



UNIVERSIDAD DE VALLADOLID

E.T.S.I. TELECOMUNICACIÓN

TRABAJO FIN DE GRADO

GRADO EN INGENIERÍA DE TECNOLOGÍAS DE TELECOMUNICACIÓN

**Datos abiertos en el contexto global actual  
y caso práctico enfocado al ámbito sanitario**

Autor:

**D. Javier González Lázaro**

Tutor:

**D. Ignacio de Miguel Jiménez**

Valladolid, 28 de febrero de 2017



---

**TÍTULO: Datos abiertos en el contexto global actual y caso práctico enfocado al ámbito sanitario**

**AUTOR: D. Javier González Lázaro**

**TUTOR: D. Ignacio de Miguel Jiménez**

**DEPARTAMENTO: Teoría de la Señal y Comunicaciones e Ingeniería Telemática**

---

## **TRIBUNAL**

---

**PRESIDENTE: Dña. Patricia Fernández del Reguero**

**VOCAL: D. Ignacio de Miguel Jiménez**

**SECRETARIO: Dña. Noemí Merayo Álvarez**

**SUPLENTE: D. Ramón J. Durán Barroso**

**SUPLENTE: D. Juan Carlos Aguado Manzano**

---

**FECHA: 28 de febrero de 2017**

**CALIFICACIÓN:**

---

## **Resumen del TFG**

Desde los comienzos de Internet y el desarrollo tecnológico de los últimos años, la cantidad de datos generados ha crecido. Aprovechar dichos datos, así como las nuevas oportunidades que ofrecen los avances técnicos para su obtención, permite el nacimiento de una nueva cultura de información y gestión.

Los datos abiertos u *Open Data*, tiene en los gobiernos e instituciones públicas sus principales valedores, que ven en esto una oportunidad para mejorar los servicios y administración de los mismos, utilizando su mayor herramienta, la información.

En un mundo cada vez más conectado, es importante mantener unas reglas adaptadas al momento histórico que sigan protegiendo derechos fundamentales, como el derecho a la privacidad, y permitiendo a su vez, el uso de información para un fin mayor.

En este trabajo, se hablará del contexto global en el que se encuentra el Open Data, así como de la legislación vigente en cuanto a transparencia de las administraciones públicas y del derecho a la privacidad, recogido en la Ley Orgánica de Protección de Datos.

En última instancia, se aplicarán modelos de clasificación estadística a un caso médico, con el fin de, por una parte, obtener un método rápido de diagnóstico y por la otra, y de manera consecuente, ver la importancia y las posibilidades que ofrece el disponer de información para la innovación y mejora de servicios.

## **Palabras clave**

Datos abiertos, administración, legislación, procesamiento de datos, aprendizaje automático, clasificación, R.

## **Abstract**

Since the inception and expansion of the Internet, and technological development, the volume of information and data has grown. Taking advantage of these data, as well as the new opportunities offered by the technical advances to obtain them, allows the birth of a new culture of information and management.

Open Data has in government and in public institutions, its main values, which provides an opportunity to improve services and administration of the same, using its greatest tool, information.

In a connected world, it's important to keep rules that adapt the present with the fundamental rights, such as the right to privacy, while allowing the use of information for a greater purpose.

In this article, we will talk about the global context in which Open Data is located, as well as the current legislation on transparency of public administrations and the right to privacy, as established by Spanish law.

Finally, a few classification models are applied to a medical case, in order to obtain a fast method of diagnosis and consequently, see the importance and possibilities which offers to get information for innovation and improvement of services.

## **Keywords**

Open data, administration, law, data processing, machine learning, classification, R.

*A quien, sin estar, estuvo.  
Y lo hizo posible.  
A todos aquellos que han estado,  
que están y que estarán.*



# ÍNDICE

1	Open Data.....	13
1.1	Contextualización.....	13
1.1.1	Normas Técnicas de Interoperabilidad.....	16
1.2	Contexto legislativo en España .....	18
1.3	Situación global.....	21
1.3.1	Estados Unidos.....	22
1.3.2	Reino Unido .....	23
1.3.3	España .....	24
2	Catálogos de datos y servicios <i>web</i> .....	28
2.1	Estados Unidos.....	29
2.2	Reino Unido .....	30
2.3	Canadá.....	31
2.4	España .....	32
2.5	Otros catálogos de datos.....	34
2.5.1	Unión Europea.....	34
2.5.2	OCDE.....	34
2.5.3	OMS .....	35
2.5.4	Otros catálogos.....	35
2.5.4.a	Temáticas .....	35
2.5.4.b	Competiciones.....	37
2.6	Servicios Web .....	39
2.6.1	CKAN .....	39
2.6.2	Socrata.....	39
2.6.3	Amazon Web Services .....	39
2.6.4	Microsoft Azure .....	40
2.6.5	Google Cloud Platform .....	40
2.6.6	IBM Cloud .....	40
3	Caso Práctico.....	42
3.1	Introducción .....	42
3.2	Contextualización del caso.....	42
3.3	Contextualización de las herramientas empleadas.....	44
3.3.1	Lenguaje R .....	44
3.3.2	R Studio.....	44

3.	Caret .....	48
4	Análisis realizado .....	52
1.	Modelos empleados.....	52
1.	Support Vector Machines (SVM).....	52
a.	Kernel Lineal.....	53
b.	Kernel Radial.....	53
2.	Redes Neuronales.....	53
3.	Regresión logística regularizada .....	54
2.	Resultados obtenidos.....	54
1.	SVM con kernel radial .....	54
2.	SVM con kernel lineal.....	56
3.	Redes Neuronales.....	57
4.	Regresión logística regularizada .....	59
3.	Comparativa entre modelos.....	61
4.	Modelo elegido y <i>test</i> .....	63
5	Conclusiones y líneas futuras .....	66
6	Referencias.....	70



# ÍNDICE DE FIGURAS

Figura 1. Aspectos principales RISP. Fuente: Guía de aplicación de las NTIs.....	13
Figura 2. Grafo de Linked Data hasta marzo de 2009. Fuente: lod-cloud.net/ .....	14
Figura 3. Grafo de Linked Data hasta agosto de 2014. Fuente: lod-cloud.net/ .....	15
Figura 4. Arquitectura de la Web Semántica. Fuente: Semantic Web - XML2000 by Tim Berners-Lee.....	16
Figura 5. Ejemplos de URIs. Fuente: Guía de aplicación de la NTI.....	17
Figura 6. Vocabularios empleados en la generación de URIs. Fuente: Guía de aplicación de NTI..	17
Figura 7. Ránking de los 10 mejores países en Open Data. Fuente: Open Data Barometer, 3ª Edición. ....	21
Figura 8. Puntuaciones Open Data. Elaboración propia. Fuente: Open Data Barometer.....	22
Figura 9. Mapa de iniciativas de datos abiertos. Fuente: Fundación CTIC .....	28
Figura 10. Mapa de iniciativas Open Data. Fuente: open data inception.....	29
Figura 11. Temáticas disponibles en el portal del gobierno de Estados Unidos. Fuente: data.gov...	29
Figura 12. Temáticas relacionadas con la salud. Fuente: healthdata.gov.....	30
Figura 13. Temáticas del portal del gobierno de Reino Unido. Fuente: data.gov.uk.....	30
Figura 14. Formatos y número de datasets disponibles en la web de Reino Unido. Fuente: data.gov.uk .....	31
Figura 15. Frecuencia de actualización de los datos del gobierno de Canadá. Fuente: open.canada.ca .....	31
Figura 16. Categorías disponibles del gobierno de España. Fuente: datos.gob.es .....	32
Figura 17. Opciones disponibles en el portal interactivo referente a población. Fuente: pestadistico.inteligenciadegestion.msssi.es.....	33
Figura 18. Evolución de la población en Castilla y León. Fuente: pestadistico.inteligenciadegestion.msssi.es/publicosns.....	33
Figura 19. Estancia media hospitalaria por países. Fuente: data.oecd.org/healthcare/length-of-hospital-stay.htm .....	35
Figura 20. Representación del funcionamiento de Quandl. Fuente: quandl.com.....	36
Figura 21. Temáticas del proyecto Copernicus. Fuente: copernicus.eu .....	36
Figura 22. Ejemplo de competición para una empresa. Fuente: crowdanalytix.com/community.....	37
Figura 23. Competición de iniciación en Data Science. Fuente: kaggle.com .....	38
Figura 24. Servicios ofrecidos por AWS. Fuente: aws.amazon.com .....	40
Figura 25. Certificaciones de Microsoft Azure. Fuente: azure.microsoft.com .....	40
Figura 26. Fases de la aparición del cáncer. Fuente: aecc.es .....	43
Figura 27. Entorno de RStudio. Ventana principal. ....	45
Figura 28. Panel Packages de RStudio.....	45
Figura 29. Parte izquierda de RStudio, dividida para la generación de scripts. ....	46
Figura 30. Hello World en RStudio. ....	47
Figura 31. Entorno con las variables creadas. ....	47
Figura 32. Ejemplo de plot generado con RStudio. ....	48
Figura 33. La imagen de la izquierda representa un modelo con problema de underfitting (o subajuste), la de la derecha presenta un problema de overfitting o sobreajuste, y la central sería un modelo adecuado.....	49
Figura 34. Izq: Posibles límites. Drcha: límite con support vector machines. Fuente: Applied Predictive Modeling. ....	52

Figura 35. Esquema de funcionamiento de las redes neuronales. Fuente: Applied Predictive Modeling .....	54
Figura 36. Visualización del resultado obtenido en el modelo SVM con kernel radial y parámetros genéricos. ....	55
Figura 37. Resultados obtenidos tras el ajuste de parámetros en el modelo SVM con kernel radial.	55
Figura 38. Resultado obtenido con el modelo SVM con kernel lineal y parámetros por defecto. ....	56
Figura 39. Resultados obtenidos con el modelo SVM con kernel lineal y parámetros ajustados. ....	56
Figura 40. Resultados obtenidos con el modelo de redes neuronales y barrido general de parámetros. ....	57
Figura 41. Resultado obtenido con el modelo de redes neuronales ajustando parámetros. ....	58
Figura 42. Resultados del modelo de regresión logística con penalización, con parámetros por defecto. ....	59
Figura 43. Resultado del ajuste para el modelo de regresión logística con penalización.....	60
Figura 44. Boxplot de los resultados. ....	62
Figura 45. Intervalo de confianza del 95% de los modelos.....	62
Figura 46. Curva ROC del modelo seleccionado. ....	64

# ÍNDICE DE TABLAS

Tabla 1. Modelos empleados y sus parámetros de ajuste.....	50
Tabla 2. Resultados obtenidos con cada modelo analizado. ....	61
Tabla 3. Comparativa de los modelos empleados en términos de precisión. ....	61
Tabla 4. Resultados con los datos de test. ....	63



# 1

## Open Data

### 1. Contextualización

Cuando hablamos de *Open Data* o datos abiertos en castellano, nos referimos principalmente a datos públicos referidos al gobierno, si bien pueden referirse también a otro tipo de información, como puede ser medioambiental, demográfica o social.

Se denominan abiertos ya que pueden ser utilizados, reutilizados y redistribuidos libremente por cualquiera, sea un particular o una empresa, con el objetivo principal de mejorar y optimizar los recursos disponibles, obteniendo beneficio o no por ello, poniéndolos a disposición de los demás, [1].

La Reutilización de la Información del Sector Público (**RISP**) es el principal objetivo de los datos abiertos. Se encarga de proporcionar los datos en bruto y en estándares abiertos que faciliten su reutilización por cualquier persona o empresa interesada en su explotación. Las estrategias **RISP** deben seguir al menos **3 aspectos principales**, visualizados en la Figura 1, [2].

- **Datos.** La materia prima.
- **Sitio web.** Medio principal de difusión.
- **SopORTE.** Elementos de supervisión y actualización de la información.

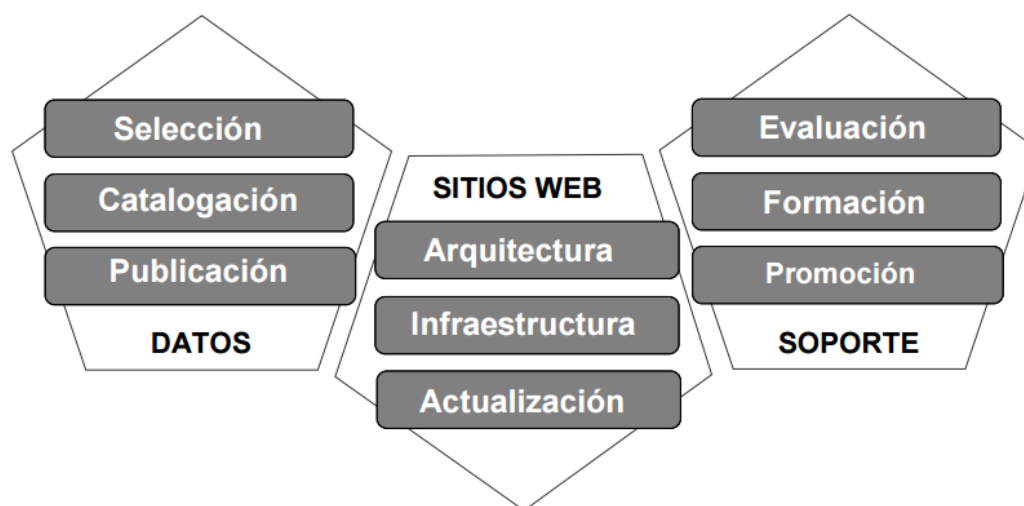


Figura 1. Aspectos principales RISP. Fuente: Guía de aplicación de las NTIs.

Los datos abiertos se centran en **información de carácter no personal**, es decir, que no identifiquen a un individuo unívocamente. Esta matización es importante, en muchos países se protegen los datos personales de sus habitantes que, en el caso de España, se recoge en la **LOPD** (Ley Orgánica de Protección de Datos), de la que hablaremos más adelante, [1].

Para garantizar la apertura y el acceso a los datos mediante estándares abiertos la *World Wide Web Consortium (W3C)* propone estándares a seguir. En el caso de datos abiertos se propone *Resource Description Framework (RDF)*, *Web Ontology Language (OWL)* y *SPARQL Protocol and RDF Query Language (SPARQL)*.

**RDF** es un marco para expresar información relacionada con recursos web, de manera que la web no sea solamente entendida por un lector humano, sino también por una máquina. Se trata pues, de adaptar la web actual hacia una **web Semántica** [3]. Se trata de no perder el significado de las palabras y enlazar (**Linked Data**) los recursos relacionados entre sí. A modo de ejemplo, en la Figura 2 podemos ver los diversos conjuntos de datos o *datasets* en formato *Linked Data* hasta marzo de 2009 y en la Figura 3 hasta agosto de 2014 [4].

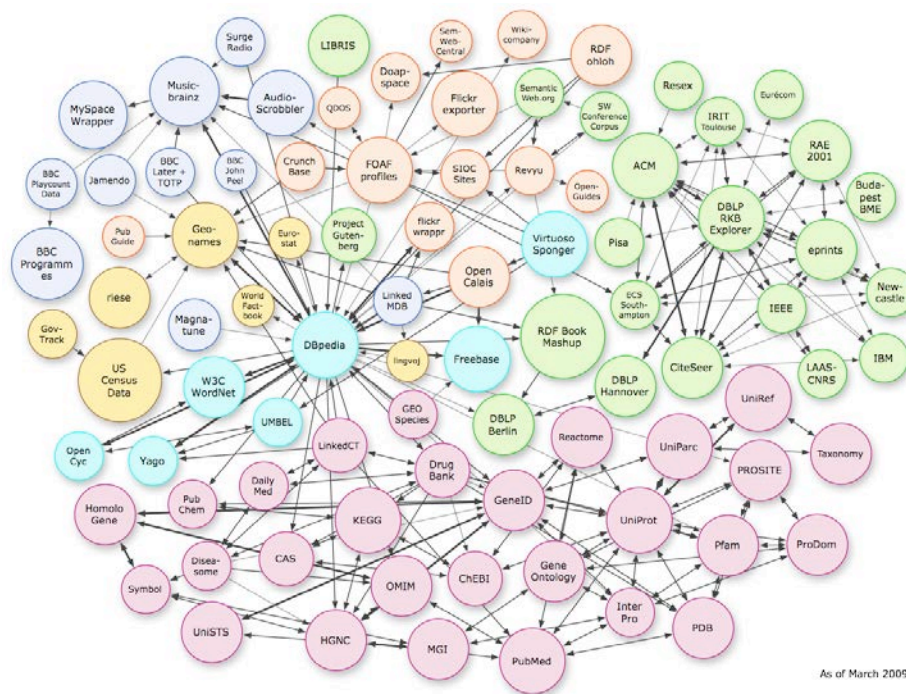


Figura 2. Grafo de Linked Data hasta marzo de 2009. Fuente: [lod-cloud.net/](http://lod-cloud.net/)

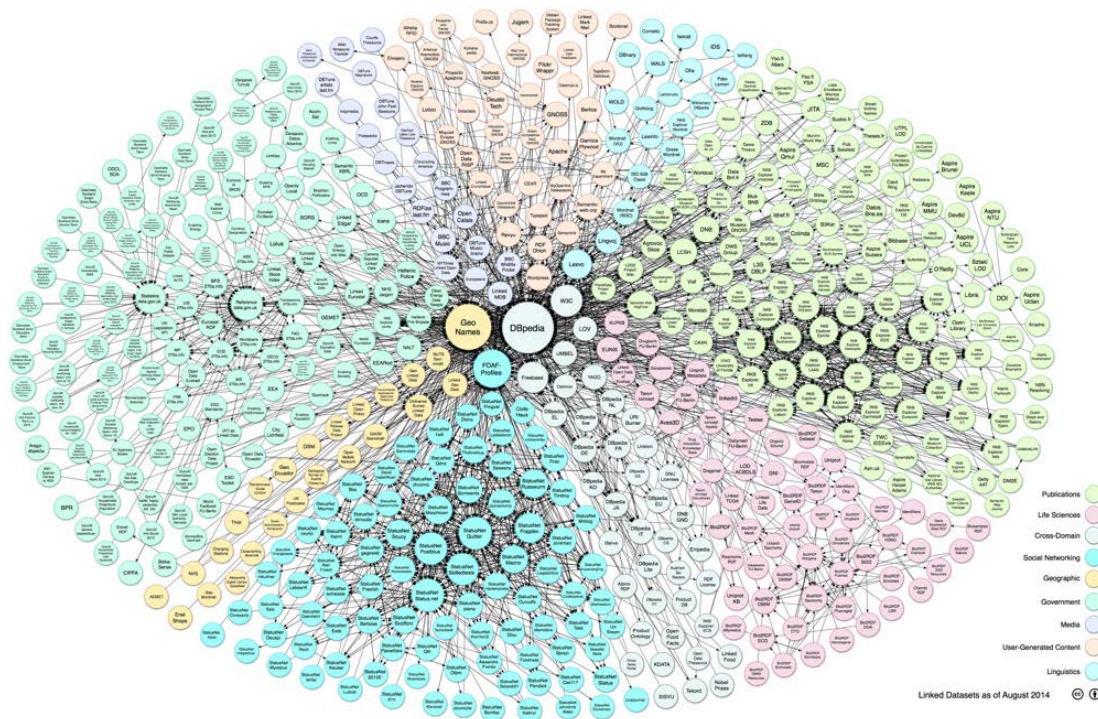


Figura 3. Grafo de Linked Data hasta agosto de 2014. Fuente: lod-cloud.net/

**OWL** está diseñado para que lo usen las máquinas a la hora de procesar recursos web. Facilita la interoperabilidad de contenido, añadiendo funcionalidades a XML, RDF y *RDF-Schema*. Se puede emplear para representar explícitamente el significado en contexto de los términos y las relaciones entre ellos. A estas relaciones entre términos se les denomina **ontología**<sup>1</sup>, [5]. Actualmente, el lenguaje **OWL** se encuentra en su segunda versión, llamada informalmente **OWL 2**, [6].

**SPARQL** es el lenguaje de consultas para RDF (*SPARQL Query Language for RDF*) y un protocolo de recuperación (*SPARQL Protocol for RDF*).

Previamente se menciona la **Web Semántica**, iniciativa de Tim Berners-Lee, creador del *World Wide Web (WWW)*, e impulsada por el **W3C** para establecer los estándares a emplear. Se busca crear una web más inteligente, que no solo liste resultados, sino que comprenda y entienda su significado. También se pretende que cada recurso posea un identificador único representado por su *Uniform Resource Identifier (URI)* que lo diferencie del resto de recursos web. Estos recursos estarán enlazados con otros empleando la capa de metadatos de **RDF** y a su vez, serán representativos para el usuario. Es decir, para el usuario será transparente estas capas de metadatos que vemos en la Figura 4, pero la **URI** del recurso será representativa y unívoca para facilitar la navegación. En España, por ejemplo, existe un Esquema Nacional de Interoperabilidad (**ENI**) [7] que tienen que seguir todas las Administraciones Públicas (AA. PP en adelante) para publicar y preservar la información en la web. Se establecen unas Normas Técnicas de Interoperabilidad (**NTI**) [8] en las que se describe detalladamente cómo se debe actuar. En la sección 1 desarrollaré más en profundidad las **NTIs**.

<sup>1</sup> Ontología: “Una ontología define los términos y las relaciones básicas para la comprensión de un área del conocimiento, así como las reglas para poder combinar los términos para definir las extensiones de este tipo de vocabulario controlado”, [13].

Continuando con la Web Semántica, muestro a continuación en la Figura 4 la arquitectura propuesta por *Tim Berners-Lee* en la que debemos observar que las capas con datos y vocabulario, van firmadas digitalmente y que, adicionalmente, hay una capa de confianza. Tiene su importancia ya que, si vamos a enlazar información (de cualquier tipo, no solo documentos), tenemos que asegurar que esos enlaces son fiables y correctos.

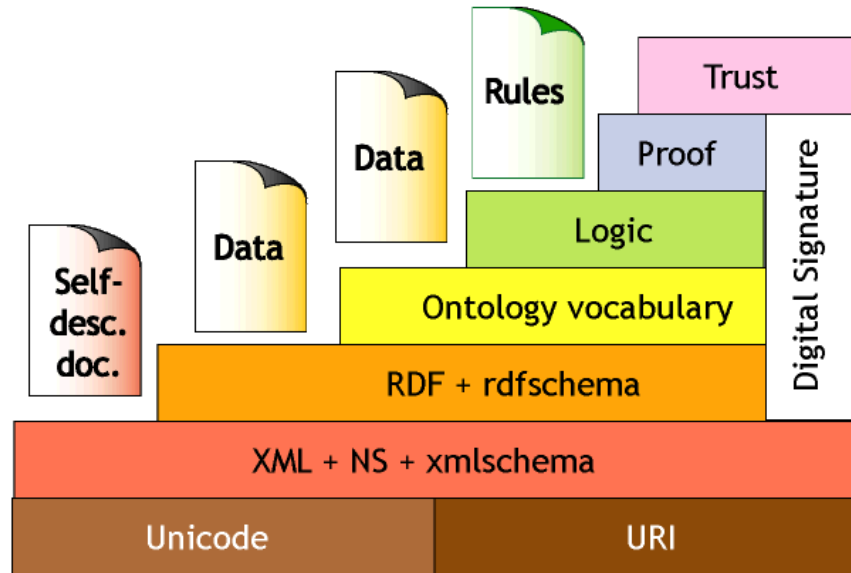


Figura 4. Arquitectura de la Web Semántica. Fuente: *Semantic Web - XML2000* by *Tim Berners-Lee*.

## 1. Normas Técnicas de Interoperabilidad

Las **NTIs** se desarrollan para cubrir las necesidades normativas para garantizar la interoperabilidad entre A.A. P.P, dotando de una serie de directrices y de requisitos mínimos a cumplir por todas las administraciones. No es en ningún caso una metodología de publicación de datos ni unas especificaciones sobre los requisitos funcionales del *software*, aunque sea importante tenerlo en cuenta a la hora de desarrollarlo.

A la hora de identificar los recursos de información, las **NTIs** dan una serie de directrices para la correcta identificación, creación y gestión de referencias y recursos. Se establece que las **URIs** deben ser unívocas e independientes de las tecnologías empleadas, de construcción común y coherente con el recurso referenciado, con información auto contenida y no deben dejar de existir y en caso de modificación o cambio, es la administración la encargada de referenciar y enlazar el contenido de forma correcta [2].

En el paradigma de la Web Semántica, los *Linked Data* deben tener **3 URIs**: Una **abstracta** que identifique el recurso de manera conceptual, otra **legible** para una persona (documento HTML) y una última de **descripción semántica** [2].

El esquema general para la construcción de **URIs** es el siguiente [2]:

**http://{base}/{carácter}/{sector}/{dominio}/{concepto}].[ext]**

Alternativamente, utilizando los identificadores de fragmento mediante la marca “#” al final de la dirección:

**http://{base}/{carácter}/{sector}/{dominio}].[ext]#{concepto}**



Como ejemplo la Figura 5 muestra ejemplos ficticios de cómo se construyen las URIs [2].

EJEMPLO	URI
<b>Catálogo de datos</b>	<i>http://{base}/catalogo</i> http://datos.gob.es/catalogo
<b>Conjunto de datos</b> correspondiente al listado de los centros sanitarios en el año 2011	<i>http://{base}/catalogo/{dataset}</i> http://datos.gob.es/catalogo/centros-sanitarios-2011
alternativa usando un URI basado en identificadores de fragmento (#)	<i>http://{base}/catalogo#{dataset}</i> http://datos.gob.es/catalogo#centros-sanitarios-2011
<b>Vocabulario</b> de centros sanitarios	<i>http://{base}/def/{sector}/{dominio}</i> http://datos.gob.es/def/salud/centros-sanitarios
<b>Clase</b> hospital definida en el vocabulario de centros sanitarios	<i>http://{base}/def/{sector}/{dominio}/{Clase}</i> http://datos.gob.es/def/salud/centros-sanitarios/Hospital
alternativa usando un URI basado en identificadores de fragmento (#)	<i>http://{base}/def/{sector}/{dominio}#{Clase}</i> http://datos.gob.es/def/salud/centros-sanitarios#Hospital
<b>Propiedad</b> especialidad clínica definida en el vocabulario de centros sanitarios	<i>http://{base}/def/{sector}/{dominio}/{propiedad}</i> http://datos.gob.es/def/salud/centros-sanitarios/especialidad
alternativa usando un URI basado en identificadores de fragmento	<i>http://{base}/def/{sector}/{dominio}#{propiedad}</i> http://datos.gob.es/def/salud/centros-sanitarios#especialidad

Figura 5. Ejemplos de URIs. Fuente: Guía de aplicación de la NTI.

Para simplificar la notación se siguen los vocabularios que establece la W3C [2]. Figura 6.

VOCABULARIO	ABR. URI
XML Schema	<b>xsd:</b> http://www.w3.org/2001/XMLSchema#
Simple Knowledge Organization System (SKOS)	<b>skos:</b> http://www.w3.org/2004/02/skos/core#
Dataset Catalog (dcat)	<b>dcat:</b> http://www.w3.org/ns/dcat#
Dublin Core Terms	<b>dct:</b> http://purl.org/dc/terms/
Dublin Core Elements	<b>dc:</b> http://purl.org/dc/elements/1.1/
W3C Time Ontology	<b>time:</b> http://www.w3.org/2006/time#
Friend Of A Friend (FOAF)	<b>foaf:</b> http://xmlns.com/foaf/0.1/

Figura 6. Vocabularios empleados en la generación de URIs. Fuente: Guía de aplicación de NTI.

## 2. Contexto legislativo en España

La regulación existente a día de hoy en España en materia de *Open Data* se agrupa en tres niveles diferentes: **comunitario**, **estatal** y **autonómico**.

En lo que respecta a las normas de carácter comunitario, podemos decir que existen **dos directivas comunitarias** en este aspecto. Para comprender bien la forma en que la regulación comunitaria tiene influencia en nuestro ordenamiento jurídico, es necesario destacar qué son las directivas europeas.

Según la propia definición proporcionada por la Unión Europea en su página web, una directiva es uno de los instrumentos jurídicos que utiliza el organismo comunitario (junto con las recomendaciones, los dictámenes y reglamentos) para que las políticas que se crean sean de aplicación en los países miembros.

El proceso es el siguiente:

1. La Unión Europea emite una directiva en la que obliga a que los países logren un resultado concreto, pero les deja margen de maniobra para que lo lleven a cabo de la forma que consideren oportuna.
2. Mediante la transposición<sup>2</sup> de la misma a la legislación nacional. Hasta que no se transpone la directiva mediante una norma propia del ordenamiento jurídico del país en el que se está incorporando, no es de obligado cumplimiento, es decir, que si alguien actúa incumpliendo lo que la Unión Europea dice que se ha de cumplir, no puede ser sancionado [9].

A continuación, aparecen recogidas las dos directivas europeas que existen en este ámbito:

- **DIRECTIVA 2003/98/CE DEL PARLAMENTO EUROPEO Y DEL CONSEJO** de 17 de noviembre de 2003 relativa a la reutilización de la información del sector público. Esta norma ha sido incorporada al ordenamiento jurídico español mediante la **Ley 37/2007**, de 17 de noviembre, sobre reutilización de la información del sector público, [10].

En rasgos generales, de esta directiva podemos destacar que tiene como objetivos:

- En lo relativo al término reutilización, establece la regulación legal básica que se debe aplicar a la misma, estableciendo además una serie de organismos destinados a que se lleve a cabo dicha reutilización.
- **DIRECTIVA 2013/37/UE DEL PARLAMENTO EUROPEO Y DEL CONSEJO** de 26 de junio de 2013 por la que se **modifica** la **Directiva 2003/98/CE** relativa a la reutilización de la información del sector público, [11]. Esta directiva se transpone al ordenamiento jurídico español mediante la **Ley 18/2015**, de 9 de julio, por la que se **modifica** la **Ley 37/2007**, de 16 de noviembre, sobre reutilización de la información del sector público. De esta directiva lo más destacable es que realiza numerosas modificaciones sobre la anterior Directiva que existía en este aspecto, y de dichos cambios los más destacables son:
  - La obligación de los estados de reutilizar la información de la que disponen sus Administraciones Públicas.
  - La posibilidad de revisar la negativa que una Administración Pública tome acerca de liberar datos. Para esta revisión, actuará un órgano imparcial, que decidirá acerca de

---

<sup>2</sup> Transposición: Proceso que consiste en crear la norma propia que regule la forma en la que se cumple la directiva europea.

la conveniencia o no de liberar esos datos. Las decisiones que estos organismos de revisión tomen serán vinculantes para la Administración.

En cuanto al segundo nivel, relativo a las normas de carácter **nacional** en materia de datos abiertos, la legislación más importante en este aspecto es la siguiente:

- **Ley Orgánica 15/1999**, de 13 de diciembre, de Protección de Datos de Carácter Personal (LOPD), de la que se hablará más adelante, [12].
- **Ley 37/2007**, de 17 de noviembre, sobre reutilización de la información del sector público, [13], que establece como objetivos fundamentales:
  - La definición, en un primer momento del término **reutilización**, que para la directiva europea consiste en la utilización de la información que las diferentes Administraciones Públicas y los organismos que de ellas dependen obtienen en el ejercicio de sus funciones, para fines diferentes que los que derivan de dichas funciones públicas.
  - La armonización de la **explotación de la información** de la que dispone el sector público, con el objetivo de que se potencie su uso y la eficacia de su empleo sea mayor para los ciudadanos.
  - **Igualdad de condiciones** para quien solicitase la reutilización de la información pública.
- **Ley 18/2015**, de 9 de julio, por la que se **modifica** la **Ley 37/2007**, de 16 de noviembre, sobre reutilización de la información del sector público, [14]. Esta Ley realiza importantes modificaciones sobre la anterior, y se hace necesaria tras aprobarse la **Directiva 2013/37/UE** del Parlamento Europeo y del Consejo, de 26 de junio de 2013 (de la que ya se ha hablado con anterioridad). De entre estas modificaciones, las más importantes son:
  - La **obligación** de las Administraciones de autorizar la **reutilización** de los datos de los que dispone (salvo aquellos que traten de materias excluidas en la Directiva del año 2013).
  - La mejora de la forma en que los datos se ponen a disposición de los ciudadanos, fomentando en la medida de lo posible su acceso en **formatos abiertos y legibles por máquina** junto con sus metadatos.
  - En lo relativo a las tarifas que se derivan de la reutilización de datos, se impulsa una mayor **transparencia**.
  - La obligación de que las licencias para utilizar este tipo de datos sean **licencias abiertas**.
  - La obligación para todos los Estados miembros de emitir un informe cada tres años dirigido a la Comisión Europea, acerca de la utilización que se está dando de la reutilización de datos.
- **Real Decreto<sup>3</sup> 1495/2011**, de 24 de octubre, por el que se **desarrolla** la **Ley 37/2007**, de 16 de noviembre, sobre reutilización de la información del sector público, para el ámbito del sector público estatal, [15]. Forma parte del conjunto de medidas que se crean en el marco del **Plan Avanza 2**, concretamente de la Estrategia llevada a cabo entre los años 2011 y 2015. Como aspectos más destacables de dicho Real Decreto, podemos citar:
  - Un principio general que establece que, cuando hablamos de actuaciones del sector público estatal, la norma establece la autorización para que se puedan reutilizar los

---

<sup>3</sup> Real Decreto: Norma con rango inferior al de la ley, que es elaborada por el poder ejecutivo, es decir, por el Gobierno, y no por las Cortes.

documentos que han sido elaborados o se encuentren en manos del ente público. Esto siempre respetando lo que dispongan otras disposiciones legales en cuanto al acceso de este tipo de información, que puede estar limitado, sin que este Real Decreto exceptúa la limitación.

- Se recoge la intención de mantener un catálogo que incluya la información pública que pueda ser reutilizada, con el objetivo de que se pueda acceder a cualquier información desde un mismo punto.
- Se impone como norma general la reutilización, o la modalidad que la facilite en mayor medida.

En el punto anterior se menciona el **Plan Avanza 2**, continuación del **Plan Avanza** que busca a través del uso de las Tecnologías de la Información y la Comunicación (**TIC**) cambiar el modelo económico para basarlo en el aumento de la productividad y competitividad. Con el Plan Avanza 2 se establece una estrategia para llevar a cabo las iniciativas para este cambio de modelo, podemos enumerar los ejes principales de actuación [16]:

- **Desarrollo TIC**
- **Capacitación TIC**
- **Servicios Públicos Digitales**
- **Infraestructuras**
- **Confianza y Seguridad**

En cuanto a la legislación **autonómica**, voy a hacer referencia a la legislación existente en Castilla y León, por ser la de ámbito más cercano a nuestro interés.

- **LEY 3/2015**, de 4 de marzo, de Transparencia y Participación Ciudadana de Castilla y León [17]. Como objetivos principales de la misma cabe destacar:
  - La **regulación** de la **transparencia** en las actuaciones de los organismos públicos, creándose para ello el **Portal de Gobierno Abierto**, que aglutina toda la información para hacer más accesible el acceso a la misma.
  - A través de ese mismo portal se permite la **participación ciudadana** en asuntos públicos.

El **Portal de Gobierno Abierto** de Castilla y León trata de poner en práctica la filosofía de datos abiertos. Para ello, sigue 3 principios o pilares fundamentales, como son, la **transparencia**, la **participación** y la **colaboración** que son indispensables en ese orden, pues, sin transparencia y acceso a los datos, no puede haber participación y mucho menos colaboración. Estos principios están condicionados a 2 factores, la **apertura de datos** y a la **comunicación** y **explicación** de los datos, para hacerlos comprensibles [18].

### 3. Situación global

Las iniciativas de datos abiertos no son algo nuevo, ya algunos países como Reino Unido y Estados Unidos, entre otros, comenzaron a publicar sus datos en 2009 [1]. Si bien es cierto que su crecimiento y expansión han ido en aumento en los últimos años. Según revelan los datos del tercer barómetro de *open data*, hay más países a lo largo del mundo entrando con fuerza en la filosofía de datos abiertos, estos países son: Francia, Canadá, Corea, México, Uruguay y Filipinas [19]. En la Figura 7 podemos ver el ránking mundial en preparación para el *Open Data*, así como su implementación [20]. Como podemos observar, la mayoría de los países son europeos.

En las dos primeras columnas vemos la posición y país que la ocupa, respectivamente. Las cuatro columnas más a la derecha son las puntuaciones obtenidas, siguiendo los criterios de metodología del tercer barómetro de datos abiertos, [20]. La columna **Score** (puntuación) refleja la puntuación total obtenida por el país acorde a los porcentajes aplicados al resto de columnas. La siguiente columna, denominada **Readiness** (preparación), comprende los datos (primarios y secundarios) de las políticas del gobierno y sus acciones además de las empresas y ciudadanos (y sociedad civil). En esta columna han añadido preguntas respecto al barómetro anterior que busca evaluar la preparación del país para cumplir los principios establecidos por la *Open Data Charter*, del que hablaremos más adelante. La siguiente, **Implementation** (implementación) comprende las evaluaciones de los conjuntos de datos en lo referido a responsabilidad, innovación y política social. Por último, **Impact** (impacto) se refiere a los datos primarios, a nivel político, económico y social [21].











Position	Country	Score	Readiness	Implementation	Impact
1	 UK	100	100	100	100
2	 USA	81.89	97	76	76
2	 France	81.65	97	76	74
4	 Canada	80.35	89	84	67
5	 Denmark	76.62	77	77	78
6	 New Zealand	76.35	87	62	87
7	 Netherlands	75.13	90	69	70
8	 Korea	71.19	95	64	58
9	 Sweden	69.26	88	60	64
10	 Australia	67.99	84	77	39
Average top 10		78.04	90.04	74.50	71.30

Figura 7. Ránking de los 10 mejores países en Open Data. Fuente: Open Data Barometer, 3ª Edición.

La tendencia global es de aumento la cantidad y calidad de datos abiertos, aunque en algunos países ha descendido en los últimos años. En la Figura 8 podemos ver un ejemplo con los países líderes mostrados en la Figura 7 junto con los emergentes mencionados anteriormente y con España. Los datos mostrados en la gráfica de la Figura 8 comprenden los barómetros de los años 2013 al 2015 de los países citados, [22]. De manera global, observamos la tendencia alcista.

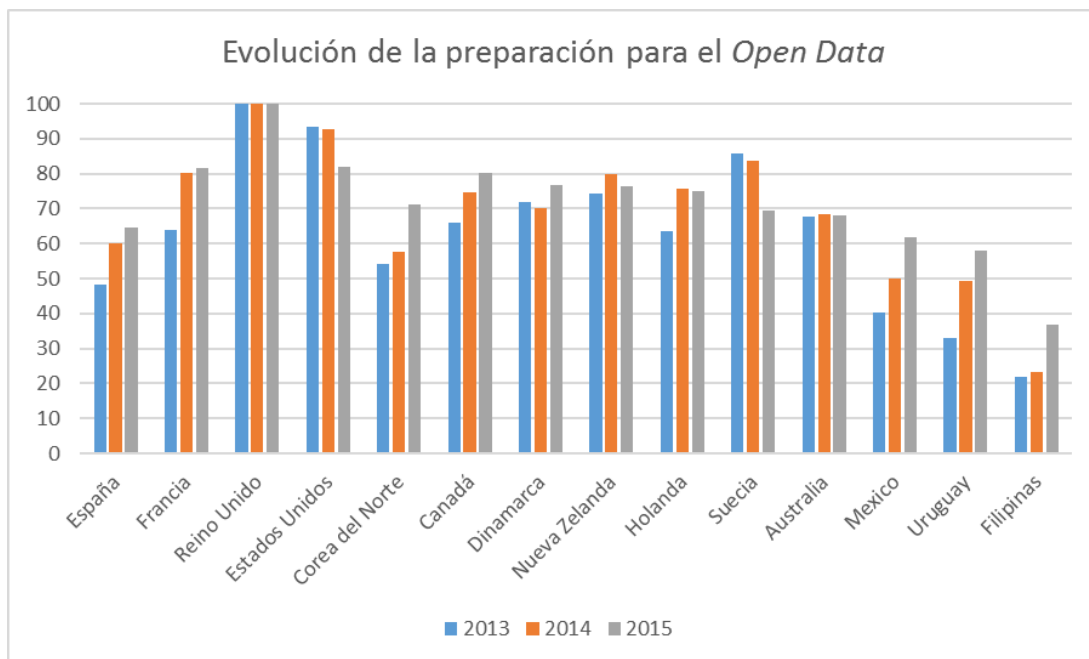


Figura 8. Puntuaciones Open Data. Elaboración propia. Fuente: Open Data Barometer.

En los apartados siguientes, veremos con más detalle alguno de los países referencia en *Open Data* como son **Estados Unidos**, primer país en lanzar una iniciativa *Open Government Data* (UGD) [23], **Reino Unido**, como referente europeo en estas iniciativas y **España**, que, por ser el país de referencia en el presente trabajo, contará con un análisis más amplio.

## 1. Estados Unidos

Como ha sido mencionado antes, Estados Unidos es el precursor de las iniciativas de *Open Government Data* con la publicación en diciembre de 2009 del “*Memorandum for the heads of executive departments and agencies*” [24] en el que se recogen los principios de **transparencia**, **participación** y **colaboración** como las bases de un gobierno abierto, en el que los ciudadanos conocen qué hace el gobierno, aportan su propio conocimiento para la mejora la eficiencia de los estamentos estatales y fomenta la colaboración entre los distintas administraciones estatales y a su vez, con iniciativas privadas. Todo ello enfocado a mejorar los servicios y el funcionamiento de la sociedad.

Para llevar a cabo esta tarea, el *Memorandum* establece 4 principios de actuación:

- Publicar la información gubernamental *online*.  
Cada administración debe publicar los datos (siempre que sea posible) para hacerlos accesibles y crear así oportunidades nuevas.
- Mejorar la calidad de la información gubernamental.  
Siguiendo las directrices de la Oficina de Administración y Presupuestos (“*Office of Management and Budget*”), se establecen una serie de criterios de calidad de información.
- Crear e institucionalizar una cultura de gobierno abierto.  
Para garantizar la cultura de gobierno abierto es necesario tener como base, los principios de **transparencia**, **participación** y **colaboración** dentro de la administración pública. Para ello es necesario un equipo multidisciplinar amplio que trabaje unido.

- Crear y habilitar un marco de políticas para el gobierno abierto.  
Cada vez son más las tecnologías que permiten más y mejor comunicación y por tanto se hace necesaria una política adaptable a estas tecnologías.

A nivel legislativo, el congreso de Estados Unidos, aprobó una ley de transparencia y responsabilidad (*DATA act.*) [25] para mejorar la disponibilidad y calidad de los datos publicados en relación al gasto público [26].

Al hablar sobre datos abiertos, se abre la **discusión** sobre los **datos de carácter personal**. En el caso de Estados Unidos encontramos la “política de divulgación inteligente” (*Smart Disclosure Policy*) [27] en la que se comenta el uso de datos personales. Aseguran ser un beneficio para el consumidor, ya sea con una bajada en la tarifa eléctrica o en la contratación de servicios financieros y, además, se garantiza la privacidad y la seguridad del consumidor y de sus propios datos.

## 2. Reino Unido

Dentro del territorio europeo, cabe destacar la actuación del Reino Unido en esta materia, que ya en el año 2012 publicó el *Open Data White Paper* [28] en el que se presupone el deber del gobierno de publicar los datos. En esta línea, se establecen también unos principios a seguir dentro del servicio público [29].

Un año después, en 2013, el Reino Unido se adhiere al *Open Data Charter* para establecer, como principal objetivo, que los datos sean abiertos por defecto [26].

El *Open Data Charter* [30] describe una serie de principios de actuación que los gobiernos que forman el G8 (grupo de economías más industrializadas del planeta, compuesto por Estados Unidos, Rusia, Canadá, Francia, Italia, Alemania, Reino Unido y Japón) para dar acceso, uso y reutilización de sus datos. Se recogen los cinco principios:

- Datos abiertos por defecto  
Dar acceso gratuito y fomentar la reutilización de datos como valor añadido para la sociedad y la economía.
- Calidad y cantidad  
La información es importante para el ciudadano, por ello se hace necesario tener unos criterios de cantidad y sobretodo de calidad. Los datos han de ser **comprensibles** y **precisos** además de estar disponibles con la mayor descripción lo antes posibles.
- Aprovechables por todos  
Los datos tienen que ser accesibles, reutilizables y gratuitos para todos y sin barreras burocráticas. Además, y en la medida de lo posible, han de estar en formatos abiertos. Para ciertos datos, los que no son de acceso gratuito, se promoverá la obtención de beneficio.
- Liberalización de datos para mejorar la gobernanza  
Se busca con estas iniciativas mejorar las políticas y hacerlas más eficientes y adaptadas a las necesidades de los ciudadanos.
- Liberalización de datos para innovación.  
Haciendo accesible los datos a más gente se promueve la creatividad y la innovación para producir mayores beneficios para la sociedad.

En materia legislativa, el Comité selectivo de la administración pública (*Public Administration Select Committee, PASC*) elaboró un informe en el que pone de manifiesto la importancia social y económica de los datos abiertos. En ese informe, la empresa *Deloitte* cifra en £1.8 billones el valor de la

información del sector público para los consumidores, empresas y el propio sector, en los años 2011-2012 [31].

En cuanto a los datos de carácter personal, en Reino Unido existe una plataforma web (*midatalab*) para acelerar el crecimiento de empresas, cediendo los datos personales, si bien aseguran estar protegidos y asegurados [32]. De igual manera que en otros países, cuenta con la legislación pertinente para garantizar la protección de la información de sus ciudadanos (*Data Protection Act 1998* [33]).

### 3. España

En España, el *open data* comienza en 2009 impulsado por el Ministerio de Industria, Energía y Turismo (que tenía las competencias en agenda digital y TIC) a través de *Red.es* y con la colaboración del Ministerio de Hacienda y Administraciones Públicas tras trasponerse en la Ley 37/2007 (ya mencionada) la Directiva 2003/98/CE sobre la Reutilización de la Información del Sector Público. Se crea para ello el **Proyecto Aporta** (enmarcado en el **Plan Avanza 2**) para situar a España en la vanguardia de los datos abiertos, creando entonces el **Catálogo de Información Pública del Sector Público** lugar de convergencia de datos de la Administración Pública. Posteriormente, el Proyecto Aporta se adhiere al portal *datos.gob.es* que pasa a ser el lugar de referencia y canalización de todos los datos de las Administraciones Públicas. A día de hoy, nos encontramos entre los 4 países con más catálogos de datos públicos y portales oficiales [34] y somos líderes en madurez de datos abiertos en 2016 de entre los países europeos [35].

En materia legislativa, se ha hablado con anterioridad de la legislación vigente actualmente en esta materia, salvo en lo referente a la protección de datos de carácter personal, que relato a continuación.

A la hora de buscar información para reutilizarla, hay que tener en cuenta que si lo que estamos buscando está relacionado con la salud y, por consiguiente, con datos personales de pacientes. Podríamos encontrarnos con la dificultad de acceder a ellos por la LODP. Es lo que ocurre, por ejemplo, en el caso práctico que analizo en el capítulo 3.Caso Práctico, el cual trata sobre la categorización (benigno/maligno) de cáncer de mama a partir del tamaño de las células cancerígenas obtenidas mediante imagen digitalizada de una aspiración por aguja delgada (FNA, *Fine Needle Aspiration*).

A continuación, muestro los puntos de mayor interés en nuestro caso, descritos en dicha ley.

Ley Orgánica 15/1999, de 13 de diciembre, de protección de datos de carácter personal, [12].

#### *Artículo 3 Definiciones.*

*A los efectos de la presente Ley Orgánica se entenderá por:*

- a) Datos de carácter personal: cualquier información concerniente a personas físicas identificadas o identificables.*
- b) Fichero: todo conjunto organizado de datos de carácter personal, cualquiera que fuere la forma o modalidad de su creación, almacenamiento, organización y acceso.*
- c) Tratamiento de datos: operaciones y procedimientos técnicos de carácter automatizado o no, que permitan la recogida, grabación, conservación, elaboración, modificación, bloqueo y cancelación, así como las cesiones de datos que resulten de comunicaciones, consultas, interconexiones y transferencias.*
- d) Responsable del fichero o tratamiento: persona física o jurídica, de naturaleza pública o privada u órgano administrativo, que decida sobre la finalidad, contenido y uso del tratamiento.*



- e) *Afectado o interesado: persona física titular de los datos que sean objeto del tratamiento a que se refiere el apartado c) del presente artículo.*
- f) *Procedimiento de disociación: todo tratamiento de datos personales de modo que la información que se obtenga no pueda asociarse a persona identificada o identificable.*
- g) *Encargado del tratamiento: la persona física o jurídica, autoridad pública, servicio o cualquier otro organismo que, sólo o conjuntamente con otros, trate datos personales por cuenta del responsable del tratamiento.*
- h) *Consentimiento del interesado: toda manifestación de voluntad, libre, inequívoca, específica e informada, mediante la que el interesado consienta el tratamiento de datos personales que le conciernen.*
- i) *Cesión o comunicación de datos: toda revelación de datos realizada a una persona distinta del interesado.*
- j) *Fuentes accesibles al público: aquellos ficheros cuya consulta puede ser realizada, por cualquier persona, no impedida por una norma limitativa o sin más exigencia que, en su caso, el abono de una contraprestación. Tienen la consideración de fuentes de acceso público, exclusivamente, el censo promocional, los repertorios telefónicos en los términos previstos por su normativa específica y las listas de personas pertenecientes a grupos de profesionales que contengan únicamente los datos de nombre, título, profesión, actividad, grado académico, dirección e indicación de su pertenencia al grupo. Asimismo, tienen el carácter de fuentes de acceso público los diarios y boletines oficiales y los medios de comunicación.*

#### *Artículo 4. Calidad de los datos.*

*Los datos de carácter personal objeto de tratamiento no podrán usarse para finalidades incompatibles con aquellas para las que los datos hubieran sido recogidos. No se considerará incompatible el tratamiento posterior de éstos con fines históricos, estadísticos o científicos.*

#### *Artículo 5. Derecho de información en la recogida de datos.*

- a) *Cuando los datos de carácter personal no hayan sido recabados del interesado, éste deberá ser informado de forma expresa, precisa e inequívoca, por el responsable del fichero o su representante, dentro de los tres meses siguientes al momento del registro de los datos, salvo que ya hubiera sido informado con anterioridad, del contenido del tratamiento, de la procedencia de los datos, así como de lo previsto en las letras a), d) y e) del apartado 1 del presente artículo.*
- b) *No será de aplicación lo dispuesto en el apartado anterior, cuando expresamente una ley lo prevea, cuando el tratamiento tenga fines históricos, estadísticos o científicos, o cuando la información al interesado resulte imposible o exija esfuerzos desproporcionados, a criterio de la Agencia de Protección de Datos o del organismo autonómico equivalente, en consideración al número de interesados, a la antigüedad de los datos y a las posibles medidas compensatorias.*

#### *Artículo 6. Consentimiento del afectado.*

- a) *El tratamiento de los datos de carácter personal requerirá el consentimiento inequívoco del afectado, salvo que la ley disponga otra cosa.*

#### *Artículo 7. Datos especialmente protegidos.*

- a) *Los datos de carácter personal que hagan referencia al origen racial, a la salud y a la vida sexual sólo podrán ser recabados, tratados y cedidos cuando, por razones de interés general, así lo disponga una ley o el afectado consienta expresamente.*

- b) *No obstante lo dispuesto en los apartados anteriores, podrán ser objeto de tratamiento los datos de carácter personal a que se refieren los apartados 2 y 3 de este artículo, cuando dicho tratamiento resulte necesario para la prevención o para el diagnóstico médicos, la prestación de asistencia sanitaria o tratamientos médicos o la gestión de servicios sanitarios, siempre que dicho tratamiento de datos se realice por un profesional sanitario sujeto al secreto profesional o por otra persona sujeta asimismo a una obligación equivalente de secreto.*

*Artículo 8. Datos relativos a la salud.*

*Sin perjuicio de lo que se dispone en el artículo 11 respecto de la cesión, las instituciones y los centros sanitarios públicos y privados y los profesionales correspondientes podrán proceder al tratamiento de los datos de carácter personal relativos a la salud de las personas que a ellos acudan o hayan de ser tratados en los mismos, de acuerdo con lo dispuesto en la legislación estatal o autonómica sobre sanidad.*

*Artículo 12. Acceso a los datos por cuenta de terceros.*

- a) *La realización de tratamientos por cuenta de terceros deberá estar regulada en un contrato que deberá constar por escrito o en alguna otra forma que permita acreditar su celebración y contenido, estableciéndose expresamente que el encargado del tratamiento únicamente tratará los datos conforme a las instrucciones del responsable del tratamiento, que no los aplicará o utilizará con fin distinto al que figure en dicho contrato, ni los comunicará, ni siquiera para su conservación, a otras personas.*

La protección de datos y la seguridad de los mismos, es un tema recurrente y remarcado y por ello se toman las medidas oportunas y necesarias para garantizar la privacidad de los ciudadanos. Si bien es cierto que las técnicas empleadas para proteger y anonimizar los datos pueden ser usadas en el sentido opuesto, lo que requiere un continuo trabajo.



# 2

## Catálogos de datos y servicios web

Una vez introducido los conceptos y objetivos de la filosofía de datos abiertos, podemos dar paso al elemento principal para conseguir estas metas, los **catálogos de datos**. Estos catálogos son el punto de encuentro entre la Administración Pública (o entidad emisora de datos abiertos) y el ciudadano/empresa solicitante de información.

Podemos encontrar una gran cantidad de estos catálogos en *Internet*, en su mayoría **entidades públicas**, pero también encontramos catálogos por **temática** concreta<sup>4</sup> (pueden ser asuntos económicos como *Quandl*) y de instituciones u **organizaciones**.

Son varias las organizaciones que se encargan de aglutinar los diferentes catálogos para simplificar las búsquedas de los usuarios. Por ejemplo, la fundación CTIC cuenta con un mapa actualizado hasta 2015 de las iniciativas globales en materia de datos abiertos, como podemos ver en Figura 9 [36].



Figura 9. Mapa de iniciativas de datos abiertos. Fuente: Fundación CTIC

Por otro lado, podemos ver en la Figura 10 el *open data inception* en el que se recoge un número mayor de fuentes de información [37].

<sup>4</sup> Si bien es cierto, que normalmente estos sitios se nutren de la información publicada por las Administraciones Públicas.

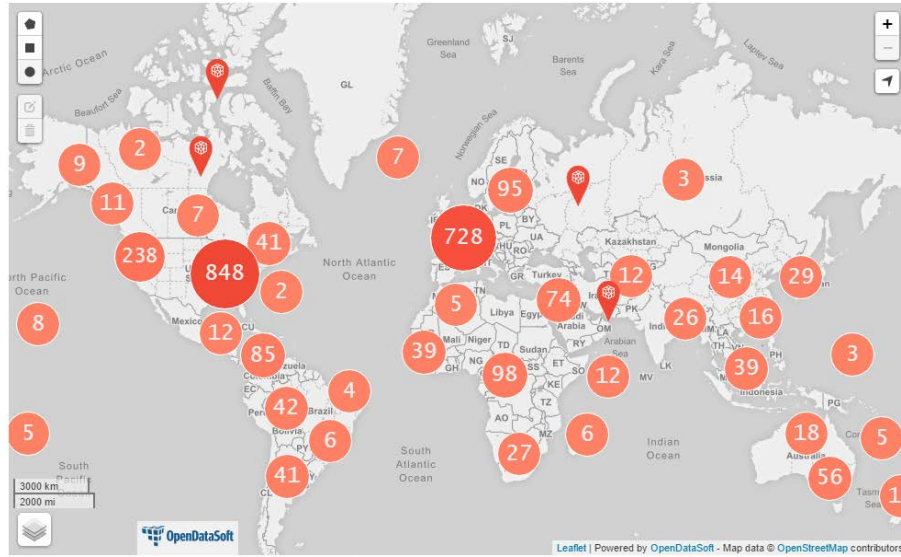


Figura 10. Mapa de iniciativas Open Data. Fuente: open data inception.

Hay que tener en cuenta que estos mapas y recopilaciones se generan gracias a la ayuda de gente y que, por tanto, no tienen por qué tener todas las fuentes de información y albergar alguna incorrecta. Es decir, estos datos hay que tomarlos como orientativos.

Si acudimos a fuentes oficiales de alguno de los países, podemos ver el estado actual de cada uno. A modo de síntesis, solo se detallará el caso de 4 países, los que tienen más catálogos de datos [34]. Estos países son: **Estados Unidos, Reino Unido, Canadá y España.**

## 1. Estados Unidos

Como ya comentamos anteriormente, Estados Unidos es uno de los principales impulsores de las iniciativas de datos abiertos y cuenta con multitud bases de datos abiertas a consultas. La *web* principal en la que consultar es **data.gov** en la cual encontramos (a fecha de febrero de 2017) cerca de 194.489 conjuntos de datos (*datasets*).

Estos conjuntos de datos se dividen en temáticas como podemos observar en la Figura 11.

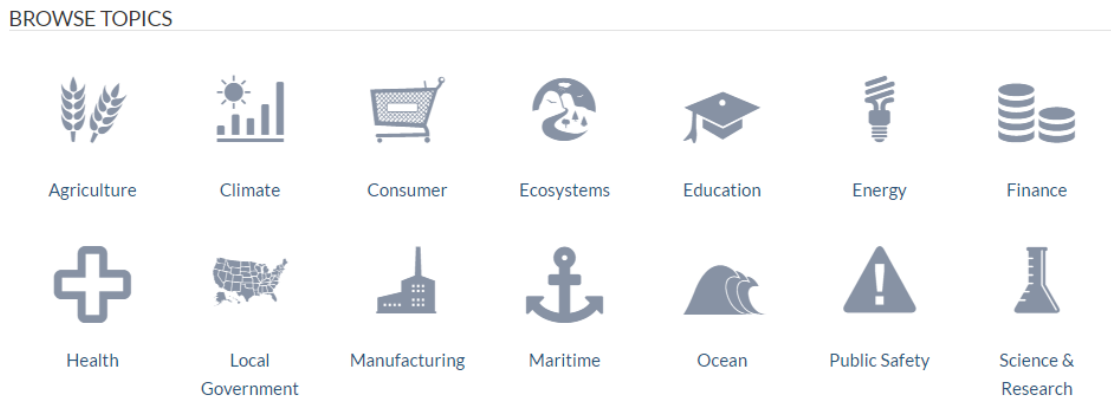


Figura 11. Temáticas disponibles en el portal del gobierno de Estados Unidos. Fuente: data.gov.

Más específicamente, en el apartado dedicado a la salud (*Health*) nos encontramos 1.981 conjuntos de datos y una *web* dedicada en exclusiva a este tipo de información, en la que hay todavía más conjuntos de datos (3.120) y subdividida en los siguientes apartados.



Figura 12. Temáticas relacionadas con la salud. Fuente: *healthdata.gov*

## 2. Reino Unido

En el caso de Reino Unido, también país destacado y precursor de las iniciativas de datos abiertos, encontramos una *web* oficial con multitud de datos para consultar. Específicamente, la *web* es *data.gov.uk* y tiene (a fecha de febrero de 2017) alrededor de 40.916 conjuntos de datos para consultar.

### Browse data by theme

#### Business and economy

Small businesses, industry, imports, exports and trade

#### Crime and justice

Courts, police, prisons, offenders, borders and immigration

#### Defence

Armed forces, health and safety, search and rescue

#### Education

Students, training, qualifications and the National Curriculum

#### Environment

Weather, flooding, rivers, air quality, geology and agriculture

#### Government

Staff numbers and pay, local councillors and department business plans

#### Government spending

Includes all payments by government departments over £25,000

#### Health

Includes smoking, drugs, alcohol, medicine performance and hospitals

#### Mapping

Addresses, boundaries, land ownership, aerial photographs, seabed and land terrain

#### Society

Employment, benefits, household finances, poverty and population

#### Towns and cities

Includes housing, urban planning, leisure, waste and energy consumption

#### Transport

Airports, roads, freight, electric vehicles, parking, buses and footpaths

Figura 13. Temáticas del portal del gobierno de Reino Unido. Fuente: *data.gov.uk*

Nuevamente, en el apartado dedicado a la salud, encontramos en este caso 1.947 conjuntos de datos listos para consultar. En este caso, más que por una subcategoría, podríamos filtrar los resultados por formato, ya que, aunque se pretende dar la información en el mayor número de formatos, no siempre es posible.

RESOURCE FORMAT
HTML (743)
CSV (558)
XLS (278)
PDF (73)
WMS (72)
XML (53)
ZIP (52)
GeoJSON (45)
Esri REST (42)
WFS (16)

Figura 14. Formatos y número de datasets disponibles en la web de Reino Unido. Fuente: data.gov.uk

### 3. Canadá

En tercer lugar, hablaremos de Canadá, uno de los países que más impulso está dando a la apertura de datos en los últimos años. Su *web* oficial es *open.canada.ca* y en ella podemos encontrar 148.203 conjuntos de datos, de los cuales 691 están relacionados con salud (y seguridad). En el caso de Canadá, dispone de un mayor número de formatos disponibles, aunque en alguno de ellos solo disponga de un documento. Quisiera destacar en este caso, la posibilidad de filtrar los resultados por la frecuencia de actualización de los mismos. Como podemos ver en la Figura 15, de los 691 conjuntos de datos referidos a la salud, casi la mitad se actualizan bajo petición y solo 1 recibe actualizaciones diarias.

▼ Maintenance and Update Frequency

- [As Needed \(308\)](#)
- [Irregular \(169\)](#)
- [Annually \(162\)](#)
- [Unknown \(25\)](#)
- [Not Planned \(18\)](#)
- [Monthly \(3\)](#)
- [Continual \(2\)](#)
- [Quarterly \(2\)](#)
- [Biannually \(1\)](#)
- [Daily \(1\)](#)

[▲ Show less](#)  
[✖ Clear All](#)

Figura 15. Frecuencia de actualización de los datos del gobierno de Canadá. Fuente: open.canada.ca

## 4. España

En el caso de España, encontramos en su *web* oficial **datos.gob.es** el número de iniciativas españolas relacionadas con los datos abiertos. En total 122 iniciativas, procedentes de todos los niveles del Estado y de 8 Universidades.

En cuanto al catálogo y como en ocasiones anteriores, vemos en la Figura 16, las diferentes temáticas disponibles. En el caso de España, contamos con 12.858 conjuntos de datos de los cuales, 611 están relacionados con el ámbito sanitario.

Cabe señalar que, se ofrecen multitud de formatos pero que sucede lo mismo que en casos anteriores, no todos los conjuntos de datos están disponibles en todos los formatos, bien por posibilidades prácticas, bien por dejadez. También veo destacable la ausencia de filtro por frecuencia de actualización de datos.



Figura 16. Categorías disponibles del gobierno de España. Fuente: *datos.gob.es*

Respecto al ámbito sanitario, podemos indagar más acudiendo a *web* del Ministerio de Sanidad, Servicios Sociales e Igualdad podemos encontrar información más detallada. Desde los años 90, por orden ministerial, se lleva a cabo una recogida de datos en lo que se conoce como Conjunto Mínimo Básico de Datos (CMBD) en el que se recogen datos estadísticos de los hospitales públicos para conocer y clasificar por Grupos Relacionados por el Diagnóstico (GRD) los ingresos hospitalarios.

De estos CMBD se realiza un barómetro sanitario cuya finalidad es la mejora del sistema sanitario y en palabras del Ministerio: “Conocer la opinión de los ciudadanos para tomar en consideración sus expectativas, como elemento importante para establecer las prioridades de las políticas de salud.” [38].

El mismo Ministerio, tiene un portal estadístico para consultas interactivas en el que no solo podemos encontrar información sobre los barómetros sanitarios, sino también, sobre los CMBD, la mortalidad, la Encuesta Nacional de Salud de España (ENSE) y la población [39].

Si entramos por ejemplo en **población**, veremos una página similar al programa *Microsoft Excel*, en el que podremos ver, bien de forma gráfica o bien en formato tabla, los datos de población. Estos datos de población se pueden ajustar nuestras necesidades a través del panel mostrado en la Figura 17.



- Medidas y Dimensiones :
- Población
  - Edad
    - Tramo de edad
  - Geografía Hospital
    - CCAA
  - Sexo
    - Sexo
  - Temporal
    - Año

Figura 17. Opciones disponibles en el portal interactivo referente a población. Fuente: [pestadistico.inteligenciadegestion.msssi.es](http://pestadistico.inteligenciadegestion.msssi.es)

A modo de ejemplo de uso, he seleccionado la población en Castilla y León, desde el año 2000 que supere los 70 años de edad y he exportado a *Excel* los resultados. Los muestro en la Figura 18.

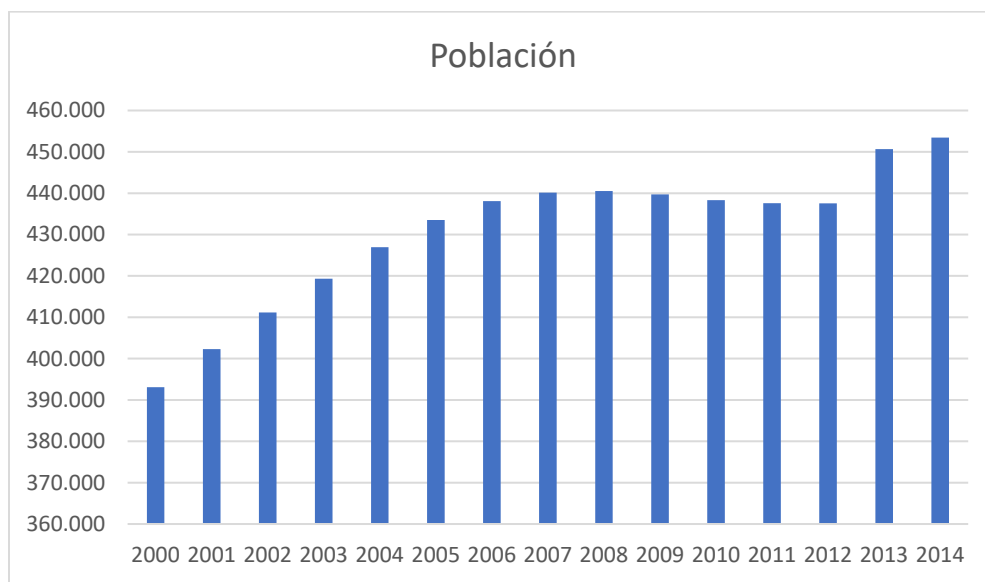


Figura 18. Evolución de la población en Castilla y León. Fuente: [pestadistico.inteligenciadegestion.msssi.es/publicosns](http://pestadistico.inteligenciadegestion.msssi.es/publicosns)

Una vez recogido algunos de los catálogos de datos oficiales de países, hablaremos de otros catálogos de datos relevantes como puede ser el portal de la **Unión Europea**, el de la Organización para la Cooperación y el Desarrollo Económico (**OCDE**) o de la Organización Mundial de la Salud (**OMS**)

## 5. Otros catálogos de datos

### 1. Unión Europea

La Unión Europea recopila datos a partir de los portales oficiales de los países miembro y los ofrece en conjunto en la *web*: [europeandataportal.eu](http://europeandataportal.eu).

Como en casos anteriores, también muestra una búsqueda por temática específica, así como filtros por formato, catálogo del que proviene o del país en concreto. Cabe señalar, que de los 606.087 conjuntos de datos de los que se compone el portal europeo, el país que **más conjuntos de datos** aporta, no es ninguno de los mencionados hasta ahora, sino que se trata de la **República Checa** seguido de cerca por **Alemania**. Veo importante resaltar este dato puesto que, la apertura de datos no es cosa de un par de países, es esfuerzo e implicación de todos, independientemente de su tamaño o capacidad. Sin embargo, en el apartado sanitario los países más destacados vuelven a ser, por orden: **Reino Unido, España, Alemania y Francia** [40].

Otro lugar de referencia es *Eurostat*, en el que se recogen estadísticas de diferentes temáticas acerca de los países europeos [41].

### 2. OCDE

La OCDE es una organización de países a lo largo del mundo, que cuenta actualmente a 35 miembros, entre ellos España, pero abierto a nuevas incorporaciones. Entre sus objetivos destaca que trata de promover la colaboración y cooperación entre países para solventar los problemas comunes y proponer políticas que sean beneficiosas económica y socialmente [42].

Cuenta con un portal estadístico ([stats.oecd.org](http://stats.oecd.org)) y otro de datos ([data.oecd.org](http://data.oecd.org)) en los que se muestra información de los países miembro y de otros que, aunque no lo son, están en vías de serlo o son de especial interés. Muestro por ejemplo en la Figura 19 la estancia media hospitalaria por países, en el caso de Japón, la más destacable, alcanza los 16 días en promedio, mientras que la media de la OCDE está en 6, media en la que se encuentra justamente España.

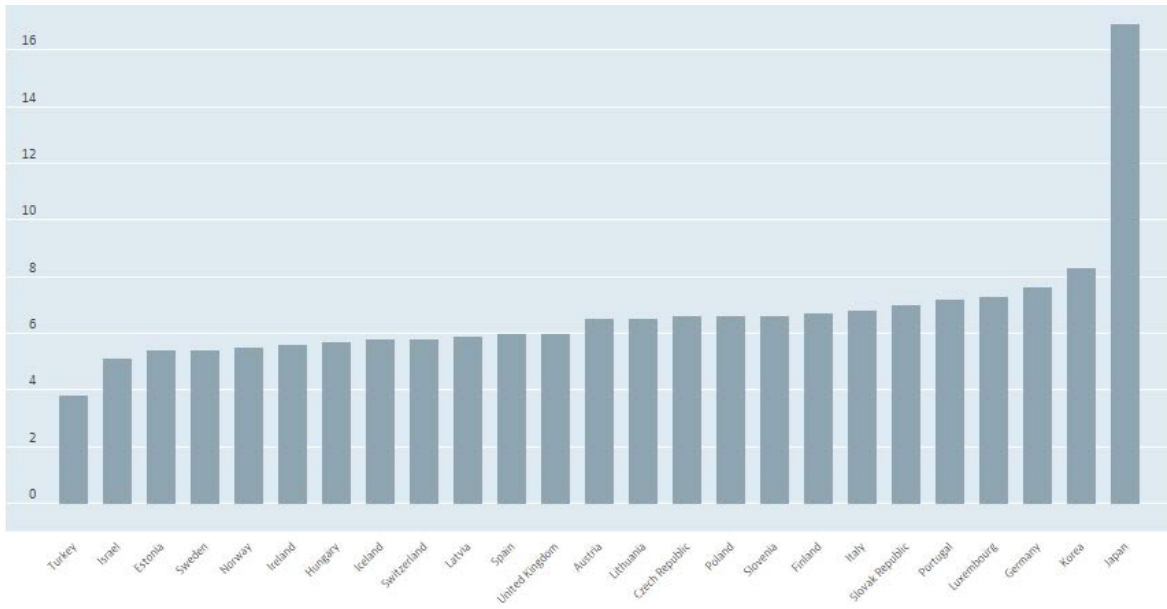


Figura 19. Estancia media hospitalaria por países. Fuente: [data.oecd.org/healthcare/length-of-hospital-stay.htm](http://data.oecd.org/healthcare/length-of-hospital-stay.htm)

### 3. OMS

La Organización Mundial de la Salud es la autoridad directiva y coordinadora de asuntos de salud a nivel mundial dentro de las Naciones Unidas, que busca el bienestar completo, no solo libre de enfermedades o afecciones de todos los pueblos del planeta, independientemente de su situación económica o social [43].

La OMS cuenta con varias bases de datos dedicadas a la salud y si por ejemplo accedemos a la sección regional europea, nos ofrece datos relacionados con los países europeos recogidos, entre otros sitios, del *Eurostat*.

### 4. Otros catálogos

En la búsqueda de catálogos de datos, nos encontramos con *KDnuggets*. Es una *web* de referencia en *Business Analytics*, *Big Data*, *Data Mining* y *Data Science*, que recopila lugares en los que encontrar estos catálogos. Creado y gestionado por un experto en la materia, Gregory Piatetsky, que además es cofundador de las conferencias *Knowledge Discovery and Data mining (KDD)*. En su origen se creó para conectar a investigadores de la materia de todo el mundo y actualmente es uno de los medios más reconocidos y premiados [44].

Cuenta con un amplio repertorio de contenidos, como  **cursos** de *Big Data* o de *Data Mining*, contenido *Software*, calendario de  **conferencias**, además de conjunto de datos (*datasets*) y enlaces a *webs* específicas. Dentro de los conjuntos de datos encontrados a través de esta *web*, vamos a diferenciar dos tipos: conjunto de datos en **temáticas generales/específicas** y conjuntos de datos para **competiciones**.

#### a. Temáticas

En este subapartado de páginas *web* encontradas en *KDnuggets* destacaré sitios no oficiales de países, para no reiterar y repetir en exceso. Por ello, he seleccionado algunos sitios *web* relevantes, no solo en el ámbito de la salud, sino también en otros ámbitos.

Como ejemplos relacionados con la **economía** y las **finanzas**, destaco *AssetMacro* que cuenta con datos históricos y en directo de datos financieros e indicadores macroeconómicos de los mercados y economías mundiales, [45]. Y, por otro lado, *Quandl*, que ofrece información financiera y económica, no solo de fuentes gubernamentales sino de agentes de los mercados financieros y los ofrecen en varios formatos para su análisis de sus clientes, [46]. Más que una *web* para distribuir conjuntos de datos, se trata de una plataforma que **ofrece herramientas** de visualización y tratamiento de datos de las fuentes de las que se nutre.



Figura 20. Representación del funcionamiento de Quandl. Fuente: quandl.com

Similar a *Quandl*, existen otras herramientas de visualización y tratamiento de datos como *Qlik*, que venden sus herramientas a profesionales para gestionar sus datos, [47]. Más adelante, hablaré sobre servicios *web* específicos y relacionados con grandes conjuntos de datos.

Continuando con otros conjuntos de datos disponibles, pongamos la vista en la tierra, o mejor dicho apuntando a ella. El proyecto *Copernicus* es una iniciativa europea para observar los cambios medioambientales del planeta y pone a disposición a través de *sentinels.copernicus.eu* de información relativa a los temas mostrados en la Figura 21.



Figura 21. Temáticas del proyecto Copernicus. Fuente: copernicus.eu

Si nos vamos con la canción a otra parte, encuentro curioso el caso de *Million Song Dataset*, que como su nombre indica, alberga información relacionada con música, tanto del artista como de la propia canción, como puede ser la duración o la frecuencia en cada momento que está sonando.

Por último y no por ello menos importante, *UCI Machine Learning Repository* que cuenta con conjuntos de datos pensados para *machine learning*, disciplina que busca el aprendizaje automático de las máquinas mediante ejemplos. Se centra en buena parte en aspectos más de investigación, ciencia e ingeniería.

### ***b. Competiciones***

Un apartado interesante dentro de los datos abiertos y de los conjuntos de datos, es la posibilidad de competir para obtener los mejores resultados y, en algunos casos, hasta obtener remuneración por ello.

Un caso destacable de estos últimos, de los remunerados, es una competición patrocinada por la NFL (*National Football League*) americana que trata de detectar de manera temprana lesiones cerebrales de sus jugadores y mejorar la protección para su seguridad [48]. La remuneración ofrecida supera los 10 millones de dólares americanos.

Otro portal para competiciones, en este caso del mundo empresarial, es *Crowd Analytix* en la que se resuelven problemas que plantean las empresas y ofrecen una remuneración a cambio [49]. En la Figura 22 podemos observar un reto ya cerrado de modelado. En este reto había que predecir el consumo de combustible de un avión en las diferentes fases de vuelo, para el cual, han colaborado 362 personas y el ganador ha obtenido 7.500 \$.

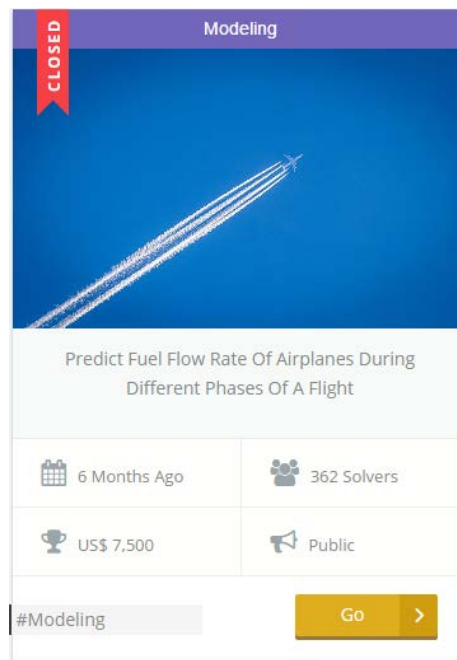


Figura 22. Ejemplo de competición para una empresa. Fuente: [crowdfunder.com/community](https://crowdfunder.com/community)

Un ejemplo más de competición, en este caso referido al ámbito de la salud, es la organizada por la IEEE EMBS (*Engineering in Medicine and Biology Society*) y la IEEE *Big Data Initiative*, en la que se facilitan unos datos ya tratados, cuya finalidad es la de conocer qué se puede aprender de esos datos.

Existen otras alternativas como *DrivenData*, *Innoventive* y *TunedIT* dentro de las competiciones de datos, pero creo conveniente centrar el interés en un último caso, el de *Kaggle* del cual se ha extraído los datos para el caso de uso detallado en los capítulos siguientes [50].

*Kaggle*, es la plataforma líder para competiciones que cuenta también con más conjuntos de datos, sin competiciones de por medio [50]. Una característica interesante que tiene, es la posibilidad de **iniciarte** en *Data Science*, con “competiciones de entrenamiento”. Esta competición, entra dentro de lo que podíamos llamar anecdótico y en cierto modo “divertido”, al tratarse del *Titanic* [51].

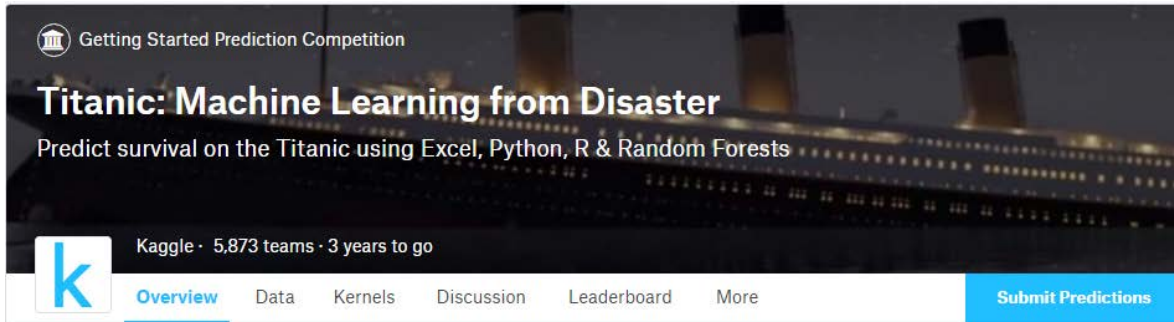


Figura 23. Competición de iniciación en Data Science. Fuente: kaggle.com

Se trata de predecir la supervivencia de los pasajeros al naufragio del Titanic.

En la competición tienes una descripción general del caso a resolver, en este caso el *Titanic*, y unos datos ya trabajados y estructurados, con su respectiva explicación para que los trabajes. Para resolverlo, puedes emplear diversos métodos y/o lenguajes, como puede ser *Excel*, *Python* o *R*.

De la *web* de *kaggle* he obtenido el conjunto de datos que, en capítulos posteriores del trabajo, analizaré y realizaré un ejemplo de uso.

Estos datos, en concreto, tratan sobre el diagnóstico del cáncer de mama mediante imagen digitalizada de aspirado por aguja fina (**FNA**, por sus siglas en inglés) y cuya finalidad es intentar predecir con el mayor porcentaje posible, si un cáncer es benigno o es maligno.

En el capítulo siguiente, pondremos en contexto con un poco más de profundidad, de que trata y que significan los datos con los que vamos a trabajar, dejando claro que se trata solo de una contextualización que no cuenta con la ayuda de ningún experto en la materia y que por tanto el nivel de profundización no será exhaustivo. Por otro lado, se detallará las herramientas utilizadas para tal análisis.

Por último, veo necesario describir una serie de servicios *web* que son fundamentales para la filosofía de datos abiertos, en el caso de *CKAN* (*Comprehensive Knowledge Archive Network*), por ejemplo, muchas de las páginas oficiales de países están basadas en este sistema abierto [52].

## 6. Servicios Web

Además de páginas *web* con repositorios de datos, encontramos también otras, destinadas a ofrecer servicios. Estos servicios son muy variados, desde alojamiento de datos, pasando por su procesado y visualización. Se caracterizan normalmente por pertenecer a empresas con grandes infraestructuras para dichos objetivos.

### 1. CKAN

Es el portal **líder** de plataformas de datos de carácter *open-source*, empleado por gobiernos e instituciones independientes. Es desarrollado por la *Open Knowledge Foundation*, ofreciendo una serie de tecnologías de integración y visualización de datos a sus clientes [53].

Entre estos clientes está, por ejemplo, el gobierno de Reino Unido o la Comisión Nacional de los Mercado y la Competencia (CNMC).

### 2. Socrata

En el caso de *Socrata*, es líder también, pero en soluciones basadas en *Cloud* para gobiernos (*Cloud-based Data Democratization solutions for Government, CDDG*). Ofrece *software-as-a-service* (*SaaS*) pero es importante destacar que se basa y sustenta en *Amazon Web Services* (*AWS*), del que hablaré a continuación [54].

Los clientes de *Socrata* son solo gubernamentales, a todos los niveles, desde la ciudad hasta el Estado completo. Centran sus soluciones en la publicación de datos en *cloud*, en el aprovechamiento y rendimiento de los mismos, en datos financieros y de seguridad.

### 3. Amazon Web Services

Es la plataforma de *Amazon* para proveer de servicios en *Cloud* a empresas y gobiernos, pionera en el sector. Ofrece una gran cantidad de herramientas, como podemos ver en la Figura 24, pensadas para externalizar servicios, esto es, en el caso de grandes volúmenes de datos, tema que estamos tratando, la infraestructura necesaria y los costes asociados al mantenimiento y soporte de la misma hacen, en muchas ocasiones, inviable los proyectos con pocos recursos, ya sean económicos o de conocimiento. Por ello, *Amazon*, que cuenta con una infraestructura a nivel mundial y tiene los conocimientos necesarios para su gestión y mantenimiento, pone a disposición (previo pago) sus instalaciones para alojar, tratar y distribuir datos, entre otras cosas [55].

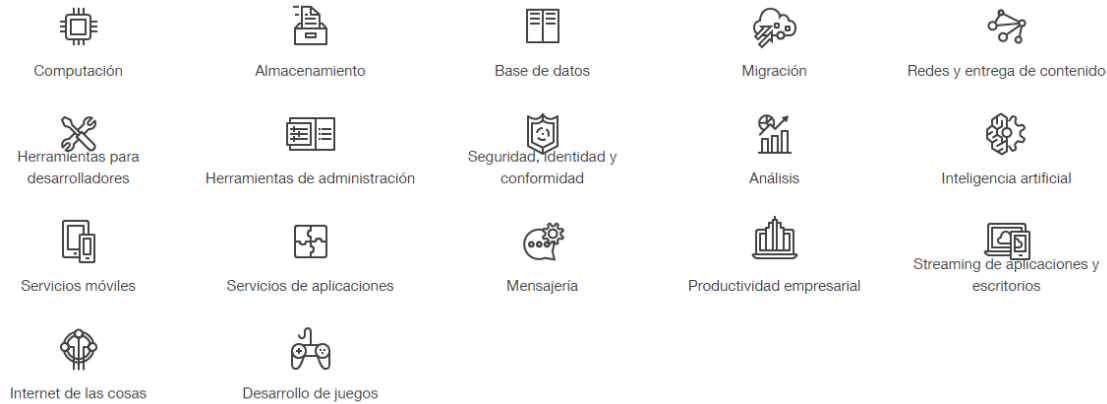


Figura 24. Servicios ofrecidos por AWS. Fuente: aws.amazon.com

#### 4. Microsoft Azure

En la misma línea que *Amazon Web Services*, *Microsoft* cuenta con un servicio propio llamado *Azure*, en el que ofrece servicios en *Cloud* muy similares y en los mismos ámbitos, con una gran red internacional como respaldo. En su caso, y para marcar diferencias, *Microsoft Azure* hace hincapié en su seguridad y la confianza otorgada por grandes empresas en sus servicios y en sus certificaciones de seguridad [56]. (Figura 25).



Figura 25. Certificaciones de Microsoft Azure. Fuente: azure.microsoft.com

#### 5. Google Cloud Platform

No podía faltar *Google* en cuanto a servicios en *Cloud* se refiere, pues goza de unas infraestructuras y una cantidad de datos a su disposición inmensas [57].

Ofrece servicios similares a los mencionados anteriormente, pero pone el foco en el futuro, para diferenciarse del resto.

#### 6. IBM Cloud

IBM ofrece también servicios en *cloud*, en el que destacan la capacidad de innovación de la empresa, de generar nuevo conocimiento y el ponerlo al servicio de las empresas y entidades públicas [58]. Cabe destacar la posibilidad de utilizar **Watson**, la conocida Inteligencia Artificial de IBM, para dotar de mayor inteligencia a los productos y aplicaciones.





# 3

## Caso Práctico

### 1. Introducción

Una vez conocido el contexto y la situación actual en lo concerniente a datos abiertos es objetivo del presente trabajo, realizar un análisis de un conjunto de datos.

En una primera idea, se pretendía emplear datos del gobierno de España. Más concretamente del Ministerio de Sanidad, Servicios Sociales e Igualdad y del Ministerio de Agricultura y Pesca, Alimentación y Medio Ambiente, en concreto de este último, para realizar un análisis sobre la incidencia de ciertas enfermedades relacionadas con las vías respiratorias en zonas con un índice de contaminación elevado.

El acceso y recopilación de los datos necesarios para tal estudio se alejaba de los tiempos efectivos para realizar un trabajo final de grado. Es por ello, que se decide escoger unos datos más sencillos y simples para realizar un análisis.

Los datos elegidos son del ámbito sanitario, más concretamente, sobre el cáncer de mama. Puede accederse a estos datos desde la *web* de *kaggle*, [59]. Estos datos son proporcionados por *UCI Machine Learning Repository* en el conjunto de datos “*Wisconsin Diagnostic Breast Cancer (WDBC)*”, [60], [61].

El objetivo del estudio es tratar de clasificar, a partir de una serie de datos referidos a los tamaños de las células cancerígenas, si el cáncer es benigno o maligno. Para ello, los datos descargados de [59], nos ofrecen, en formato de *Excel*, 569 filas correspondientes a pacientes, de los cuales 357 tienen resultado benigno (62,74%) y los restantes 212, maligno (37,26%). En cuanto a las columnas, en ellas disponemos, primero de un número de identificación (que eliminaremos al no ser de utilidad), de una columna de **diagnóstico** en formato de **factor**, que será nuestra columna de **predictores**, y contaremos con información referida 3 valores (media, error estándar y peor caso) de 10 magnitudes referidas a **dimensión de las células**, incorporadas como variables **numéricas**.

Con estos datos, pretendo obtener, mediante comparación de modelos de clasificación, una predicción lo más fiable posible para el diagnóstico del cáncer de mama.

### 2. Contextualización del caso

Antes de comenzar a trabajar con los datos, conozcamos un poco su significado. El cáncer de mama afecta mayoritariamente a mujeres, pero es una enfermedad que todos podemos padecer. Para entender qué es el cáncer, y en particular el de mama, debemos hablar primero de las células del cuerpo.

Cualquier órgano del cuerpo está formado por un conjunto de células, que se dividen automáticamente, de forma regular, para poder reemplazar a las ya envejecidas. Este mecanismo se regula solo, pero existen ocasiones en las que se altera el proceso, en una o varias células, provocando una división incontrolada que originará la existencia de un tumor.

No siempre tiene que surgir un tumor maligno, sino que para que adquiera esta caracterización debe cumplirse que la célula o células, además de dividirse de forma constante, invadan a otros tejidos u órganos del cuerpo.

En el caso del cáncer de mama, surge en los tejidos de la glándula mamaria, y para que sea maligno debe causar la invasión de tejidos sanos de su entorno, como reza el párrafo anterior.

En la Figura 26 podemos apreciar gráficamente cómo es este proceso [62].

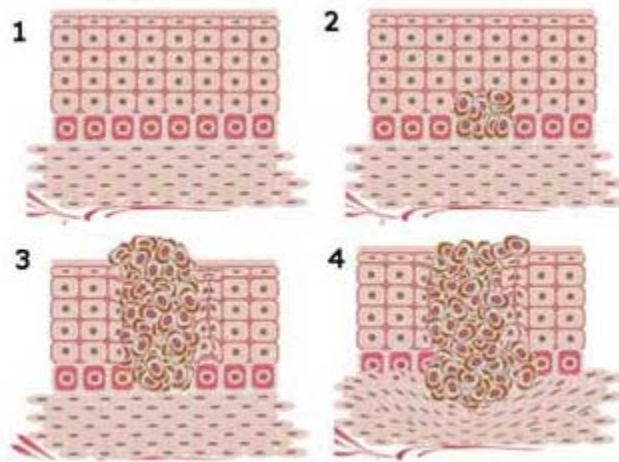


Figura 26. Fases de la aparición del cáncer. Fuente: *aecc.es*

En la descripción del caso, se habla de un procedimiento para obtener los tamaños de las células cancerígenas conocido como **FNA** (*Fine Needle Aspirate*) o aspiración por aguja delgada.

Este procedimiento es una de las formas que existen en la actualidad para examinar tejidos a la hora de examinar tumores consiste en la realización de una biopsia, que consiste en la extracción de una parte de tejido para su posterior examen.

Existen diferentes tipos de biopsias que, en el caso del **FNA**, consiste en recoger una pequeña muestra de tejido de la región mamaria, con el objetivo de estudiar si la tumoración es o no maligna. La ventaja de este tipo de biopsia con respecto a otros tipos es que es menos invasiva, y conlleva menos riesgos para el paciente, sobre todo derivados de infecciones o sangrados, habituales al aplicarse otras técnicas [63].

Una vez conocido un poco el tema que vamos a tratar, veamos que relevancia tienen los datos aportados.

En el conjunto de datos que vamos a analizar posteriormente, tenemos 569 pacientes y de estos, 3 valores distintos (media, error estándar y peor caso) para 10 datos sobre el tamaño de las células cancerígenas, entre los que se encuentra el radio, la textura o el perímetro.

Podemos preguntarnos por qué es importante conocer el tamaño de las células y en que puede ayudar en el diagnóstico, pues bien, en el cáncer se diferencian una serie de **grados**, estos grados indican la

rapidez con la que probablemente crecerá y se extenderá el tumor. Unas células cancerígenas similares en forma y en estructura a las normales tienen a extenderse más despacio que las que tienen una estructura más asimétrica y diferenciada de las normales.

No hay que confundir grado con **estadio** del cáncer, pues esto último se refiere a la expansión del cáncer por el cuerpo, siendo el peor caso (*estadio IV*) cuando se encuentra en zonas alejadas del origen primario del cáncer [64].

Lo expuesto en los dos últimos párrafos justifica el análisis de estos datos, los tamaños de las células cancerígenas, pues, si el diagnóstico es precoz, las posibilidades de superación de la enfermedad aumentan.

### 3. Contextualización de las herramientas empleadas.

Una vez contextualizado el caso, en esta sección detallaré las herramientas utilizadas para tratar y analizar estos datos.

La herramienta principal será **R Studio**, un entorno de desarrollo *software* pensado para programar en **lenguaje R**. Y en concreto, se usará el paquete *caret* (`()`), que aglutina y simplifica la creación de modelos predictivos. Comenzaremos describiendo brevemente qué es cada uno de los elementos mencionados.

#### 1. Lenguaje R

R es un lenguaje (y un entorno de desarrollo) para estadística y visualización de datos. Nació como un proyecto de *software* libre, con licencia de proyecto GNU<sup>5</sup> y que se asemeja a un lenguaje anterior, llamado S, resolviendo carencias de este.

Ofrece una gran variedad de opciones estadísticas predefinidas, como modelos lineales o de clasificación o *test* clásicos, así como herramientas de visualización como *plots*, *boxplots* e histogramas. Además, es posible usarlo en diferentes plataformas como *Windows*, UNIX (Linux, por ejemplo) y macOS [65].

Cuenta con manuales de iniciación y documentación necesaria para comenzar a trabajar con R y sus paquetes por defecto.

#### 2. R Studio

Es un entorno de desarrollo de libre distribución, aunque cuenta con una versión empresarial, que ofrece un *IDE* (*Integrated Development Environment*) completo además de paquetes adicionales como *Shiny* y *R Markdown* para creación de aplicaciones *web* y generación de documentos, respectivamente o por ejemplo *ggplot2* que ofrece herramientas de visualización más completas y mejores que *lattice* [66].

A continuación, mostraré las partes principales del entorno de desarrollo, comentando a su vez, sus funciones. En la Figura 27 podemos ver la ventana principal de *RStudio*<sup>6</sup> en la que podemos diferenciar **3 zonas**. En la parte **izquierda**, el área más amplia la ocupa la **consola**, lugar en el que podremos ejecutar nuestro código en R y visualizar los resultados. La parte **derecha** de la pantalla,

---

<sup>5</sup> Proyecto GNU: Filosofía de *software* libre que busca dar al usuario libertad y control sobre el *software* que usa.

<sup>6</sup> El fondo de la consola se ha cambiado para facilitar la visualización.

se divide en dos partes, en la **superior**, encontramos el **entorno** y el **historial**. El entorno nos mostrará las variables y los datos que estemos manejando y el historial, guardará cronológicamente, los comandos que hayamos ido utilizando. En la parte **inferior** derecha de la pantalla tenemos 5 paneles, el primero es **File** en la cual podremos ver la jerarquía de ficheros y carpetas disponibles en nuestro directorio de trabajo. El siguiente panel **Plots** en la que se mostrarán las gráficas correspondientes cuando se llame a una función que genere gráficas (como `plot()`, `hist()` ...). **Packages** por su parte, permite descargar, instalar y usar paquetes adicionales a las que posee *RStudio* por defecto, en la Figura 28 muestro la visualización de ese panel, en el que se aprecia la posibilidad de instalar (en caso de no estar en el ordenador se descarga primero), actualizar y seleccionar los paquetes que necesitemos utilizar. En **Help** obtendremos la documentación necesaria sobre aquello que necesitemos, ya sea información sobre un paquete o sobre una función determinada. Por último, **Viewer** permite visualizar como queda el contenido en versión *web* cuando se llama a esa función.

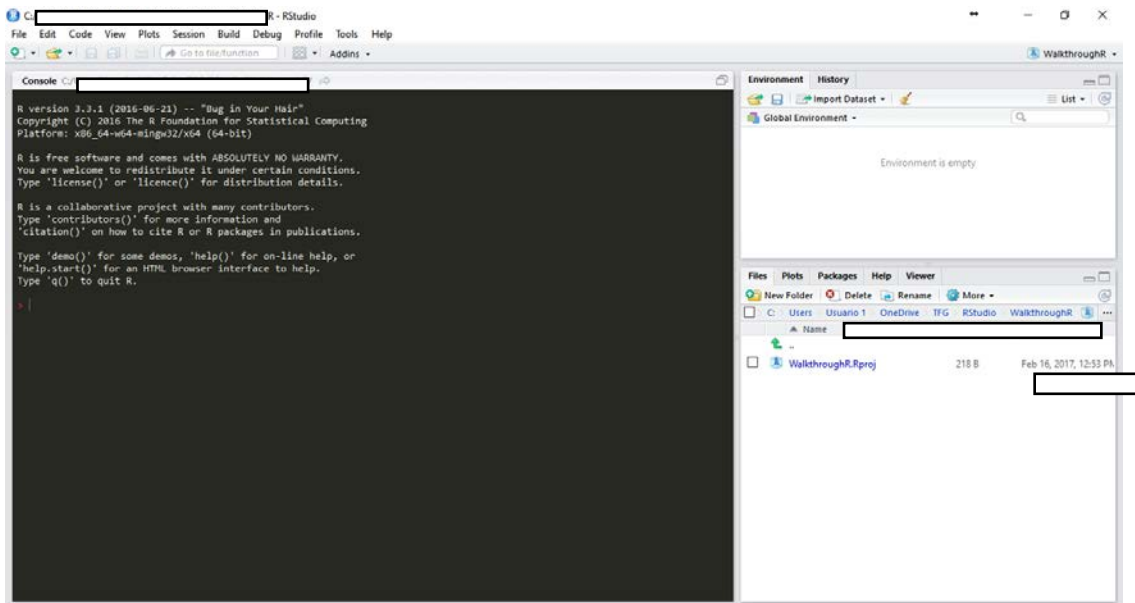


Figura 27. Entorno de RStudio. Ventana principal.

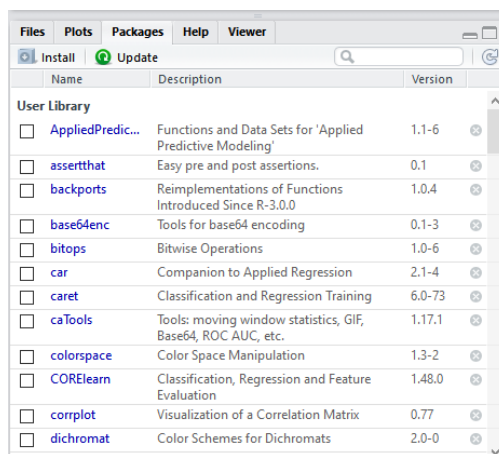


Figura 28. Panel Packages de RStudio.

Ya hemos visto el entorno básico de desarrollo, pero en la mayoría de ocasiones, a la hora de escribir código, necesitaremos reutilizar lo escrito anteriormente y posiblemente varias líneas de código antes de comprobar los resultados. Para esta tarea, la **consola**, aunque es funcional, no nos es demasiado útil, para ello emplearemos los **scripts**, pequeños programas que nos permitirán ejecutar código de una manera mucho más sencilla.

Cuando creamos un *script* en R, la parte izquierda del panel de la Figura 27, se divide en 2 partes, la **inferior** con la consola, ya comentada, y la **superior**, en la que escribiremos nuestro *script*<sup>7</sup>.

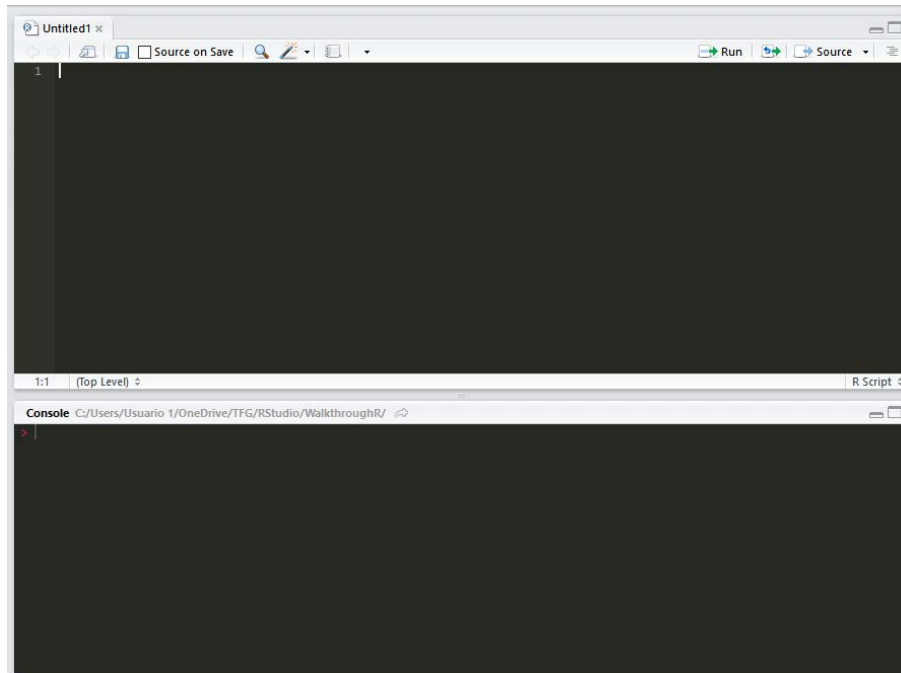
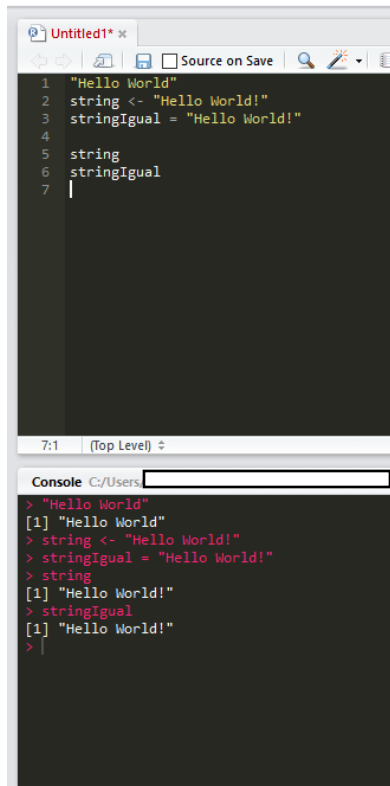


Figura 29. Parte izquierda de RStudio, dividida para la generación de scripts.

Como ejemplo ilustrativo de uso, mostraré a continuación, varias formas de escribir *Hello World* y lo que se muestra en el entorno. En la Figura 30 podemos ver varias formas de implementar el conocido *Hello World*. En la línea 1, bastan unas comillas para que cuando se ejecute el *script* aparezca en consola (2 primeras líneas de la consola). En las líneas 2 y 3, tenemos otra manera de proceder. En esos casos, se guarda en una variable. Ambos casos son idénticos, en uno se emplea '`<-`' y en el otro '`=`' para asignar "*Hello World!*" a la variable de su izquierda. Sin embargo, si observamos la consola, **no** se muestra por pantalla directamente. Tenemos que nombrar (líneas 5 y 6) a la variable en cuestión para que se muestre. Estas variables quedan almacenadas, como muestra la Figura 31.

---

<sup>7</sup> Se puede trabajar también a pantalla completa o sin mostrar la consola, depende del desarrollador elegir la opción más cómoda y acorde a sus gustos.



```
1 "Hello World"
2 string <- "Hello World!"
3 stringIgual = "Hello World!"
4
5 string
6 stringIgual
7
```

```
> "Hello World"
[1] "Hello World"
> string <- "Hello World!"
> stringIgual = "Hello World!"
> string
[1] "Hello World!"
> stringIgual
[1] "Hello World!"
>
```

Figura 30. Hello World en RStudio.

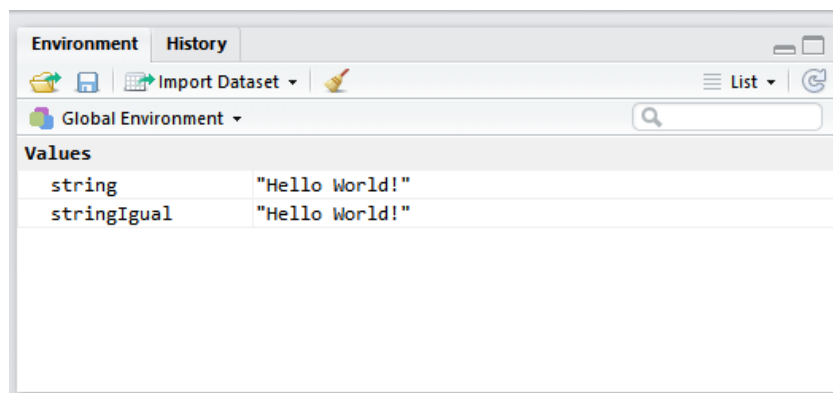


Figura 31. Entorno con las variables creadas.

Un último ejemplo básico será mostrar un *plot*, para ello se ha utilizado la llamada a la función *plot()*, pasándole como parámetros otra función llamada *seq()* que genera una secuencia de puntos siguiendo la siguiente sintaxis: *seq(valorInicial, valorFinal, numeroPuntos)*. Adicionalmente se ha añadido un título a la cabecera (*main*) del *plot* para hacerlo más vistoso y claro.

```
plot(seq(0, 10, length =10), main = "Plot secuencia de puntos")
```

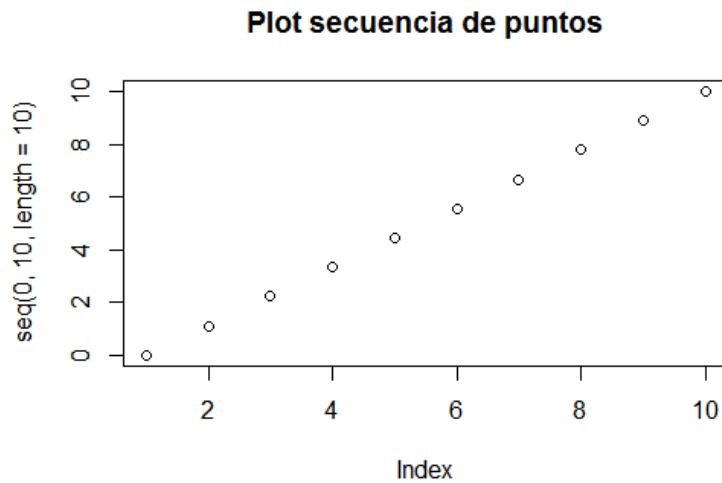


Figura 32. Ejemplo de plot generado con RStudio.

### 3. Caret

*Caret* (*Classification and REgression Training*) es un paquete pensado para simplificar el proceso de creación de modelos predictivos que añade algunas funcionalidades pero que su característica principal es la de simplificar la notación a la hora de programar, es decir, por debajo llama a la función pertinente con los parámetros que ha recibido de manera transparente para el usuario, simplificando la programación y la cantidad de código necesario para obtener el mismo resultado. En otras palabras, *caret* es una envoltura que provee al programador de una estructura sintáctica uniforme a la hora de modelar diferentes algoritmos [67].

Entre las funcionalidades que tiene *caret* se encuentra la **separación de datos**, herramientas de **pre-procesado** o **ajuste de parámetros** mediante re-muestreo [68].

Veamos que realizan algunas de estas funcionalidades. En primer lugar, la separación de datos o *data splitting* [69] es uno de los primeros conceptos que es importante destacar y tener en cuenta.

Dado un conjunto de datos, quiero evaluar el comportamiento de cierto modelo con esos datos. Para ello, es necesario que **pruebe** con ellos diferentes versiones del modelo hasta que consiga un resultado que consideremos apropiado. Esto que a priori es sencillo, trae consigo un problema, conocido como **Overfitting** o sobreajuste, que consiste en un **ajuste desmesurado** del modelo a los datos introducidos, en otras palabras, que el modelo puede acabar ajustando hasta el ruido, determinista o estocástico, asociado a esos datos (Figura 33).



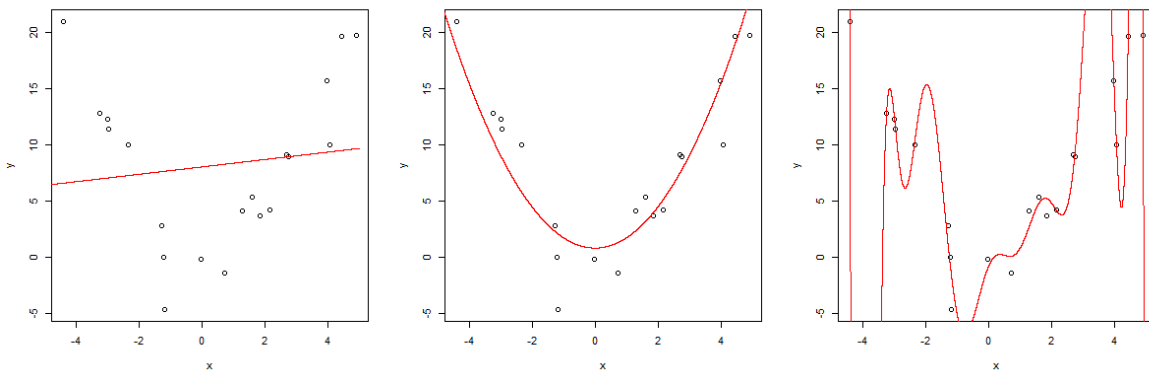


Figura 33. La imagen de la izquierda representa un modelo con problema de *underfitting* (o *subajuste*), la de la derecha presenta un problema de *overfitting* o *sobreajuste*, y la central sería un modelo adecuado.

Como podemos ver, el modelo de la izquierda es demasiado simplista y no se ajusta bien a los datos. Por su parte, el de la derecha del todo, es un modelo muy complejo que se ajusta muy bien a los datos de entrenamiento, pero esto ocasiona que, ante nuevos datos, no se ajuste en absoluto (ya que el modelo ha ajustado hasta el ruido presente en los datos de entrenamiento y por ello no generaliza bien). En cambio, en el centro de la Figura 33, vemos un modelo mucho más adecuado, con un compromiso entre complejidad y ajuste.

Entendido este concepto, la **separación de datos** se hace muy sencillo de entender. Antes de comenzar a tratar los datos, **reservaremos** una **parte** de los mismos para comprobar al **final** si el modelo empleado es bueno o no.

Este reparto, se hace en *caret* mediante la función *createDataPartition()* que detallaré en el análisis del caso. En función de los datos que tengamos la separación entre datos de **entrenamiento** y datos de **test** (que **no** tocaremos en absoluto) se hace en mayor o menor medida (80%-20%, 75%-25%).

Otra funcionalidad que requiere atención es el **pre-procesado**. En este caso, y ya después de separar los datos, debemos tener en cuenta la naturaleza de la información de la que dispongamos. Pongamos un ejemplo.

Tenemos datos de varios países, que se distribuyen en población, altura media de sus habitantes. La población se mide en millones de habitantes y la altura en metros o centímetros, que, de igual manera, estamos comparando cifras del entorno de  $10^6$ , con cifras alrededor a 1.

Esto puede suponer un problema al aplicar ciertos modelos y darle un peso mucho mayor a la población respecto a la altura o incluso, despreciar esta última. Por ello, el **pre-procesado** [70] nos sirve para ajustar los datos. La manera más usada en este ajuste es el **centrado** y **escalado** de los datos [71]. Centrar los datos se refiere a extraer el valor promedio a todos los datos. De esta manera, los datos tendrán media cero. Para escalar, dividimos los datos por su desviación estándar.

Un elemento no mencionado hasta ahora, son los **outliers**, datos que se desvían en gran medida del entorno en el que se mueven los datos. Empleando las alturas de las personas como ejemplo, si en una clase de 20 personas, 19 tienen alturas comprendidas entre los 165 cm y los 175 cm y la restante mide 210 cm, esta persona es el **outlier**.

Es importante determinar si en nuestros datos tenemos **outliers** y cuantos tenemos. Una vez conocido este dato tenemos que tener cuidado, que se salga de la media de los datos no siempre significa que sea un dato erróneo o que podamos prescindir de él [72]. En ocasiones lo que podemos tener delante

es un subgrupo poco muestreado pero válido. Siguiendo el ejemplo de la clase y las alturas, si ahora entran en clase 15 personas más y todas ellas tienen una altura superior a los 200 cm ya no podríamos contar como *outlier* a la persona que mide 210 cm.

Brevemente, y antes de explicar el ajuste de parámetros, comentaré una opción disponible (no siempre) para agilizar el procesado de datos. Se trata del *parallel processing* [73], que nos permite emplear varios núcleos del procesador a la vez.

En cuanto al **ajuste de parámetros**, introduzco ahora los modelos que posteriormente emplearé en el análisis para comentar de que se trata.

<b>Modelo</b>	<b>Tipo</b>	<b>Biblioteca</b>	<b>Parámetros de ajuste</b>
<i>Support Vector Machines (Linear Kernel)</i>	Clasificación/Regresión	<i>Kernlab</i>	C
<i>Support Vector Machines (Radial Basis Function Kernel)</i>	Clasificación/Regresión	<i>Kernlab</i>	C, sigma
<i>Red Neuronal</i>	Clasificación/Regresión	<i>Nnet</i>	Size, Decay
<i>Regresión logística regularizada</i>	Clasificación/Regresión	<i>Glmnet, Matrix</i>	Alpha, Lambda

Tabla 1. Modelos empleados y sus parámetros de ajuste.

Sin entrar en detalles todavía de los parámetros en sí, apuntar que, en estos modelos, podemos **probar** diferentes valores y ver cuál da mejor resultado.

Existen varios métodos para hacer este ajuste, en el caso de este trabajo, he empleado **10-Fold Cross-Validation** [74] que consiste en particionar en **10** partes los datos (los de entrenamiento) y en el primer ajuste coger todos los datos, excepto el primer subgrupo, en la siguiente iteración, todos salvo el segundo y así hasta terminar con todos. En cada iteración se evalúa la tasa de acierto (*Accuracy*) conseguido en el subgrupo no empleado en esa iteración para generar el modelo.

Finalmente se seleccionará aquella combinación de parámetros que den lugar a una mayor tasa de acierto medio (al promedio de las 10 iteraciones).

En el último capítulo del trabajo, se detallarán los pasos y los ajustes necesarios para realizar el análisis.



# 4

## Análisis realizado

En esta parte analizaremos en detalle los datos seleccionados para tal efecto, siguiendo una serie de algoritmos mencionados en la Tabla 1. El objetivo final es seleccionar el modelo que mejor se ajuste a los datos y que, por tanto, nos ofrezca el mejor resultado a la hora de clasificar un tumor en **benigno** o **maligno**.

### 1. Modelos empleados

#### 1. *Support Vector Machines (SVM)*

Es uno de los métodos más utilizados. Nace en los 60, de la mano de Vladimir Vapnik. En principio nace para clasificación, pero acaba evolucionando y es empleado también en regresión.

La idea básica del SVM es tratar de separar mediante un hiperplano las distintas clases. En el caso más sencillo, con 2 clases (como es el caso de **benigno/maligno**), vamos a suponer que inicialmente que son linealmente separables tal y como muestra el ejemplo de la Figura 34 Izq. Habría, en ese caso, infinitas posibilidades para solventar esto. Vlapnik introduce un *margen* que básicamente es medir la distancia de la muestra más cercana a la línea divisoria entre las clases a separar según vemos en la Figura 34 Drcha. A las muestras que cumplen esta definición se las denomina vectores de soporte [75].

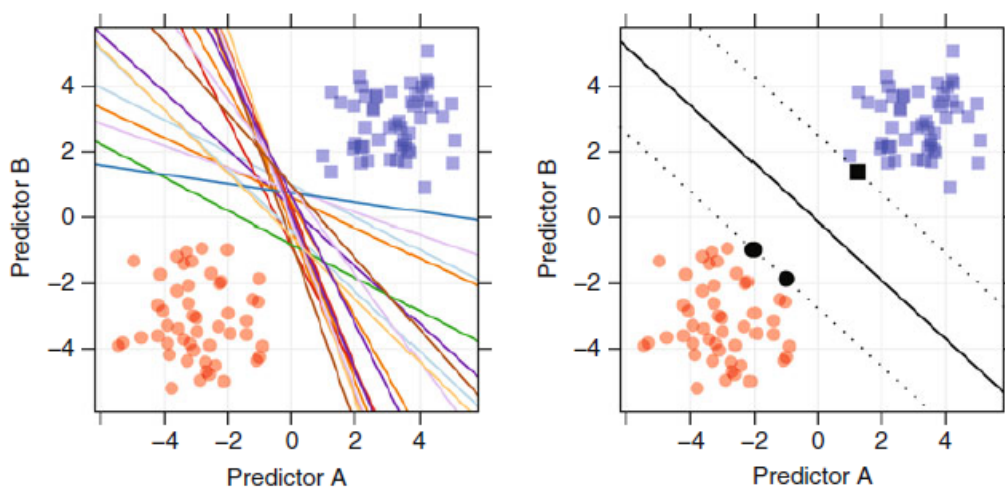


Figura 34. Izq: Posibles límites. Drcha: límite con support vector machines. Fuente: Applied Predictive Modeling.

Sin entrar mucho en la matemática que subyace a los SVM,  $D(\mathbf{u})$  es el criterio de decisión de una muestra nueva. Si el resultado es positivo se determina que pertenece a una clase y si es negativo a la otra.

$$D(\mathbf{u}) = \beta_0 + \sum_{i=1}^n y_i \cdot \alpha_i \cdot x'_i \cdot \mathbf{u} \quad (1)$$

$$D(\mathbf{u}) = \beta_0 + \sum_{i=1}^n y_i \cdot \alpha_i \cdot K(x'_i, \mathbf{u}) \quad (2)$$

Dónde  $x'_i$  son los vectores de soporte,  $y'_i$  la clase a la que pertenecen,  $\alpha_i$  parámetros de valor  $> 0$  encontrados al resolver un problema de optimización de programación cuadrática,  $n$  el número de vectores de soporte y  $\beta_0$  una constante cuyo valor puede calcularse fácilmente a partir de los datos anteriores.

El término  $K(x'_i, \mathbf{u})$  se refiere al *Kernel*, esta matización tiene su sentido en el momento en el que las componentes que queramos clasificar no sean linealmente separables, que, de serlo, con aplicar una definición **lineal** para *Kernel* bastaría.

En el momento en el que podemos cometer un error, es importante introducir un parámetro de penalización **C** denominado **coste** que, a cuanto más elevado sea, peor será cometer un error y, además, más fácil cometer *overfitting*.

#### a. *Kernel Lineal*

Es el caso más sencillo de SVM, en que se hace el producto  $x'_i \cdot \mathbf{u}$ . Esta operación explicada en términos sencillos, lo que hace es ver a que distancia está la nueva muestra ( $\mathbf{u}$ ) respecto de los vectores de soporte para decidir una clase u otra.

Este *Kernel* solo tiene un parámetro de ajuste, el ya mencionado coste **C**.

#### b. *Kernel Radial*

En este caso, el *Kernel*,  $K(x'_i, \mathbf{u})$ , busca en un espacio de dimensión superior para resolver casos no lineales. Esta función tiene forma de campana, siendo el tamaño de esta, uno de los parámetros de ajuste de este *Kernel*.

Como parámetros de ajuste tendremos el coste **C** y uno adicional, **sigma** ( $\sigma$ ) que define el tamaño de la campana. Si  $\sigma$  es bajo, la campana será ancha y viceversa [76].

## 2. *Redes Neuronales*

Las redes Neuronales son técnicas muy potentes de clasificación y regresión no lineal y se sustentan en las teorías de cómo funciona el cerebro humano [77].

El esquema mostrado en la Figura 35 muestra su esquema de funcionamiento [78].

Las unidades ocultas (*hidden units*) son combinaciones lineales de los predictores originales, que será uno de los parámetros de ajuste disponibles, denominado *size*. Cabe decir que, cuanto más sencillo sea el modelo, ajustándose suficientemente bien a los datos, mejor [77].

Este modelo tiene tendencia al *overfitting* por lo que se introduce otro parámetro de ajuste, el *weight decay*. Este parámetro atenúa el tamaño de los parámetros estimados, pudiendo conseguir una mejora en los resultados obtenidos.

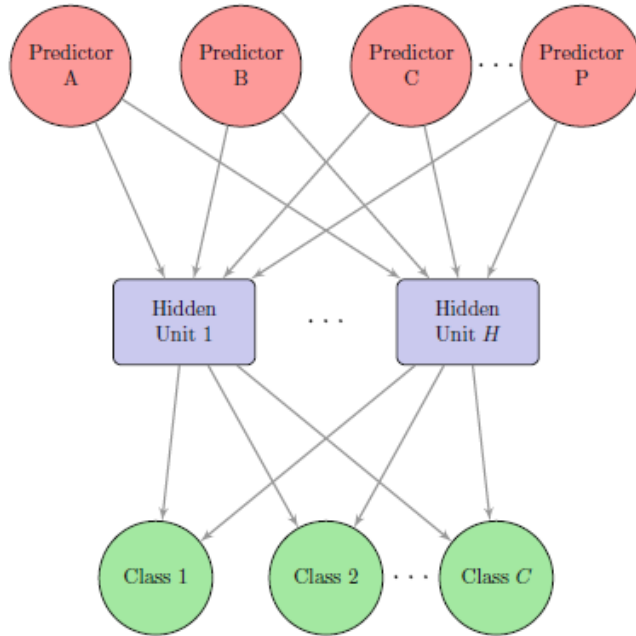


Figura 35. Esquema de funcionamiento de las redes neuronales. Fuente: Applied Predictive Modeling

### 3. Regresión logística regularizada

Es un modelo lineal, de los llamados **generalizados**, muy popular debido a su sencillez y que por defecto establece límites lineales. A este modelo se le puede añadir unos parámetros de ajuste,  $\lambda$  y  $\alpha$ . En el caso de  $\lambda$ , nos indica la cantidad de penalización que vamos a introducir en nuestro modelo y  $\alpha$  es un término de mezclado, esto es, al modelo sin penalización se le introducen 2 términos, uno cuadrático y otro del valor absoluto de los coeficientes de la regresión. El parámetro  $\alpha$  determina la importancia de la penalización debida a cada uno de esos dos términos [79].

$$\log L(p) - \lambda \cdot \left[ (1 - \alpha) \cdot \frac{1}{2} \cdot \sum_{j=1}^P \beta_j^2 + \alpha \cdot \sum_{j=1}^P |\beta_j| \right] \quad (3)$$

## 2. Resultados obtenidos

En esta sección mostraré y explicaré los resultados obtenidos tras la ejecución del código correspondiente para cada modelo elegido.

La primera parte del análisis consta de la adecuación de los datos, en la que debemos eliminar datos innecesarios y **separar** en parte de **entrenamiento** de la parte de **test**. En nuestro caso será un 70% a entrenamiento y un 30% para el *test*.

A continuación, mostraré los resultados por modelo, marcando los resultados finales. Estos resultados se compararán posteriormente para elegir el mejor modelo.

### 1. SVM con kernel radial

En el primer caso empleo el modelo *svmRadial*, dejando que el propio paquete *caret* estime un valor apropiado para el parámetro sigma y pruebe con 10 valores distintos del parámetro coste.

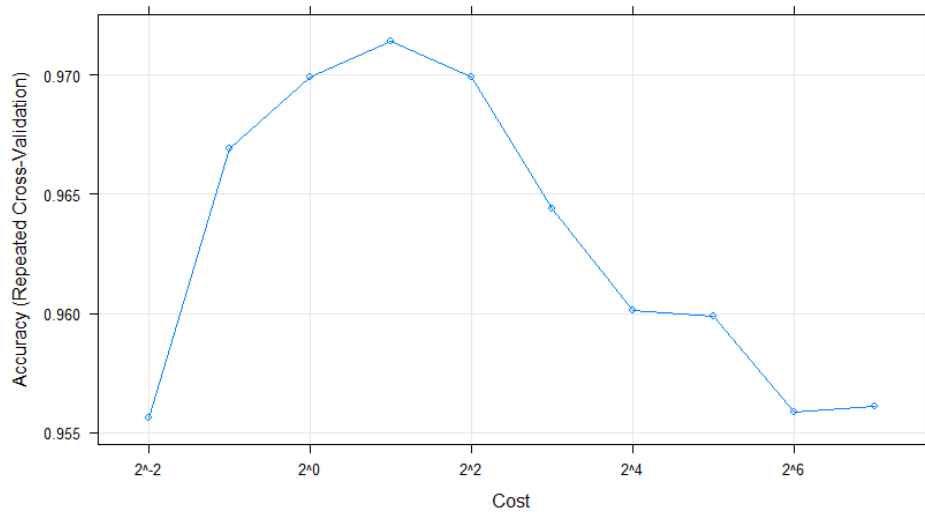


Figura 36. Visualización del resultado obtenido en el modelo SVM con kernel radial y parámetros genéricos.

Observando la Figura 36, vemos que alcanza el máximo de precisión para **coste 2** y **sigma 0.04386279**

Para afinar más, también he probado con una rejilla de parámetros, esto es, probar todas las combinaciones de coste y sigma posibles, tomando los **costes** 10 valores comprendidos entre **[0, 4]** y **sigma** valores comprendidos entre **[2<sup>-6</sup>, 2<sup>-4.8</sup>]**. Figura 37.

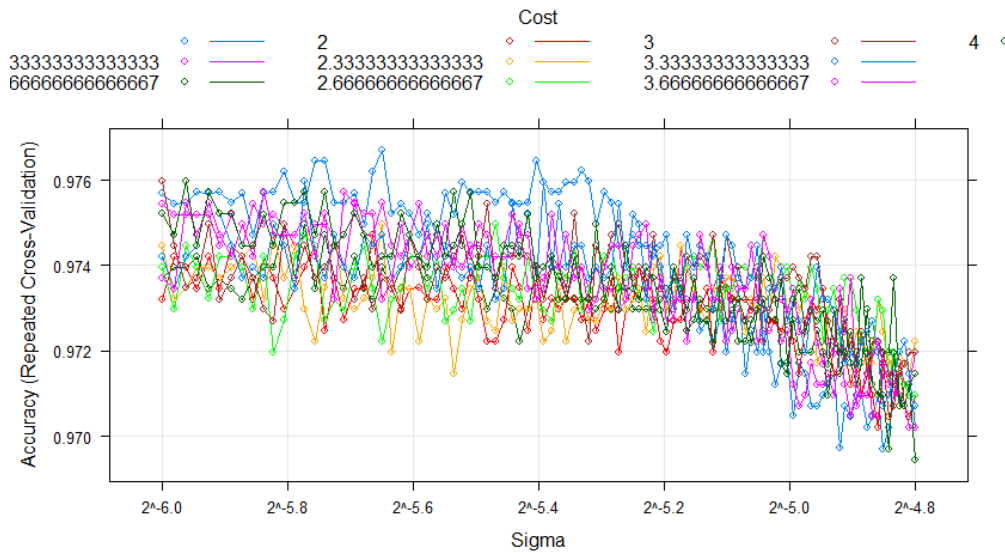


Figura 37. Resultados obtenidos tras el ajuste de parámetros en el modelo SVM con kernel radial.

Observamos como los resultados son muy parejos, siendo los parámetros finales: **C: 1** y  **$\sigma$ : 0.01992508**

## 2. SVM con kernel lineal

En este caso, al tratarse de un *kernel* lineal, solo tendremos un parámetro para ajustar, el coste **C**. En el primer barrido, elijo unos valores de coste comprendidos entre  $[2^{-10}, 2^5]$  con los que probar el modelo *svmLinear*, obtenemos lo representado en la Figura 38.

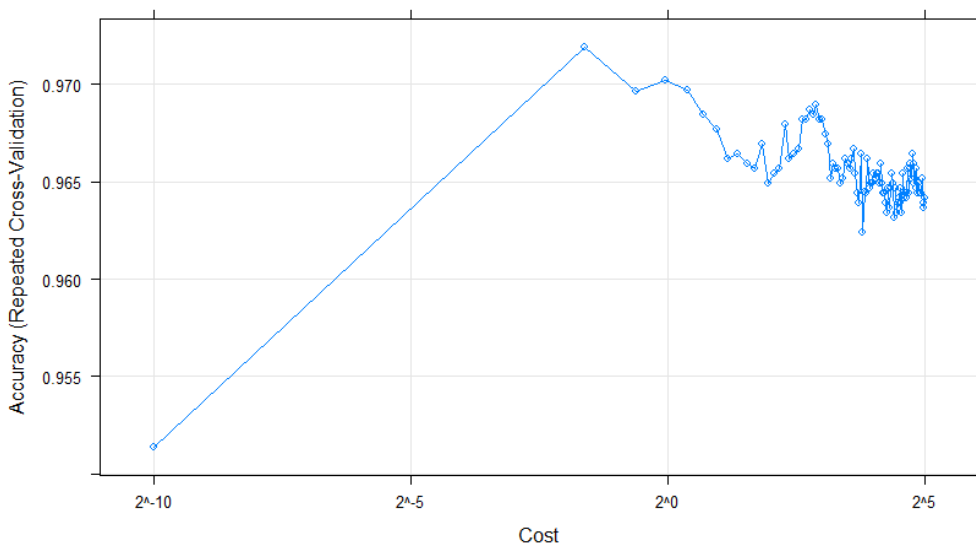


Figura 38. Resultado obtenido con el modelo SVM con kernel lineal y parámetros por defecto.

Ajustando los límites que recorre el parámetro **C**, obtenemos la siguiente gráfica.

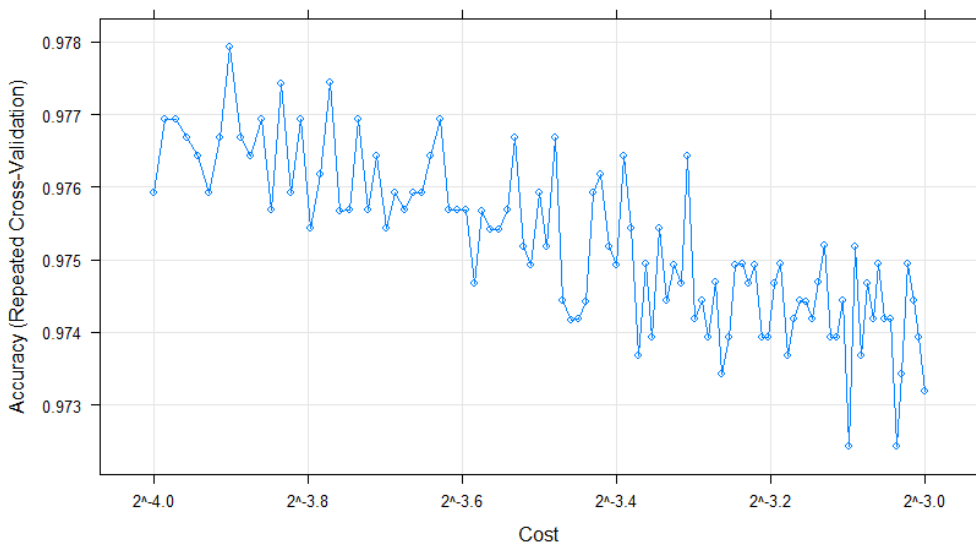


Figura 39. Resultados obtenidos con el modelo SVM con kernel lineal y parámetros ajustados.

El parámetro final es: **C: 0.06691919**



### 3. Redes Neuronales

En cuanto a las redes neuronales, utilizo el paquete *nnet*. Contamos con dos parámetros para ajustar, *size* y *decay*. En una primera aproximación, empleo un barrido amplio para ver la tendencia de los resultados. En el caso de las unidades de la capa oculta, probaremos con hasta 10 unidades. Para comprobar cómo evoluciona el parámetro *weight decay* tomaremos 100 valores comprendidos entre  $[0, 10]$ . Figura 40.

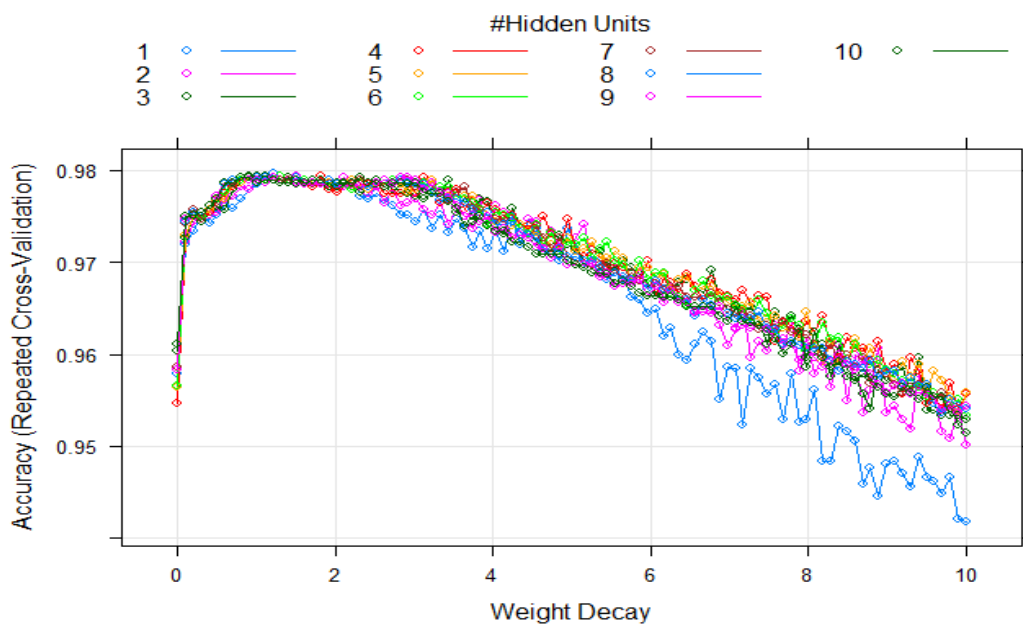


Figura 40. Resultados obtenidos con el modelo de redes neuronales y barrido general de parámetros.

Podemos apreciar una caída en la tasa de acierto a partir de 3, independientemente del número de unidades ocultas que utilicemos. Para acotar estos resultados, limitaremos el *weight decay* tomando de nuevo 100 puntos, pero esta vez entre  $[0, 2]$ . En el caso del *size* usaremos hasta 5 unidades ocultas al ver resultados muy parejos.

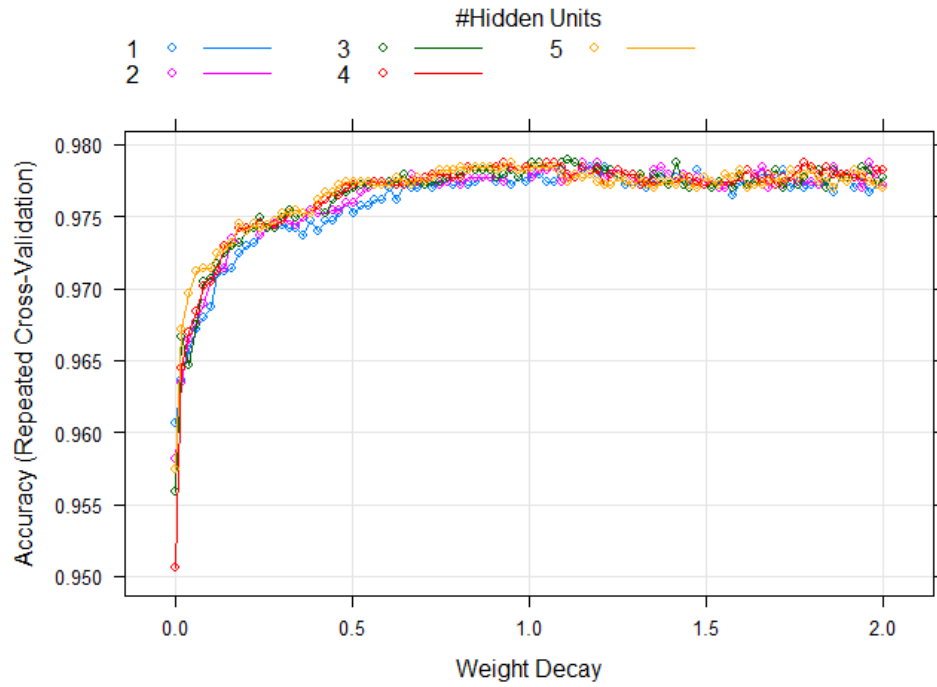


Figura 41. Resultado obtenido con el modelo de redes neuronales ajustando parámetros.

Como podemos apreciar, los resultados son muy parejos en todas las combinaciones. Obtenemos los siguientes resultados: **size: 3** y **decay: 1.111111**

## 4. Regresión logística regularizada

Por último, se analizó utilizando una regresión logística regularizada, empleando *glmnet*.

Se probó con una rejilla formada por 10 valores para cada uno de los dos parámetros (parámetro de regularización y porcentaje de mezcla), elegidos automáticamente por *caret*. Figura 42.

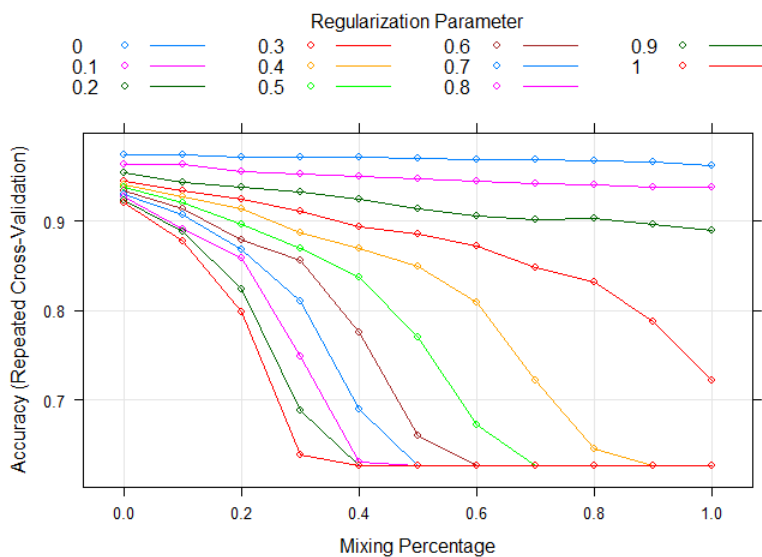


Figura 42. Resultados del modelo de regresión logística con penalización, con parámetros por defecto.

Podemos apreciar como disminuye la tasa de acierto cuando el porcentaje de mezclado aumenta. Por tanto, en el ajuste emplearemos el parámetro de regularización  $< 0.1$ . Concretamente, los parámetros que se van a emplear, tanto para la regularización como para el porcentaje de mezclado son  $\{0.1, 0.01, 0.001\}$  obteniéndose el resultado de la Figura 43.

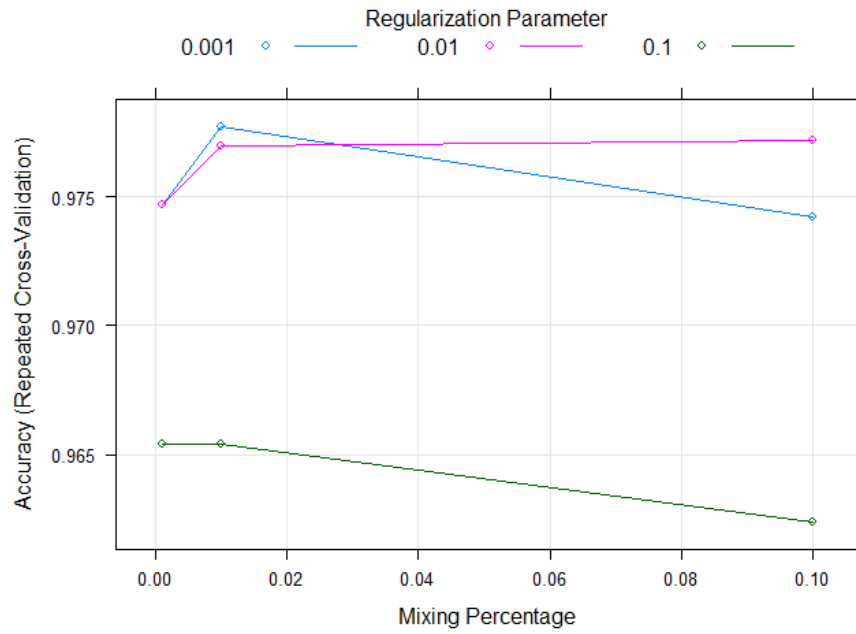


Figura 43. Resultado del ajuste para el modelo de regresión logística con penalización.

Podemos apreciar la diferencia de resultados en cuanto el parámetro de regularización es  $< 0.1$ .

Resultados finales:  $\alpha$ : **0.01** y  $\lambda$ : **0.001**.

### 3. Comparativa entre modelos

En la Tabla 2 recojo los parámetros finalmente seleccionados en cada uno de los modelos evaluados.

Modelo	Parámetros finales
SVM con <i>kernel</i> radial	C: 1 $\sigma$ : 0.01992508
SVM con <i>kernel</i> lineal	C: 0.06691919
Red Neuronal	<i>size</i> : 3 <i>decay</i> : 1.111111
Regresión logística regularizada	$\alpha$ : 0.01 $\lambda$ : 0.001

Tabla 2. Resultados obtenidos con cada modelo analizado.

Para comparar los resultados obtenidos en el proceso de validación cruzada, empleamos la función *resamples()* con la lista de modelos a comparar y visualizamos con *summary()* los resultados. En la Tabla 3 podemos ver los resultados obtenidos.

<i>Accuracy</i>							
	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.	NA's
<i>svmRadial</i>	0.9250	0.9500	0.975	0.9767	1	1	0
<i>svmLineal</i>	0.9231	0.9550	0.975	0.9779	1	1	0
RedNeuronal	0.9250	0.9750	0.975	0.9794	1	1	0
<i>glmNet</i>	0.9000	0.9744	0.975	0.9777	1	1	0

Tabla 3. Comparativa de los modelos empleados en términos de precisión.

Los valores representados en la Tabla 3 reflejan la composición del *boxplot*<sup>8</sup> de la Figura 44, que nos ayuda en la decisión a tomar.

<sup>8</sup> *Boxplot*: Elemento visual para resaltar la distribución de un conjunto de datos.

### Resultado de los modelos empleados

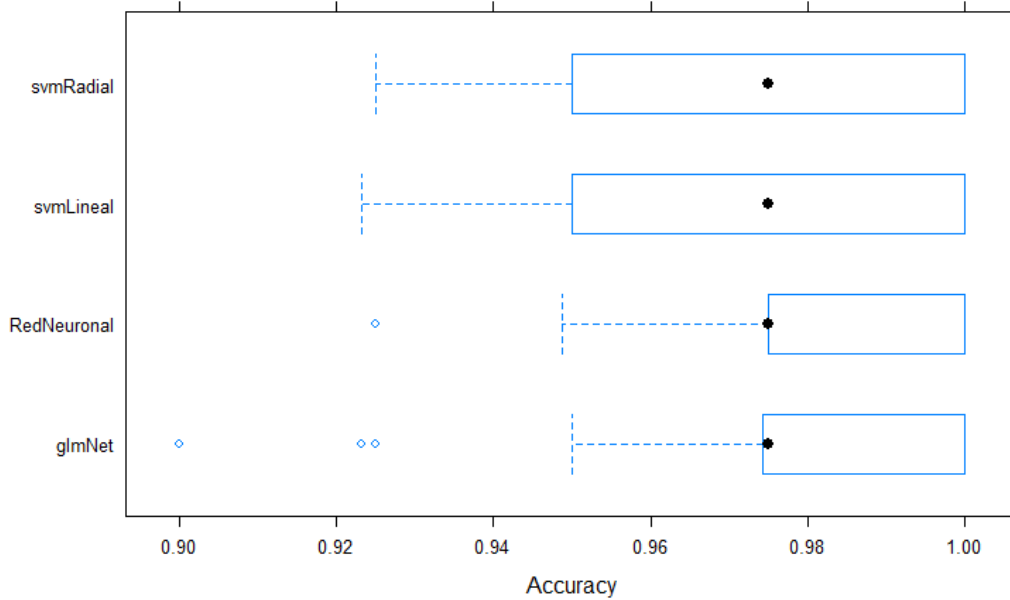


Figura 44. Boxplot de los resultados.

Tal y como podemos ver e interpretar junto con la Tabla 3, los modelos ofrecen unas prestaciones similares, destacando ligeramente la Red Neuronal, que posee una media un poco superior a la del resto de modelos y todos sus valores están más hacia la derecha. Para asegurarnos, podemos ver un *dotplot()* (Figura 45), donde se muestra la tasa de acierto media y el intervalo de confianza del 95%. Por tanto, el modelo elegido será la Red Neuronal, configurada con los parámetros mencionados ( $size = 3$  y  $decay = 1.111$ ).

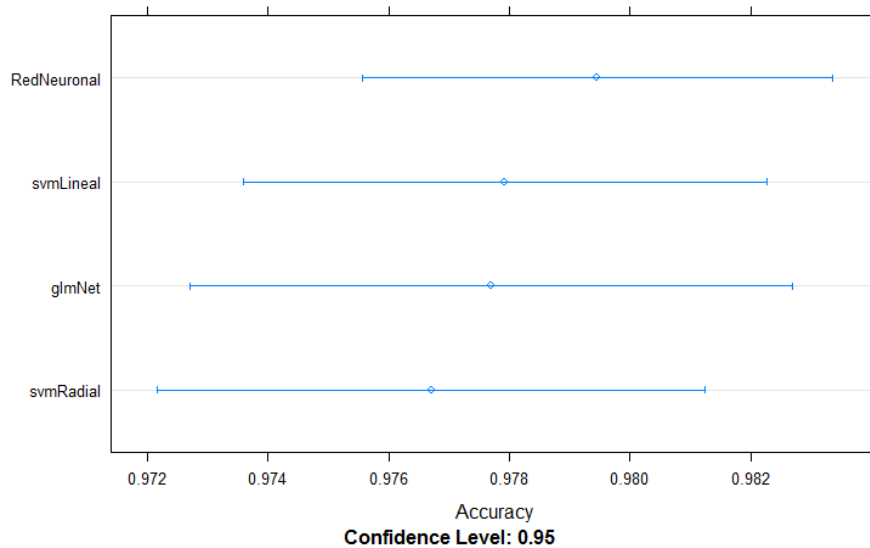


Figura 45. Intervalo de confianza del 95% de los modelos.

## 4. Modelo elegido y test

Una vez seleccionado el modelo basado en Red Neuronal, recordemos, unidades ocultas 3 y *weight decay* 1.111111, comprobaremos los resultados con los datos de *test*, esto es, un conjunto de datos que no se ha utilizado en ningún momento a la hora de generar el modelo.

Emplearemos la función *predict()* y *confusionMatrix()* para obtener los resultados (Tabla 4). Podemos apreciar una tasa de acierto del **97,67%**, un valor similar al estimado en el proceso de validación cruzado, lo cual da confianza en la validez del modelo generado al aplicarlo a nuevos datos.

Ahora bien, la matriz de confusión da información adicional. Muestra que hay 61 positivos (tumores malignos) correctamente identificados (TP, *True Positives*); 2 falsos positivos, esto es, identificar a un tumor como maligno cuando realmente no lo es (FP, *False Positives*); 105 negativos identificados correctamente (TN, *True Negatives*), es decir, tumores identificados correctamente como benignos; y 2 falsos negativos, esto es, identificar a un tumor como benigno cuando realmente es maligno (FN, *False Negatives*). A partir de estos valores pueden calcularse otros parámetros. La sensibilidad (*sensitivity*) es la proporción de positivos (tumores malignos) que se detectan correctamente, esto es,  $TP/(TP+FN) = 61/(61+2) = 96,83\%$ . Por otro lado, la especificidad (*specificity*) es la proporción de negativos (tumores benignos) que se identifican correctamente, es decir,  $TN/(TN+FP) = 105/(105+2) = 98,13\%$ .

<i>Matriz de confusión</i>		
<i>Predicción</i>	<i>Referencia</i>	
	<b>B</b>	<b>M</b>
<b>B</b>	105	2
<b>M</b>	2	61
B: Benigno / M: Maligno		
<i>Accuracy</i>	<b>0.9767</b>	
<i>Sensitivity</i>	<b>0.9683</b>	
<i>Specificity</i>	<b>0.9813</b>	

Tabla 4. Resultados con los datos de test.

A continuación, muestro la curva ROC (*Receiver Operating Characteristic*) del modelo, con un área debajo de la curva de 0.9907. Esta curva relaciona los aciertos y los fallos a través de la sensibilidad y de la especificidad. En nuestro caso, se acerca al valor ideal de 1.

La red neuronal es capaz de dar una estimación de la probabilidad de que un tumor sea maligno. Para etiquetarlo finalmente como maligno o no, se compara con un umbral (por defecto 0.5). Este umbral puede modificarse. Si, por ejemplo, lo reducimos, se etiquetará a un tumor como maligno, aunque el modelo no esté “muy seguro” de que realmente lo sea. De esta forma se consigue disminuir el número de falsos negativos (identificar a un tumor como benigno cuando realmente es maligno) y aumenta la sensibilidad. La contrapartida viene porque aumentará el número de tumores benignos que serán etiquetados como malignos (disminuyendo la especificidad). La curva ROC muestra precisamente la especificidad y la sensibilidad que se consiguen en función del umbral utilizado para decidir. A modo de ejemplo, en la gráfica se muestran también tres umbrales concretos (0.75, 0.5 y 0.05) y los valores de especificidad y sensibilidad que se obtienen en cada caso.

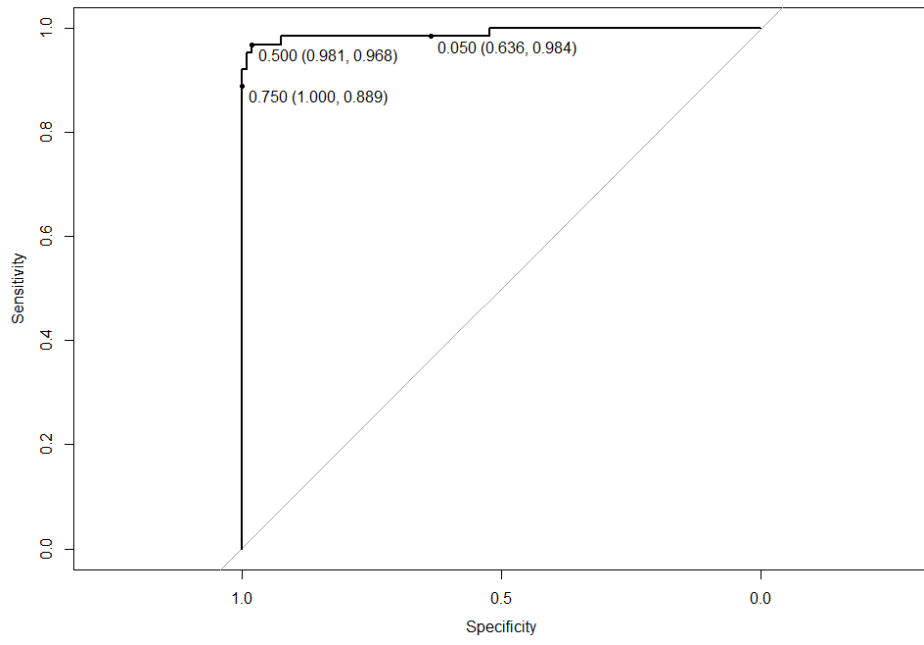


Figura 46. Curva ROC del modelo seleccionado.





# 5

## Conclusiones y líneas futuras

Una vez concluido el trabajo, creo conveniente resaltar la importancia que tienen todas las ramas del conocimiento y el esfuerzo necesario para integrar y completar unas áreas con otras. Vivimos en un mundo cada día más conectado, más sensorizado y por ello, más complejo y difícil de comprender si no es con ayuda de otros, colaborando e innovando nuevas formas de actuación y nuevas tecnologías que nos impulsen hacia el futuro.

Desde el sector TIC, tenemos una labor fundamental a la hora de encajar las piezas que compondrán la sociedad. El *Open Data* se puede considerar una herramienta de base con la que construir los cimientos de una sociedad eficiente, que gestione los recursos acordes a las necesidades de sus ciudadanos y que permita además lograr avances científico-técnicos en disciplinas muy variadas.

No hay que olvidar la responsabilidad que conlleva la tecnología, que, con cada avance, se producen inherentemente algunos riesgos adicionales, ya sean de seguridad o de transgresión de derechos fundamentales. Una buena regulación, ágil y funcional, es necesaria para el buen hacer y para la previsión de estas situaciones.

En los primeros capítulos del presente trabajo, se ha mostrado una imagen del estado actual del *Open Data* en el mundo y con más detalle en España. Pese a que la labor de recopilación y búsqueda de datos conlleva su tiempo, el poder facilitar fuentes de información, así como la manera de publicarlos de manera ordenada puede asentar las bases de futuros trabajos en la materia. Además, conocer comunidades de soporte, así como empresas que ofrecen sus servicios e infraestructuras, abre posibilidades a aquellos que no dispongan de los recursos necesarios para trabajar con datos abiertos.

En lo que respecta al apartado más técnico, en la resolución de un caso práctico, quisiera destacar la importancia de las herramientas de trabajo, que facilitan día a día la obtención rápida y fiable de resultados. Puesto que no contaba con conocimientos iniciales de aprendizaje automático ni del lenguaje R, una parte importante del esfuerzo dedicado a la elaboración de este trabajo ha estado dirigido a la formación en estos dos aspectos.

Concretamente se han aplicado diversos modelos de aprendizaje automático para resolver un problema de clasificación orientado a identificar tumores como malignos o benignos, utilizando un conjunto de datos publicado de forma abierta. Aunque todos los modelos ofrecían prestaciones similares, se ha selecciona finalmente una red neuronal (convenientemente parametrizada), la cual conseguía una tasa de acierto del 97,94% en la validación cruzada y del 97,67% cuando se aplicaba a un conjunto de datos de test.

Después de este breve y sencillo caso de uso, cabe plantearse futuras líneas de análisis. Por un lado, realizar este mismo estudio contando con información abierta de un gobierno, lo que permitiría recabar un número mayor de datos y con ello, la posibilidad de ajustar mejor los modelos. Por otro

lado, retomar la idea inicial, emplear datos del gobierno de España, para realizar un análisis sobre la incidencia de ciertas enfermedades relacionadas con las vías respiratorias en zonas con un índice de contaminación elevado.





# 6

## Referencias

- [1] Open Knowledge International, «El manual del Open Data,» [En línea]. Disponible en: <http://opendatahandbook.org/guide/es/introduction/>. [Último acceso: 15 01 2017].
- [2] Datos Gobierno de España, «Guía de aplicación de la NTI: reutilización de recursos de Información.,» [En línea]. Disponible en: [http://datos.gob.es/sites/default/files/20160726\\_guia\\_de\\_aplicacion\\_de\\_la\\_nti\\_reutilizacion\\_recursos\\_de\\_informacion\\_1.pdf](http://datos.gob.es/sites/default/files/20160726_guia_de_aplicacion_de_la_nti_reutilizacion_recursos_de_informacion_1.pdf). [Último acceso: 02 02 2017].
- [3] Hipertexto, «La Web Semántica,» [En línea]. Disponible en: [http://www.hipertexto.info/documentos/web\\_semantica.htm](http://www.hipertexto.info/documentos/web_semantica.htm). [Último acceso: 01 02 2017].
- [4] Hasso Plattner Institut y University of Mannheim, «The Linking Open Data cloud diagram,» [En línea]. Disponible en: <http://lod-cloud.net/>. [Último acceso: 02 02 2017].
- [5] W3C, «OWL Web Ontology Language,» [En línea]. Disponible en: <https://www.w3.org/TR/owl-features/>. [Último acceso: 01 02 2017].
- [6] W3C, «OWL 2 Web Ontology Language,» [En línea]. Disponible en: <https://www.w3.org/TR/owl2-overview/>. [Último acceso: 01 02 2017].
- [7] Administración electrónica del Gobierno de España, «Portal administración electrónica,» [En línea]. Disponible en: <https://administracionelectronica.gob.es/es/ctt/eni#.WJLoBPnhDDf>. [Último acceso: 02 02 2017].
- [8] Administración electrónica del Gobierno de España, «Portal administración electrónica. Normas Técnicas de Interoperabilidad,» [En línea]. Disponible en: [https://administracionelectronica.gob.es/pae\\_Home/pae\\_Estrategias/pae\\_Interoperabilidad\\_inicio/pae\\_Normas\\_tecnicas\\_de\\_interoperabilidad.html#DOCUMENTOELECTRONICO](https://administracionelectronica.gob.es/pae_Home/pae_Estrategias/pae_Interoperabilidad_inicio/pae_Normas_tecnicas_de_interoperabilidad.html#DOCUMENTOELECTRONICO). [Último acceso: 02 02 2017].
- [9] European Union Law, «European Union law,» [En línea]. Disponible en: <http://eur-lex.europa.eu/legal-content/ES/TXT/?uri=uriserv%3A114527>. [Último acceso: 10 02 2017].
- [10] Boletín Oficial del Estado, «L00090-00096,» 17 11 2013. [En línea]. Disponible en: <https://www.boe.es/doue/2003/345/L00090-00096.pdf>. [Último acceso: 20 02 2017].

- [11] Boletín Oficial del Estado, «L00001-00008,» 27 06 2013. [En línea]. Disponible en: <https://www.boe.es/doue/2013/175/L00001-00008.pdf>. [Último acceso: 20 02 2017].
- [12] Boletín Oficial del Estado, «BOE 298 de 14/12/1999,» 14 12 1999. [En línea]. Disponible en: <https://www.boe.es/boe/dias/1999/12/14/pdfs/A43088-43099.pdf>. [Último acceso: 20 02 2017].
- [13] Boletín Oficial del Estado, «BOE 276 de 17/11/2007,» 17 11 2007. [En línea]. Disponible en: <https://www.boe.es/boe/dias/2007/11/17/pdfs/A47160-47165.pdf>. [Último acceso: 20 02 2017].
- [14] Boletín Oficial del Estado, «Disposición 7731 del BOE núm. 164 de 2015,» 10 07 2015. [En línea]. Disponible en: <https://www.boe.es/boe/dias/2015/07/10/pdfs/BOE-A-2015-7731.pdf>. [Último acceso: 20 02 2017].
- [15] Boletín Oficial del Estado, «Disposición 17560 del BOE núm. 269 de 2011,» 08 11 2011. [En línea]. Disponible en: <https://www.boe.es/boe/dias/2011/11/08/pdfs/BOE-A-2011-17560.pdf>. [Último acceso: 20 02 2017].
- [16] Gobierno de España, «Agenda digital,» [En línea]. Disponible en: <http://www.agendadigital.gob.es/agenda-digital/planes-anteriores/Paginas/plan-avanza2.aspx>. [Último acceso: 07 02 2017].
- [17] Junta de Castilla y León, «BOCyL n.º 49 12-marzo-2015,» 12 03 2015. [En línea]. Disponible en: <http://bocyl.jcyl.es/boletines/2015/03/12/pdf/BOCYL-D-12032015-3.pdf>. [Último acceso: 20 02 2017].
- [18] Junta de Castilla y León, «Portal de Gobierno Abierto,» [En línea]. Disponible en: [http://www.gobiernoabierto.jcyl.es/web/jcyl/GobiernoAbierto/es/Plantilla100/1284247727568/\\_/\\_/\\_](http://www.gobiernoabierto.jcyl.es/web/jcyl/GobiernoAbierto/es/Plantilla100/1284247727568/_/_/_). [Último acceso: 11 02 2017].
- [19] World Wide Web Foundation, «Open Data Barometer,» [En línea]. Disponible en: <http://opendatabarometer.org/3rdEdition/report/#rankings>. [Último acceso: 09 02 2017].
- [20] World Wide Web Foundation, «ODB Global Report Third Edition,» [En línea]. Disponible en: <http://opendatabarometer.org/doc/3rdEdition/ODB-3rdEdition-GlobalReport.pdf>. [Último acceso: 09 02 2017].
- [21] Open Data Barometer, «3rd Edition Methodology,» [En línea]. Disponible en: <http://opendatabarometer.org/doc/3rdEdition/ODB-3rdEdition-Methodology.pdf>. [Último acceso: 21 02 2017].
- [22] Open Data Barometer, «Rankings & Data,» [En línea]. Disponible en: [http://opendatabarometer.org/data-explorer/?\\_year=2015&indicator=ODB&lang=en](http://opendatabarometer.org/data-explorer/?_year=2015&indicator=ODB&lang=en). [Último acceso: 20 02 2017].
- [23] F. B. y. M. Kaltenböck, «1. FROM OPEN DATA TO LINKED OPEN DATA. Open Government & Open (Government) Data,» de *Linked Open Data: The Essentials. A Quick Start Guide for Decision Makers*, edición mono, 2012, p. 9.

- [24] The White House, «Memorandum for the heads of executive departments and agencies,» 08 12 2009. [En línea]. Disponible en: <https://www.treasury.gov/open/Documents/m10-06.pdf>. [Último acceso: 15 01 2017].
- [25] Congreso de Estados Unidos, «Library of Congress,» 2014. [En línea]. Disponible en: <https://www.congress.gov/bill/113th-congress/senate-bill/994>. [Último acceso: 20 01 2017].
- [26] Ministerio de Hacienda y Administraciones Públicas, «Buenas prácticas Open Data,» 09 2014. [En línea]. Disponible en: [http://datos.gob.es/sites/default/files/bestpractices\\_opendata\\_sep2014\\_1\\_1.pdf](http://datos.gob.es/sites/default/files/bestpractices_opendata_sep2014_1_1.pdf). [Último acceso: 20 01 2017].
- [27] Gobierno de Estados Unidos, «An Introduction to Smart Disclosure Policy,» [En línea]. Disponible en: <https://www.data.gov/introduction-smart-disclosure-policy/>. [Último acceso: 20 01 2017].
- [28] Gobierno de Reino Unido, «Open Data White Paper,» 06 2012. [En línea]. Disponible en: [https://www.gov.uk/government/uploads/system/uploads/attachment\\_data/file/78946/CM8353\\_acc.pdf](https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/78946/CM8353_acc.pdf). [Último acceso: 20 01 2017].
- [29] Gobierno de Reino Unido, «Public Data Principles for Data.Gov,» 06 2012. [En línea]. Disponible en: <https://data.gov.uk/library/public-data-principles>. [Último acceso: 20 01 2017].
- [30] Gobierno de Reino Unido, «G8 Open Data Charter,» 06 2013. [En línea]. Disponible en: [https://www.gov.uk/government/uploads/system/uploads/attachment\\_data/file/207772/Open\\_Data\\_Charter.pdf](https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/207772/Open_Data_Charter.pdf). [Último acceso: 20 01 2017].
- [31] Public Administration Select Committee, «Statistics and Open Data: Harvesting unused knowledge, empowering citizens and improving public services,» 17 03 2014. [En línea]. Disponible en: <http://www.publications.parliament.uk/pa/cm201314/cmselect/cmpublicadm/564/564.pdf>. [Último acceso: 20 01 2017].
- [32] midata innovation lab, «What will happen to my data – is it safe?,» [En línea]. Disponible en: <http://www.midatalab.org.uk/who-will-be-able-to-see-my-data/>. [Último acceso: 20 01 2017].
- [33] Gobierno de Reino Unido, «Data Protection Act 1998,» [En línea]. Disponible en: <http://www.legislation.gov.uk/ukpga/1998/29/contents>. [Último acceso: 14 02 2017].
- [34] Red.es, «Proyecto Aporta,» [En línea]. Disponible en: <http://www.red.es/redes/actuaciones/administracion-en-linea/aporta>. [Último acceso: 14 02 2017].
- [35] Administración electrónica. Gobierno de España, «PAe,» [En línea]. Disponible en: [https://administracionelectronica.gob.es/pae\\_Home/pae\\_Actualidad/pae\\_Noticias/Anio2016/Noviembre/Noticia-2016-11-25-Espa-a-lider-europeo-en-madurez-open-data-en-2016.html#.WKLU8fnhDDd](https://administracionelectronica.gob.es/pae_Home/pae_Actualidad/pae_Noticias/Anio2016/Noviembre/Noticia-2016-11-25-Espa-a-lider-europeo-en-madurez-open-data-en-2016.html#.WKLU8fnhDDd). [Último acceso: 05 02 2017].



- [36] CTIC, «Public Dataset Catalogs Faceted Browser,» [En línea]. Disponible en: <http://datos.fundacionctic.org/sandbox/catalog/faceted/>. [Último acceso: 20 02 2017].
- [37] Open Data Inception, «Open Data Portals Around the World,» [En línea]. Disponible en: <https://opendatainception.io/>. [Último acceso: 20 02 2017].
- [38] Ministerio de Sanidad, Servicios Sociales e Igualdad., «Barómetro Sanitario 2015,» [En línea]. Disponible en: [https://www.msssi.gob.es/estadEstudios/estadisticas/docs/BS\\_2015/PresentacionWebBS\\_2015.pdf](https://www.msssi.gob.es/estadEstudios/estadisticas/docs/BS_2015/PresentacionWebBS_2015.pdf). [Último acceso: 15 02 2017].
- [39] Ministerio de Sanidad, Servicios Sociales e Igualdad, «Consultas Interactivas del SNS,» [En línea]. Disponible en: <http://pestadistico.inteligenciadegestion.msssi.es/publicoSNS/comun/DefaultPublico.aspx>. [Último acceso: 15 02 2017].
- [40] Portal Europeo de Datos, «Salud. European Data Portal,» [En línea]. Disponible en: <https://www.europeandataportal.eu/data/es/group/health>. [Último acceso: 15 02 2017].
- [41] Eurostat, «your key to european statistics,» [En línea]. Disponible en: <http://ec.europa.eu/eurostat/web/main/home>. [Último acceso: 20 02 2017].
- [42] OCDE, «About the OCDE,» [En línea]. Disponible en: <http://www.oecd.org/about/>. [Último acceso: 15 02 2017].
- [43] Organización Mundial de la Salud, «Acerca de la OMS,» [En línea]. Disponible en: <http://www.who.int/about/mission/es/>. [Último acceso: 15 02 2017].
- [44] KDnuggets, «About KDnuggets,» [En línea]. Disponible en: <http://www.kdnuggets.com/about/index.html>. [Último acceso: 13 02 2017].
- [45] Asset Macro, «Macroeconomic Indicators - Financial data - Market data,» [En línea]. Disponible en: <https://www.assetmacro.com/>. [Último acceso: 20 02 2017].
- [46] Quandl, «Financial and Economic Data,» [En línea]. Disponible en: <https://www.quandl.com/>. [Último acceso: 20 02 2017].
- [47] Qlik, «Business Intelligence | Herramientas de visualización de datos,» [En línea]. Disponible en: <http://www.qlik.com/es-es>. [Último acceso: 20 02 2017].
- [48] Nine Sights, «Head Health Challenges,» [En línea]. Disponible en: <https://ninesights.ninesigma.com/web/head-health>. [Último acceso: 15 01 2017].
- [49] Crownd Analytix, «Crownd Analytix home,» [En línea]. Disponible en: <https://www.crowdanalytix.com/home>. [Último acceso: 20 02 2017].
- [50] KDnuggets, «Competitions,» [En línea]. Disponible en: <http://www.kdnuggets.com/competitions/index.html>. [Último acceso: 21 02 2017].

- [51] Kaggle, «Machine Learning from Disaster,» [En línea]. Disponible en: <https://www.kaggle.com/c/titanic>. [Último acceso: 20 02 2017].
- [52] CKAN, «The open source data portal software,» [En línea]. Disponible en: <https://ckan.org/>. [Último acceso: 21 02 2017].
- [53] CKAN, «About CKAN,» [En línea]. Disponible en: <https://ckan.org/about/>. [Último acceso: 10 01 2017].
- [54] Socrata, «What Is Socrata? Learn all about the Company,» [En línea]. Disponible en: <https://socrata.com/company-info/>. [Último acceso: 21 02 2017].
- [55] Amazon Web Services, «AWS | Cloud Computing - Servicios de informática en la nube,» [En línea]. Disponible en: <https://aws.amazon.com/es/>. [Último acceso: 21 02 2017].
- [56] Microsoft Azure, «Microsoft Azure: Plataforma de informática en la nube,» [En línea]. Disponible en: <https://azure.microsoft.com/es-es/>. [Último acceso: 21 02 2017].
- [57] Google, «Cloud Computing, servicios de alojamiento y APIs de Google Cloud,» [En línea]. Disponible en: <https://cloud.google.com/>. [Último acceso: 21 02 2017].
- [58] IBM, «Cloud Computing for Builders & Innovators,» [En línea]. Disponible en: <https://www.ibm.com/cloud-computing/>. [Último acceso: 20 02 2017].
- [59] Kaggle, «Breast Cancer. Wisconsin Data,» [En línea]. Disponible en: <https://www.kaggle.com/uciml/breast-cancer-wisconsin-data>. [Último acceso: 21 02 2017].
- [60] UCI Machine Learning Repository, «Breast-cancer-wisconsin,» [En línea]. Disponible en: <https://archive.ics.uci.edu/ml/machine-learning-databases/breast-cancer-wisconsin/wdbc.names>. [Último acceso: 21 02 2017].
- [61] O. L. M. y. W. H. Wolberg, Cancer diagnosis via linear programming, Volume 23, Number 5, SIAM News, 1990, p. 1 y 18.
- [62] Asociación Española Contra el Cáncer, «Cáncer de mama,» [En línea]. Disponible en: <https://www.aecc.es/SobreElCancer/CancerPorLocalizacion/CancerMama/Paginas/quees.aspx>. [Último acceso: 17 02 2017].
- [63] American Cancer Society, «Biopsia del seno,» [En línea]. Disponible en: <https://www.cancer.org/es/cancer/cancer-de-seno/pruebas-de-deteccion-y-deteccion-temprana-del-cancer-de-seno/biopsia-del-seno.html>. [Último acceso: 17 02 2017].
- [64] National Cancer Institute, «Grado de un tumor,» [En línea]. Disponible en: <https://www.cancer.gov/espanol/cancer/diagnostico-estadificacion/pronostico/hoja-informativa-grado-tumor>. [Último acceso: 17 02 2017].
- [65] R project, «About R,» [En línea]. Disponible en: <https://www.r-project.org/about.html>. [Último acceso: 16 02 2017].

- [66] ggplot2, «ggplot2,» [En línea]. Disponible en: <http://ggplot2.org/>. [Último acceso: 16 02 2017].
- [67] GitHub, «The caret package,» [En línea]. Disponible en: <http://topepo.github.io/caret/index.html>. [Último acceso: 21 02 2017].
- [68] GitHub, «The caret Package,» [En línea]. Disponible en: <http://topepo.github.io/caret/index.html>. [Último acceso: 16 02 2017].
- [69] M. K. y. K. Johnson, «4.3 Data Splitting,» de *Applied Predictive Modeling*, p. 67.
- [70] M. K. y. K. Johnson, «3. Data Pre-processing,» de *Applied Predictive Modeling*, Springer, pp. 27-49.
- [71] M. K. y. K. Johnson, «3.2 Data Transformations for Individual Predictors. Centering and Scaling.,» de *Applied Predictive Modeling*, Springer, pp. 30-31.
- [72] M. K. y. K. Johnson, «3.3. Data Transformations for Multiple Predictors,» de *Applied Predictive Modeling*, Springer, pp. 33-34.
- [73] Caret, «Parallel processing,» [En línea]. Disponible en: <http://topepo.github.io/caret/parallel-processing.html>. [Último acceso: 16 02 2017].
- [74] M. K. y. K. Johnson, «4.4 Resampling Techniques,» de *Applied Predictive Modeling*, Springer, p. 69.
- [75] M. K. y. K. Johnson, «13.4 Support Vector Machines,» de *Applied Predictive Modeling*, Springer, p. 344.
- [76] UNIVERSITY OF WISCONSIN–MADISON, «The Radial Basis Function Kernel,» [En línea]. Disponible en: <http://pages.cs.wisc.edu/~matthewb/pages/notes/pdf/svms/RBFKernel.pdf>. [Último acceso: 17 02 2017].
- [77] M. K. y. K. Johnson, «7.1 Neural Networks,» de *Applied Predictive Modeling*, Springer, p. 141.
- [78] M. K. y. K. Johnson, «13. Nonlinear Classification Models,» de *Applied Predictive Modeling*, Springer, p. 334.
- [79] M. K. y. K. Johnson, «12.5 Penalized Models,» de *Applied Predictive Modeling*, Springer, pp. 302-303.
- [80] Hipertexto, «Ontologías,» [En línea]. Disponible en: <http://www.hipertexto.info/documentos/ontologias.htm>. [Último acceso: 01 02 2017].
- [81] Open Knowledge International, «Global Open Data Index,» [En línea]. Disponible en: <http://index.okfn.org/place/>. [Último acceso: 09 02 2017].

[82] M. K. y. K. Johnson, *Applied Predictive Modeling*, New York: Springer, 2016.