR_FUEL_EMISSIONS en el cual existen numerosos indicadores para realizar predicciones. Luego la dimensión CAR nos describe la marca, modelo y descripción de los coches. En ENGINE, que representa la dimensión del motor, existe mucha variabilidad entre sus atributos y se ha reducido. La dimensión TIME representa el año del modelo de coche. La dimensión TRANSMISSION describe la transmisión y el tipo de transmisión que utiliza un coche. La dimensión FUEL tiene todos los tipos de combustible. La dimensión EURO describe el estándar europeo.

Más tarde, por problemas de memoria se ha tenido que realizar una selección de registros del conjunto original para una correcta ejecución de todas las técnicas.

En este conjunto, no existen valores a cero sino vacíos. Es decir, no debemos realizar la técnica desarrollada de rellenar con ceros para mejorar el resultado de los algoritmos. Además, se quiere calcular la media de consumo de CO^2 por lo tanto, se realiza la media aritmética en vez de la función de agregación sum. En todos los otros conjuntos de datos, utilizamos Sum, ya que es posible realizar un rellenado de ceros y por tanto, mejorar los resultados finales de la ejecución.

B.1.1 DISEÑO

Se muestra en la ilustración 11 un diagrama entidad relación para ver de forma detallada la estructura y todos los campos del conjunto de datos.

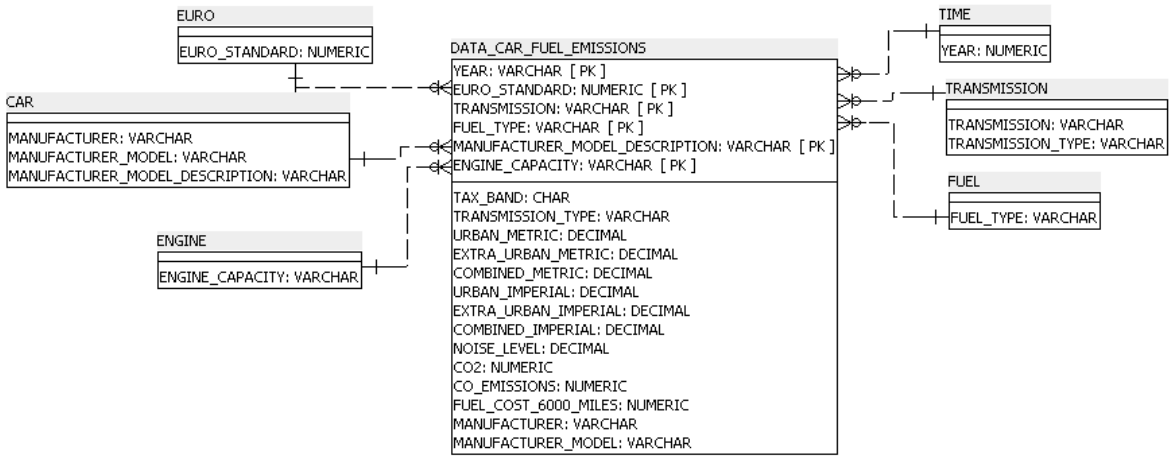

Ilustración 11: Diagrama entidad relación Data Car Fuel Emission

B.1.2 PREPARACIÓN

Para la adaptación de este conjunto de datos al script MDHM que estamos utilizando, se han realizado diferentes modificaciones en el conjunto para que facilitara la ejecución de las técnicas predictivas. Algunas de las modificaciones son:

- Establecer los tipos de campos adecuados para permitir la indexación y el análisis (por ejemplo, los campos numéricos).
- Normalizar fabricante y modelo descripciones.
- Recorte del exceso de espacios en blanco.
- Fijar codificación de caracteres especiales.
- Eliminación de 10 registros que contenían valores a N/A en el atributo TRANSMISSION.
- Eliminación de atributos no necesarios para el estudio, en las cuales el valor más frecuente es el NULL.
- Creación de una columna MANUFACTURER_MODEL con el resultado de la concatenación entre MANUFACTURER y MODEL. Se ha realizado para

deshacernos de una relación N a N y obtener como resultado una relación 1 a N.

- Creación de una nueva columna MANUFACTURER_MODEL_DESCRIPTION que es el resultado de la concatenación de MANUFACTURER, MODEL y DESCRIPTION. Se ha realizado para obtener una relación 1 a N.
- Categorización del atributo ENGINE que contenía valores 800,1100,2400,7000… a los siguientes rangos:
    - <1000
    - >=1000 and <2000
    - >=2000 and <3000
    - >=3000 and <4000
    - >=4000 and <5000
    - >=5000 and <6000
    - >6000

También, por la gran cantidad de variabilidad de MANUFACTURER que existían y gran volumen de registros, hemos tenido que realizar una limpieza de registros importante eliminando las marcas con menor frecuencia.

Como dato informativo se muestra en la tabla 13 los manufacturer que han sido eliminados del conjunto original, obteniendo un conjunto de datos más reducido que hemos llamado **Car Fuel Emissions Selection**.

MANUFACTURER	REGISTROS
Micro Compact Car	85
Vauxhall	4139
Bentley Motors	96
Abarth	6
Rolls-Royce	65
Jaguar Cars	453
Saab	902
Cadillac	172
Smart	268
Land Rover	302
Dodge	63
Chrysler Jeep	606
Chevrolet	442
Honda	1212
Daihatsu	169

Subaru	557
Mazda	613
Lotus	89
Mini	348
Mitsubishi	766
Morgan Motor Company	55
MG Rover Group	470
Porsche	606
Infiniti	47
Rover Group Limited	59
Perodua	61
Skoda	1349
Volvo	1877
Mercedes-Benz	5742
BMW	2769
Hyundai	641
Metrocab	22
Kia	721
Volkswagen	3578
Daewoo Cars	156
Audi	2794
Alfa Romeo	464
Fiat	698
Citroen	1138
Maserati	85
Aston Martin Lagonda	126
Renault	1848
Ford	2873

Tabla 13: Manufacturers descartados (CAR FUEL EMISSION)

También, se adjunta una relación en la tabla 14 con los manufacturers que sí que se utilizan para el experimento.

MANUFACTURERS
Toyota
Hummer
MG Motors Uk
Lexus
Peugeot
McLaren
Seat
Ferrari
Corvette
Lamborghini
Dacia
Tata

Tabla 14: Manufacturer considerados para experimento (CAR FUEL EMISSION)

B.1.3 DETALLE

Se describen en las siguientes tablas las dimensiones y hecho del conjunto de datos Car Fuel Emissions.

YEAR			
COLUMNA	**TIPO**	**DESCRIPCIÓN**	**RANGO VALORES**
YEAR	NUMERIC	Año del coche / modelo	[2000,2013]

Tabla 15: Dimensión YEAR (CAR FUEL EMISSION)

FUEL			
COLUMNA	**TIPO**	**DESCRIPCIÓN**	**RANGO VALORES**
FUEL_TYPE	VARCHAR	Tipo de combustible	Petrol, Gasoil…

Tabla 16: Dimensión FUEL (CAR FUEL EMISSION)

ENGINE			
COLUMNA	**TIPO**	**DESCRIPCIÓN**	**RANGO VALORES**
ENGINE	VARCHAR	Rango definido de potencia del motor.	○ <1000 ○ >=1000 and <2000 ○ >=2000 and <3000 ○ >=3000 and <4000 ○ >=4000 and <5000 ○ >=5000 and <6000 ○ >6000

Tabla 17: Dimensión ENGINE (CAR FUEL EMISSION)

TRANSMISSION			
COLUMNA	**TIPO**	**DESCRIPCIÓN**	**RANGO VALORES**
TRANSMISSION	VARCHAR	Transmisión	A5, A4, M4, M5…
TRANSMISSION_TYPE	VARCHAR2	Tipo de transmission.	Automatic, Manual…

Tabla 18: Dimensión TRANSMISSION (CAR FUEL EMISSION)

EURO			
COLUMNA	**TIPO**	**DESCRIPCIÓN**	**RANGO VALORES**
EURO_STANDARD	NUMERIC	Estandard Europeo	[1,5]

Tabla 19: Dimensión EURO (CAR FUEL EMISSION)

CAR			
COLUMNA	**TIPO**	**DESCRIPCIÓN**	**RANGO VALORES**
MANUFACTURER_MODEL_DESCRIPTION	VARCHAR	Descripción completa del coche, fabricante, modelo y descripción junta.	Peugeot-307-1.4 (75 bhp)
MANUFACTURER_MODEL	VARCHAR	Fabricante - Modelo	Peugeot-307
MANUFACTURER	VARCHAR	Fabricante	Peugeot

Tabla 20: Dimensión CAR (CAR FUEL EMISSION)

CAR_FUEL_EMISSIONS			
COLUMNA	**TIPO**	**DESCRIPCIÓN**	**RANGO VALORES**
year	NUMERIC	Año	[2000,2013]
manufacturer-model-description	VARCHAR	Fabricante – Modelo - Descripción	Peugeot-307
euro_standard	NUMERIC	Fabricante	3,4…
tax_band	CHAR	Tasas	A,D,L
transmission	VARCHAR	Transmisión	A5, M5…
transmission_type	VARCHAR	Tipo de transmisión	Manual,etc.
engine_capacity	VARCHAR	Capacidad del motor	<1000
fuel_type	VARCHAR	Tipo de combustible	Petrol
urban_metric	DECIMAL	Consumo urbano	11.7
extra_urban_metric	DECIMAL	Consumo extraurbano	4.9
combined_metric	DECIMAL	Consumo combinado	6.8
urban_imperial	DECIMAL	Consumo urbano	21.7
extra_urban_imperial	DECIMAL	Consumo extraurbano	24.9
combined_imperial	DECIMAL	Consumo combinado	26.8
noise_level	DECIMAL	Nivel de ruido	71.9
co2	NUMERIC	CO^2 emitido	247
co_emissions	NUMERIC	Emisiones de CO	270
manufacturer	VARCHAR	Fabricante	Ford
manufacturer-model	VARCHAR	Fabricante- Modelo	Ford-Fiesta
fuel_cost_6000_miles	NUMERIC	Coste del combustible a las 6000 millas	612

Tabla 21: Hecho Car Fuel Emissions (CAR FUEL EMISSION)

B.1.4 JERARQUÍAS

Las jerarquías necesarias para la ejecución del experimento con el conjunto Car Fuel Emission son las descritas en las tablas a continuación:

CAR
Jerarquía del coche, dónde se describe el fabricante, modelo y descripción.
hierarchyCAR: manufacturer.model, manufacturer, manufacturer.model.description, ROLLED_UP_LEVEL

Tabla 22: Jerarquía CAR (CAR FUEL EMISSION)

EURO
Jerarquía del estándar europeo.
hierarchyEURO: euro_standard, ROLLED_UP_LEVEL

Tabla 23: Jerarquía EURO (CAR FUEL EMISSION)

TRANS
Jerarquía de las transmisiones y tipos de transmisiones.
hierarchyTRANS: transmission, transmission_type, ROLLED_UP_LEVEL

Tabla 24: Jerarquía TRANS (CAR FUEL EMISSION)

ENGINE
Jerarquía del motor.
hierarchyENGINE: engine_capacity, ROLLED_UP_LEVEL

Tabla 25: Jerarquía ENGINE (CAR FUEL EMISSION)

FUEL
Jerarquía del tipo de combustible.
hierarchyFUEL: fuel_type, ROLLED_UP_LEVEL

Tabla 26: Jerarquía FUEL (CAR FUEL EMISSION)

PROMOTION
Jerarquía de los años de los coches.
hierarchyTIME: year

Tabla 27: Jerarquía PROMOTION (CAR FUEL EMISSION)

B.1.5 CUBOS DE DATOS

Los cubos de datos que se han generado a partir de las jerarquías definidas, es el producto de 4x2x3x2x2x1 cubos que hacen un total de 96, listados en la tabla 28.

Nº	CUBOS CAR FUEL EMISSIONS
1	year, manufacturer.model.description, engine_capacity, euro_standard, transmission, fuel_type
2	year, manufacturer.model.description, engine_capacity, euro_standard, transmission, ---
3	year, manufacturer.model.description, engine_capacity, euro_standard, transmission_type, fuel_type
4	year, manufacturer.model.description, engine_capacity, euro_standard, transmission_type, ---
5	year, manufacturer.model.description, engine_capacity, euro_standard, ---, fuel_type
6	year, manufacturer.model.description, engine_capacity, euro_standard, ---, ---
7	year, manufacturer.model.description, engine_capacity, ---, transmission, fuel_type
8	year, manufacturer.model.description, engine_capacity, ---, transmission, ---
9	year, manufacturer.model.description, engine_capacity, ---, transmission_type, fuel_type
10	year, manufacturer.model.description, engine_capacity, ---, transmission_type, ---
11	year, manufacturer.model.description, engine_capacity, ---, ---, fuel_type
12	year, manufacturer.model.description, engine_capacity, ---, ---, ---
13	year, manufacturer.model.description, ---, euro_standard, transmission, fuel_type

14	year, manufacturer.model.description, ---, euro_standard, transmission, ---
15	year, manufacturer.model.description, ---, euro_standard, transmission_type, fuel_type
16	year, manufacturer.model.description, ---, euro_standard, transmission_type, ---
17	year, manufacturer.model.description, ---, euro_standard, ---, fuel_type
18	year, manufacturer.model.description, ---, euro_standard, ---, ---
19	year, manufacturer.model.description, ---, ---, transmission, fuel_type
20	year, manufacturer.model.description, ---, ---, transmission, ---
21	year, manufacturer.model.description, ---, ---, transmission_type, fuel_type
22	year, manufacturer.model.description, ---, ---, transmission_type, ---
23	year, manufacturer.model.description, ---, ---, ---, fuel_type
24	year, manufacturer.model.description, ---, ---, ---, ---
25	year, manufacturer.model, engine_capacity, euro_standard, transmission, fuel_type
26	year, manufacturer.model, engine_capacity, euro_standard, transmission, ---
27	year, manufacturer.model, engine_capacity, euro_standard, transmission_type, fuel_type
28	year, manufacturer.model, engine_capacity, euro_standard, transmission_type, ---
29	year, manufacturer.model, engine_capacity, euro_standard, ---, fuel_type
30	year, manufacturer.model, engine_capacity, euro_standard, ---, ---
31	year, manufacturer.model, engine_capacity, ---, transmission, fuel_type
32	year, manufacturer.model, engine_capacity, ---, transmission, ---
33	year, manufacturer.model, engine_capacity, ---, transmission_type, fuel_type
34	year, manufacturer.model, engine_capacity, ---, transmission_type, ---
35	year, manufacturer.model, engine_capacity, ---, ---, fuel_type
36	year, manufacturer.model, engine_capacity, ---, ---, ---
37	year, manufacturer.model, ---, euro_standard, transmission, fuel_type
38	year, manufacturer.model, ---, euro_standard, transmission, ---
39	year, manufacturer.model, ---, euro_standard, transmission_type, fuel_type
40	year, manufacturer.model, ---, euro_standard, transmission_type, ---
41	year, manufacturer.model, ---, euro_standard, ---, fuel_type
42	year, manufacturer.model, ---, euro_standard, ---, ---
43	year, manufacturer.model, ---, ---, transmission, fuel_type
44	year, manufacturer.model, ---, ---, transmission, ---
45	year, manufacturer.model, ---, ---, transmission_type, fuel_type
46	year, manufacturer.model, ---, ---, transmission_type, ---
47	year, manufacturer.model, ---, ---, ---, fuel_type
48	year, manufacturer.model, ---, ---, ---, ---
49	year, manufacturer, engine_capacity, euro_standard, transmission, fuel_type
50	year, manufacturer, engine_capacity, euro_standard, transmission, ---
51	year, manufacturer, engine_capacity, euro_standard, transmission_type, fuel_type
52	year, manufacturer, engine_capacity, euro_standard, transmission_type, ---
53	year, manufacturer, engine_capacity, euro_standard, ---, fuel_type
54	year, manufacturer, engine_capacity, euro_standard, ---, ---
55	year, manufacturer, engine_capacity, ---, transmission, fuel_type
56	year, manufacturer, engine_capacity, ---, transmission, ---

57	year, manufacturer, engine_capacity, ---, transmission_type, fuel_type
58	year, manufacturer, engine_capacity, ---, transmission_type, ---
59	year, manufacturer, engine_capacity, ---, ---, fuel_type
60	year, manufacturer, engine_capacity, ---, ---, ---
61	year, manufacturer, ---, euro_standard, transmission, fuel_type
62	year, manufacturer, ---, euro_standard, transmission, ---
63	year, manufacturer, ---, euro_standard, transmission_type, fuel_type
64	year, manufacturer, ---, euro_standard, transmission_type, ---
65	year, manufacturer, ---, euro_standard, ---, fuel_type
66	year, manufacturer, ---, euro_standard, ---, ---
67	year, manufacturer, ---, ---, transmission, fuel_type
68	year, manufacturer, ---, ---, transmission, ---
69	year, manufacturer, ---, ---, transmission_type, fuel_type
70	year, manufacturer, ---, ---, transmission_type, ---
71	year, manufacturer, ---, ---, ---, fuel_type
72	year, manufacturer, ---, ---, ---, ---
73	year, ---, engine_capacity, euro_standard, transmission, fuel_type
74	year, ---, engine_capacity, euro_standard, transmission, ---
75	year, ---, engine_capacity, euro_standard, transmission_type, fuel_type
76	year, ---, engine_capacity, euro_standard, transmission_type, ---
77	year, ---, engine_capacity, euro_standard, ---, fuel_type
78	year, ---, engine_capacity, euro_standard, ---, ---
79	year, ---, engine_capacity, ---, transmission, fuel_type
80	year, ---, engine_capacity, ---, transmission, ---
81	year, ---, engine_capacity, ---, transmission_type, fuel_type
82	year, ---, engine_capacity, ---, transmission_type, ---
83	year, ---, engine_capacity, ---, ---, fuel_type
84	year, ---, engine_capacity, ---, ---, ---
85	year, ---, ---, euro_standard, transmission, fuel_type
86	year, ---, ---, euro_standard, transmission, ---
87	year, ---, ---, euro_standard, transmission_type, fuel_type
88	year, ---, ---, euro_standard, transmission_type, ---
89	year, ---, ---, euro_standard, ---, fuel_type
90	year, ---, ---, euro_standard, ---, ---
91	year, ---, ---, ---, transmission, fuel_type
92	year, ---, ---, ---, transmission, ---
93	year, ---, ---, ---, transmission_type, fuel_type
94	year, ---, ---, ---, transmission_type, ---
95	year, ---, ---, ---, ---, fuel_type
96	year, ---, ---, ---, ---, ---

Tabla 28: Cubos (CAR FUEL EMISSION)

B.2 CONJUNTO DE *DATOS AROMADB*

AROMADB es una base de datos (IBM, 2014) con información ficticia que la hemos obtenido de la página web de IBM. Representa las ventas realizadas en un grupo de supermercados, contando la cantidad de productos y dólares vendidos.

Como hecho principal del conjunto de datos AROMDB, se tiene SALES en el cual se pueden sacar resultados agregando su información para obtener la cantidad total de ventas y de dólares que se han vendido en todos los niveles.

Los datos ya tenían una estructura de datamart multidimensional, con las dimensiones PERIOD que responde a cuando se ha realizado una venta, la dimensión PROMO que describe cómo se ha vendido, la dimensión STORE que nos dice dónde se ha vendido el producto, y las dimensiones CLASS y PRODUCT que describen que se ha vendido.

Para la construcción, estudio y generación de la vista necesaria para el estudio, como la base de datos estaba en formato DB2 (es el tipo de base de datos de IBM), se ha tenido que montar una base de datos y a partir de ahí realizar consultas de bases de datos para obtener la vista deseada.

B.2.1 DISEÑO

La representación del diagrama multidimensional del conjunto AROMADB, que podemos ver en la ilustración 12 deja claramente que la estructura se compone del hecho SALES, las medidas QUANTITY y DOLLARS y sus 6 dimensiones PRODUCT, PERIOD, CLASS, MARKET, STORE y PROMOTION.

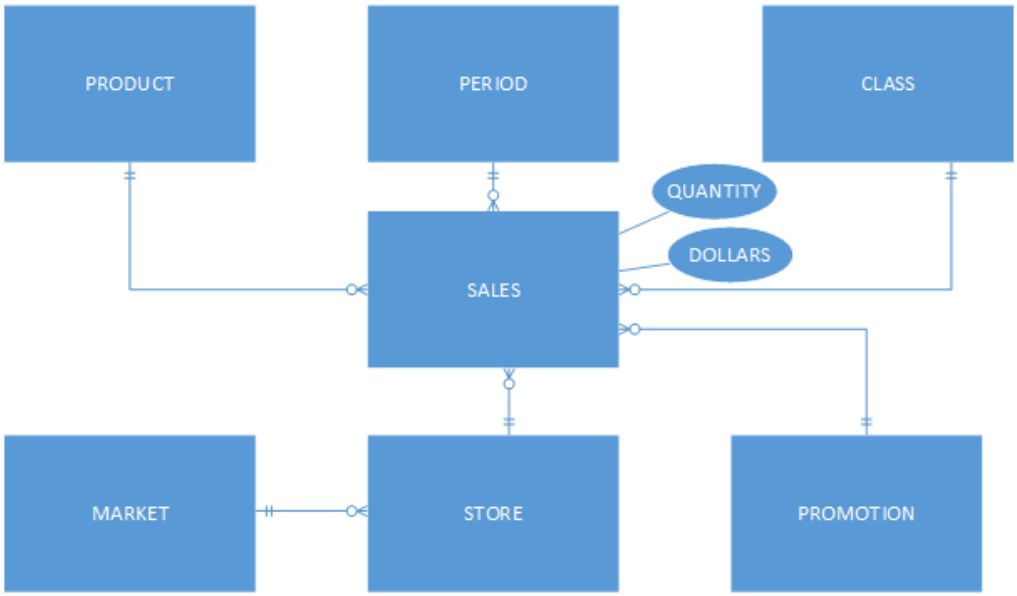

Ilustración 12: Diagrama multidimensional AROMADB

Ya entendido el diagrama multidimensional, pasamos a describir el mismo conjunto pero con un diagrama entidad relación en la ilustración 13.

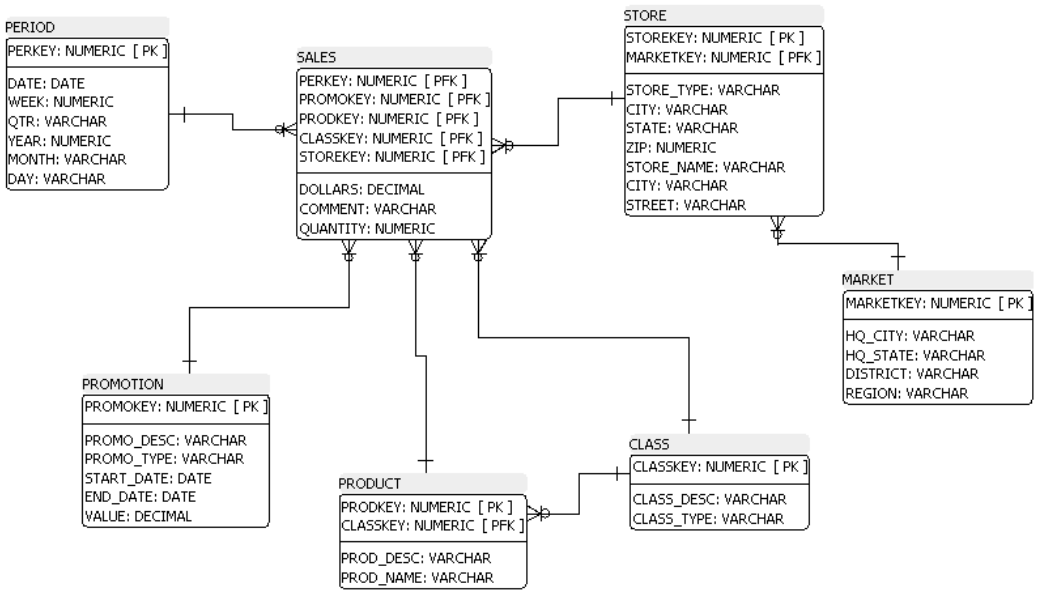

Ilustración 13: Diagrama Entidad Relación AROMADB

B.2.2 DETALLE

Para tener la disponibilidad de AROMADB se ha tenido que crear una vista global del hecho SALES, que se puede ver de forma detallada más abajo. En todo el conjunto de datos tenemos 11 tablas de las cuales nos quedamos con 7 y son las que están detalladas en la tabla 29.

TABLA	FILAS
AROMA.CLASS	9
AROMA.DEAL	9
AROMA.LINE_ITEMS	182
AROMA.MARKET	17
AROMA.ORDERS	27
AROMA.PERIOD	821
AROMA.PRODUCT	59
AROMA.PROMOTION	194
AROMA.SALES	69941
AROMA.STORE	18
AROMA.SUPPLIER	9

Tabla 29: Todas las tablas del conjunto AROMADB

PERIOD			
COLUMNA	**TIPO**	**DESCRIPCIÓN**	**RANGO VALORES**
Perkey	NUMERIC	Identificador único de un periodo.	1, 2, 3,4…
Date	DATE	Día, mes y año de un periodo.	YYYY-MM-DD. De 2004-01-01 a 2006-03-31
Day	VARCHAR	Día de la semana	FR, MO,SA, SU,TH ,TU,WE
Week	NUMERIC	Semana	[1,53]
Month	VARCHAR	Mes	JAN, FAB…
Qtr	VARCHAR	Cuatrimestre	Q1_04, Q1_05, Q1_06, Q2_04, Q2_05, Q3_04, Q3_05, Q4_04, Q4_05
Year	NUMERIC	Año	2004,2005,2006

Tabla 30: Dimensión PERIOD (AROMADB)

PRODUCT			
COLUMNA	**TIPO**	**DESCRIPCIÓN**	**RANGO VALORES**
Classkey	NUMERIC	Identificador de la clase.	1, 2, 3,…
Prodkey	NUMERIC	Identificador del producto.	1, 2, 3,…
Prod_Name	VARCHAR	Nombre asignado al producto.	Veracruzano,…
Pkg_Type	VARCHAR	Tipo de empaquetado.	Aroma designer box, Gift box, No pkg, One-pound bag, Original box, Qtr-pound bag…

Tabla 31: Dimensión PRODUCT (AROMADB)

CLASS			
COLUMNA	**TIPO**	**DESCRIPCIÓN**	**RANGO VALORES**
Classkey	NUMERIC	Identificador de la clase.	1, 2, 3…
Class_Type	VARCHAR	Tipo de la clase.	Bulk_beans, Bulk_spice, Bulk_tea, Clothing, Gifts, Hardware, Pkg_coffee, Pkg_spice, Pkg_tea
Class_Desc	VARCHAR	Descripción de la clase.	

Tabla 32: Dimensión CLASS (AROMADB)

STORE			
COLUMNA	**TIPO**	**DESCRIPCIÓN**	**RANGO VALORES**
Storekey	NUMERIC	Identificador de la tienda.	
Mktkey	NUMERIC	Identificador del mercado.	
Store_Type	VARCHAR	Tipo de tienda.	small, medium, large
Store_Name	VARCHAR	Nombre de la tienda.	
Street	VARCHAR	Calle.	
City	VARCHAR	Ciudad.	
State	VARCHAR	Estado.	
Zip	NUMERIC	Código postal.	

Tabla 33: Dimensión STORE (AROMADB)

MARKET			
COLUMNA	**TIPO**	**DESCRIPCIÓN**	**RANGO VALORES**
Mktkey	NUMERIC	Identificador del mercado.	
Hq_city	VARCHAR	Barrio de la ciudad.	Atlanta, Miami, New Orleands
Hq_state	VARCHAR	Estado de la zona de la ciudad.	GA, FL, LA, TX…
District	VARCHAR	Distrito.	Atlanta, Boston, Chicago , Los Angeles , Minneapolis, New Orleans, New York, San Francisco
Región	VARCHAR	Región.	Central , North South , West

Tabla 34: Dimensión MARKET (AROMADB)

PROMOTION			
COLUMNA	**TIPO**	**DESCRIPCIÓN**	**RANGO VALORES**
Promokey	NUMERIC	Identificador de la promoción.	1, 2,3…
Promo_type	NUMERIC	Tipo de la promoción.	1, 100,200, 300, 400, 900
Promo_Desc	VARCHAR	Descripción de la promoción.	No Promotion, Aroma catalog…
Value	DECIMAL	Valor aplicado al precio para la promoción.	0.00, 1.00, 2.00, 5.00
Start_Date	DATE	Fecha cuando empieza la promoción.	[2004-01-01,9999-01-01]
End_Date	DATE	Fecha cuando termina la promoción.	[2004-01-31,9999-01-01]

Tabla 35: Dimensión PROMOTION (AROMADB)

SALES		
COLUMNA	**TIPO**	**DESCRIPCIÓN**
Perkey	NUMERIC	Identificador del periodo.
Classkey	NUMERIC	Identificador de la clase.
Prodkey	NUMERIC	Identificador del producto.
Storekey	NUMERIC	Identificador de la tienda.
Promokey	NUMERIC	Identificador de la promoción.
Quantity	NUMERIC	Cantidad de producto.
Dollars		Coste total en Dólares
Comments	VARCHAR	Comentarios.

Tabla 36: Hecho SALES (AROMADB)

Una vez ya conocemos con detalle el conjunto AROMADB, lo que se ha hecho para una mejor legibilidad de los registros del hecho, ha sido crear una vista con toda la información que se necesita para realizar el estudio. La consulta que se ha ejecutado para crear dicha vista, es la siguiente sentencia SQL de la tabla 37:

SALES VIEW

```
select
        sales.prodkey as key_prod,
        product.prod_name as prod_name,
        product.pkg_type as pkg_type,
        sales.quantity  as quantity,
        sales.dollars as dollars,
        sales.perkey as key_period,
        PERIOD.date as date,
        PERIOD.day as day,
        PERIOD.week as week,
        PERIOD.month as month,
        PERIOD.qtr as qtr,
        PERIOD.year as year,
        sales.classkey as key_class,
        class.class_type as class_type,
        class.class_desc as class_desc,
        market.mktkey as key_mkt,
        market.hq_city as mkt_hq_city,
        market.hq_state as mkt_hq_state,
        market.district as mkt_district,
        market.region as mkt_region,
        sales.storekey  as key_store,
        store.store_type as store_type,
        store.store_name as store_name,
        store.street  as store_street,
        store.city as store_city,
        store.state as store_state,
        store.zip as store_zip,
        sales.promokey  as key_promo,
        promotion.promo_type  as promo_type,
        promotion.promo_desc as promo_desc,
        promotion.value as promo_value,
        promotion.start_date as promo_start_date,
        promotion.end_date as promo_end_date
from aroma.sales sales
left join aroma.period period on sales.perkey = period.perkey
left join aroma.product product on sales.prodkey = product.prodkey and sales.classkey =
product.classkey
left join aroma.class class on product.classkey = class.classkey
left join aroma.store store on sales.storekey = store.storekey
left join aroma.market market on store.mktkey = market.mktkey
left join aroma.promotion promotion on sales.promokey = promotion.promokey ;
```

Tabla 37: SQL Vista SALES (AROMADB)

En la tabla 38 se describen todas las columnas obtenidas a partir de la vista formada con la consulta de base de datos anterior.

VISTA SALES		
COLUMNA	**TIPO**	**DESCRIPCIÓN**
KEY_PROD	NUMERIC	Id del producto.
PROD_NAME	VARCHAR	Nombre del producto.
PROD_PKG_TYPE	VARCHAR	Tipo de paquete del producto.
SAL_QUANTITY	NUMERIC	Cantidad de producto vendido.
SAL_DOLLARS	NUMERIC	Cantidad total de dólares de la venta producida.
KEY_PERIOD:	NUMERIC	Id del periodo en el que se ha realizado la venta.
SAL_DATE	DATE	Fecha de venta del producto.
SAL_DAY	VARCHAR	Día de la semana que se ha vendido el producto.
SAL_WEEK	NUMERIC	Semana del año en la que se ha vendido el producto.
SAL_MONTH	VARCHAR	Mes en el que se ha vendido el producto.
SAL_QTR	VARCHAR	Cuatrimestre en el que se ha vendido el producto.
SAL_YEAR	NUMERIC	Año en el que se ha vendido el producto.
KEY_CLASS	NUMERIC	Clase del producto vendido.
CLASS_TYPE	VARCHAR	Tipo de clase del producto vendido.
CLASS_DESC	VARCHAR	Descripción de la clase del producto vendido.
KEY_MKT	NUMERIC	Id del Market dónde se ha vendido el producto.
MKT_HQ_CITY	VARCHAR	Barrio de la ciudad
MKT_HQ_STATE	VARCHAR	Estado en el que se ha vendido el producto.
MKT_DISTRICT	VARCHAR	Distrito en el que se ha vendido.
MKT_REGION	VARCHAR	Región en el que se ha vendido.
KEY_STORE	NUMERIC	Store en el que se ha vendido.
STORE_TYPE	VARCHAR	Tipo de Store en el que se ha vendido.
STORE_NAME	VARCHAR	Nombre del Store dónde se ha vendido.
STORE_STREET	VARCHAR	Calle del Store dónde se ha vendido.
STORE_CITY	VARCHAR	Ciudad del Store dónde se ha vendido.
STORE_STATE	VARCHAR	Estado del Store dónde se ha vendido.
STORE_ZIP	NUMERIC	Código postal del Store dónde se ha vendido.
KEY_PROMO	NUMERIC	Id de la promoción que se ha aplicado a la venta.
PROMO_TYPE	NUMERIC	Tipo de promoción aplicada a la venta.
PROMO_DESC	VARCHAR	Descripción de la promoción aplicada a la venta.
PROMO_VALUE	DECIMAL	Tanto por cien de la promoción aplicada a la venta.
PROMO_START_DATE	DATE	Fecha de comienzo de la promoción.
PROMO_END_DATE	DATE	Fecha de finalización de la promoción.

Tabla 38: Detalle vista SALES (AROMADB)

B.2.3 JERARQUÍAS

Las jerarquías necesarias para los experimentos que componen el conjunto AROMADB se definen en las siguientes tablas.

PROMOTION
Jerarquía de las promociones de ventas. Se compone del identificador de la promoción y el tipo de promoción.
hierarchyPROMO : KEY_PROMO, PROMO_TYPE, ROLLED_UP_LEVEL

Tabla 39: Jerarquía PROMOTION (AROMADB)

CLASS
La jerarquía CLASS, se compone del identificador y del tipo. Como tiene relación entre ellos 1 a 1, al final nos quedamos sólo con el identificador de la clase.
hierarchyCLASS : KEY_CLASS, ROLLED_UP_LEVEL

Tabla 40: Jerarquía CLASS (AROMADB)

PRODUCT
PRODUCT se compone del identificador del product. Por temas de rendimiento y tiempo de ejecución, se ha descargado para el estudio.
hierarchyPRODUCT : KEY_PROD, ROLLED_UP_LEVEL

Tabla 41: Jerarquía PRODUCT (AROMADB)

PERIOD
PERIOD, por ser la jerarquía que contiene el campo por el que vamos a particional el conjunto para obtener el train y test, se ha dejado sólo el campo que vamos a utilizar para hacerlo, el año de la venta.
hierarchyPERIOD : SAL_YEAR

Tabla 42: Jerarquía PERIOD (AROMADB)

STORE
Se compone del identificador de la tienda y del mercado con toda la información asociada.
hierarchySTORE: KEY_STORE, KEY_MKT, MKT_HQ_CITY, MKT_HQ_STATE, MKT_DISTRICT, MKT_REGION, ROLLED_UP_LEVEL

Tabla 43: Jerarquía STORE (AROMADB)

B.2.4 CUBOS DE DATOS

Se han generado 3x2x1x7 cubos de datos que hacen un total de 42 para el conjunto de datos AROMADB que se listan en la tabla 44.

Nº	CUBOS AROMADB
1	KEY_PROMO, KEY_STORE, KEY_CLASS, SAL_YEAR
2	KEY_PROMO, KEY_STORE, ---, SAL_YEAR
3	KEY_PROMO, KEY_MKT, KEY_CLASS, SAL_YEAR
4	KEY_PROMO, KEY_MKT, ---, SAL_YEAR
5	KEY_PROMO, MKT_HQ_CITY, KEY_CLASS, SAL_YEAR
6	KEY_PROMO, MKT_HQ_CITY, ---, SAL_YEAR
7	KEY_PROMO, MKT_HQ_STATE, KEY_CLASS, SAL_YEAR
8	KEY_PROMO, MKT_HQ_STATE, ---, SAL_YEAR
9	KEY_PROMO, MKT_DISTRICT, KEY_CLASS, SAL_YEAR
10	KEY_PROMO, MKT_DISTRICT, ---, SAL_YEAR
11	KEY_PROMO, MKT_REGION, KEY_CLASS, SAL_YEAR
12	KEY_PROMO, MKT_REGION, ---, SAL_YEAR
13	KEY_PROMO, ---, KEY_CLASS, SAL_YEAR
14	KEY_PROMO, ---, ---, SAL_YEAR
15	PROMO_TYPE, KEY_STORE, KEY_CLASS, SAL_YEAR
16	PROMO_TYPE, KEY_STORE, ---, SAL_YEAR
17	PROMO_TYPE, KEY_MKT, KEY_CLASS, SAL_YEAR
18	PROMO_TYPE, KEY_MKT, ---, SAL_YEAR
19	PROMO_TYPE, MKT_HQ_CITY, KEY_CLASS, SAL_YEAR
20	PROMO_TYPE, MKT_HQ_CITY, ---, SAL_YEAR
21	PROMO_TYPE, MKT_HQ_STATE, KEY_CLASS, SAL_YEAR
22	PROMO_TYPE, MKT_HQ_STATE, ---, SAL_YEAR
23	PROMO_TYPE, MKT_DISTRICT, KEY_CLASS, SAL_YEAR
24	PROMO_TYPE, MKT_DISTRICT, ---, SAL_YEAR
25	PROMO_TYPE, MKT_REGION, KEY_CLASS, SAL_YEAR
26	PROMO_TYPE, MKT_REGION, ---, SAL_YEAR
27	PROMO_TYPE, ---, KEY_CLASS, SAL_YEAR
28	PROMO_TYPE, ---, ---, SAL_YEAR
29	---, KEY_STORE, KEY_CLASS, SAL_YEAR
30	---, KEY_STORE, ---, SAL_YEAR
31	---, KEY_MKT, KEY_CLASS, SAL_YEAR
32	---, KEY_MKT, ---, SAL_YEAR
33	---, MKT_HQ_CITY, KEY_CLASS, SAL_YEAR
34	---, MKT_HQ_CITY, ---, SAL_YEAR
35	---, MKT_HQ_STATE, KEY_CLASS, SAL_YEAR
36	---, MKT_HQ_STATE, ---, SAL_YEAR
37	---, MKT_DISTRICT, KEY_CLASS, SAL_YEAR
38	---, MKT_DISTRICT, ---, SAL_YEAR
39	---, MKT_REGION, KEY_CLASS, SAL_YEAR
40	---, MKT_REGION, ---, SAL_YEAR
41	---, ---, KEY_CLASS, SAL_YEAR
42	---, ---, ---, SAL_YEAR

Tabla 44: Cubos (AROMADB)

B.3 CONJUNTO DE DATOS *GEN VARATIONS*

El conjunto de datos **Gen Variations** (Pastor, y otros, 2012) y (Celma & Martin, 2011) es un conjunto de datos proporcionado por el centro de investigación en métodos de producción de software (PROS). Se compone por un hecho principal "GENO VARIATIONS" que contiene la información de las variaciones genéticas. Respecto a las dimensiones, el conjunto ha sido transformado a un esquema multidimensional, con las siguientes dimensiones. La dimensión DATE responde a cuándo ha sucedido la variación. Cómo ha sucedido la variación la describe la dimensión SPEC. La dimensión GENOTYPE responde a dónde ha sucedido la variación. Qué ha variado se describe en la dimensión PHENOTYPE y dónde la dimensión DBANK.

B.3.1 DISEÑO

En la ilustración 14 se representa en un diagrama multidimensional el conjunto de datos Gen Variations.

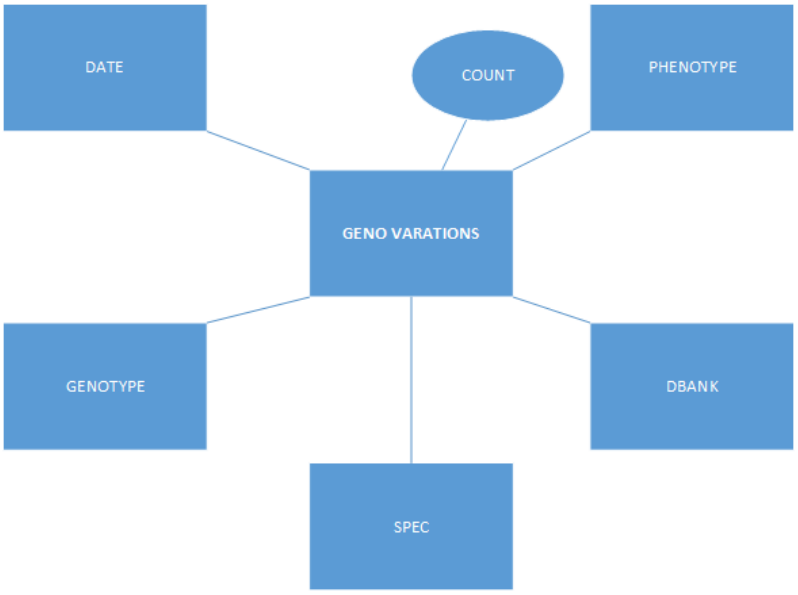

Ilustración 14: Diagrama multidimensional GENO-VARIATIONS

Para entender mejor la estructura multidimensional del conjunto de datos y qué información aporta este, se describe detalladamente con todos sus atributos en la ilustración 15.

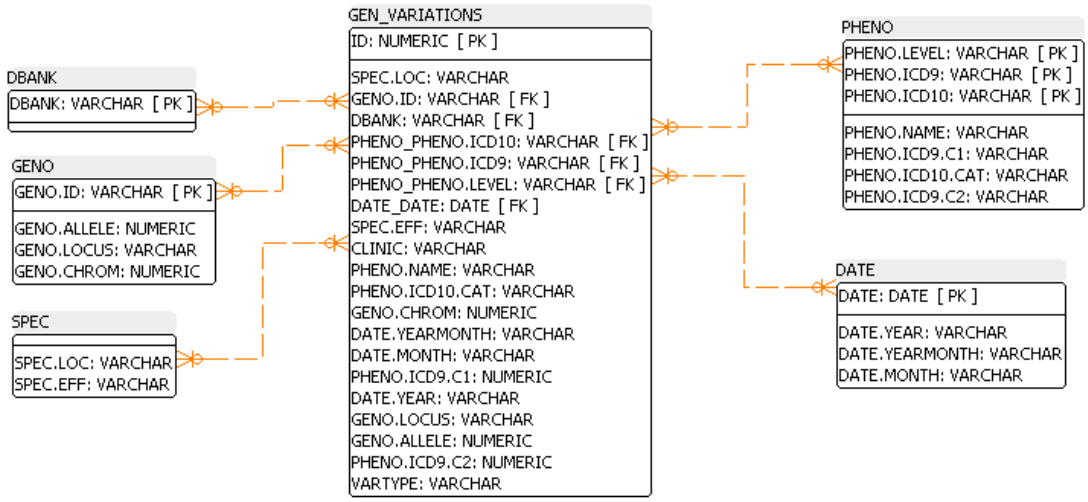

Ilustración 15: Diagrama entidad relación GEN-VARIATIONS

B.3.2 DETALLE

Se describe cada uno de los atributos del conjunto Gen Variations con tablas estructuradas por dimensiones y terminando con el hecho.

SPEC			
COLUMNA	**TIPO**	**DESCRIPCIÓN**	**RANGO VALORES**
SPEC.Loc	VARCHAR	Localización de la especialización.	G
SPEC.Eff	VARCHAR	Efecto de la especialización.	M,U

Tabla 45: Dimensión SPEC GEN-VARIATIONS

DBANK			
COLUMNA	**TIPO**	**DESCRIPCIÓN**	**RANGO VALORES**
DBANK	VARCHAR	Data Bank	

Tabla 46: Dimensión DBANK GEN-VARIATIONS

PHENO			
COLUMNA	**TIPO**	**DESCRIPCIÓN**	**RANGO VALORES**
PHENO.Level	VARCHAR	Nivel de fenotipo	D,S,NULL
PHENO.Name	VARCHAR	Es cualquier característica o rasgo observable de un organismo, como su morfología, desarrollo, propiedades bioquímicas, fisiología y comportamiento	Osteogenesis imperfecta II,…
PHENO.ICD9	VARCHAR	Conjunto codificaciones de diagnósticos.	[179,…,USH], NULL
PHENO.ICD9.C1	VARCHAR	Subconjunto de codificaciones ICD9 de diagnósticos.	[1,…,U],NULL
PHENO.ICD9.C2	VARCHAR	Subconjunto de codificaciones ICD9 de diagnósticos.	[1,…,8,NULL
PHENO.ICD10	VARCHAR	Conjunto codificaciones de diagnósticos.	C22.9,…,UNK, NULL
PHENO.ICD10.Cat	VARCHAR	Subconjunto de codificaciones ICD10 de diagnósticos	C,D,…,UNK

Tabla 47: Dimensión PHENO GEN-VARIATIONS

DATE			
COLUMNA	**TIPO**	**DESCRIPCIÓN**	**RANGO VALORES**
DATE	DATE	Fecha	[1970,2012] , NULL
DATE.Month	NUMERIC	Mes	[1,12], NULL
DATE.Year	NUMERIC	Año	[1970,2012], NULL
DATE.YearMonth	VARCHAR	AñoMes	[1970-1,2012-3], NULL

Tabla 48: Dimensión DATE GEN-VARIATIONS

GENO			
COLUMNA	**TIPO**	**DESCRIPCIÓN**	**RANGO VALORES**
GENO.ID	NUMERIC	Identificador del gen	BRCA1, BRCA2,…,USH2A,NULL
GENO.Chrom	NUMERIC	Cromosoma	[1,17] , NULL
GENO.Locus	VARCHAR	Locus	10q21.1,…,9q32,NULL
GENO.Allele	NUMERIC	Alelo	947,…,961,NULL

Tabla 49: Dimensión GENO GEN-VARIATIONS

GEN-VARIATIONS		
COLUMNA	**TIPO**	**DESCRIPCIÓN**
ID	NUMERIC	Identificación de variación.
SPEC.Loc	CHAR	Localización de la especialización.
SPEC.Eff	CHAR	Efecto de la especialización
CLINIC	VARCHAR	Información clínica importante.
DBANK	VARCHAR	Banco de datos
PHENO.Level	CHAR	Nivel de fenotipo
PHENO.Name	VARCHAR	Fenotipo
PHENO.ICD9	VARCHAR	ICD-9
PHENO.ICD9.C1	VARCHAR	ICD9-PrimCat
PHENO.ICD9.C2	VARCHAR	ICD9-SecCat
PHENO.ICD10	VARCHAR	ICD-10
PHENO.ICD10.Cat	VARCHAR	ICD-10-Cat
DATE	DATE	Fecha
DATE.Month	NUMERIC	Mes
DATE.Year	NUMERIC	Año
DATE.YearMonth	VARCHAR	AñoMes
GENO.ID	VARCHAR	Identificador del gen
GENO.Chrom	NUMERIC	Cromosoma
GENO.Locus	VARCHAR	Locus
GENO.Allele	NUMERIC	Allele
VARTYPE	VARCHAR	Tipo de variación.

Tabla 50: Hecho GEN-VARIATIONS

B.3.3 JERARQUÍAS

Para los experimentos realizados, es necesario tener unas jerarquías definidas para la generación de los cubos, las jerarquías asociadas al hecho Gen Variations se detallan a continuación.

SPEC
Especialización
hierarchySPEC: SPEC.Eff, ROLLED_UP_LEVEL

Tabla 51: Jerarquía SPEC (GEN VARIATIONS)

DBANK
Data Bank
hierarchyDBANK: DBANK, ROLLED_UP_LEVEL

Tabla 52: Jerarquía DBANK (GEN VARIATIONS)

PHENOTYPE
Fenotípo
hierarchyPHENO: PHENO.Name, PHENO.ICD10, PHENO.ICD10.Cat, ROLLED_UP_LEVEL

Tabla 53: Jerarquía PHENOTYPE (GEN VARIATIONS)

GENOTYPE
Genotípo
hierarchyGENO: GENO.ID, GENO.Chrom, ROLLED_UP_LEVEL

Tabla 54: Jerarquía GENOTYPE (GEN VARIATIONS)

DATE
Fecha
HierarchyDATE: DATE, DATE.Year, ROLLED_UP_LEVEL

Tabla 55: Jerarquía DATE (GEN VARIATIONS)

B.3.4 CUBOS DE DATOS

Ya conocidas las jerarquías del conjunto definidas para el experimento, el siguiente listado de la tabla 56 especifica todos los cubos de datos que se obtienen con la ejecución del conjunto de datos Geno Variations. En concreto, 2x2x4x3x1 cubos que generan un total de 48 cubos.

Nº	CUBOS GENO VARIATIONS
1	SPEC.Eff, DBANK, PHENO.Name, DATE.Year, GENO.ID
2	SPEC.Eff, DBANK, PHENO.Name, DATE.Year, GENO.Chrom
3	SPEC.Eff, DBANK, PHENO.Name, DATE.Year, ---
4	SPEC.Eff, DBANK, PHENO.ICD10, DATE.Year, GENO.ID
5	SPEC.Eff, DBANK, PHENO.ICD10, DATE.Year, GENO.Chrom
6	SPEC.Eff, DBANK, PHENO.ICD10, DATE.Year, ---
7	SPEC.Eff, DBANK, PHENO.ICD10.Cat, DATE.Year, GENO.ID
8	SPEC.Eff, DBANK, PHENO.ICD10.Cat, DATE.Year, GENO.Chrom
9	SPEC.Eff, DBANK, PHENO.ICD10.Cat, DATE.Year, ---
10	SPEC.Eff, DBANK, ---, DATE.Year, GENO.ID
11	SPEC.Eff, DBANK, ---, DATE.Year, GENO.Chrom
12	SPEC.Eff, DBANK, ---, DATE.Year, ---
13	SPEC.Eff, ---, PHENO.Name, DATE.Year, GENO.ID
14	SPEC.Eff, ---, PHENO.Name, DATE.Year, GENO.Chrom
15	SPEC.Eff, ---, PHENO.Name, DATE.Year, ---
16	SPEC.Eff, ---, PHENO.ICD10, DATE.Year, GENO.ID
17	SPEC.Eff, ---, PHENO.ICD10, DATE.Year, GENO.Chrom

18	SPEC.Eff, ---, PHENO.ICD10, DATE.Year, ---
19	SPEC.Eff, ---, PHENO.ICD10.Cat, DATE.Year, GENO.ID
20	SPEC.Eff, ---, PHENO.ICD10.Cat, DATE.Year, GENO.Chrom
21	SPEC.Eff, ---, PHENO.ICD10.Cat, DATE.Year, ---
22	SPEC.Eff, ---, ---, DATE.Year, GENO.ID
23	SPEC.Eff, ---, ---, DATE.Year, GENO.Chrom
24	SPEC.Eff, ---, ---, DATE.Year, ---
25	---, DBANK, PHENO.Name, DATE.Year, GENO.ID
26	---, DBANK, PHENO.Name, DATE.Year, GENO.Chrom
27	---, DBANK, PHENO.Name, DATE.Year, ---
28	---, DBANK, PHENO.ICD10, DATE.Year, GENO.ID
29	---, DBANK, PHENO.ICD10, DATE.Year, GENO.Chrom
30	---, DBANK, PHENO.ICD10, DATE.Year, ---
31	---, DBANK, PHENO.ICD10.Cat, DATE.Year, GENO.ID
32	---, DBANK, PHENO.ICD10.Cat, DATE.Year, GENO.Chrom
33	---, DBANK, PHENO.ICD10.Cat, DATE.Year, ---
34	---, DBANK, ---, DATE.Year, GENO.ID
35	---, DBANK, ---, DATE.Year, GENO.Chrom
36	---, DBANK, ---, DATE.Year, ---
37	---, ---, PHENO.Name, DATE.Year, GENO.ID
38	---, ---, PHENO.Name, DATE.Year, GENO.Chrom
39	---, ---, PHENO.Name, DATE.Year, ---
40	---, ---, PHENO.ICD10, DATE.Year, GENO.ID
41	---, ---, PHENO.ICD10, DATE.Year, GENO.Chrom
42	---, ---, PHENO.ICD10, DATE.Year, ---
43	---, ---, PHENO.ICD10.Cat, DATE.Year, GENO.ID
44	---, ---, PHENO.ICD10.Cat, DATE.Year, GENO.Chrom
45	---, ---, PHENO.ICD10.Cat, DATE.Year, ---
46	---, ---, ---, DATE.Year, GENO.ID
47	---, ---, ---, DATE.Year, GENO.Chrom
48	---, ---, ---, DATE.Year, ---

Tabla 56: Cubos GEN-VARIATIONS

B.4 ENLACES WEB DE CONJUNTOS DE DATOS

En la siguiente tabla se listan todos los sitios web visitados donde se han buscado conjuntos de datos que pudieran ser válidos para los experimentos, incluyendo los que se han utilizado en los experimentos.

ENLACE	DESCRIPCIÓN
http://en.wikipedia.org/wiki/Open_data	Enlaces que pueden ser interesantes sobre directorios y páginas web interesantes sobre open data.
http://gcmd.gsfc.nasa.gov/KeywordSearch/Metadata.do?Portal=GCMD&KeywordPath=&EntryId=CDIAC_FACE_FACTS2_WISC&MetadataView=Data&MetadataType=0&lbnode=mdlb6	Directorio de informaciones open data enfocada al cambio global en la tierra. Este conjunto en concreto, hacer referencia a un experimento que se hace con el CO_2. Conjunto de datos con posible dimensión de tiempo entre otras.
http://datamarket.com/data/set/148w/inflation-consumer-prices-annual#!display=line	Base de datos en la cual se puede hacer las sql online además de descargar la información deseada. Representa la inflación población por países de todo el mundo. Dónde se puede realizar comparaciones, ver gráficos, entre más cosas.
https://explore.data.gov/profile/Data-gov-Program-Management-Office/e8ug-wzay	Directorio de bases de datos públicas de USA
https://explore.data.gov/Other/2013-Federal-Register-in-XML/mpm3-k649	Registro federal USA público con 80 MB de base de datos en Excel.
http://quickfacts.census.gov/qfd/download_data.html	Contiene información de la población de los estados de USA, como porcentajes de personas menores de 5 años, personas hispánicas, etc. La información está en varios ficheros txt.
http://www.cms.gov/Research-Statistics-Data-and-Systems/CMS-Information-Technology/DataAdmin/Downloads/DimensionalDataDesign.pdf	Documentación sobre desarrollo de datamarts.
http://freedatasheets.com/	Directorio de conjuntos de datos de elementos de conductores eléctricos
http://www.boconline.co.uk/en/sheq/safety-data-sheets/index.html	Algunos conjuntos de la industria del gas UK.
http://www.phac-aspc.gc.ca/lab-bio/res/psds-ftss/index-eng.php	Documentos sobre seguridad patógena y sus riesgos, no he encontrado como descargar conjuntos de esta web.

http://www.dshs.state.tx.us/chs/datalist.shtm	Departamento de salud de Texas, es necesaria solicitud de licencia para obtener el acceso a sus conjuntos abiertos al público.
http://www.dse.vic.gov.au/property-titles-and-maps/spatial/spatial-events-and-user-groups/local-government-vicmap-helpdesk/library/spatial-datamart	Conjuntos de datos que representan el efecto ambiental en diferentes aspectos.
http://www.dpi.vic.gov.au/earth-resources/investment-and-trade/licences-permits/licences-permits	Conjuntos con datos geotérmicos y petróleo…
http://www.doc.ic.ac.uk/~pjm/databases/datasets.html	En esta web, hay diferentes enlaces a distintos conjuntos, unos en txt y otros en xml. Bases de datos de ejemplos con diferentes enfoques de información. Ningún hecho interesante.
http://pgfoundry.org/frs/?group_id=1000150.	Bases de datos con información sobre ciudades del mundo.
http://www.dbis.informatik.uni-goettingen.de/Mondial/#SQL	Mondial DB, tablas con poca cantidad de datos. No existente hechos ni fechas que indiquen cuando ha pasado algo.
http://www.icsm.gov.au/publications/index.html#cadastral_databases	Bases de datos con datos catastrales en AU.
http://datawarehouse.hrsa.gov/HPSADownload.aspx	Base de datos bastante completa del sector sanitario de EEUU. Tiene varias dimensiones y además un hecho. Este hecho no contiene ni estados ni fechas, no se puede detectar ningún movimiento, con lo cual no es válido para el estudio ya que no contiene un histórico.
http://databaseanswers.org/data_models/index.htm	Muchos diseños de modelos de base de datos que podrían ser útiles a la hora de montarse una base de datos para lo que sea.
http://toxbank.net/about/datawarehouse	Grupo de investigación de bioinformática.
http://data.worldbank.org/data-catalog/world-development-indicators	Indicadores del desarrollo mundial. Es una base de datos de indicadores, no hay ni dimensiones ni hechos. Por tanto, tampoco existe ningún histórico.
https://data.cityofnewyork.us/Business/NYC-Jobs/kpav-sd4t	Base de datos de ofertas de trabajo de NYC pero no tiene prácticamente ninguna dimensión ni ningún cambio de estado de la oferta de trabajo, con lo cual no se visualiza ningún histórico.
http://www.ine.es/ss/Satellite?L=es_ES&c=Page&cid=1259942408928&p=125994240	Bases de datos públicas del instituto nacional de estadística. Por ejemplo,

8928&pagename=ProductosYServicios%2FPYSLayout	estadísticas sobre la renta Española.
http://digital.csic.es/bitstream/10261/8740/1/DE%20TESIS%20DOCTORALES.pdf	Documento de cómo buscar recursos relacionados con tesinas.
http://dissexpress.umi.com/dxweb#results	Directorio de tesinas con sus respectivos recursos ordenadas por distintos filtros.
http://apps.ams.usda.gov/USDAMIB/Main/TableDescriptions.aspx	Base de datos recursos agrarios, cómo por ejemplo, de leche producida.
http://www.slavevoyages.org/tast/database/download.faces	Base de datos de viajes trans-atlánticos. Poca información de la que se pueda usar. Se utiliza para realizar comparaciones con distintos años.
http://blog.lib.uiowa.edu/transitions/?cat=13	Directorio de recursos open data de la University of Lowa
http://atlantides.org/downloads/pleiades/dumps/	Base de datos con dumps diarios de información geográfica de la comunidad sobre lugares en todo el mundo. No existen hechos, sólo información almacenada.
http://www.scidb.org/about/history.php	Base de datos del gran telescopio de rastreos sinópticos. Es necesario registro en web/foro para la descarga de esta.
https://github.com/jaberg/skdata/wiki/Data-Set-Modules	Varios modelos de open source para predicción de datos que pueden ser útiles como por ejemplo, para la herramienta kaggle.
https://opendata.socrata.com/	Directorio de bases de datos open source, Socrata es una empresa de software en la nube con sede en Seattle, se centró exclusivamente en la democratización del acceso a los datos del gobierno de EEUU.
https://data.edmonton.ca/	Directorio open source de la ciudad de Edmonton EE.UU.
http://linkeddata.org/data-sets	Página web con conjuntos que usan para conectar información que no se ha relacionado nunca con otra.
http://linkedopencommerce.com/#__h2	Datasets open source para el ecommerce.
http://www.product-open-data.com/	Base de datos open source, con la representación de líneas de productos como por ejemplo de bebidas de la empresa Coca-Cola, con toda su información. Interesante por la cantidad de datos que tiene pero no contiene ningún hecho, sólo información.
https://ir.radford.edu/common_data_set/	El objetivo combinado de esta colaboración es el de mejorar la calidad y

	la exactitud de la información proporcionada a todos los involucrados en la transición de un estudiante en la educación superior, así como para reducir la carga de información sobre los proveedores de datos.
http://productdb.org/	ProductDB pretende ser fuente más completa y abierta del mundo de datos del producto.
http://apps.fs.fed.us/fiadb-downloads/datamart.html	Forest Inventory and Analysis National Program, ficheros Access 2010 preparados para todo (limpios, tablas predefinidas y preparadas para importar)
http://www.epa.gov/ttn/airs/aqsdatamart/basic_info.htm	Sistema de calidad de aire. Se dispone de una base de datos en la cual se registra toda la información que se genera del sistema.
https://data.colorado.gov/	Diferentes conjuntos con información open source del colorado.
https://github.com/jaberg/skdata/wiki/Data-Set-Modules	Scripts escritos en Phyton, los cuales realizar un tratamiento y descarga de diferentes conjuntos.
http://www.nrel.gov/	Future Automotive Systems Technology Simulador. FASTSim evalúa el impacto de las mejoras tecnológicas en la eficiencia de los vehículos, el rendimiento, coste y duración de la batería de los vehículos convencionales y avanzadas.
http://proyectocolibri.es/apps/	De los pocos proyectos Open Source en España. Proyecto Colibri dónde tienen unos conjuntos con información del registro diario del congreso.
http://vizual-statistix.tumblr.com/	Diferentes conjuntos de todos los ámbitos.
http://fontanon.github.com/us_bank_failures/	Un conjunto de transformaciones de datos y visualizaciones sobre la base de la Federal Deposit Insurance Corporation (FDIC) lista del banco de Fallas.
http://datahub.io/	Directorio de bases de datos open source.
http://www.statsci.org/	Diferentes conjuntos del ámbito de investigaciones. Algunos de los enlaces ya no funcionan.
http://apps.ecmwf.int/datasets/	Diferentes conjuntos del ámbito de climatología en tiempo real.
http://www.stccmop.org/datamart/cruises/sample_inventory	Registro de muestras de agua de diferentes estaciones meteorológicas.
https://www.lib.ncsu.edu/data/socialscie	Conjuntos de datos de Ciencias Sociales y

nceandhumsets.html	Humanidades.
http://lib.stat.cmu.edu/	Sistema de distribución de software y bases de datos estadísticas.
http://opendata.stackexchange.com/	Foro donde gente comparte orígenes de datos que pueden ser útiles.

Tabla 57: Enlaces web de conjuntos de datos

C SCRIPT MDHM

MDHM es un script en R de jerarquías de modelos multidimensionales. Este lee datamarts preparados para la ejecución y luego organiza los atributos en forma de jerarquías en un conjunto de dimensiones para crear todos los cubos posibles a partir de la combinación de las jerarquías obtenidas del esquema. En concreto:

- Se entrena un modelo para el nivel más bajo (*lowest level*) para todas las dimensiones, donde se obtiene el máximo detalle y por tanto se tiene mucha más precisión. A esto le llamamos ll-model (*lowest level model*).
- Se entrena un modelo para cada cubo con diferentes técnicas.
- Se aplica el *ll-model* para cada cubo agregando esta predicción y comparando los resultados con el modelo entrenado y aplicado al cubo.
- Por ahora, sólo soporta jerarquías lineales.

Para realizar los experimentos sobre los diferentes conjuntos de datos utilizados, se han utilizado los scripts generados MDHM_funcions.r (detalle en apéndice C.1) y MDHMv2.5.r (detalle en apéndice C.2). Están estructurados de forma que en el MDHM_functions es un script auxiliar el cual se compone de distintas funciones que se utilizan en el script principal MDHMv2.5.r.

Para introducir las funciones de los scripts, a continuación se describen las cabeceras:

- **LoadDataset** <- function (DATAMART, ROLLED_UP_LEVEL, INDICATOR)
- **CreateDatasetWithAllZeros** <- function (all_data, hierarchies, ROLLED_UP_LEVEL)
- **TST_Learning** <- function (TECHNIQUE, my_aggdata, train_my_aggdata, aggregation_names, NumHiers, my_model, CHECK_LEVELS_DURING_TEST)
- **TRA_Learning** <- function (TECHNIQUE, my_aggdata, aggregation_names, l)
- **winlosedraw** <- function(x, y)
- **reldif** <- function(x,y)
- **CreateAggrDataJoin** <- function()
- **no.dimnames** <- function(a)
- **MDHM** <- function (DATAMART, INDICATOR, TECHNIQUE, verbose=TRUE)

Para comprender mejor los scripts se pasa a describir detalladamente cada una de las funciones que contienen los siguientes apéndices (C1 y C2).

C.1 FUNCIONES MDHM

Se detallan a continuación las funciones del Script MDHM que están implementadas en un script R auxiliar llamado MDHM_functions.r

LoadDataset <- function (DATAMART, ROLLED_UP_LEVEL, INDICATOR)

Realiza la carga en memoria de los conjuntos de datos pasándole como argumento qué conjunto se quiere cargar, el contenido del atributo ROLLED_UP_LEVEL que indica el tope de la jerarquía (en el caso de este trabajo "- -") y el atributo a agregar. Es posible realizar la carga del conjunto AROMADB con sus indicadores QUANTITY y DOLLARS, GENOMICS con una columna añadida que se le asigna un conteo, TOY con INDICATOR y CARS con CO^2.

CreateDatasetWithAllZeros <- function (all_data, hierarchies, ROLLED_UP_LEVEL)

La función de rellenar con ceros, lo que hace es detectar en todo el conjunto cargado en memoria, qué atributos contienen valores no numéricos o nulos y rellenarlos con un 0 para que las agregaciones y operaciones que se realizaran posteriormente no descarten dichos registros por no tener valor.

TST_Learning <- function (TECHNIQUE, my_aggdata, train_my_aggdata, aggregation_names, NumHiers, my_model, CHECK_LEVELS_DURING_TEST)

Realiza la aplicación de la técnica pasada como argumento al conjunto generado como test para obtener predicciones. Las técnicas que se pueden ejecutar son KNN, MEAN, LRW, SVM y M5P.

TRA_Learning <- function (TECHNIQUE, my_aggdata, aggregation_names, l)

Para el conjunto de datos de entrenamiento se aplica el aprendizaje de la técnica escogida.

winlosedraw <- function(x, y)

Calcula de dos valores según la precisión establecida en la función (0.0000001) si son iguales, si una es mayor o si otra es menor. Realiza la resta entre x e y, si el resultado es mayor que 0.0000001 se devuelve WIN, si es menor LOSE y si es igual DRAW.

reldif <- function(x,y)

Calcula la diferencia de los valores x e y, dividiendo el resultado por el valor de x.

CreateAggrDataJoin <- function()

Crea un conjunto de datos my_aggdata0 como resultado de una join de dos conjuntos.

no.dimnames <- function(a)

Elimina todos los nombres de las dimensiones para sacar por pantalla de una forma más compacta la información.

C.2 SCRIPT MDHM

MDHM <- function (DATAMART, INDICATOR, TECHNIQUE, verbose=TRUE)

La función principal del script es MDHM, que contiene como argumentos el DATAMART a cargar, el INDICATOR a agregar, la TECHNIQUE a aplicar y como último argumento si se quiere o no sacar por consola los comentarios sobre el seguimiento de la ejecución. El funcionamiento del script principal es el siguiente: se le pasa un conjunto de datos con un indicador y una técnica a aplicar, lo primero que realiza es cargar en memoria el conjunto a partir de un fichero que se encuentra en el directorio de trabajo definido previamente o mediante el script se le puede indicar. Esta carga en memoria se realizará con una llamada a la función *LoadDataset*.

Ya cargado el conjunto de datos en memoria y estructurado las jerarquías con sus atributos correspondientes, se procede a rellenar con ceros todos los valores nulos o que son NA del conjunto llamando a la función *CreateDatasetWithAllZeros*.

Una vez ya se tiene preparado el conjunto de datos, se procede a realizar para cada cubo:

1. La agregación del indicador agrupando por una lista de atributos que corresponde a los del cubo. Para todos los experimentos se ha utilizado la función de agregación sum, cosa que para el conjunto de CARS se ha utilizado la función de agregación mean, ya que no tiene sentido realizar el sumatorio del consumo de CO^2 de los coches si no la media del consumo.
2. Se realiza una partición de la información agregada del cubo en dos conjuntos, el de entrenamiento y el de prueba, a partir del año que se ha definido para realizar la partición.
3. Una vez ya particionada la información agregada, se procede a aplicar la técnica ejecutada a dicha partición (la del conjunto de entrenamiento) para entrenar el modelo llamando a la función *TRA_Learning*.
4. Luego, se coge la información agregada particionada como test y se le aplica el modelo generado y entrenado a partir de la partición del conjunto de entrenamiento.
5. Se realiza la predicción a nivel más bajo (*lowest level*) con todos los atributos de las jerarquías del conjunto de datos (esto solo se calcula una vez).
6. Luego, se realizan dos predicciones para cada cubo, una a bajo nivel: se coge los atributos del cubo y se predice el indicador utilizando las predicciones obtenidas por el cubo *lowest level*, y otro: se realiza la predicción sólo con los atributos del cubo.

A partir de todos estos cálculos y cubos generados, tanto al nivel más bajo (*lowest level*) como al nivel del cubo (*same level*), se han hecho los siguientes cálculos para cada uno de los cubos que se volcán a un fichero resultado para cada técnica:

- **mse**: Error cuadrático medio, es la diferencia de las predicciones realizadas al conjunto de test y lo que se espera que tiene que dar.

- **evar**: la variancia de la diferencia de las predicciones realizadas al conjunto de test y lo que se espera que se obtenga.
- **mse_ll:** Error cuadrático medio para el nivel más bajo (*lowest level*), es la diferencia de las predicciones realizadas al conjunto de test y lo que se espera que tiene que dar.
- **evar_ll**: la variancia de la diferencia de las predicciones realizadas al conjunto de test a nivel más bajo (*lowest level*) y lo que se espera que se obtenga.
- **ll_best** : se obtiene cuál de los dos errores cuadráticos es mayor, si devuelve un LOW es que gana el error cuadrático a nivel más bajo, si devuelve un WIN es que gana el error cuadrático a nivel del cubo y si devuelve DRAW es que los errores cuadráticos son iguales.
- **ll_reldif** : Diferencia de los errores cuadráticos al mismo nivel y al nivel más bajo .
- **resolution**: lista de cubos obtenidos a partir de las jerarquías definidas.
- **train**: cantidad de registros para el conjunto de entrenamiento.
- **test**: cantidad de registros para el conjunto de test.
- **total**: cantidad total de registros, sumando los registros del conjunto de entrenamiento y del conjunto de test para cada cubo.

D R-PROJECT

R es un entorno y un lenguaje (Project, 2014) para el cálculo estadístico y generación de gráficos. Se ofrece un lenguaje de programación completo con el que permite el añadirle nuevas funcionalidades. Actualmente, R es uno de los lenguajes más utilizados en investigación en estadística.

Está disponible libremente bajo la Licencia Pública General de GNU, y pre-compilado. Para los distintos sistemas operativos R utiliza una interfaz de línea de comandos, sin embargo, varias interfaces gráficas de usuario están disponibles para su uso con R, como la que se ha utilizado en esta tesis, el RStudio.

RStudio (R Studio, 2014) es el entorno de desarrollo integrado para R. Está disponible en código abierto y en ediciones comerciales y se puede utilizar en las plataformas Windows, Mac y Linux. También dispone de RStudio server que se puede utilizar desde la web.

D.1 ¿POR QUÉ R?

Se ha escogido la herramienta R por el hecho de que es una de las herramientas más potentes para realizar cálculos estadísticos (comparado con otras herramientas similares, R utiliza poca memoria) además de ser muy completa al tener infinidad de paquetes que se pueden descargar e instalar en la aplicación según necesidad. Para los estudios realizados, es necesario tener una herramienta estadística la cual proporcione un potente motor de cálculos y a la vez eficiente. También un aspecto positivo a destacar es que existe mucha documentación en la red, en la cual en algunos momentos ha sido útil. Proporciona una facilidad de lectura de ficheros alta, además de soportar muchos tipos de formatos.

Cabe destacar que es una herramienta gratuita y que tiene un IDE para hacer el trabajo más fácil además de ser gratuito, como el que hemos utilizado en los experimentos, el RStudio. El RStudio proporciona una interfaz gráfica muy potente y a la vez fácil de usar. Ayuda visualmente en todos los aspectos, por ejemplo:

- Los ficheros leídos los podemos visualizar en cualquier momento en forma de tabla.
- Se puede realizar una depuración exhaustiva, viendo en cada momento el estado de las variables, dataframes, vectores…
- Realiza una estructuración de los scripts la cual hacer que sea mucho más fácil entender los scripts.

D.2 OBTENCIÓN E INSTALACIÓN DE R

Depende del sistema operativo, pero todo se puede encontrar en la web oficial de R (Project, 2014). Descargar el ejecutable necesario según el sistema operativo que se vaya a usar. Si escogemos por ejemplo, para plataforma Windows, al ejecutar el archivo se instalará el sistema base y los paquetes recomendados. Si por lo contrario, escogemos la versión de GNU/Linux, existen dos opciones:

1. Obtener el R-x.y.z.tar.gz y compilar desde las fuentes. También bajar los paquetes adicionales e instalarlos.
2. Obtener binarios (ej.,*.deb para Debian,*.rpm para RedHat,SuSE, Mandrake).

D.3 PAQUETES

R consta de un sistema base y de paquetes adicionales que extienden su funcionalidad.

Para los experimentos realizados, hemos utilizado los siguientes paquetes:

- KKNN
- RWEKA
- RWEKAJARS
- E1071
- Entre otros que se han utilizado pero que al final han sido descartados por falta de recursos de memoria, como puede ser: IBK, LAZY, LBR…

Para instalar un paquete, al ejecutar el siguiente comando en la línea de comandos, R se encargara automáticamente de descargarlo e instalarlo:

install.packages('kknn',dependencies=TRUE)