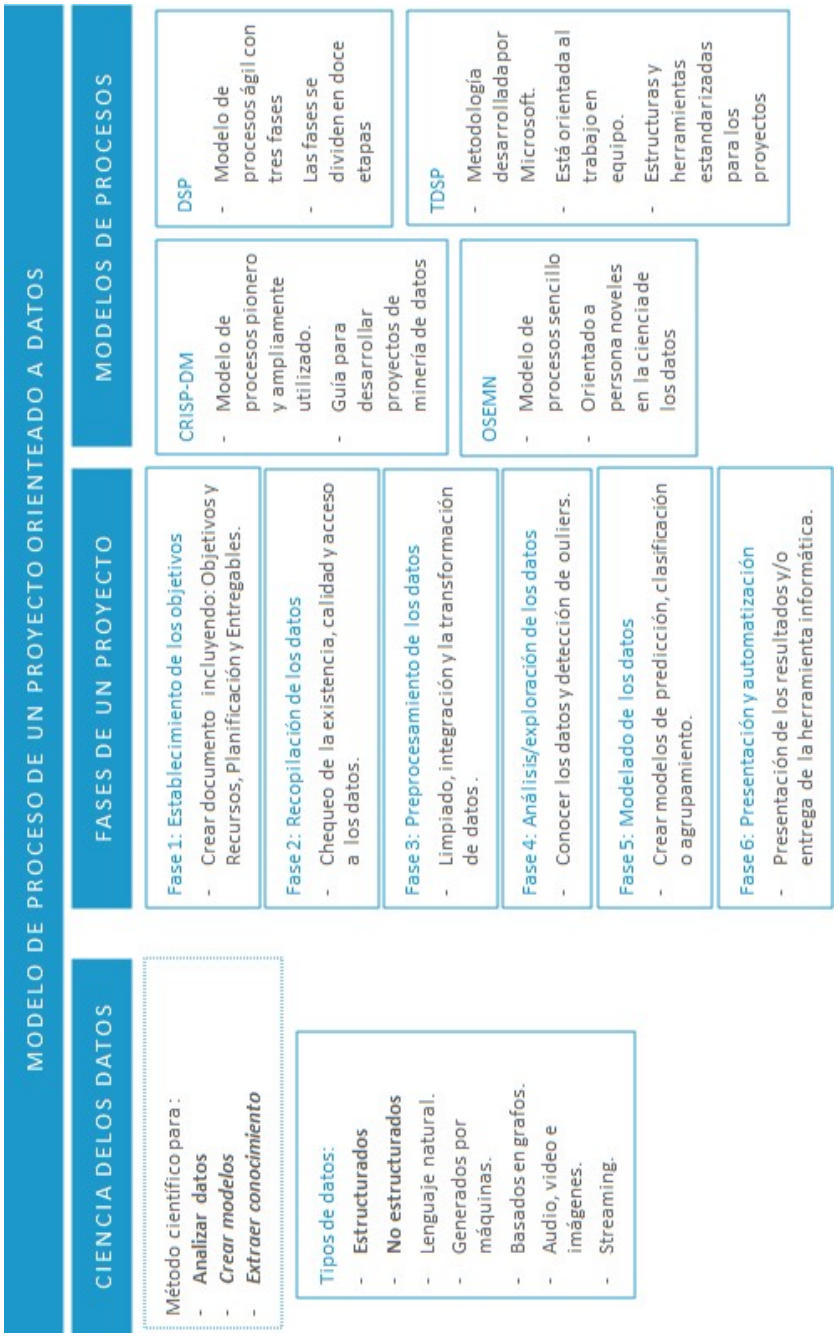
Análisis de Datos Masivos para el Negocio

Tema 1. Modelo de proceso de un proyecto orientado a datos

Índice

Esquema



1.1. Introducción y objetivos

El objetivo básico de este tema es introducir al alumno en la denominada ciencia de los datos. Se proporciona una definición de ciencia de los datos, se presentan qué tipo de datos se puede encontrar hoy en día en el entorno empresarial y se dan ejemplos de aplicación de la ciencia de los datos. Así mismo se presenta un conjunto de modelos de procesos o metodologías para el desarrollo de proyectos orientados a datos. Se han seleccionado modelos contrastados y muy utilizados y modelos propuestos por empresas orientados a mejorar la productividad de los equipos de trabajo.

Objetivos que se pretenden conseguir:

- Definir ciencia de los datos.
- Conocer los tipos de datos presentes en las empresas.
- Conocer qué es y para qué sirve un proyecto orientado a datos.
- Conocer las fases de un proyecto en la ciencia de los datos.
- Conocer los modelos de procesos más utilizados en los proyectos.

1.2. La ciencia de los datos

La ciencia de los datos es un término que engloba al conjunto de métodos científicos utilizados para *analizar* grandes cantidades de datos, crear *modelos* y *extraer conocimiento* contenidos en esos datos. Ese conocimiento se puede utilizar para la toma de decisiones. El término *big data* se asocia a cualquier conjunto de datos tan grande o complejo que dificulta su procesamiento mediante técnicas clásicas de gestión de datos.

Por tanto, se puede considerar que el *big data* es la materia prima a la que se le aplica la ciencia de los datos para obtener un producto elaborado. **La ciencia de los datos** es una evolución de la estadística y de los métodos tradicionales de gestión de datos (por ejemplo, las bases de datos relaciones y los lenguajes de consulta), incluso incluye técnicas de ambas, pero puede considerarse una nueva disciplina pues incluye nuevas técnicas (por ejemplo, aprendizaje automático y bases de datos no relaciones).

Al *big data* se le puede caracterizar con las denominadas tres V, esto es:

- Volumen. ¿Cuántos datos hay?
- *Variedad. ¿Cómo de diversos son los tipos de datos disponibles?*
- *Velocidad. ¿A qué velocidad se generan los nuevos datos?*

También se puede añadir una cuarta V que es la Veracidad, esto es, ¿cómo de preciso son los datos disponibles? Estas **cuatro propiedades son las que hacen diferente al *big data*** de los datos utilizados en los sistemas tradicionales de gestión de datos. Estas características incluyen retos en algunos de los siguientes aspectos: recopilación, almacenamiento, búsqueda, compartición, transferencia o visualización de datos. Además, *big data* necesita de técnicas especializadas para extraer el conocimiento incluido en los datos.

Las principales características que diferencian a un científico de datos de estadístico clásico son la capacidad de trabajar con *big data*, técnicas de aprendizaje automático, computación y programación de algoritmos.

Ejemplos de utilización de la ciencia de los datos y *big data*

Los ejemplos de utilización de las técnicas de la ciencia de los datos son amplios, tanto en empresas comerciales como en otro tipo de organizaciones.

Las empresas comerciales e industriales utilizan la ciencia de los datos y *big data* para mejorar el conocimiento acerca de sus clientes, procesos, empleados y productos. Dichas compañías utilizan la ciencia de los datos para ofrecer a los clientes una mejor experiencia, realizar ventas cruzadas (*cross-selling*), ventas mejoradas (*up-selling*) y en definitiva personalizar sus ofertas. Por ejemplo, Google AdSense acumula datos de los usuarios de internet, de forma que se puede personalizar los mensajes comerciales a la persona que navega por internet. MaxPoint es otro ejemplo de publicidad personalizada en tiempo real. Los profesionales de recursos humanos utilizan análisis de personas (*people analytics*) y minería de texto (*text mining*) para buscar candidatos, monitorizar el estado de los empleados y estudiar las conexiones informales entre los compañeros de trabajo. Las instituciones financieras utilizan la ciencia de los datos para predecir el stock de los mercados, determinar el riego en inversiones o aprender cómo atraer nuevos clientes a sus servicios.

En las organizaciones gubernamentales, un científico de datos puede trabajar en proyectos de detección de fraude u otro tipo de actividad delictiva o la optimizaciónde procesos. Organizaciones no gubernamentales también utilizan la ciencia de los datos para incrementar los resultados de sus campañas. Las universidades también la utilizan para mejorar la experiencia con sus estudiantes.

Categoría de los datos

A nivel generallas principales categorías de datos son:

- Estructuradas.
- No estructuradas.
- Lenguaje natural.
- Generados por máquinas.
- Basados en grafos.
- Audio, vídeo e imágenes.
- *Streaming*.

Datos estructurados

Un dato estructurado es un dato que **reside en un campo fijo que está etiquetado.** Los datos estructurados tienen exactamente fijadas su longitud, formato y tamaño. Los datos estructurados han permitido tradicionalmente que los computadores gestionaran y procesaran grandes cantidades de datos. Sin embargo, en el mundo real los datos no aparecen de forma estructurada, ese orden se crea con el fin de facilitar su procesado en máquinas. Además, los datos estructurados permiten responder fácilmente a preguntas del tipo: ¿cuántos productos se han vendido? ¿Qué clientes han realizado pedidos?, etc.

La forma más sencilla de almacenar datos estructurados es una tabla (ver figura 1) como las que se utilizan en una hoja de cálculo o bien en una base de datos relacional. Una tabla de datos se compone de filas y columnas. Las columnas están etiquetadas indicando el tipo de datos que incluyen. Las filas contienen los datos ordenados según indican las etiquetas de las columnas.

	Indicator ID	Dimension List	Timeframe	Numeric Value	Missing Value Flag	Confidence Inte
2	214390830	Total (Age-adjusted)	2008	74.6%		73.8%
3	214390833	Aged 18-44 years	2008	59.4%		58.0%
4	214390831	Aged 18-24 years	2008	37.4%		34.6%
5	214390832	Aged 25-44 years	2008	66.9%		65.5%
6	214390836	Aged 45-64 years	2008	88.6%		87.7%
7	214390834	Aged 45-54 years	2008	86.3%		85.1%
8	214390835	Aged 55-64 years	2008	91.5%		90.4%
9	214390840	Aged 65 years and over	2008	94.6%		93.8%
10	214390837	Aged 65-74 years	2008	93.6%		92.4%
11	214390838	Aged 75-84 years	2008	95.6%		94.4%
12	214390839	Aged 85 years and over	2008	96.0%		94.0%
13	214390841	Male (Age-adjusted)	2008	72.2%		71.1%
14	214390842	Female (Age-adjusted)	2008	76.8%		75.9%
15	214390843	White only (Age-adjusted)	2008	73.8%		72.9%
16	214390844	Black or African American only (Age-adjusted)	2008	77.0%		75.0%
17	214390845	American Indian or Alaska Native only (Age-adjusted)	2008	66.5%		57.1%
18	214390846	Asian only (Age-adjusted)	2008	80.5%		77.7%
19	214390847	Native Hawaiian or Other Pacific Islander only (Age-adjusted)	2008	DSU		
20	214390848	2 or more races (Age-adjusted)	2008	75.6%		69.6%

Figura 1. Ejemplo de datos estructurados. Fuente: Cielen et al., 2016.

Datos no estructurados

Son datos cuyo contenido es dependiente del contexto o bien variable. No tienen un formato específico. Un ejemplo podría ser un simple correo electrónico (ver figura 2). El correo electrónico contiene información pero no estructurada en campos reconocibles. Aunque un correo electrónico tiene elementos estructurados como el remitente, título y cuerpo del correo, es difícil, por ejemplo, calcular el número de correo que se declaran descontentos de cierto servicio pues hay muchas formas de expresar dichas quejas. Por tanto, una característica importante de los datos no estructurados es **la dificultad que presentan en su procesamiento.** El mundo está lleno de datos no estructurados (ver figura 3).

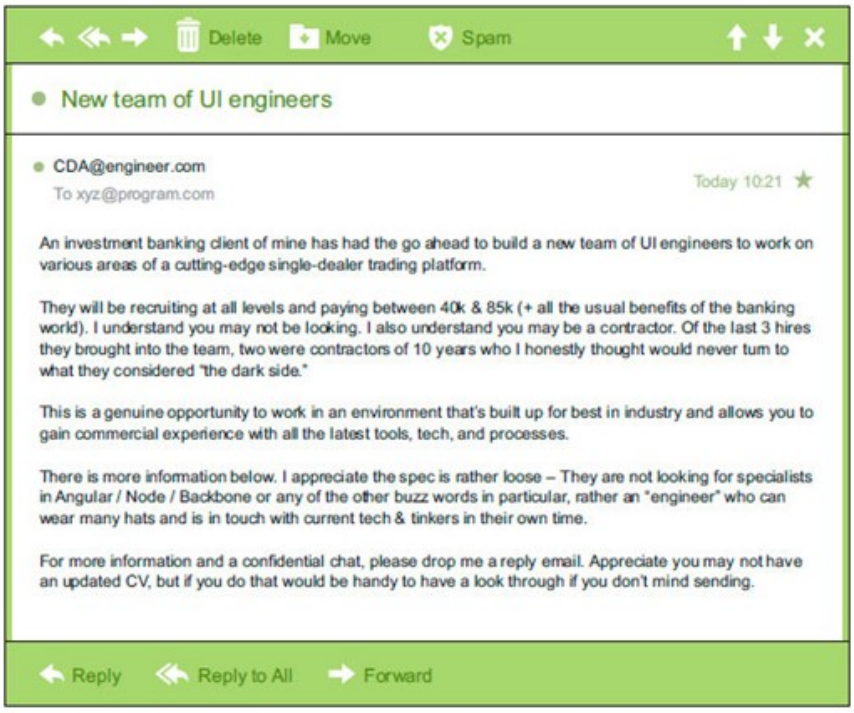


Figura 2. Ejemplo de datos no estructurados. Fuente: Cielen et al., 2016.

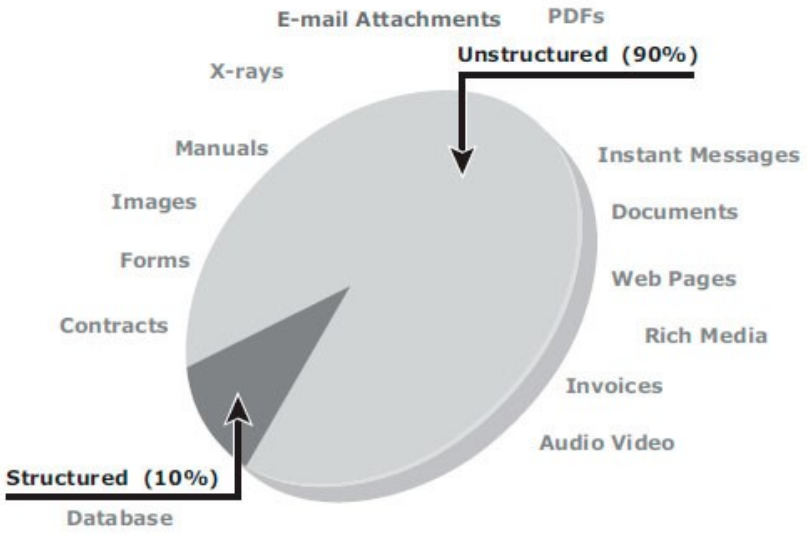


Figura 3. Reparto porcentual entre datos estructurados y no estructurados Fuente: Gnanasundaramy Shrivastava, 2012.

Lenguaje natural

Representan un tipo de dato no estructurado y su procesamiento es todo un reto ya que requiere conocimiento específico de técnicas de la ciencia de los datos. El lenguaje natural **está presente en correos electrónicos, página web, redes sociales, documentos electrónicos, audios, vídeos, etc.** Las técnicas de procesamiento del lenguaje natural han tenido cierto éxito en el reconocimiento de entidades (personas, empresas, organizaciones, etc.) presentes en un texto, reconocimientos de los temas tratados en un texto, resumen de un texto, resumen de textos y análisis de sentimientos. Sin embargo, actualmente dichas técnicas no generalizan bien a todos los contextos y no son capaces de descifrar el significado de todas las partes del texto.

Datos generados por máquinas

Los datos generados por máquinas es información automáticamente creada por un computador, un proceso, una aplicación o cualquier máquina sin la intervención del ser humano. Esta forma de generación de datos es la que está generando el mayor volumen de ellos. La conexión de sensores en red, maquinaria compleja y software de adquisición de datos, referido como el *internet de las cosas*, supondrá que en un futuro cercano el número de dispositivos conectados será mucho mayor que el de personas.El análisis de ese tipo de datos requiere herramientas altamente escalables, esto es, que se adapten bien a un gran crecimiento en el volumen y velocidad de los datos. Ejemplos típicos de sistemas que crean este tipo de datos son **los ficheros de registro de los servidores webs (ficheros log) y de las incidencias en redes, la telemetría, los eventos de los centros de llamadas, etc.**

Datos basados en grafos o redes

Un grafo o red es una estructura matemática que modela relaciones entre objetos. Son el núcleo de la teoría de grafos o redes de las matemáticas y se estudian

mediante el análisis de grafos o redes. **Los componentes de un grafo son los nodos** que modelan los objetos y aristas que modelan las relaciones entre objetos. Los datos basados en grafos representan de forma natural las redes sociales y su estructura permite calcular métricas específicas tales como la influencia que ejerce cierta persona o el camino más corto entre dos personas. Como ejemplo para entender un grafo, se podría considerar a las personas como nodos y las aristas corresponderían a relaciones de amistad o familiar entre dichas personas (ver figura 4).

También podríamos diseñar otro grafo donde las aristas fuesen relaciones de negocio o trabajo y un tercer grafo donde los nodos fuesen personas y películas y las aristas el interés por las películas de las personas. Realizando un estudio de esos tres grafos se podrían obtener interesantes respuestas a diversas preguntas. Esos grafos se pueden construir utilizando la información de diversas aplicaciones de redes sociales o servicios por *streaming*. La información gráfica que proporciona un grafo se puede codificar fácilmente en una tabla y ser por tanto almacenada.Existen bases de datos especializadas en almacenar grafos que incluyen un lenguaje de interrogación especializado como puede ser SPARQL.

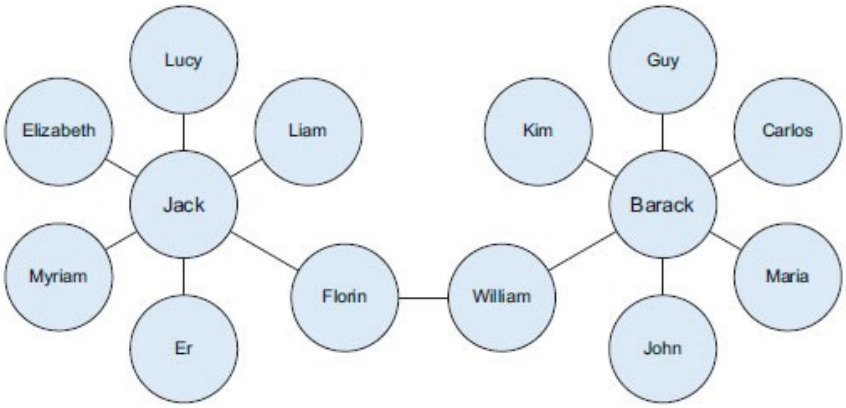


Figura 4. Ejemplo de grafo o red. Fuente: Cielen et al., 2016.

Audio, imagen y vídeo

Uno de los grandes retos para un científico de datos es el procesado de datos procedentes de audio, imágenes o vídeo. Tareas que son sencillas para los humanos, tales como, reconocimiento de objetos en fotografías son un gran reto para los computadores. La acumulación de datos basados en audios, vídeos o imágenes por las empresas es ingente y su gestión y aprovechamiento una tarea compleja y, sin embargo, necesaria para el futuro.

Un apartado especial requieren los datos en *streaming*. En este caso los datos fluyen al sistema cuando el evento está ocurriendo. No es información almacenada previamente. Ejemplos de este tipo de procesado son averiguar las tendencias actuales en redes sociales, deporte en directo o eventos musicales, o bien, averiguar stocks de mercado.

1.3. Fases de un proyecto orientado a datos

En la sección precedente se comentaron ejemplos prácticos para los cuales compañías utilizan la ciencia de los datos. En el fondo es la creación de un sistema informático que permita a las empresas,por ejemplo,ofrecer a los clientes una mejor experiencia, realizar ventas cruzadas (*cross-selling*), ventas mejoradas (*up-selling*) o en definitiva personalizar sus ofertas. Ese sistema informático debe ser creado a partir de lo que se denomina **un proyecto orientado a datos.** Esta sección presenta las fases por la que transcurre un proyecto empresarial orientado a datos (ver figura 5). Se centra en los aspectos tecnológicos generales y no en la lógica empresarial concreta. Un proyecto de este tipo comienza estableciendo los objetivos a conseguir, se debe realizar una recopilación, preparación y análisis de los datos a procesar. A continuación, se creará un modelo basado en los datos que consiga alcanzar los objetivos fijados al principio del proyecto y finalmente se realizará una automatización del proceso.

A continuación, se presenta una pequeña explicación de cada una de las fases de un proyecto orientado a datos. Notar que el desarrollo extenso y detallado de dichas fases se realizará a lo largo de los temas que componen esta asignatura.

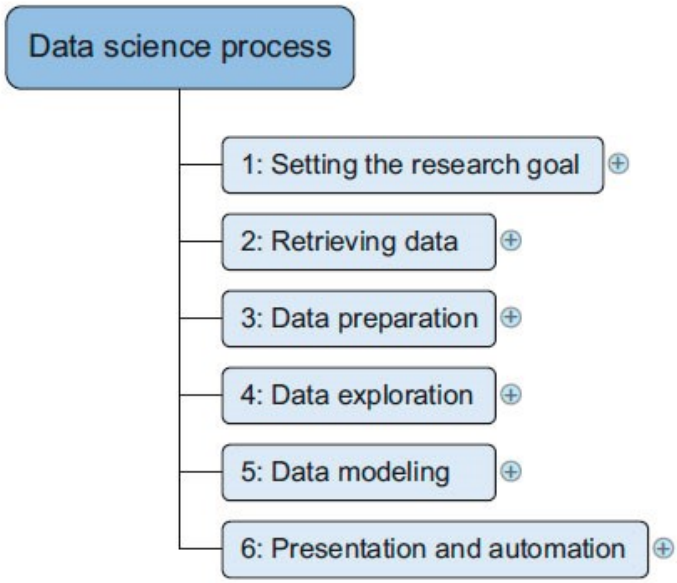


Figura 5. Fases de un proyecto orientado a datos. Fuente: (Cielen y otros, 2016)

Fase 1: Establecimiento de los objetivos

Un proyecto orientado a datos se desarrolla generalmente en el contexto de una organización o empresa. El primer paso es la creación de un *project charter* o documento inicial. Este documento contiene información acerca de cuál es el objeto del proyecto o estudio, cómo la organización se beneficiará del estudio, qué datos y recursos son necesarios, una planificación temporal y qué conjunto de entregables se incluyen. Nótese que el proyecto puede ser un estudio de consultoría en la que los entregables serán informes o bien puede ser un proyecto de ejecución en los que algunos entregables serán las aplicaciones informáticas desarrolladas.

Fase 2: Recopilación de los datos

El segundo paso en el proyecto es la recolección de los datos necesarios. En el documento o memoria inicial del proyecto se debe haber establecido qué datos son necesarios y dónde se pueden encontrar. En este paso se debe asegurar que se pueden utilizar los datos, lo que significa el chequeo de la existencia, calidad y acceso a dichos datos. Los datos también pueden ser proporcionados por terceras

compañías y pueden tomar diferentes formatos como archivos de texto, hojas de cálculo, conjuntos de bases de datos, etc.

Fase 3: Preparación/preprocesamiento de los datos

La fase 2 proporciona un conjunto de datos los cuales pueden contener ciertos errores. La fase 3 del proyecto está dedicada a mejorar la calidad de los datos y prepararlos para su uso en las siguientes fases. Se pueden distinguir tres pasos en esta fase, el *limpiado de datos* que elimina valores falsos o inconsistentes de la fuente de datos, la *integración de datos* que enriquece la información disponible combinando información de varias fuentes y la *transformación de datos* que asegura que los datos están en un formato adecuado para su posterior utilización. Las fases 2 y 3 se desarrollan en profundidad en el tema «Técnicas estadísticas de análisis de datos» de esta asignatura.

Fase 4: Análisis/exploración de los datos

Esta fase se centra en alcanzar un conocimiento profundo de los datos. Se intenta entender cómo las distintas variables interactúan unas con otras, las distribuciones de los datos y si existen datos extraños también denominados *outliers*. Para conseguir esos objetivos se utilizan técnicas de la estadística descriptiva, técnicas visuales y modelado de datos simple (por ejemplo, técnica de regresión lineal). Esta fase se desarrolla en profundidad en el tema «Técnicas estadísticas de análisis de datos» y en el tema «Series temporales» de esta asignatura.

Fase 5: Modelado de los datos

En esta fase se utilizan técnicas de la estadística, aprendizaje automático o de la investigación operativa para construir modelos en el dominio del conocimiento que busquen en los datos preparados previamente las respuestas a las preguntas formuladas como objetivos del proyecto. Estas respuestas pueden incluir entre otras, predicciones sobre ventas, almacenaje, beneficios, etc., clasificaciones de clientes, proveedores, activos, etc., o agrupamiento de productos, clientes o empresas. La

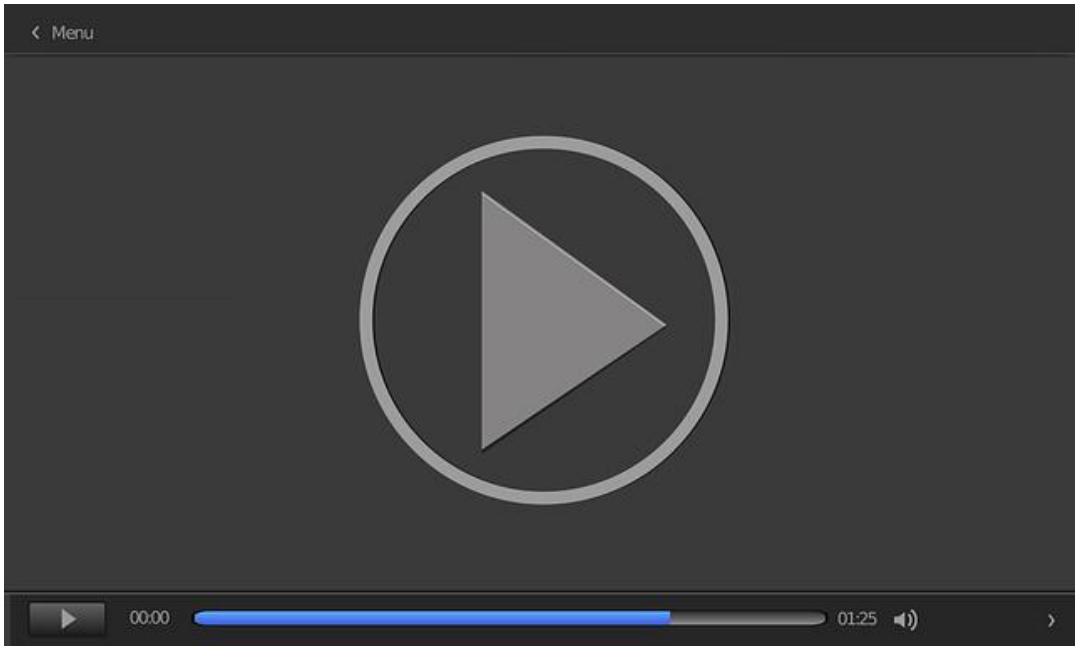
construcción de un modelo es un proceso iterativo que incluye la selección de variables y la ejecución y valoración del modelo. Las técnicas usadas en esta fase se desarrollan en profundidad en algunos temas de esta asignatura.

Fase 6: Presentación y automatización

La última fase del proyecto es la presentación de los resultados. Los resultados se pueden mostrar mediante informes o bien a través de presentaciones audiovisuales. En los proyectos de ejecución la fase final también incluye la entrega de la aplicación informática que automatiza el proceso de generación de los resultados. Esta fase se desarrolla en profundidad en la asignatura de Visualización Avanzada de Datos.

1.4. Modelos de proceso de un proyecto orientado a datos

En el vídeo *Modelo de proceso de un proyecto orientado a datos* se muestran algunos de los modelos de proceso más usados a la hora de desarrollar proyectos orientados a datos. Estos modelos de proceso se basan en las fases presentadas en la sección 1.3, aunque cada una de ellas presenta sus pequeñas peculiaridades.



Accede al vídeo:

https://unir.cloud.panopto.eu/Panopto/Pages/Embed.aspx?id=aaf7f602-03e3-4843-ac24-b15d00aa3443

Es importante en este punto aclarar cierta confusión que existe entre varias terminologías. A veces es complicado diferenciar entre un modelo de proceso, una metodología y un ciclo de vida. Un **modelo de proceso** es el conjunto de tareas a realizar para desarrollar un determinado proyecto, de elementos que producen dichas tareas (salidas) y de los elementos necesarios para realizar las tareas

(entradas). El objetivo del modelo de proceso es hacer el proceso repetible, manejable y medible. Una **metodología** se puede definir como una instancia de un modelo de procesos donde se definen las tareas, entradas y salidas y se especifican cómo realizar dichas tareas indicando técnicas y posibles herramientas a utilizar.

Por tanto, una metodología se puede ver cómo una particularización de un modelo de proceso. Finalmente, un modelo de **ciclo de vida** determina en qué orden se deben realizar las distintas actividades. Un modelo de ciclo de vida es la descripción de las diferentes maneras de desarrollar un proyecto.

CRoss-Industry Standard Process for Data Mining (CRISP-DM)

CRISP-DM, ver (Chapman y otros,2000) establece qué tareas se deben realizar para finalizar exitosamente un proyecto orientado a datos. Por tanto, es un modelo de proceso que incluye un modelo de ciclo de vida. CRISP-DM también incluye ciertas recomendaciones sobre cómo realizar ciertas tareas por lo que también contiene cierto componente metodológico. Sin embargo, como estas recomendaciones se limitan a proponer otras tareas y no se proporciona una guía clara de cómo realizarlas, CRISP-DM es principalmente un modelo de proceso.

CRISP-DM es una guía para desarrollar proyectos de minería de datos (Data Mining, DM) y descubrimiento de conocimiento (Knowledge Discovery, KD) desarrollada por un conjunto de empresas (Teradata, SPSS–ISL, Daimler-Chrysler and OHRA) en los años 90 del siglo pasado. Dicha guía se confeccionó tras estudiar el conjunto de problemas con que se habían encontrado proyectos de DM & KD. Los conceptos DM y KD podemos relacionarlos directamente con lo que ahora se denomina ciencia de los datos. CRISP-DM es independiente de proveedores por lo que se puede utilizar con las herramientas y aplicar a los problemas que el personal del proyecto considere. CRISP-DM define las distintas fases por las que debe transitar un proyecto de DM y también define qué entregables se asocian a cada tarea. Las seis **fases en las que CRISP-DM divide un proyecto** son (ver figura 6):

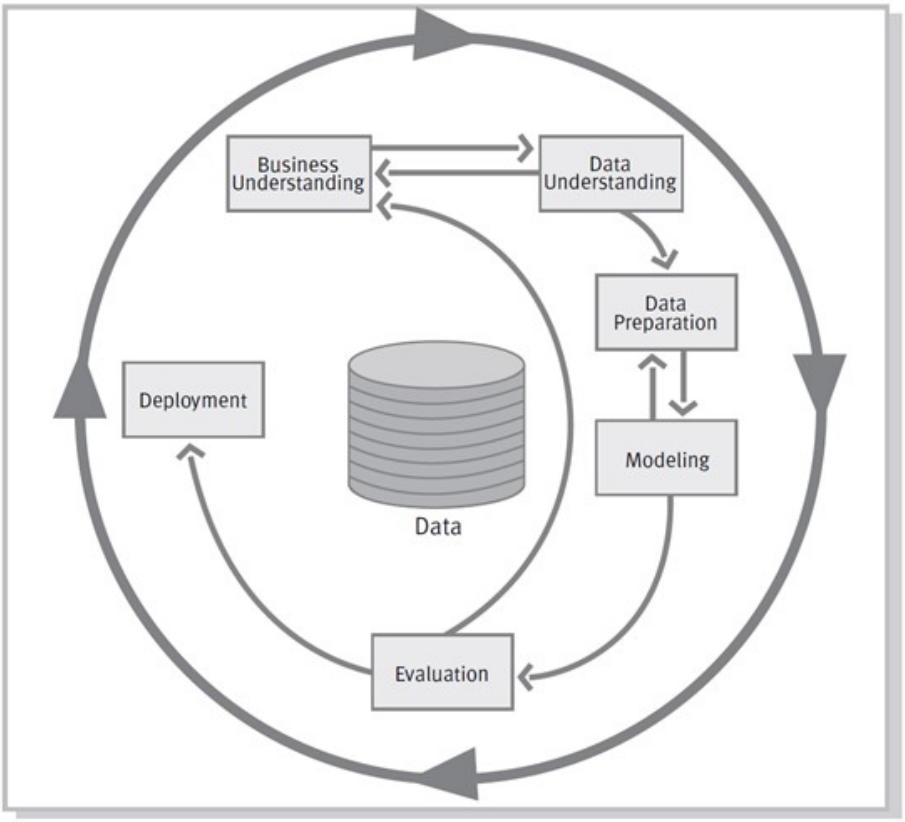


Figura 6. Fases del modelo CRISP-DM. Fuente: (Chapman y otros 2000)

Comprensión del negocio. Esta fase se centra en entender los objetivos del proyecto y los requerimientos desde una perspectiva empresarial y convertir ese objetivo empresarial en la definición de un problema de DM y un plan preliminar para alcanzar dichos objetivos.

Comprensión de los datos. Esta fase comienza con una recopilación inicial de datos y continúa actividades dirigidas a familiarizarse con los datos, identificar problemas de calidad en los datos, descubrir conocimiento oculto en los datos o detectar subconjuntos interesantes para formar hipótesis sobre información oculta.

Preparación de los datos. Esta fase cubre todas las actividades requeridas para construir el conjunto de datos a utilizar por el resto de las fases partiendo del conjunto inicial de datos.

Modelado. En esta fase, se seleccionan y aplican varias técnicas de modelado. También se procede a calibrar los parámetros asociados a cada técnica usada. Es típico que existan varias técnicas aplicables al mismo tipo de problema DM. Algunas técnicas tienen ciertos requerimientos en el formato de los datos, por lo que si es necesario se puede volver a la fase de preparación para adecuarlos.

Evaluación. Esta fase evalúa el modelado cuidadosamente y se revisa los pasos dados para comprobar que efectivamente se alcanzan de forma adecuada los objetivos empresariales buscados. Al final de esta fase se debe tomar una decisión acerca de cómo usar los resultados obtenidos.

Despliegue. La construcción del modelo no es generalmente el fin del proyecto. Incluso si el propósito del modelo es incrementar el conocimiento sobre los datos, ese conocimiento adquirido debe ser organizado y presentado de forma que el cliente pueda usarlo.

En la figura 7 se pueden observar tareas genéricas (negrita) y salidas (itálicas) del modelo de referencia CRISP-DM.

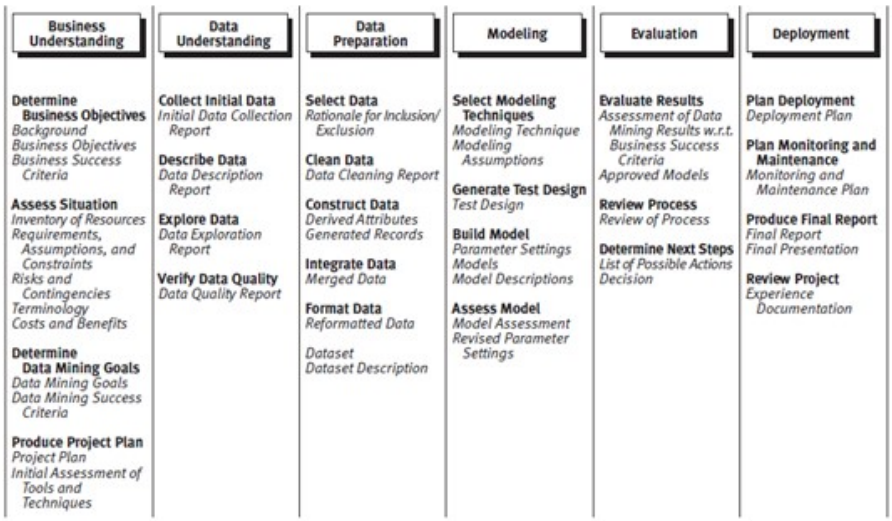


Figura 7. Tareas y salidas del modelo CRISP-DM. Fuente: (Chapman y otros 2000).

Data Science Process

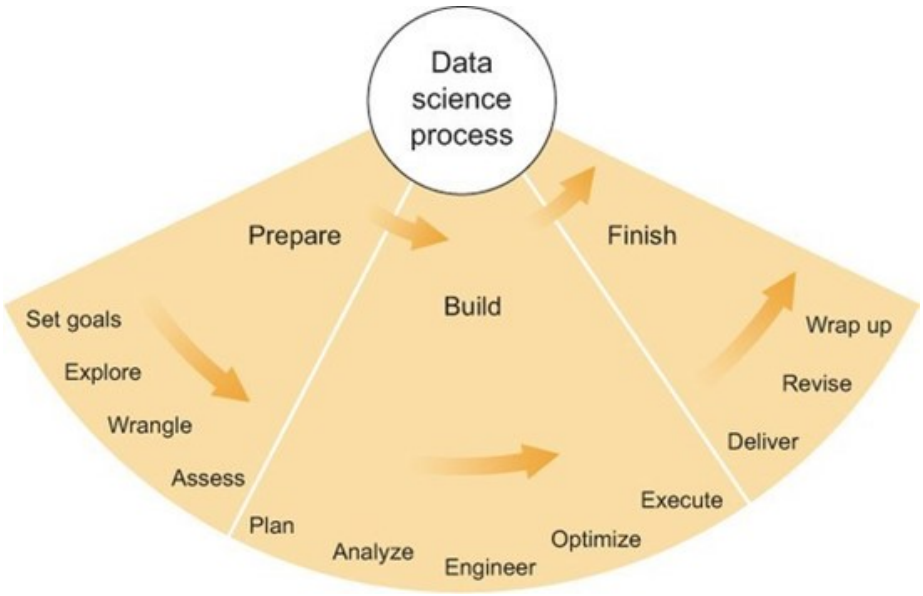


Figura 8. Fases del modelo DSP. Fuente: (Godsey B., 2017).

En la referencia (Godsey B., 2017) el autor propone un modelo de proceso para la ciencia de los datos. Incluye, además, ciertos toques metodológicos al hablar de qué herramientas pueden ser útiles en las diversas tareas. La propuesta divide un **proyecto en tres fases** (ver figura 8):

- **Preparación.** Fase para establecer los objetivos del proyecto y recopilar los datos.
- **Construcción.** Esta fase trata de la construcción del producto desde su planificación hasta su ejecución, utilizando todos los conocimientos adquiridos en la fase anterior y todas las herramientas estadísticas y software necesarios.
- **Finalización.** Esta fase incluye el despliegue del producto, realimentación por parte de los clientes, realización de posibles revisiones, soporte del producto y acabado del proyecto.

Cada una de las fases anteriores se dividen en tareas, a continuación, se proporciona una breve descripción de cada tarea.

Fase de preparación

Establecimiento de los objetivos. Todo proyecto en la ciencia de los datos tiene un cliente con conciertas expectativas. Es preciso realizar un conjunto de buenas preguntas al cliente, disponer de datos relevantes y realizar un análisis en profundidad para establecer los objetivos del proyecto.

*Exploración de datos.*Acceso a los datos disponibles. Se puede acceder a los datos a través de un fichero (texto estructurado csv o tsv, HTML, XML, JSON), de una base de datos (relacional o no relacional) o bien detrás de una API (application programming interface).

«*Doma*» *de datos.* El proceso de tomar datos e información en formatos no estructurados o complicados o de diversas fuentes y convertirlos a un formato convencional y usable. No existe un conjunto de pasos predefinidos para realizar esta tarea, cada caso es diferente y sí existen un conjunto de herramientas que se pueden utilizar. Esta fase también incluye la denominada *limpieza de datos.*

Evaluación de los datos. Esta fase consiste en el análisis de los datos mediante técnicas de estadística descriptiva. Se utiliza para detectar posibles datos extraños (*outliers*), derivas (bias), falta de precisión, especificidad u otros aspectos de los datos que puedan tener relevancia.

Fase de construcción

Desarrollo del plan. Una vez adquirido un gran conocimiento sobre el proyecto en la fase anterior, esta tarea se dedica al desarrollo de una planificación del resto del proyecto.

Análisis de los datos. Esta tarea es el equivalente al modelado de los datos en otros procesos. Se utilizarán técnicas estadísticas o de aprendizaje automático para obtener modelos de datos. El objetivo es analizar los datos y obtener conclusiones.

Ingeniería del producto. Esta tarea es la encargada de construir el software producto del proyecto. Se puede hacer uso de herramientas software disponibles como hojas de cálculo, SPSS, Stata, SAS, Minitab o MATLAB. Algunas de esas herramientas disponen de un lenguaje de programación propio. Otra opción es utilizar directamente de lenguajes de programación como R, Python, JAVA, etc., que incluyen librerías con utilidades. A la hora de elegir dichas herramientas se debe tener en cuenta los métodos implementados que incluyen, la flexibilidad, información y documentación disponible, compatibilidad con otras herramientas, y las licencias de uso.

Optimización. Esta tarea consiste en estudiar la optimización del producto a obtener con software complementario de almacenamiento, gestión y mejora de la capacidad de computación.

Plan de ejecución. El último paso o tarea de esta fase es la propia construcción del producto, o sea la ejecución del plan desarrollado en la primera tarea de esta fase.

Fase de finalización

Entrega del producto. Generalmente la entrega del proyecto consiste en un informe final, la entrega de una herramienta de análisis utilizable por el cliente, o bien la entrega de una aplicación software que realice más funciones.

Realización de revisiones. Una vez que el cliente ha revisado o empezado a usar el producto es posible obtener realimentación para realizar ciertas revisiones que mejoren la versión actual del producto.

Acabado del proyecto. Incluye la documentación del proyecto y el almacenamiento de datos y software implicados.

OSEMN

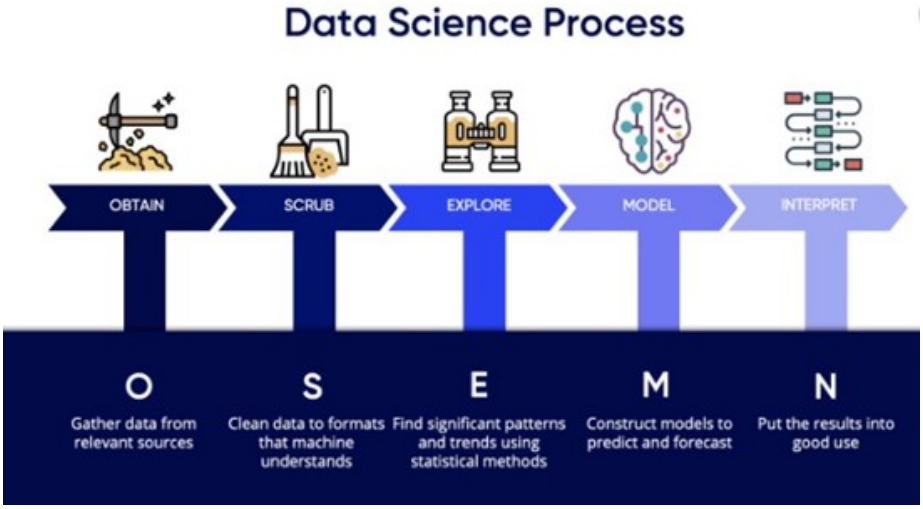


Figura 9. Fases del modelo OSEMN. Fuente: https://towardsdatascience.com/5-steps-of-a-data-science-project-lifecycle-26c50372b492

Este modelo de proceso tiene su origen en la referencia (Mason & Wiggins, 2010). El **modelo de proceso OSEMN se divide en los siguientes pasos** (ver figura 9): (1) obtención de los datos, (2) depuración de los datos, (3) exploración de los datos, (4) modelado de los datos e (5) interpretación de los datos. A continuación se da una breve explicación de cada paso.

Aunque los cinco pasos se presentan de forma lineal e incremental, en la práctica es común retroceder o realizar algunos de ellos en paralelo. El proceso en la ciencia de los datos es iterativo y no lineal. Por ejemplo, una vez modelados los datos y estudiado los resultados, es posible retroceder a la fase de depuración para ajustar ciertas características del conjunto de datos.

Obtención de los datos

El primer paso es obtener el conjunto de datos, para lo que se tendrá que realizar una o varias de las siguientes acciones: a) Descargar los datos de alguna localización remota (por ejemplo, un servidor o alguna página web); b) interrogar

alguna base de datos o API (por ejemplo, MySQL o alguna red social); c) extraer datos de algún fichero (por ejemplo, una hoja de cálculo o ficheros HTML) o d) generar datos (por ejemplo, mediante sensores o realizando resúmenes).

Depuración de los datos

Normalmente, los datos obtenidos tienen valores perdidos, inconsistencias, errores, caracteres extraños o incluso conjuntos de datos sin interés. En ese caso hay que realizar una limpieza de los datos. Algunas de las operaciones de limpiado más frecuentes incluyen: filtrado de líneas de datos, eliminación de ciertas columnas de datos, reemplazo de valores, extracción de palabras, gestión de valores perdidos o conversión de formato.

Es normal que más de un 60% de la duración de un proyecto se lo lleven estas dos primeras fases.

Exploración de los datos

Esta fase incluye las siguientes operaciones: revisar los datos, derivar estadísticas de los datos y realizar visualizaciones de los datos.

Modelado de los datos

Esta fase consiste en construir un modelo de los datos basado en métodos estadísticos o bien aprendizaje automático. Las técnicas para crear un modelo incluyen la regresión, clasificación, agrupamiento o reducción de la dimensión.

Interpretación de los datos

La fase final del modelo OSEMN es la interpretación de los datos. Esta fase incluye obtener conclusiones de los datos, evaluar qué significado tienen los resultados obtenidos y finalmente comunicar dichos resultados.

The Team Data Science Process

El proceso de ciencia de datos en equipo (TDSP) es una metodología de ciencia de datos ágil e iterativa desarrollada por Microsoft para proporcionar soluciones de análisis predictivo y aplicaciones inteligentes de manera eficiente. TDSP ayuda a mejorar la colaboración en equipo y el aprendizaje. Incluye los mejores procedimientos y estructuras usadas por Microsoft y otros fabricantes del sector para facilitan la correcta implementación de iniciativas de ciencia de datos.

El TDSP consta de los siguientes componentes clave:

Una definición de ciclo de vida de ciencia de datos.

Una estructura de proyecto estandarizada.

Infraestructura y recursos para proyectos de ciencia de datos.

Herramientas y utilidades para la ejecución de proyectos.

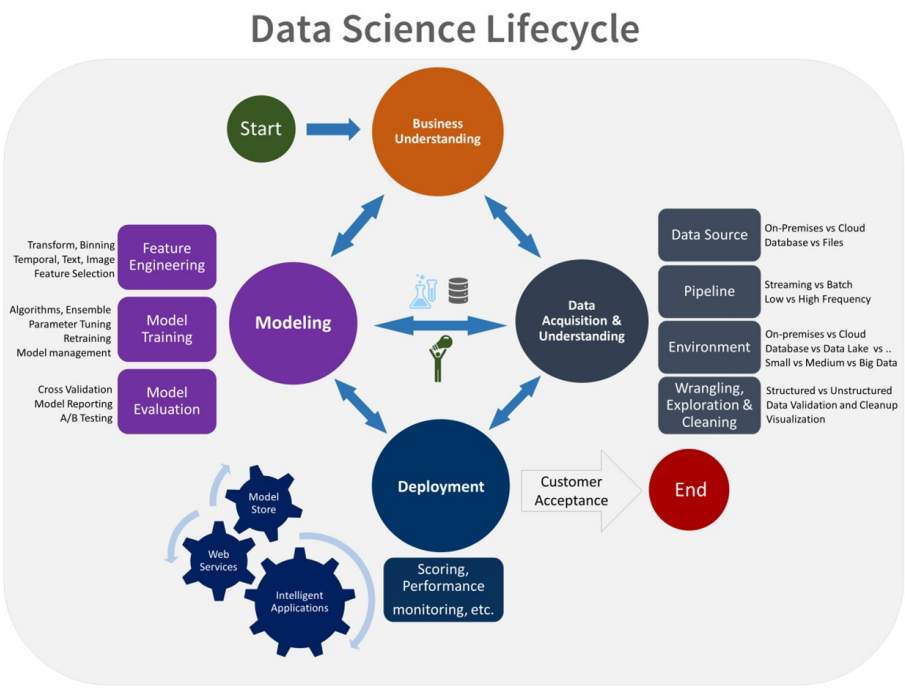


Figura 10. Fases del modelo TDSP. Fuente: https://docs.microsoft.com/es-es/azure/machine-learning/team-data-science-process/

Ciclo de vida

El ciclo de vida de TDSP se compone de cinco fases principales que se ejecutan de forma iterativa. Estas fases incluyen (ver figura 10): a) Conocimiento del negocio; b) adquisición y comprensión de los datos; c) modelado; d) implementación y e) aceptación del cliente. El TDSP describe los objetivos, las tareas y la documentación resultante de cada fase del ciclo de vida.

Estas tareas y documentación están asociados con personas que representan ciertos roles en el proyecto: arquitecto de soluciones, jefe de proyecto, científico de datos, responsable de proyecto.

En la figura 11 se proporciona una vista de cuadrícula de las tareas (en azul) y los artefactos/documentos (en verde) asociados con cada fase del ciclo de vida (eje horizontal) de estos roles (eje vertical).

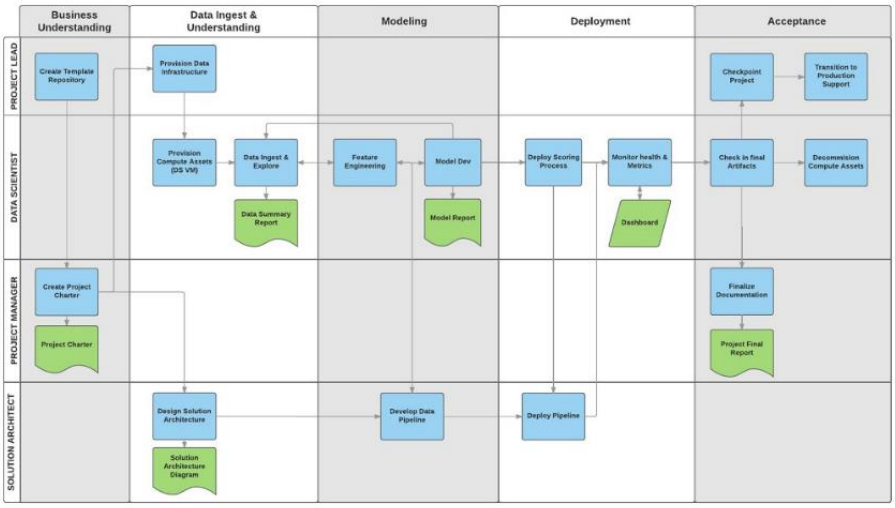


Figura 11. Fases del modelo OSEMN. Fuente: https://docs.microsoft.com/es-es/azure/machine-learning/team-data-science-process/

Estructura de proyecto estandarizada

Se propone utilizar en todos los proyectos la misma estructura de directorio y el mismo conjunto plantillas para clasificar y crear los documentos. El objetivo es facilitar a los miembros del equipo encontrar información sobre sus proyectos. Todo el código y los documentos se almacenan en un sistema de control de versiones (VCS), como Git, TFS o Subversion para permitir la colaboración en equipo. La estructura estandarizada para todos los proyectos ayuda a crear conocimiento institucional en toda la organización. Se proporcionan plantillas para la estructura de carpetas y los documentos necesarios en ubicaciones estándar. Esta estructura de carpetas organiza los archivos que contienen código para la exploración de datos, la extracción de características y los que registran las iteraciones de los modelos.

Infraestructura y recursos para los proyectos de ciencia de datos

TDSP proporciona recomendaciones para la gestión de análisis de datos compartidos e infraestructura de almacenamiento. Por ejemplo, se recomiendan sistemas de archivos en la nube para almacenar conjuntos de datos; bases de datos; clústeres de macrodatos (Hadoop o Spark); servicio de aprendizaje automático.

Herramientas y utilidades para la ejecución de proyectos

En la mayoría de las organizaciones la introducción de procesos presenta ciertos desafíos. Las herramientas proporcionadas para implementar el proceso y el ciclo de vida de ciencia de datos ayudan a reducir las barreras a su adopción y la normalizan. TDSP proporciona un conjunto inicial de herramientas y scripts para impulsar la adopción de TDSP dentro de un equipo. También ayuda a automatizar algunas de las tareas comunes del ciclo de vida de ciencia de datos, como la exploración de datos y el modelado de línea de base. Existe una estructura bien definida que se proporciona a los individuos para que contribuyan con herramientas y utilidades compartidas al repositorio de código compartido de su equipo. Estos recursos se pueden aprovechar luego en otros proyectos dentro del equipo o en la organización.

Temas interesantes: componentes y roles de las personas en los proyectos. Diferencias entre un científico de datos y un ingeniero de datos.

1.5. Referencias bibliográficas

Chapman, P., Clinton, J., Kerber, R., Khabaza, T., Reinartz, T., Shearer, C. y Wirth, R. (2000).*CRISP-DM 1.0 Step-by-step data mining guide*. The CRISP-DM consortium. Recuperado de: https://the-modeling-agency.com/crisp-dm.pdf

Cielen D., Meysman A. D. B. y Ali M. (2016) *Introducing data science. Big data, machine learning, and more, using python tools*. Manning Publications Co.

Gnanasundaram, S. y Shrivastava, A. (2012). *Information Storage and Management: storing, managing, and protecting digital information in classic, virtualized, and cloud environments* (2º ed.). John Wiley & Sons, Inc.

Godsey B. (2017). *Think like a data scientist. Tackle the data science process step-by-step*. Manning Publications.

Janssens J. (2015). *Data science at the command line*. USA: O'Reilly Media.

Mason, H. y Wiggins, C. H. (2010). *A taxonomy of data science*. Recuperado de: http://www.dataists.com/2010/09/a-taxonomy-of-data-science.

TDSP

Microsoft Azure (s. f.). *Documentación del proceso de ciencia de datos en equipo.*

Explicación en profundidad del proceso de ciencia de datos en equipo (TDSP). Representa un ejemplo de herramienta propietaria (Microsoft Azure) aplicada a la ciencia de los datos. Es especialmente interesante por el énfasis que hace en intentar incrementar la productividad del equipo desarrollo del proyecto proponiendo estandarizar muchas fases del proyecto.

Accede al documento a través del aula virtual o desde la siguiente dirección web: https://docs.microsoft.com/es-es/azure/machine-learning/team-data-science-process/

Test

1. La principal utilidad de la ciencia de los datos es:

A. Analizar, crear modelos y extraer conocimiento de datos.

B. Estudiar el contenido de cantidades masivas de datos.

C. Consolidar la infraestructura de datos de una empresa.

D. Detectar posibles clientes con problemas de pago.

2. ¿Qué frase es la más correcta?

A. Ciencia de los datos y *big data* son términos sinónimos y por tanto intercambiables.

B. La ciencia de los datos define un modelo de proceso de datos y *big data* se refiere a la disponibilidad de una cantidad masiva de datos para su procesado.

C. La ciencia de los datos solo es aplicable a conjuntos de datos *big data*.

D. La ciencia de los datos incrementa la productividad de transacciones en línea para el caso de *big data*.

3. Las tablas incluidas en las bases de datos son:

A. Datos estructurados.

B. Datos no estructurados.

C. Lenguaje natural.

D. Datos en formato texto.

4. La posible eliminación de valores falsos o inconsistentes de la fuente de datos se realiza en la fase:

A. Establecimiento de objetivos.

B. Preprocesamiento de los datos.

C. Análisis de datos.

D. Modelado de los datos.

5. Las fases de un modelo de proceso en la ciencia de los datos:

A. Es un proceso secuencial que no admite vuelta atrás.

B. Es un proceso iterativo que admite vuelta atrás.

C. Se realizan en paralelo para aumentar la productividad.

D. Se deben organizar según el hardware disponible en el proyecto.

6. ¿Cuál de las siguientes categorías corresponde a datos generados automáticamente por sensores o máquinas?

A. Datos en lenguaje natural.

B. Datos estructurados.

C. Datos generados por máquinas.

D. Datos de redes sociales.

7. ¿Cuál es una diferencia principal entre CRISP-DM y TDSP?

A. CRISP-DM es una metodología y TDSP es un modelo de datos.

B. CRISP-DM es una propuesta reciente, mientras que TDSP es obsoleto.

C. CRISP-DM se enfoca en fases generales y TDSP enfatiza la estandarización y colaboración en equipo.

D. CRISP-DM requiere el uso de herramientas Microsoft, mientras que TDSP es independiente de proveedor.

8. ¿Qué modelo de proceso propone las fases OSEMN?

A. TDSP.

B. Data Science Process (DSP).

C. CRISP-DM.

D. Mason & Wiggins.

9. ¿En qué fase del modelo CRISP-DM se evalúa si el modelo cumple con los objetivos de negocio?

A. Comprensión del negocio.

B. Modelado.

C. Evaluación.

D. Despliegue.

10. ¿Cuál es una característica del modelo Data Science Process (DSP) propuesto por Godsey?

A. El modelo DSP no considera herramientas específicas.

B. Divide el proceso en tres fases principales: preparación, construcción y finalización.

C. Se limita exclusivamente a proyectos académicos.

D. Es el único modelo orientado a datos en tiempo real.