UNIVERSITAT POLITÈCNICA DE VALÈNCIA

Escuela Técnica Superior de Ingeniería Informática

Herramientas de calidad del dato. Comparativa y metodología de selección.

Trabajo Fin de Grado

Grado en Ciencia de Datos

AUTOR/A: Maté Martínez, Jorge

Tutor/a: Cuenca González, María Llanos

CURSO ACADÉMICO: 2023/2024

UNIVERSITAT POLITÈCNICA DE VALÈNCIA

etsinf

Escola Tècnica Superior d'Enginyeria Informàtica

Escola Tècnica Superior d'Enginyeria Informàtica
Universitat Politècnica de València

Herramientas de calidad del dato
Comparativa y metodología de selección

TRABAJO FIN DE GRADO

Grado en Ciencia de Datos

Autor: Jorge Maté Martínez

Tutor: M. Llanos Cuenca González

Curso 2023-2024

Resumen

En tiempos donde producir y propagar información está al alcance de cualquiera, su calidad debería alzarse como un referente que nunca deberíamos soslayar.

La desinformación generada por un uso erróneo o tendencioso de las fuentes de datos, tan desgraciadamente común hoy en día, es un veneno potente que puede alterar la percepción de la realidad e inducir a tomar partido u optar por decisiones perjudiciales o puntos de vista sesgados pero aparentemente justificables por ese pecado original.

Esto nos puede afectar a nivel personal o lo que sería peor, ser utilizado para que terceros adoptasen decisiones condicionadas por el mal tratamiento –voluntario o no–, de cualquier tipo de datos.

Este trabajo pretende ser una humilde contribución en aras de estas afirmaciones. La veracidad –en gran parte fruto de la calidad de los datos subyacentes–, debería imponerse siempre a cualquier interés general o particular por muy convencidos que estemos de la necesidad de su adopción. No olvidar este precepto es lo que dota de fuerza moral y verdadera calidad a cualquier tarea que afrontemos desde los conocimientos adquiridos en el ámbito de la ciencia de datos.

La duda no es una presunción intelectual: debería ser el estímulo mismo que nos motive en la búsqueda de la verdad. A su vez, no deberíamos perder de vista y asumir nuestras limitaciones, y como afirma el viejo dicho, contemplar que la estadística no ha de utilizarse como los borrachos utilizan las farolas: sino para iluminarse y no para apoyarse en ellas.

Circunscritos a la índole de este estudio, pretendemos realizar una síntesis y una exploración somera de la normativa implicada, de algunas de las herramientas disponibles y de algunas técnicas, –luego aclararemos con más precisión estos conceptos–, que nos puedan ayudar a profundizar en la calidad del dato o al menos guiarnos para avanzar en la dirección correcta.

A la vez, transmitir al lector la importancia de que valore adecuadamente sus propias necesidades y que seleccione unos requerimientos de calidad acordes a las mismas basándose en criterios propios pero objetivos.

«Si tuviera más tiempo, hubiese escrito una carta más corta». *Blaise Pascal*

Palabras clave: Calidad del dato, Confiabilidad, Exactitud, *Big Data*

Índice general

Índice de figuras

CAPÍTULO 1
Introducción

Calidad, del lat. *qualitas*, según nuestro diccionario normativo[1], es la "propiedad o conjunto de propiedades inherentes a algo, que permiten juzgar su **valor**". También nos interesa su otra acepción como la "**adecuación** de un producto o servicio a las características especificadas".

Dato, del lat. *datum* "lo que se da", es "**información** sobre algo concreto que permite su conocimiento exacto o sirve para deducir consecuencias derivadas de un hecho".

A lo largo de un grado de ciencia de datos se proponen muy diversas técnicas para transformar datos en información, y esta en conocimiento que aporte valor.

Desde este punto de partida, parece oportuno afirmar que la calidad de un conjunto de datos representa la base de todo lo que se construya sobre ellos: Si está mal asentada, cualquier proceso o análisis posterior estarán condicionados por esta carencia inicial.

En pleno auge de la era del *big data*[2], la **gestión de la calidad del dato** (*DQM*)[3] ha asumido un papel crítico en pos de esta solidez. *DQM* incluye acciones, metodologías y técnicas que permiten comprobar que los datos tratados se ajustan a unos requerimientos específicos de calidad, vinculando de este modo los dos términos anteriormente descritos.

Así pues, podríamos intentar definir la *DQM* como un marco sistemático con el objeto de producir datos precisos, válidos y suficientes, mediante un proceso continuo que ajuste fuentes de datos, verifique la calidad de la información que proporcionan e implemente mecanismos que eliminen o minimicen la propagación de posibles errores, sin olvidar adecuar todo ello al *corpus* normativo aplicable en cada contexto y a la vez intentando asumir los estándares necesarios que garanticen la interoperabilidad durante su desarrollo, todo ello sujeto a unos costes asumibles.

Es una definición compleja que abarca muchos términos y que vamos a diseccionar y evaluar desde distintas ópticas, pero sin obviar su necesaria integridad.

Durante el camino que pretendemos recorrer, intentaremos obtener un mapa lo más preciso posible de un mundo en continua evolución, tal vez por ello impreciso, pero por eso mismo totalmente necesario.

Ante de entrar en materia, nos parece oportuno hacer un comentario sobre las fuentes consultadas para la redacción de este texto y las referencias a las mismas: Hace no demasiado tiempo, cualquier trabajo académico contendría citas a antecedentes simila-

[1] Diccionario de la lengua española de la RAE

[2] Este es un concepto bastante genérico, pero que aquí podemos asimilar al masivo incremento de las tres uves en el ámbito del tratamiento de datos: Velocidad, Volumen y Variedad.

[3] o *Data Quality Management*, en su acepción en inglés

res editados en un clásico formato impreso o a publicaciones científicas[4] que sirvieran para fundamentar sus afirmaciones, justificar la necesidad de su desarrollo o proceder a la refutación de las fuentes expuestas.

Esta aproximación ya no es tan factible hoy en día: La rápida evolución de las tecnologías involucradas, la dispersión y la variedad de los formatos y los soportes en los que se encuentra la información de interés, –*webs* institucionales o comerciales, publicaciones digitales, referencias cruzadas, hiperenlaces...– difuminan y confunden las fronteras de las fuentes empleadas y dificultan o cuanto menos no facilitan su trazabilidad.

En nuestro trabajo abundan las citas de autores externos. Cuando Los términos expuestos explican con solvencia y concisión el asunto tratado, se opta por incluir y citar la fuente original en vez de reelaborar su significado, persiguiendo la concisión pero intentando mostrar siempre su correcta atribución.

Como diría Roland Barthes, podemos incurrir en un «tejido zurcido con las citas provenientes de otros textos, con las innumerables influencias procedentes de otras fuentes...»

Aun a pesar de lo expuesto asumimos el compromiso de hacer un riguroso esfuerzo por mantener este espíritu de claridad tradicional y facilitar la correcta atribución de los contenidos que necesitemos citar.

Este texto pretende ser accesible –e interesar– a individuos sin cualificación en la materia, pero a la vez, no puede evitar incluir múltiples referencias técnicas en el ámbito informático: cierto conocimiento de la nomenclatura inherente al tema por parte del lector le facilitará enormemente su comprensión.

Como complemento a esta afirmación, también ponemos en tela de juicio la infalibilidad de cualquier información vertida en estas páginas. Instamos a quien pueda aportar alguna rectificación o consideración válida que pueda corregir alguna de las informaciones reseñadas que se ponga en contacto con el autor para su posible subsanación.

Aunque en gran parte este trabajo ha de ser recopilatorio y comparativo, condicionado incluso por la índole del título que precede a su contenido, pretendemos, en la medida de nuestro alcance y posibilidades, hacer un absoluto ejercicio de creatividad sin recurrir a asistentes ni fuentes de composición sintéticas como las proporcionadas por los tan en boga modelos de IA generativa. La pasión por el lenguaje es un atributo que suscribimos y que creemos sigue siendo por el momento humano.

1.1 Motivación

Existen distintas soluciones comerciales que mediante infinidad de técnicas enumerables: limpieza, *deduplicación*[5], normalización... –por citar solo algunas de ellas–, preparan los datos para ajustarlos a requerimientos específicos, establecen jerarquías o taxonomías y comprueban su grado de consistencia. Una revisión preliminar de las mismas puede proporcionar alguna pista sobre el marco teórico en el que se asientan y las necesidades que pretenden satisfacer.

Adicionalmente, unos datos de calidad pueden ser la fuente en la que se basen modelos y herramientas de análisis potentes, pero no hay que obviar el hecho de que también pueden estar sujetos a restricciones ineludibles.

[4] Los llamados *papers*.

[5] Este término no está recogido en nuestro diccionario normativo pero es muy utilizado en el contexto de la gestión de datos, por lo que hemos decidido incluirlo tal cual.

Uno de los principales usos de la *DQM* es el cumplimiento de normativas regulatorias y la verificación de su conformidad con las mismas.

En ámbitos donde se gestiona información personal existen códigos de obligado cumplimiento, –citemos solo el RGPD[6]–, que ha sido articulado para proteger los derechos y libertades de los individuos en lo concerniente al tratamiento de sus datos personales.

Otros, que obligan a identificar, clasificar y documentar la información gestionada, y otras normativas, como la inminente Ley de Inteligencia Artificial de la UE[7], que clasifican las aplicaciones de la IA[8] en categorías de riesgo, y que llegan a considerar algunas de ellas como inaceptables, como por ejemplo los sistemas de puntuación social gestionados por algunos gobiernos.

A título particular, el autor de este trabajo lleva años construyendo y proporcionando servicios de acceso a bases de datos de registro de actos médicos y agregación de información tanto estructurada como sin estructurar y de registros clínicos, concediendo accesos adecuados y protegiendo su contenido, por lo que ha experimentado en primera instancia la necesidad de atenerse a requerimientos legales, operativas y casos de uso particulares que garanticen la confidencialidad y la accesibilidad a los datos gestionados.

Muchos de estos problemas son abordados, resueltos y auditados por las modernas herramientas de evaluación y gestión de la calidad del dato que intentan resolver la operativa expuesta de un modo sistemático: Este hecho propicia abordar el tema tratado con sumo interés y expectativas.

A nivel más general, a lo largo de nuestra etapa como estudiantes se nos dota de conocimiento y se nos instruye en el uso de herramientas poderosas con la intención de convertirnos en científicos, lo que también nos otorga una responsabilidad: No debemos olvidar que nuestra obligación es dejar de lado nuestros prejuicios y mantener un perfil crítico que no nos aparte de la veracidad de los hechos, aunque estos vayan en contra de nuestras propias convicciones o intuición original. Si mantenemos esta premisa estaremos haciendo bien aquello para lo que se nos ha educado.

Establecida la importancia y la necesidad de la *DQM* y expuestas estas razones, se nos ocurren algunas cuestiones preliminares asumiendo el precepto de que formular preguntas certeras es casi tan importante como poder contestarlas.

- ¿Cómo adecuamos el concepto de calidad al ámbito de la ciencia de datos?
- ¿Qué herramientas son las más apropiadas para cada tarea?
- ¿Podemos definir y cuantificar métricas que sean indicativas de la calidad de un conjunto de datos?
- En caso afirmativo, ¿Cómo se pueden implementar con el menor coste posible?

Vamos a intentar dar respuesta a alguno de estos interrogantes o al menos proporcionar una guía para intentar ayudar al lector cuando pretenda seleccionar las herramientas que más de adecúen a sus necesidades.

Lo haremos desde dos enfoques: mediante una revisión de la información disponible en los ámbitos citados en el resumen previo[9], –con la intención de que el lector sea capaz

[6] Reglamento general de protección de datos, o *GDPR General Data Protection Reglament*, en su acepción en inglés.

[7] *EU Artificial Intelligence Act*, cuyo borrador final completo data de fechas tan recientes como el 21 de enero del 2024.

[8] A partir de aquí usaremos los términos IA e Inteligencia artificial indistintamente.

[9] Normativa, herramientas, técnicas.

de comprender y construir sus propios requerimientos de calidad–, y proporcionando algún método práctico que le permita ponderar los mismos y tomar decisiones en base a ello.

1.2 Objetivos

1.2.1. Generales

Hacer un breve compendio de la normativa, tan importante en nuestro caso, aplicable al dominio de la calidad del dato. Normativa como concepto referido tanto en el ámbito legal como en el técnico. (Normas de calidad y estandarización).

Revisar las principales herramientas comerciales disponibles en el mercado que ofrecen prestaciones en este ámbito de la calidad y gobernanza del dato[10] haciendo un breve resumen de las ventajas y desventajas de cada una.

Prestaremos especial atención a las que involucren elementos propios de la ciencia de datos o la inteligencia artificial en su operativa. En algunos casos solicitaremos al fabricante o proveedor del servicio una licencia o periodo de evaluación.

Hacer un repaso del concepto y las etapas relacionadas con el ciclo de vida inherente a la calidad del dato.

Encontrar y comentar trabajos académicos previos que nos hayan precedido, con la intención de confirmar su actualidad o constatar como lo postulado en los mismos ha sido superado por perspectivas más modernas.

En resumen, pretendemos generar un *vademecum*[11] relacionado con el título del trabajo.

1.2.2. Específicos

- Discernir algún tipo de criterio que nos permita **tomar decisiones en función de los datos a tratar y de su estructura**.

  Como veremos más adelante, la índole misma de los datos a tratar y su nivel de estructura van a condicionar su proceso y la definición de sus requisitos propios de de calidad. En la sección 2.4, que trata del ciclo de vida de los datos, nos referiremos a esta afirmación.

- Profundizar en la **revisión de las herramientas comerciales que parezcan más relacionadas con técnicas de la ciencia de datos**, indicando en qué aspectos.

  Para ello es necesario analizarlas en detalle y prestar más atención y comentar las que ofrezcan ese tipo de prestaciones. Esto se lleva a cabo a lo largo de la sección relacionada (3).

- Extraer de la herramientas anteriormente descritas unas dimensiones que podamos evaluar, ponderar o priorizar en función de necesidades específicas.

- Generalizar e inferir un marco teórico que sea deducible de sus propuestas y prestaciones.

  Esto se tratará principalmente en la sección 3.3, **características evaluables**.

[10] Este es un término sobre el que profundizaremos más adelante.

[11] En el sentido de «libro de poco volumen y fácil manejo que contiene las nociones y datos básicos de una disciplina». RAE. Diccionario panhispánico de dudas.

Todas abordan el marco de la calidad desde distintos enfoques. ¿**Podemos extraer generalizaciones?** En caso afirmativo, ¿hasta qué punto? ¿nos pueden ayudar la revisión de trabajos más académicos al respecto? La sección 4 explora esta perspectiva.

- **Plantear algún experimento que muestre cómo la calidad de los datos**, en alguna de sus vertientes, **afecta a las prestaciones de un modelo**.

  Una de las conclusiones a las que llegaremos es que actuaciones puntuales pueden ser muy beneficiosas para mejorar algún aspecto de la calidad de nuestros datos. Como contrapunto a esta premisa pretendemos evaluar cómo la variación de la calidad de los datos puede afectar al desempeño de un modelo. En (5) trataremos este asunto.

1.3 Estructura de la memoria

La presente memoria se articula comenzando con una breve introducción para poner en contexto el asunto tratado y continúa con motivaciones particulares y otras más generales que inducen al autor a interesarse por este campo.

Tras el establecimiento de estas premisas iniciales, consistentes en una breve descripción de los conceptos más genéricos del título del trabajo y la razón de la necesidad de implementar controles, calidad y gobernanza en nuestros datos, haremos un repaso del *corpus* normativo principal:

En fundamentos teóricos revisaremos tanto de la *normativa regulatoria* que juzgamos más afín al ámbito de la gestión de datos como de las *normas técnicas y de calidad* vinculadas con nuestro objeto de estudio. Introduciremos también el concepto de ciclo de vida en la gestión de la calidad de datos, donde abundaremos en este tipo de cuestiones.

En el siguiente apartado, revisión y propuesta de selección, es donde enumeramos y empezamos a desgranar las herramientas evaluadas y extraer información, conceptos y prestaciones diferenciales de las mismas:

Compendiaremos las distintas dimensiones más comunes del concepto de calidad inherente a un conjunto de datos y su posible clasificación, así como las prestaciones, que referidas a los mismos, ofrezcan distintos paquetes comerciales diseñados para la evaluación y mejora de la calidad de los datos de una organización.

Sin que esta evaluación sea exhaustiva, pretendemos proporcionar orientación sobre las principales virtudes y carencias de los mismos, y a la vez ofrecer una guía práctica y de consulta y referencia rápida.

Adoptaremos un enfoque *bottom-up*[12] ya que nos parece lógico primero explorar el uso y la taxonomía de los términos empleados por herramientas comerciales de alto nivel que ofrezcan aplicar criterios de calidad, y después, buscar generalizaciones a reglas formulables o principios más básicos que nos permitan aplicar el conocimiento compilado.

Al margen de herramientas comerciales, hemos encontrado diversos trabajos académicos que ya han abordado el tema que nos incumbe y seleccionado algunos de ellos por su interés sintetizando su contenido cuando nos ha parecido relevante.

No podríamos tampoco dejar de intentar aplicar algunas de la técnicas aprendidas: La última sección, calidad y ciencia de datos, pretende abordar esta cuestión.

En los apartados adjuntos se incluye el código utilizado.

[12] En el sentido de obtener reglas genéricas de casuísticas particulares.

Notas adicionales

Hemos recurrido a fuentes comerciales, intentando en todo caso en no incurrir en violaciones de derechos de autor o *copyrights*. En caso de que alguno de esas fuentes alegase algo en contrario expresamente manifestamos nuestra disposición a revisar cualquier contenido afectado.

Algunos de los gráficos –referidos como figuras en este trabajo–, están inspirados en fuentes externas, pero hemos reelaborado su mayoría para incluirlas y adecuarlas a nuestro discurso. Aún a pesar de ello, en dos casos los hemos incluido tal cual atribuyéndolos a sus autores y ambos incluyen el símbolo de ©.

En muchos contextos se hace necesaria una traducción a nuestra lengua de infinidad de términos ya empleados corrientemente en su versión original, pero que son usados por regla general sin la debida precisión, y que en ciertos momentos nos convendrá remarcar o señalar.

Recurriremos a notas a pie de página cuando la puntualización a realizar sea lo suficientemente escueta para no romper el hilo del discurso. En caso de necesitar hacer disquisiciones más completas recurriremos a artículos separados.

CAPÍTULO 2
Fundamentos teóricos

2.1 Normas y Calidad del Dato

Las normas podrían definirse como «Unas fórmulas que describen la mejor manera de hacer algo, y que son el resultado de un acuerdo internacional entre expertos.» [1]

La Organización Internacional de Normalización, o ISO[1] es la encargada establecer estos estándares internacionales. Algunas de ellas se centran en nuestro objeto de estudio, por lo que consideramos ineludible hacer al menos un somero repaso de las mismas.

2.1.1. ISO 8000

Empezamos citando la familia de normas ISO 8000 ya que está específicamente destinada a definir un marco para la gestión de la calidad de los datos y datos maestros, e incluir la definición de conceptos clave y especificaciones para la calidad de los datos.

Se subdivide en dos bloques diferenciados: el primero reúne toda la información relativa a los procesos de gestión de calidad (ISO 8000-6X) y el segundo todo lo que representa la gestión de datos maestros (ISO 8000-1x0). [2]

Cabe reseñar que este segundo bloque establece los roles de proveedor, consumidor o cliente y custodio de datos. Las distintas partes del estándar proponen usar un formato específico para el intercambio de mensajes entre las aplicaciones que hagan uso de datos maestros, y cuyos términos o atributos correspondientes sean almacenados en un diccionario.

Nos parece oportuno incidir en el hecho de que propone un **ciclo de mejora continua** de los procesos de gestión de la calidad del dato como el reflejado en el diagrama 2.1 adjunto.

También nos parece importante señalar que el ciclo arranca con un aprovisionamiento de recursos en el que se da importancia al *factor humano* en base a su formación.

2.1.2. ISO/IEC 25012

La norma ISO/IEC[2] 25012 establece el término de **modelo** de calidad de datos como «las características que se han de tener en cuenta a la hora de evaluar las propiedades de un producto determinado de datos». [4] Así pues, la calidad del producto de datos se puede entender como la adecuación del mismo a este modelo establecido.

[1] International Organization for Standarization.
[2] IEC *International Electrotechnical Commision.*

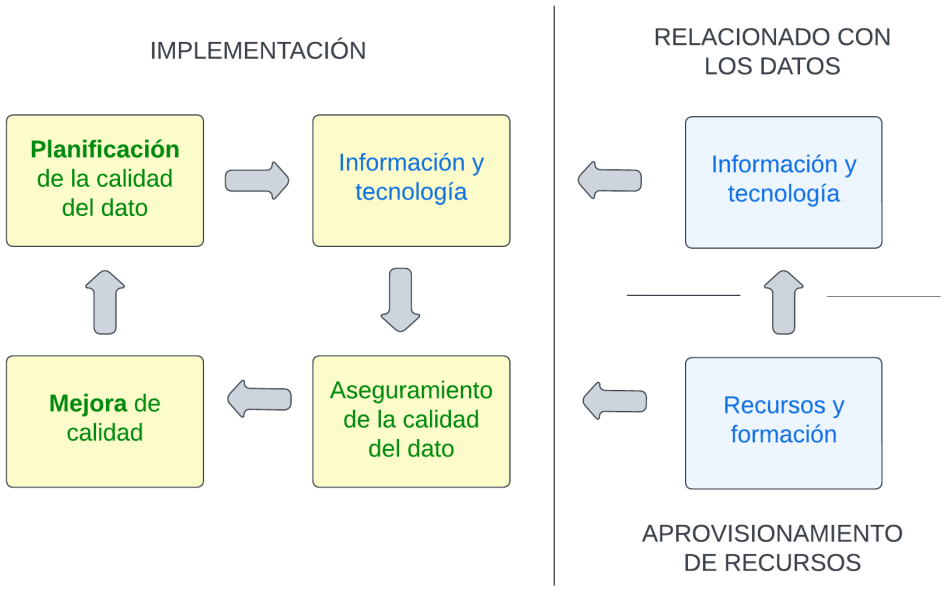

Figura 2.1: ISO 8000-61. Etapas del ciclo de implementación de gestión de calidad de los datos. Esquema adaptado de [3]

Clasifica las características de **calidad de datos** en dos categorías diferenciadas:

- Calidad de datos **inherente**.
- Calidad de datos **dependiente del sistema**.

La primera se refiere al potencial intrínseco de los datos tratados cuando se utilicen en condiciones concretas para satisfacer unas necesidades definidas referidas a su dominio y restricciones, (entendidas como *business rules*[3] requeridas) y su consistencia, (entendida como relaciones entre sus posibles valores y posibles metadados[4]).

La segunda al grado de calidad de datos alcanzado mediante el uso de algún sistema informático en particular, que lógicamente estará condicionada por sus prestaciones y buen desempeño a la hora de poder realizar, por ejemplo, copias de respaldo (que garanticen la recuperabilidad de los datos), o migración de los mismos a otro entorno para facilitar portabilidad.

Estas características son importantes y merece la pena definir su significado:

Exactitud. En sus vertientes sintáctica y semántica[5], entendida como el valor correcto del atributo referenciado.

Completitud. Grado en el que todos los atributos e instancias de los mismos tienen valores asignados.

Consistencia. Grado de coherencia de los datos en un contexto específico (que estén libres de contradicciones).

Credibilidad. Veracidad de la información representada por los datos.

Actualidad. O edad de los datos, con su necesaria adecuación a su contexto específico.

[3] O reglas de negocio.
[4] Un metadato es un dato sobre los datos.
[5] Función y significado.

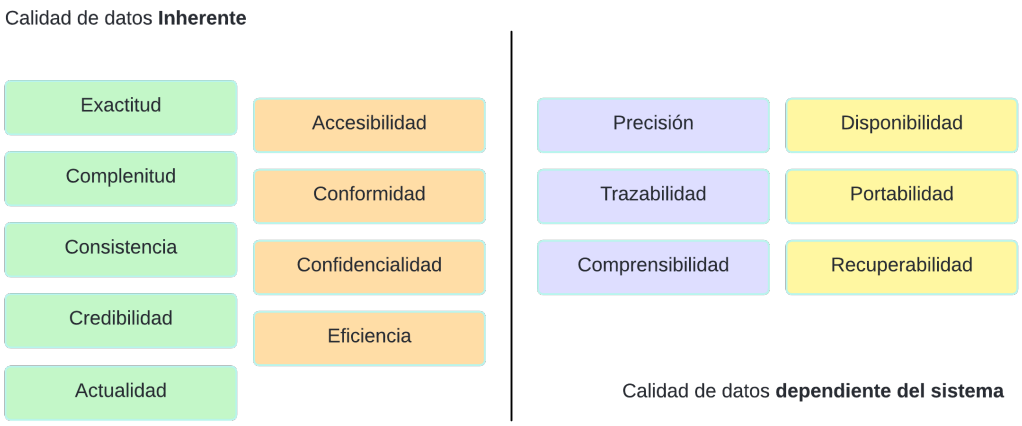

Figura 2.2: ISO 25012. Características de calidad de datos.
Esquema adaptado de [4]

Accesibilidad. Grado en que los datos pueden ser accedidos en su contexto, dando especial relevancia a su disponibilidad para usuarios con alguna discapacidad o limitación.

Conformidad. Nivel de adecuación de los datos a la normativa o regulaciones a los que puedan estar sujetos.

Confidencialidad. Grado de garantía de que los datos solo puedan ser accedidos por sus usuarios autorizados o legítimos.

Eficiencia. Nivel de adecuación de los datos para su proceso y capacidad de proporcionar un rendimiento satisfactorio durante el mismo mediante un uso razonable de recursos.

Precisión. Nivel de exactitud de los atributos de los datos.

Trazabilidad. Capacidad de los datos para generar un camino de acceso, recuperación y autoría a cualquier cambio generado.[6]

Comprensibilidad. Grado de interpretación de los datos por sus posibles usuarios.

Obviamos la definición de las características de **Disponibilidad**, **Portabilidad** y **Recuperabilidad** por estar suficientemente autocontenidas en su propio término.

La figura 2.2 incluye todas. Es interesante destacar que el grado de pertenencia a cada una de las dos categorías mencionadas no es absoluta. Cuanto más se acerque a los extremos cada característica enumerada, mayor será su grado de inclusión en la categoría correspondiente. Las características de las dos columnas interiores comparten cierto grado de pertenencia a ambas categorías –podríamos definirlas como híbridas–.

2.1.3. ISO/IEC 25024:2015

Es una norma perteneciente a la familia ISO 25000, que de un modo genérico, trata sobre la calidad del software.

«ISO/IEC 25024:2015 define las medidas de calidad del dato para una medición cuantitativa de la calidad del dato en términos de las características definidas en la norma anterior (ISO/IEC 25012)».

[6] Este concepto toma especial relevancia en contextos de datos sensibles, como datos médicos o clínicos.

ISO/IEC 25024:2015 contiene lo siguiente:

- Un conjunto básico de medidas de calidad del dato para cada característica.
- Un conjunto base de entidades, que son objeto de las medidas de calidad aplicables durante el ciclo de vida de los datos
- Una explicación de como aplicar las medidas de calidad del dato.
- Una guía para que las organizaciones definan sus propias medidas de requerimientos de calidad y evaluación.

Es reseñable que esta norma no define rangos de valores de medidas de calidad para puntuar niveles o grados porque estos valores vienen condicionados por cada sistema, por su naturaleza dependiente a su vez del contexto y por las necesidades de los usuarios.

Este estándar internacional pretende ser lo bastante genérico como para ser aplicado a cualquier conjunto de datos mantenidos en cualquier formato estructurado en un sistema computacional y usado por cualquier clase de aplicaciones.

Las personas involucradas en gestionar datos y servicios que incluyan datos son las principales beneficiarias de las medidas de calidad: Está destinado a usuarios que necesiten producir o implementar medidas de calidad del dato durante el ejercicio de sus funciones.

Nos interesa este componente de la norma porque incumbe al factor humano. Se definen los siguientes **roles** al respecto:

- **Adquiriente**. Individuo u organización que adquiere u obtiene datos de un proveedor.
- **Evaluador**. Individuo u organización que realice una una evaluación. Podría ser un laboratorio de pruebas, el departamento de calidad de una organización, una entidad gubernamental o un mero usuario.
- **Desarrollador**. Individuo o entidad que realiza actividades de desarrollo incluyendo especificación de requerimientos, análisis, diseño, implementación o pruebas a los datos durante su ciclo de vida.
- **Encargado de mantenimiento**, Quien opere (individuo u organización) o realice operaciones de mantenimiento sobre los datos.
- **Proveedor**. Quien mediante un contrato provea de datos o servicios relacionados al adquiriente de los mismos.
- **Usuario**. Quien haga uso de los datos con alguna función específica.
- **Gerente**[7] de calidad. Quien realice un control sistemático de los datos.
- **Propietario**. Quien asume la responsabilidad de la gestión de los datos y su valor económico, con autoridad legal y responsabilidad para establecer evaluaciones, recopilar, acceder, propagar, almacenar, e implementar medidas de seguridad y cancelación sobre los datos.

Cabe reseñar que este estándar –según su propia fuente–, contribuye al objetivo de desarrollo sostenible número 9: el relacionado con la industria, la innovación y las infraestructuras. [5]

[7] hemos traducido *manager* por gerente.

2.2 Normativa regulatoria

En esta sección hacemos una revisión de los códigos y regulaciones, en muchos casos de **obligado cumplimiento** y que afectan a la gestión de datos, en cuanto que pueden constituir una restricción a su recopilación, uso, cesión o cualquier otro tratamiento que nos planteemos realizar.

No olvidamos que el *leitmotiv* de este trabajo es la calidad del dato, pero asumimos que esta nunca deberá escapar de su marco regulatorio, sino que por el contrario habrá de armonizarse con él.

Aunque los contenidos que a continuación citamos son bien conocidos y de fácil acceso, consideramos útil hacer una breve reseña de los mismos por la relación que tienen con nuestro trabajo: Como hemos mencionado, la calidad de los datos ha de estar supeditada a su adecuación a la normativa vigente que siempre habrá de tenerse en cuenta.

Hagamos un somero repaso:

2.2.1. RGPD, Reglamento (UE 2016/679)

Lo citamos en primer lugar por ser el de más antigüedad.[8] Es el de contenido más genérico para nuestros propósitos pero lo incluimos porque incumbe a los derechos de las personas.

El Reglamento General de Protección de Datos[9] está referido a las personas físicas en lo que respecta al tratamiento de sus datos personales y a la libre circulación de estos.

Datos personales son «cualquier información relacionada con una persona identificada o identificable, también denominada *el Interesado*». [6] Entre ellos podríamos citar –por ejemplo, y sin ser un listado exhaustivo–, su perfil cultural, sus ingresos, dirección IP, datos de filiación, datos médicos...

Es de especial interés reseñar que el RGPD prohíbe expresamente el tratamiento de algunas categorías especiales de datos:

- Los de origen racial o étnico.
- Orientación sexual.
- Opiniones políticas.
- Convicciones religiosas o filosóficas.
- Afiliación sindical.
- Datos biométricos, genéticos o sanitarios.[10]

Otra característica digna de mención de este mismo reglamento es una clara distribución de roles: Se establecen las figuras de responsable, el encargado del tratamiento de los datos y el delegado de protección de datos, importante figura que se instaura como interlocutor con las autoridades de protección de datos, supervisor del tratamiento e instructor de los posibles usuarios.

[8] Abril 2016.

[9] *General Data Protection Regulation* GDPR en su acepción en inglés.

[10] Salvo lógicamente por razones médicas o de interés público esencial y siempre que obre consentimiento explícito.

2.2.2. Actas de Servicios Digitales (DSA) y de Marketing digital (DMA)

(UE 2022/2065 2022/1925)

La DSA y la DMA[11] constituyen un único conjunto de muy recientes reglas aplicables en el ámbito de la Unión Europea[12] con dos objetivos principales: [7]

- Crear un espacio digital más seguro en el cual los derechos más fundamentales de los usuarios de servicios digitales sean protegidos.
- Establecer unas reglas de juego que promuevan la innovación, el crecimiento y la competitividad tanto en el mercado europeo como globalmente.

Hemos dudado a la hora de hacer mención a estas normativas, pero la introducción de conceptos tan interesantes (en la DSA) como la desinformación, los patrones oscuros[13] o el contenido ilegal, –comunes hoy en día y tan dependientes de los datos vinculados que los sustentan–, y la definición y el establecimiento de los nuevos roles de coordinador de servicios digitales o de detectores de confianza,[14] nos han inducido a hacerlo.

Es reseñable también el establecimiento de reglas específicas para plataformas *on line* o motores de búsqueda que tengan más de 45 millones de usuarios mensuales en la UE. Éstas deben cumplir las obligaciones más estrictas del acta: El tamaño importa.

Por su parte, la DSA establece el también interesante concepto de *Gatekeepers*[15]. Estos serían grandes plataformas digitales que provean los llamados servicios centrales del plataforma[16], como por ejemplo motores de búsqueda en linea, distribuidores de *apps*, o servicios de mensajería.

Estos *Gatekeepers* estarán especialmente sujetos a cumplir con las obligaciones y las limitaciones que la DSA estipula.

2.2.3. Ley de Inteligencia Artificial de la UE

(Corrigendum, publicación prevista junio/julio 2024)

Nos parece también oportuno revisar la reciente acta de la Unión Europea sobre la inteligencia artificial. Su borrador final data del 1 de enero de 2024. Es importante reseñar que aún a pesar de la publicación de este borrador, el acta está actualmente todavía en desarrollo.

Es interesante en lo concerniente al nexo entre calidad y riesgo atribuido a los datos gestionados por la IA y es la más directamente relacionada con todos los aspectos o vertientes de la ciencia de datos.

¿Así pues, qué es la *EU IA Act*?

[11] *Digital Services Act* y *Digital Marketing Act*.
[12] Publicadas en octubre y septiembre de 2022 respectivamente.
[13] Los patrones oscuros o *dark patterns* son las estrategias que implementan algunos sitios *web* para inducir al usuario a realizar alguna acción que en principio no pretendía hacer. Están basados en el engaño, la ocultación de información o la manipulación. Como ejemplos se pueden citar descargar un programa, darse de alta en un servicio o completar un formulario.
[14] Traducido libremente de *trusted flagger*, que es quien posea una particular capacidad -otorgada por el coordinador se servicios digitales-, y experiencia para detectar identificar y notificar contenidos fraudulentos.
[15] Curioso término que podríamos traducir como Guardianes del Portal.
[16] *Core Platform Services*.

«El acta de la IA es una propuesta europea de regulación sobre la Inteligencia Artificial. Es la primera propuesta de regulación integral propuesta en cualquier lugar por un gran regulador». [8]

El acta asigna las aplicaciones de IA a tres posibles categorías de riesgo:

- Las de riesgo inaceptable.

  Como las ya implementadas en algunos países que por ejemplo hacen un uso indiscriminado y masivo del reconocimiento facial u otorgan un carné de puntos social[17]

- Las de alto riesgo.

  Como por ejemplo las que podrían explorar y calificar de un modo automático los CVs[18] o las solicitudes de candidatos a un puesto laboral, que están sujetas a requerimientos legales específicos.

- No reguladas.

  Que serían todas aquellas que quedasen fuera de las dos categorías anteriores.

Serían también reseñables los siguientes puntos concernientes a esta futura ley:

- La obligaciones impuestas por la ley atañen principalmente a los proveedores (o desarrolladores) de los sistemas considerados como de alto riesgo, independientemente de que tengan su sede o no en la UE mientras sus productos se utilicen en la misma.

- En dicha ley, el término de usuario se refiere a quienes desplieguen dichos productos, no a sus usuarios finales.

- Los modelos denominados GPAI[19] deberán estar documentados, proporcionar instrucciones de uso, publicar un resumen de los contenidos utilizados para su entrenamiento y cumplir con los derechos de autor.

  En función de su nivel de riesgo, si este es bajo y son de licencia libre bastará con que respeten los derechos de autor y publiquen un resumen de sus datos de entrenamiento. En caso contrario (riesgo alto) deberán realizar evaluación de sus modelos, pruebas adversarias[20] [9] y monitorizar e informar sobre incidentes o brechas detectadas en la seguridad del sistema.

En su página *web* oficial, ofrece una interesante utilidad para comprobar el grado de cumplimiento de esta nueva normativa por parte de herramientas de IA que cualquier usuario pretenda utilizar. Lógicamente como la normativa está todavía en desarrollo, esta se compromete a adoptar los subsiguientes cambios que se produzcan en la misma. [10] También incluye un calendario de de aplicación.

A su vez, la UE crea un nuevo regulador, *la Oficina Europea de la IA* para controlar, supervisar y hacer cumplir la futura ley de la IA.

[17] En el sentido de que un ciudadano puede llegar a perder derechos sociales, padecer restricciones de libertad e incluso estar sujeto a castigos por intentar ejercer ciertas libertades restringidas.

[18] Currículos.

[19] *General Purpose Artificial Intelligence*, o Inteligencia Artificial de Propósito General.

[20] Técnicas consistentes en el proceso de extracción de información sobre el comportamiento y las características de un sistema de *Machile Learning*.

2.2.4. Regulación de datos NO personales (UE 2023/2854)

El reglamento UE 2023/2854, conocido simplemente como Reglamento de Datos[21], complementa de un modo natural al RGPD al abarcar el tratamiento de datos no personales y concierne a datos generados y tratados por productos y servicios digitales.

Está previsto que su entrada en vigor sea en enero de 2025 y afectará a proveedores de servicios, desarrolladores y a productos digitales.

Este reglamento fomenta la compartición y el acceso a datos no personales y el que los usuarios puedan acceder libremente a la información proporcionada por ellos mismos para el uso de los productos o servicios anteriormente mencionados, promoviendo así la transparencia en la gestión de datos.

Este planteamiento, que postula el acceso a datos por defecto por parte de los individuos, obligará a los actores implicados cuanto menos a adecuar sus políticas de gestión y protección de datos.

2.2.5. El reglamento de identidad europea *eIDAS2* (UE 2024/1182)

Hemos dejado para el final la normativa que consideramos más ambiciosa desde una perspectiva técnica.

«El Reglamento (UE) 910/2014, conocido como reglamento *eIDAS*, es el marco legal vigente en Europa para la identificación electrónica y servicios de confianza en transacciones electrónicas. La Comisión Europea propuso en septiembre del año 2020 actualizar este reglamento, (pasando a denominarse coloquialmente *eIDAS2*).» [11]

Esta nueva legislación pretende redefinir y simplificar la identificación personal en cualquier transacción relacionada con bienes o servicios a través de una aplicación digital proporcionada por los gobiernos.

Se intenta así obviar la actual necesidad de recopilar datos personales reiteradamente: *eIDAS2* pretende facilitar y automatizar el intercambio seguro de datos personales para simplificar el modo en el que optemos a servicios digitales como por ejemplo solicitar a una beca, un seguro o un préstamo. El proveedor de dichos servicios se verá obligado a adaptarse a esta nueva operativa abandonando prácticas como cumplimentar formularios u obtener los datos del solicitante por correo electrónico.

«Integrando conceptos fundamentales del reglamento anterior, *eIDAS2* abraza la firma electrónica, el certificado cualificado de autenticación de sitio web, el sello electrónico y el servicio de entrega electrónica certificada. Estos elementos garantizan la autenticidad, integridad y seguridad en las transacciones digitales, marcando una nueva era en la interacción digital en la UE». [12]

Una vez más consideramos que lo pretendido por esta propuesta y operativa nos obliga a incluirla en nuestra relación: Los datos subyacentes a este tipo de servicios deberán fluir y compartirse bajo esta nueva operativa, a cuya adecuación se podrían atribuir nuevas métrica de calidad.

2.3 Algunos conceptos clave

La revisión de algunos conceptos es necesaria para facilitar la comprensión de los entresijos de la calidad del dato.

[21] Publicado en diciembre de 2023.

Sin ser una lista exhaustiva aquí se reseñan algunos de los que el autor entiende como principales y a los que se hace referencia en este texto.

Adicionalmente, al explorar las prestaciones de las herramientas evaluadas con posterioridad en otra sección (3), aparecerán otros términos importantes que en este caso se definirán en su contexto.

- Estandarización

  El formato de los datos ha de atenerse a criterios específicos que faciliten su proceso, almacenamiento y comprensión. Un claro ejemplo universal es el de las direcciones postales.[22]

- Integridad

  Referida a la fiabilidad de los datos durante todo su ciclo de vida, –su captura, almacenamiento, conversión, transmisión o integración–. En un contexto más amplio representaría la capacidad de los datos de mantener la información necesaria para su consistencia y adecuación al contexto por el que existen.

- Datos Maestros

  Los datos maestros[23], son los datos que cimentan el funcionamiento de cualquier organización o entidad y están referidos a sus proveedores, clientes, artículos gestionados y cualquier elemento clave o principal de su actividad.

- Registro de oro[24]

  Entendido como un concepto ideal en la gestión de datos, definido como la fuente única de verdad y que capture toda la información necesaria sobre la entidad a la que se refiere con una precisión absoluta. Aun a pesar de lo teórico del término, no interesa el concepto porque representaría el extremo opuesto a los registros duplicados, incorrectos o faltantes.

- Perfilado de datos[25]

  Consistiría en la «revisión de la fuente del dato para entender su estructura, contenido y relaciones para así identificar su potencial». [14] Podríamos entender y equiparar este concepto como asimilable al del análisis exploratorio de la ciencia de datos.

- Gobernanza de datos

  Término amplio que abarca las políticas y los procedimientos que se implementan para garantizar que los datos de una organización sean precisos desde su origen, –y que luego se gestionen adecuadamente–, mientras se recopilan, almacenan, manejan, accedan o eliminen.

- Datos abiertos

  Son datos que cualquiera puede usar, distribuir o reutilizar con la única condición de manifestar su fuente o atribuir su autoría.

  Es un concepto relacionado con el de de *portal de transparencia*.

[22] Direcciones erróneas o imprecisas pueden causar la devolución de envíos de correo tradicional. Tan solo en un año, el servicio postal del los E.E.U.U. gestionó más de 6.500 millones de envíos devueltos por incorrecciones en la dirección de los destinatarios. [13]

[23] del inglés *Master Data*.

[24] del inglés: *Golden Record*.

[25] *Data profiling*.

A nuestro entender una fuente de datos bien explotada puede y debe contribuir al beneficio general o comunitario:

Creemos en la necesidad de blindar ciertos tipos de datos contra un posible uso o abuso de los mismos con fines no lícitos o lucro de unos pocos, la necesidad del respeto a la propiedad intelectual y la adecuada retribución del coste de su recolección, pero en general y mientras no se vulnere ninguno de estos supuestos, abogamos por el beneficio de proporcionar acceso a datos de un modo público y transparente.

Los mencionados portales de transparencia se constituyen como lugar donde –entre otros servicios–, se proporciona acceso a los datos abiertos por parte de sus tenedores.

2.4 El ciclo de vida y la calidad del dato

2.4.1. ETL y ELT

Como parte del flujo de trabajo para su generación, la transformación de datos se ha considerado tradicionalmente como el proceso previo de preparar, formatear e integrar datos en un repositorio o almacén de datos para su posterior uso y análisis. Este proceso clásico se ha venido denominando como proceso de ETL[26], o extracción transformación y carga.

Hoy en día, gracias al incremento de las capacidades de proceso y almacenamiento propiciadas por la computación en la nube y el surgimiento de nuevos paradigmas, como los no tan novedosos *data warehouses*[27], o por el contrario, los que sí lo son: los *data lakes* –término que equivaldría a nuestra acepción de lago de datos–, ha surgido una nueva forma de abordar este proceso de preparación: la ELT[28] o extracción, carga y transformación.

Esta permutación de los términos no es trivial e implica cambios sustanciales de enfoque y procedimiento: Para el lector avezado resultará evidente que la diferencia está en el momento donde se produce el proceso de transformación.

« Para almacenes de datos que estén orientados a técnicas de *SQL (Structured Query Language)*, un esquema de datos será siempre necesario para interrogar a nuestras fuentes de datos, por lo que la transformación es realizada antes de la carga de los datos.»[29]

«Por el contrario, la nueva aproximación ELT entra en juego cuando el esquema, o el patrón de preguntas a los que someter a nuestros datos, nos es desconocido de antemano, por lo que la transformación es diferida a un momento posterior a la carga de los datos.» [15]

Según la descripción general de los *data lakes* [16] que ofrece la plataforma *Google Cloud*, «Un *data lake* es un repositorio centralizado diseñado para almacenar, procesar y proteger grandes cantidades de datos estructurados, semiestructurados o no estructurados. Puede almacenar datos en su formato nativo y procesar cualquier variedad de datos, ignorando los límites de tamaño.»[30]

«...proporciona una plataforma escalable y segura que permite a las empresas realizar las siguientes tareas: transferir cualquier dato desde cualquier sistema y a cualquier

[26] *Extract / Transform / Load*.
[27] Almacenes de datos.
[28] *Extract / Load / Transform*.
[29] *SQL*: Lenguaje de consulta estructurado para recuperar información de bases de datos relacionales.
[30] Es una cita literal completa, que decidimos incluir para analizarla en detalle.

velocidad (incluso si los datos provienen de sistemas que son locales, de la nube o de procesamiento perimetral); almacenar cualquier tipo o volumen de datos con fidelidad absoluta; procesar datos en tiempo real o en modo por lotes; y analizar datos mediante *SQL*, *Python*, *R* o cualquier otro lenguaje, datos de terceros o aplicaciones de estadísticas».

Cualquier sistema, cualquier velocidad, cualquier tipo, cualquier volumen, *batch/on-line*[31], cualquier lenguaje, cualquier aplicación estadística... Son muchas ausencias de restricciones, pero en cualquier caso parece un hilo interesante a seguir.

Vamos a intentar hacerlo en la siguiente sección, (2.4.2) pero no queremos concluir esta sin al menos mencionar otro término –más reciente–, que pretende fusionar las virtudes de los *data warehouses* y los *data lakes*: Se trata de los llamados *Data Lakehouses*[32].

«Un *Data Lakehouse* proporciona la flexibilidad y eficiencia en el coste de un *data lake* con las capacidades contextuales y de consulta rápida proporcionada por los *data warehouses*»

«Esto permite a sus usuarios utilizar el modelo de repositorio único propio de los *data warehouses* para un almacenamiento unificado, sin sacrificar la flexibilidad analítica proporcionada por los *data lakes* permitiendo así a los *data Lakehouses* sobresalir en las tareas analíticas y de *machine learning*». [17]

2.4.2. Herramientas de IA y calidad de datos

Por la familiaridad del autor con el lenguaje *Python* y mejor conocimiento de la plataforma *AWS*, vamos a revisar someramente las posibilidades de esta última en cuanto a gestión de *data lakes*.

Aparte de las más genéricas y conocidas librerías de *Python*, *boto3* [18] como *SDK*[33] de *AWS*[34], y *pandas* [19] por sus prestaciones para leer y escribir datos desde y hacia distintas fuentes, conviene detenerse en *AWS data wrangler* (ahora *AWS SDK for pandas* [20]), también implementada para *Python*.

Este es un «servicio *open source* que extiende la funcionalidad de las librerías pandas a *AWS*, conectando *DataFrames* y servicios de analítica de datos y de datos *AWS*» y «ofrece funciones abstractas para ejecutar tareas ELT ordinarias en carga y descarga de datos de almacenes, lagos o bases de datos a cualquier escala».

Destaca su capacidad para suministrar información sobre los datos tratados y su calidad, posibilitando el entrenamiento automático de modelos durante el flujo de datos.

AWS proporciona también la interesante utilidad *SageMaker*.

En la figura 2.3 mostramos un diagrama genérico del flujo de trabajo típico de esta potente utilidad. Nos interesa porque muestra que en los datos de entrenamiento se pueden combinar tanto los proporcionados por el usuario como los de modelos previamente entrenados por la plataforma.

En la documentación oficial de la misma, [21] se propone una completa lista de algoritmos y dominios de aprendizaje, tanto supervisados como no supervisados, y para el procesamiento de imágenes o de textos.

Incluye una interesante tabla de asignación de casos de uso a sus algoritmos adecuados con la siguiente estructura:

[31] Proceso por lotes on en linea.

[32] Es un término que intenta combinar las dos acepciones de las que proviene y de traducción muy forzada a nuestro idioma. Si lo intentásemos sería algo parecido a casas (o almacenes) de lagos de datos

[33] *Software Development Kit* o herramientas de desarrollo de software.

[34] *Amazon Web Services*.

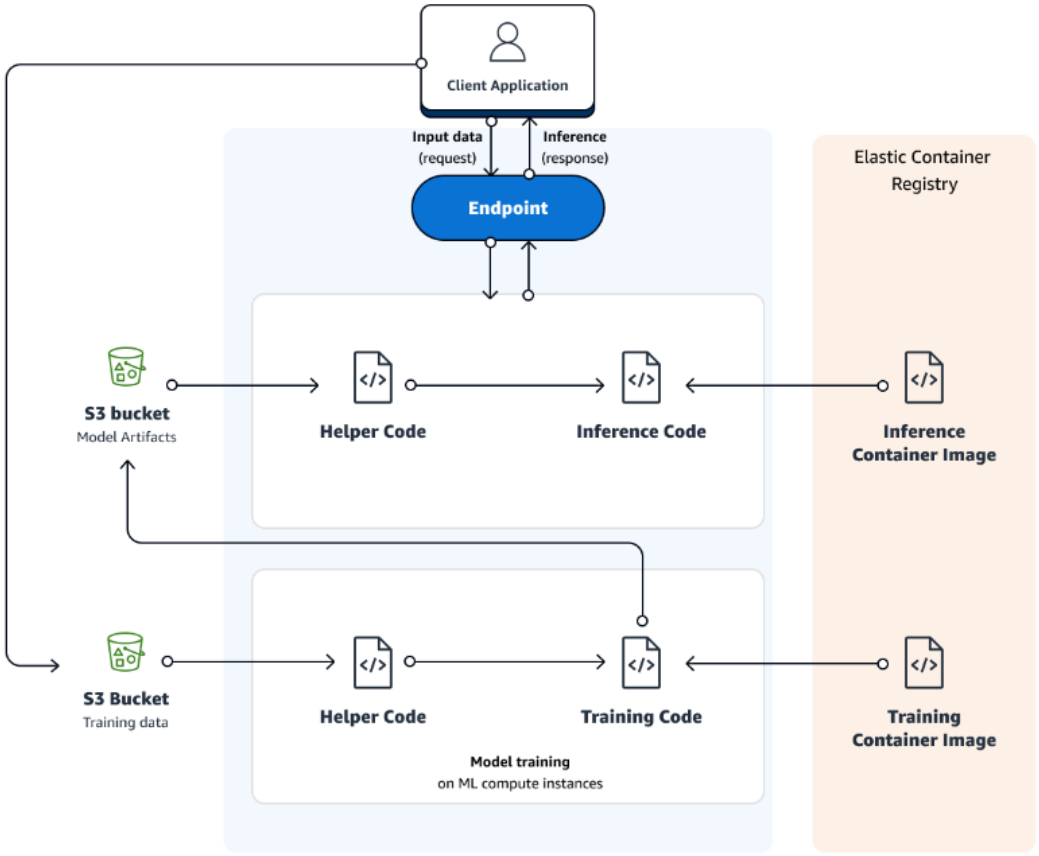

Figura 2.3: Ejemplo de implementación de un modelo con © *AWS Sagemaker*
Imagen tomada directamente de *AWS*.

1. Ejemplo de problema y casos de uso.
2. Paradigma o dominio de aprendizaje.
3. Tipos de problema.
4. Formato de datos de entrada.
5. Algoritmos integrados.

Ejemplo uno:

1. Asignar categorías predefinidas a los documentos de un *corpus*[35]: clasificar los libros de una biblioteca en disciplinas académicas.
2. Análisis textual.
3. Clasificación de textos.
4. Texto.
5. Algoritmo *BlazingText*[36], *Clasificación de textos - TensorFlow*[37].

[35]Según la RAE: Conjunto lo más extenso y ordenado posible de datos o textos científicos, literarios, etc., que pueden servir de base a una investigación.
[36]Este algoritmo es una actualización muy optimizada de los algoritmos de clasificación de texto y *Word2vec*.
[37]Plataforma y biblioteca de código abierto creada por *Google* para construir modelos de aprendizaje automático

Ejemplo dos:

1. Mejorar la incrustación de datos de los objetos de alta dimensión: identificar las incidencias de asistencia duplicadas o encontrar la ruta correcta en función de la similitud del texto en las incidencias.
2. Aprendizaje supervisado.
3. Incrustaciones: convertir objetos de alta dimensión en espacios de baja dimensión.
4. Tabular.
5. Algoritmo *Object2Vec*.[38]

Se ofrecen también modelos preentrenados para cada una de las propuestas de modelado y ofrece conectividad con *Apache Spark*[39], también con el propósito de entrenar modelos.

Se introduce también el uso de *Apache Arrow* [22] como plataforma de desarrollo de análisis en memoria[40] que contiene un conjunto de tecnologías que permiten a los sistemas de *big data* procesar y mover datos con rapidez.

Como conclusión, Si decidiésemos auditar y aplicar procesos de calidad a la gestión del dato, estaríamos muy interesados en incorporar todo esto a nuestra caja de herramientas.

También como contrapartida negativa pero de necesaria inclusión, habría que evaluar muy bien los costes de uso. Estos servicios tan especializados no suelen ser económicos.

2.4.3. Etapas del ciclo de vida de los datos

Al margen de cuestiones cualitativas, los datos tienen un recorrido propio que es necesario considerar.

Vamos a asumir como punto de partida el esquema básico proporcionado por el artículo publicado en la sección de tecnología del boletín del ESIC:[41] [23] en negrita las fases incluidas en dicha clasificación.

Extenderemos el contenido propuesto, que representa una visión muy general pero que tomaremos como base, para precisar o incluir en cada uno de sus puntos o intercalados, otros procesos que creemos importantes.

Creación de datos. En esta primera fase los datos se generan o se obtienen y se compilan. Puede requerir la extracción de características de algún proceso concreto y es muy importante alcanzar determinado grado de precisión como medio de garantizar la calidad de la información generada.

Limpieza. Necesaria localización y corrección o eliminación de registros de datos con errores en los datos adquiridos. Esta fase se suele englobar un proceso más amplio (*preparación*) que conlleva la adecuación de los registros a un mínimo necesario para su posterior almacenamiento.

Almacenamiento y organización. De un modo genérico, lo datos han de ser almacenados de una forma segura y accesible. Una estructuración y clasificación adecuadas

[38] Algoritmo de integración neuronal de uso general.
[39] Motor multilenguaje para ejecutar ingeniería o ciencia de datos y *machine learning* en máquinas de un solo nodo o *clusters* de varias de ellas
[40] *In memory analytics*.
[41] esic *Business & Marketing School*.

facilitarán cualquier proceso posterior. Los repositorios de datos pueden ser de infinidad de tipos con muy distintas prestaciones en función de su necesidad.

Procesamiento y análisis. Término amplio cuyo objetivo es extraer información de los datos, una información que a su vez aporte valor. En esta fase se pueden aplicar desde técnicas de minería de datos, –entendida como la detección de patrones, tendencias o conocimiento que pueda ser útil en la toma de decisiones–, hasta otras relacionadas con el aprendizaje automático.

Incluimos estas acciones dentro de su posible procesamiento:

Perfilado.[42], Acto de revisión del dato en el contexto de su fuente, entendimiento de su estructura, contenido y relaciones para identificar su potencial. Puede incluir procesos de análisis atómicos como la detección de valores faltantes, de valores únicos o análisis de rangos.

Clasificación y agrupamiento. «La clasificación de datos permite determinar y asignar valor a los datos de la organización y proporciona un punto de partida común para la gobernanza. El proceso de clasificación de datos clasifica los datos por confidencialidad e impacto empresarial a fin de identificar los riesgos. Cuando los datos están clasificados, se pueden administrar de formas que protegen aquellos confidenciales o importantes frente al robo o la pérdida» [24]

Deduplicación. La deduplicación, –en este caso y en oposición a ejemplos anteriores de este término *a nivel de conjuntos de datos*–, es un proceso consistente en la eliminación de copias excesivas de datos para reducir así la capacidad de almacenamiento. Ha de conjugarse con la necesidades de retención y copia expuestas más adelante. También puede entenderse como una compresión de los datos para eliminar copias duplicadas de datos repetidos.

Fusión. «La fusión de datos consiste en combinar datos de distintas fuentes en un solo conjunto de datos para realizar análisis posteriores o para almacenarlos en un almacén de datos. Las herramientas de fusión de datos ayudan a combinar datos y, generalmente, requieren preparación y estandarización antes de que se puedan fusionar los datos». [25]

Distribución y acceso. Implica la entrega y disponibilidad de la información procesada mediante la obtención de informes o visualización de resultados, o de cualquier otro modo que permita su acceso a sus debidos destinatarios. Este punto nos conduce a otros como son los de la privacidad y la seguridad de la información.

Implementación de *interoperabilidad* y *certificación*. La inmensa variedad de estándares disponibles para la gestión de datos es un factor en contra de su integración. Como medidas paliativas se diseñan normas técnicas de interoperabilidad, de obligado cumplimiento en diferentes ámbitos[43] en el que se hace hincapié en modelos de datos comunes. [26]

La interoperabilidad mencionada facilitará el diseño de conexiones vía *APIs* que brinden conectividad a los usuarios de los datos facilitando su distribución y acceso.

Retención y copia de seguridad. Aunque las modernas tecnologías de almacenamiento parecen obviar –por redundantes y seguras– el concepto de copia de seguridad, las política de seguridad y conservación de los datos no deben omitirse. Estas deben también incluir razones legales, regulatorias y operativas sobre la conservación de los conjuntos de datos.

Archivado y gestión de datos históricos. Como en el punto anterior, los modernos paradigmas de almacenamiento propuestos más arriba (*data lakes*, etc.) diluyen la separa-

[42] *Data profiling*.
[43] En España lo es en el de las administraciones públicas, que cuentan con su propio esquema nacional.

ción entre datos y datos históricos, pero esa misma razón impone también en este caso la necesidad de establecer políticas bien definidas de archivado y gestión de datos históricos que satisfagan los requerimientos necesarios.

Eliminación. (segura). Los datos obsoletos o descartados como no relevantes deberían ser purgados por razones no solo de economía de almacenamiento, sino también por regulaciones de seguridad. Algunos ámbitos –sanidad, finanzas, social–, son especialmente sensibles este tipo a de normativas e imponen un protocolo de eliminación segura que garantice la destrucción de los datos de un modo que no puedan ser recuperados de ningún modo.

CAPÍTULO 3
Revisión y propuesta de selección

Hemos dudado a la hora de titular de este capítulo *estado del arte*, ya que no comprende una revisión de *todas* las herramientas disponibles que podríamos haber evaluado.[1]

En nuestro caso hemos seleccionado, –sin ser un listado exhaustivo–, las que nos parecen ser las principales actualmente disponibles en el mercado que ofrecen gestión y análisis de calidad de datos.

Para confeccionar esta lista no nos hemos basado por un único patrón, criterio o proveedor, sino que la hemos elaborado explorando distintas fuentes y haciendo uso de nuestro propio criterio de selección: La publicidad de cualquier producto suele magnificar sus prestaciones o características pero en principio ha sido nuestra única fuente de orientación. Profundizando en la misma hemos elegido las que nos parecen más afines a nuestro interés por la ciencia de datos.

- Open Refine La citamos en primer lugar por ser de código abierto. Permite identificar, aplicar transformaciones y corregir problemas de calidad. Especialmente interesante porque aplica técnicas de *clustering*[2] para resolver inconsistencias.
- Talend También de código abierto. Combina integración, gobernanza y calidad del dato en una plataforma sin necesidad de generar código que presume de funcionar bajo cualquier arquitectura o cualquier fuente de datos.
- Astera Ofrece prestaciones de validación de datos, que incluyen limpieza de datos, creación de perfiles de errores y reglas de calidad de dato.
- IBM InfoSphere Plataforma de integración de datos con funciones de limpieza, supervisión y transformación de datos con el objetivo de mejorar la calidad de los mismos.
- Data Ladder Enfocado en mejorar la calidad de datos distribuidos en distintas fuentes utilizando algoritmos de coincidencia propios.
- Experian Aperture Herramienta para el perfilado, preparación, goberanza y auditorías sobre regulación de datos.
- attacama ONE Según sus propias fuentes, unifica la gobernanza, la calidad y la gestión de datos en una única estructura impulsada por IA en plataformas en la nube o híbridas.

[1] Nos parece oportuno utilizar este término para tipificarlas.

[2] Término original en inglés de no siempre fácil transcripción a nuestro idioma en el contexto adecuado, pero asimilable aquí a la agrupación de entidades por sus características similares.

- Informatica Este producto conjuga la flexibilidad en cuanto a arquitectura disponible por parte de sus usuarios con la asistencia de la IA para minimizar el uso de código en proceso de su implementación.

Metaherramientas de evaluación

La tipificamos así por ser *per se* un modelo clasificatorio de herramientas disponibles.

- Gartner Magic Quadrant La clasificación proporcionada por el cuadrante mágico de Gartner es una forma original de evaluar características deseables en tecnologías de la información, en nuestro caso la calidad del dato.

Este es un producto propiedad de la empresa de consultoría *Gartner Inc* [28] que proporciona una representación gráfica de la posición relativa de los proveedores de tecnología en un mercado o sector determinado.[3]

Pertenece a una categoría de evaluaciones de tecnología más amplia y que se actualiza con frecuencia.

Nos parece interesante reseñarla por su poder visual y su no vinculación a los actores evaluados, que son clasificados por su visión de la tecnología y su capacidad para ejecutarla.

Estos son divididos en *Líderes*, *Retadores*, *Visionarios*. y *Actores de Nicho*. Los términos nos parecen bien elegidos por su capacidad de autodefinición. El cuadrante viene representado en la figura 3.1 y proporciona al demandante de esas tecnologías una primera orientación que le permita explorar la solución que mejor se adapte a sus necesidades.

Queremos remarcar el hecho de que nuestra selección tampoco ha venido condicionada por su contenido, pero si hemos elegido los tres productos que aparecen clasificados como *Líderes* para incluirlos en nuestra evaluación particular.

Herramientas comerciales para la calidad del dato

Evaluemos más a fondo los elementos del listado anterior por separado: Vamos a intentar incidir en las características diferenciales de cada una de las soluciones exploradas a la vez que recopilar puntos comunes que nos permitan ponderar las distintas prestaciones o servicios ofrecidos por cada una de ellas.

La mayoría proporciona un periodo de prueba o evaluación. En algunos casos vamos a hacer uso de él, –suele requerir un registro previo y tiene prestaciones mucho más limitadas que el producto al que representa– para verificar su funcionalidad, facilidad de uso y prestaciones.

En nuestro caso despiertan especial interés las que ofrezcan técnicas relacionadas con el aprendizaje automático o el entrenamiento de modelos dado que ambos aspectos son una parte fundamental en la ciencia de datos.

Hemos de destacar que el grado de sofisticación, variedad y extensión de las prestaciones ofrecidas por estas herramientas nos ha inducido a la evaluación de algunas de ellas mediante su uso para obtener una experiencia mínimamente certera de sus prestaciones y características en un entorno real. En algunos otros casos más limitados por el fabricante, al menos hemos podido entrever como funciona su interfaz.

[3] La autoría de este gráfico corresponde a la empresa mencionada. Su inclusión no tiene la intención de primar unas herramientas analizadas sobre otras, sino mostrar el modo en que plasma su clasificación.

Figura 3.1: *Gartner Magic Quadrant*para herramientas de la calidad del dato. 2024 ©
Imagen tomada directamente de [27]

Esta mínima experiencia de uso también ha venido condicionada por la facilidad que proporciona cada fabricante a la hora de evaluar su producto.

En apartados anteriores hemos mencionado el típico ciclo de vida de los datos (2.4.3), que abarca desde su creación a su eliminación segura. Las herramientas evaluadas, con sus propias particularidades, aportan soluciones para cada una de las etapas incluidas en el mismo. Al final de cada revisión, señalamos las fases en las que la herramienta evaluada parece hacer mayor incidencia o destacar.

A su vez, en 3.2.1 extraemos nuestra propia lista de dimensiones evaluables, –nótese que en este caso no de los datos, sino de las propias herramientas de evaluación–. Con ellas haremos lo mismo: al final de cada herramienta mencionamos aquellas que parecen mejor satisfechas por la misma.

3.1 Herramientas Comerciales

3.1.1. Open refine

Se instala como servidor web local en indistintas plataformas (*MS Windows*, *Linux*, *Mac*) que cuenten con una distribución de *Java*[4] disponible y es accesible mediante cualquier navegador principal operativo en las anteriores plataformas.

Es un producto cuyo funcionamiento se centra en proyectos que suelen inicializarse importando datos existentes, ya que *Open refine* no permite crear conjuntos de datos arbitrarios.

[4]Nos referimos a una máquina virtual Java.

Bajo este supuesto permite importar datos de muy diversas fuentes (ficheros locales, enlaces *web*, bases de datos mediante consultas *SQL*, hojas de cálculo...) y en múltiples formatos (*XML, JSON, ODS, ficheros de texto, CSVs...*) que también admiten extensiones particulares.

Exploración. Tras la ingesta de datos, se realiza una primera exploración en dos ámbitos: La *asignación a uno de los tipos* atómicos de datos de los cuatro disponibles: *tira, numérico, lógico o de fecha* o a su *etiquetado como nulo o erróneo* dada la posible incertidumbre de su asignación.

Diferencia entre *filas*, entendidas como un conjunto de celdas representando cada una a una columna, y *registros*, definidos en el caso de esta plataforma como un conjunto de filas.

El siguiente ámbito o concepto de interés es el de *facetado*, entendido como la posibilidad de encontrar patrones o tendencias en la varianza de una columna determinada. Este es un término en cierto modo asimilable a la agregación que proporciona el lenguaje SQL.

Transformación. *Open refine* permite permutar el orden de filas o columnas, editar el contenido de celdas de columnas determinadas, intercambiar filas por columnas, dividir o fusionar columnas o añadir nuevas columnas basadas en datos existentes nuevos o generados mediante una reconciliación,[5] o convertir las filas en registros (multifila).

Las fuentes de datos disponibles para comparar y reconciliar pueden provenir de bibliotecas, archivos, museos, organizaciones académicas, instituciones científicas, sin ánimo de lucro o grupos de interés.

Como contrapartida al uso del lenguaje *Python*, *Open refine* está implementado con *Java*, por lo que para nuestros intereses y desde nuestra perspectiva como científicos de datos (mucho más familiarizados con otro tipo de lenguajes como podría ser *Python*), hemos de recurrir a *Jython* [29] como implementación de *Python* en Java.

Tras el procesado, se pueden exportar los nuevos datos generados en multiplicidad de distintos formatos: *tabular, SQL, JSON, ODS, XML, CSV...*

A nuestro juicio, *Open Refine* puede ser una buena herramienta para la adquisición e integración de datos y su post-procesado, que sin ser demasiado ambiciosa en cuestiones semánticas o lógicas, nos permite realizar un buen nivel de transformación de cualquier fuente de datos que originalmente presente un formato estructurado.

Posteriormente contemplaremos –sin ser esta una afirmación especialmente ligada a este producto–, como el grado de estructura de los datos a evaluar condiciona enormemente la capacidad de evaluar su nivel de calidad.

1. **Etapas del ciclo de vida mejor caracterizadas:**
   Creación (obtención de datos), transformación, perfilado.

2. **Características destacadas:**
   Plataforma local, Versión de evaluación[6], Funciones de clasificación.

3.1.2. Talend

Talend es una plataforma de código abierto implementada en *Java* que pretende proporcionar cualquier funcionalidad necesaria para el manejo de datos.

[5] Reconciliación es el proceso de comparar y ajustar nuestro conjunto de datos con una fuente fiable externa.

[6] En este caso la herramienta es también *Open Source* o de código abierto.

Según su propia *web*, «combina integración, calidad y gobernanza de datos en una única única plataforma con bajo nivel de exigencia de codificación que funciona virtualmente con cualquier fuente o arquitectura de datos».

Su arquitectura gira en torno al concepto de *Data Fabric*[7], término que define en esencia la creación de un hilo conductor (de ahí tejido) que conecte y unifique distintas fuentes de datos distribuidos en diferentes ubicaciones.

Este sistema pretende armonizar los conceptos de integración de datos, *APIs*[8] y de aplicaciones con los de integridad y gobernanza de datos, con la intención (según su *web* corporativa), de «potenciar la lealtad y el entusiasmo de sus usuarios[9], mejorar la eficiencia operacional y ahorrar costes, reducir riesgos y asegurar el cumplimiento de normativas aplicables y modernizar infraestructuras *IT*».[10]

Nos producen especial interés sus afirmaciones respecto a sus prestaciones sobre privacidad de datos y cumplimiento de normativas. Para ello, propone:

- «Crear inventarios de datos de un modo automático recolectando datos de cualquier fuente: Sistemas heredados[11], *shadow IT*[12], sistemas de *CRM*[13], sensores de dispositivos, aplicaciones digitales, redes sociales... capturando y mapeando elementos de datos críticos a lo largo de distintos conjuntos de datos, registrando y trazando la procedencia de los datos y su uso final.»
- Diseñar y establecer puntos de control a lo largo de los *data pipelines*[14], anonimizando o pseudo-anonimizando los datos sensibles con enmascaramiento de los datos. Todo esto facilitaría el cumplimiento de por ejemplo el GDPR (2.2.1) con procesos establecidos y controles para cualquier dato de carácter personal.
- Gestión sencilla de consentimientos: A lo largo de cualquier aplicación que recoja datos sensibles[15], proporcionando derecho al olvido, de acceso y de rectificación a todos los posibles sujetos con datos presentes, facilitando portabilidad de datos mediante cumplimiento de la *HIPAA*[16] y haciendo posible el acceso de las personas –que no olvidemos son los últimos propietarios de sus propios datos–, a la información que en su nombre se pueda estar procesando.

Nos parece interesante profundizar en el mencionado concepto de *shadow IT*, considerado como un compromiso entre seguridad y flexibilidad:

Ventajas. Proporciona los beneficios de mejorar la satisfacción y retención el usuario final otorgándole la posibilidad de elegir qué herramientas son más adecuadas para sus

[7] La traducción literal de este término sería la de *Tejido de Datos*.

[8] Un Interfaz de Programación de Aplicaciones, o *API*, es un conjunto de procedimientos que permite la integración de distintos sistemas permitiendo que otra aplicaciones puedan reutilizar sus funcionalidades.

[9] Esta es una afirmación que debería ser mensurable.

[10] Tecnologías de la información.

[11] Traducimos así *Legacy Systems*.

[12] Se introduce este interesante concepto referido a los programas, proyectos o sistemas implementados fuera de los habituales departamentos de IT o de seguridad de la información por los propios usuarios finales.

[13] *Custom Relationship Management*, o sistemas de gestión de relación con los clientes.

[14] Podemos traducir este importante concepto como canal de datos, englobando bajo el mismo cualquier conjunto de procesos automatizados que permitan el flujo de los datos desde su fuente a un destino específico durante su proceso de extracción, transformación o carga.

[15] Entendiendo como tales aquellos sujetos a especial protección.

[16] La *HIPAA, Health Insurance Portability and Accountability Act*, o Ley de transferencia y responsabilidad de seguros médicos en los EE.UU. es una importante ley federal que establece requisitos de privacidad y seguridad de los datos para organizaciones que gestionen información médica sobre individuos. Podría ser equiparada a nuestras leyes de protección de datos personales pero enfocada a datos clínicos.

propósitos, reduce la carga de los departamentos *IT* –permitiéndoles concentrarse en tareas de más enjundia– y ahorra tiempo a ese mismo usuario final al permitirle hacer uso de herramientas de su propia elección en lugar de pedir ningun diseño específico de las mismas.

Desventajas. La compartición de ficheros es una práctica común de la *shadow IT* sujeta a múltiples vulnerabilidades. La integración libre de software de terceros no sujeto a sus debidas actualizaciones podría abrir también una brecha de acceso y ataque a los sistemas de una organización.

Se pueden aplicar distintas estrategias respecto al uso de *shadow ITs*:

- Reforzar las medidas de seguridad en la organización con vistas a limitar el acceso a ciertas aplicaciones de riesgo potencial, (mediante el uso de cortafuegos corporativos o auditorías de uso de aplicaciones),
- Practicar la indulgencia en el uso de aplicaciones opcionales por parte del usuario final, pero reforzando otros aspectos de seguridad como pueden ser un encriptado más robusto de los datos corporativos o limitar el acceso a datos sensibles.
- Una solución de compromiso entre ambos extremos.

1. **Etapas del ciclo de vida mejor caracterizadas:**
   Almacenamiento y organización, funciones de *ELT*.
2. **Características destacadas:**
   Plataforma local/cloud/híbrida, Versión de evaluación, Conexión vía *APIs*.

3.1.3. Astera

Astera tiene por bandera ser un entorno[17] de gestión de datos sin la necesidad de escribir código. Ofrece prestaciones de:

- Preparación de datos
- Almacenaje[18]
- Integración de datos
- Gestión de datos no estructurados
- Diseño y gestión de APIs
- Gobernanza de datos
- Gestión de intercambio electrónico de datos[19]

En referencia al último de los puntos mencionados, el conjunto de normas de documentos EDI [30] es amplio y concerniente al intercambio de documentos comerciales sujeto a estrictas normas de formato, pero merece ser tenido en cuenta o al menos mencionado en nuestro propósito de sistematizar todo lo relacionado con la calidad del dato.

[17] Traducimos *stack* en este caso como entorno.
[18] Referido a la acepción inglesa *Data Warehousing*.
[19] EDI o *Electronic Document Interface*.

Como prestaciones reseñables al margen de no tener que implementar código, el producto ofrece automatización de tareas, múltiples soluciones en una única plataforma, transformaciones de datos prediseñadas, facilidad de uso y (según el fabricante) un soporte técnico confiable.

Mención especial merece el uso de técnicas de IA que esta plataforma utiliza con varios fines:

- Extracción de datos de documentos sin estructurar[20]. Permite la creación de plantillas para la extracción de datos estructurados mediante técnicas de minería de datos[21].
- Auto-mapeo de datos. En este caso entendido como el proceso de hacer coincidir campos de datos de una fuente con los de otra mediante la interpretación de relaciones semánticas.
- Inferencia de relaciones. Para deducir automáticamente relaciones complejas entre entidades.
- Selección de entidades de categorización incierta, para agilizar el proceso de modelado de datos.

1. **Etapas del ciclo de vida mejor caracterizadas:**
   Creación (obtención de datos), Limpieza, Almacenamiento y organización.
2. **Características destacadas:**
   Funciones de *ELT*, Herramientas de AI, Conexión vía *APIs*.

3.1.4. IBM InfoSphere Information Server

De todas las soluciones comerciales exploradas esta el la más opaca en cuando a funcionalidad y prestaciones. En su sitio web corporativo se define como una plataforma de integración de datos para comprender, limpiar, supervisar y transformar los datos.

Cita la posibilidad de ser aplicada a datos en las propias instalaciones del usuario[22], o en un entorno de *cloud*[23] privado, público o híbrido. En cuanto a calidad de datos expone la posibilidad de proporcionar reglas de análisis integradas y una plataforma escalable.

Su oferta final de producto parece vinculada al tamaño de la solución requerida, citando según sus propias fuentes:

- InfoSphere *Information Server (IS) Enterprise Edition*
- InfoSphere *IS for Data Integration*
- InfoSphere *Data Quality*[24]
- InfoSphere *IS on cloud* como modelo de integración y gobernanza de datos completa.

[20] Después aclararemos este concepto.
[21] O *Data Mining*. Podríamos definir este concepto como el proceso que intenta descubrir patrones en grandes volúmenes de conjuntos de datos.
[22] Hemos intentado traducir así el término *on premises*.
[23] Referido a *cloud computing* o computación en la nube.
[24] Para según sus propias fuentes implementar un proceso que convierte los datos en información de confianza mediante un proceso de limpieza y supervisión de calidad continuos.

El punto de nuestro interés es el referente a *Data Quality*, que ofrece:

- **Calidad y gestión de datos**, estableciendo planes de corrección y métricas de evaluación.
- **Estandarización y validación**, con funciones de estandarización, enriquecimiento y limpieza de datos a medida y una configuración flexible de reglas de validación.
- **Funciones de clasificación**, con la interesante característica de ofrecer la posibilidad de señalar dónde se almacena la **información personal identificable** o *IPI*. [25], los datos sensibles y otras clases de datos.

  Permite también identificar el tipo de datos contenidos en una columna mediante el uso de varias docenas de datos predefinidas. Adicionalmente, permite crear tres tipos de clases da datos: Listas de valores válidos, expresiones regulares[26] y clases de *Java*.
- **Certificaciones**, «Admite perfiles de datos, clasificación, investigación, estandarización, coincidencia, supervivencia, verificación de direcciones y supervisión directamente en un clúster de *Apache Hadoop*»[27] [31]

1. **Etapas del ciclo de vida mejor caracterizadas:**
   Creación (obtención de datos), Limpieza, Almacenamiento y organización, Procesamiento y análisis.
2. **Características destacadas:**
   Identificación de datos sensibles, Interoperabilidad y certificación.

3.1.5. Data Ladder

Según su fabricante, *Data Ladder* « ofrece un motor de coincidencia y calidad de datos de extremo a extremo para mejorar la confiabilidad y precisión del ecosistema de datos empresariales sin fricciones»

Hace hincapié en el uso de algoritmos de coincidencia propios que permiten mejorar la calidad de datos distribuidos en fuentes dispersas, por lo que parece una herramienta especialmente indicada para tareas de verificación, conciliación y comprobación de coincidencias.

El mismo nombre de la plataforma, *Data Ladder*[28] sugiere un proceso iterativo o secuencial que ofrece las siguientes prestaciones que constituyen, según la misma fuente, todas las partes integrantes del proceso de Gestión de la Calidad del Dato (*DQM*).

- **Importación de datos**. Conectando fuentes de datos dispares para limpiar, comparar, deduplicar y fusionar o purgar datos, con los propósitos finales de estandarizar datos resolviendo sus diferencias sintácticas y semánticas y a nuestro entender lo que es más importante: construir una única fuente de verdad[29].

[25] La IPI es el conjunto de datos que puede utilizarse para identificar unívocamente a una persona.
[26] *Regex*.
[27] *Apache Hadoop* es un entorno de trabajo de código abierto que permite el almacenamiento y procesamiento distribuido de grandes volúmenes de datos en *clusters* de ordenadores.
[28] Escalera de datos.
[29] Esto viene entroncado con el concepto de *Golden record* o registro maestro como un ideal de información de absoluta completitud y exactitud.

- **Perfilado**. Entendido como el proceso de descubrir detalles ocultos sobre la estructura y el contenido de los conjuntos de datos tratados. «Por ejemplo, si se quiere mejorar la calidad de los datos, un perfil de datos ayuda a identificar posibles oportunidades de limpieza de datos y a evaluar el grado de mantenimiento de sus datos en relación con las dimensiones de calidad de los mismos».

- **Limpieza**. Definido com el proceso de corregir la información incorrecta o inválida para conseguir una visión coherente obtenida de fuentes dispares, eliminando valores incorrectos y validando el formato y el patrón de los valores de datos utilizando sus tipos adecuados.

  Este proceso contribuiría también a alcanzar estándares de cumplimiento de datos tales como como GDPR, HIPAA o CCPA[30]

- **Concordancia**.[31] Nos parece una prestación especialmente útil y citamos textualmente su definición por parte de la plataforma por su concisión y claridad: «El cotejo de datos es el proceso de comparar valores de datos y calcular el grado de similitud entre ellos. Este proceso es útil para eliminar los duplicados de registros que suelen formarse con el tiempo, especialmente en las bases de datos que no contienen identificadores únicos o claves primarias y externas adecuadas».

  «En estos casos, se utiliza una combinación de atributos no únicos (como el apellido, el nombre de la empresa o la dirección) para cotejar los datos y encontrar la probabilidad de que dos registros sean similares.» [32]

- **Deduplicación**[32]. Esta deseable eliminación de duplicados de cualquier base o fuente de datos puede alcanzarse aplicando algoritmos de cotejo como los mencionados en el punto anterior.

- **Fusión y purga**. Volvemos a citar textualmente al fabricante: «Un software de purga de fusiones examina todos los registros de datos que residen en múltiples fuentes de datos, reconoce los registros duplicados y le permite crear reglas de supervivencia que fusionan o eliminan automáticamente los duplicados».

  «Dado que la empresa media utiliza más de 65 fuentes de datos diferentes para almacenar registros de entidades relacionados con clientes, proveedores o productos, un software de depuración de fusiones puede añadir un enorme valor al proceso de creación de una única fuente de verdad para su organización.» [33]

Para finalizar en lo referente a esta plataforma, ofrece también una interesante categorización de soluciones por característica, casos de uso o industria aplicable.

1. **Etapas del ciclo de vida mejor caracterizadas:**
   Creación (obtención de datos), limpieza, Deduplicación, Almacenamiento y organización, Eliminación segura.

2. **Características destacadas:**
   Versión de evaluación, Identificación del tipo de datos, Conexión via *APIs*.

[30] La CCPA es la legislación más importante de los EE.UU. en lo referente a la privacidad de los datos y ha entrado en vigor después del Reglamento general de la protección de datos (RGPD) europeo, que tuvo lugar en mayo en 2018.
[31] de *Matching*, en inglés.
[32] En este caso hace referencia a la localización de uno o varios registros duplicados.

3.1.6. Experian Aperture

Experian Aperture Data Studio se auto define como «una plataforma inteligente de autoservicio para la calidad y enriquecimiento del dato» [34]

Como característica reseñable, permite construir flujos de trabajo que incorporan algoritmos de aprendizaje automático con el fin de proceder al etiquetado de datos, a la vez que utiliza conjuntos de datos propios seleccionados globalmente para contribuir al enriquecimiento mencionado.

Como las expectativas parecen interesantes, decidimos subscribir una versión de evaluación (solo válida durante 2 semanas) que consta de las siguientes prestaciones básicas:

- un único usuario
- 3 fuentes de datos disponibles
- un límite de 35.000 filas[33]

Este paquete de *software* permite, entre otras funcionalidades:

- La creación de espacios de trabajo, que son áreas lógicas estancas, en el sentido en que permiten implementar controles y gestión de acceso.
- La creación de paneles[34] como contenedores de *widgets*[35] en forma de gráficos, vistas o conjuntos de datos.
- El uso de conjuntos de datos[36] para almacenar el esquema (o diseño de columnas) de los datos a cargar incluídas sus opciones de análisis, configuración, etiquetas y detalle del lote de datos a cargar.

  Este enfoque nos permite definir conjuntos de datos vacíos y proceder a su carga de un modo parcial o por lotes, facilitando la gestión y la prueba de conjuntos de datos de gran tamaño.
- La creación de flujos de trabajo como secuencia de pasos conectados que defina un proceso de transformación o gestión de datos.
- La creación de funciones específicas para la transformación, filtrado o validación de datos.
- Gestión de versiones. Importante prestación aplicable a los flujos de trabajo, vistas de datos o funciones permitiendo tener un control sobre los cambios históricos de las mismas, revertir a versiones previas o asegurarse de que los posibles usuarios comparten el uso de la versión correcta en cada caso.

1. **Etapas del ciclo de vida mejor caracterizadas:**
   Creación (obtención de datos), Preparación, Perfilado, Interoperabilidad y certificación.
2. **Características destacadas:**
   Versión de evaluación, Permite generar certificaciones, Funciones de clasificación.

[33] *grid row limit*.
[34] *Dashboards*.
[35] Término de difícil traducción asimilable al de artilugio o pequeñas aplicaciones o programas que permiten el acceso a las funciones más usadas por una aplicación.
[36] O *Datasets*.

3.1.7. Attacama ONE

Según el sitio *web* del fabricante, *Attacama ONE* es una «única plataforma modular que incorpora utilidades de gobernanza, calidad y gestión de datos maestros en un único producto».

Presenta una estructura modular diferenciada en tres bloques:

- La adquisición de datos, tanto de sistemas *SaaS* o *PaaS*[37], desde cualquier sistema alojado en la nube o desde las propia infraestructura local del usuario.
- Implementación de calidad, gobernanza y gestión de datos maestros mediante el uso de herramientas y técnicas propias de la *AI*.
- La satisfacción de distintos requerimientos, como pueden ser el de aumento de beneficios, minimización de riesgos, incremento de la efectividad operativa o de mejoras de las expectativas de los usuarios finales.

Como características reseñables, ofrece:

- Un ajuste específico para las necesidades de cada perfil de usuario: científicos de datos, responsables de gobernanza, analistas o ingenieros de datos...
- Funciones da catalogación, calidad, integración de datos, *data stories*[38]
- Integración de un modo nativo con las mayores plataformas de *big data*: AWS, Google y Azure[39], *Spark*[40], *Hadoop*, y *Databricks*[41] entre otras.
- Ser una plataforma de grado empresarial al proporcionar seguridad basada en roles de usuarios, auditabilidad y alta disponibilidad (siempre según sus propias fuentes).
- Gestión automática de la calidad del dato, con asignación automática de *business rules*[42] para todos los datos gestionados, automatización de detección en linea de anomalías en los datos.
- Gestión de datos maestros y catálogo de datos automatizado.
- Observabilidad de los datos, entendida como un modo de notificar sobre incidencias o problemas con los datos antes de que produzcan efectos adversos.

En este caso y dadas las prestaciones descritas, también se ha decidido contactar con el fabricante del producto para solicitar una licencia de evaluación.

[37] *SaaS* es un modelo de software basado en la nube accesible a sus usuaros finales mediante un navegador mientras que *PaaS* es un modelo más amplio ofreciendo una plataforma que permite a los desarrolladores crear y gestionar aplicaciones.

[38] o *Data Storytelling*. Concepto muy ligado con la ciencia de datos: Metodología para comunicar información de un modo comprensible y mediante una narrativa adecuada a cualquier audiencia específica. Suele ser la conclusión de todo análisis de datos.

[39] *Amazon Web Services, Google Could* y *MS Azure* están consideradas como las tres *majors* a nivel mundial en lo concerniente al tratamiento de datos en la nube.

[40] Esta herramienta es descrita someramente más adelante.

[41] Herramienta creada por los mismos autores que *Spark* y que propone aplicar técnicas de inteligencia artificial al uso y la gestión de análisis de datos.

[42] Reglas de negocio. Hemos incluido el término original en inglés por su uso extendido en este contexto.

1. **Etapas del ciclo de vida mejor caracterizadas:**
   Creación (obtención de datos), Preparación, Perfilado, Clasificación y agrupamiento.

2. **Características destacadas:**
   Distintos perfiles de usuario, Plataforma *multi-cloud*, Conexión via *APIs*, Identificación del tipo de datos.

3.1.8. Informatica

Informatica pretende potenciar la productividad de cualquier negocio adecuando sus datos gestionados mediante técnicas de IA que según sus promotores «se emplean en todo momento y de extremo a extremo de la plataforma, para conectar, gestionar y unificar sus datos en virtualmente cualquier entorno *multi-cloud* o híbrido».

En esta linea, ofrece:

- Flexibilidad, a la hora de adaptarse a cualquier arquitectura de referencia.[43]
- un sistema *multi-cloud* o híbrido, que entendemos induce la migración desde datos *on-premises* a cualquiera de estos entornos elegible por el cliente.
- Ausencia o baja necesidad de generar codificación particular.
- un motor de IA propio[44] que según su documentación, aúna el poder de la IA con el uso de *machine learning*. Como característica, conlleva la modalidad de pago por uso.

  Dicho motor se interpone entre los productores y consumidores de datos proporcionando utilidades de catalogación e integración de datos, integración de *APIs* y aplicaciones, supervisión de la calidad y otras funciones de gobernanza privacidad y compartición de datos.
- Otras funciones de perfilado y transformación, estas últimas dirigidas a estandarizar, enriquecer, deduplicar o validar los datos mediante reglas de negocio predefinidas.
- Experiencia de uso orientada al individuo y enfocado en sus habilidades e intereses.
- Escalado en función de la carga de trabajo para adaptarse a procesos de datos en tiempo real, de servicios *web*, por lotes[45] o el el ámbito del *Big Data*.

Mención especial merece su habilidad para desplegar calidad de datos en distintos ámbitos de uso –siempre según su propia documentación–, entre los que se encuentran la gobernanza, la gestión de datos maestros o la gestión de riesgos.

Otra prestación ofrecida es la de la automatización de tareas críticas –como ejemplo cita el descubrimiento de datos[46]–.

A nuestro juicio parece uno de los productos más ambiciosos y ofrece una experiencia de uso previa a su adquisición con un particular nivel de escalado:

[43] Entendemos como tales las de mayor uso.
[44] Denominado CLAIRE ©.
[45] Batch processing.
[46] Término relacionado con la tecnología de inteligencia empresarial y orientado al usuario que conlleva extraer y evaluar datos de distintas fuentes.

Hay dos niveles gratuitos que permiten la carga e integración de datos (en el segundo caso con un límite de capacidad de almacenamiento y de tiempo de proceso). Por encima de ellos hay otros dos niveles de prestaciones incrementadas que ya requieren de suscripción y pago.

Reiteramos que dada la complejidad de todos estos sistemas, cualquier facilidad que permita evaluar su uso previo ha de ser un factor a tener muy en cuenta antes de su adopción.

1. **Etapas del ciclo de vida mejor caracterizadas:**
   Creación (obtención de datos), Preparación, Perfilado, Clasificación y agrupamiento.

2. **Características destacadas:**
   Distintos perfiles de usuario, Plataforma *multicloud*, Conexión via *APIs*, Identificación del tipo de datos.

3.2 Características y su evaluación

Proponemos un método que involucra unos sencillos pasos y que creemos útil y aplicable en la toma de decisiones a la hora de seleccionar la herramienta de calidad que mejor se ajuste a nuestras necesidades particulares.

1. **Definición de características adecuadas a nuestras propias necesidades.**

2. **Ponderación de las características en función de su importancia.**

3. **Extracción y valoración de características de las herramientas consideradas.**

4. **Cómputo de valoraciones y selección del valor óptimo ponderado.**

3.2.1. Definición de características adecuadas

De las fuentes comerciales anteriores hemos ido extrayendo las siguientes características, que podemos denominar *características evaluables* y que según nuestro criterio particular, nos podrían permitir implementar algún mecanismo de decisión sobre cual sería la mejor solución u herramienta a adoptar en función de las necesidades de un ámbito –lo más genérico posible–, a abordar.

Este primer paso es muy importante y debería personalizarse en función de las necesidades del agente evaluador que han de ser previamente definidas y priorizadas. Esta definición de características ha de contar con la suficiente granularidad –o nivel de detalle–, para satisfacer nuestros requerimientos de un modo lo más preciso posible.

Así pues, como paso previo a la hora de elegir cualquiera de las herramientas revisadas, se debería hacer una lista de las características de necesaria implementación (en función de la índole de nuestros datos, sus requerimientos de seguridad, equipo de proceso disponible, o cualquier otro que se ajustase a nuestras necesidades) que sirviese como guía para decidir cuales de ellas cumplen con los requisitos deseados.

En nuestro (genérico) caso:

Características evaluables sobre el concepto de calidad del dato en plataformas comerciales:

- Tipo (respecto a ubicación) de *plataforma*.[47]
- Posibilidad de implementar distintos *perfiles* en función del papel de cada *usuario*.
- Necesidad de *codificación*. Tipo de lenguaje(s) principal(es).
- Disponible Versión de evaluación[48]
- Identificación de *datos sensibles*.
- Implementación de *identificación del tipo de datos*.
- Permite generar *certificaciones*.
- Herramientas de *ELT/ETL*[49].
- Herramientas de *AI*.
- Funciones de *estandarización y validación*.
- Funciones de *clasificación*.
- *Interoperabilidad* y *certificación*.
- Conexión vía *APIs*.

En la tabla de la figura 3.2 mostramos una matriz cuyas columnas asociamos a las características que en nuestro caso hemos seleccionado. En sus filas, disponemos cada una de las herramientas evaluadas y en las casillas de intersección vamos a incluir la puntuación (o ausencia/presencia) de dichas características para cada herramienta evaluada.

Es necesario incidir en que **esta extracción de características es subjetiva y puede modificarse para ser más acorde a las necesidades del demandante** de datos de calidad.

3.2.2. Ponderación de las características

Las características (que también podemos denominar *dimensiones*) deseables **se deberían ponderar** en función de nuestros requerimientos particulares otorgando mayor valor a aquellas más prioritarias.

Dichas necesidades se pueden puntuar y tipificar como binarias (1 ó 0 para SI o NO) o graduadas en una escala entera, en la que por convenio mayor puntuación implicaría mayor adecuación.

El modelo propuesto permitiría sofisticaciones adicionales, tales como la implementación de una función de optimización que contemplase tanto variables binarias –para reflejar los requerimientos como ineludibles o prescindibles– así como numéricas, (para reflejar un grado de adecuación): El producto evaluado que presentase un valor óptimo –mayor puntuación en nuestro caso–, debería ser el seleccionado.

Dicha función podría contemplar como restricciones las presencia de características ineludibles o un valor mínimo a satisfacer por ciertas características.

Volvemos a recordar la necesidad de **optar por las dimensiones que más se ajusten a nuestros requerimientos para cada uno de nuestros análisis** como paso previo a su correcta ponderación.

[47] En las instalaciones del propio usuario, en la nube...

[48] Esta, más que una característica inherente a la herramienta evaluada, sería un factor que nos facilitaría considerar su posible uso en función de que el resto de las características deseables estuviesen representadas por la opción.

[49] Incidiremos en estos términos más adelante.

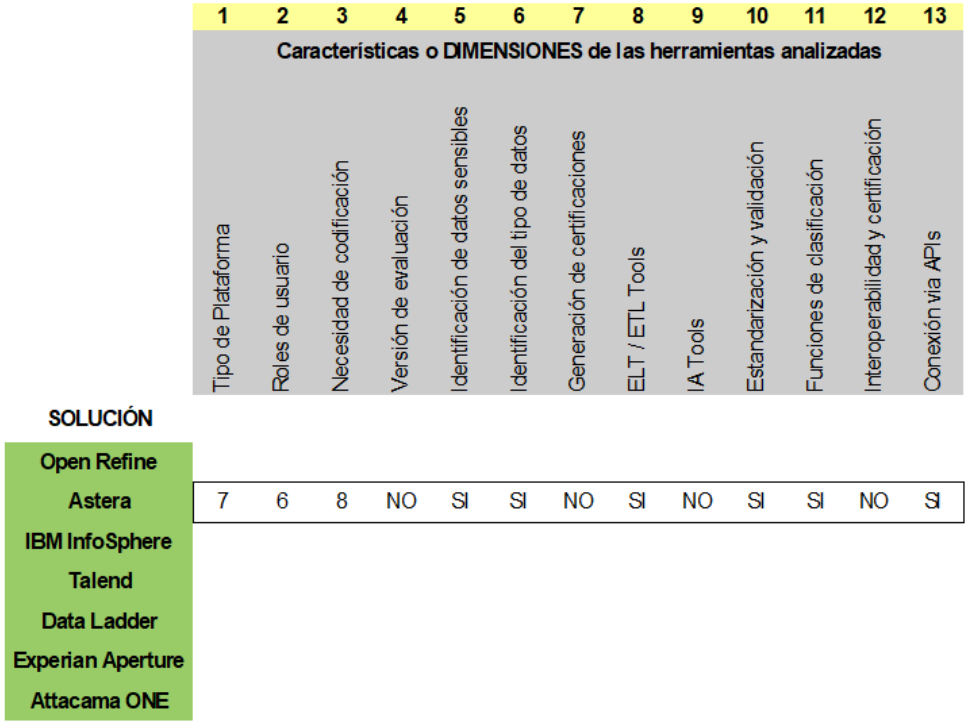

Figura 3.2: Definición y extracción de características
Diagrama de elaboración propia.

3.2.3. Extracción y valoración de características

Tras haber definido y ponderado nuestras necesidades es el momento de extraer el grado de adecuación a las mismas de las herramientas a evaluar.

Esta es con creces la tarea más compleja y costosa en tiempo requerido y debería fundamentarse en dos acciones:

1. La **revisión de la información** que el proveedor de la herramienta pone a disposición general del público en su web promocional (mejor en su apartado técnico) para verificar si la herramienta ofrece algo relacionado con la prestación requerida.

   Esto constituiría una primera criba.

   Como nota adicional deseamos señalar que sería interesante explorar hasta que punto las herramientas disponibles de *open AI*[50] podrían asistirnos –creando *prompts* específicos para ello–, a discriminar de un modo automático las herramientas que proporcionen los requisitos requeridos. Por razones de brevedad no hemos incurrido en este desarrollo.

2. La **experimentación y uso** de la prestación directamente en el prototipo ofrecido por el fabricante o su versión de evaluación.

   Este proceso se puede refinar recurriendo al servicio de soporte de la herramienta evaluada para recabar información adicional.

El objetivo final de estas dos acciones es disponer de información suficiente para puntual cada herramienta: A más adecuación al requerimiento evaluado deberíamos asignar una puntuación mayor.

[50] Tipo *ChatGPT* u otras.

Dimensiones	1	2	3	4	5	6	7	8	9	10	11	12	13	Grado de satisfacción a nuestras necesidades
Solución														
Open Refine														
Astera	1,4	0,3	0,8	0	0,15	0,025	0	0,05	0	0,05	0,025	0	0,15	2,95
IBM InfoSphere														
Talend														
Data Ladder														
Experian Aperture														
Attacama ONE														

		Para cada producto...	Puntuación (1)	Peso (2)	Elegimos el de máxima puntuación ponderada
Dimensiones	1	Tipo de plataforma	7	0,2	1,4
	2	Roles de usuario	6	0,05	0,3
	3	Necesidad de codificación. Lenguaje(s).	8	0,1	0,8
	4	Versión de evaluación	0	0,05	0
	5	Identificación de datos sensibles	1	0,15	0,15
	6	Identificación del tipo de datos	1	0,025	0,025
	7	Generación de certificaciones	0	0	0
	8	ELT / ETL Tools	1	0,05	0,05
	9	AI Tools	0	0,1	0
	10	Funciones de estandarización y validación	1	0,05	0,05
	11	Funciones de clasificación	1	0,025	0,025
	12	Interoperabilidad y certificación	0	0,05	0
	13	Conexión vía APIs	1	0,15	0,15
				1	2,95

(1) **De 0 a 10, por ejemplo, en función de la satisfacción de las necesidades del evaluador.**
Posibilidad de utilizar variables binarias o enteras dentro de una escala.

(2) **En % o proporción de satisfacción de los requerimientos necesarios. También definido por el evaluador.**

Figura 3.3: Modelo de evaluación de herramientas para la calidad del dato
Diagrama de elaboración propia.

3.2.4. Cómputo de valoraciones y selección del valor óptimo ponderado

Habiendo obtenido el grado de ajuste de cada característica –por su valor en la escala definida o su valor binario–, en la figura 3.3 columna **Puntuación (1)**, y multiplicándola por el valor que damos a esa característica como proporción de su importancia en nuestros requerimientos, en la misma figura columna **Peso (2)**, obtenemos un valor ponderado para cada característica. Sumando el de todas ellas obtenemos un valor final que puede ser utilizado para clasificar por orden las herramientas evaluadas.

En el ejemplo de la figura 3.3, el valor total ponderado que obtendríamos sería el de 2,95. Dicho valor no tiene dimensión ni unidades de medida pero nos sirve para nuestro propósito: Dotar de un orden de adecuación de los productos evaluados a nuestra necesidad.

Adicionalmente, y una vez generada esa lista ordenando las herramientas por su grado de adecuación, se podrían considerar otros condicionantes importantes (como los económicos) que nos permitiesen tomar una decisión final.

Explícitamente hemos decidido no incluir el coste de implementación como una dimensión más, sino dejarlo como una consideración final al margen. El motivo de este proceder es comprobar si las soluciones técnica más adecuadas a nuestras necesidades están al alcance de nuestros recursos destinados para su implementación. En caso contrario, podemos considerar incrementarlos.

El las figuras 3.4 y 3.5 mostramos, respectivamente, un gráfico de tipo radar que podemos utilizar para visualizar de un modo gráfico el nivel de adecuación de cada dimensión (con valor entero) a nuestros requisitos y en el otro el flujo de acciones[51] del proceso propuesto.

[51] O *work-flow*.

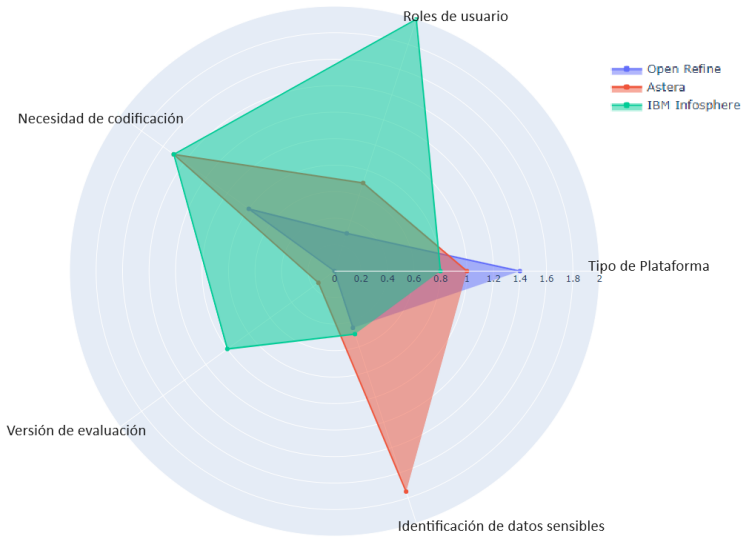

Figura 3.4: Gráfico comparativo de la puntuación asignada a cada solución evaluada como adecuación a nuestros requisitos definidos
Figura de elaboración propia.

De un modo objetivo, el polígono de 3.4 asociado a cada solución que tenga la mayor superficie es el que más se adecuaría a nuestros intereses particulares.

En este caso mostramos el soporte de seis dimensiones concretas para tres herramientas evaluadas. El número de dimensiones y herramientas puede alterarse para hacerlo extensible a la cantidad adecuada de las mismas.

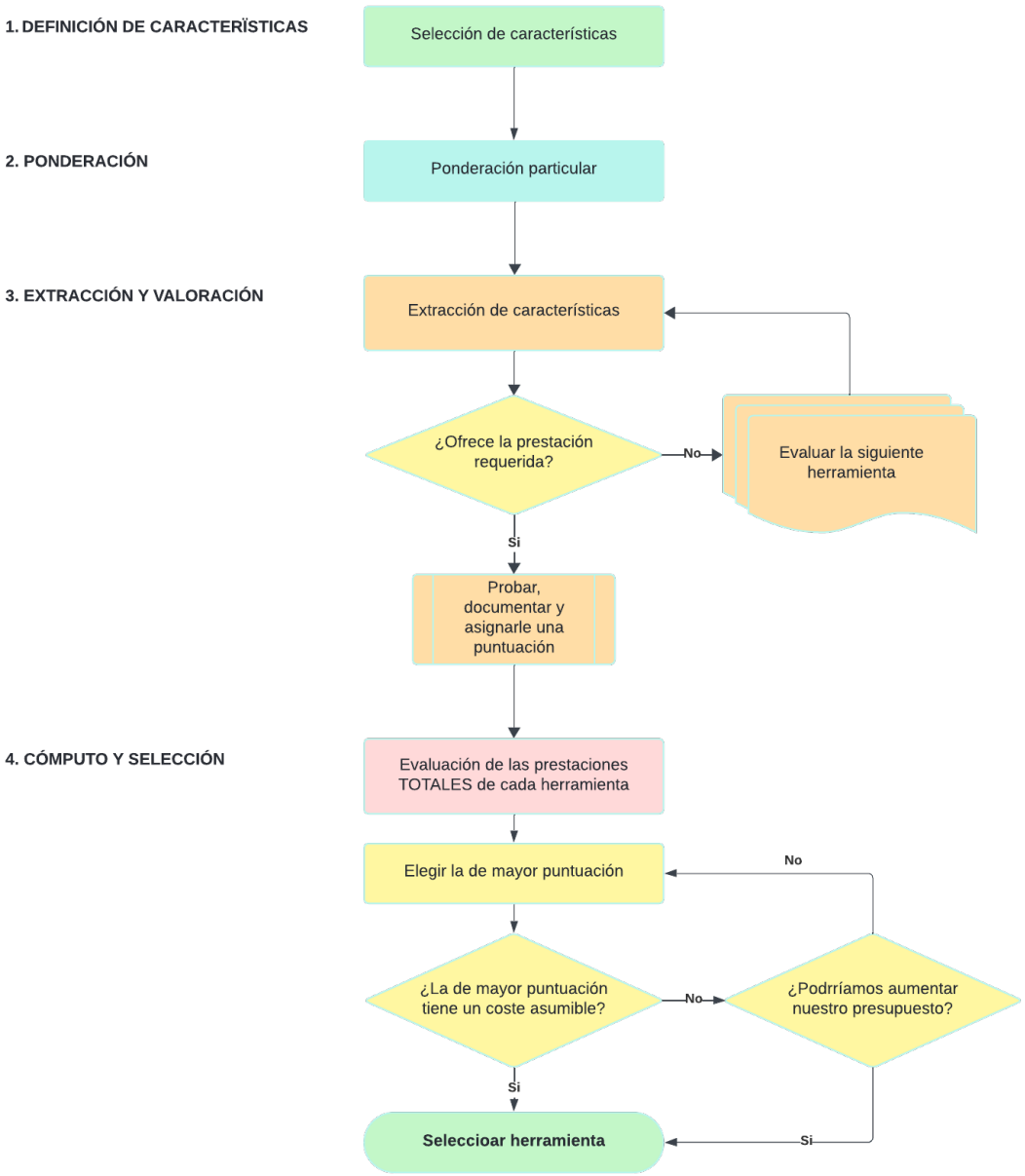

Figura 3.5: Diagrama de flujo del proceso de selección de una herramienta de calidad
Figura de elaboración propia.

CAPÍTULO 4
Propuestas no comerciales

Se han revisado algunas propuestas que pretenden implementar metodologías propias desde el punto de vista académico o dentro de un marco originalmente no comercial. Hacemos una breve reseña de algunas de ellas:

4.1 Atlan

Atlan [35] ha sido originalmente el mayor *data lake* gubernamental existente hoy en día y está considerada como la plataforma nacional de la India[1].

Se desarrolló a gran escala para conectar 42 bases de datos pertenecientes a 52 ministerios diferentes en un tiempo récord de 12 meses, constituyendo el mayor despliegue global de este tipo.

Nos parece digna de mención y propone una metodología de once pasos para articular la construcción de un marco completo de implementación de la calidad del dato.

1. Definir una métricas aplicables a la calidad del datos.
2. Identificar fuentes de datos y *APIs*.[2]
3. Estandarizar formatos de datos y protocolos.
4. Gestión de errores y mensajes (notificaciones).
5. Detección de limitaciones y cuellos de botella.[3]
6. Autenticación y autorización.[4]
7. Observación y registro en tiempo real.[5]
8. Versionado y gestión de cambios.
9. Colaboración entre partícipes.[6]
10. Mejora continua.

[1] Ahora también denominada *Bharat* en fuentes más modernas.
[2] Volvemos a recurrir al concepto de *Application Program Interfaces* que posibiliten la interacción con el programador y los desarrolladores de código.
[3] Traducido libremente de *Rate limitting & throttling*.
[4] Es conveniente comprender y diferenciar estos conceptos.
[5] Traducido de *Real-time monitoring and logging*. No consideramos el término "monitoreo" como adecuado.
[6] Traducción que hacemos de *Stakeholders*.

11. Entrenamiento y formación.

Nos parece interesante precisar la diferencia entre los términos autenticación y autorización: El primero se refiere a verificar la verdadera identidad de un usuario, mientras que el segundo determina a qué tiene acceso cada usuario garantizando que obtenga los permisos adecuados para ello.

4.2 Framework para la gestión de la calidad de datos

El trabajo citado [36] en esta sección pretende realizar una síntesis de las muy distintas técnicas y metodologías perfiladas en este estudio y «en base a estos se propone un marco general que incluya lo más relevante de los *frameworks* existentes».

Su implementación empieza con la definición de los componentes del *framework* [7] con el propósito de garantizar su independencia respecto al dominio de aplicación:

1. Equipo responsable de la implementación.
2. Modelo de madurez de la calidad de datos.
3. Proceso de mejora de la calidad de datos.

El equipo responsable está compuesto por unos roles bien definidos:

Responsable de calidad, a cargo de documentar las actividades realizadas y asignar tareas. **Experto de negocio**, Es quien conoce los aspectos fundamentales de la actividad relacionada con los datos a tratar y es capaz de documentar los requisitos de dicha actividad. **Analista de calidad de datos**, responsable de los aspectos de calidad y la identificación de problemas relacionados con la misma y la definición de métricas relevantes. **Técnico de calidad de datos**, responsable de la implementación de los recursos técnicos necesarios y los métodos asociados a las métricas requeridas. Un **Experto técnico**, que lo sea en los datos generados por los sistemas de información adoptados y con conocimientos en la representación de los datos gestionados.

Y finalmente un **Usuario de procesos** que es cualquier persona que participe en la creación o tratamiento de los datos. Este rol es importante en la medida en que se verá afectado por cualquier cambio propiciado por los agentes anteriores.

El **modelo de madurez** es una «herramienta para que las organizaciones tengan la posibilidad de conocer el estado actual de su gestión de calidad de datos, para luego, generar planes de mejora».

El modelo adoptado, y propuesto por [37] consta de seis niveles «utilizados para caracterizar a la organización en relación a un conjunto de atributos de distintas áreas de proceso».

Dichos atributos de caracterización considerados por los autores citados son:

1. Conocimiento y comunicación (AC, *Awareness and communication*).
2. Políticas, planes y procedimientos (PSP, *Policies, plans and procedures*).
3. Herramientas y automatización (TA, *Tools and automation*).
4. Habilidades y experiencia (SE, *Skills and expertise*).

[7] Decidimos utilizar el término original *framework* en vez de marco de trabajo.

5. Responsabilidad y rendición de cuentas (RA, *Responsibility and accountability*).

6. Definición y medición de objetivos (GSM, *Goal and measurement*).

Nos parece oportuno enumerarlos en detalle porque son un pilar del método propuesto: cada atributo es evaluado en cada nivel del modelo de madurez, indicando el cambio a implementar para alcanzar un nivel de madurez superior.

Los **niveles** o estadios **de madurez** se califican de menor a mayor nivel de optimización:

1. Inexistente. No se conoce ningún tipo de proceso relativo a la gestión de datos

2. Inicial. Se asumen problemas relativos a la gestión de datos

3. Repetitivo pero intuitivo. Diferentes agentes realizan la misma tarea mediante procedimientos similares.

4. Definido. La organización ha implementado políticas de estandarización y documentación pero que pueden estar sujetos a desviaciones no detectadas.

5. Gestionado y medible. A este nivel se realiza un seguimiento de los niveles de adecuación de los procedimientos a las políticas definidas y se toman medidas correctoras implementando una mejora continua. Consta cierno nivel de automatización y uso de herramientas específicas.

6. Optimizado. Los procesos alcanzan un nivel de refinado en el que se propician las buenas prácticas basadas en la mejora continua y el uso de herramientas adecuadas.

Como ya hemos mencionado, se trata de realizar un proceso iterativo de análisis y aplicación de medidas que conduzca a la implementación de estrategias de mejora partiendo de la identificación del nivel de calidad existente, el análisis y la aplicación de distintas estrategias.

Incluimos un gráfico proporcionado por los autores del trabajo que clarifica este concepto. Ver figura 4.1.

La primera etapa incluye la definición de roles mediante el uso de una matriz R.A.C.I.[8] (Figura 4.2) en la que se cruza cada rol con cada etapa del ciclo indicando la responsabilidad de cada rol en cada etapa.

Traducidos a nuestro idioma, estos roles (o papeles) tienen las siguientes funciones:

- *Responsible*: En este caso encargado. La persona encargada de realizar o llevar a cabo el proceso o la tarea asignada.

- *Accountable*: En este caso responsable. La persona finalmente responsable de que la tarea o proceso se realice adecuadamente de un modo apropiado.

- *Consulted*: La persona o persona que no están directamente involucrada en realizar una tarea pero son consultadas para ello.

- *Informed*: Aquellos que reciben información sobre el proceso realizado cuando hay necesidad de ello.

[8] *Reponsible - Accountable - Consulted - Informed.*

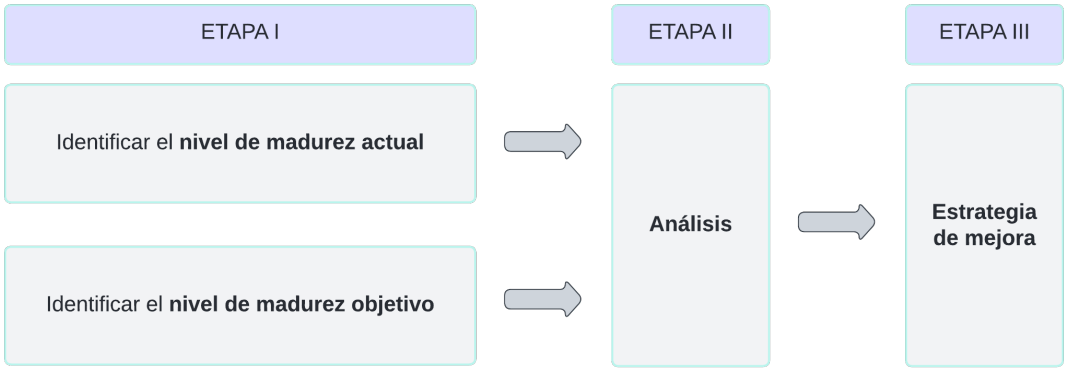

Figura 4.1: Etapas de implementación del modelo de madurez
Diagrama adaptado del trabajo relacionado.

Fases / Roles	Rol 1	Rol 2	Rol 3	Rol 4
Creación	R	A	A	I
Procesamiento	R	C	C	I
Almacenamiento	A	C		I
Distribución				A

R Responsible
A Accountable
C Consulted
I Informed

Figura 4.2: Matriz RACI.
Diagrama adaptado del trabajo relacionado.

Los dos primeros roles se prestan a confusión pero son significativamente diferentes. El primero ejecuta la tarea mientras que el segundo es el responsable de su ejecución.

Tras esta definición de roles se debe identificar el nivel de madurez actual, el nivel (o los niveles) objetivo y las áreas de proceso incluidas.

Establecidos los roles y los niveles, en una segunda etapa se evalúan los niveles actuales de madurez y el nivel de adecuación de los mismos a los objetivos deseados. Los autores proponen una útil herramienta visual, –un diagrama de tipo radar–, para visualizar las distancias entre los niveles actuales de calidad y los objetivos a alcanzar.

Esta idea nos ha inspirado para nuestros propios fines, (ver figura anterior 3.4): hemos hecho uso de la misma para nuestro propio modelo de evaluación anterior descrito más arriba y sintetizado en la figura 3.5.

En lugar de implementar este tipo de gráfico para lo que hacen los autores de este trabajo –solapar niveles actuales con objetivos de calidad–, nosotros lo utilizamos para mostrar la adecuación de distintas soluciones comerciales analizadas a nuestros requerimientos propios.

Por último, junto con el equipo responsable y el modelo de calidad, el tercer componente principal propuesto es el **proceso de mejora de calidad**.

Dicho proceso está compuesto por dos etapas diferenciadas, la Evaluación[9] y la Mejora[10]. Durante la evaluación se propone:

[9] *Assesment*.
[10] *Improvement*.

- **Definición del dominio**. «Incluye identificar reglas y procesos del negocio, fuentes de datos, políticas de la organización, leyes y tipos de usuarios de los sistemas involucrados, así como también su vinculación con los procesos de negocio».

- **Análisis**. Que sería una primera aproximación a los problemas de calidad de los datos de la organización que permitan perfilar un plan de mejora. Esta etapa ya debería proporcionar un informe de *Data Profiling*, un documento de problemas de calidad detectados y un documento de requisitos de calidad a alcanzar.

- **Definición del modelo de calidad de datos**. Basado en los resultados proporcionados por los pasos anteriores, se desarrolla el *software* (o los procedimientos) que permitan implementar las métricas definidas en el modelo de calidad.

  Se propone un modelo muy atómico en el que a cada variable del modelo de datos se le asignan dimensiones (como su unicidad o completitud, por ejemplo) factores, métricas, –número de valores nulos, por ejemplo–, y procedimientos de medición, basados en sencillas consultas SQL.

- **Mejora**. Como última etapa, se enfoca en la resolución de los problemas de calidad detectados para adaptarse a los requisitos plateados previamente mediante un plan de acción.

  En este plan se asigna un identificador a cada uno de los problemas detectados, se documenta el problema y se propone una estrategia de resolución.

Tras la descripción y justificación del todo el marco teórico anterior, los autores proponen una implementación práctica basada en el paquete de software libre *Eclipse Process Framework Composer*[11], ya bastante obsoleto y cuya última actualización data del año 2018.

Dicho *framework* consta de un gestor de meta-modelos, que permite estructurar métodos y procesos mediante diagramas y archivos *XML* asociados, y un marco adicional de herramientas de procesos extensibles que porporcionan *APIs* de funcionalidad básica para crear métodos, procesos, gestión de bibliotecas y publicación de resultados.

Sin entrar en mayor detalle, recalcar que este marco sistematiza la asignación de roles y genera listas de comprobación para cada uno de los niveles de calidad a evaluar y mejoras a implementar.

4.3 Guía para evaluar la calidad de datos basada en ISO/IEC 25012

En este caso [38] se propone una guía de medición basada en la norma ISO/IEC 25012 tomando como caso de estudio una sencilla base de datos de acceso libre.

Ya hemos revisado con anterioridad algunas de las normas de calidad concernientes a la calidad del dato, en particular algunas de las normas de la familia 25000: la 25012 que describe un modelo general de calidad de datos para aquellos que cuenten con un formato estructurado y permite detallar requisitos, métricas y acciones de evaluación y la 25024, más enfocada en el detalle de la definición de esas métricas que requieren cierta capacidad de análisis basada en el contexto de los datos.

Lo que los autores plantean aquí es también una sistematización en la aplicación de dichas normas pero apoyándose en una adicional, la 25040, que proporciona, citando

[11] Eclipse foundation, https://projects.eclipse.org/projects/technology.epf.

literalmente «requisitos, recomendaciones y guías para llevar a cabo el proceso de evaluación de cualquier producto de software».

Como es la primera vez que mencionamos esta última nos parece oportuno desglosar el nivel de actividades propuesto en la misma:

1. Establecer los requisitos de evaluación, Consistente en establecer el propósito de la evaluación, identificar las partes interesadas, los posibles riesgos y el modelo de calidad a utilizar.
2. Especificar la evaluación. Definir métricas, herramientas y técnicas y criterios de decisión.
3. Diseñar la evaluación. Consistente en el establecimiento de un plan de tareas.
4. Ejecutar la evaluación. Obteniendo métricas aplicando criterios de evaluación.
5. Conclusión. Presentando un informe de resultados y conclusiones en función de los valores obtenidos.

Para cada una de las características definidas por la norma 25012 (anteriormente mostradas en la figura 2.2) el método propone la compilación de los siguientes datos:

- Tipo (Inherente o dependiente del sistema).
- Documentación previa requerida: Para realizar la medición.
- Método (forma de realizar la medición).
- Variables: Valores a recabar.
- Fórmula: para obtener el valor final objeto de la métrica.
- Escala: Si es necesario para categorizar los datos obtenidos.

Tomando como ejemplo alguna característica de la norma (en nuestro caso la de completitud), dicho método establecería la obtención de los siguientes valores:

- Tipo: Inherente
- Documentación previa requerida: Por cada atributo a evaluar, indicar si es obligatorio
- Método para la medición: Verificar para cada atributo si hay valores vacíos o en blanco
- Variables: X = Porcentaje de datos del atributo que se encuentran completos
- Fórmula: Valor = X/100 (%)
- Escala: (1) X<=10 %, (2) 10 % <X <= 45 %, (3) 45 % <X <= 85 %

Que serían extensibles al resto de las características de la norma. Como ya hemos mencionado y como ejemplo de implementación, se aplica esta evaluación a todas las variables definidas por la norma a una única tabla de ejemplo con diez atributos de de una base de datos, en la que para cada uno de ellos se establecen ciertos niveles de aceptación:

- Inaceptable.
- Mínimamente aceptable.
- Rango objetivo.
- Excede los requerimientos .

Obviamos el detalle de dicha implementación por estar asociada a un ejemplo muy trivial y sobre el que ya se ha descrito el método propuesto.

4.4 Methodologies for Data Quality Assessment and Improvement

De todas las metodologías no comerciales revisadas para la evaluación de la calidad del dato, esta, [39] aun a pesar de su antigüedad (fue publicada en el 2009) nos parece la más completa y exhaustiva.

Los autores reconocen la gran variedad de técnicas disponibles para evaluar y mejorar la calidad de los datos, y debido a su diversidad y complejidad, proponen metodologías que ayuden a seleccionar, personalizar y aplicar alguna de ellas.

Todo ello en función de las dimensiones de calidad requeridas, los tipos de datos e incluso los sistemas de información sobre los que se sustenten.

Dado lo ambicioso del trabajo, este comienza haciendo una comparativa de las diferentes perspectivas desde las que se pueden analizar distintas metodologías de la calidad del dato:

1. Las fases y los pasos constitutivos de cada metodología.
2. Las estrategias y técnicas adoptadas por cada una de ellas.
3. Las dimensiones y métricas elegidas para detectar y mejorar los niveles de calidad.
4. tipos de costes asociados a cuestiones relacionadas con la calidad de datos, incluyendo:
   a) Costes asociados a la pobre calidad de los datos originales, que denomina *costes indirectos*.
   b) costes asociados a la estimación y a actividades de mejora de los datos, *costes directos*.
5. Los tipos de datos tratados.
6. Los tipos de los *sistemas de información* que usan, modifican o gestionan los datos considerados.
7. Las *organizaciones* involucradas en el proceso de creación o actualización de los datos.
8. Los *procesos* que crean o actualizan datos con el objetivo de generar los servicios requeridos.
9. Los *servicios* producidos por los procesos considerados por la metodología.

Nos ha parecido interesante hacer una reseña completa porque introduce unos actores no contemplados en otro tipo de estudios proporcionando una visión más amplia y dando cabida a conceptos como costes –imprescindibles en cualquier planteamiento serio de la materia al intentar introducir algún control de calidad–, tipo de organización, de proceso o de servicio proporcionado por el sistema de datos objeto de evaluación.

Estrategias y técnicas

En su intento de mejora, las diferentes metodologías pueden adoptar dos tipos de estrategias generales: Guiadas por datos o enfocadas a procesos.

Las *estrategias guiadas por datos* intentan mejorar la calidad de los datos modificando su valor, por ejemplo, datos obsoletos pueden ser sustituidos por datos más actualizados.

Por el contrario, las *estrategias enfocadas a procesos* intentan mejorar la calidad alterando los procesos que crean o modifican datos, por ejemplo obligando a que adopten formatos concretos antes de su almacenaje.

Entre las técnicas del primer grupo (orientadas a datos) podrían citarse:

1. Adquisición de nuevos datos (de mayor calidad, para sustituir a los presentes).
2. Estandarización y normalización (sustitución de abreviaturas o apodos por datos completos, por ejemplo).
3. Vinculación de registros, que identifiquen las representaciones de datos en más de una tabla que referencien el mismo objeto real.
4. Integración de datos, en una vista unificada de datos proporcionados por fuentes heterogéneas.
5. Confiabilidad de la fuente de datos, primando aquellas que proporcionen mayor calidad.
6. Localización y corrección de errores detectando registros que no satisfagan determinadas reglas de calidad.
7. Optimización de costes, definiendo acciones de mejora de calidad a lo largo de un conjunto de dimensiones.

Las técnicas enfocadas en los procesos se basan en dos estrategias principales:

- Control de procesos, que inserta puntos de control en las fases de creación, actualización o acceso a los datos.
- Rediseño de procesos, basada en el rediseño de los procesos para eliminar o mitigar las causas de la pobre calidad de los datos o implementen nuevas acciones que generen datos de mayor calidad.

Dimensiones y métricas

Común a todas la metodologías se presentan los conceptos de dimensión y métrica para evaluar cualquier conjunto de datos.

Aunque los autores mencionan que «las dimensiones de calidad pueden referirse tanto a la extensión de los datos como a sus valores (su intensión o características)»[12], la mayoría de dimensiones de calidad y sus métricas referidas a los datos se centran en considerar los valores de los mismos en lugar de su esquema o estructura.

Sin que exista un consenso sobre dichas dimensiones a las que aplicar métricas y el significado exacto de cada una de ellas, se reseñan expresamente las de precisión, completitud, consistencia, junto con otras tres relacionadas con el tiempo, que es un aspecto importante relacionado con los datos: Su nivel de actualización, volatilidad y oportunidad.[13]

Como estos términos son novedosos respecto a la literatura anteriormente revisada y nos parecen interesantes vamos a definirlos:

- El nivel de actualización sería el grado en el que un dato está actualizado.
- La volatilidad describiría el tiempo durante el cual un dato sería válido en el mundo real.
- Su oportunidad tienen distintas acepciones, siendo las dos principales el promedio de edad de los datos de una fuente, y la extensión durante la cual los datos estarían lo suficientemente actualizados para un propósito concreto.

Costes

No parece oportuno revisar en detalle este concepto en el marco del trabajo comentado.

Según los autores, «Los costes de la calidad del dato son la suma de los costes de evaluar su calidad, la de sus procedimientos de mejora –también referido como el coste del programa de la calidad de datos–, junto con el coste asociado a la pobre calidad de los datos».

El coste generado por una baja calidad de los datos puede ser reducido implementando programas de calidad de datos más efectivos, lo que a su vez incrementa el coste, por lo que de un modo recíproco al incrementar el coste de los programas de calidad se reduce el coste de la pobre calidad de los datos, lo que a nuestro entender nos induce a buscar un punto de equilibrio entre ambos extremos.

Por su parte, los costes –siempre según los autores–, pueden clasificarse en:

- *Costes de proceso*, que serían los asociados a volver a ejecutar procesos completos debido a errores en los datos.
- *Costes de oportunidad*, debidos a la pérdida de beneficios propiciada por una mala calidad de los datos.

Cabe reseñar que al contrario que los costes dependientes de la implementación de un programa de calidad, los costes de la pobre calidad de los datos son fuertemente dependientes del contexto, lo que hace su evaluación especialmente difícil: El mismo dato con un determinado nivel de calidad tiene diferente impacto según el uso que deba darle su destinatario.

[12] Nos parece oportuno recordar que la extensión sería el número de sujetos o miembros a los que se aplica un término o predicado y la intensión es su conjunto de sus características o rasgos definitorios.

[13] Hemos traducido los términos *currency*, *volatility* y *timeliness*.

Tipos de datos

Tradicionalmente se acepta una clasificación general de los datos como *estructurados*, *semiestructurados* y *no estructurados*.

Las tablas relacionales o los datos estadísticos caerían en el primer grupo, mientras que cuando incurren en cierto nivel de flexibilidad [14] pertenecerían al segundo. En la tercera categoría, la de datos no estructurados, se verían incluidos los que presentasen cualquier secuencia de símbolos (lenguaje natural) extraída de cuestionarios de respuesta libre.

Esta definición nos interesa pues conlleva a la conclusión –defendida por los autores y que nos parece lógica– de que las técnicas de implementación de calidad del dato se vuelven más complejas (y costosas) cuando los datos pierden estructura.

Así pues, la mayoría de las técnicas de calidad del dato son aplicables a los dos grupos caracterizados por su mayor nivel de estructura.

Consideraciones adicionales

Como el trabajo evaluado nos parece el de mayor interés, hemos incluido en el anexo 9.1 otras consideraciones tratadas en el mismo referentes a **Sistemas de Información**, **Metodologías** consideradas y **Cuestiones pendientes** propuestas por los autores.

Ya para recapitular, a lo largo de esta sección hemos abordado distintos enfoques sobre la evaluación y mejora de la calidad del dato desde ópticas no comerciales:

La sistematización que nos proponen pueden ayudarnos a extraer conclusiones más elaboradas sobre nuestros propios requisitos y que podrían ser válidos a la hora de optar por alguna de las soluciones disponibles en el mercado abordadas anteriormente.

También podrían ser el punto de partida para elaborar nuestro propio plan de calidad y flujo de trabajo, –incidiendo en acciones y controles– a implementar sobre nuestros datos asignando los recursos necesarios para ello.

[14] Los autores los refieren como *schemaless* o *self-describing*, (sin esquema o autodescriptivos), citando como ejemplo los que presentasen una estructura tipo *XML*.

CAPÍTULO 5

Algunas técnicas de Ciencia de Datos relacionadas con la calidad del dato

5.1 Ciencia de Datos y calidad

Parece oportuno vincular y aplicar algunas de las técnicas estudiadas en un grado de ciencia de datos a la solución de algunos de los problemas que son objeto de estudio y cuantificación por parte de las herramientas de evaluación de la calidad del dato expuestas con anterioridad.

Muchas de las herramientas consideradas proponen una analítica descriptiva, o cuanto más diagnóstica, sobre el conjunto de datos evaluados. Otras realizan acciones de integración o correctoras. ¿Por qué no dar un paso más allá e intentar paliar, con alguna metodología válida, las posibles carencias que se hayan podido detectar en la calidad de nuestros datos?

Brindaríamos así la posibilidad de reforzar un análisis prescriptivo, al menos a algunas cuestiones puntuales, cuando las necesidades de uso de los datos tratados lo hiciesen necesario.

Bajo esta nuevo enfoque proponemos algunos experimentos que se exponen y justifican a continuación.

Los mismos están basados en modelos de datos públicos, sencillos y conocidos con la intención de que esa sencillez facilite el juicio sobre las razones expuestas.

5.2 Imputación de datos incompletos: Aproximación por *kNN*

Suele ser habitual encontrar conjuntos de datos que durante su recolección, proceso o transmisión generen datos faltantes.

En capítulos anteriores hemos hablado de la ***completitud*** de los datos como una dimensión evaluable de su calidad. Vamos a ver como el grado de la misma afecta a un modelo diseñado para imputar esos datos en función de otros que si se conozcan con exactitud.

Para ello, vamos a tomar como ejemplo un caso típico[1] presente en los algoritmos de análisis de expresiones genéticas. [40]

[1] *Missing value estimation methods for DNA microarrays*

Los algoritmos para el análisis de expresión genética requieren de matrices completas de datos como entrada. En estas matrices la filas representan los distintos genes, y las columnas la expresión de dichos genes bajo diferentes condiciones.

Estas matrices, por diferentes razones (métodos ineficientes de recogida, degradado de las muestras...) suelen presentar valores faltantes. Se impone pues la necesidad de implementar métodos que imputen estos valores faltantes con la mayor precisión posible.

El trabajo citado y revisado en [40] evalúa tres modos de abordar el problema:

- Una descomposición singular de valores *(SVD)*
- K-vecinos más próximos (*kNN*) ponderados
- Media de filas

En el caso planteado, estos métodos se aplican a muestras con un rango de valores faltantes del 1 al 20 % y los autores concluyen que el método *kNN* es el más robusto y preciso para imputar valores faltantes en el ámbito de estos experimentos.

kNN[2] (o técnica de los K vecinos más próximos) es una técnica de clasificación basada en distancias.

Para reproducir algo parecido, vamos a utilizar un conocido repositorio de datos. Se llama *load digits* y está disponible en la librería de *Python scikit-learn*.

Sus características están muy bien documentadas, pero en resumen se trata de dígitos escritos a mano y después pixelados en matrices de 8 x 8 en los que cada posición de la cuadrícula representa un tono de gris. Dichos dígitos está debidamente etiquetados con el valor, de 0 a 9, al que representan.

Disponemos pues de unos 1.800 vectores de 64 componentes cada uno, absolutamente balanceados entre las clases que representan, y a los que aleatoriamente vamos a privar de alguno de sus 64 elementos para simular valores faltantes.[3]

En la figura 5.1 observamos valores muy similares (7 ú 8) para el número k de vecinos a considerar para realizar la imputación de los valores faltantes con la máxima precisión posible (en nuestro caso con el menor valor para el error cuadrado medio o *MSE*) y para distintos porcentajes de datos faltantes, en el rango de 5, 10, 15 y 20.

Esto nos induce a afirmar que esta técnica, empleable para imputar valores faltantes, es robusta e independiente de diferentes grado de completitud de nuestros datos.

En el anexo 9.2 proponemos un ejemplo propio para la implementación y la aplicación de este método.

5.3 El efecto de los datos erróneos

Hemos tipificado también la ***exactitud*** como otra dimensión evaluable de la calidad de los datos.

Vamos a hacer un pequeño experimento para intentar averiguar la sensibilidad de otro modelo a los errores a los que puedan estar sujetos los datos en los que sustente.

[2] K *nearest neighbors* en inglés

[3] en el sentido estadístico *MCAR*, o *Missing Completly At Random*, lo que implica que las causas que producen los datos faltantes no están relacionados con los datos.

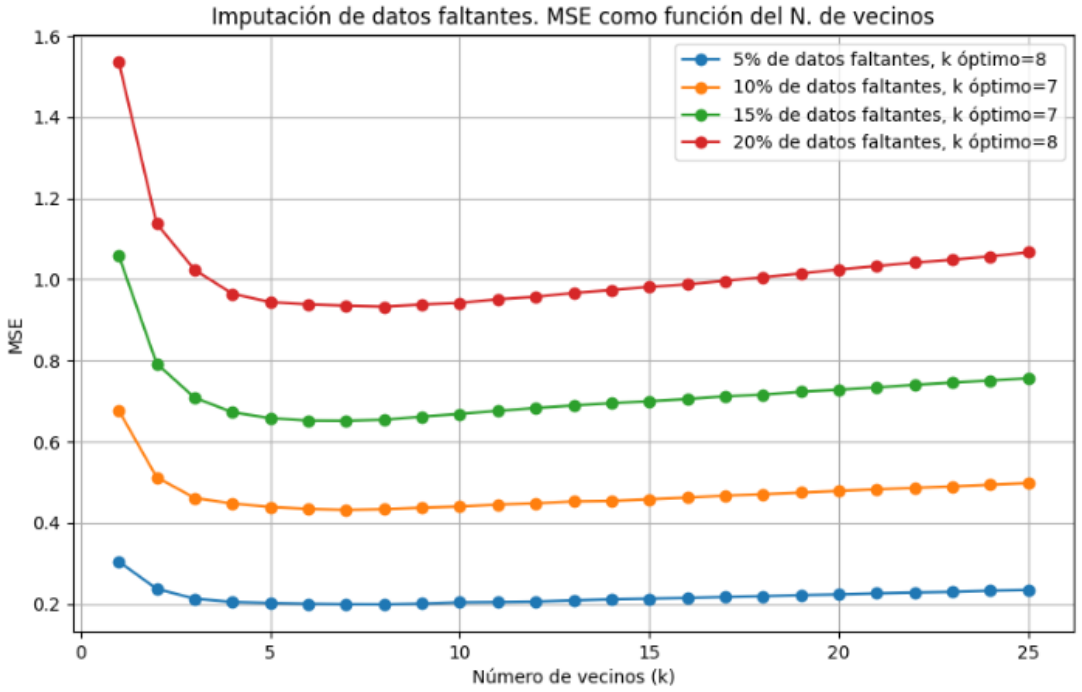

Figura 5.1: Imputación de datos faltantes por *KNN*

En este caso vamos a usar otro conocido conjunto de datos: *California Housing Prices*.[4] Contiene nueve variables predictoras que nos permiten establecer un sencillo modelo de regresión lineal para predecir el precio de las viviendas.

Vamos a descartar (por simplicidad) las de su geolocalización (dos variables) y proximidad a la costa (una variable).

Sometemos a cada una de las variables restantes por separado a un gradual deterioro en sus valores, (gradual en el sentido de incrementar su porcentaje de alteración), e intentamos visualizar el efecto de esa alteración en la capacidad de predicción del modelo.

Dicha capacidad de predicción sería inversamente proporcional al incremento de sus métricas de error, (*MSE* y *MAE*, por ejemplo). Cualquiera de ellos nos sirve, y las siguientes afirmaciones son extensibles a ambos.

En la figura 5.2 hemos ido alterando los valores de las diferentes variables predictoras del modelo en un porcentaje cada vez mayor en un rango de 0 a 20.

Apreciamos un efecto notable: Una variable, *median income*[5] es absolutamente sensible a la introducción de errores a la hora de ser buena predictora del valor del inmueble; cuando su porcentaje de errores aumenta, también lo hace el error de estimación del modelo.

En la figura 5.3 se ha hecho lo propio pero con una diferencia: La magnitud del error introducido.

Cuando en este caso particular nos referimos a la introducción de errores, lo que estamos haciendo es alterar el valor original del dato. En cada caso, el mismo porcentaje de registros de cada variable predictora es alterado. ¿En qué medida?

[4] *kaggle*, Precios medios de las casas en distintos distritos de California extraídos del censo de 1990.

[5] Las variables consideradas *son median income, total bedrooms, total rooms, population, households, housing median age*, cuyo significado se corresponde a valores de ingresos medios, n. de dormitorios, n. total de habitaciones, población, n. de hogares y edad de la vivienda en el barrio donde se ubica.

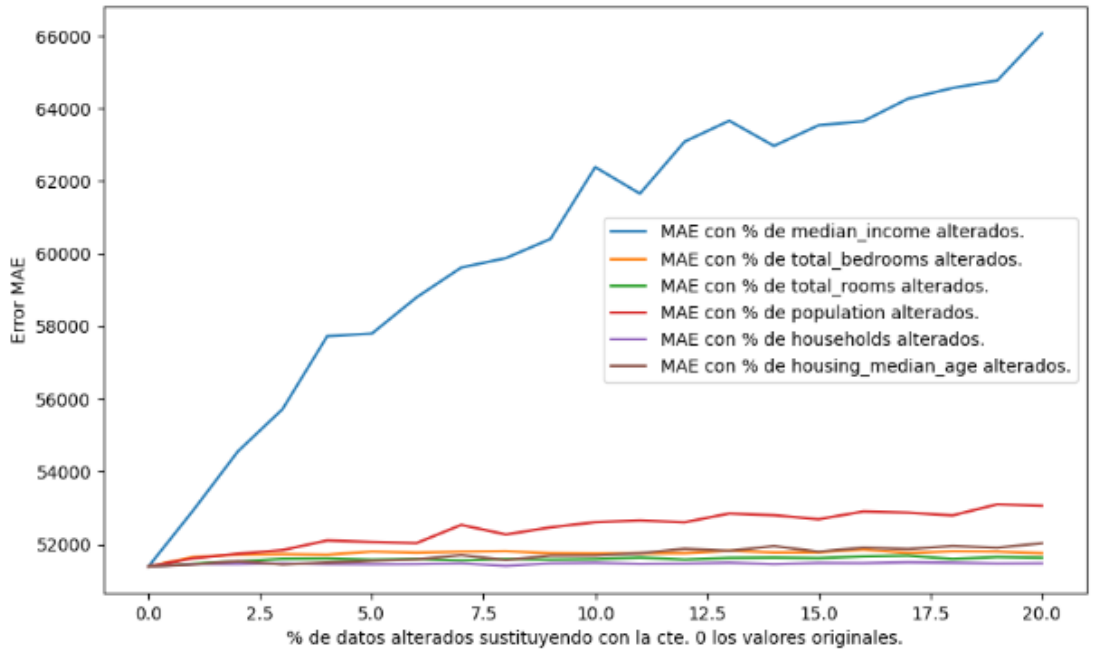

Figura 5.2: Porcentaje da datos erróneos y evolución del *MAE* del modelo. (1)

Tras distintas tentativas hemos optado por dos enfoques: En el primero consideramos oportuno que cada dato soporte una alteración proporcional a su valor, y en el segundo, que todas las variable sustituidas lo sean por un mismo valor arbitrario.

Los gráficos reseñados muestran el efecto de esas alteraciones aplicados a distintos porcentajes de ocurrencia sobre todas las distintas variables predictoras del modelo por separado.

El grado de precisión del modelo solamente es sensible a la introducción de errores en una variable (*median income*) y en mucha menor medida en otra (*population*). Los mismos porcentajes de datos alterados en otras variables parecen no tener efecto en el rendimiento.

Más experimentos de esta índole, evaluando la sensibilidad de distintos modelos a la introducción de errores medibles es posible que nos proporcionase algún método indirecto de análisis no disponible en otras circunstancias.

En este caso particular, estamos tentados a afirmar que:

- Esas variables sensibles son los componentes principales que explican el modelo.
- El comportamiento de todas las demás podría denotar un alto grado de colinealidad (por su poco efecto predictivo).

Parece este un buen modo indirecto de obtener los mismos resultados que en un análisis de componentes principales a partir de la sensibilidad de las distintas variables de un modelo de regresión lineal a la existencia de datos inexactos.

Se observa otro fenómeno significativo: Cuando la magnitud del error introducido aumenta, todas las variables alcanzan su umbral de máximo error con un porcentaje de errores introducido muy pequeño y luego se mantiene constante.

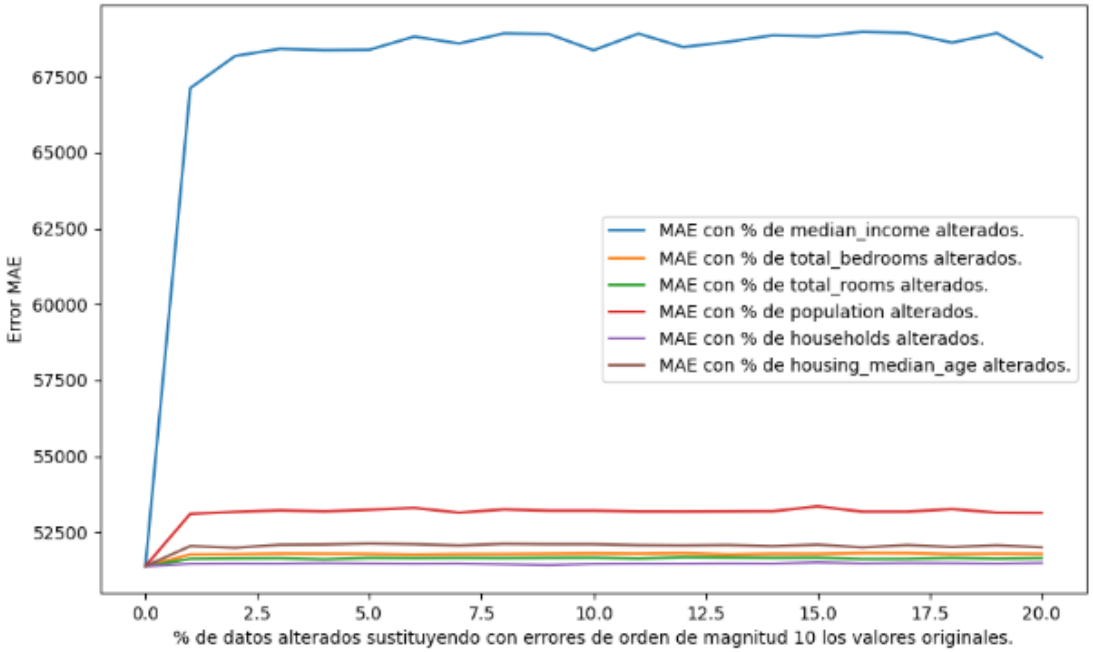

Figura 5.3: Porcentaje da datos erróneos y evolución del *MAE* del modelo. (2)

5.4 El efecto de la precisión y la actualidad

En este caso vamos a utilizar un conjunto de datos con una componente temporal. Dicho atributo representa la fecha cuando se registraron el resto de las variables que integran cada uno de los registros compilados.

El modelo de partida también es de sobra conocido: *Bike sharing: Times Series Analysis*[6]. Este conjunto de datos almacena ocurrencias de alquileres de bicicletas con datos sobre la fecha y las circunstancias meteorológicas existentes en el momento de cada registro. El tipo de día (laborable o festivo) y la estación del año también están registrados como variables.

Contamos pues con una serie temporal de datos, que como tal, presenta una tendencia un nivel, un ruido y una estacionalidad asociada.

Se pueden utilizar diversas técnicas para entrenar un modelo que pretenda predecir el número de alquileres para una fecha concreta teniendo en cuenta la naturaleza de los datos disponibles.

Para tratar el atributo fecha, que como tal presenta un formato original del tipo *día - mes - año*, este se sustituye por un valor numérico que represente los días transcurridos desde la fecha del primer registro. En este caso mantendríamos la granularidad de los datos originales respecto a este atributo.

A renglón seguido, aplicaremos el mismo modelo seleccionado a *datos menos precisos en este atributo*. Para ello, sustituiremos el número de días transcurridos por el de semanas (también como atributo entero) para ver hasta que punto se ve afectada la calidad de la precisión de nuestro modelo.

Hemos implementado un modelo del tipo *random forest* en el lenguaje *R* que nos permite predecir el posible valor del número de bicicletas alquiladas a partir de las demás

[6]kaggle, Bikesharing in Washington D.C. dataset.

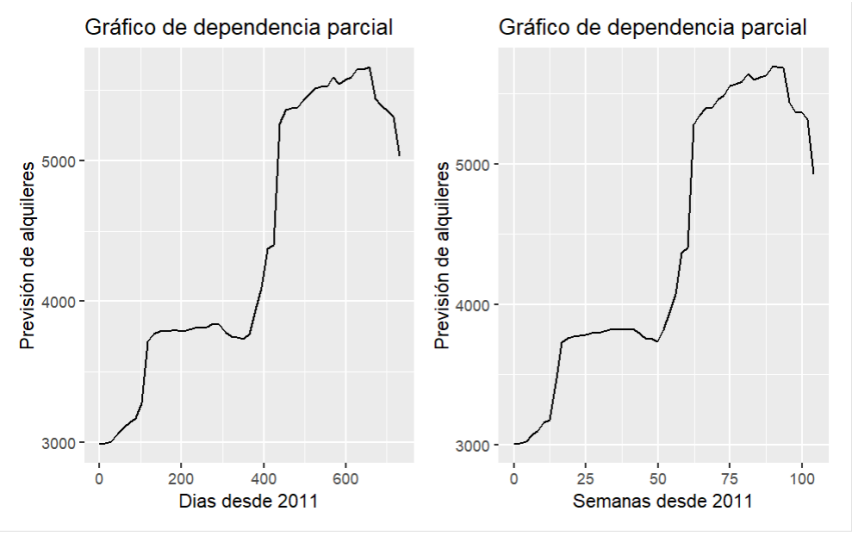

Figura 5.4: Gráficos de dependencia parcial de la variable temporal expresada en semanas y días.

variables predictoras y hemos aplicado dicho modelo ambos conjuntos de datos, uno que especifica el día preciso de alquiler y otro solo la semana en que se produjo

En la figura 5.4 mostramos los gráficos de dependencia parcial [7], que representan el ejemplo marginal de la característica temporal en el resultado predicho por el modelo.

Son casi idénticos: Ambos muestran una evidente tendencia a crecer con el paso del tiempo y cierta estacionalidad. El que expresa la variable tiempo en semanas, tal vez presenta un muy ligero menor nivel de definición y cambios más abruptos.

¿Hasta que punto realizamos peores predicciones con datos temporales menos precisos?

Realizamos una validación cruzada de 10 pliegues sobre los datos disponibles (731 registros) obteniendo los siguientes valores para el MSE[8] en ambos modelos (con datos diarios y con datos semanales) contemplando siempre el mismo número de variables adicionales. [9]

En ambos casos obtenemos resultados muy similares al evaluar la importancia discriminante de cada variable (Ver figura 5.5).

Vamos a ver como este cambio puede afectar al error en la predicción.

La función *rfcv* del paquete *randomForest* en R puede ser útil para esto. Esta función realiza una validación cruzada y elimina las variables menos importantes, permitiendo apreciar cómo el error cambia a medida que se eliminan las variables.

La Impureza del Nodo Incrementada[10] es una medida de cuánto mejora la predicción del modelo cada variable. Se calcula sumando las disminuciones en la impureza del nodo para cada variable sobre todos los árboles en el bosque. La impureza del nodo se mide generalmente usando el índice de Gini o la entropía (el primero en este caso), que son medidas de cuán mezcladas están las clases en un nodo.

[7] O *Partial dependency Plots*.

[8] Que en este caso representaría el valor al cuadrado de número de bicicletas que erramos en cada predicción. Por cuestiones de interpretabilidad podríamos obtener el RMSE que es simplemente la raíz cuadrada del anterior.

[9] *Ceteris paribus*.

[10] *IncNodePurity*, en el eje horizontal de la figura 5.5

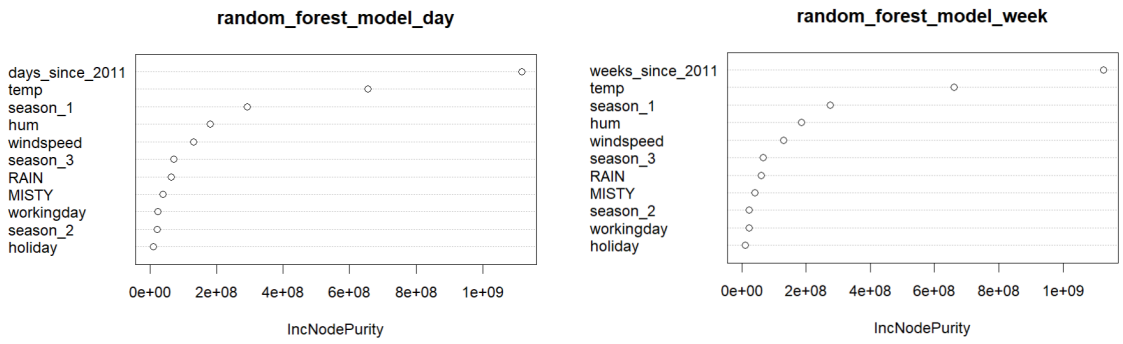

Figura 5.5: Índice de impureza para cada una de las variables en ambos modelos.

```
$error.cv
     11       6       3       1
436920.4 458289.5 519041.8 782185.3

                IncNodePurity
workingday           21755192
holiday              10060584
season_1            292147900
season_2             21227742
season_3             69958139
MISTY                38489235
RAIN                 62889086
temp                655452781
hum                 179958446
windspeed           130435058
days_since_2011    1117596968
```

```
$error.cv
     11       6       3       1
446452.1 498512.9 535087.4 919525.7

                IncNodePurity
workingday           22037455
holiday               9736565
season_1            274671491
season_2             22517102
season_3             65338783
MISTY                40017933
RAIN                 59387530
temp                661577045
hum                 186022564
windspeed           129425712
weeks_since_2011   1126653731
```

Figura 5.6: Error en función del número de variables seleccionadas e índice de impureza de cada una.

La variable temporal (expresada en días o en semanas) es con creces la más determinante de nuestro modelo independientemente de su nivel de precisión.

El error cuadrado medio de la predicción es ligeramente superior cuando la variable temporal es menos precisa, pero a su vez, el índice de impureza de la misma (su valor discriminante) es mayor.

5.5 Muestreo de bases de datos no relacionales

Las bases de datos *NoSQL* [11] representan un paradigma de almacenamiento y tratamiento de grandes cantidades de datos distinto al de las bases de datos relacionales clásicas. Ya sean del tipo clave-valor, documental o grafo, ofrecen características y prestaciones (redundancia, flexibilidad de esquema, escalado horizontal, facilidad de uso...) que las hacen ser muy atractivas hoy en día como primera opción de uso.

Sin embargo, la especial concepción de su arquitectura las distancian de los tradicionales esquemas relacionales en ciertos aspectos que pueden ser desventajas: Recorrer todos sus registros, –para por ejemplo hacer un conteo, o verificar ciertos requerimientos de calidad que deba cumplir algunos de sus atributos–, puede ser tan costoso en términos de coste computacional, que puede hacerlo inviable desde el punto de vista económico.

[11] *Not only SQL*: No solo de lenguaje de consulta estructurado.

Para soslayar ese inconveniente, podríamos hacer uso de algún método estadístico de muestreo, como por ejemplo el *método Montecarlo*, que es un método estadístico que puede ser utilizado para aproximar expresiones matemáticas costosas de evaluar con exactitud.

Supongamos una tabla en una base de datos *Dynamo* [12] que contenga 10^7 registros. Si estamos realizando el modo de pago por solicitud bajo demanda, –en el que el precio por lectura puede ser de $0,00013 por registro en la región *EU-WEST-1*.–, y el tamaño de un registro medio sea, digamos de 1 Kb, el coste de lectura de nuestra tabla sería:

$$\text{costo de lectura} = \text{N}^{\text{o}} \text{ de registros} * \text{tamaño de registro} * \text{coste/registro}$$

$$10^7 \text{ registros} * 1\text{Kb} * \$0.00013/\text{registro} = 1.300,- \$. \ [13]$$

Si por el contrario:

- Definimos la característica o el atributo de interés
- Hacemos un muestreo aleatorio de digamos, 10^3 registros
- Basándonos en los datos de la muestra, calculamos la proporción de registros totales que cumplen (o no) con la característica deseada, lo que nos daría una estimación de la proporción en toda la base de datos
- Calculamos los intervalos de confianza de nuestra estimación
- Iteramos repitiendo el proceso de muestreo y cálculo varias veces (esta es una de las posibilidades del *método Montecarlo*), podemos calcular la media y la mediana, –por ejemplo– de nuestras estimaciones para obtener una estimación final más robusta.

En este caso, el coste de nuestra evaluación sería el de aplicar la formula anterior al tamaño de la muestra multiplicado por el número de iteraciones:

$$10^3 \text{ registros de muestra} * 1\text{Kb} * \$0.00013/\text{registro} * 10 \text{ iteraciones} = 0{,}13 \$$$

Hemos visto así que estableciendo unos intervalos de confianza adecuados para la índole de nuestros datos a la vez que su tolerancia al posible error, podemos evaluar con toda la aproximación requerida la adecuación de nuestros datos *NoSQL* a unos requerimientos dados de calidad en cualquiera de sus atributos obteniendo una considerable reducción de coste.

[12] https://aws.amazon.com/es/dynamodb/

[13] Este es un coste aproximado que puede variar en función de muchos factores, pero nos interesa reseñarlo por su magnitud.

CAPÍTULO 6
Resultados

Hemos hecho un repaso al concepto de la calidad del dato desde distintas ópticas y a la vez una síntesis de metodologías y herramientas disponibles, deteniéndonos en mayor medida en los aspectos que el autor, subjetivamente, ha juzgado de mayor interés.

Esta apreciación de las herramientas existentes actualmente nos ha permitido sugerir un método –a nuestro entender neutral–, que permite evaluarlas bajo la premisa de que previamente sepamos considerar bien nuestras necesidades y darles un peso.

Al tratarse de un aspecto tecnológico y en continua evolución condicionado por múltiples factores, resulta difícil hacer afirmaciones u obtener verdades categóricas y aplicables de un modo axiomático, aunque esta conclusión ya pueda, *per se*, considerarse como tal.

Adicionalmente, la variabilidad y el carácter a veces estocástico del mundo de los datos nos obligan a ser muy cuidadosos en nuestras afirmaciones referidas a los mismos.

Sin embargo y como hemos podido entrever en los pequeños experimentos planteados, nuestras técnicas de ciencia de datos pueden ser un buen aliado a la hora de evaluar la calidad de un conjunto de datos o paliar la carencias en la misma.

A nuestro entender también constatamos que incertidumbre no implica ausencia de certeza, –o al menos no hasta el punto de impedir la toma de decisiones–, mientras establezcamos unos márgenes de confianza aceptables y podamos asumir el coste de incurrir en el error.

Podemos implementar pequeñas acciones puntuales que tengan un buen impacto en el desempeño de cualquier modelo.

Como punto no menos importante, también hemos comprendido que el buen conocimiento del dominio tratado y al que se refieran los datos es una condición que nos va a permitir hacer buenas consideraciones sobre la calidad de los mismos.

Con todo esto, queremos significar que implementar medidas que conlleven a mejorar el resultado de nuestras decisiones basadas en técnicas que minimicen la posibilidad de equivocarnos siempre debería parecernos, en cualquier sentido del término, una buena inversión.

CAPÍTULO 7
Conclusiones

Tras exponer la necesidad, analizado las prestaciones de distintas soluciones, revisado la normativa y distintas tentativas académicas para sistematizar la mejora de calidad del dato, constatamos que como concepto general es un tema intrínsecamente complejo y de difícil generalización por muy distintos factores:

- La índole de los datos a tratar y su estructura: ambos condicionan el enfoque de su análisis, que a su vez se ve sujeto por la normativa y las restricciones aplicables dada su particular naturaleza.
- La infraestructura disponible para alojar y tratar esos mismos datos, lo que nos obliga a recurrir a herramientas acordes y adecuadas compatibles con la misma.
- Los actores implicados en su tratamiento, con distintos requerimientos de accesibilidad y privacidad.
- Un entorno tecnológico rápidamente cambiante y en continua evolución en el que unos paradigmas se ven rápidamente superados por otros.
- Los medios o recursos disponibles para abordar el problema no siempre permiten realizar, por meras razones económicas, todas las acciones deseables.

Sin embargo, podemos extraer algunas conclusiones y recomendaciones generales válidas que nos son de utilidad.

- La calidad del dato ha de contemplarse como un **proceso iterativo y continuo** que tienda a su consecución, que en ningún caso podemos tipificar como absoluta.
- La **acumulación de buenas prácticas** en todas las fases del ciclo de vida de los datos tienden a reforzar la calidad de los mismos y a minimizar el impacto de los aspectos negativos de su ausencia.
- Una buena visión general previa y **conocimiento del dominio** tratado puede propiciar que tomemos decisiones que faciliten unos aceptables niveles de calidad.
- **Pequeñas soluciones** a problemas muy concretos o actuaciones en cualquier estadio de la gestión de datos, pueden tener un **gran impacto** en estadios posteriores mitigando la propagación de errores generados.
- La implementación de **políticas de calidad** ha de realizarse *por diseño* en las fases más tempranas de la creación de nuestras estructuras y flujos de datos.

Por otro lado, hemos constatado que tenemos a nuestra disposición herramientas muy precisas para modelar datos en ámbitos muy específicos: Si ajustamos el nivel de granularidad de nuestros requerimientos de calidad al uso de las mismas podemos realizar mejoras concretas.

Si encadenamos estas actuaciones podemos crear un ciclo de mejora continua.

La automatización robótica de procesos [1], podría ser también una buena aliada a la hora de encarar tareas de calidad dada la complejidad de muchos de los procesos implicados, que por otro lado y al estar tan estructurados, se prestan a automatización.

[1] *Robot Process Automation* o *RPA*

CAPÍTULO 8
Trabajos futuros

Nos parecería lógico profundizar en algunos de los aspectos de mayor calado que hemos esbozado a lo largo de estas páginas y que no hemos abordado por razones de concisión.

Entre ellos podemos citar:

- Crear una tabla exhaustiva de características asimilables a cada una de las herramientas comerciales evaluadas y utilizarla para calcular decisiones en distintos casos de negocio simulados.

  Esto podría facilitar el tomar partido por alguna de ellas en función de los datos y el entorno a tratar.

  Sin embargo, la rápida evolución de las versiones ofrecidas, sus prestaciones y disponibilidad harían rápidamente obsoleto cualquier intento de categorización o clasificación.

- Crear alguna función de optimización que contemplase las restricciones de alguna necesidad completa de implementación y nos sirviese para seleccionar una opción evaluando las características ponderadas de las mismas.

- En la misma linea, generar un árbol de decisión que nos permitiese optar por alguna de ellas en función de preguntas discriminantes de distinto nivel. Este intento también se vería afectado por la circunstancia expuesta en el tercer párrafo del primer punto de esta lista.

- Participar en algún programa institucional, relacionado con cualquiera de la normativa revisada, que brinde la posibilidad de ser un actor activo en la materia.

  Algunas de la legislaciones europeas son tan novedosas y están todavía en estadio tan embrionario que sus promotores brindan la posibilidad de colaborar en su desarrollo mediante sugerencias o contribuyendo a su difusión.

  En los nuevos textos legales se describen nuevos e importantes roles relacionados con la calidad del dato y la información. Optar a asumir alguno de ellos.

- Centrarse en las normas de calidad y desarrollar un completo manual de calidad de alguna de ellas para algún supuesto concreto.

- Evaluar más dimensiones de calidad desde de la ciencia de datos incidiendo en el desarrollo de ejemplos más complejos. Esto por si mismo ya podría constituir un trabajo independiente.

- Algunos de los trabajos académicos revisados, aun a pesar de su calidad, incurren en la obsolescencia. Una actualización de los mismos o dar continuidad a los interrogante abiertos desde circunstancias y técnicas más actuales sería también un buen objeto de estudio.
- Hemos mencionado la automatización robotizada de procesos: Implementar algún tipo de operativa enfocada en la calidad del dato que haga uso de ella sería otro buen punto de partida.
- El uso también de otras herramientas específicas para la gestión de las versiones de datos [1] podrían ayudarnos a arrojar luz sobre el efecto de alteraciones en la calidad de los datos sobre el desempeño de cualquier modelo propuesto.

[1] DVC o *Data Version Control.*

CAPÍTULO 9
Anexos

9.1 Method. for Data Quality Assessment & Improvement (II)

Sistemas de información

Las metodología de calidad del dato también se ven condicionadas por el tipo de sistema de información de utilizado por la organización. «La literatura proporciona el concepto de arquitectura del sistema de información para definir el modelo de coordinación soportado por el sistema de información de una organización».[1]

Nos interesa este concepto porque los autores afirman que la complejidad de la evaluación de la calidad de los datos es inversamente proporcional al nivel de integración del sistema de información: A menor nivel de integración mayor nivel de complejidad.

Se distinguen los siguientes modelos en función a su nivel de integración:

- Sistemas monolíticos: Son los que constan de aplicaciones de un solo nivel y no proporcionan servicios de acceso a datos. Dichas aplicaciones no comparten datos (aunque pueden compartir la misma base de datos subyacente) y pueden generar duplicaciones afectando a todas las dimensiones de calidad.
- Almacenes de datos[2]: Definidos en este caso como una colección centralizada de datos accesibles por múltiples bases de datos. Dichos datos centralizados son actualizados periódicamente por las bases de datos originales y por procedimientos que automatizan la extracción y agregación de datos, lo que implica un cierto nivel de integración antes de su almacenamiento final.
- Sistemas de información distribuidos: Contemplados como una colección de aplicaciones coordinadas por un flujo de trabajo. Los datos pueden almacenarse en distintas bases de datos pero su interoperabilidad es garantizada por por la integración lógica de sus aplicaciones.
- Sistemas cooperativos de información[3]: Son sistemas de información a gran escala que interconectan organizaciones autónomas con un objetivo común. En este caso los datos no presentan integración lógica -se almacenan en bases de datos separadas-, pero las aplicaciones incorporan transformación de datos y procedimientos de intercambio que permiten la cooperación entre procesos, lo que implica que la integración se realiza a nivel de proceso.

[1] *Information system architecture* o *IS architecture*.
[2] Los ya mencionados *Data Warehouses*.
[3] *Coopetative information systems CIS*.

- Sistemas de información *Web*[4]: Que sería un sistema de información que adoptase tecnologías *web*, que desde una perspectiva técnica representa el uso de aplicaciones cliente/servidor[5].
- Sistemas de información *entre pares*[6]: Que a diferencia del caso anterior no distingue entre servidores y clientes y que está formado por nodos idénticos que comparten datos y aplicaciones con el objeto de satisfacer los requerimientos de los usuarios de un modo colectivo.

  Muchas bases de datos no relacionales (NoSQL) presentan un arquitectura similar distribuyendo propagando la información entre distintos nodos con el objeto de garantizar su consistencia, disponibilidad y resilencia a fallos.

Metodologías consideradas

El núcleo del trabajo evaluado en esta sección se centra en desglosar las características de las distintas metodologías consideradas en pos de la calidad del dato.

Nos parece oportuno incluir la tabla presentada en el trabajo original. Sus tres columnas (figura 9.1) muestran el acrónimo asociado a la metodología, su nombre completo y su referencia principal.

No abundamos en ellas porque los autores lo hacen de un modo exhaustivo relacionando cada una de ellas con los fundamentos teóricos y clasificatorios que previamente exponen.

Sin embargo, si nos parece útil mencionar las cuestiones abiertas que plantean al final de su trabajo como interesantes objetivos para un estudio posterior:

Cuestiones pendientes en el trabajo analizado

Los autores del trabajo hacen una buena recapitulación de posibles lineas de desarrollo ulterior:

1. La identificación más precisa de correlaciones estadísticas, probabilísticas y funcionales entre la calidad de datos y de procesos, poniendo el foco en la validación empírica de modelos y la extensión del análisis a un rango mayor de dimensiones y tipos específicos de procesos.
2. La validación de las metodologías, ya que algunas son propuestas sin experimentación a gran escala o con escasas o nula herramientas de apoyo que las hagan factibles.
3. La extensión de las guías metodológicas a un mayor conjunto de dimensiones (los autores proponen el rendimiento, la disponibilidad, la seguridad y la accesibilidad entre otras). Técnicas como la minería de datos pueden ser también de gran apoyo para este propósito.
4. En sistemas de información *web* y almacenes de datos, los datos se gestionan con distintos niveles de agregación. La calidad de su composición también debería ser verificada para obtener información de la calidad de los agregados a partir de las métricas de los datos elementales.

[4] *Web Information System* o *WIS*.
[5] Modelo de diseño de software que reparte las tareas entre proveedores de recursos (servidores) y los demandantes de los mismos (clientes).
[6] O *peer-to-peer*, *P2P*.

Methodology Acronym	Extended Name	Main Reference
TDQM	Total Data Quality Management	Wang 1998
DWQ	The Datawarehouse Quality Methodology	Jeusfeld et al. 1998
TIQM	Total Information Quality Management	English 1999
AIMQ	A methodology for information quality assessment	Lee et al. 2002
CIHI	Canadian Institute for Health Information methodology	Long and Seko 2005
DQA	Data Quality Assessment	Pipino et al. 2002
IQM	Information Quality Measurement	Eppler and Münzenmaier 2002
ISTAT	ISTAT methodology	Falorsi et al 2003
AMEQ	Activity-based Measuring and Evaluating of product information Quality (AMEQ) methodology	Su and Jin 2004
COLDQ	Loshin Methodology (Cost-effect Of Low Data Quality	Loshin 2004
DaQuinCIS	Data Quality in Cooperative Information Systems	Scannapieco et al. 2004
QAFD	Methodology for the Quality Assessment of Financial Data	De Amicis and Batini 2004
CDQ	Comprehensive methodology for Data Quality management	Batini and Scannapieco 2006

Figura 9.1: Metodologías consideradas
Tabla extraída directamente de [39]

9.2 Imputación de datos faltantes. Código Python

Exponemos un sencillo ejemplo de reconstrucción de valores faltantes. Para ello utilizamos la librería *sklearn.impute*, para aplicar la técnica *KNN*, o de vecinos más próximos, a unos datos originales, pertenecientes al clásico repositorio *digits* que previamente hemos degradado ofuscando un porcentaje dado de los componentes de cada individuo.

Aplicamos el método citado y exploramos distintos rangos de vecinos para averiguar cual es el que proporciona un menor error de clasificación.

En el código se obviado la inclusión del trazado del gráfico.

Código *Python*.

```
1
2       import numpy as np
3       import matplotlib.pyplot as plt
4       import scipy.stats as stats
5       from sklearn.impute import KNNImputer
6       from sklearn.metrics import mean_squared_error
7       from sklearn.neighbors import KNeighborsClassifier
8       from sklearn.model_selection import cross_val_score
9
10      np.random.seed(0)
11
12      # Datos completos
13      from sklearn.datasets import load_digits
14      digits = load_digits()
15      X_completo = digits.data[:]
16      y = digits.target
17
18      def ofuscar_componentes(vector, porcentaje=0.1):
19          vector_resultante = vector.copy()
20          num_reemplazar = int(len(vector) * porcentaje)
21          indices = np.arange(len(vector))
22          np.random.shuffle(indices)
23          indices_reemplazar = np.sort(indices[:num_reemplazar])
24          vector_resultante[indices_reemplazar] = np.nan
25          return vector_resultante
26
```

```
    def imputar_valores(valores, k=1):
        knn_imputer = KNNImputer(missing_values = np.nan, n_neighbors=k,
            weights='distance', metric='nan_euclidean')
        X_imputados = knn_imputer.fit_transform(valores)
        return X_imputados

    porcentajes = [5, 10, 15, 20]
    plt.figure(figsize=(10, 6))
    k_optimos = []

    for porcentaje in porcentajes:
        X_faltantes = np.apply_along_axis(ofuscar_componentes, axis=1, arr=
            X_completo, porcentaje=porcentaje/100)
        mse_list = []
        for K in range(1, 26):
            X_imputados = imputar_valores(X_faltantes, k=K)
            mse = mean_squared_error(X_completo, X_imputados)
            mse_list.append(mse)
        k_optimo = np.argmin(mse_list) + 1
        k_optimos.append(k_optimo)
```

9.3 Introducción de datos erróneos. Código Python

Aplicamos un pequeño preproceso de datos:

Para una regresión lineal (y predecir una variable en función de otras) eliminamos los registros con datos faltantes.

Convertimos la variable categórica en entera.

Aplicamos el misto tipo de alteración a un porcentaje determinado de cada uno de los registros de cada variable predictora, y trazamos la evolución del MSA en función del incremento de registros erróneos.

Aplicamos el método citado y exploramos distintos rangos de vecinos para averiguar cual es el que proporciona un menor error de clasificación.

En el código se obviado la inclusión del trazado del gráfico.

Código *Python*.

```

    import pandas as pd
    from sklearn.model_selection import train_test_split
    from sklearn.linear_model import LinearRegression
    from sklearn.metrics import mean_squared_error
    from sklearn.metrics import mean_absolute_error
    import numpy as np

    from sklearn.preprocessing import LabelEncoder

    # Cargar el conjunto de datos
    data = pd.read_csv("housing.csv")
    # data.info()

    # Sumarizamos datos faltantes
    nan_counts = data.isna().sum()
    # print(nan_counts)

    # Los suprimimos
    data = data.dropna()

    # Convertir la variable categorica 'ocean_proximity' a numerica
```

```
        # print(data["ocean_proximity"].value_counts())

        le = LabelEncoder()
        data['ocean_proximity'] = le.fit_transform(data['ocean_proximity'])

        # print(data.head())
        print(data.describe())

        data_errores = data
        np.random.seed(42)  # Fijamos la semilla para reproducibilidad

        MSE = []
        MAE = []
        percentages = range(0,21,1)
        factor_alteracion = 0

        variables_alteradas = ['median_income', 'total_bedrooms', 'total_rooms'
            , 'population', 'households', 'housing_median_age']

        import matplotlib.pyplot as plt

        # Inicializar listas para almacenar los errores de todas las variables
        MSEs = []
        MAEs = []

        for variable_alterada in variables_alteradas:
            MSE = []
            MAE = []

           for percentage in percentages:
                size = int(len(data_errores) * percentage / 100)
                data_errores = data.copy()

                indices_aleatorios = np.random.choice(data_errores.index, size,
                     replace=False)  # Seleccionamos indices aleatorios
                data_errores.loc[indices_aleatorios, variable_alterada] =
                    data_errores.loc[indices_aleatorios, variable_alterada] *
                    factor_alteracion  # Introducimos errores

                # Dividir el conjunto de datos
                X = data_errores.drop(columns=['median_house_value'])
                y = data_errores['median_house_value']
                X_train, X_test, y_train, y_test = train_test_split(X, y,
                    test_size=0.2, random_state=42)

                # Entrenar un modelo predictivo
                model = LinearRegression()
                model.fit(X_train, y_train)
                y_pred = model.predict(X_test)

                mse_with_errors = mean_squared_error(y_test, y_pred)
                mae_with_errors = mean_absolute_error(y_test, y_pred)

                MSE.append(mse_with_errors)
                MAE.append(mae_with_errors)

            # Almacenar los errores de la variable actual
            MSEs.append(MSE)
            MAEs.append(MAE)
```

9.4 Precisión y actualidad. Código R

Adjuntamos el código correspondiente a eata sección.

Código *R*.

```r

# Cargamos los datos
data <-read.csv("day.csv")

columns = c("cnt")
bikes = data.frame(matrix(nrow=731, ncol=length(columns)))
colnames(bikes) = columns

bikes$cnt = data$cnt # Variable a predecir

# Predictores...

bikes$workingday <- data$workingday
bikes$holiday <- data$holiday

# One-hot codificacion para la estacion
bikes$season_1 <- ifelse(data$season == 1, 1, 0)
bikes$season_2 <- ifelse(data$season == 2, 1, 0)
bikes$season_3 <- ifelse(data$season == 3, 1, 0)

# Extraccion de caracteristicas de niebla (MISTY) y lluvia (RAIN)
bikes$MISTY <- ifelse(data$weathersit == 2, 1, 0)
bikes$RAIN <- ifelse(data$weathersit == 3 | data$weathersit == 4, 1, 0)

# Denormalizar temperatura, humedad, y velocidad del viento
bikes$temp <- data$temp * 47 + -8
bikes$hum <- data$hum * 100
bikes$windspeed <- data$windspeed * 67

# Crear la caracteristica dias desde (days_since_2011)
bikes$days_since_2011 <- as.numeric(difftime(as.Date(data$dteday), as.Date(
    "2011-01-01"), units = "days"))

# Crear la caracteristica semanas desde (days_since_2011)
bikes$weeks_since_2011 <- trunc(as.numeric(difftime(as.Date(data$dteday),
    as.Date("2011-01-01"), units = "weeks")))

library(randomForest)
set.seed(123) # por reproductibilidad

random_forest_model_day <- randomForest(cnt ~ workingday + holiday + season
    _1 + season_2 + season_3 + MISTY + RAIN + temp + hum + windspeed + days
    _since_2011, data = bikes)

random_forest_model_week <- randomForest(cnt ~ workingday + holiday +
    season_1 + season_2 + season_3 + MISTY + RAIN + temp + hum + windspeed
    + weeks_since_2011, data = bikes)

# Crear 2 graficos de dependecias parciales con ejes personalizados

days_since_2011_pdp <- partial(random_forest_model_day, pred.var = "days_
    since_2011", plot = TRUE, plot.engine = "ggplot2") +
  ggtitle("Grafico de dependencia parcial") +
  xlab("Dias desde 2011") +
  ylab("Prevision de alquileres")

weeks_since_2011_pdp <- partial(random_forest_model_week, pred.var = "weeks
    _since_2011", plot = TRUE, plot.engine = "ggplot2") +
```

```
    ggtitle("Grafico de dependencia parcial") +
    xlab("Semanas desde 2011") +
    ylab("Prevision de alquileres")

# Realizar la validacion cruzada
cv_results_days <- rfcv(
  trainx = bikes[, !names(bikes) %in% c("cnt", "weeks_since_2011")],
  trainy = bikes$cnt,
  cv.fold = 10
)

# Imprimir los resultados
print(cv_results_days)

# Obtiene la importancia de las variables
importance_days <- importance(random_forest_model_day)

# Imprime la importancia de las variables
print(importance_days)

# Visualiza la importancia de las variables
varImpPlot(random_forest_model_day)

# Realizar la validacion cruzada
cv_results_weeks <- rfcv(
  trainx = bikes[, !names(bikes) %in% c("cnt", "days_since_2011")],
  trainy = bikes$cnt,
  cv.fold = 10
)

# Imprimir los resultados
print(cv_results_weeks)

# Obtiene la importancia de las variables
importance_weeks <- importance(random_forest_model_week)

# Imprime la importancia de las variables
print(importance_weeks)

# Visualiza la importancia de las variables
varImpPlot(random_forest_model_week)
```

9.5 Objetivos de Desarrollo Sostenible

Hemos revisado el contenido de los Objetivos de Desarrollo Sostenible propuestos por la Organización de las Naciones Unidas en la Agenda 2030 para averiguar el nivel de adecuación de este trabajo a los mismos.

En un documento independiente y adjunto al final de esta memoria incluimos unas reflexiones al respecto.

Bibliografía

[1] ISO [Internet]. [citado 27 de mayo de 2024]. ISO - Normas. Disponible en: https://www.iso.org/es/normas

[2] Normas ISO 8000 [Internet]. [citado 27 de mayo de 2024]. Disponible en: https://iso8000.es/normas-iso-8000

[3] Normas Técnicas para alcanzar la Calidad del Dato | datos.gob.es [Internet]. [citado 27 de mayo de 2024]. Disponible en: https://datos.gob.es/es/blog/normas-tecnicas-para-alcanzar-la-calidad-del-dato

[4] ISO 25012 [Internet]. [citado 27 de mayo de 2024]. Disponible en: https://iso25000.com/index.php/normas-iso-25000/iso-25012

[5] iso.org. ISO. [citado 27 de mayo de 2024]. ISO/IEC 25024:2015. Disponible en: https://www.iso.org/standard/35749.html

[6] Your Europe [Internet]. [citado 27 de mayo de 2024]. Protección de Datos conforme al reglamento RGPD. Disponible en: https://europa.eu/youreurope/business/dealing-with-customers/data-protection/data-protection-gdpr/index_es.htm

[7] The Digital Services Act package | Shaping Europe's digital future [Internet]. 2024 [citado 27 de mayo de 2024]. Disponible en: https://digital-strategy.ec.europa.eu/en/policies/digital-services-act-package

[8] Ley de Inteligencia Artificial de la UE | Avances y análisis actualizados de la Ley de Inteligencia Artificial de la UE [Internet]. [citado 27 de mayo de 2024]. Disponible en: https://artificialintelligenceact.eu/es/

[9] Artificial Intelligence: Adversarial Machine Learning | NCCoE [Internet]. [citado 28 de mayo de 2024]. Disponible en: https://www.nccoe.nist.gov/ai/adversarial-machine-learning

[10] EU AI Act Compliance Checker | Ley de Inteligencia Artificial de la UE [Internet]. [citado 27 de mayo de 2024]. Disponible en: https://artificialintelligenceact.eu/es/evaluacion/comprobador-del-cumplimiento-de-la-ley-de-ai-de-la-ue/

[11] La Presidencia española cierra el acuerdo del reglamento para crear una identidad digital europea única y segura [Internet]. [citado 27 de mayo de 2024]. Disponible en: https://www.lamoncloa.gob.es/serviciosdeprensa/notasprensa/asuntos-economicos/Paginas/2023/081123-reglamento-identidad-digital-europea.aspx

[12] Herrera Herrera, F. Transformación digital en la Unión Europea: eIDAS2 y el Reglamento de Datos | Legal | Cinco Días [Internet]. [citado 27 de mayo de 2024]. Disponible en: https://cincodias.elpais.com/cincodias/2024/01/10/legal/1704887264_047481.html

[13] Elahi E. Address standardization guide: What, why, and how? [Internet]. Data Ladder. 2022 [citado 27 de mayo de 2024]. Disponible en: https://dataladder.com/address-standardization-guide/

[14] Data profiling: qué es y cómo ayuda a mejorar la calidad de los datos [Internet]. [citado 27 de mayo de 2024]. Disponible en: https://blog.powerdata.es/el-valor-de-la-gestion-de-datos/data-profiling-que-es-y-como-ayuda-a-mejorar-la-calidad-de-los-datos

[15] ETL vs. ELT: What's the Difference? [Internet]. [citado 27 de mayo de 2024]. Disponible en: https://www.snaplogic.com/blog/etl-vs-elt-whats-the-difference

[16] Google Cloud [Internet]. [citado 27 de mayo de 2024]. ¿Qué es un data lake? | Google Cloud. Disponible en: https://cloud.google.com/learn/what-is-a-data-lake?hl=es-419

[17] Dynatrace [Internet]. [citado 28 de mayo de 2024]. Data Lakehouse. Disponible en: https://www.dynatrace.com/monitoring/platform/data-lakehouse/

[18] Boto3 1.34.113 documentation [Internet]. [citado 27 de mayo de 2024]. Disponible en: https://boto3.amazonaws.com/v1/documentation/api/latest/index.html

[19] pandas documentation — pandas 2.2.2 documentation [Internet]. [citado 27 de mayo de 2024]. Disponible en: https://pandas.pydata.org/docs/

[20] awswrangler: Pandas on AWS. [Internet]. [citado 27 de mayo de 2024]. Disponible en: https://aws-sdk-pandas.readthedocs.io/

[21] Usa algoritmos SageMaker integrados de Amazon o modelos previamente entrenados - Amazon SageMaker [Internet]. [citado 27 de mayo de 2024]. Disponible en: https://docs.aws.amazon.com/es_es/sagemaker/latest/dg/algos.html

[22] apache/arrow [Internet]. The Apache Software Foundation; 2024 [citado 27 de mayo de 2024]. Disponible en: https://github.com/apache/arrow

[23] Ciclo de vida de los datos: qué es y etapas | ESIC [Internet]. [citado 27 de mayo de 2024]. Disponible en: https://www.esic.edu/rethink/tecnologia/ciclo-vida-datos-c

[24] Stephen-Sumner. Introducción a la gobernanza - Cloud Adoption Framework [Internet]. 2024 [citado 27 de mayo de 2024]. Disponible en: https://learn.microsoft.com/es-es/azure/cloud-adoption-framework/govern/

[25] ¿Qué es la fusión de datos? | Herramientas de fusión de datos - Zoho DataPrep [Internet]. [citado 27 de mayo de 2024]. Disponible en: https://www.zoho.com/es-xl/dataprep/what-is-data-blending.html

[26] PAe - Normas Técnicas [Internet]. [citado 27 de mayo de 2024]. Disponible en: https://administracionelectronica.gob.es/pae_Home/pae_Estrategias/pae_Interoperabilidad_Inicio/pae_Normas_tecnicas_de_interoperabilidad.html

[27] Learn about the big changes in the 2024 Gartner Magic Quadrant for Data Quality Solutions | Ataccama [Internet]. [citado 27 de mayo de 2024]. Disponible en: https://www.ataccama.com/blog/what-s-new-in-the-gartner-magic-quadrant-for-data-quality-solutions-2024

[28] Gartner [Internet]. [citado 27 de mayo de 2024]. Gartner Magic Quadrant & Critical Capabilities. Disponible en: https://www.gartner.com/en/research/magic-quadrant

[29] The Definitive Guide to Jython — Definitive Guide to Jython latest documentation [Internet]. [citado 27 de mayo de 2024]. Disponible en: https://jython.readthedocs.io/en/latest/

[30] Aspectos básicos del EDI [Internet]. [citado 27 de mayo de 2024]. Normas de documentos EDI | Conceptos básicos de EDI. Disponible en: https://www.edibasics.com/es/recursos-edi/normas-documentales/

[31] IBM® InfoSphere Information Server for Data Quality: descripción general | [Internet]. [citado 27 de mayo de 2024]. Disponible en: https://www.ibm.com/es-es/products/infosphere-info-server-for-datamgmt

[32] Software de comparación de datos. Data Ladder. [citado 27 de mayo de 2024]. Disponible en: https://dataladder.com/es/software-de-comparacion-de-datos-calificado-como-el-mejor-de-su-clase-con-una-precision-de-coincidencia-del-96/

[33] Software de purga de fusiones | Utilizar reglas de supervivencia incorporadas y personalizadas [Internet]. Data Ladder. [citado 27 de mayo de 2024]. Disponible en: https://dataladder.com/es/software-de-purga-de-fusiones-utilizar-reglas-de-supervivencia-incorporadas-y-personalizadas/

[34] Aperture Data Studio | Customer Data Platform | Experian Business [Internet]. [citado 27 de mayo de 2024]. Disponible en: https://www.experian.co.uk/business/platforms/aperture-data-studio

[35] 11 steps to build an API-driven data quality framework [Internet]. [citado 27 de mayo de 2024]. Disponible en: https://atlan.com/api-driven-data-quality/

[36] Oliveira Rizzo D, Toledo Olivera N. Framework para la gestión de calidad de datos. 2022 [citado 27 de mayo de 2024]; Disponible en: http://repositorioslatinoamericanos.uchile.cl/handle/2250/4986627

[37] Kurniati A, Surendro K. Designing IQMM as a maturity model for information quality management. En Citeseer; 2010. p. 277-90. [citado 27 de mayo de 2024]. Disponible en: Disponible en: https://citeseerx.ist.psu.edu/document?repid=rep1&type=pdf&doi=4f60d3fb9a07a8e6380053ea53c9d6f5a29c2548

[38] Calabrese J, Esponda S, Pasini AC, Boracchia M, Pesado PM. Guía para evaluar calidad de datos basada en ISO/IEC 25012. En 2019 [citado 27 de mayo de 2024]. Disponible en: http://sedici.unlp.edu.ar/handle/10915/91086

[39] Batini C, Cappiello C, Francalanci C, Maurino A. Methodologies for data quality assessment and improvement. ACM computing surveys (CSUR). 2009;41(3):1-52. [citado 27 de mayo de 2024]. Disponible en: https://dl.acm.org/doi/abs/10.1145/1541880.1541883

[40] Troyanskaya O, Cantor M, Sherlock G, Brown P, Hastie T, Tibshirani R, et al. Missing value estimation methods for DNA microarrays. Bioinformatics. 2001;17(6):520-5. Disponible en: Disponible en: https://academic.oup.com/bioinformatics/article/17/6/520/272365?login=false

ANEXO

OBJETIVOS DE DESARROLLO SOSTENIBLE

Grado de relación del trabajo con los Objetivos de Desarrollo Sostenible (ODS).

Objetivos de Desarrollo Sostenible	**Alto**	**Medio**	**Bajo**	**No procede**
ODS 1. **Fin de la pobreza.**				X
ODS 2. **Hambre cero.**				X
ODS 3. **Salud y bienestar.**		X		
ODS 4. **Educación de calidad.**	X			
ODS 5. **Igualdad de género.**				X
ODS 6. **Agua limpia y saneamiento.**				X
ODS 7. **Energía asequible y no contaminante.**				X
ODS 8. **Trabajo decente y crecimiento económico.**		X		
ODS 9. **Industria, innovación e infraestructuras.**	X			
ODS 10. **Reducción de las desigualdades.**			X	
ODS 11. **Ciudades y comunidades sostenibles.**				X
ODS 12. **Producción y consumo responsables.**				X
ODS 13. **Acción por el clima.**				X
ODS 14. **Vida submarina.**				X
ODS 15. **Vida de ecosistemas terrestres.**				X
ODS 16. **Paz, justicia e instituciones sólidas.**	X			
ODS 17. **Alianzas para lograr objetivos.**				X

Hemos revisado el contenido de los Objetivos de Desarrollo Sostenible propuestos por la Organización de las Naciones Unidas en la Agenda 2030 para averiguar si hay algún nivel de adecuación de este trabajo a los mismos.

El objetivo 9º, "Industria, innovación e infraestructuras" ha sido seleccionado en primer lugar porque algunas de las normas de calidad expuestas y revisadas (ISO/IEC 2524:2015) [1] mencionan expresamente eso.

A lo largo de su desarrollo, hemos incidido en cómo unos datos imprecisos e inexactos pueden alterar nuestra percepción de la realidad y hacernos tomar decisiones —o partido por causas erróneas— en base a información que podría estar sesgada para que se amolde al mensaje a transmitir.

Este hilo conductor podría entroncar en cierta manera con algunos de los objetivos propuestos:

El 4º, "Educación de calidad" ha sido seleccionado porque creemos que la mejora en la calidad de los datos puede también contribuir a su consecución. La elección del 16º, "Paz, justicia e instituciones sólidas" es que estas podrían estar avaladas por la veracidad exigible a datos e información de calidad.

En menor medida, nos ha parecido también oportuna la posibilidad de mencionar otros objetivos. Aunque no están directamente vinculados con nuestro trabajo, se verían propiciados si se verificase el cumplimiento de los dos objetivos anteriormente expuestos:

3º "Salud y bienestar", 8º, "Trabajo decente y crecimiento económico" y finalmente el 10º, "Reducción de las desigualdades".

La salud, el bienestar y el trabajo decente son el fruto de sociedades donde el establecimiento de tanto la justicia como la paz favorezcan su desarrollo.

Empecemos también a entender crecimiento como la mejora de la eficiencia económica en vez de como una sucesión creciente de números. Por primera vez, nuestra demografía global deja de ser ascendente: podríamos considerar aprovechar esta inminente nueva circunstancia para plantearnos una mejor distribución de los recursos y conseguir reducir desigualdades, desequilibrios u ofrecer mejores oportunidades:

Los retos son enormes, pero nuestra determinación y voluntad pueden serlo también.

[1] https://www.iso.org/standard/35749.html.

ETS Enginyeria Informàtica
Camí de Vera, s/n, 46022, València
T +34 963 877 210
F +34 963 877 219
etsinf@upvnet.upv.es - www.inf.upv.es