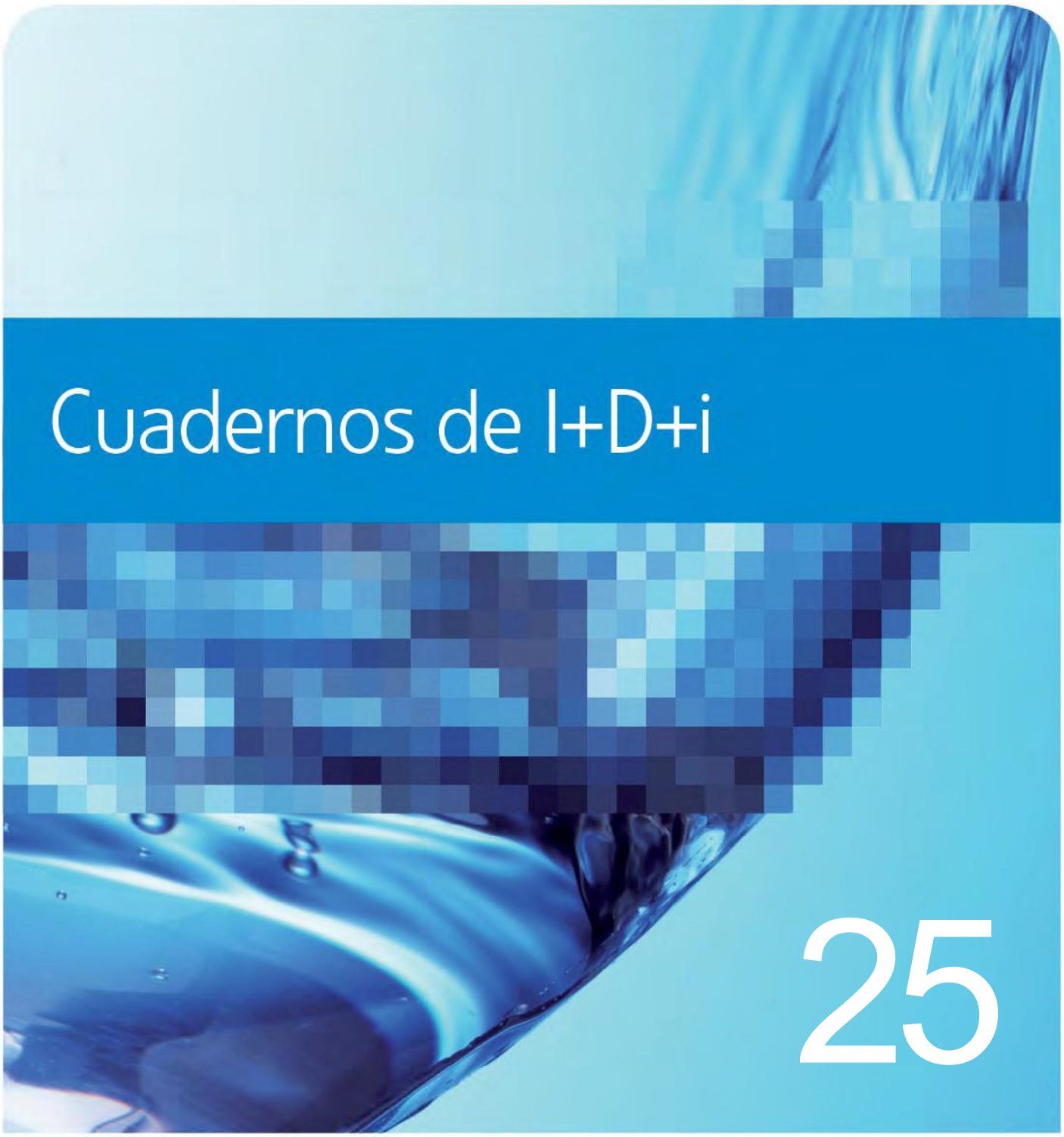


# Cuadernos de I+D+i

25

Canal de Isabel II

## Sistema de reconocimiento de patrones para identificación de usos finales del agua en consumos domésticos




**Autores**

*Pedro Luis Peñalver, Pablo García Rubí, Vanesa Pérez Salas*

*José Antonio Sánchez del Rivero, Roberto Díaz Morales, Julia Lastra García, Sergio García Caso*

**Dirección del estudio**

*Juan Carlos Ibáñez Carranza*

ISSN de la edición impresa: 2254-8955

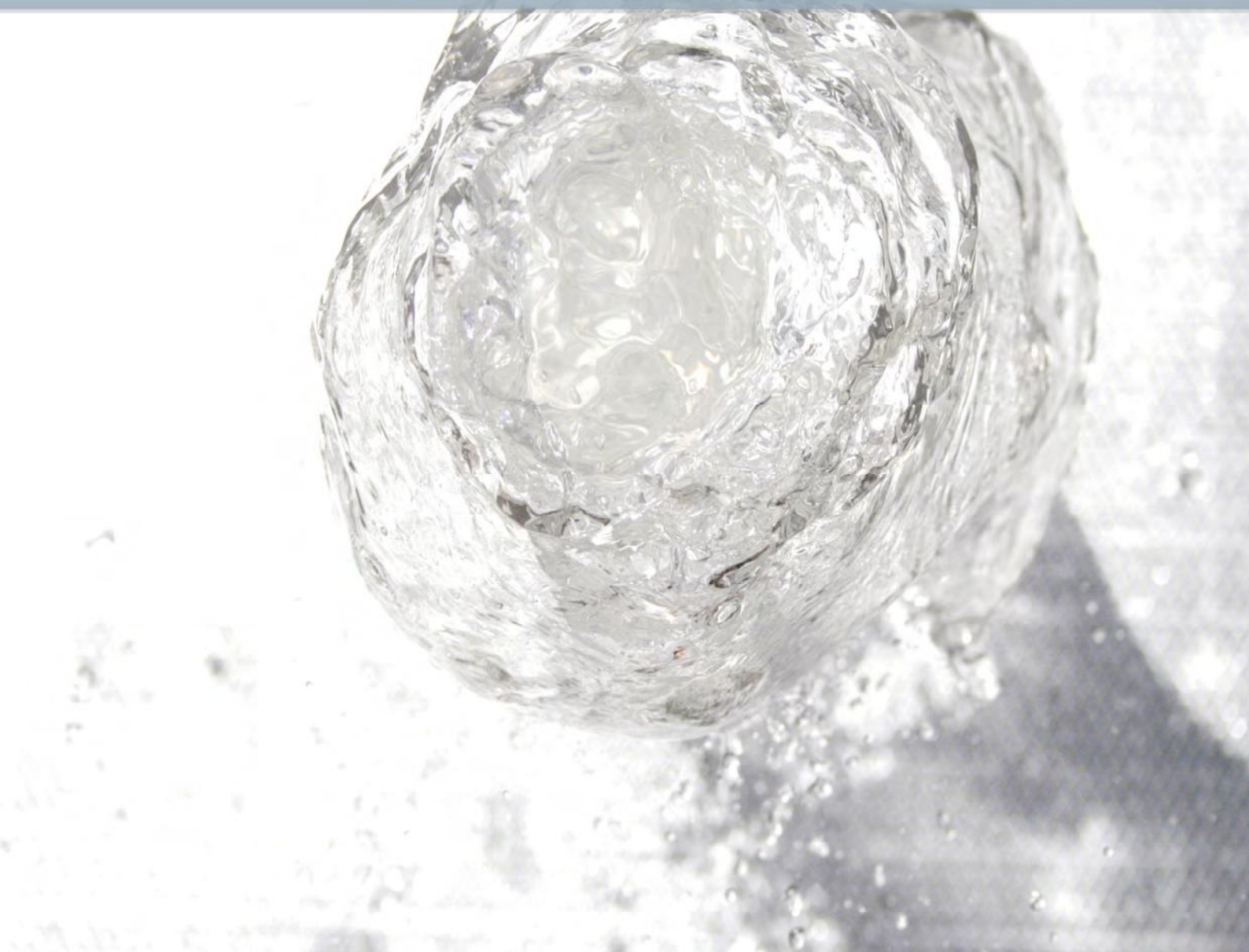
ISSN de la edición en soporte electrónico: 2340-1818

Depósito Legal: M-28367-2017

Canal de Isabel II

# 25

# Sistema de reconocimiento de patrones para identificación de usos finales del agua en consumos domésticos



# Exclusión de Responsabilidad

Las afirmaciones recogidas en el presente documento reflejan la opinión de los autores y no necesariamente la de Canal de Isabel II.

Tanto Canal de Isabel II como los autores de este documento declinan todo tipo de responsabilidad sobrevenida por cualquier perjuicio que pueda derivarse a cualesquiera instituciones o personas que actúen confiadas en el contenido de este documento, o en las opiniones vertidas por sus autores.

# Presentación

Los cuadernos de I+D+i de Canal de Isabel II forman parte de la visión sobre gestión del conocimiento de la empresa y del desarrollo de su Estrategia de I+D+i 2017-2020.

Son elemento de difusión de proyectos e iniciativas desarrollados y auspiciados desde la Empresa para la innovación en las áreas relacionadas con el servicio de agua en el entorno urbano.

Exponen las diferentes problemáticas abordadas en cada proyecto junto con los resultados obtenidos. La intención al difundirlos mediante estas publicaciones es compartir las experiencias y conocimientos adquiridos con todo el sector de servicios de agua, con la comunidad científica y con cuantos desarrollan labores de investigación e innovación. La publicación de estos cuadernos pretende contribuir a la mejora y eficiencia de la gestión del agua y, en consecuencia, a la calidad del servicio prestado a los ciudadanos.

Los títulos aparecidos en la colección de Cuadernos de I+D+i son los que figuran en la tabla siguiente.

## TÍTULOS EN LA COLECCIÓN DE CUADERNOS DE I+D+I

| *Nº colección* | *Año* | *Cuadernos Investigación, Desarrollo e Innovación publicados* |
|---|---|---|
| 1 | 2007 | Transferencias de derechos de agua entre demandas urbanas y agrarias. El caso de la Comunidad de Madrid |
| 2 | 2008 | Identificación de rachas y tendencias hidrometeorológicas en el ámbito del sistema d Canal de Isabel II |
| 3 | 2009 | Participación de Canal de Isabel II en el Proyecto Internacional de Eficiencia en la de Isabel II (IDMF) |
| 4 | 2008 | Microcomponentes y factores explicativos del consumo doméstico de agua en la Comunidad de Madrid |
| 5 | 2008 | El agua virtual y la huella hidrológica en la Comunidad de Madrid |
| 6 | 2008 | Estudio de potenciales de ahorro de agua en usos residenciales de interior |
| 7 | 2008 | Investigación sobre potenciales de eficiencia con el empleo de lavavajillas |
| 8 | 2010 | Precisión de la medida de los consumos individuales de agua en la Comunidad de Madrid |
| 9 | 2010 | Proyecto de investigación para la definición y evaluación de la aplicabilidad de un bioensayo para la determinación de la toxicidad del agua utilizando embriones de pez Cebra |
| 10 | 2010 | Eficiencia en el uso del agua en jardinería en la Comunidad de Madrid |
| 11 | 2010 | Técnicas de teledetección y sistemas de información geográfica para l evaluación de la demanda de agua para usos de exterior en la Comunidad de Madrid |
| 12 | 2010 | Estudio sobre la dinámica de cianotoxinas en dos embalses de abastecimiento de Canal de Isabel II |
| 13 | 2011 | Desarrollo de un sistema de validación, estimación ypredicción de consumos horarios por sectores para la red de distribución de Canal de Isabel II |
| 14 | 2011 | Seguimiento de la consolidación del desarrollo urbano en la Comunidad de Madrid mediante técnicas de teledetección |
| 15 | 2012 | Experiencias para la recuperación del fósforo de las aguas residuales en forma de estruvita en Canal de Isabel II |
| 16 | 2012 | Integración de la predicción meteorológica en los módulos de gestión del sistema de abastecimiento de Canal de Isabel II, mediante modelos de aportación diaria |
| 17 | 2012 | Mejora de la capacidad de pronóstico de aportaciones mensuales y estacionales en el ámbito de Canal de Isabel II |
| 18 | 2013 | Aportación de nutrientes desde la cuenca al embalse de Pinilla. Incidencia en el proceso de eutrofización |
| 19 | 2013 | Un nuevo criterio para el cálculo del caudal de agua residual urbana |
| 20 | 2014 | Gestión de Ideas en Canal de Isabel II de Isabel II: la experiencia*GENYAL* |

| Nº colección | Año | Cuadernos Investigación, Desarrollo e Innovación publicados |
|---|---|---|
| 21 | 2014 | Investigación sobre técnicas para la medición de subsidencias relacionadas con la explotación de acuíferos |
| 22 | 2015 | Régimen de precipitaciones en la Cuenca del Lozoya y adyacentes |
| 23 | 2016 | Estudio de observabilidad para la estimación del estado hidráulico de la redsectorizada de abastecimiento |
| 24 | 2016 | Estudio de casuística y modos de fallo en tuberías, acometidas y conjuntos de medida de la Comunidad de Madrid |

# ÍNDICE DE CONTENIDOS

# Resumen Ejecutivo

# Ficha Técnica

| | |
|---|---|
| ***Título del proyecto*** | **Sistema de reconocimiento de patrones para identificación de usos finales del agua en consumos domésticos** |
| ***Línea de investigación*** | **Aseguramiento del equilibrio disponibilidades / demandas** |
| ***Unidades de Canal de Isabel II implicadas*** | Subdirección I+D+i |
| ***Participación externa*** | Exeleria, Treelogic |
| ***Objeto y justificación del proyecto*** | Desarrollar un sistema automático de identificación de los usos finales del agua en las distintas aplicaciones domésticas, a partir de las señales registradas por contadores de precisión, utilizando para ello metodologías avanzadas de reconocimiento de patrones y clasificación supervisada de señales, tales como redes neuronales artificiales (RNA), métodos estadísticos u otros. |
| ***Contribución al estado del arte*** | Presenta un algoritmo matemático propio para la transformación automática de pulsos volumétricos en caudales instantáneos y la identificación de eventos asociados a diferentes usos domésticos, permitiendo la caracterización y cuantificación de cada uno de ellos.<br><br>Desarrolla dos metodologías para la clasificación de eventos, una basada en máquinas de vector soporte (SVM) y otra en redes neuronales artificiales (RNA), analizando y comparando los resultados obtenidos por ambos métodos.<br><br>La aplicación informática desarrollada permite el tratamiento masivo de datos de diferentes contadores y fechas, reduciendo la intervención del técnico operador a la selección de los datos a procesar. |
| ***Resumen del desarrollo del proyecto e hitos relevantes*** | • Recopilación de las lecturas de 375 contadores volumétricos con emisor de pulsos, con precisión de lectura de 1 y 0,1 litros.<br>• Formulación de un algoritmo matemático para la transformación de lecturas de pulsos en series temporales de caudales.<br>• Desarrollo de una aplicación informática (VBA sobre*Access*) para el tratamiento masivo de datos de lecturas de contadores, para la transformación de pulsos en caudales.<br>• Creación de una metodología para identificación de eventos a partir de las series de caudales obtenidas y desarrollo de un segundo módulo informático en VBA para su automatización.<br>• Desarrollo de dos procedimientos informáticos para la clasificación masiva de eventos (etiquetado) según usos finales, uno basado en Máquinas de Vector Soporte (SVM) y otro en Redes Neuronales Artificiales (RNA), a partir de un etiquetado previo de un determinado número de eventos realizados por operador que permite crear modelos de clasificación específicos para cada contador (modelos individuales). Todo ello, diferenciando según precisión del contador (1 ó 0,1 litros).<br>• Para los dos procedimientos de clasificación, RNA y SVM, se han desarrollado sendos métodos para clasificar eventos identificados a partir de lecturas de nuevas instalaciones, es decir, contadores de los que no se dispone etiquetado previo, creando modelos generales que suplen la inexistencia de modelos individuales para estas nuevas instalaciones.<br>• Diseño de informes gráficos de resultados<br>• Integración de los diferentes módulos en una única aplicación informática que contemple todo el procedimiento. |

| | |
|---|---|
| ***Resumen de resultados obtenidos*** | • Para contadores de 1 litro de precisión, en términos globales, es decir, considerando el conjunto de los eventos identificados con las lecturas de todos los contadores analizados, la clasificación de estos eventos mediante modelos individuales basados en RNA presenta un porcentaje global de acierto, en términos de volumen, del 86%, frente al 63% de los basados en SVM.<br>• Para contadores de 0,1 litros de precisión, los porcentajes de acierto alcanzan el 91% (RNA) y el 85% (SVM).<br>• Con los modelos generales, la precisión global disminuye en comparación con los modelos individuales, siendo del 82% para RNA, y del 75% para SVM, también en términos de volumen. |
| ***Líneas de Investigación abiertas para la continuación de los trabajos*** | • Optimización del proceso de etiquetado automático y aplicación a datos masivos.<br>• Evaluación del impacto de campañas de ahorro de agua mediante la herramienta de etiquetado automático.<br>• Aplicabilidad del etiquetado automático a grandes patrones de consumo.<br>• Generación de cuadros de mando integral ligados al etiquetado automático. |

# Resumen Ejecutivo

## OBJETIVO

El objetivo principal de este estudio queda resumido en el propio título y no es otro que el desarrollo de un sistema automático de identificación de los usos finales del agua en las distintas aplicaciones domésticas, a partir de las señales registradas por contadores de precisión, utilizando para ello metodologías avanzadas de reconocimiento de patrones y clasificación supervisada de señales, como son las redes neuronales artificiales (**RNA**) y métodos basados en máquinas de vectores soporte (**SVM**).

Los contadores utilizados en este trabajo no miden caudales directamente, sino que están equipados con un emisor digital de pulsos, el cual emite una señal (pulso) cada vez que se consume un volumen determinado (1 litro ó 0,1 litros, según la precisión del aparato). En consecuencia, ha sido necesario idear previamente una manera de transformar dichos registros de pulsos en caudales.

Cabe destacar que los métodos desarrollados en este proyecto también podrían aplicarse sobre otro tipo de registradores como, por ejemplo, los de caudal en cuyo caso no ha lugar la transformación de pulsos en caudales.

Todos estos procedimientos han sido automatizados y programados mediante una aplicación informática que permite el tratamiento masivo de datos procedentes de las lecturas de multitud de contadores y amplios periodos de tiempo, sin necesidad de contar con la participación de un operario.

## MÉTODO

La información de partida de los trabajos aquí presentados incluye los datos correspondientes a los pulsos registrados por 375 contadores, desde enero de 2008 a julio de 2015, lo que supone el procesamiento de los registros de unos 20.340 meses, aproximadamente, que incluyen más de 34,65 millones de eventos de uso de agua. Estos registros fueron clasificados anteriormente mediante un método de identificación automática de usos desarrollado por Canal de Isabel II, basado en redes bayesianas. Para cada contador existen al menos dos meses de datos clasificados manualmente por operador, que sirvieron como datos de entrenamiento para las redes bayesianas. Este etiquetado previo (clasificación) realizado de forma manual se considera representativo de la realidad para cada contador y es el que se ha utilizado para, mediante un proceso de entrenamiento, la generación de modelos RNA y SVM, basados en los cuales se clasifican nuevos eventos de manera automática.

Las clases de usos utilizados en esta clasificación o etiquetado han sido las siguientes:

- ✓ Grifos
- ✓ Cisternas
- ✓ Duchas, que incluye las bañeras
- ✓ Lavadora
- ✓ Lavavajillas
- ✓ Piscina
- ✓ Riego
- ✓ Fugas

A partir de esta información se han realizado los trabajos que se relacionan a continuación, agrupados por módulos:

- Módulo 1. Transformación de pulsos en caudales
- Módulo 2. Identificación de eventos
- Módulo 3. Desarrollo de los modelos de redes neuronales artificiales (RNA)
- Módulo 4. Desarrollo de modelos basados en máquinas de vector soporte (SVM)
- Módulo 5. Asignación de usos finales
- Módulo 6. Creación de nuevos modelos para futuras instalaciones

Como trabajo previo a estos módulos se ha realizado una revisión del estado del arte en investigación sobre reconocimiento de patrones y clasificación automática de usos finales del agua.

La transformación de las lecturas de pulsos en caudales y la identificación de eventos ha sido realizada mediante sendos algoritmos matemáticos ideados*ad hoc*, de manera que permitan esta transformación de una forma automática, que no dependa de los criterios, más o menos subjetivos, de un operador.

Para generar los modelos, tanto los basados en RNA como los basados en SVM, es necesario el etiquetado previo (arriba mencionado) de un cierto número de eventos, y sobre esa base “aprender” y establecer los algoritmos necesarios para poder asignar automáticamente la etiqueta correspondiente al resto de los eventos.

Los contadores utilizados son de dos tipos, en cuanto a su precisión: de 1 litro y de 0,1 litros. Esta precisión corresponde al volumen consumido que da lugar a la emisión de un pulso y se ha tenido en cuenta a la hora de formular los diferentes algoritmos matemáticos, tanto de transformación de pulsos en caudales como de generación de modelos RNA y SVM.

Todos estos trabajos han requerido el desarrollo de diferentes aplicaciones informáticas, específicas para cada uno de los módulos. Para los dos primeros módulos –transformación de pulsos en caudales e identificación de eventos– se ha utilizado *Visual Basic for Applications* (VBA), bajo entorno *Access*[1], mientras que, en el resto de los módulos, el entorno de producción utilizado ha sido *Miniconda*, una versión reducida de *Anaconda*[2], distribución y gestor de paquetes basados en *Python*[3], que sólo contiene *Conda* y *Python*.

Como colofón de los trabajos realizados, se integran todas estas aplicaciones informáticas en una sola, desarrollada bajo entorno *Access*, que incluye todo el procedimiento necesario para la identificación y clasificación de los usos finales del agua en consumos domésticos.

Esta aplicación permite clasificar los eventos identificados a partir de nuevas lecturas de los contadores para los que ya se ha generado un modelo específico de clasificación (modelo individual), y el desarrollo de nuevos modelos para contadores de otras instalaciones, siempre que se disponga de un conjunto de datos de entrenamiento (clasificados manualmente). Si no se dispone de esos datos de entrenamiento, pueden utilizarse los modelos genéricos construidos en base a los eventos tratados en este estudio.

## RESULTADOS Y CONCLUSIONES

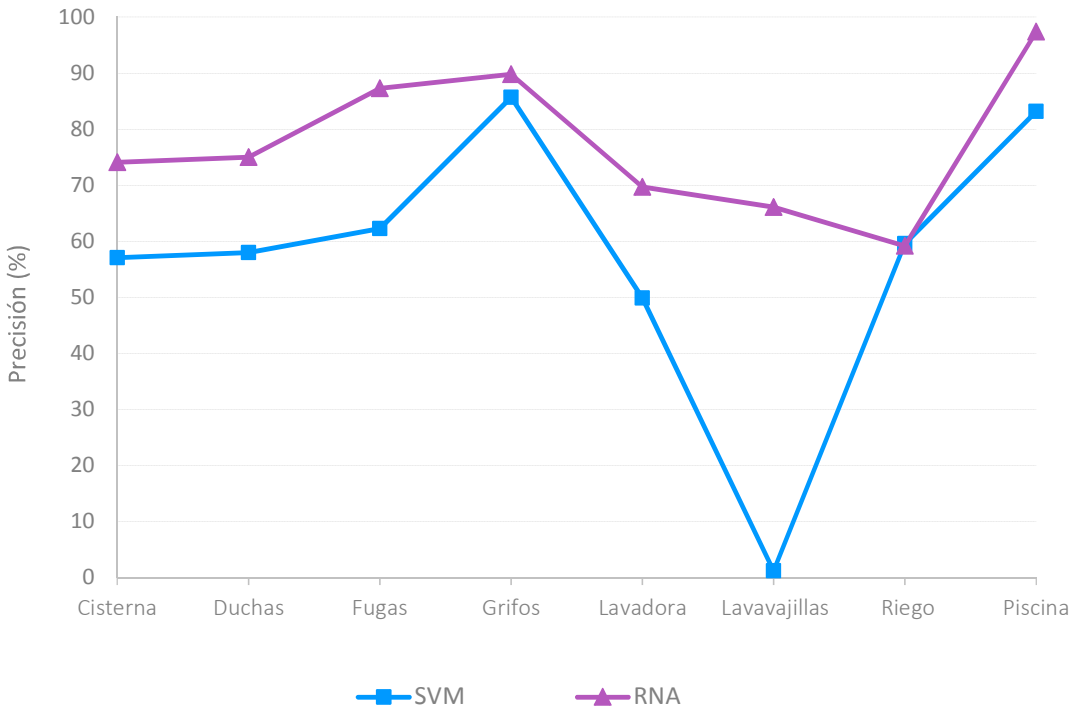
Para contadores de precisión de 1 litro los resultados obtenidos mediante la Red Neuronal son significativamente mejores que los obtenidos con las SVM, tanto en términos globales como a nivel de contador.

En efecto, si se tienen en cuenta todos los resultados obtenidos con ambos métodos de clasificación, se tiene que el grado de acierto con las SVM es del 67,41%, mientras que el con la Red Neuronal es del 81,78%, en cuanto al número de eventos correctamente clasificados se refiere. En términos de volumen correctamente clasificado, el grado de acierto es muy similar, siendo del 63,41% para las SVM y del 85,76% para la Red Neuronal.

Observando los resultados de todos los contadores de este tipo y distinguiendo según los usos, la precisión obtenida con la Red Neuronal es siempre mayor que la obtenida con las SVM. La Figura 1 muestra la diferencia en la precisión considerando cada tipo de uso por separado.

[1] *Access: ©Microsoft*
[2] *Conda: Open Source y Con Nueva Licencia BSD*
[3] *Phyton: Lenguaje de Programación con Licencia PSFL*

**FIGURA 1. PRECISIÓN DE LOS ALGORITMOS SEGÚN EL USO EN CONTADORES DE 1 LITRO**



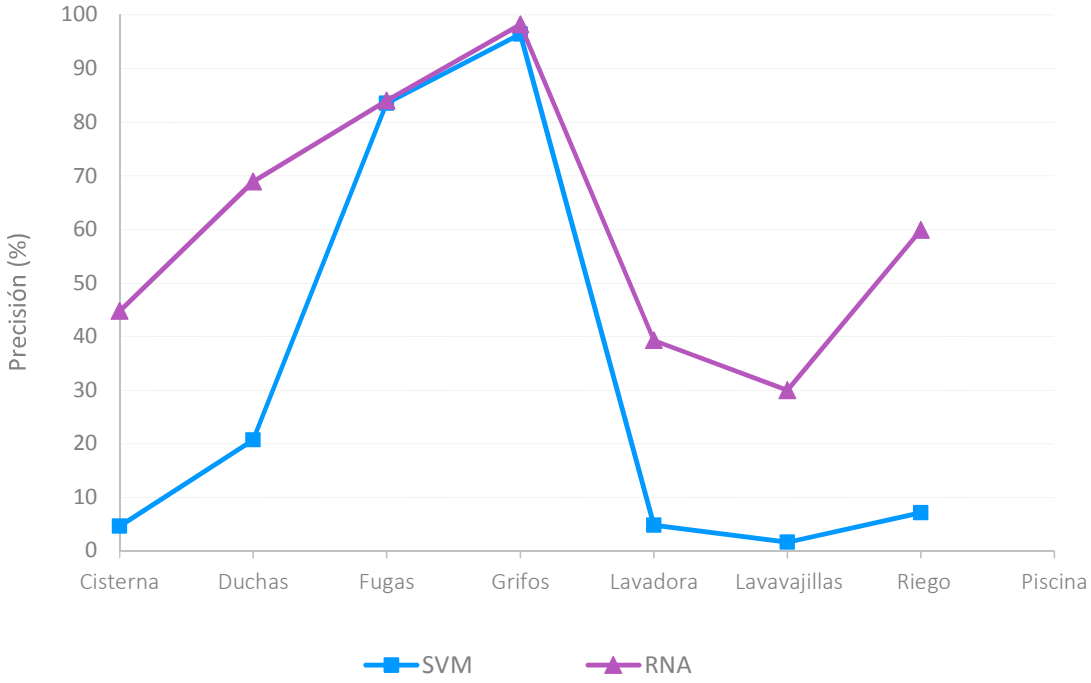

A nivel de contadores individuales, no es posible encontrar ninguno para el cual los resultados de las SVM sean mejores que los de la Red Neuronal.

El número de eventos correctamente clasificados con la Red Neuronal aumenta de media un 22,35%, llegando a superar el 40% de aumento en 19 ocasiones.

Por otro lado, para los 19 **contadores de 0,1 l de precisión** analizados, al igual que con los contadores de 1 litro, los resultados obtenidos mediante la Red Neuronal son mejores que los obtenidos con las SVM, tanto en términos globales como a nivel de contador. El grado de acierto medio de las SVM y de la Red Neuronal es de 84,78% y de 91,19%, respectivamente. Si se cuantifica el porcentaje en volumen, estos valores bajan a 73,5 % y 85,9%. Esto se debe a que la mayoría de los eventos son de tipo *Grifos,* uso con volúmenes medios bajos, que clasifican muy bien ambos métodos.

La Figura 2 refleja muy bien esta observación. La precisión en la clasificación de los *Grifos* es superior al 95% con los dos algoritmos. Como la mayoría de los eventos son de este tipo, se producen muchas clasificaciones erróneas, pues los métodos reconocen el sesgo en la distribución de eventos y tienden a clasificar como *Grifos* eventos de otra naturaleza. A pesar de ello, se aprecia una mejoría notable en la precisión de la clasificación de eventos tipo *Cisternas, Duchas, Lavadora, Lavavajillas* y *Riego*.

**FIGURA 2. PRECISIÓN DE LOS ALGORITMOS DE CLASIFICACIÓN, SEGÚN EL USO EN CONTADORES DE 0,1 LITROS**




Asimismo, por lo que respecta al contador, tampoco es posible encontrar ningún evento para el que los resultados de las SVM sean significativamente mejores que los de la Red Neuronal.

El número de eventos correctamente clasificados con la Red Neuronal aumenta de media un 8,5%, llegando a superar un aumento del 20%, en 3 ocasiones.

Para los **modelos generales**, la precisión global del método es del 75,18%, con las SVM y del 82,17% con la Red Neuronal.

Comparando los modelos individuales con los generales se concluye que, como era esperable, los resultados son mejores aplicando los modelos en contadores individuales.

En las tablas y gráficos siguientes se muestran los resultados obtenidos mediante los dos métodos de clasificación, RNA y SVM, para la totalidad de los datos, procesados en el periodo que comprende desde enero de 2008, a julio de 2015.

En concreto, la Tabla 1 y las figuras 3 a 6 reflejan los resultados de distribución según usos del consumo total en el periodo enero de 2008, a julio de 2015, para el método de clasificación RNA. De la misma manera, la Tabla 2 y las figuras 7 a 10 muestran los resultados de clasificación para el método de SVM según usos del consumo total, durante el periodo referido.

**TABLA 1. RESULTADOS DE LA CLASIFICACIÓN MEDIANTE RNA. DISTRIBUCIÓN SEGÚN USOS DEL CONSUMO TOTAL EN EL PERIODO ENERO-2008 A JULIO-2015**

| Uso | Consumo total (m³) | | Consumo medio mensual (m³) | Nº de eventos total | | Nº de eventos, media mensual | Consumo medio, por evento (l) |
|---|---|---|---|---|---|---|---|
| Grifos | 54.198 | 41% | 595,58 | 21.028.397 | 61% | 231.081 | 2,58 |
| Cisternas | 13.662 | 10% | 150,14 | 2.614.989 | 8% | 28.736 | 5,22 |
| Duchas | 43.323 | 33% | 476,08 | 2.338.834 | 7% | 25.701 | 18,52 |
| Lavadora | 9.133 | 7% | 100,36 | 1.291.398 | 4% | 14.191 | 7,07 |
| Lavavajillas | 1.951 | 1% | 21,68 | 1.883.072 | 5% | 20.923 | 1,04 |
| Piscina | 88 | 0% | 0,99 | 1.608 | 0% | 18 | 54,56 |
| Riego | 2.751 | 2% | 30,57 | 48.657 | 0% | 541 | 56,54 |
| Fugas | 5.543 | 4% | 60,91 | 5.442.828 | 16% | 59.811 | 1,02 |
| **Total** | **130.649** | **100%** | **1.436,30** | **34.649.783** | **100%** | **381.003** | **3,77** |

**FIGURA 3. RESULTADOS DE LA CLASIFICACIÓN MEDIANTE RNA.DISTRIBUCIÓN CONSUMO TOTAL M³**

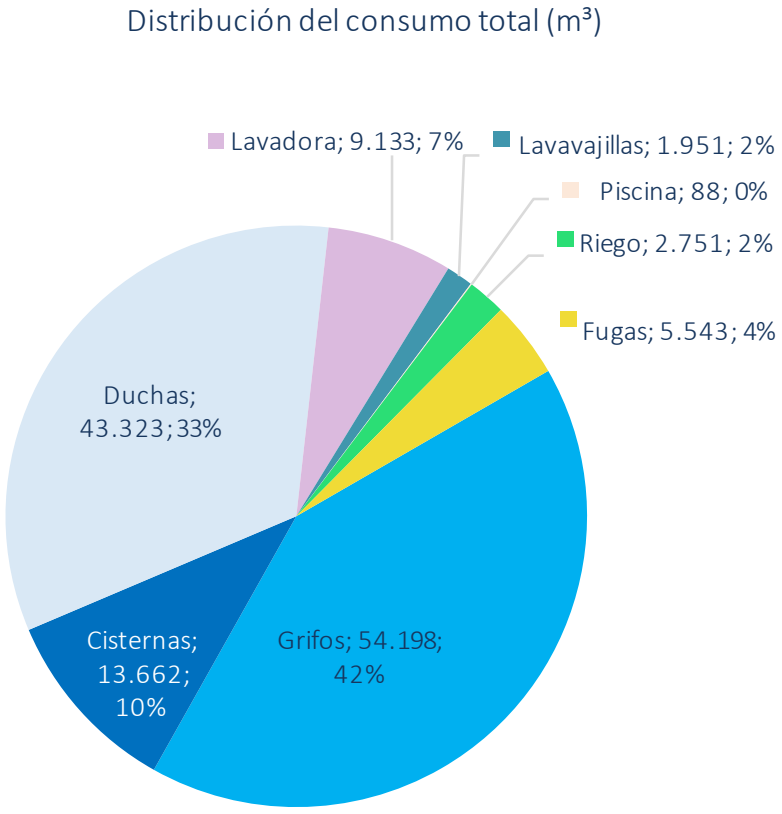


**FIGURA 4. RESULTADOS DE LA CLASIFICACIÓN MEDIANTE RNA. DISTRIBUCIÓN DEL NÚMERO TOTAL DE EVENTOS**

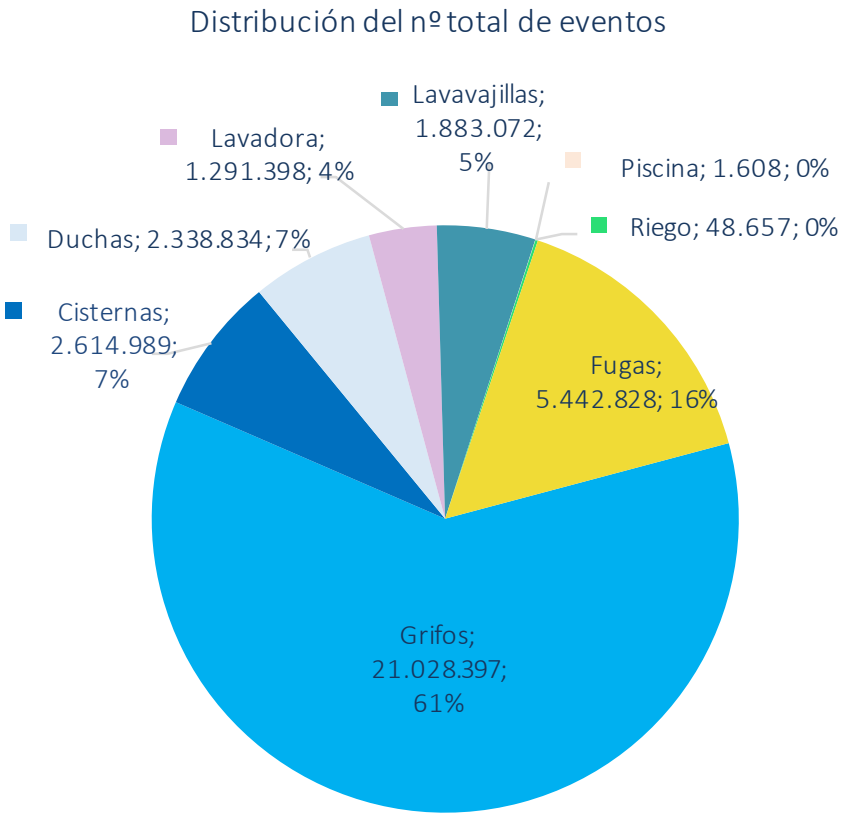



**FIGURA 5. RESULTADOS DE LA CLASIFICACIÓN MEDIANTE RNA. CONSUMO MEDIO POR EVENTO (L)**

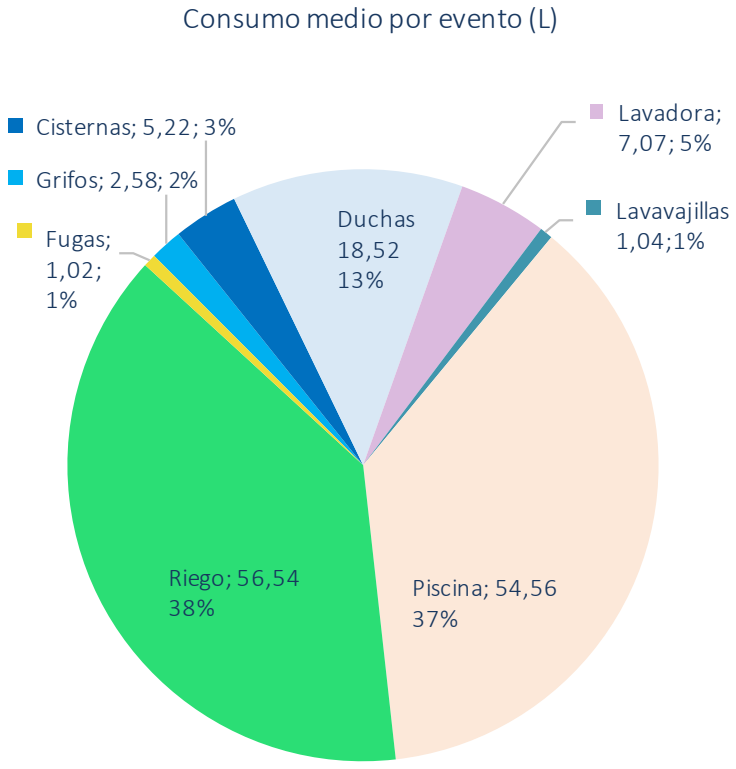


**FIGURA 6.** RESULTADOS DE LA CLASIFICACIÓN MEDIANTE RNA. DISTRIBUCIÓN DEL NÚMERO DE EVENTOS MEDIO MENSUAL

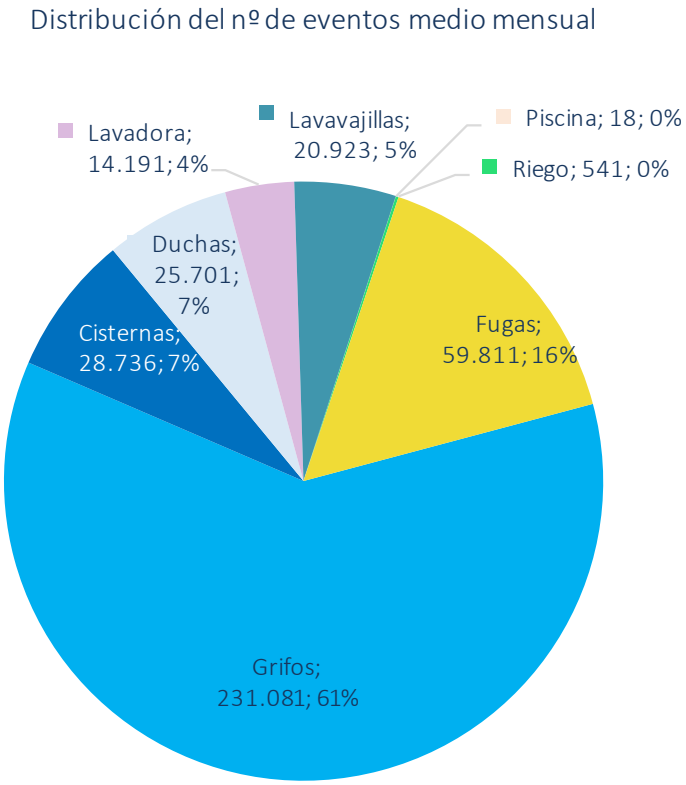



**TABLA 2.** RESULTADOS DE LA CLASIFICACIÓN MEDIANTE SVM. DISTRIBUCIÓN SEGÚN USOS DEL CONSUMO TOTAL EN EL PERIODO ENERO-2008 A JULIO-2015

| Uso | Consumo total (m³) | | Consumo medio mensual (m³) | Nº de eventos total | | Nº de eventos, medio mensual | Consumo medio por evento (l) |
|---|---|---|---|---|---|---|---|
| Grifos | 59.561 | 46% | 654,51 | 24.915.468 | 72% | 273.796 | 2,39 |
| Cisternas | 15.244 | 12% | 167,52 | 2.379.525 | 7% | 26.149 | 6,41 |
| Duchas | 39.316 | 30% | 432,04 | 1.900.028 | 5% | 20.879 | 20,69 |
| Lavadora | 9.535 | 7% | 104,78 | 1.205.685 | 3% | 13.249 | 7,91 |
| Lavavajillas | 152 | 0% | 1,69 | 78.946 | 0% | 877 | 1,93 |
| Piscina | 57 | 0% | 7,11 | 755 | 0% | 94 | 75,32 |
| Riego | 2.688 | 2% | 29,86 | 50.218 | 0% | 558 | 53,52 |
| Fugas | 4.097 | 3% | 45,02 | 4.119.158 | 12% | 45.265 | 0,99 |
| **Total** | **130.649** | **100%** | **1.442,53** | **34.649.783** | **100%** | **380.869** | **3,77** |

**FIGURA 7.** RESULTADOS DE LA CLASIFICACIÓN MEDIANTE SVM.DISTRIBUCIÓN CONSUMO TOTAL M̂

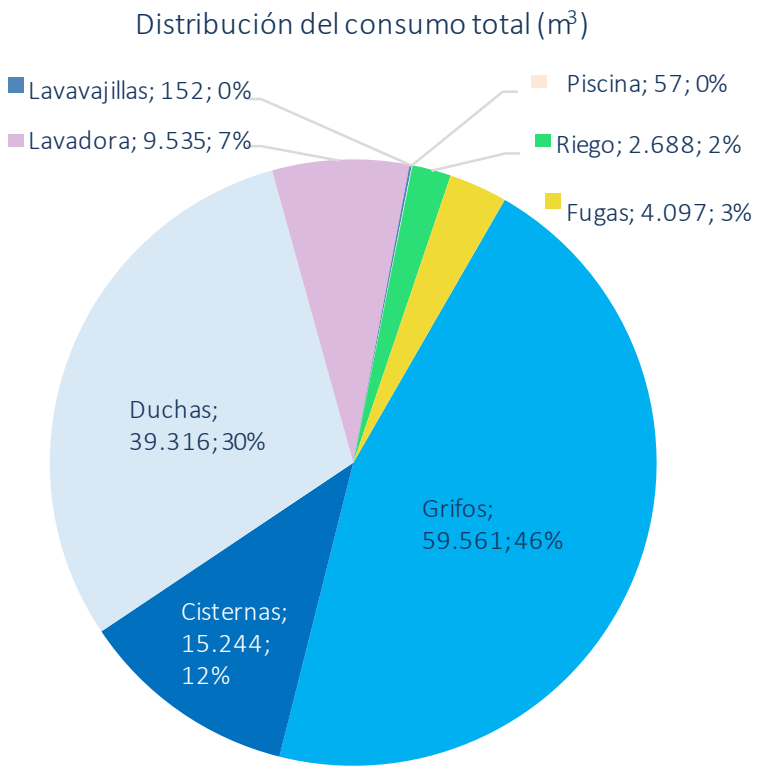



**FIGURA 8.** RESULTADOS DE LA CLASIFICACIÓN MEDIANTE SVM. DISTRIBUCIÓN DEL NÚMERO TOTAL DE EVENTOS

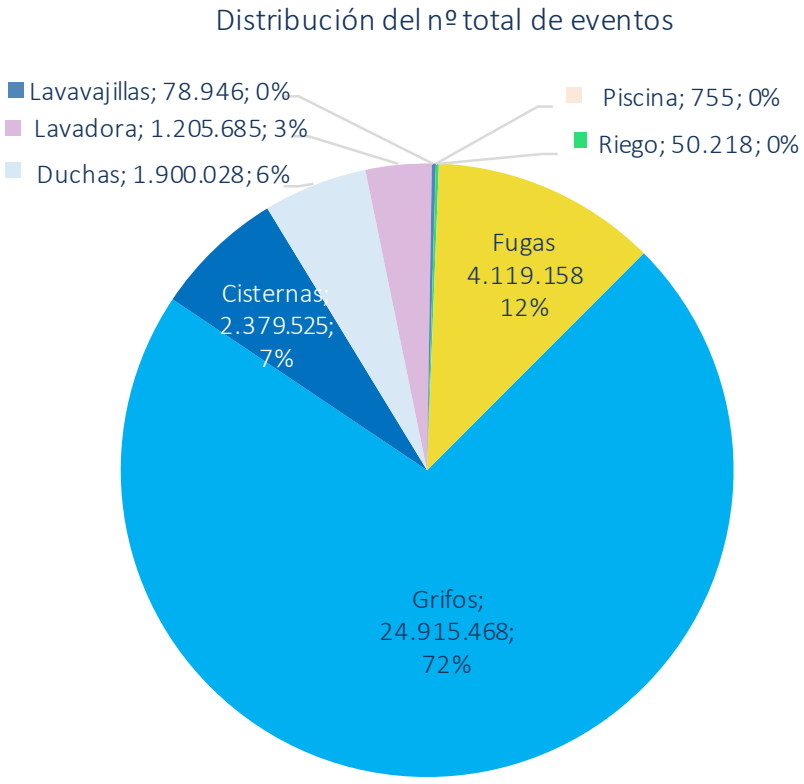


**FIGURA 9. RESULTADOS DE LA CLASIFICACIÓN MEDIANTE SVM. CONSUMO MEDIO POR EVENTO (L)**

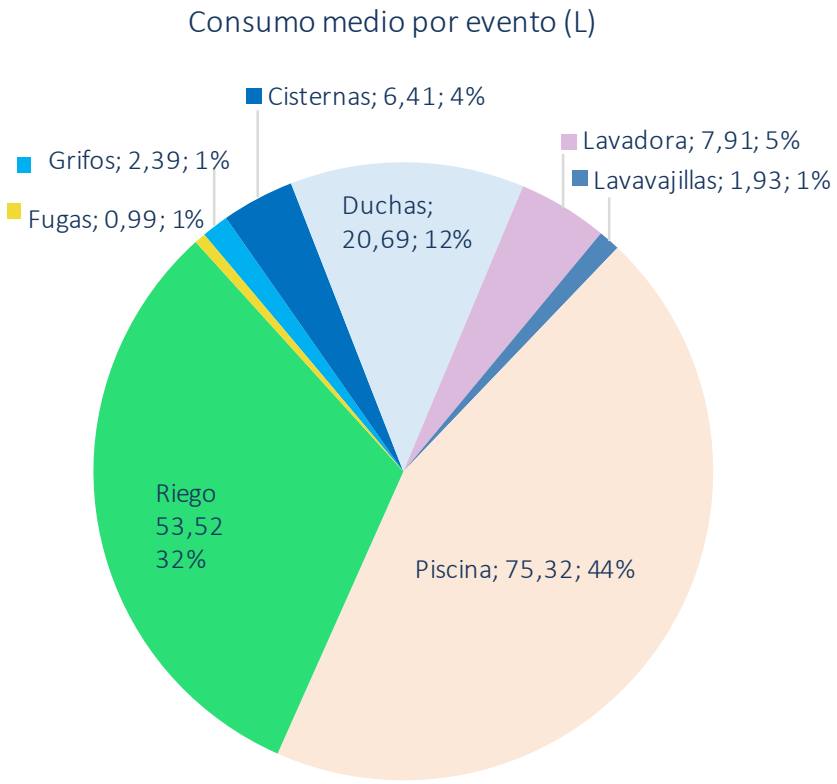
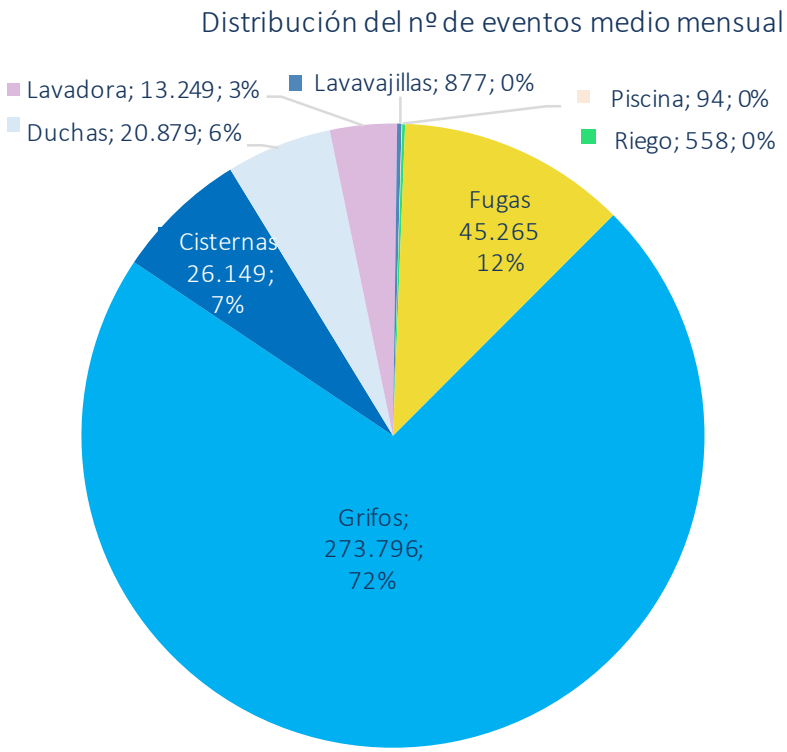



**FIGURA 10. RESULTADOS DE LA CLASIFICACIÓN MEDIANTE SVM. DISTRIBUCIÓN DEL NÚMERO DE EVENTOS MEDIO MENSUAL**



# 1. Introducción

Desde comienzos de 2008 Canal de Isabel II lleva monitorizando una muestra de viviendas repartidas por toda la región, que comenzó con contadores volumétricos clase C con emisor de pulsos con precisión de un litro; a partir del año 2011 los contadores instalados fueron de clase D con precisión de un decilitro.

Desde el inicio del proyecto y hasta la fecha se han monitorizado más de quince millones de horas de consumo, y se han contabilizado del orden de 140 millones de litros de agua, en un total de 375 instalaciones diferentes.

Para la identificación de los usos finales, a partir de los caudales (pulsos) registrados, se utilizó el software comercial *Trace Wizard*®, de la empresa *Aquacraft*. Esta aplicación realiza una clasificación preliminar de los distintos *eventos* detectados, que debe ser supervisada por un operador, el cual decide la asignación final de usos. Este método requiere una gran cantidad de horas de operador, estimada en 1-2 horas por cada semana de registro de consumos en una vivienda, lo que lo hace prácticamente inviable para la monitorización de una muestra de un tamaño considerable, durante varios años,

Por ello, a principios de 2009 se desarrolló una metodología de identificación automática de usos, basada en métodos bayesianos, y heurística de mínima entropía para discretización de las series de datos de las muestras. El método se basa en el análisis estadístico de series de datos, de entre dos y tres meses de duración, de cada usuario, procesada por operador, que se utiliza como fase de entrenamiento. Este análisis estadístico, consiste en calcular la probabilidad de que un determinado evento, dadas sus características de: duración, caudal punta y caudal medio, corresponda a cada uno de los distintos usos tipificados, adoptando el que presenta una probabilidad mayor.

Por tanto, en la actualidad, para cada una de las 375 instalaciones, existe una serie histórica de como mínimo dos meses de duración, de información procesada por operador, y el resto con el clasificador bayesiano.

En el proyecto que se describe en este documento se ha abordado el desarrollo de métodos más avanzados de clasificación de los eventos del consumo doméstico de agua, que mejoren las prestaciones de las técnicas estadísticas actuales, aumentando su fiabilidad, versatilidad y facilidad de uso, y que puedan ser exportados a otros sistemas de abastecimiento de agua.

El proceso de clasificación de eventos requiere un aprendizaje o entrenamiento previo que permita el reconocimiento de los patrones de los diferentes usos. Este aprendizaje se puede realizar de forma automática aplicando diferentes modelos, que se suelen clasificar en cinco categorías, a saber:

- redes neuronales
- aprendizaje basado en casos o instancias
- algoritmos genéticos
- inducción de reglas
- aprendizaje por analogía.

Las recientes comparaciones de las que han sido objeto estos modelos hace que hoy se acepten como muy similares todos ellos, al menos en cuanto a capacidad de aprendizaje se refiere. En aplicaciones específicas pueden ser más eficientes unos que otros, pero las diferencias que se manejaban hace unos años están en entredicho. Cada vez está más aceptado, por ejemplo, que el aprendizaje llevado a cabo por las redes neuronales, calificado de aprendizaje subsimbólico, no es más subsimbólico que las reglas de clasificación, aunque se admite que hay aspectos que los diferencian y que hacen que cada uno de ellos tenga aplicaciones específicas.

Hay muchas aplicaciones del **aprendizaje automático**, siendo la **minería de datos** para la clasificación una de las más significativas. Establecer relaciones entre múltiples características es un proceso complicado y candidato al fallo, lo que hace difícil el diseño de soluciones para este tipo de problemas. Las estrategias de **aprendizaje automático** ofrecen una solución eficiente ante este problema. Estas técnicas se demuestran válidas para la problemática de la clasificación (asignación final de un valor) teniendo en cuenta ciertas características (valores de entrada).

Cada registro en un conjunto de datos usado por los algoritmos de **aprendizaje automático** se representa usando las mismas características. Estas características pueden ser continuas, discretas o binarias. Si cada registro viene etiquetado con su valor de salida correcto (etiqueta), se denominan “estrategias de aprendizaje supervisado”. En contra, las estrategias que utilizan registro sin etiquetar se denominan “estrategias de aprendizaje no supervisado”.

La problemática de clasificación que incumbe el ámbito de este proyecto se centra en las estrategias de aprendizaje supervisado. Estas estrategias se engloban dentro del concepto de aprendizaje inductivo supervisado. La inducción en un sentido amplio consiste en encontrar propiedades comunes a un subconjunto finito de elementos de un cierto dominio y considerar esas propiedades extensibles a cualquier elemento del dominio. Una buena inducción tratará de preservar la verdad, para ello obviará aquellos elementos que contengan errores produciendo ruido en la información que proporcionan, y hasta puede contemplar excepciones al conocimiento extraído.

# 2. Objetivos

El objetivo principal de este estudio queda resumido en el propio título y no es otro que el desarrollo de un sistema automático de identificación de los usos finales del agua en las distintas aplicaciones domésticas, a partir de las señales registradas por contadores de precisión, utilizando para ello metodologías avanzadas de reconocimiento de patrones y clasificación supervisada de señales, como son las redes neuronales artificiales (RNA) y métodos basados en máquinas de vectores soporte (SVM). Para lograr este objetivo, ha sido necesario desarrollar diferentes algoritmos y procedimientos que permitan automatizar las diferentes fases de este proceso, cuales son:

- Transformar en caudales las lecturas de pulsos de los contadores
- Identificar eventos
- Generar modelos de aprendizaje
- Clasificar eventos

Si bien los contadores utilizados en este trabajo no miden caudales directamente sino que están equipados con un emisor digital de pulsos –emite una señal (pulso) cada vez que se consume un volumen determinado (1 litro ó 0,1 litros, según la precisión del aparato)–, la metodología desarrollada también podría aplicarse a otro tipo de registros de precisión, como puede ser la medición directa de caudales, ya que los pulsos en sí no forman parte de la entrada de datos de los modelos, sino la información desprendida de los caudales asociados a los registros continuos de pulsos.

Los diferentes usos utilizados en esta clasificación o etiquetado han sido los siguientes:

- Grifos
- Cisternas
- Duchas (que incluye las bañeras)
- Lavadora
- Lavavajillas
- Piscina
- Riego
- Fugas

Cada una de estas fases ha dado lugar a los correspondientes módulos informáticos, que finalmente han sido agrupados en una única aplicación.

Dicha aplicación informática permite el tratamiento masivo de datos procedentes de las lecturas de multitud de contadores y amplios periodos de tiempo sin necesidad de contar con la participación de un operario.

En este estudio se ha trabajado con los datos correspondientes a los pulsos registrados por 375 contadores desde enero de 2008 a julio de 2015, lo que supone el procesamiento de los registros de 15 millones de horas y un volumen total contabilizado de 140 millones de litros aproximadamente, en unos 34,65 millones de eventos de uso de agua.

# 3. Estado del arte

La clasificación automática de los usos finales del agua en el ámbito doméstico es un problema de interés actual cuya solución pasa por la realización de precisos esquemas de gestión del agua y por una medición adecuada del consumo de la misma. Las importantes mejoras en esta medición y la implantación de técnicas de análisis de datos han permitido abordar este problema con resultados positivos. El deseo de mejorar la monitorización y el análisis del consumo de agua ha permitido la posibilidad de coleccionar los datos de consumo a través de un sistema de medición inteligente y transferirlos y almacenarlos en un repositorio para su posterior análisis. Pero para llevar a cabo este análisis del consumo se requiere información filtrada y procesada, es decir, datos de consumo desagregados según su uso. Por tanto, la clave para desarrollar cualquier herramienta analítica, o para que las empresas encargadas de la gestión del agua amplíen el conocimiento de sus clientes, es necesario el desarrollo de algoritmos de reconocimiento de patrones que sean capaces de categorizar flujos y/o eventos en usos domésticos habituales (lavadora, lavavajillas, etc.).

## 3.1. ESTADO DEL ARTE EN LOS SISTEMAS DE MEDICIÓN DE CONSUMOS

Las primeras referencias en medición de consumos de agua domiciliarios, para la determinación de usos finales del agua, datan de finales de siglo XX. En la actualidad se siguen desarrollando este tipo de estudios encaminados a una mejor comprensión de las pautas de consumo domiciliario.

Canal de Isabel II inició trabajos de investigación en este campo en 2001, con distintas fases de toma de datos entre 2001 y 2003, y a lo largo de 2006 (Cuaderno I+D+i nº 4[4]). Actualmente, desde 2008, mantiene una muestra estable de unas 300 viviendas, monitorizada de forma continua.

De los 22 estudios que se han localizado y analizado, más de la mitad corresponden a trabajos realizados en ciudades australianas, entre los años 2005 y 2014; cuatro son de ciudades de EEUU, tres en California (años 2004 y 2010) y uno en Seattle (año 2000); tres en España, el de Canal de Isabel II en la Comunidad de Madrid (2001 – 2006), otro centrado en la ciudad de Zaragoza (año 2010) y el tercero en varias ciudades (año 2002); y los otros dos estudios restantes se realizaron en varias ciudades de Nueva Zelanda (2007) y en Abu Dhabi, Emiratos Árabes (2013). En cuanto al número de viviendas involucradas, era variable, aunque lo más recurrente es abarcar entre 100 y 300 instalaciones, llegando en algún caso a 384 viviendas (Zaragoza, 2010) e incluso 474 viviendas (Townsville, Australia, 2007). Respecto al tipo de contador, en todos los estudios, excepto en uno, se han utilizado contadores volumétricos con emisor de pulsos, y precisiones de al menos un decilitro, es decir, un pulso se emite cada vez que se consumen 0,1 litros; hay estudios con contadores de precisiones de hasta 0,014 litros.

Como se ve, Australia es uno de los países que más ha trabajado en el análisis de patrones de uso final del agua a partir de contadores inteligentes. Todos los estudios realizados en esta línea se basan en algoritmos de análisis de patrones predefinidos y los parámetros que caracterizan estos patrones deben ser revisados y ajustados de forma manual en cada caso, lo que genera un cierto grado de incertidumbre, principalmente, cuando se produce una superposición de eventos.

[4] Cuaderno de I+D+i 4, Microcomponentes y factores explicativos del consumo doméstico de agua en la Comunidad de Madrid

Para tratar de complementar estos métodos, la agencia de investigación australiana CSIRO está desarrollando un sistema prototipo que automáticamente determina el uso final del agua residencial sin la necesidad de un registro diario de eventos por parte de los propietarios de la vivienda. El prototipo incluye sensores capaces de identificar el uso de dispositivos interiores mediante el registro de las señales acústicas que se generan durante un evento. Cuando esta información es combinada con un registro de caudal mediante contador inteligente, el sistema permite definir con alta precisión el caudal de agua en cada uno de los dispositivos de una forma desagregada.

Para probar el prototipo se utilizó un estudio realizado por la Universidad de Griffith[5] durante 3 años, sobre uso final del agua en viviendas. Para ello se estudiaron los datos analizados sobre una muestra de 320 hogares en cuatro zonas de Australia (Ipswich, Brisbane, Gold Coast y Sunshine Coast). El dispositivo se instaló tan solo en una vivienda, como parte del estudio general de uso final del agua y con el objetivo de comprobar el sistema.

En esta vivienda se obtuvieron, por tanto, tres tipos de análisis para cada uso final del agua, según la procedencia de los datos registrados:

- Datos diarios obtenidos por los propios usuarios
- Datos obtenidos mediante *Trace Wizard*
- Datos del sensor acústico CSIRO

El dispositivo utiliza las señales de los sensores acústicos que coinciden con eventos de caudal del contador inteligente para asignar el uso final. Los sensores acústicos añaden dos dimensiones adicionales al análisis de uso final, a saber:

- Identificación espacial: permite segmentar la señal por la zona donde se ha producido la señal acústica (cocina, cuarto de baño...).

- Clasificación de uso: las características de la señal acústica se pueden relacionar con un tipo de dispositivo en particular; en la actualidad, esta evaluación se realiza manualmente, aunque es factible su automatización mediante un proceso de testeo más ambicioso.

Los resultados de la prueba de campo demostraron de manera concluyente que los sensores acústicos pueden determinar con precisión el uso final, cuando se utiliza junto con los datos de caudal obtenidos de contador inteligente. Los resultados fueron más precisos que los obtenidos utilizando el programa *Trace Wizard*. No obstante, se detectaron varios problemas: errores de sincronización de datos; la pérdida frecuente de las comunicaciones de la red Wi-Fi o el módem 3G; los sensores prototipo diseñados no tenían memoria interna por lo que se generaron lagunas de información. La próxima versión de estos dispositivos incorporará tarjeta de datos y se solucionará el problema del error de sincronización.

La prueba de campo demostró que los dispositivos acústicos podrían remplazar el uso del *Trace Wizard*. El reto ahora es automatizar el proceso de interpretación de datos.

[5] Roger O'halloran, Michael Best y Nigel Goodman, Urban Water Security Research Alliance. Technical Report No. 91. 2012

## 3.2. CARACTERIZACIÓN DE LOS USOS DOMÉSTICOS DEL AGUA

Diferentes investigaciones de Estados Unidos[6], España[7] y Brasil[8] y [9] proponen distintas metodologías para caracterizar el consumo de agua en hogares. En estos trabajos de investigación, se utilizaron medidores de agua por pulsos y *dataloggers* para la adquisición de datos y se desarrollaron distintas metodologías para su análisis.

Para permitir la identificación de señales o pulsos es necesario conocer previamente un conjunto de características de cada dispositivo (caudal, volumen, duración de uso, etc.). Esto permite que el programa pueda distinguir entre el uso de un grifo y de una cisterna, por ejemplo. Si estos parámetros no están bien ajustados, no es posible llevar a cabo una identificación correcta.

Cuando tres o más eventos ocurren de forma simultánea puede no ser factible diseccionar de forma precisa todos los usos finales. El procedimiento habitual pasa por convertir los pulsos en caudales (litro/segundo) para elaborar gráficos de consumo (caudal vs. tiempo). Estas señales (representación de señales de consumo de agua) están correlacionados con el tiempo empleado en la información de cada dispositivo (proporcionado por usuarios). Esta correlación permite identificar, de forma precisa, las señales de algunos de los usos, pero por lo general, debido a la baja precisión de la información proporcionada, algunos de los consumos tienen que ser estimados. En el caso incluido en el Cuaderno nº 4 de I+D+i[7], la información proporcionada por las señales de los pulsos se comparó con la caracterización previa de su amplitud y patrones temporales, lo que permitió identificar el uso en cada momento.

Los estudios de Fernandes[9] y Barreto[10] utilizaron un medidor en cada dispositivo hidráulico y en el suministro de tuberías de entrada para garantizar una buena precisión en la fase de caracterización. A pesar de la precisión de la información, esta metodología resulta difícil de aplicar en edificios de viviendas por el trabajo necesario en la instalación en la red de suministro. Este método no tiene viabilidad ni técnica, ni económica en la mayoría de las situaciones.

---

6 Mayer, P. Water. Energy savings from high efficiency fixtures and appliances in single family homes. USEPA, Combined Retrofit Report 1, 2005.

7 Cubillo, F., Moreno, T., Ortega, S.: Microcomponentes y factores explicativos del consumo doméstico de agua en la Comunidad de Madrid. Cuaderno de I+D+i nº 4. Canal de Isabel II, 2008.

8 Almeida, G.A., Kiperstok, A., Dias, M., Ludwig, O.*Metodologia para Caracterização de consumo de água doméstico por equipamento hidráulico*. Anais Do Silubesa/Abes. Figueira da Fo. 2006.

9 Fernandes, B.C.: Construção de um sistema eletrônico de monitoramento de consumo de água residencial. Projeto de graduação apresentado ao departamento de Engenharia Elétrica. P. 65 Centro Tecnológico da Univ. Federal Do Espírito Santo, 2007.

10 Barreto, D. *Perfil do consumo residencial e usos finais da água* Ambiente Construído, Porto Alegre 8(2), 23–40 (2008) ISSN 1678-8621; © 2008, Associação Nacional de Tecnologia do Ambiente Construído, April/June 2008.

## 3.3. ESTADO DEL ARTE EN EL RECONOCIMIENTO DE PATRONES Y CLASIFICACIÓN DE CONSUMOS DE AGUA

En las líneas que siguen se exponen las diferentes técnicas, referidas en la literatura, para el reconocimiento de patrones, en el consumo del agua en distintos dispositivos de los hogares, utilizando los datos de los medidores.

### 3.3.1. Técnica 1. Clasificador lineal robusto multicategoría

Esta metodología se inspira en la aplicación de la clasificación de usos finales del consumo doméstico eléctrico y de gas, aunque existe una diferencia fundamental entre ambos: en la clasificación de los usos finales del agua no se puede suponer que el caudal de consumo es constante y, además, los caudales máximo y mínimo pueden tener una amplia gama de valores posibles para un único dispositivo; tampoco existe un consumo base regular, tal y como existe en el uso de la electricidad[11].

Un *evento* está descrito por cuatro características físicas: volumen, duración, caudal máximo y caudal más frecuente. Además de esas características, el instrumento de análisis empleado utiliza la hora de inicio, la hora final y la frecuencia del caudal más frecuente durante la duración del *evento*. Toda esta información puede extraerse de la cuantificación de los datos que vienen del medidor inteligente.

#### *Planteamiento del problema*

Los datos utilizados para desarrollar este modelo vienen de la utilización de medidores inteligentes, en 74 hogares, durante dos semanas. Los datos de flujos se almacenaban cada segundo y con una resolución de 100 y 250 pulsos por litro, dependiendo del tipo de medidor. Los datos se remuestrearon en una resolución de 1 litro. Los datos con usos finales identificados contienen la siguiente información en cada *evento* identificado: accionamiento de la instalación, hora de inicio, duración, volumen, caudal medio y caudal máximo.

El nivel de cuantificación, es decir, la resolución tomada por los investigadores (1 litro), está adaptada a los medidores que están actualmente disponibles en el mercado. El tiempo de muestreo (1 segundo) está también adaptado al nivel de cuantificación y está significativamente por delante del tiempo de muestreo actual usado en el mercado. El caudal registrado en los hogares no es muy elevado (por ejemplo, el caudal máximo de todos los usos identificados que se ha registrado es de 0,64 litros/segundo) lo que implica que:

- En 1 segundo sólo podrá registrarse, como mucho, un pulso (pues la resolución tomada es de 1 pulso por litro).
- El tiempo mínimo entre dos registros consecutivos, no nulos, es de 2 segundos.

[11] Vasak M., Banjac G., Novak H.: *Water use disaggregation based on classification of feature vectors extracted from smart meter data, Procedia Engineering, 119, 1381 – 1390, 3th Computer Control for Water Industry Conference, CCWI 2015.*

El principal problema de la cuantificación es la gran pérdida de información: los datos procesados tienen solamente una dimensión disponible, es decir, instante en el que ocurre un pulso. Esto implica que todas las características asociadas a un uso del agua sólo se pueden extraer de la información contenida en los tiempos de los registros correspondientes. Sin embargo, llevar los registros de un medidor inteligente a un uso de agua no es una tarea sencilla. Por ejemplo, cuando el tiempo transcurrido entre dos pulsos consecutivos es de 60 segundos, no es posible asegurar que el segundo pulso es resultado de un caudal elevado durante los últimos segundos o un caudal bajo continuo durante los 60 segundos. Por otro lado, puede ser resultado de dos usos consecutivos (posiblemente diferentes) donde el volumen del primero es inferior a 1 litro.

Por eso es necesaria la hipótesis de que sólo un elemento es usado a un tiempo –considerar eventos simultáneos complicaría mucho el problema con una resolución tan pequeña.

Para convertir los registros en eventos, los investigadores definen un parámetro de tiempo máximo basado en la decisión de si dos pulsos consecutivos corresponden al mismo uso. Si el tiempo transcurrido entre dos pulsos consecutivos es menor que el tiempo máximo, se les asigna el mismo grupo. Sólo se considerarán aquellos grupos que contengan dos o más registros. Si el parámetro tiempo máximo es demasiado pequeño, puede ser que se divida un uso en dos o más usos. Por otro lado, si el parámetro es muy grande, dos o más usos de agua pueden unirse en un único uso. Si los usos unidos son distintos, se tendrá que elegir uno de los usos. Elegir estos parámetros no es sencillo. Los investigadores han fijado el parámetro tiempo máximo entre pulsos en 150 segundos. Los eventos largos (ciclo del lavavajillas, ciclo de la lavadora y ducha) están formados por ciclos intermitentes que se caracterizan como eventos separados si el medidor de agua está inactivo por más de 150 segundos, de modo que pueden identificarse los programas de los dispositivos. A los usos simultáneos se les asigna el dispositivo de uso dominante.

Las horas inicial y final estimadas usando los registros procesados no coinciden con las estimadas usando los registros de alta resolución. Además, el volumen total estimado usando los registros procesados sólo puede ser múltiplo de 1 litro.

### *Datos de interés*

En los datos del estudio, el 29% de los eventos presentaban solapamiento en el tiempo, el 20% del tiempo del solapamiento era de distintos dispositivos y los solapamientos suman un 13% del consumo total del agua.

Los *eventos* considerados en este estudio son 12: fregadero, lavadora, cisterna 1, cisterna 2, cisterna 3, lavavajillas, ducha, ducha a presión, grifos, bañera, descalcificador de agua y grifo exterior. Grifos, lavadora y cisterna (cisterna 1, cisterna 2 y cisterna 3) suman más del 92% del total de usos, el 69% de la duración y el 62% del volumen total consumido. Por otro lado, aunque se usan muy poco (2.20% del número total de usos), la ducha suma alrededor del 19% de la duración total y consume mucha agua en los hogares, el 23,06% del volumen total. Considerando sólo las etiquetas sin unir dispositivos similares, se obtiene que los grifos, la lavadora, la cisterna 1, cisterna 2 y la ducha suman el 93,09% del total de usos identificados, con el 84,6% de la duración y el 83,1% del volumen consumido.

### *Vector de características de los usos finales del agua*

Como se ha dicho anteriormente, un *evento* puede ser descrito por unas características físicas: volumen, duración, caudal máximo y caudal más frecuente, además de otras características que pueden ser aportadas: instante inicial, instante final y frecuencia del caudal más frecuente. Se plantean también dos funciones armónicas, $sin$ y $cos$, que pueden usarse para representar ese momento del día.

Estas funciones son continuas, periódicas y sus valores determinan, unívocamente, el momento del día. Así, todo uso final estará representado por el vector de características:

$$x = \left[v, d, q_{max}, q_{freq}, N_{freq}, \sin\frac{2\pi t}{T}, \cos\frac{2\pi t}{T}\right] \in \boldsymbol{R}^7$$

Donde $v$ es el volumen, $d$ es la duración, $q_{max}$ es el caudal máximo, $q_{freq}$es el caudal más frecuente, $N_{freq}$es la frecuencia del caudal más frecuente durante el evento,$t$ es el número de minutos transcurridos desde medianoche y $T = 24 \cdot 60$ es el total de minutos en un día.

Cada vector de características $x$ está asociado con la caracterización del uso final y se le puede aplicar la técnica de clasificación lineal robusta en varias categorías (M-RLP).

### *Método de clasificación*

Este trabajo utiliza un clasificador lineal, que supone que dos tipos de datos pueden separarse por una frontera lineal. Este tipo de técnicas no son suficientemente descriptivas y en realidad puede existir una frontera no lineal que separe mejor los datos. Para hacer trabajar a los modelos lineales en conjuntos no lineales, el vector de características se llevó a dimensiones superiores donde sea posible aplicar técnicas de clasificación no lineal. Así, el vector original de características $x = [x_1, \dots, x_7]$ irá a un espacio de dimensión superior utilizando la siguiente transformación:

$$\phi: x \longrightarrow [x_1, \dots, x_7, x_1^2, \dots, x_7^2, x_1x_2, \dots, x_1x_7, \dots, x_6x_1, \dots, x_6x_7] \in R^{35},$$

Los cuadrados de cada característica, y los productos entre las características se han añadido al vector original. Por otro lado, la complejidad de un modelo de este tipo aumenta mucho en términos de número de parámetros necesarios en el modelo.

### *RESULTADOS*

Los usos finales correspondientes a los distintos dispositivos están etiquetados, tal y como se ha visto. Existen un total de 12 etiquetas. Los resultados obtenidos han sido poco satisfactorios: sólo 5 de 12 de los dispositivos tienen una precisión distinta de 0. La caracterización de grifo, ducha a presión y descalcificador de agua han tenido una precisión por encima del 50%, pero el uso agregado de la ducha a presión y del descalcificador suman menos del 0.5% del total de usos, lo que hace que el clasificador no sea práctico.

Los resultados de la clasificación del modelo común con 7 características no son muy satisfactorios. Esto se debe a que el problema no es linealmente separable con esas características y se emplea un clasificador lineal. El *fregadero* representa dos tercios de los usos totales del agua y la precisión de su clasificación está por debajo del 5%. La cisterna tiene una precisión de 0%. Cuando el espacio de características se extiende con características adicionales la precisión para algunos usos mejora, aunque la clasificación para la cisterna 2 empeora muchísimo.

Este estudio muestra que al generar un modelo diferente por hogar la precisión de la clasificación mejora significativamente cuando la comparamos con la obtenida en con el modelo de clasificación común.

### 3.3.2. Técnica 2. Sistema de inferencia neuro difusa adaptativa (*Anfis*)

En los trabajos realizados en 2008 por Corona*et al.*[12] se proponía la identificación automática de los usos del agua en un hogar mediante un modelo*Anfis* y un *clustering* difuso (*fuzzy clustering*).

#### *Planteamiento del problema*

Los datos utilizados para la elaboración de este modelo venían de una casa en la que habitaban tres personas y sobre cuyos usos domésticos del agua se elaboró un conjunto de datos con 1000 ejemplos, 100 para cada una de las cinco clases y 100 para cada una de las cinco subclases. Las clases son: W.C., grifo, ducha, lavavajillas y lavadora. La ducha se divide en tres subclases debido a la diferencia en los hábitos de ducha de las tres personas y la lavadora en otras dos como resultado de la intervención del usuario al principio del ciclo de lavado.

Estos investigadores consideraron que el problema de la clasificación de los usos domésticos del agua variaba mucho de una casa a otra, pues de ello depende el número de*outputs*, instalaciones hidráulicas, tipos de dispositivos utilizados, hábitos de consumo, etc. Utilizaron entonces un modelo para el clasificador buscando una fácil interpretación por humanos y la posibilidad de incluir conocimiento proporcionado por usuarios o expertos: modelo neuro difuso, en particular el modelo*Anfis.*

#### *Modelos neuro difusos*

Los modelos neuro difusos se caracterizan porque utilizan lo mejor de las redes neuronales y lo mejor de los modelos de lógica difusa: por un lado, proporcionan la capacidad de aprendizaje y generalización de las redes neuronales y por otro lado, el razonamiento lógico basado en reglas de inferencia.

[12] Corona-Nakamura M. A., Ruelas R., Ojeda-Magaña B., Andina D.: Classification of Domestic Water Consumption Using an Anfis Model; Conference paper, Automation Congress, 2008.

### *Sistema de inferencia neuro difusa adaptativa (Antis)*

La arquitectura neuro difusa utilizada para la clasificación de los usos finales del agua es red neuronal difusa adaptativa, llamada *Anfis*. Esta arquitectura es equivalente a un sistema de inferencia difusa que puede construirse a partir de las relaciones entre los valores de entrada y de salida de un dataset. En este sistema de inferencia, *Anfis* ajusta las funciones utilizadas durante el proceso de entrenamiento del modelo. Las funciones y las reglas difusas que deben adaptarse al problema han de ser definidas antes de entrenar el modelo *Anfis*. Para la estimación inicial de esos parámetros es posible usar un algoritmo de *clustering*, como el *Fuzzy c-means* (FCM), el *Mountain Method*, el método substractivo o, simplemente, con el conocimiento del experto. Las reglas difusas están basadas en el método de inferencia de Takagi-Sugeno y las conclusiones son funciones polinómicas.

### *Método de Clustering subtractivo (SCM)*

Este método suele utilizarse con frecuencia cuando se conoce el número de clúster, pero no sus centros. Su objetivo es precisamente estimar esos centros. Se trata de un método rápido y está basado en una idea similar a la del *Mountain Method,* pues ambos dividen el espacio característico de los datos en una cuadrícula en la que las intersecciones son un conjunto de candidatos que pertenecen a un clúster dado. El cálculo del centro del clúster se basa en la densidad del conjunto. A pesar de que el*Mountain Method* es simple y efectivo, su coste computacional crece exponencialmente con la dimensión del problema, como resultado de la evaluación de la función de densidad en todos los puntos de la cuadrícula.

### *Clasificación*

Debido a las grandes diferencias en consumo en distintas localizaciones y actividades, el espacio característico presenta grandes regiones sin datos. Es por ello que se han seleccionado funciones no acotadas, de modo que cubran todo el espacio y el clasificador tiene la posibilidad de reconocer puntos más allá de la frontera de los datos disponibles. La función seleccionada para representar los conjuntos difusos del modelo es una función Gaussiana.

Una vez que el clasificador está entrenado habrá que cambiar algunas de las funciones del modelo*Anfis* para obtener un mejor reconocimiento de los nuevos datos, sobre todo cuando están entre clases. En este caso se sustituye una Gaussiana por una Gaussiana acotada (Gaussiana2), una campana o una función triangular.

El número de reglas difusas a partir de los resultados de clasificación o*clustering* coincide, en general, con el número de clases o clúster.

### *RESULTADOS*

Los resultados obtenidos con esta metodología han sido positivos: en el peor de los casos el porcentaje de aciertos superaba el 91%, pero al haberse entrenado con datos de un único hogar, no es posible garantizar estos resultados sobre una muestra más amplia y variada.

### 3.3.3. Técnica 3. Modelo híbrido de filtrado, red neuronal artificial y modelo oculto de *Markov*

En los trabajos realizados por Nguyen, Zhang y Stewart en el año 2012, se sugiere un modelo híbrido de filtrado y reconocimiento de patrones para la categorización de usos domésticos del agua[13].

Esta técnica es la más sofisticada e interesante de las recogidas.

#### *Planteamiento del problema*

La base de datos utilizada para el estudio realizado por estos autores era muy amplia, con medidores de alta resolución, almacenando 0,014 litros/pulso en intervalos de 5 segundos, en más de 500 hogares durante 3 años. De estos datos, se extrajeron manualmente los eventos, que se clasificaron utilizando el software *Trace Wizard*. Las salidas posibles de este proceso de identificación eran siete: ducha (*shower*), grifo (*faucet*), lavavajillas (*dishwasher*), lavadora (*clothes washer*), cisterna (*toilet*), bañera (*bathtub*) y riego (*irrigation*).

Se utilizó una muestra de aproximadamente 100.000 eventos clasificados: unos 83.000 para entrenamiento y 16.000 de verificación o *test.*

Las técnicas de reconocimiento de patrones utilizadas en este estudio son: modelo oculto de *Markov* (HMM), red neuronal artificial (ANN) y un algoritmo de deformación dinámica (DTW). Una combinación híbrida de estas técnicas es la que finalmente eligieron como la más adecuada y precisa para el sistema de reconocimiento de patrones.

#### *Modelo Oculto de Markov (HMM)*

En este estudio se utilizó este modelo como uno de los clasificadores para el uso final del agua, basándose en la forma del evento. Sin embargo, la debilidad de esta técnica reside en que no clasifica adecuadamente aquellas categorías que son altamente dependientes del comportamiento del usuario. Tales usos son altamente variables, de modo que, a veces, presentan características que se parecen mucho a las características de otras categorías.

La ducha, la bañera y el riego pueden presentar patrones similares pese a ser categorías distintas. Por eso, es necesario incluir una técnica adicional que pueda inspeccionar las características físicas de esos eventos: la red neuronal artificial.

#### *Red Neuronal Artificial (ANN)*

Los autores y desarrolladores de este método han utilizado una red de compensación con un algoritmo de entrenamiento de propagación hacia atrás como técnica principal para aprender el patrón típico de cada categoría en términos de características físicas (por ejemplo, volumen, duración, caudal máximo, etc.).

[13] Nguyen, Zhang y Stewart. Analysis of simultaneous water end use events using a hybrid combination of filtering and pattern recognition techniques. International Congress on Environmental Modelling and Software (2012).

***Algoritmo de deformación dinámica (DTW)***

Se trata de un popular método para medir la similitud entre dos series temporales de distinta longitud. En general, esta tarea se realiza encontrando un alineado óptimo entre dos series temporales con determinadas restricciones. Las series se extienden o se acortan en el tiempo para determinar su similitud, independientemente de ciertas variaciones no lineales. El objetivo es encontrar un mapeo con la mínima distancia. Esta técnica se ha utilizado con frecuencia en el reconocimiento de patrones o en la búsqueda de palabras mediante imágenes. En la herramienta desarrollada por los investigadores era la clave para encontrar ciclos de agua enlazados relacionados con un evento particular para mecanizar eventos (lavadora y lavavajillas principalmente) que no se clasificaban correctamente con el HMM ni con la ANN. La lavadora y el lavavajillas tienen unos ciclos de uso del agua asociados con una selección de programas hecha por el consumidor, que puede ser reconocida con el DTW.

***RESULTADOS***

Para las categorías que presentan patrones claramente definidos como el lavavajillas y la lavadora la precisión fue más alta (en torno al 90%), mientras que las que están más influenciadas por el comportamiento humano tenían tasas de acierto en torno al 60%-80%.

### 3.3.4. Técnica 4. Otras técnicas

En 2011 se llevaron a cabo pruebas experimentales en tres ciudades alemanas, entre ellas Berlín[14], con condiciones controladas para explorar distintas posibilidades para preparar distintas instalaciones y para testear algoritmos.

Los pasos para las pruebas experimentales fueron:

- construir una instalación experimental
- calibrar el sistema hidráulico
- adquirir las señales
- almacenar las señales
- procesar los datos y
- desarrollar algoritmos para el reconocimiento de patrones en el consumo de agua.

Herramientas de procesamiento de señal digital como los algoritmos de convolución se utilizaron en el procesamiento de datos, intentando resolver el solapamiento de señales que venían del uso simultáneo de distintos dispositivos en el sistema hidráulico. Después, implementaron y testearon técnicas para la extracción de características y clasificadores para la identificación de las señales de cada clase de caudal en su respectivo dispositivo. Se propuso entonces la extracción de características y un clasificador de reconocimiento de patrones para desarrollar el algoritmo dependiendo del dispositivo estudiado. Estas características pueden extraerse en el ámbito del tiempo o de frecuencia.

[14] Almeida G., Vieira J., Marques J., Cardoso A. Pattern Recognition of the Household Water Consumption through Signal Analysis. In: Camarinha-Matos L.M. (eds) Technological Innovation for Sustainability. DoCEIS 2011. IFIP Advances in Information and Communication Technology, vol 349. Springer, Berlin, Heidelberg. 2011.

Se espera que las características temporales, o frecuenciales, de la respuesta transitoria del sistema hidráulico a la activación de dispositivos hidráulicos varíe, según las características y posición en el sistema de suministro hidráulico. Esto justifica la necesidad de utilizar distintos dispositivos de medición, aunque después sólo se utilice la señal del caudal en el suministro de tuberías como*input* del clasificador. El algoritmo compara la señal con el prototipo generado para cada dispositivo hasta que el proceso de identificación esté completo.

# 4. Planteamiento metodológico

Los trabajos desarrollados se han agrupado en los siguientes módulos:

- Módulo de Transformación de pulsos en caudales
- Módulo de Identificación de eventos
- Módulo de Clasificación de eventos
- Módulo de Consultas y generación de gráficos

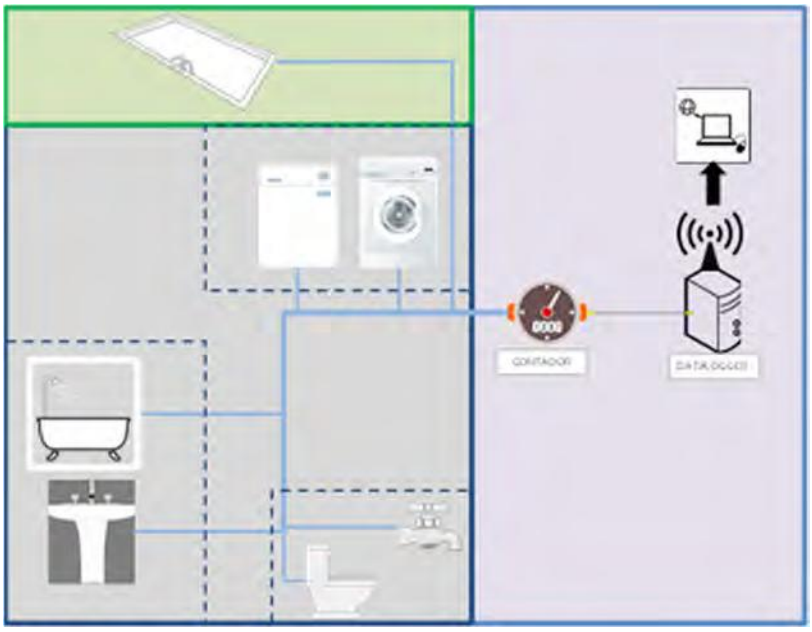
Todos estos módulos han sido integrados en una aplicación informática desarrollada en*Visual Basic for Applications (VBA)*, bajo entorno *Access (©Microsoft)*, que permite el tratamiento masivo y automatizado de las lecturas de los contadores (pulsos), a partir de las cuales y mediante la metodología que se describe en los apartados siguientes se determinan los usos finales de los diferentes consumos domésticos.

## 4.1. TRANSFORMACIÓN DE PULSOS EN CAUDALES

### 4.1.1. Información de partida

Los datos de partida disponibles, en lo que se refiere a datos de consumo, proceden de registros de contadores volumétricos con emisor de pulsos (ver Figura 11). Este tipo de contador emite un pulso cada vez que se consume un determinado volumen, dado por la precisión del contador (1 ó 0,1 litros, según el tipo de dispositivo), registrándose de manera automática el instante en que se produce, con una precisión de un segundo. En los contadores de mayor precisión (0,1 litros), con relativa frecuencia se producen dos y hasta tres pulsos en un segundo.

**FIGURA 11. ESQUEMA DE REGISTRO DE LECTURAS DE CONTADOR CON EMISOR DE PULSOS**



*Fuente: Elaboración propia a partir de imágenes publicadas por pixabay.com, libres de derechos de autor bajo la licencia Creative Commons CC0*

De esta manera cada contador genera una serie temporal de pulsos acumulados.

Toda esta información de lecturas de contadores se encuentra almacena en bases de datos tipo*Access*, agrupadas en diferentes carpetas según las fechas de los registros.

### 4.1.2. Algoritmo de cálculo. Medias móviles

La transformación de las lecturas de los contadores (instante-pulsos acumulados) en series temporales (instante-caudal) resulta bastante más complicado de lo que en un principio pudiera parecer, y más si se pretende automatizar mediante un algoritmo matemático. En efecto, un pulso aislado correspondiente a un consumo, por ejemplo, de 1 litro, puede obedecer a infinitas combinaciones de caudales ocurridos desde el registro del anterior pulso, y cada una de ellas daría lugar a una serie instante-caudal diferente según el criterio adoptado. Ese volumen consumido ¿responde a un caudal constante desde que se produjera el anterior pulso? ¿se han producido varios usos que han dado lugar a un caudal intermitente, que finalmente totalizan 1 litro? ¿cuánto tiempo han durado esos consumos? Cuando el tiempo transcurrido entre varios pulsos es reducido, del orden de unos pocos segundos, suponer que aquel litro se distribuye de una forma constante entre los instantes en que se producen dos pulsos consecutivos puede ser una simplificación hasta cierto punto asumible, pero cuanto mayor es ese tiempo transcurrido dicha suposición resulta cada vez menos plausible; no es creíble que durante 4 horas se haya mantenido un consumo constante de 0,25 l/h, salvo que se trate de una fuga –éste es precisamente un patrón que permite identificarlas–. En definitiva, el procedimiento que se elija para esta transformación de pulsos en caudales puede dar lugar a resultados que se alejen demasiado de la realidad que se quiere reproducir.

Para evaluar la bondad del método elegido, la serie de caudales calculada se ha de comparar con la serie de caudales original que ha dado lugar a la serie de pulsos utilizada en el procedimiento. Después de tantear diferentes procedimientos, finalmente se ha optado por un algoritmo matemático basado en el cálculo de medias móviles, según se detalla a continuación:

1º) A partir de la serie de lecturas de pulsos acumulados, se construye la serie de pulsos no acumulados, por simple diferencia de lecturas consecutivas en el contador ($\Delta L$).

2º) Construcción de una serie temporal regular: Puesto que el consumo doméstico de agua no es regular en el tiempo, el intervalo de tiempo entre dos registros consecutivos tampoco lo es, resultando una serie con intervalos de tiempo irregulares. Se trata, por tanto, de confeccionar una serie temporal regular, con incrementos temporales de un segundo.

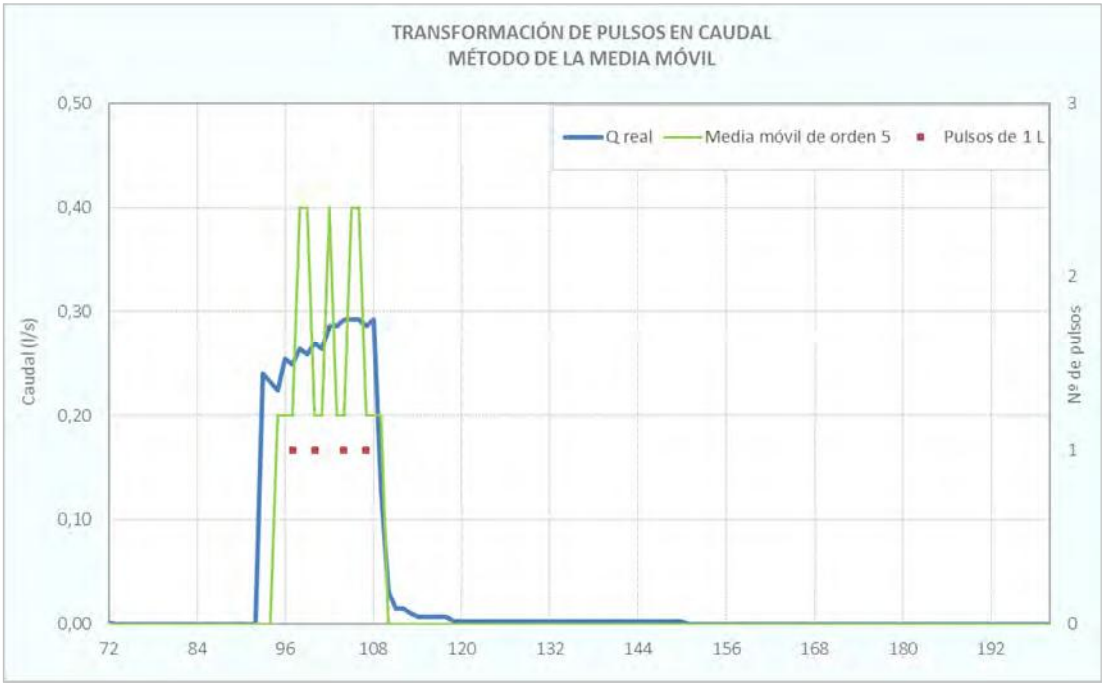
3º) Se supone, como primera aproximación, que en el momento de la emisión del pulso se consume el volumen dado por $\Delta L \cdot P$, siendo P la precisión del contador (equivalente al volumen de cada pulso).

4º) Se asigna a cada segundo un caudal igual a la media móvil de los caudales anteriormente calculados.

5º) Para afinar el resultado, se vuelve a calcular una segunda media móvil de las anteriores medias. El orden de estas medias móviles es un parámetro a ajustar, como se detallará más adelante.

La Figura 12 ilustra el resultado obtenido siguiendo este proceso con los datos de un contador con emisor de pulsos de 1 litro.

**FIGURA 12.** TRANSFORMACIÓN DE PULSOS EN CAUDAL. PRIMERA MEDIA MÓVIL



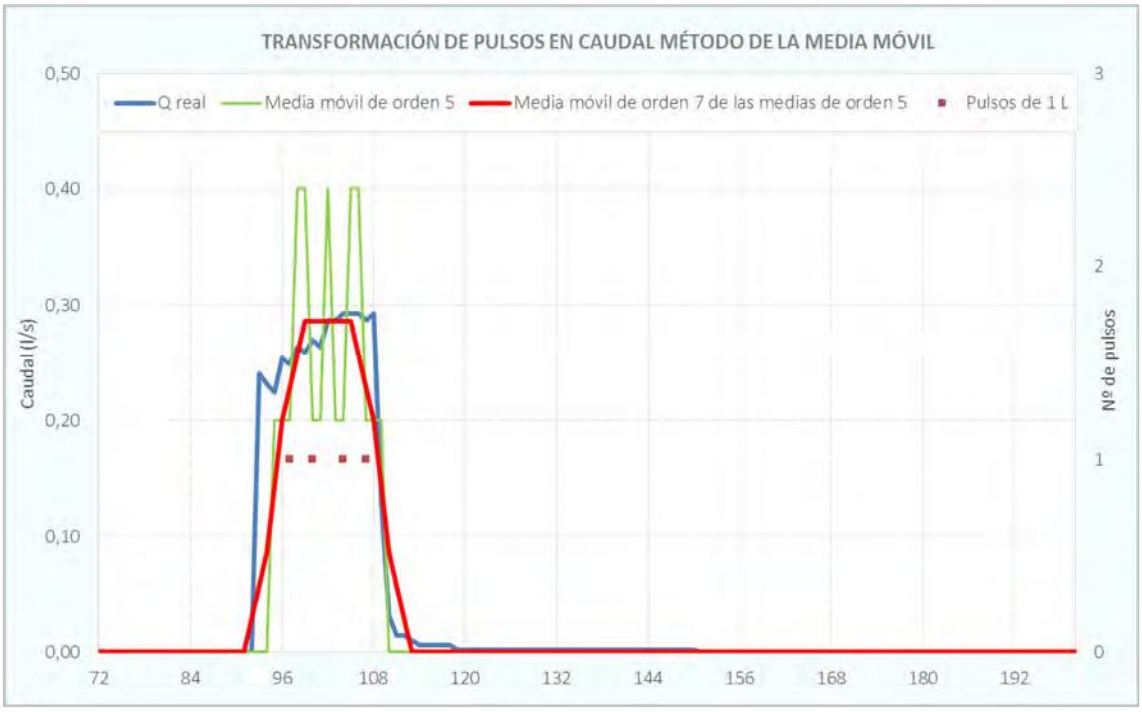

La curva de color **azul** corresponde al caudal real que habrían producido los pulsos registrados, señalados con ■ , mientras que la línea **verde** representa el caudal calculado, según el método propuesto (que a modo ilustrativo se ha representado para media móvil de orden 5).

Si se repite el proceso calculando una nueva media móvil de la serie calculada, el resultado se ajusta mucho mejor, como puede verse en el ejemplo que se incluye, aplicando esta vez las medias móviles de orden 7 de la serie de medias móviles de orden 5 calculadas previamente.

En la Figura 13 se representan estos valores calculados, y se observa un ajuste mucho mejor entre la curva de caudales calculados (en **rojo**) y la curva de caudales reales (**azul**) que la anteriormente calculada (**verde**).

Se ha comprobado que la adopción de un valor u otro para el orden de las medias móviles tiene una gran influencia en la calidad de ajuste de los resultados obtenidos, por lo que se trata de un parámetro que debe ser ajustado. En el apartado siguiente se detalla el procedimiento a seguir para ello.

**FIGURA 13.** TRANSFORMACIÓN DE PULSOS EN CAUDAL. SEGÚN MEDIA MÓVIL




### 4.1.3. Ajuste de los parámetros de cálculo

El ajuste del método utilizado se centra en la determinación del orden de las medias móviles a aplicar para que el ajuste de la curva de caudales calculados a la curva de caudales reales sea óptimo.

Para evaluar la bondad de los resultados obtenidos con esta metodología, se analiza el coeficiente de correlación y el error típico entre una curva de caudales generada artificialmente*a priori*, que será considerada como curva de caudales "reales", y la curva de caudales calculados aplicando el método de las medias móviles descrito.

El procedimiento de ajuste propuesto es el siguiente:

1º) Se parte de una serie temporal de datos de caudales instantáneos conocida, procedente de un usuario determinado, ya sea real o ficticio.

2º) Se calcula la serie temporal de pulsos que generaría dicha curva de caudales instantáneos conocida, en caso de que existiese un contador de pulsos de 1 litro de precisión.

3º) A partir de esta serie de pulsos, se aplica el método para diferentes supuestos de pares de valores correspondientes a los órdenes de la primera y de la segunda media móvil, con objeto de seleccionar los valores que proporcionan mejores resultados, según el coeficiente de correlación y error típico.

4º) Para cada par de valores, se identifican los diferentes "episodios" de consumo de agua, entendiendo por episodio aquel periodo de tiempo, con caudal distinto de cero, comprendido entre un instante con caudal nulo y el siguiente[15].

5º) Se comparan las series de caudales reales y calculados, episodio a episodio, determinando el coeficiente de correlación y el error típico.

El coeficiente de correlación se define como la relación entre la covarianza de las dos series de datos, (X e Y) y el producto de las respectivas desviaciones estándar:

Coeficiente de correlación:

$$\sigma_{xy} = \frac{Cov\ (X,Y)}{\sigma_x \cdot \sigma_y}$$

Siendo la covarianza:

$$Cov\ (X,Y) = \frac{\sum(x-\bar{x})\cdot(y-\bar{y})}{n-1}$$

Y las desviaciones típicas de cada serie:

$$\sigma_x = \sqrt{\frac{\sum(x-\bar{x})}{n-1}}$$

$$\sigma_y = \sqrt{\frac{\sum(y-\bar{y})}{n-1}}$$

Siendo $\bar{X}$, e $\bar{y}$ las medias respectivas de las dos series de datos.

El error típico es una medida de la cuantía de error en el pronóstico del valor de ***y*** para un valor individual de ***x*** y viene dado por la siguiente expresión:

$$\sqrt{\frac{1}{n-2}\left[\sum(y-\bar{y})^2 - \frac{[\sum(x-\bar{x})(y-\bar{y})]^2}{\sum(x-\bar{x})^2}\right]}$$

[15] No confundir el término "episodio" con el de "evento" tratado en el Módulo 2, y que se refiere a un periodo de tiempo en el que se produce un uso doméstico determinado. Un episodio puede resultar de la combinación de diferentes eventos solapados o comprender un único evento aislado. Y viceversa, un evento puede traducirse en diferentes episodios de caudal, como es el caso de una lavadora con varios ciclos de llenado y vaciado: cada llenado da lugar a un episodio diferente.

6º) Por último, una vez analizados los resultados que se obtengan para los diferentes pares de valores, se seleccionará aquel par que proporcione la combinación óptima en relación al coeficiente de correlación mayor y al error típico menor.

Comoquiera que se dispone de dos tipos de contadores con diferentes precisiones (1 y 0,1 litros) y esta precisión interviene directamente en el método utilizado, el orden de las medias móviles, óptimo a aplicar, en uno, u otro caso, también es diferente. La determinación de los valores óptimos de este parámetro se ha establecido comprobando los resultados con diferentes pares de valores sobre series de datos reales y sintéticas.

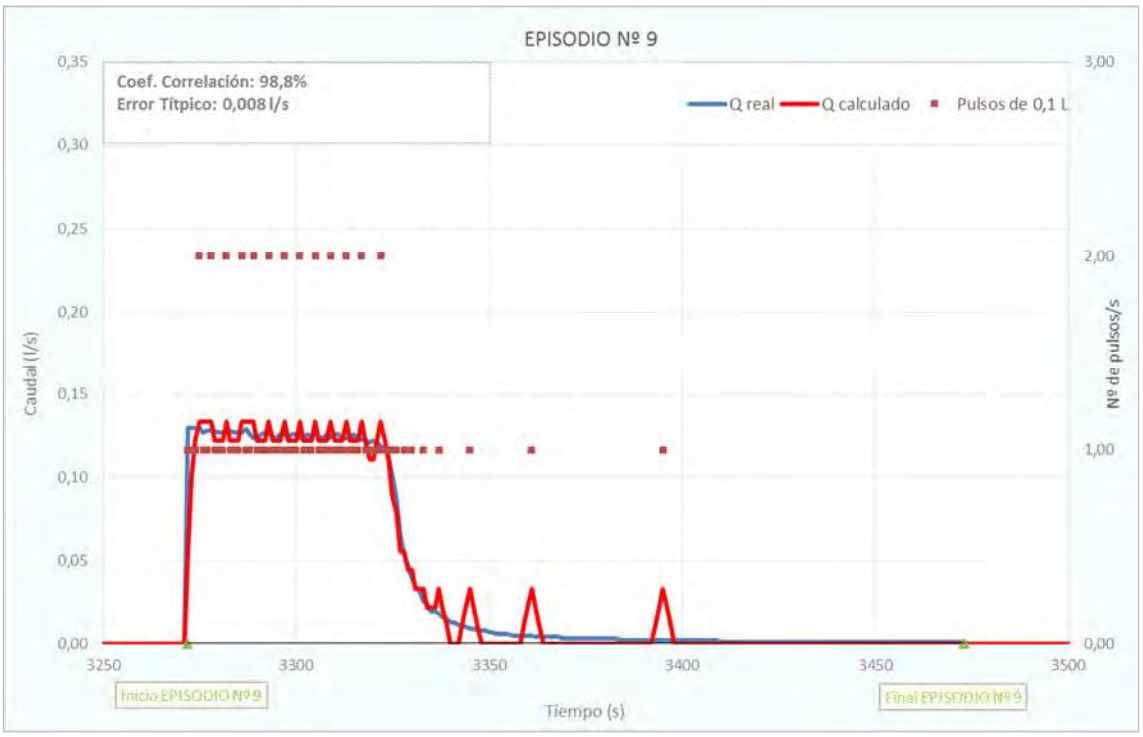
Finalmente, se han adoptado los órdenes de medias móviles siguientes:

- 9-9 para contadores de 1 litro de precisión
- 3-3 para contadores de 0,1 litro de precisión

Las figuras 14 y 15 siguientes ilustran algunos resultados de este ajuste.

**FIGURA 14. AJUSTE DE LOS ÓRDENES DE MEDIAS MÓVILES PARA CONTADORES DE 1 LITRO DE PRECISIÓN. RESULTADOS OBTENIDOS CON MEDIAS MÓVILES DE ORDEN 9 Y 9, PARA UNA SERIE SINTÉTICA DE CAUDALES**

**FIGURA 15. AJUSTE DE LOS ÓRDENES DE MEDIAS MÓVILES PARA CONTADORES DE 0,1 LITROS DE PRECISIÓN. RESULTADOS OBTENIDOS CON MEDIAS MÓVILES DE ORDEN 3 Y 3, PARA UNA SERIE REAL DE CAUDALES**

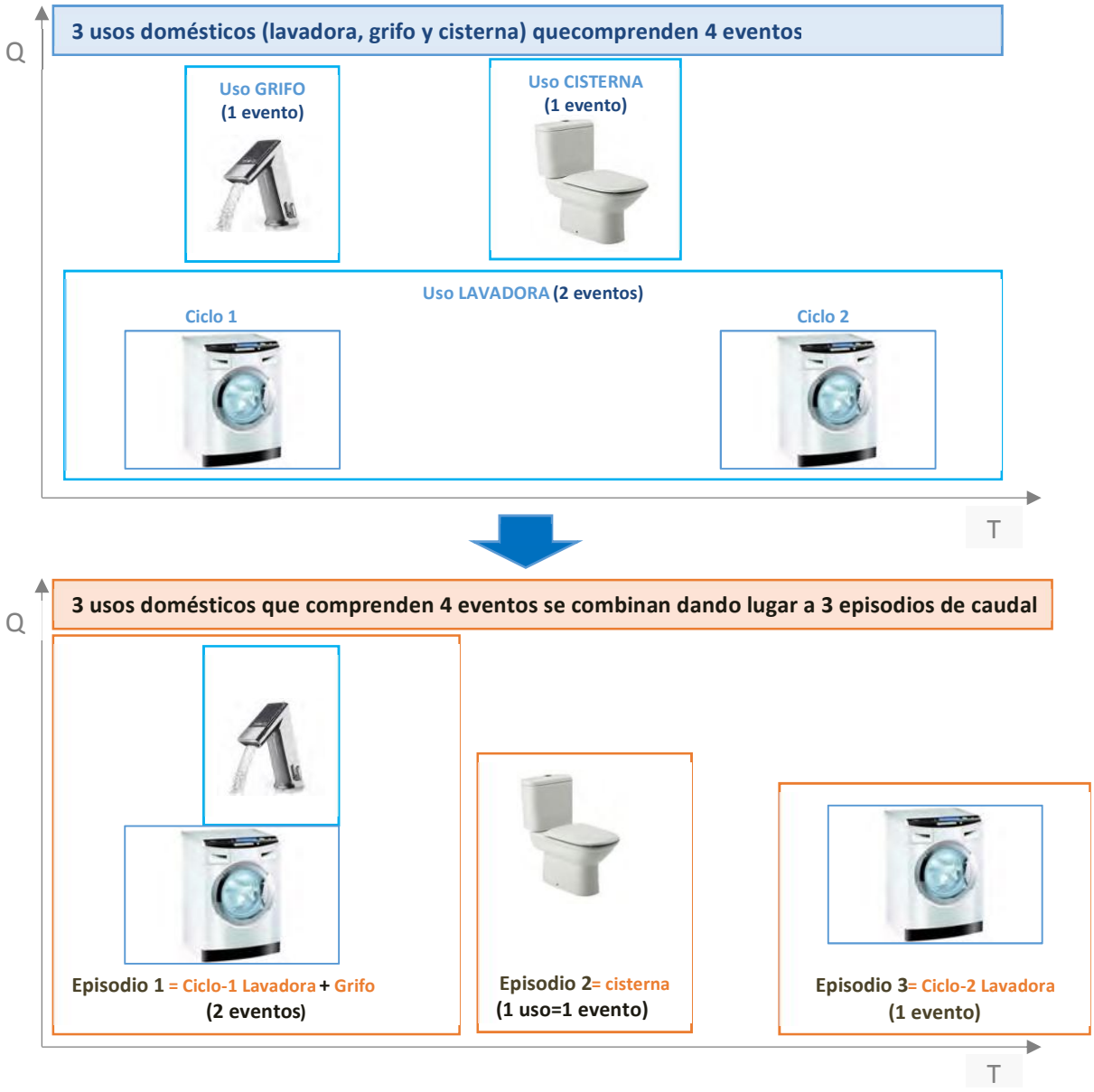



## 4.2. IDENTIFICACIÓN DE EVENTOS

Los episodios de caudales obtenidos previamente obedecen a diferentes combinaciones de usos domésticos cotidianos, como pueden ser la apertura de un grifo, la puesta en marcha de una lavadora o el uso de la cisterna, cuyos patrones de consumo podrían ser reconocidos más fácilmente si se presentaran de forma individual.

Por otro lado, cada uso concreto puede traducirse en una o varias unidades elementales de consumo o eventos, definidos como aquellos periodos de tiempo de duración suficiente en los que el caudal instantáneo se mantiene claramente diferenciable del resto. Así pues, por ejemplo, el uso de un electrodoméstico como una lavadora, con un programa de varios ciclos de lavado y aclarado, daría lugar a diferentes eventos diferenciables y separados unos de otros. En un momento dado, alguno de estos ciclos podría coincidir en el tiempo con el uso, por ejemplo, de grifo o una cisterna, dando lugar a un episodio complejo (Figura 16).

**FIGURA 16.** EVENTOS Y EPISODIOS DE CAUDAL GENERADOS POR DIFERENTES USOS DOMÉSTICOS

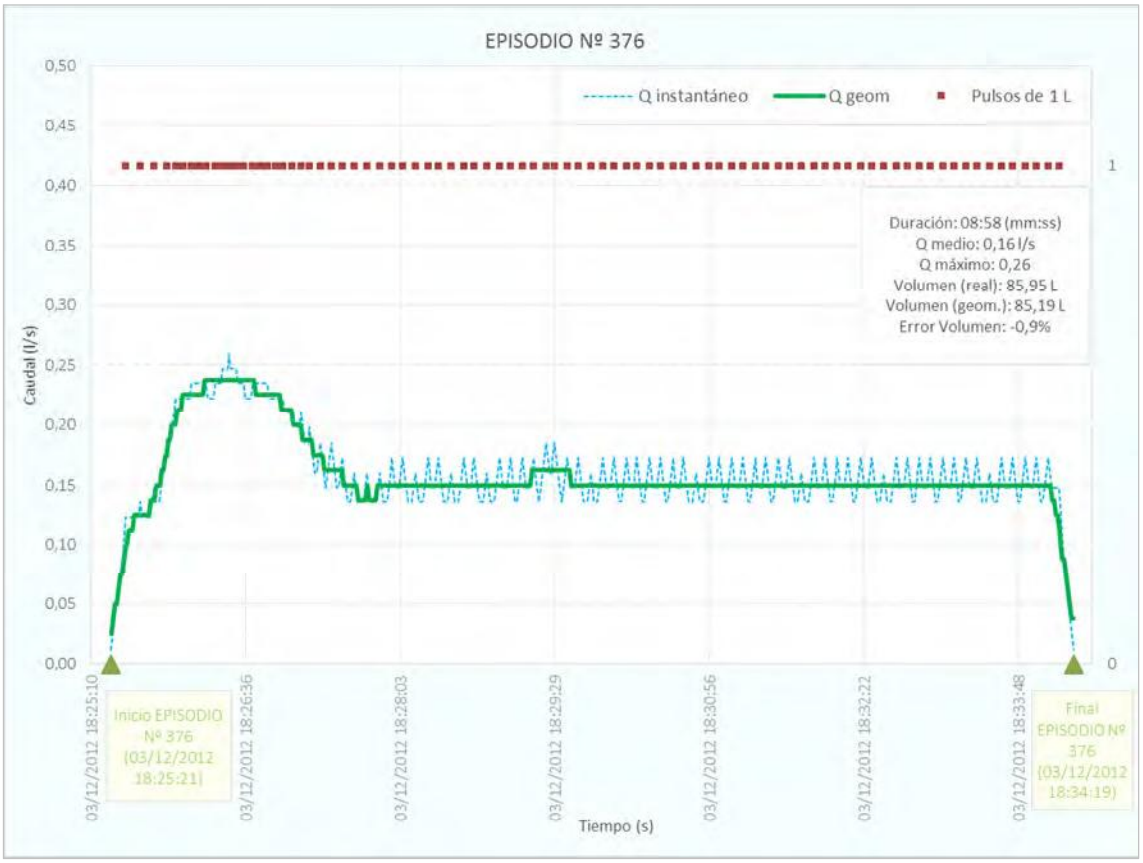



### 4.2.1. Geometrización de episodios

Los episodios obtenidos con la metodología descrita anteriormente presentan fluctuaciones de caudal, más o menos frecuentes, aunque poco significativas ("ruido"), que obedecen más al método aplicado que a un consumo real con constantes cierres y aperturas. Estas fluctuaciones hacen prácticamente inviable una caracterización automática en función de parámetros fácilmente cuantificables.

La metodología propuesta simplifica estos episodios de caudales reduciéndolos a formas geométricas compuestas por elementos más sencillos (*eventos*) asimilables a figuras más o menos rectangulares o trapezoidales. Este proceso de simplificación se ha denominado ***geometrización*** y su resultado puede visualizarse en la Figura 17, en la que se representa a modo de ejemplo el resultado obtenido de un episodio concreto (episodio nº 376 de diciembre de 2012, del contador 006).

**FIGURA 17. IDENTIFICACIÓN DE EVENTOS. PROCESO DE GEOMETRIZACIÓN**




Las pequeñas fluctuaciones de caudal (línea azul de la Figura 17), ruido producto del algoritmo de transformación de pulsos en caudales, se eliminan con la *geometrización* propuesta (línea verde en la misma figura), permitiendo un análisis más sencillo y automatizable del episodio.

El algoritmo matemático que proporciona esta *geometrización* consiste en el cálculo de la media móvil de orden 20 de los caudales instantáneos, redondeando esta media al valor más próximo múltiplo de 0,0125. Ambos valores, orden de la media móvil y redondeo, se han establecido experimentalmente tras varias pruebas con diferentes valores hasta obtener una *geometrización* óptima.

### 4.2.2. Identificación de eventos

Como ha quedado dicho anteriormente, se entiende por*evento* cada una de las unidades elementales de consumo ocurridas en un periodo de tiempo de duración suficiente, en el que el caudal instantáneo se mantiene claramente diferenciable del resto. Un uso doméstico concreto puede estar formado por uno o varios eventos, cuya eventual combinación con los eventos de otros usos dan lugar a un episodio de caudal más o menos complejo.

Es preciso señalar que la *geometrización* de los caudales instantáneos no es más que una herramienta ideada para facilitar la identificación de los instantes de inicio y final de los eventos que forman parte de un episodio de caudal registrado. En modo alguno los caudales *geometrizados* sustituyen a los instantáneos, calculados previamente, sino que únicamente se trata de un artificio para poder identificar eventos y asignarles una serie de parámetros, los cuales permiten su consiguiente etiquetado –asignación de un determinado uso doméstico–, durante el desarrollo del Módulo 3.

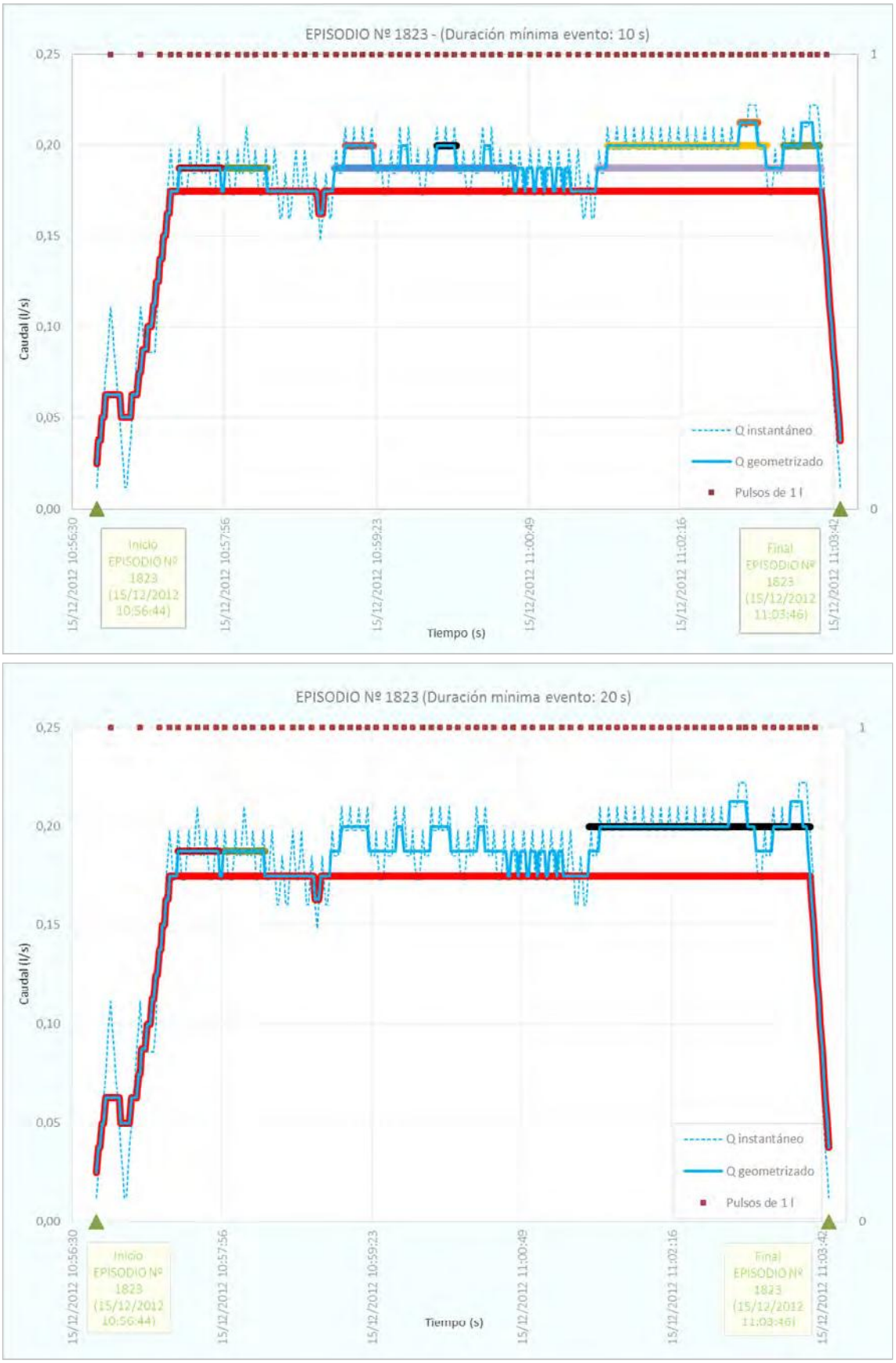
Los episodios son tratados como una superposición de eventos, “apilados” unos sobre otros a modo de escalera, que son identificados por las huellas de los escalones que resultan de la*geometrización.*

El criterio adoptado para discriminar eventos es el siguiente: se ha considerado que mientras se mantenga constante un caudal durante un determinado tiempo, o si una eventual variación de caudal no se mantiene –huella del escalón de una mínima duración–, se trata de un evento único. La cuantificación de esta duración mínima se ha establecido de forma empírica, comparando los resultados para diferentes duraciones, habiéndose optado por un valor de 20 segundos. Es decir, se considerará un cambio de evento si se produce un cambio a un nuevo caudal que se mantiene, al menos, durante 20 segundos.

Las siguientes figuras muestran la diferencia entre considerar una duración mínima de 10 segundos y una duración mínima de 20 segundos. Según se ve en el primer caso (Figura 18) el episodio queda dividido en 10 eventos, mientras que en el segundo se reducen a 4 (Figura 19). Los resultados se han comparado con los obtenidos en su día en el marco de un estudio similar, con estos mismos contadores, pero aplicando otra metodología, concretamente, con el modelo*Trace Wizard*, habiéndose verificado que los resultados obtenidos por dicho modelo son del mismo orden que los obtenidos con la duración de 20 segundos.

En definitiva, la división en eventos efectuada con una duración mínima de evento de 20 segundos resulta adecuada.

**FIGURA 18. IDENTIFICACIÓN DE EVENTOS CONSIDERANDO UNA DURACIÓN MÍNIMA DE 10 Y 20 SEGUNDOS**



Volviendo al episodio nº 376 tomado como ejemplo, tal como refleja la Figura 19, el caudal 0,15 l/s marca claramente un escalón que define un primer evento, que puede considerarse que se extiende a lo largo de todo el episodio. Superpuesto a éste, se diferencia un segundo escalón de mucha menor duración, pero suficiente para considerarse un evento –se mantiene este caudal durante 22 segundos–. Y, por último, un tercer evento, "apoyado" también sobre el primero, termina de completar el episodio.

**FIGURA 19. IDENTIFICACIÓN DE EVENTOS, EPISODIO 376**


Una vez descompuestos los episodios en eventos elementales, procede la caracterización de éstos mediante los parámetros que se detallan en el apartado siguiente.

### 4.2.3. Parámetros para la caracterización de eventos

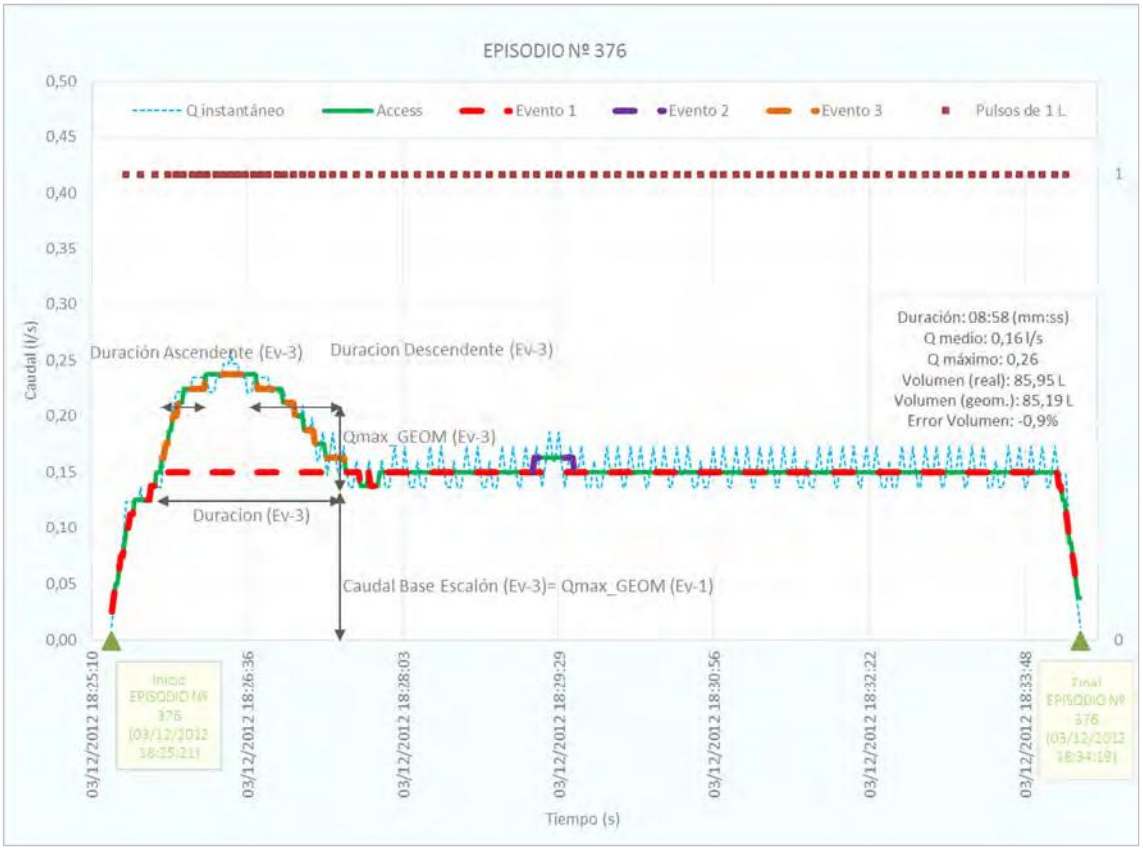
La *geometrización* de los episodios y su descomposición en eventos permiten la identificación de los instantes inicial y final de éstos, y con ellos la definición de los parámetros necesarios para su caracterización, los cuales se refieren tanto a datos procedentes de los caudales instantáneos como a datos procedentes de los caudales geometrizados.

A continuación, se describe cada uno de los parámetros que se han considerado y, a modo de ejemplo, los valores que toman para el caso concreto del susodicho **evento 3 del episodio 376 del mes de diciembre de 2012, del contador 006.**

- ***IdContador***: identificador del contador.
- ***Mes***: mes en el que se ha producido el evento.
- ***Episodio***: número correlativo para la identificación del episodio de caudal al que pertenece el evento.
- ***Evento:*** número correlativo del evento dentro del episodio
- ***Inicio***: instante en el que empieza el episodio
- ***InstanteFinal***: instante en el que termina el episodio.
- ***Duracion***: tiempo de duración del evento.
- ***DuracionAscendente***: tiempo de duración de la rama ascendente
- ***DuracionDescendente***: tiempo de duración de la rama descendente
- ***GradienteAscendente***: relación entre el caudal máximo y la duración de la rama ascendente.
- ***GradienteDescendente***: relación entre el caudal máximo y la duración de la rama descendente.
- ***Volumen***: volumen del evento.
- ***Qmax_GEOM***: caudal máximo del evento *geometrizado.*
- ***NumEventosSimul***: número de eventos que se registran simultáneamente y que se sitúan por debajo del evento en cuestión; indica el nivel que ocupa en la "escalera de eventos geometrizados".
- ***CaudalBaseEscalon***: caudal sobre el que se "apoya" el evento
- ***Simultaneidad***: indica los eventos que se producen simultáneamente junto con el evento en cuestión.

Para una mejor compresión, en la Figura 20 se representan gráficamente los parámetros *Duracion, DuracionAscendente, DuracionDescendente CaudalBaseEscalon y Qmax_GEOM*, del evento 3 de dicho episodio 376.

**FIGURA 20.** PARÁMETROS PARA LA CARACTERIZACIÓN DE EVENTOS




## 4.3. CLASIFICACIÓN DE EVENTOS

Una vez identificados y caracterizados los eventos según se ha expuesto más arriba, llega el momento de su clasificación, asignándoles el uso doméstico final que corresponda. Este proceso de "clasificación" requiere un aprendizaje previo, que se realiza tomando como referencia eventos ya clasificados, que sirven de patrón para los diferentes usos domésticos a considerar.

Los eventos clasificados previamente mediante operador son los que se han utilizado como patrón de aprendizaje para la clasificación automática de todos los eventos identificados en el módulo anterior.

La clasificación se ha realizado considerando dos metodologías diferentes, según el método de aprendizaje, a saber:

- Clasificación de eventos mediante *Redes Neuronales Artificiales* con técnicas de aprendizaje profundo.
- Clasificación de eventos mediante *Máquinas de Vectores Soporte.*

Las clases de usos utilizadas en esta clasificación o etiquetado han sido las siguientes:

*-Grifos*
*-Cisternas*
*-Duchas, que incluye las bañeras*
*-Lavadora*
*-Lavavajillas*
*-Piscina*
*-Riego*
*-Fugas*

### 4.3.1. Etiquetado de eventos por operador

La metodología que se ha empleado para el etiquetado automático de eventos ha sido la utilización de algoritmos de aprendizaje supervisado. Este tipo de algoritmos parte de un conjunto de datos de entrenamiento, previamente clasificado, y sobre esa base aprende a clasificar nuevos datos. Para generar dicho conjunto de entrenamiento es necesario procesar un periodo de tiempo, en el que se conoce*a priori* el uso final de los eventos detectados en dicho periodo, e identificar automáticamente los eventos mediante la metodología descrita en el apartado anterior.

Los periodos de tiempo seleccionados son aquéllos en los que previamente se ha realizado una identificación manual de eventos por un operador. En paralelo se ha procesado también con el procedimiento automático desarrollado para la identificación de eventos.

Para la asignación de las etiquetas de los datos de entrenamiento se ha buscado, para cada evento detectado, el equivalente en etiquetado por operador y se le ha asignado dicha etiqueta. La relación de eventos entre ambas plataformas no es unívoca, por lo que el proceso de emparejamiento de eventos se ha realizado de la siguiente forma:

1. Para cada usuario se compara cada evento resultante del procedimiento automático (módulo de identificación de eventos) con aquellos eventos etiquetados por operador que coincidan total o parcialmente en el tiempo, con el fin de seleccionar aquellos eventos etiquetados por operador cuyo instante de inicio sea anterior al final del evento del procedimiento automático y su instante final posterior al momento de inicio de éste.

2. Una vez seleccionados los eventos coincidentes se analizan sus características para ver a cuál de ellos se parece más. Para ello, se sigue un método basado en dos comparaciones:

    a. El Índice de Jaccard

    b. El volumen de los eventos

    El índice de Jaccard es un coeficiente que mide el grado de similitud entre dos conjuntos, independientemente de los elementos que contengan. Esta similitud se mide como el número de elementos presentes en la intersección de ambos conjuntos dividido por el número total, es decir:

$$J(A,B) = \frac{|A \cap B|}{|A \cup B|}$$

Por tanto, el índice de Jaccard se calcula como el tiempo (segundos) en que el evento del módulo 2 y el evento etiquetado por el operador coinciden entre el tiempo (segundos) del periodo que va desde el mínimo de los instantes de inicio de los dos eventos al máximo de los instantes finales, es decir:

$$J(ev_{TW}, ev_{M2}) = \frac{duración\ (ev_{TW}\ \cap\ ev_{M2})}{duración\ (ev_{TW}\ \cup\ ev_{M2})}$$

El evento seleccionado será aquel que tenga un mayor índice de Jaccard, aunque este debe ser ratificado por la comparación de volúmenes. La comparación del volumen es necesaria porque la división de eventos no se realiza con el mismo criterio en el módulo de identificación de eventos que por el operador y, en consecuencia, el tiempo en el que los eventos coinciden puede ser engañoso. Esta comparación es una tolerancia encubierta, ya que ratifica o deshace la asignación basada en el índice de Jaccard. Si alguno de los dos volúmenes es el triple que el otro, es decir, si:

$$\frac{Vol(ev_{TW})}{Vol(ev_{M2})} < \frac{1}{3} \quad o \quad \frac{Vol(ev_{TW})}{Vol(ev_{M2})} > 3$$

Entonces, se busca entre el resto de candidatos cuál tiene un volumen más similar.

### 4.3.2. Variables de entrada

El conjunto de datos de entrada del algoritmo está formado por**37 variables**. De estas variables, algunas han sido previamente calculadas por el módulo de identificación de eventos (8 variables). El resto de ellas (29 variables) se obtienen en tiempo real, con el fin de disponer de más información útil a la hora de clasificar y asignar el uso doméstico de cada evento, permitiendo así crear modelos que clasifiquen los eventos de una manera satisfactoria.

Para ciertos algoritmos de clasificación es necesario llevar a cabo una selección de variables debido a la complejidad algorítmica que supone tener un conjunto de datos (variables y observaciones) relativamente grande. En el marco de este proyecto ha sido necesaria esa selección de variables para la clasificación basada en máquinas de vectores soporte, pero no lo ha sido para la clasificación mediante redes neuronales, pues su gran potencia de cálculo y aprendizaje no sólo ha permitido usar todas las variables sino también más observaciones.

Cuando se tiene una cantidad suficientemente grande de datos, las redes neuronales son capaces de seleccionar de forma automática las variables que les son útiles para clasificar un problema. Esto es posible ya que pueden multiplicar cada variable de entrada por un peso, y este será cercano a cero en las variables que no aporten información útil.

En el caso de las máquinas de vectores soporte, sí que es necesaria una selección acertada de variables para el correcto funcionamiento del algoritmo. Esta selección se ha realizado de forma independiente para los contadores de 0,1 y 1 litros y el procedimiento utilizado ha sido el que comúnmente se conoce como “voraz”, que consiste en un proceso iterativo en el que se selecciona la variable más útil ejecutando el algoritmo utilizando cada una de las variables de forma individual; a continuación, para seleccionar la segunda variable, se vuelve a ejecutar utilizando la variable ganadora y cada una de las otras variables. Este proceso continúa de forma iterativa hasta que no mejoran los resultados al incluir nuevas variables.

La Tabla 3 recoge las variables seleccionadas en cada tipo de contador. Se puede observar que en los contadores de 0,1 litros se han seleccionado más variables. Esto se debe a que, al utilizar contadores de 1 litro, la falta de precisión en la medición de caudales hace que muchas variables no aporten información útil.

**TABLA 3. VARIABLES DE ENTRADA PARA EL PRE-PROCESAMIENTO DE EVENTOS**

| *Variable* | *Descripción* | *Contador tipo 1 litro* | *Contador tipo 0,1 litro* |
|---|---|---|---|
| Duración | Duración (segundos) del evento | ✓ | |
| DuracionAscendente | Duración (segundos) de la rama ascendente del evento | ✓ | ✓ |
| DuracionDescendente | Duración (segundos) de la rama descendente del evento | ✓ | ✓ |
| GradienteAscendente | Gradiente de la rama ascendente | ✓ | ✓ |
| GradienteDescendente | Gradiente de la rama descendente | ✓ | ✓ |
| Volumen | Volumen del evento en litros | | ✓ |
| Qmax_GEOM | Caudal máximo (litros por segundo) geometrizado | ✓ | ✓ |
| CaudalBaseEscalon | Caudal máximo (litros por segundo) del evento con el que se superpone | ✓ | ✓ |
| HoraEpisodio | Hora de inicio del episodio en el que ocurre el evento | | ✓ |
| VolumenEpisodio | Volumen total (litros) del episodio en el que ocurre el evento | ✓ | ✓ |
| DuracionEpisodio | Duración total (segundos) del episodio en el que ocurre el evento | ✓ | ✓ |
| Volumen de los cuatro episodios anteriores | 4 variables que representan el volumen (litros) de cada uno de los cuatro episodios anteriores al episodio en el que ocurre el evento | | ✓ |
| Duración de los cuatro episodios anteriores | 4 variables que representan la duración (segundos) de cada uno de los cuatro episodios anteriores al episodio en el que ocurre el evento | | ✓ |
| Tiempo transcurrido desde los cuatro episodios anteriores | 4 variables que representan el tiempo transcurrido (segundos) desde el fin de cada uno de los cuatro episodios anteriores hasta el inicio del episodio en el que ocurre el evento | | ✓ |
| Distancia a los cuatro episodios posteriores | 4 variables que representan el tiempo transcurrido (segundos) desde el fin del episodio en el que ocurre el evento hasta el inicio de cada uno de los cuatro episodios posteriores | | ✓ |
| Volumen del episodio con un volumen superior a 1 litro más cercano en el tiempo | Volumen (litros) del episodio más cercano en el tiempo de entre los 5 episodios anteriores y 5 posteriores que tienen un volumen superior a 1 litro | | ✓ |
| Tiempo transcurrido entre el episodio con un volumen superior a 1 litro más cercano en el tiempo | Tiempo transcurrido (segundos) del episodio en el que ocurre el evento al episodio más cercano de entre los 5 episodios anteriores y 5 episodios posteriores que tienen un volumen superior a 1 litro | | ✓ |

### 4.3.3. Normalización de variables

Las variables de entrada de los algoritmos han sido sometidas a procesos de normalización como parte del procedimiento habitual del análisis de datos, previo al desarrollo de modelos de clasificación y predicción.

La normalización consiste en la transformación de los datos de modo que todas las variables estén medidas en la misma escala. Así es posible hacer comparaciones entre ellas, que sean independientes de la unidad de medida empleada en cada una y se evita que,*a priori*, unas vayan a tener más peso que otras en los modelos que se desarrollan más adelante.

Existen diversas técnicas para la normalización de variables invariantes de escala. En el modelo de clasificación basado en máquinas de vectores soporte desarrollado en el marco de este proyecto, la transformación elegida convierte a las variables de entrada en variables con media0 y desviación típica 1, siguiendo una **puntuación estándar**:

$$\tilde{X} = \frac{X - \mu}{\sigma}$$

donde $\tilde{X}$ es la variable $X$ normalizada, $\mu$ y $\sigma$ son la media y la desviación típica de la variable$X$.

Sin embargo, para la clasificación mediante redes neuronales la normalización de variables las convierte en el intervalo $[0 - \ 0{,}1]$ . Para ello, utiliza la siguiente expresión:

$$X' = X - min(\mathrm{X})$$

$$\tilde{X} = \frac{X'}{max(X')} * 0.1$$

donde $\tilde{X}$ es la variable $X$ normalizada.

### 4.3.4. Clasificación de eventos mediante Redes Neuronales Artificiales con técnicas de aprendizaje profundo

Una red neuronal es una herramienta matemática que modela, de forma simplificada, el funcionamiento del cerebro. En términos simplistas, es una serie de operaciones matemáticas sobre un vector de entrada que da como resultado otro vector de salida.

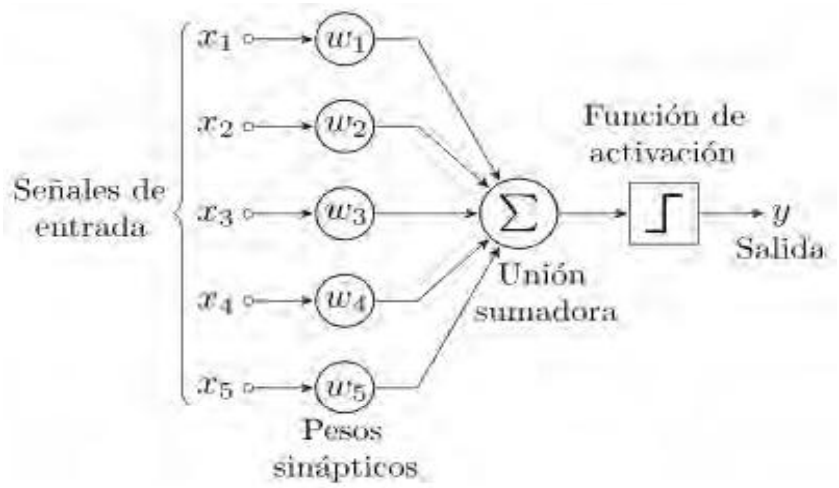
#### *El perceptrón*

El elemento básico de computación se llama habitualmente**nodo** o **neurona**. Pese a que existen múltiples tipos de neuronas (perceptrón, RBF, mapas auto-organizados, recurrentes,…) la más extendida (ya que es de carácter general y aplicable a todo tipo de problemas) es el**perceptrón**.

El perceptrón recibe como entrada una serie de variables de una fuente externa de datos. Cada*input* tiene un peso asociado $w$, que se va modificando durante el proceso deaprendizaje.

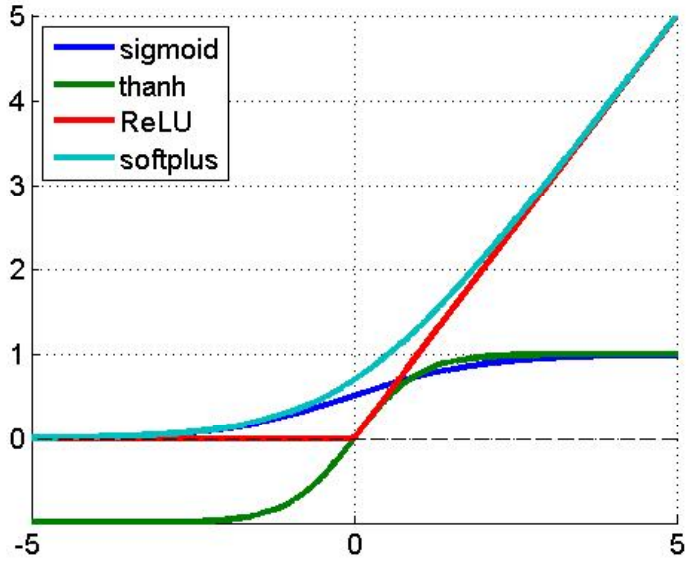
Cada unidad aplica una **función de activación** $f$ de la suma de las entradas ponderadas por los pesos. La Figura 21 muestra la estructura de un perceptrón:

**FIGURA 21. REDES NEURONALES ARTIFICIALES. ESTRUCTURA DE UN PERCEPTRÓN**




En cuanto a funciones de activación existen múltiples alternativas: sigmoide, tangente hiperbólica, lineal rectificada (ReLU), *softplus*, etc. En la Figura 22 se muestran algunos ejemplos de función de activación representando el valor de la salida en función de la entrada.

**FIGURA 22. REDES NEURONALES ARTIFICIALES. FUNCIONES DE ACTIVACIÓN**



### *El perceptrón multicapa*

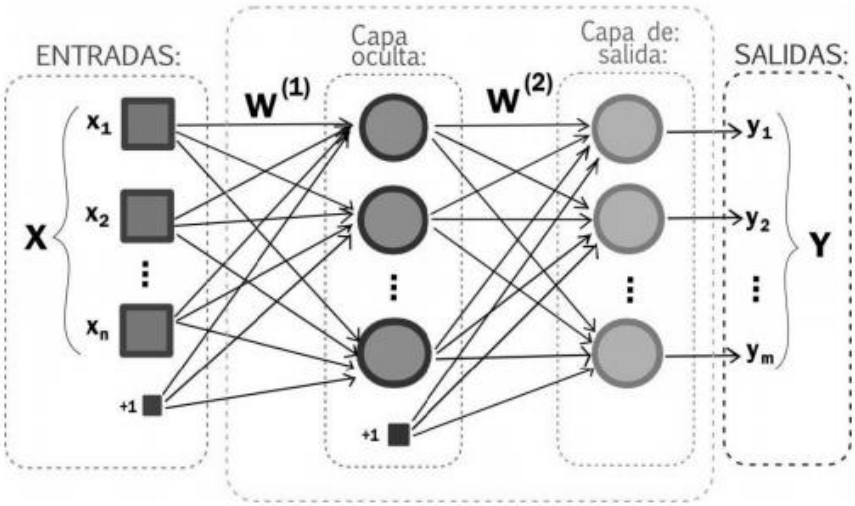
Si se utiliza un perceptrón para clasificar datos, solo se podrán resolver problemas linealmente separables en el espacio de variables de entrada. Para no tener esta limitación se puede crear una **red neuronal artificial** (ANN, "*Artificial Neural Network*") formada por múltiples capas[16] [17]. Esto permite crear una estructura que resuelva problemas no lineales (Figura 23).

Cuando una red neuronal se construye con neuronas del tipo perceptrón se le denomina **perceptrón multicapa**, donde las salidas de las neuronas de una capa son las entradas de las neuronas de la capa siguiente. Atendiendo a la posición de las capas se pueden clasificar en tres tipos:

- **Entrada**: reciben los parámetros de entrada (vector entrada).
- **Salida**: producen los valores resultados (vector salida).
- **Oculta**: contienen los cálculos intermedios.

Aquí se muestra un ejemplo típico de red neuronal:

**FIGURA 23. EJEMPLO DE RED NEURONAL ARTIFICIAL. PERCEPTRÓN MULTICAPA**




Una red neuronal típica está formada por una capa de entrada con un número de neuronas igual al número de variables que tiene el conjunto de datos, una capa de salida con un número de neuronas igual a la cantidad de categorías en las que pueden clasificarse los datos y varias capas ocultas encargadas de procesar los datos.

Para obtener el valor de los pesos de cada neurona de la red se utiliza el algoritmo de **propagación hacia atrás.**

16 Bishop, C. M. (2006). *Pattern Recognition and Machine Learning*, Information Science and Statistics, Springer-Verlag New York. Inc. Secaucus, NJ, USA.

### *Propagación hacia atrás*

La propagación hacia atrás de errores (*backpropagation*) es un algoritmo de aprendizaje supervisado que se usa para entrenar redes neuronales artificiales.

El algoritmo emplea un ciclo propagación–adaptación de dos fases. Una vez que se ha aplicado un patrón a la entrada de la red como estímulo, éste se propaga desde la primera capa a través de las capas superiores de la red, hasta generar una salida. La señal de salida se compara con la salida deseada y se calcula una señal de error para cada una de las salidas.

El error se propaga hacia atrás, partiendo de la capa de salida, hacia todas las neuronas de la capa oculta que contribuyen directamente a la salida. Sin embargo, las neuronas de la capa oculta solo reciben una fracción de la señal total del error, basándose aproximadamente en la contribución relativa que haya aportado cada neurona a la salida original. Este proceso se repite, capa por capa, hasta que todas las neuronas de la red hayan recibido una señal de error que describa su contribución relativa al error total.

La importancia de este proceso consiste en que, a medida que se entrena la red, las neuronas de las capas intermedias se organizan a sí mismas, de tal modo que las distintas neuronas aprenden a reconocer distintas características del espacio total de entrada. Después del entrenamiento, cuando se les presente un patrón arbitrario de entrada que contenga ruido o que esté incompleto, las neuronas de la capa oculta de la red responderán con una salida activa si la nueva entrada contiene un patrón semejante a los que se han aprendido durante su entrenamiento.

### *Aceleración del entrenamiento*

Normalmente, antes de empezar el entrenamiento de la red neuronal, los pesos de las neuronas se inicializan al azar.

Un problema de los algoritmos de propagación hacia atrás es que el error se va diluyendo de forma exponencial a medida que atraviesa capas en su camino hasta el principio de la red. Esto es un inconveniente porque en una red muy profunda (con muchas capas ocultas), sólo las últimas capas se entrenan, mientras que las primeras apenas sufren cambios. Por este motivo, en el pasado compensaba utilizar redes con pocas capas ocultas que contengan muchas neuronas, en lugar de redes con muchas capas ocultas que contengan pocas neuronas. Esto ha sido así hasta que se desarrollaron mejoras para entrenar redes con múltiples capas ocultas.

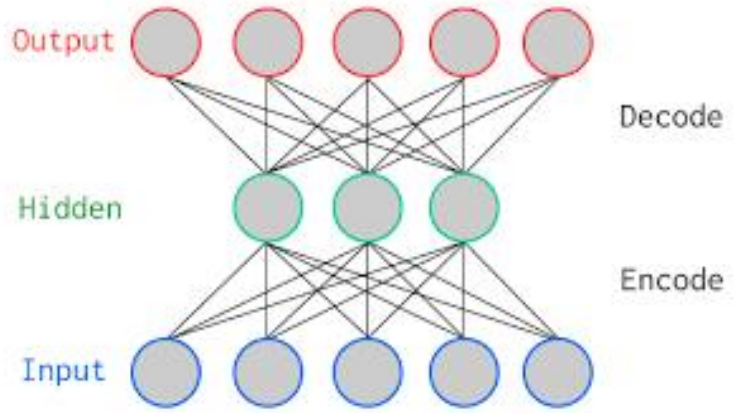
### *Autocodificadores apilados*

Esta técnica se ha utilizado en este proyecto para mejorar las prestaciones y acelerar el entrenamiento de las redes neuronales profundas. Los autocodificadores apilados son una técnica que se utiliza para poder inicializar los pesos de las redes neuronales, de forma que el entrenamiento sea más rápido y se puedan obtener redes con muchas capas ocultas *(Deep Learning)* en un tiempo razonable de entrenamiento[18].

[18] Vincent, P., Larochelle, H., Lajoie, I., Bengio, Y., & Manzagol, P. A. (2010).*Stacked Denoising Autoencoders: Learning useful representations in a deep network with a local denoising criterion*.Journal of Machine Learning Research, 11 (Dec), 3371-3408

Un **autocodificador** es una red neuronal con una única capa oculta que aprende a producir a la salida exactamente la misma información que recibe a la entrada. Por eso, las capas de entrada y salida siempre deben tener el mismo número de neuronas. Por ejemplo, si la capa de entrada recibe los píxeles de una imagen, se espera que la red aprenda a producir en su capa de salida exactamente la misma imagen que le hemos introducido, algo que a primera vista carece de utilidad. En la Figura 24 se puede apreciar la estructura:

**FIGURA 24. RED NEURONAL ARTIFICIAL. AUTOCODIFICADORES APILADOS**




La clave está en la capa oculta. Supongamos que tenemos un auto-codificador con menos neuronas en la capa oculta que en las capas de entrada y salida. Dado que se exige a esta red que produzca a la salida el mismo resultado que recibe a la entrada, y la información tiene que pasar por la capa oculta, la red se verá obligada a encontrar una representación intermedia de la información en su capa oculta usando menos números. Por tanto, al aplicar unos valores de entrada, la capa oculta tendrá una versión comprimida de la información, pero además será una versión comprimida que se puede volver a descomprimir para recuperar la versión original a la salida.

De hecho, una vez entrenada se puede dividir la red en dos: una primera red que utiliza la capa oculta como capa de salida y una segunda red que utiliza esa capa oculta como capa de entrada. La primera red sería un compresor y la segunda un descompresor.

Precisamente por este motivo, este tipo de redes se denominan autocodificadores: son capaces de descubrir por sí mismos una forma alternativa de codificar la información en su capa oculta. Una de sus ventajas es que no necesitan a un supervisor que les muestre ejemplos de cómo codificar información.

### Añadir ruido a los datos de entrada

Existe una técnica que hace que la red no se limite únicamente a copiar la información cuando la capa oculta tiene neuronas de sobra. Esta técnica, en lugar de indicar el mismo valor para las entradas y las salidas, introduce **ruido en el vector de entrada** y deja el de salida sin ruido al componer los ejemplos. De esta forma, la red está forzada a generalizar porque tendrá varios ejemplos de entrada ligeramente diferentes que producen la misma salida.

La representación intermedia en la capa oculta tendrá que focalizarse en las características comunes de todas las versiones del mismo dato con diferentes ruidos.

### Apilación de autocodificadores

Un solo autocodificador puede detectar características fundamentales en la información de entrada. Sin embargo, si se quiere que las máquinas detecten conceptos más complejos será necesaria más potencia.

Por ejemplo, a partir de información cruda, sin significado (por ejemplo, píxeles de imágenes) un autocodificador es capaz de detectar características simples (líneas y curvas). Si al resultado codificado en esa capa oculta se le aplica otro autocodificador será capaz de encontrar características más complejas (como círculos, arcos, ángulos rectos, etc.). Si se repite este proceso se obtendrá una jerarquía de características cada vez más complejas junto con una pila de codificadores. Siguiendo el ejemplo de las imágenes, dada una profundidad suficiente e imágenes de ejemplo suficientes, se conseguirá alguna neurona que se active cuando la imagen tenga un rostro, y sin necesidad de que ningún supervisor le explique a la red cómo es un rostro.

La idea de *Deep Learning* mediante autocodificadores apilados es precisamente esa: usar varios codificadores y entrenarlos, uno a uno, usando cada codificador entrenado para entrenar el siguiente.

Tras inicializar la red neuronal con autocodificadores apilados, cuando ésta se entrena mediante el algoritmo de propagación hacia atrás, el entrenamiento finaliza en un tiempo mucho menor y con la necesidad de un número mucho menor de datos de entrenamiento.

## *Ensamblaje con otros algoritmos*

Una técnica muy utilizada para acelerar el entrenamiento de algoritmos complejos de aprendizaje automático es añadir variables que aporten información útil de cara a la clasificación[19]. Por ejemplo, entrenando algoritmos mucho más sencillos y rápidos e incluyendo su salida como una nueva variable de entrada en la red neuronal.

En el marco de este proyecto, se ha utilizado un algoritmo rápido y sencillo llamado **árboles potenciados por gradiente** *(Gradient Boosted Trees)*[20]. Este algoritmo construye una cadena de árboles de decisión en los que cada árbol trata de resolver los errores que comete el anterior.

La predicción de este algoritmo se utiliza como una nueva entrada en la red neuronal. De esta forma, la red parte con este conocimiento y sus esfuerzos se centran en aprender a clasificar los problemas más complejos.

---

[19] Zhang, C., & Ma, Y. (Eds.). (2012). *Ensemble Machine Learning: Methods and Applications* Springer Science & Business Media.

[20] Chen, T., & Guestrin, C. (2016). Xgboost:*A Scalable Tree Boosting System*. Proceedings of the KDD 2016 Conference.

### *Criterios de entrenamiento y test*

Con el objetivo de validar correctamente las prestaciones de las redes neuronales en estas bases de datos se ha realizado un proceso de entrenamiento y test que permite ver el rendimiento al procesar datos que no han sido vistos para entrenar las redes.

Antes de procesar todos los contadores con los parámetros definitivos, se ha realizado un proceso previo que ha permitido seleccionar los parámetros con los que mejores prestaciones se consiguen.

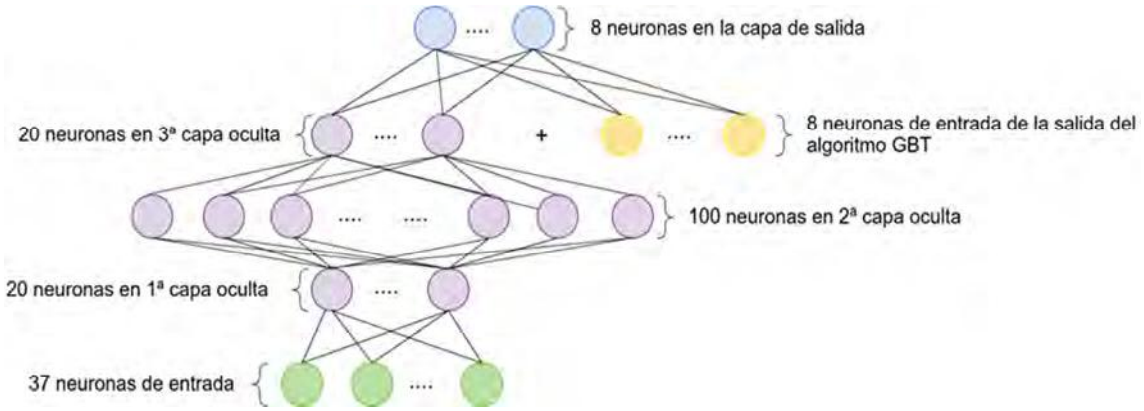
Se ha comenzado haciendo pruebas para seleccionar los parámetros que indican la profundidad del árbol y el número de estimadores de los árboles potenciados por gradiente. A la vista de los resultados se ha decidido mantener fijo el número de estimadores a 100 y la profundidad del árbol a 11 niveles.

El número de épocas de entrenamiento indica el número de veces que el conjunto total de eventos de entrenamiento va a ser usado por la neurona para actualizar los pesos y es uno de los parámetros de la red neuronal.

En lo que se refiere a las redes neuronales, se ha hecho una ejecución de todos los contadores validando el número de épocas (con valores 20, 50,100) y el número de capas y neuronas de la red neuronal. En los resultados se ha visto que la arquitectura de la red neuronal no era lo más relevante, sino que lo resultados mejoraban a medida que se aumentaba el número de épocas de entrenamiento. Es por ello que se ha decidido fijar el número de épocas a 100 y el número de capas y neuronas como sigue (ver Figura 25):

- 37 neuronas de entrada. Una por cada variable de entrada.
- 20 neuronas en la primera capa oculta.
- 100 neuronas en la segunda capa oculta.
- 20 neuronas, en la tercera capa oculta + 8 neuronas de entrada de los árboles potenciados por gradiente.
- 8 neuronas en la capa de salida. Una por cada tipo de evento.

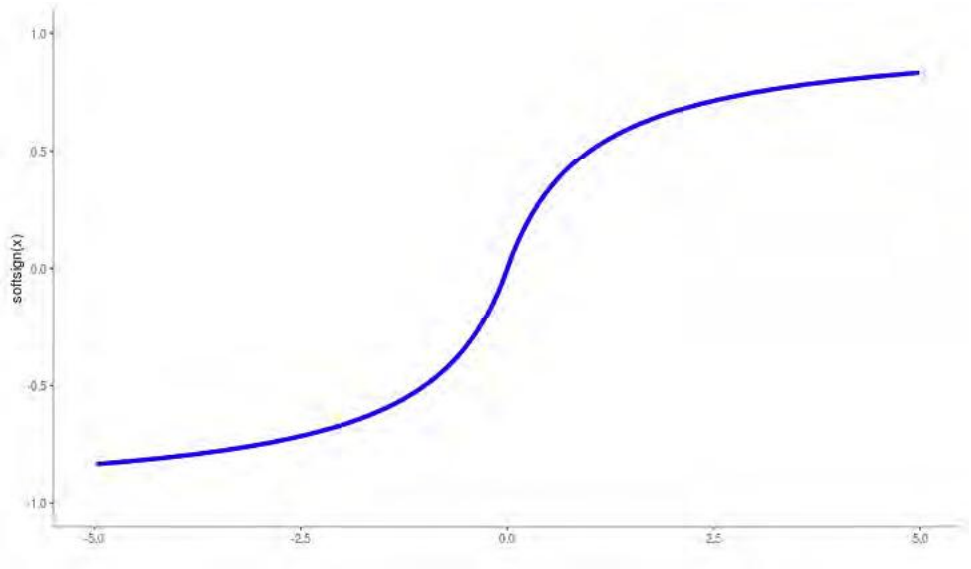
**FIGURA 25. REDES NEURONALES ARTIFICIALES. PROCESO DE ENTRENAMIENTO**



Tras hacer pruebas con varios tipos de funciones de activación en las capas ocultas se ha utilizado la función ***softsign***, que se representa en la Figura 26, ya que es la que mejor resultado ha obtenido:

$$f(x) = \frac{x}{1 + |x|}$$

**FIGURA 26. FUNCIÓN DE ACTIVACIÓN EN LAS CAPAS OCULTAS: SOFTSIGN**




En el caso de las neuronas de salida, la función de activación ha sido ***softmax***, que es utilizada en problemas multiclase donde la salida sólo debe pertenecer a una única categoría. Esta función tiene la expresión:

$$f(x_i) = \frac{e^{x_i}}{\sum_{j=1}^{N} e^{x_j}}$$

Donde $x_i$ es la salida de la neurona i en un total de N neuronas de salida.

### a) Datos de Test

Al cargar los datos que harán que se generen cada uno de los modelos se ha dejado fuera un 30% de los datos con el objetivo de medir las prestaciones del método en eventos que no han sido previamente utilizados para entrenar, es decir, se ha formado un conjunto de test con un 30% de las observaciones.

Con los resultados de los modelos entrenados con las variables normalizadas y seleccionadas, se calculan la tasa de error, la tasa de precisión en litros y la matriz de confusión en la asignación de usos finales.

### b) Datos de Entrenamiento

Debido a que las redes neuronales no tienen un coste computacional demasiado alto, se ha limitado a 100.000 el número de eventos utilizados para el entrenamiento y el test, es decir, para la mayoría de contadores se cuenta con la totalidad de eventos. Del número total de eventos, el 70% se utiliza para entrenar la red neuronal.

En la primera fase del entrenamiento se entrenan los autocodificadores para inicializar los pesos de la red neuronal. A cada una de las capas de entrada de los autocodificadores se le añade ruido Gaussiano con una desviación estándar de 0,01.

A continuación, se procesan los eventos con los árboles potenciados por gradiente. La predicción de este algoritmo se añade como entrada a la red neuronal.

### c) Desarrollo e implementación

Los algoritmos descritos se han desarrollado e implementado en el lenguaje de programación *Python* 2.7[21], software desarrollado mediante licencia*Open Source* que permite su libre uso y distribución. Se ha utilizado *Anaconda,*[22] instalador de *Python* con un conjunto de paquetes orientados al aprendizaje automático y análisis de datos, también con licencia*Open Source*.

Los paquetes utilizados para el desarrollo de los modelos de redes neuronales han sido:

- *Scikit-learn*[23] es una librería ***Open Source de Aprendizaje Automático*** para entornos de programación en *Python*. Incluye la implementación de los algoritmos más importantes del estado del arte actual en clasificación, regresión, métodos bayesianos, etc. Es altamente eficaz gracias a la utilización de las bibliotecas numéricas y científicas*NumPy* y *SciPy*, fuertemente optimizadas para la ejecución de cálculos matemáticos. Se ha utilizado la**versión 0.17.1** de esta librería.

- *NumPy*[24]es la librería encargada de dar soporte matemático y vectorial de alto nivel para operar con vectores y matrices. La **versión** utilizada es la **1.11.1**.

- *Pyodbc*[25] es la librería encargada de gestionar la conexión con la base de datos Access mediante el uso del estándar ODBC. La **versión** utilizada es la **3.0.10**.

- *Theano*[26] es la librería de *Python* que permite definir, optimizar y evaluar eficientemente expresiones matemáticas que implican *arrays* multidimensionales. La **versión** utilizada es la **0.8.2**.

  Entre las características de *Theano* encontramos:

  - ✓ Integración con *Numpy.*
  - ✓ Uso de GPU transparente para el usuario. Realiza cálculos intensivos 140 veces más rápido que utilizando la CPU.

[21] https://www.python.org
[22] https://www.continuum.io/anaconda-overview
[23] http://scikit-learn.org/
[24] http://www.numpy.org/
[25] https://pypi.python.org/pypi/pyodbc
[26] http://deeplearning.net/software/theano/

- ✓ Diferenciación simbólica eficiente. *Theano* realiza derivadas con una o varias entradas.
- ✓ Velocidad y estabilidad en las optimizaciones. Obtiene el resultado correcto para $log(1 + x)$ incluso cuando la $x$ es realmente pequeña.
- ✓ Generación dinámica de código en **C**. Evalúa las expresiones más rápido.
- ✓ Extensa unidad de pruebas y auto verificación. Detecta y diagnostica varios tipos de errores.

- *Keras*[27] es una biblioteca de redes neuronales minimalista, altamente modular, escrita en *Python* y puede ser ejecutada en la parte superior de *Theano*. Fue desarrollada con el objetivo de facilitar al investigador llegar al resultado en el menor tiempo posible. La **versión** utilizada es la versión **1.0.8**.

  Las características principales de *Keras* son:

  - ✓ Permite la creación de prototipos, fácil y rápidamente, (a través de la modularidad total, el minimalismo, y extensibilidad).
  - ✓ Soporta tanto redes convolucionales, como redes recurrentes, así como combinaciones de las dos.
  - ✓ Soporta diferentes tipos de conexión entre neuronas (incluyendo multi-entrada y la formación de múltiples salidas).
  - ✓ Se ejecuta sin problemas en la CPU y la GPU.

- *XGBoost*[28] es una librería de código abierto compatible con *Python* que implementa GBT. La **versión** utilizada es la **0.6.**

### 4.3.5. Clasificación de eventos mediante Máquinas de Vectores Soporte

Las Máquinas de Vectores Soporte[29], conocidas por sus siglas en inglés **SVM** (*Support Vector Machines*) son un conjunto de algoritmos de aprendizaje automático que se usan para resolver problemas de clasificación o de análisis de regresión.

Para entender la motivación de las SVM basta con imaginar un conjunto de puntos donde cada uno de ellos pertenece a una única clase (si, por ejemplo, se tienen dos clases: clase roja y clase azul, un punto sólo podrá ser, o bien de la clase roja, o bien de la clase azul). La SVM se encargará de encontrar la frontera que separe correctamente ambos tipos de datos (línea negra en la Figura 27).

Algo tan sencillo de explicar conceptualmente es verdaderamente complejo en la práctica pues las fronteras entre clases no están tan claras como en la Figura 27. Muchas veces no es posible encontrar la frontera que separe correctamente las clases, o simplemente, tener más de dos clases añade mucha complejidad al problema.

---

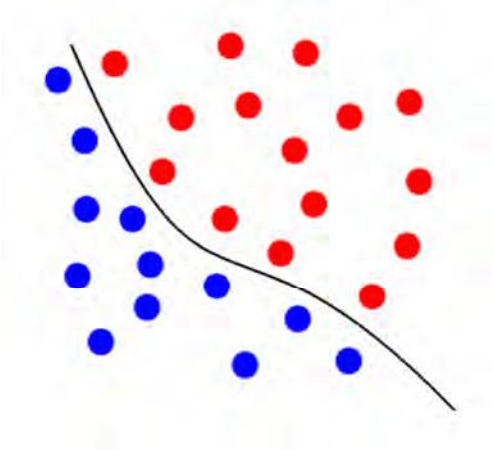
27 https://keras.io/

28 http://xgboost.readthedocs.io/

29 Vapnik, V. *The nature of statistical learning theory*. Springer Science & Business Media, 2013

A continuación, se explica de forma técnica la formulación de las SVM y el algoritmo elegido para resolver el problema de clasificación.

**FIGURA 27. MOTIVACIÓN DE LAS SVM**



## *Máquinas de Vectores Soporte para clasificación binaria*

Las SVM fueron diseñadas originalmente para resolver problemas de clasificación binaria (aquellos problemas en los que los datos pueden pertenecer a dos posibles clases o categorías): una es considerada como la positiva ($y = 1$) y la otra como la negativa ($y = -1$).

Así, se tendrá un conjunto de entrenamiento $\mathcal{D}$ formado por $n$ observaciones en un espacio de $p$ variables:

$$\mathcal{D} = \left\{(\boldsymbol{x_i}, y_i)|\boldsymbol{x_i} \in \mathbb{R}^p, y_i \in \{-1,1\}\right\}_{i=1}^{n}$$

donde $y_i$ es la clase de la observación $\boldsymbol{x_i}$.(la notación en negrita $\boldsymbol{x_i}$ indica que se trata de un vector).

### a) Clasificador de máximo margen

Las SVM parten originalmente del clasificador de máximo margen. Una alternativa cuando se tiene un problema que es perfectamente clasificable por un plano es encontrar el plano cuya distancia a los datos sea mayor.

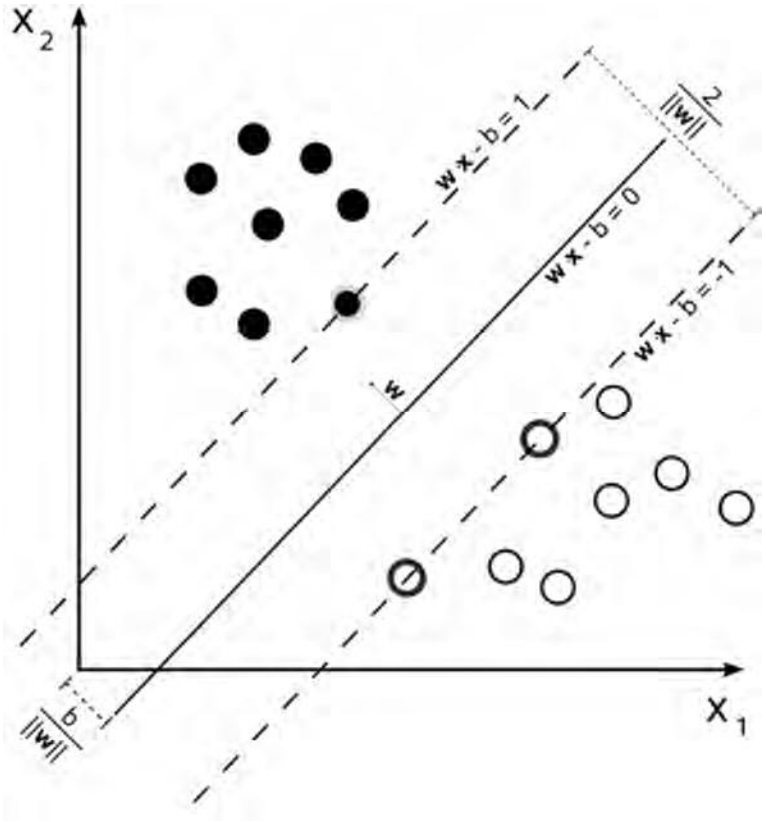
La formulación[30] del problema a resolver es la siguiente: encontrar un hiperplano de dimensión$m - 1$ que separe los ejemplos etiquetados, es decir,

$$Maximiza: \frac{1}{2}\|\boldsymbol{w}\|^2$$
$$Sujeto\ a: y_i(\boldsymbol{w}\boldsymbol{x_i} + b) \geq 1, i = 1,2,\dots,n$$

[30] Vapnik, V. *Pattern recognition using generalized portrait method* Automation and remote control, 1963, 24:774–780

El plano que separa ambas clases es $\boldsymbol{wx_i} + b = 0$ y los planos paralelos con máximo margen son$\boldsymbol{wx_i} + b = \pm 1$ respectivamente, de modo que el problema se reduce a buscar los valores de $W$ y $b$ que maximicen el margen (distancia entre los planos paralelos) y que garanticen que se clasifican correctamente todos los ejemplos del conjunto de entrenamiento, (ver Figura 28).

**FIGURA 28.** **MÁQUINAS DE VECTOR SOPORTE. CLASIFICADOR DE MÁXIMO MARGEN**




## b) SVM lineal de margen blando

Como la mayor parte de los problemas no son separables por un plano, este clasificador se mejoró incluyendo lo que llaman "margen blando" que permite que haya ciertos datos que no cumplan las condiciones de estar separados por el clasificador de máximo margen.

Para cada observación $(\boldsymbol{x}_i, y_i)$, se permite una pérdida u holgura $\xi_i$, relajándose las restricciones de lo que es considerado un ejemplo *bien clasificado*. El ejemplo de entrenamiento $(\boldsymbol{x}_i, y_i) \in \mathrm{D}$ se considera como bien clasificado si se verifica:

$$w \cdot x_i + b \geq +1 - \xi_i \quad para\ y_i = +1$$

$$w \cdot x_i + b \leq -1 + \xi_i \quad para\ y_i = -1$$

$$con\ \xi_i \geq 0\ \forall i \in \{1, \ldots, p\}$$

La cantidad de pérdidas sobre el conjunto de entrenamiento se ajusta usando el parámetro $C$:

- Con un valor de $\boldsymbol{C}$ alto el algoritmo se esforzará en clasificar correctamente las observaciones del conjunto de entrenamiento, pero correrá el riesgo de no poder generalizar y clasificar correctamente observaciones nuevas.

- Por el contrario, para valores bajos de $\boldsymbol{C}$ el sistema obtendrá soluciones generales que funcionarán tanto para observaciones vistas durante el entrenamiento como para las nuevas, pero correrá el riesgo de obtener una solución muy sencilla de bajas prestaciones.

Al añadir estas nuevas condiciones la formulación del problema pasa a ser la siguiente:

$$Maximiza: \frac{1}{2}\|\boldsymbol{w}\|^2 + C\sum_{i=1}^{n}\xi_i$$
$$Sujeto\ a: \begin{matrix} y_i(\boldsymbol{w}\boldsymbol{x_i} + b) \geq 1, i = 1,2,\dots,n \\ \xi_i \geq 0, i = 1,2,\dots,n \end{matrix}$$

Este es un problema de optimización convexo, lo que garantiza una solución única.

En ocasiones, la complejidad en la formulación de un problema conlleva la búsqueda de una formulación alternativa cuya resolución sea conocida. Es lo que se conoce como formulación dual. En este caso, puede hacerse como fórmula de Lagrange pudiendo ser resuelto mediante programación cuadrática:

$$Maximiza\ F(\boldsymbol{\alpha}) = \sum_{i=1}^{n}\alpha_i - \sum_{i=1}^{n}\sum_{j=1}^{n}\alpha_i\alpha_j\, y_i y_j \boldsymbol{x}_i\boldsymbol{x}_j$$
$$Sujeto\ a: \begin{cases} \sum_{i=1}^{n}\alpha_i y_i = 0 \\ 0 \leq \alpha_i \leq C \end{cases}$$

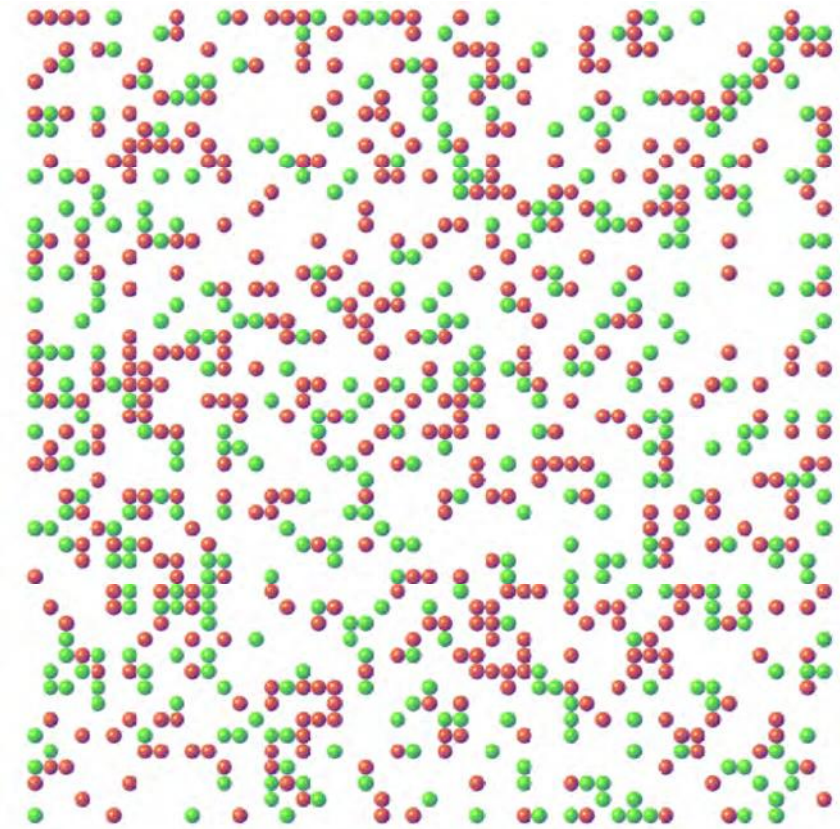
Donde $\alpha_i$ es el multiplicador de Lagrange asociado a la muestra $\boldsymbol{x}_i$.

### c) SVM para la clasificación no lineal

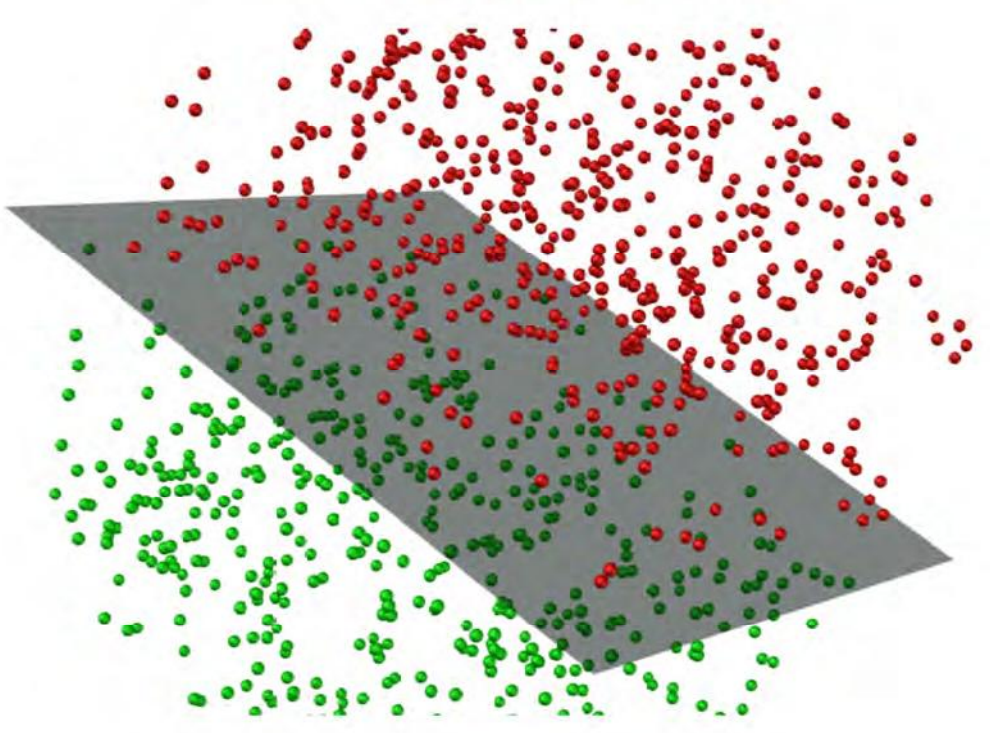
A la hora de clasificar, la utilización de un simple plano para separar ambos tipos de datos, no es la mejor opción, ya que en la vida real hay problemas de naturaleza muy compleja. De modo que es necesario encontrar métodos para obtener separadores no lineales, según se ilustra en la Figura 29.

**FIGURA 29. MÁQUINAS DE VECTOR SOPORTE. SEPARADORES NO LINEALES Y ESPACIO DE CARACTERÍSTICAS**

*a. Separadores No Lineales*



*b. Espacio de Características*



La idea que subyace detrás de estos métodos es la de transformar los ejemplos de entrenamiento a un espacio vectorial de alta dimensión ($N >> n$) (denominado **espacio de características**), donde sea posible la separación lineal. Es lo que se conoce como el**método del núcleo**[31], que sustituye el producto vectorial de la formulación por una función no lineal en el espacio de entrada original:

$$Maximiza\ F(\boldsymbol{\alpha}) = \sum_{i=1}^{n} \alpha_i - \sum_{i=1}^{n}\sum_{j=1}^{n} \alpha_i \alpha_j\, y_i y_j K(\boldsymbol{x}_i, \boldsymbol{x}_j)$$

Las **funciones de núcleo** (funciones encargadas de hacer la transformación, ilustradas en la Figura 29*b.*) más utilizadas son:

- **Función de base radial** (Gaussiana): $K(\boldsymbol{x}_i, \boldsymbol{x}_j) = e^{-\gamma\|x_i - x_j\|^2}$
- **Polinómica**: $K(\boldsymbol{x}_i, \boldsymbol{x}_j) = (\boldsymbol{x}_i \boldsymbol{x}_j + 1)^p$
- **Tangente hiperbólica**: $K(\boldsymbol{x}_i, \boldsymbol{x}_j) = \tanh(k\boldsymbol{x}_i \boldsymbol{x}_j + c)$

### *Formulación de las SVM para problemas multiclase*

Tal y como se ha mencionado al comienzo de esta sección, las máquinas de vector soporte han sido creadas originalmente como clasificadores binarios. En la práctica, el número de categorías en las que se puede clasificar una observación es, con cierta frecuencia, mayor que dos. Para abordar la clasificación de más de dos clases es necesario encontrar un método que transforme un problema de clasificación multiclase en múltiples problemas de clasificación binaria.

1. Método **uno contra todos** (*one-against-all*). Construye $k$ modelos de clasificación binaria, siendo $k$ el número de clases. El $m$-ésimo modelo se entrena con todos los elementos de la clase $m$ como pertenecientes a la clase positiva y el resto a la negativa. Una vez resuelto este problema, se tienen tantas funciones de decisión como clases.

A la hora de clasificar, se utilizan los $k$ modelos y se le asigna la clase cuyo modelo ha obtenido una mayor puntuación.

2. Método **uno contra uno** (*one against one*). Construye $k(k-1)/2$ clasificadores y cada uno de ellos se entrena con datos pertenecientes a dos clases diferentes, es decir, dados los datos de entrenamiento de las clases $i$ y $j$, se resuelve el siguiente problema:

[31] Boser et al., 1992] Boser, B. E., Guyon, I. M., And Vapnik, V. N. (1992).*A Training Algorithm for Optimal Margin Classifiers*. In Proceedings of The Fifth Annual Workshop on Computational Learning Theory, 1992, Pages 144–152. Acm

$$Minimiza: \frac{1}{2}\|w^{ij}\|^2 + C\sum_{t=1}^{n} {\xi_t}^{ij}$$
$$(w^{ij})^T\phi(x_t) + b^{ij} \geq 1 - {\xi_t}^{ij}, \qquad si\ y_t = i$$
$$(w^{ij})^T\phi(x_t) + b^{ij} \leq -1 + {\xi_t}^{ij}, \qquad si\ y_t = j$$
$${\xi_t}^{ij} \geq 0$$

Una vez construidos los clasificadores, para elegir la clase a la que pertenece el elemento se sigue una estrategia de votación: si el elemento $x$ pertenece a la clase $i$ se suma uno a la puntuación de la clase $i$. Si por el contrario pertenece a $j$, es la puntuación de esta clase la que aumenta un punto. Al final se escoge la clase cuyo valor es mayor y, si hay empate entre varias de ellas, se elige la que tenga menor índice. Esto se conoce como estrategia Max Wins.

La estrategia a utilizar en este proyecto es *uno contra uno* debido a que el tiempo de entrenamiento es menor que con *uno contra todos*. El tiempo de entrenamiento de las máquinas de vectores soporte depende del número de elementos que tenga el conjunto de entrenamiento. Utilizando *uno contra uno* se entrenan más modelos, pero con menos datos que utilizando *uno contra todos*[32].

### *Resolución de la SVM: algoritmo SMO*

Para resolver la formulación de la SVM vista en el apartado anterior es necesario resolver un **problema de optimización** (encontrar el máximo de una función) con restricciones (condiciones que tienen que cumplir los datos). Existen numerosas alternativas para ello, entre las que destacan:

- la **optimización mínima secuencial** (*Sequential Minimal Optimization, SMO*)[33],
- el **método del punto interior** (*Interior Point Method, IPM*)[34] y
- el **método iterativo de los mínimos cuadrados ponderados** (*Iterative Re-Weighted Least Squares, IRWLS*)[35] y [36].

[32] Hsu, C. W., & Lin, C. J. *A comparison of methods for multiclass support vector machines* IEEE transactions on Neural Networks, 2002, 13(2), 415-425.

[33] Platt, J. Et al. *Fast Training of Support Vector Machines Using Sequential Minimal Optimization* Advances in Kernel Methods-Support Vector Learning, 1999, 3.

[34] Karmarkar, N. A *New Polynomial-Time Algorithm for Linear Programming* In Proceedings of the Sixteenth Annual ACM Symposium on Theory of Computing, 1984, Pages 302–311. ACM

[35] Pérez-Cruz, F., Bousoño-Calzón, C., and Artés-Rodriguez, A. Convergence of the IRWLS procedure to the support vector machine solution. Neural Computation, 2005, 17(1):7–18.

[36] Díaz-Morales, R. and Navia-Vázquez, A. *Improving the efficiency of IRWLS SVMS using parallel Cholesky factorization*. Pattern Recognition Letters.

Para el presente módulo se ha optado por el método SMO ya que por un lado no requiere una gran cantidad de memoria RAM, pudiendo ser ejecutado en un PC de gama media y por otro lado es el método más extendido y utilizado[37].

SMO es un algoritmo desarrollado en *Microsoft Research*, divide el problema en subproblemas del menor tamaño posible. Para ello, en cada iteración selecciona dos de los coeficientes $\alpha_i$ que aparecen en la formulación dual explicada en el apartado anterior "b) SVM lineal de margen blando" y los actualiza para optimizar la función de coste. Al trabajar únicamente con dos variables cada vez, puede resolver el problema analíticamente de forma sencilla.

### *Criterios de validación y test*

Con el objetivo de validar correctamente las prestaciones de las SVM en estas bases de datos, se realiza un proceso justo de validación y test que permite ver el rendimiento al procesar datos que no han sido vistos para entrenar la SVM.

#### a) Test

El conjunto de datos utilizado para crear los modelos ha sido dividido en dos: conjunto de entrenamiento (formado por el 70% de las observaciones) y conjunto de test (compuesto por el 30% restante). Es necesario tener un conjunto de test para medir las prestaciones del método en eventos que no han sido previamente utilizados para entrenar.

El resultado final de las prestaciones del algoritmo se calculan evaluándose sobre este conjunto de Test.

#### b) Entrenamiento

Debido al coste computacional de las SVM, de complejidad $O(n^3)$ –cada vez que se duplica el número de datos de entrenamiento se multiplica por 8 su tiempo de ejecución– y al alto número de contadores a procesar, es conveniente limitar el valor máximo de datos de entrenamiento. Este valor ha sido fijado por defecto a 10.000 eventos.

El objetivo de la primera fase es obtener el valor de los parámetros de la SVM:

- $C$: parámetro de la función de coste.
- $\gamma$: parámetro de la función de núcleo del tipo función de base radial.

[37] Chang, C.-C. and Lin, C.-J. (2011). *Libsvm: A Library for Support Vector Machines*. ACM Transactions on intelligent systems and technology (TIST), 2011, 2(3):27.

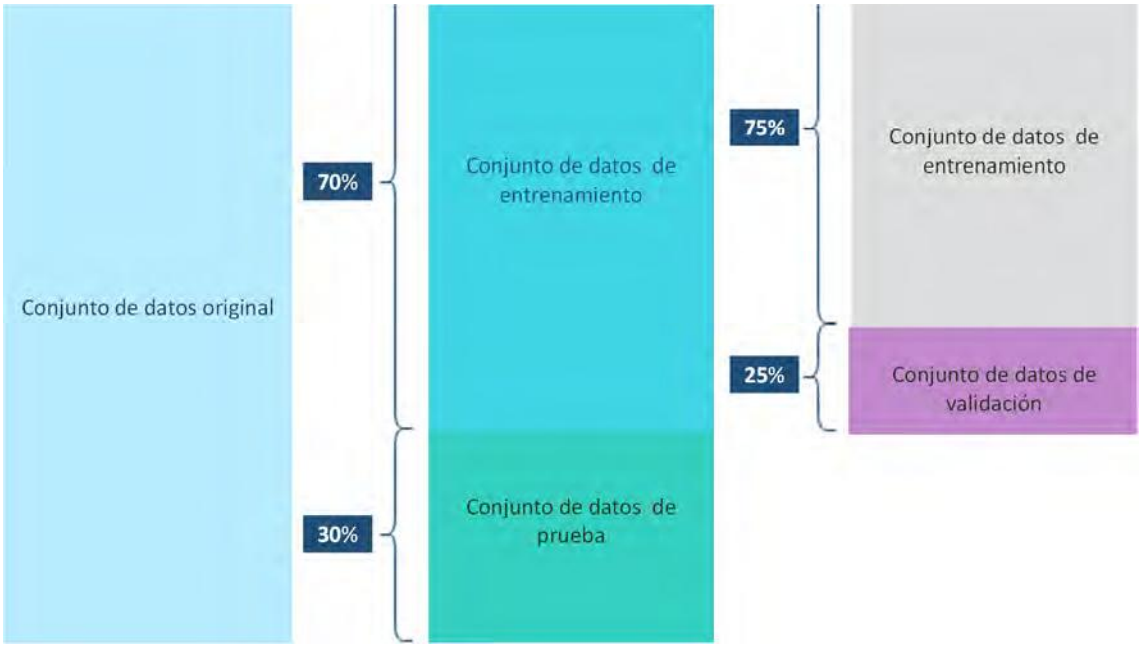
Para obtener el mejor valor de estos parámetros, para un entrenamiento se utiliza el 75% de los datos de entrenamiento y el 25% restante para validar el resultado. Los parámetros que obtienen la mejor tasa de acierto de validación son los que se seleccionan.

Tras unas pruebas iniciales donde se probaron valores de los parámetros iguales a $10^{n}$, donde **n** tomaba valores enteros entre -4 y 4, se vio que para los distintos contadores los valores rondaban los valores de C igual a 100 o 1.000 y de γ igual a 10, 100 o 1.000. Estos son los valores que se validan de forma individual para cada contador.

Tras la obtención de los valores de dichos parámetros se realiza un entrenamiento con la totalidad de los datos (limitada a 10.000 datos, como hemos comentado previamente).

La Figura 30 muestra la división del conjunto de original de datos en datos de entrenamiento, validación y test.

**FIGURA 30. DIVISIÓN DEL CONJUNTO DE DATOS EN ENTRENAMIENTO, VALIDACIÓN Y TEST**




## c) Desarrollo e implementación

Al igual que para el desarrollo de las redes neuronales, para las SVM se ha utilizado el lenguaje de programación *Python 2.7*, con *Anaconda*.

Los paquetes utilizados han sido *Scikit-learn*, *NumPy* y *Pyodbc*, descritos en el apartado anterior.

# 5. Resultados

## 5.1. MODELOS DE CONTADORES

En esta sección se resumen y comparan los resultados de la clasificación realizada con los distintos modelos: modelos individuales (contadores de 1 litro y contadores de 0,1 litros) y modelos generales con las dos metodologías utilizadas.

Los modelos generales predicen a partir de la información de todos los contadores, mientras que los individuales tienen en cuenta únicamente la información relativa al contador para el que van a predecir. Al final de esta sección, se incluye una breve comparación de los métodos desarrollados en este proyecto y los métodos estadísticos usados hasta el momento por Canal de Isabel II.

El desarrollo de modelos generales viene motivado porque en un futuro pudiera no disponerse de datos de entrenamiento individualizados por vivienda y sería necesario aplicar clasificadores entrenados con datos de otras viviendas.

Como se ha mencionado anteriormente, se ha limitado a 10.000 el número de datos de las SVM y a 100.000 el de las redes neuronales y un 30% de dichos datos se ha destinado a los test de evaluación.

### 5.1.1. Modelos de contadores de 1 litro

Para la comparación de los resultados en este apartado se han tenido en cuenta un total de 239 contadores de tipo 1 litro. Se han entrenado los algoritmos de clasificación sobre el 70% del histórico de cada contador y se ha llevado a cabo el test sobre el 30% restante.

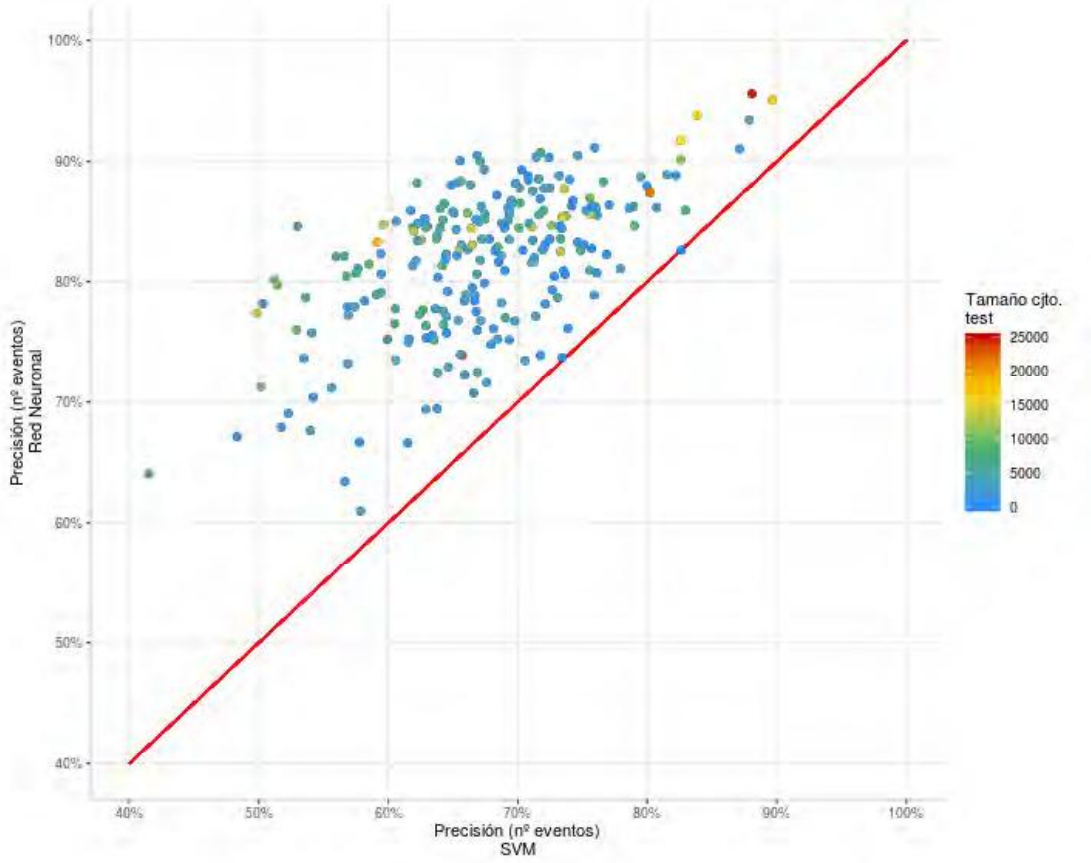
Los resultados obtenidos mediante la Red Neuronal son significativamente mejores que los obtenidos con las SVM, tanto en términos globales como a nivel de contador.

Si se tienen en cuenta todos los resultados obtenidos con ambos métodos de clasificación, es decir, si se suman todos los casos clasificados en todos los conjuntos de test (i.e., todos los contadores de este tipo) se tiene que la precisión media de las SVM es del 67,41%, mientras que la de la Red Neuronal es del 81,78%. Esta precisión ha sido calculada teniendo en cuenta el número de eventos correctamente clasificados. La precisión media en términos de litros correctamente clasificados es muy similar, siendo del 63,41% para las SVM y del 85,76% para la Red Neuronal.

Si se desgranan los resultados a nivel de contador, no es posible encontrar ninguno para el que los resultados de las SVM sean mejores que los de la Red Neuronal. En la Figura 31, cada punto representa uno de los contadores analizados y su posición en el plano es el resultado de enfrentar la precisión obtenida con los dos métodos. Así, si el punto está por encima de la recta de color rojo significa que la precisión obtenida para ese contador con la Red Neuronal es mayor que la obtenida con las SVM. Si está sobre la recta, es que ambas precisiones son iguales. Si estuviera por debajo de la recta, las SVM habrían producido mejores resultados que la Red Neuronal.

El color de cada punto indica el número de eventos con los que se ha ensayado. La mayoría de los conjuntos de test supera los 500 eventos, aunque se han encontrado 11 contadores (aproximadamente el 5% del total de contadores de tipo 1 litro) cuyos conjuntos de test tienen un tamaño por debajo de esa cifra. En este punto podría plantearse la hipótesis de que, a mayor tamaño del conjunto de test, mayor precisión, pero sería necesario comprobarlo con una muestra mayor y con otra distribución de tamaños.

**FIGURA 31. COMPARACIÓN DE LA PRECISIÓN DE LOS CLASIFICADORES. CONTADORES DE PRECISIÓN 1 L**




El número de eventos correctamente clasificados con la Red Neuronal aumenta de media un 22,35%, llegando a superar el 40% de aumento en 19 ocasiones.

### 5.1.2. Modelos de contadores de 0,1 litros

El total de contadores de este tipo que se han tenido en cuenta para la comparación de resultados es de 19. Se han entrenado los algoritmos de clasificación sobre el 70% del histórico de cada contador y se ha llevado a cabo el test sobre el 30% restante.

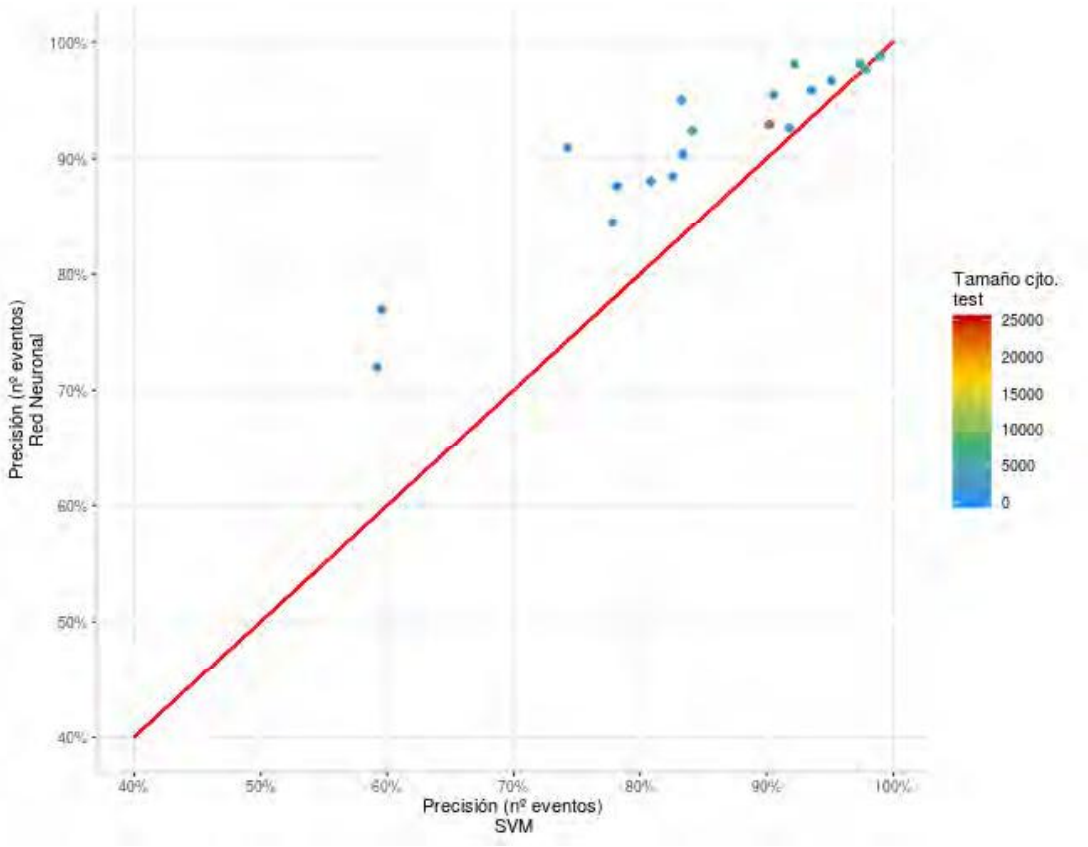
Al igual que ocurría con los contadores de tipo 1 litro, los resultados obtenidos mediante la Red Neuronal son mejores que los obtenidos con las SVM, tanto en términos globales como a nivel de contador.

La precisión media de las SVM y de la Red Neuronal es de 84,78% y de 91,19%, respectivamente. Si se cuantifica el porcentaje de litros correctamente clasificados éstos bajan a 73,5 % y 85,9%.

Desgranando los resultados a nivel de contador, no es posible encontrar alguno para el que los resultados de las SVM sean significativamente mejores que los de la Red Neuronal.

En la Figura 32, cada punto representa uno de los contadores analizados y su posición en el plano es el resultado de enfrentar la precisión obtenida con los dos métodos. Así, si el punto está por encima de la recta de color rojo significa que la precisión obtenida para ese contador con la Red Neuronal es mayor que la obtenida con las SVM. Si está sobre la recta, es que ambas precisiones son iguales. Si estuviera por debajo de la recta, las SVM habrían producido mejores resultados que la Red Neuronal.

**FIGURA 32. COMPARACIÓN DE LA PRECISIÓN DE LOS CLASIFICADORES. CONTADORES DE PRECISIÓN 0,1 L**




El número de eventos correctamente clasificados con la Red Neuronal aumenta de media un 8,5%, llegando a superar el 20% de aumento en 3 ocasiones.

### 5.1.3. Modelos generales

Se entienden por modelos generales aquellos que han sido entrenados con datos de todos los contadores, diferenciando si éstos son de precisión de 1 litro o de precisión de 0,1 litros. Estos modelos se han probado para comprobar si era posible entrenar con datos de otros contadores distintos, sin perder precisión en los resultados.

El límite de entradas para entrenar estos modelos es de 100.000 eventos en el caso de la Red Neuronal, y de 10.000 en el caso de las SVM.

También se limita el número mínimo de tipos de uso distinto en 4, es decir, se tomarán eventos de aquellos usuarios que hayan registrado al menos 4 usos de naturaleza distinta (ej., Duchas, Grifos, Fugas y Lavadora).

La precisión global del método es del 75.18% con las SVM y del 82.17% con la Red Neuronal.

Comparando los modelos individuales con los generales se concluye que los resultados son mejores aplicando los modelos en contadores individuales.

### 5.1.4. Comparación con modelos estadísticos

Este apartado contiene un análisis comparativo de los métodos de clasificación implementados en este proyecto y el método estadístico de clasificación automática existente.

No ha sido posible llevar a cabo un análisis de la precisión de los métodos en la línea que se ha seguido en los apartados anteriores (ofreciendo porcentaje de clasificación correcta en número de eventos y litros), ya que no estaba disponible la clasificación llevada a cabo por un operador en los periodos en los que se ha clasificado con el método estadístico bayesiano.

Dada la distinta naturaleza de los usos disponibles para todos los usuarios, se ha seleccionado una muestra de resultados de los tres métodos comparados (Bayesiano, SVM y RNA) para la evaluación global del funcionamiento a partir de la distribución del volumen asignado a cada tipo de uso. Es decir, se ha seleccionado una muestra y se ha cuantificado el total de litros asignado a cada tipo de uso para compararlo después con la distribución del volumen en los datos que han sido etiquetados por operador.

**Nota:**

*La muestra para la comparación global ha sido extraída sobre los contadores de tipo **1 litro** para evitar el sesgo sobre los eventos de tipo **Grifos** detectado en los contadores de tipo 0,1 litros. También se han eliminado de la muestra los eventos que no están presentes en la mayoría de los usuarios (Piscinas y Riego).*

La Tabla 4 y la Figura 33 muestran la distribución del volumen según el método de clasificación empleado. Cabe destacar que las SVM tienen problemas en la detección de *Fugas y Lavavajillas*.

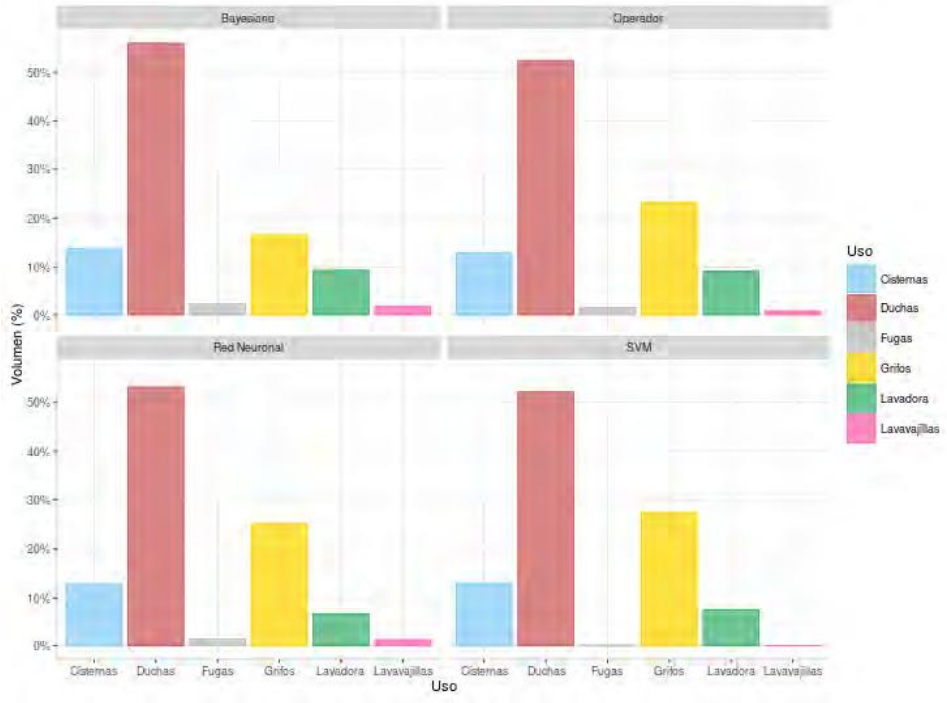
**TABLA 4. DISTRIBUCIÓN DEL VOLUMEN CONSUMIDO SEGÚN EL MÉTODO DE CLASIFICACIÓN EMPLEADO**

| *Uso* | *(%) Volumen operador* | *(%) Volumen Bayesiano* | *(%) Volumen SVM* | *(%) Volumen RNA* | *Media Vol. (litros)* | *Desviación típica Vol. (litros)* |
|---|---|---|---|---|---|---|
| Cisternas | 12,97% | 13,82% | 13,07% | 12,79% | 6,15 | 1,84 |
| Duchas | 52,37% | 55,96% | 52,14% | 53,14% | 21,33 | 32,87 |
| Fugas | 1,47% | 2,51% | 0,08% | 1,34% | 1 | 0,02 |
| Grifos | 23,21% | 16,61% | 27,43% | 25,12% | 3,06 | 3,83 |
| Lavadora | 9,21% | 9,42% | 7,28% | 6,47% | 3,75 | 3,40 |
| Lavavajillas | 0,77% | 1,69% | 0,01% | 1,14% | 1 | 0 |

Las diferencias observadas al comparar los métodos con el operador, deben interpretarse con cautela ya que éstos no han sido ejecutados en los mismos períodos, de modo que parte de la variación podría ser explicada mediante cambios de hábitos cotidianos de los usuarios (aumento o disminución del número de habitantes de la vivienda, un cambio de jornada laboral, etc.).

En términos generales, el método Bayesiano tiende a confundir eventos de tipo*Ducha* con eventos de tipo *Grifos*, mientras que la Red Neuronal clasifica con precisión este tipo de eventos.

**FIGURA 33. DISTRIBUCIÓN DEL VOLUMEN POR TIPO DE USO SEGÚN LOS DISTINTOS MÉTODOS DE CLASIFICACIÓN**




## 5.2. APLICACIÓN INFORMÁTICA

La aplicación implementada integra los diferentes paquetes informáticos, desarrollados en una única aplicación realizada en VBA para*Access*, e incluye todo el procedimiento necesario para identificación de los usos finales del agua en consumos domésticos. El aspecto que presenta el menú principal de esta aplicación se muestra en la Figura 34.

Para un determinado contador el procedimiento sigue los siguientes pasos:

1º) **Transformación de pulsos en caudales e identificación de eventos** A partir de las lecturas de pulsos acumulados, se calculan las series temporales de caudales con intervalos de un segundo. Genera un archivo *Access* con estas series temporales de caudales (episodios de caudal). Partiendo de estos caudales, se identifican y caracterizan los diferentes eventos que conforman los diferentes episodios de caudal, generando una nueva base de datos*Access*, que incluye los parámetros característicos de cada uno de los eventos identificados.

2º) **Etiquetado inicial de entrenamiento**. Asigna a cada evento la etiqueta (uso final) que en su día fue asignada a cada evento identificado por un operador. La metodología utilizada ha sido desarrollada en los Módulos 3 y 4. El resultado de este etiquetado se concreta en un nuevo campo que es añadido a la base de datos de eventos.

3º) **Generación de modelos**. El etiquetado inicial de entrenamiento permite la generación de un modelo individual específico para el contador. Dado que se han desarrollado dos metodologías, una basada en redes neuronales artificiales (RNA) y otra en máquinas de vector soporte (SVM), la aplicación permite generar sendos modelos.

Para nuevas instalaciones, de las que no se disponga de etiquetado inicial con operador, se ha desarrollado una variante de modelos que utilizan los etiquetados iniciales de los contadores que sí disponían de este etiquetado previo para generar los denominados Modelos Generales, uno para cada metodología (RNA o SVM). Estos modelos generales dependen, además de la metodología, de la precisión de los contadores (1 ó 0,1 litros), resultando cuatro tipos de modelos, uno por cada metodología y precisión de contador.

4º) **Clasificación de usos**. Con los modelos generados se procede a clasificar los eventos identificados en el primer paso, asignando una nueva etiqueta según el modelo empleado.

5º) **Informe de resultados**. Por último, se presenta un informe en formato de hoja de cálculo*Excel*, con tablas y gráficos que recogen los datos más relevantes de la clasificación realizada.

**FIGURA 34. APLICACIÓN INFORMÁTICA. MENÚ GENERAL**

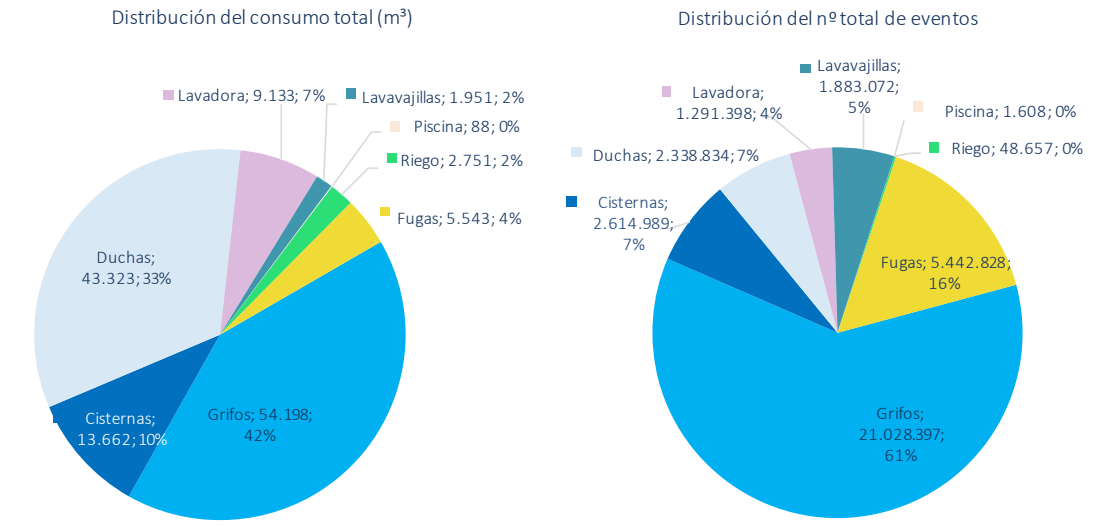
## 5.3. RESULTADOS DE LA CLASIFICACIÓN

Las tablas y gráficos que se incluyen a continuación se refieren a resultados globales, elaborados con la totalidad de los datos registrados, desde enero de 2008 a julio de 2015.

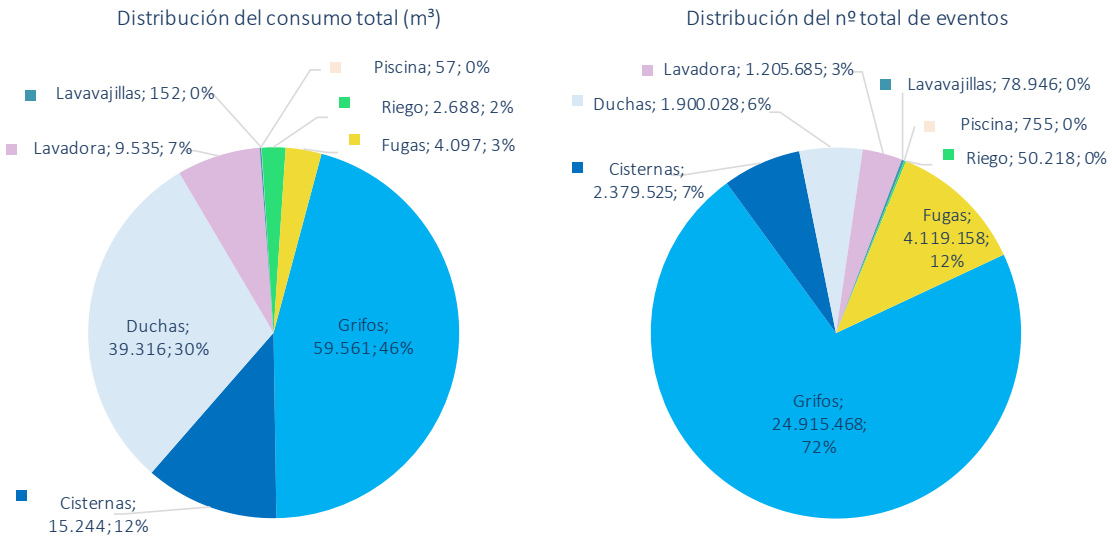
En primer lugar, se incluye una comparación de los resultados obtenidos de las clasificaciones realizadas con ambas metodologías, RNA y SVM, apreciándose ciertas diferencias entre un método y otro. Así, los usos mayoritarios, *Cisternas, Duchas y Grifos*, según la metodología RNA suponen el 10%, 33% y 42%, respectivamente; mientras que según la metodología SVM los valores, para los mismos usos, son del 12%, 30% y 46%, según se puede apreciar en la Figura 35.

**FIGURA 35. COMPARATIVA. CLASIFICACIÓN BASADA EN RNA Y EN SVM DE LA DISTRIBUCIÓN SEGÚN USOS DEL CONSUMO TOTAL, EN EL PERIODO 2008/01 A 2015/07. RESULTADOS GLOBALES DE LA TOTALIDAD DE LOS DATOS PROCESADOS**

Clasificación basada en la metodología **RNA**

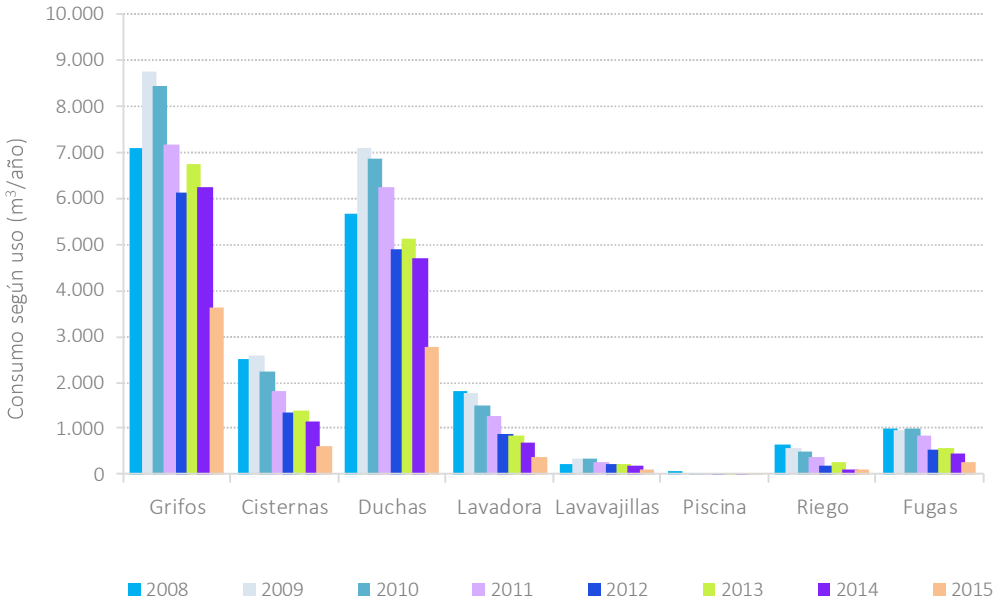



Clasificación basada en la metodología **SVM**



La Figura 36 muestra la evolución de los consumos a lo largo de los años del estudio, según los usos, apreciándose una tendencia a la baja en todos ellos.

**FIGURA 36. RESULTADOS DE LA CLASIFICACIÓN. EVOLUCIÓN DEL CONSUMO DURANTE EL PERIODO 2008 A 2015 (M³/AÑO)**




*Nota: el año 2015 sólo incluye el consumo hasta el mes de julio*

Las figuras 37 y 38 representan estos mismos resultados año a año, indicándose el porcentaje que representa cada uso en el consumo total anual.

**FIGURA 37. RESULTADOS DE LA CLASIFICACIÓN. DISTRIBUCIÓN POR USOS DEL VOLUMEN TOTAL CONSUMIDO EN LOS AÑOS 2008 Y 2009 (M³)**

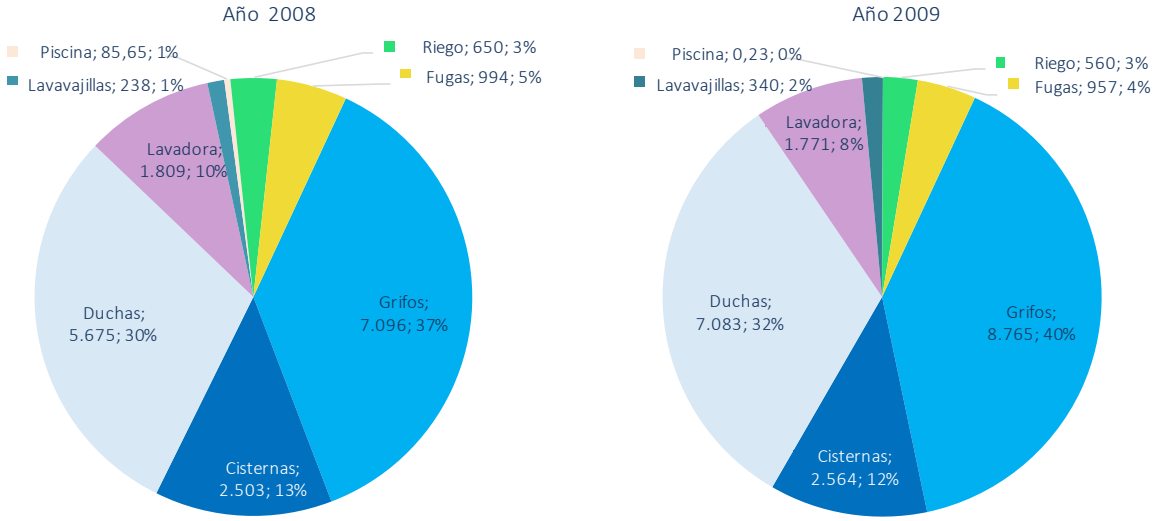
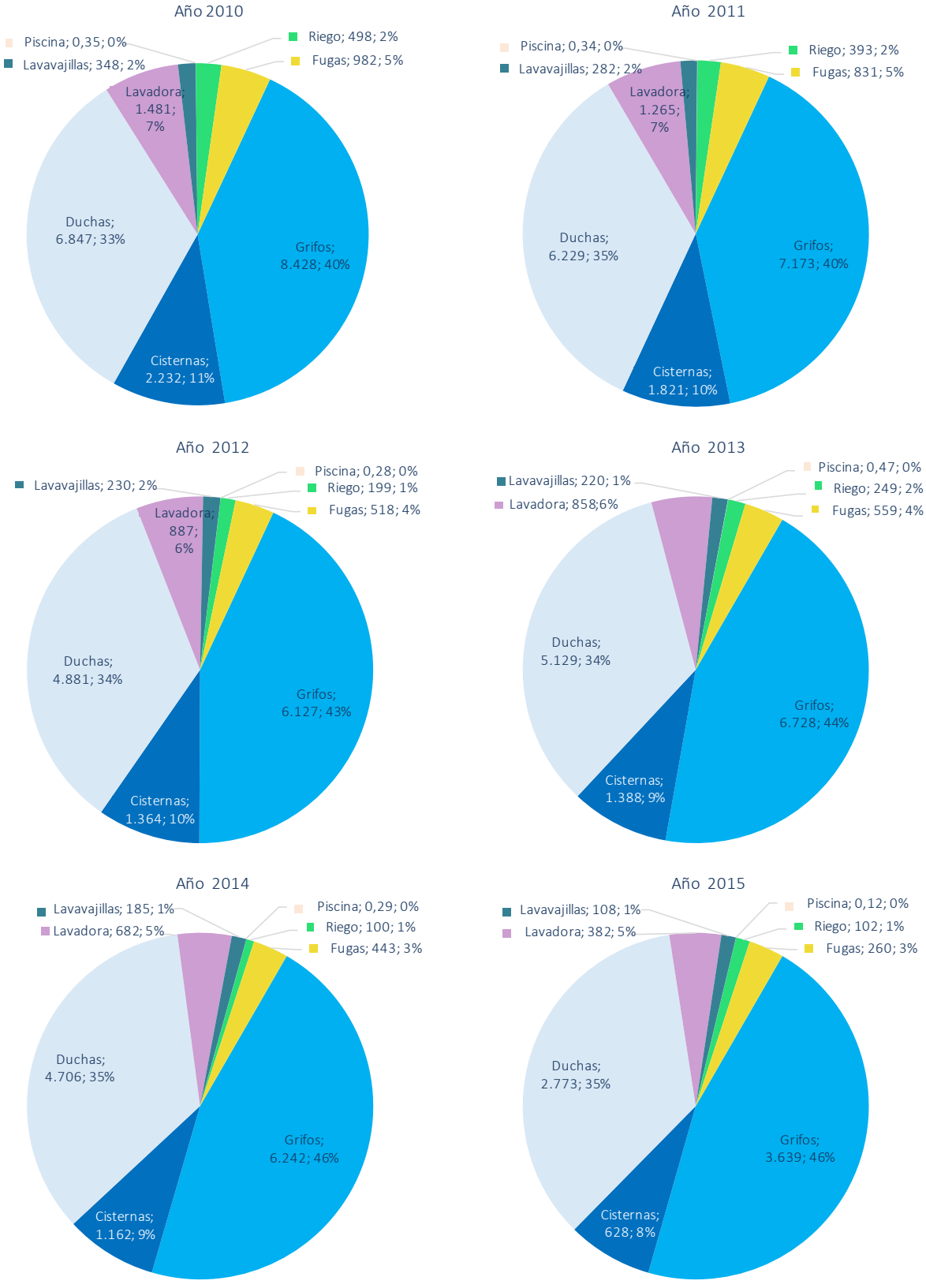


**FIGURA 38.** RESULTADOS DE LA CLASIFICACIÓN. DISTRIBUCIÓN POR USOS DEL VOLUMEN TOTAL CONSUMIDO EN LOS AÑOS 2010 A 2015 ($M^3$)



# 6. Resumen y Conclusiones

En este documento se han presentado los trabajos desarrollados para automatizar, por un lado, el proceso de identificación de eventos de uso de agua en consumos domésticos, a partir de las lecturas de contadores con emisor de pulsos de 1 y 0,1 litros de precisión y, por otro, la clasificación de estos eventos según el uso final del agua.

La identificación de eventos pasa, en primer lugar, por desarrollar un procedimiento matemático que transforme los pulsos registrados por los contadores en caudales consumidos. Este procedimiento se basa en el cálculo de medias móviles de diferentes órdenes por el cual se consigue distribuir a lo largo del tiempo el volumen registrado en un instante determinado (pulso). Los episodios de caudal así obtenidos pueden responder a un único evento o a una combinación de dos o más eventos. Mediante un segundo algoritmo matemático se ha desarrollado una metodología que permite *discretizar* aquellos episodios en eventos, para proceder a su caracterización según determinados parámetros (volumen, caudal máximo, duración total, duración de la rama ascendente, de la rama descendente, etc.).

Sobre la base de un etiquetado previo, realizado por operador, de una serie de eventos registrados en un número importante de instalaciones (375 contadores) y, por comparación con los eventos identificados mediante el algoritmo matemático, se procede a etiquetar estos eventos de manera que sirvan como patrón de aprendizaje para la generación de modelos basados en dos técnicas diferentes: **RNA** (Redes Neuronales Artificiales) y **SVM** (Máquinas de Vector Soporte). Con los modelos generados para cada instalación de medida es posible etiquetar nuevos eventos identificados y caracterizados según los algoritmos matemáticos ideados, a partir de nuevas lecturas de estos contadores.

La metodología desarrollada permite la creación de modelos de redes neuronales o máquinas de vectores soporte para otras instalaciones diferentes de las analizadas en este trabajo, siempre que se cuente con un periodo de entrenamiento de eventos clasificados previamente (de forma manual, por un operador).

Para otras instalaciones, en el caso de que no se cuente de etiquetado previo de eventos mediante operador, se han creado modelos generales a partir de los eventos que sí han sido etiquetados previamente, de manera que aplicando estos modelos generales se puedan clasificar los eventos registrados por estás nuevas instalaciones. Se debe señalar que la precisión obtenida con estos modelos generales es, lógicamente, inferior a la de los modelos desarrollados específicamente para cada contador, y que presumiblemente esta precisión podrá disminuir notablemente si se trasladan estos modelos generales a otro contexto de usos domésticos de agua muy diferentes de la Comunidad de Madrid, donde han sido entrenados.

Para contadores de precisión de 1 litro los resultados obtenidos mediante la Red Neuronal son significativamente mejores que los obtenidos con las SVM, tanto en términos globales como a nivel de contador. En efecto, si se tienen en cuenta todos los resultados obtenidos con ambos métodos de clasificación se tiene que el grado de acierto en la clasificación de las SVM es del 67,41%, mientras que el de la Red Neuronal es del 81,78%. Este grado de acierto ha sido calculado teniendo en cuenta el número de eventos correctamente clasificados. En términos de volumen correctamente clasificado, es muy similar, siendo del 63,41% para las SVM y del 85,76% para la Red Neuronal.

Si se desgranan los resultados a nivel de contador, no es posible encontrar ninguno para el que los resultados de las SVM sean mejores que los de la Red Neuronal.

El número de eventos correctamente clasificados con la Red Neuronal aumenta de media un 22,35%, llegando a superar el 40% de aumento en 19 ocasiones.

Por otro lado, para los 19 contadores de 0,1 litros de precisión analizados, al igual que con los contadores de tipo 1 litro, los resultados obtenidos mediante la Red Neuronal son mejores que los obtenidos con las SVM, tanto en términos globales como a nivel de contador. La precisión media de las SVM (Máquinas de Vector Soporte) y de la Red Neuronal es de 84,78% y de 91,19%, respectivamente. Si se cuantifica el porcentaje en volumen estos valores bajan a 73,5 % y 85,9%.

Asimismo, a nivel de contador, tampoco es posible encontrar ninguno para el que los resultados de las SVM sean significativamente mejores que los de la Red Neuronal.

El número de eventos correctamente clasificados con la Red Neuronal aumenta de media un 8,5%, llegando a superar un aumento del 20% en 3 ocasiones.

Para los modelos generales, la precisión global del método es del 75,18% con las SVM y del 82,17% con la Red Neuronal.

Comparando los modelos individuales con los generales se verifica que los resultados son mejores aplicando los modelos en contadores individuales, como era de esperar.

Todos estos procesos han sido programados y compilados en una aplicación desarrollada al efecto que permite el tratamiento masivo de datos. Los resultados de la clasificación realizada se presentan en forma de tablas y gráficos que resumen los valores más significativos en cuanto a volúmenes, duraciones de los eventos y sus distribuciones mensuales y horarias. Todo ello discriminando según el tipo de uso.

# 7. Pasos siguientes

Como posibles líneas de investigación que den continuidad a los trabajos presentados en este cuaderno se plantean las siguientes:

- Optimización del proceso de etiquetado automático y aplicación a datos masivos. La aplicación informática se ha desarrollado en VBA, sobre *Access*. En aras a una mayor agilidad y capacidad de procesamiento se propone adaptarla a otro sistema con mejores prestaciones, tipo SQL Server u Oracle.

- Evaluación del impacto de campañas de ahorro de agua mediante la herramienta de etiquetado automático, pudiendo analizar de forma objetiva la repercusión de estas campañas en los diferentes usos del agua.

- Aplicabilidad del etiquetado automático a grandes patrones de consumo, para el estudio de patrones de consumo por sectores, barrios o distritos, como herramienta para la actualización de modelos de la red.

- Generación de cuadros de mando integral ligados al etiquetado automático, para obtener informes automatizados de gestión por usos finales del agua.

- Desarrollo del sistema de etiquetado automático como sistema de prelocalización de fugas. Desarrollo de un sistema de alerta temprana de fugas interiores en viviendas como elemento de eficiencia hídrica y aporte de valor al usuario final por el ahorro económico que conlleva.

# ANEXOS

# ANEXO 1. REFERENCIAS BIBLIOGRAFICAS

**Almeida, G.A., Kiperstok, A., Dias, M., Ludwig, O.**

Metodologia para caracterização de consumo de água doméstico por equipamento hidráulico. Anais do Silubesa/ Abes. Figueira da Fo. 2006

**Almeida G., Vieira J., Marques J., Cardoso A.**

Pattern recognition of the household water consumption through signal analysis. Camarinha-Matos L.M. (eds) Technological Innovation for Sustainability. DoCEIS 2011. IFIP Advances in Information and Communication Technology, vol. 349. Springer, Berlin, Heidelberg. 2011

**Barreto, D.**

Perfil do consumo residencial e usos finais da água. Ambiente Construído, Porto Alegre 8(2), 23–40 (2008) ISSN 1678-8621; © 2008, Associação Nacional de Tecnologia do Ambiente Construído, April/June 2008

**Bishop, C. M.**

Pattern recognition and machine learning. Information Science and Statistics, Springer-Verlag New York. Inc. (2006) Secaucus, NJ, USA

**Boser et al., 1992] Boser, B. E., Guyon, I. M., and Vapnik, V. N.**

A training algorithm for optimal margin classifiers. In Proceedings of the fifth annual workshop on Computational learning theory, 1992, pages 144–152. ACM

**Chang, C.-C. and Lin, C.-J.**

(2011). Libsvm: A library for support vector machines. ACM Transactions on Intelligent Systems and Technology (TIST), 2011, 2(3):27

**Chen, T., & Guestrin, C.**

(2016). Xgboost: A scalable tree boosting system. Proceedings of the KDD 2016 conference

**Corona-Nakamura M. A., Ruelas R., Ojeda-Magaña B., Andina D.**

Classification of domestic water consumption using an *Anfis* model. Conference Paper Automation Congress, 2008

**Cubillo, F., Moreno, T., Ortega, S.**

Microcomponentes y factores explicativos del consumo doméstico de agua en la Comunidad de Madrid. Cuaderno de I+D+i nº 4. Canal de Isabel II, 2008, Madrid

**Díaz-Morales, R. and Navia-Vázquez, A**

Improving the efficiency of IRWLS SVMs using parallel Cholesky factorization. Pattern Recognition Letters

**Fernandes, B.C.**

Construção de um Sistema Eletrônico de Monitoramento de Consumo de Água Residencial. Projeto de Graduação apresentado ao Departamento de Engenharia Elétrica. p. 65 Centro Tecnológico da Univ. Federal do Espírito Santo, 2007

**Hsu, C. W., & Lin, C. J.**

A comparison of methods for multiclass support vector machines. IEEE transactions on Neural Networks, 2002, 13(2), 415-425

**Karmarkar, N.**

A new polynomial-time algorithm for linear programming. Proceedings of the sixteenth annual ACM symposium on Theory of computing, 1984, pages 302–311. ACM

**Mayer, P.**

Water energy savings from high efficiency fixtures and appliances in single family homes. USEPA — Combined Retrofit Report 1, 2005

**Nguyen, Zhang, Stewart**

Analysis of simultaneous water end use events using a hybrid combination of filtering and pattern recognition techniques. International Congress on Environmental Modelling and Software (2012)

**Platt, J. et al.**

Fast training of support vector machines using sequential minimal optimization. Advances in kernel methods-support vector learning, 1999, 3

**Pérez-Cruz, F., Bousoño-Calzón, C., and Artés-Rodriguez, A.**

Convergence of the IRWLS procedure to the support vector machine solution. Neural Computation, 2005, 17(1):7–18

**Roger O'Halloran, Michael Best y Nigel Goodman**

Urban Water Security Research Alliance. Technical Report No. 91. 2012

**Vapnik, V.**

The nature of statistical learning theory. Springer Science & Business Media, 2013

**Vapnik, V.**

Pattern recognition using generalized portrait method. Automation & remote control, 1963, 24:774- 780

**Vasak M., Banjac G., Novak H.**

Water use disaggregation based on classification of feature vectors extracted from smart meter data. Procedia Engineering, 119, 1381 – 1390, 3th Computer Control for Water Industry Conference, CCWI 2015

**Vincent, P., Larochelle, H., Lajoie, I., Bengio, Y., & Manzagol, P. A**

Stacked denoising autoencoders: Learning useful representations in a deep network with a local denoising criterion. Journal of Machine Learning Research, 11 (Dec. 2010), 3371-3408

**Zhang, C., & Ma, Y.**

Ensemble machine learning: methods and applications. Springer Science & Business Media. (Eds. 2012)

## ANEXO 2. ÍNDICE DE FIGURAS

## ANEXO 3. ÍNDICE DE TABLAS



Santa Engracia, 125. 28003 Madrid
www.canaldeisabelsegunda.es