Universidad de Sonora

Reconocimiento de patrones: Introducción

Ramón Soto C. (rsotoc@moviquest.com)

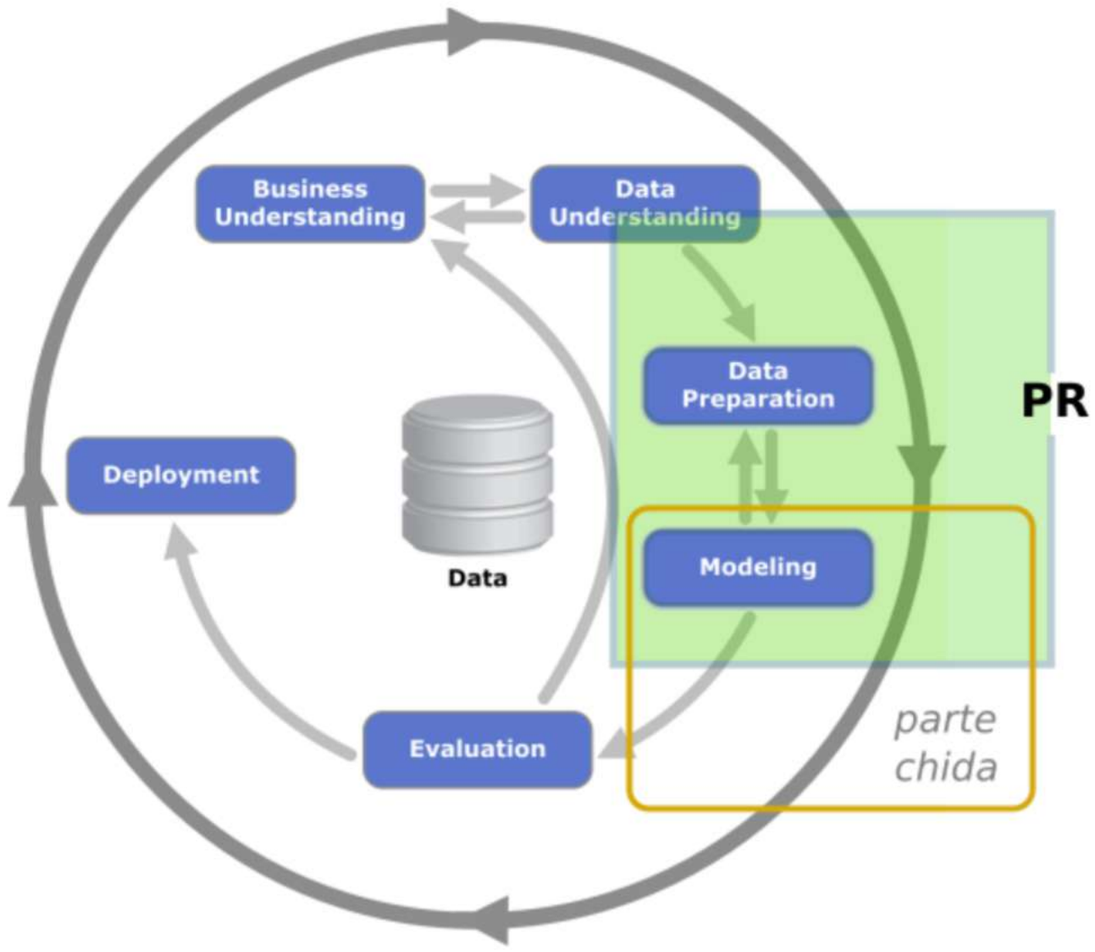
orandum est ut sit
mens sana in corpore sano

CRISP-DM

El crecimiento del interés de la industria en las metodologías de inteligencia artificial, particularmente para el caso de análisis de datos, ha venido acompañado de una formalización del proceso de desarrollo de soluciones.

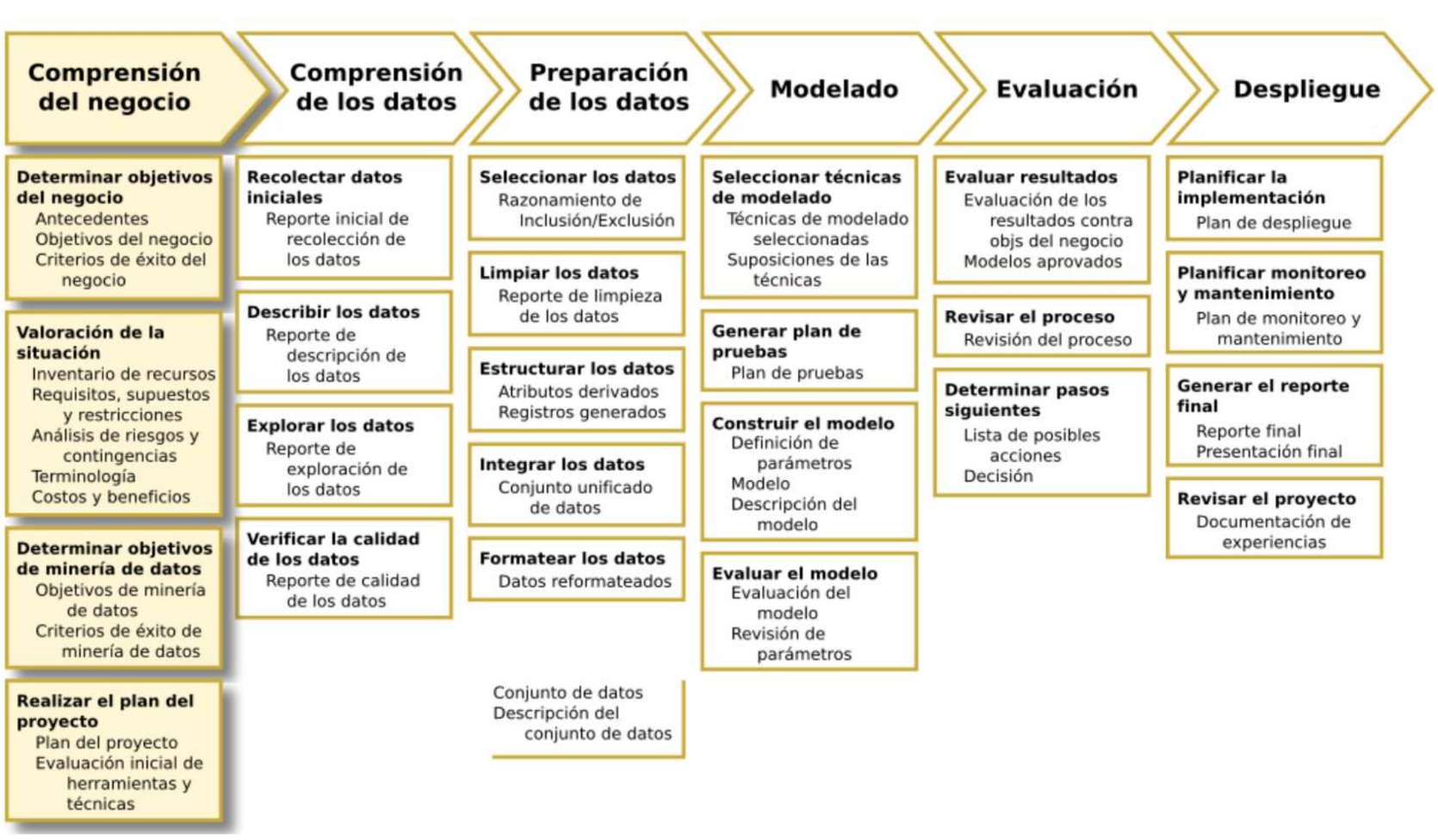
El modelo más conocido y utilizado en la industria para el desarrollo de proyectos de innovación para el análisis inteligente de datos es **CRISP-DM** (*Cross-Industry Standard Process for Data Mining*). Aunque este modelo ha sido criticado debido a diversas limitantes y ha venido siendo reemplazado por otras metodologías (principalmente TDSP,pero también otras como KDD o SEMMA), el modelo más reconocido, suficientemente simple y suficientemente poderoso y general para fines de este curso, sigue siendo CRISP-DM.

CRISP-DM es un marco de referencia que permite planificar el desarrollo de un proyecto de minería de datos (y asociados) a partir de 6 fases: 1) *Comprensión del negocio*, 2) *Comprensión de los datos*, 3) *Preparación de los datos*, 4) *Modelado*, 5) *Evaluación* y 6) *Despliegue* (implementación/puesta en marcha).



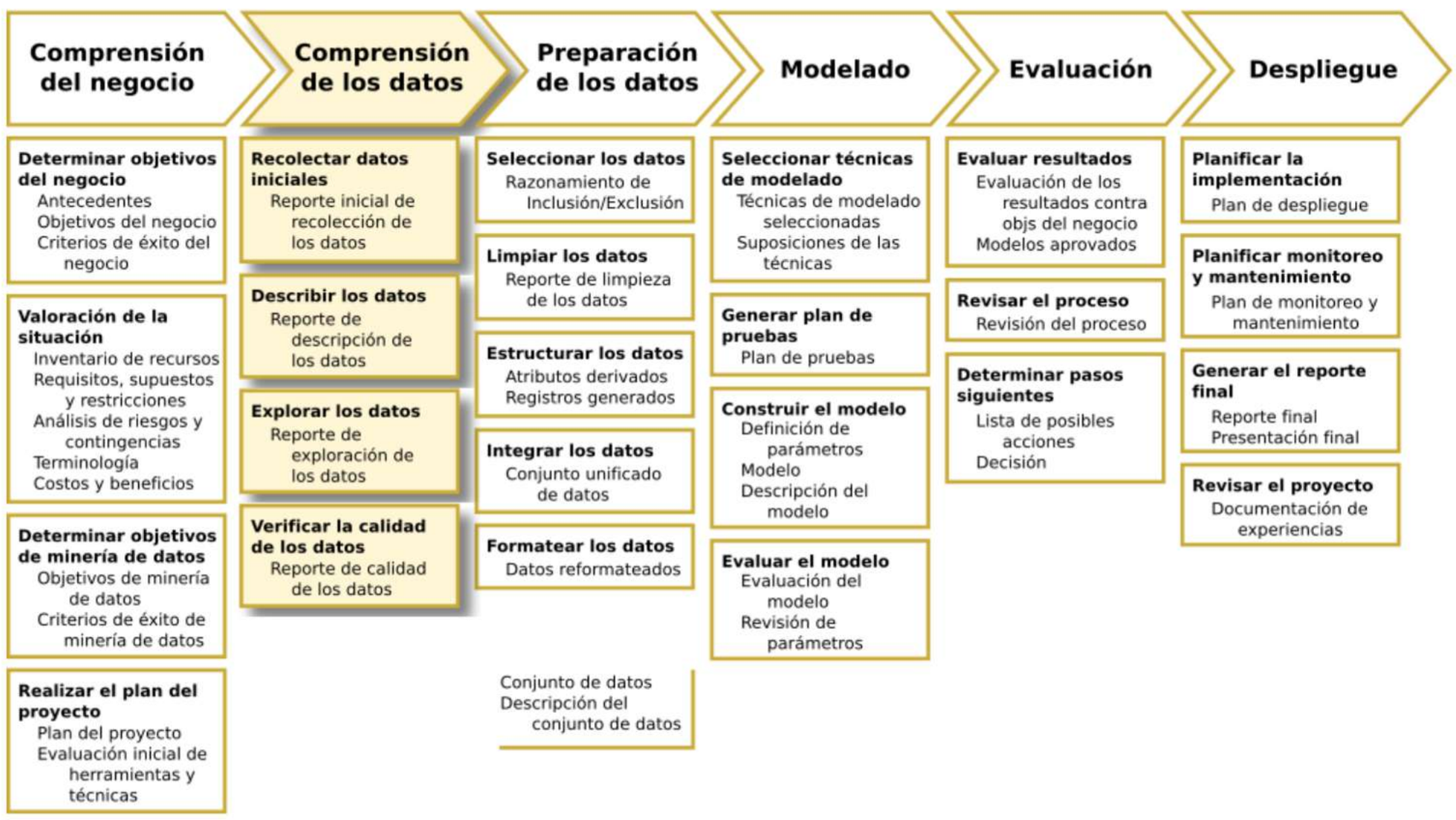
- **Comprensión del negocio**
  La fase inicial de un proyecto de minería de datos (y similares) se enfoca en entender los objetivos y requerimientos del proyecto, desde la perspectiva del *negocio*: *a*) identificar las características del problema que se quiere resolver, *b*) identificar cuáles son las necesidades prioritarias que el *cliente* quiere satisfacer y *c*) cuáles son los costos que el cliente está dispuesto a *pagar*. El resultado de esta fase debe ser la definición de un problema en términos técnicos (un problema de minería de datos, por ejemplo) y un plan preliminar de como alcanzar los objetivos.

  Las actividades en esta fase con sus respectivos resultados esperados son:
  - *Establecimiento de los objetivos del negocio*. El objetivo de este conjunto de actividades es conocer las características del cliente: ¿Cuáles son sus antecedentes?, ¿Cuales son sus objetivos de negocio?, ¿Cuáles son los criterios de éxito del negocio?
  - *Evaluación de la situación*. Con estas actividades se busca conocer la situación de la empresa para determinar su capacidad para explotar los datos con que cuenta. Entre las preguntas que hay que responder se encuentran las siguientes: ¿Cuáles son los términos más importantes para describir el negocio?, ¿Con qué recursos humanos y materiales cuenta la empresa para completar el proyecto?, ¿Qué tipos de datos están disponibles para el proyecto?, ¿Cuáles son los principales factores de riesgo?, ¿Cuáles son los planes de contingencia para cada factor de riesgo?, ¿Cuáles son los potenciales costos y beneficios del proyecto?
  - *Establecimiento de los objetivos de la minería de datos*. Aquí se busca determinar los objetivos del proyecto de minería de datos (¿qué se espera obtener con el proyecto?: ¿Una nueva herramienta o servicio? ¿Información para planificación estratégica?) y los criterios que permiten evaluar el éxito del proyecto.
  - *Generación del plan del proyecto*. La fase de comprensión del negocio debe concretarse en una determinación de intervención (hasta este punto, ¿se considera viable la realización del proyecto?) y en caso de ser positiva, en un plan de como realizar las siguientes fases de intervención (recursos a utilizar, compromisos, indicadores de avances, etc.).



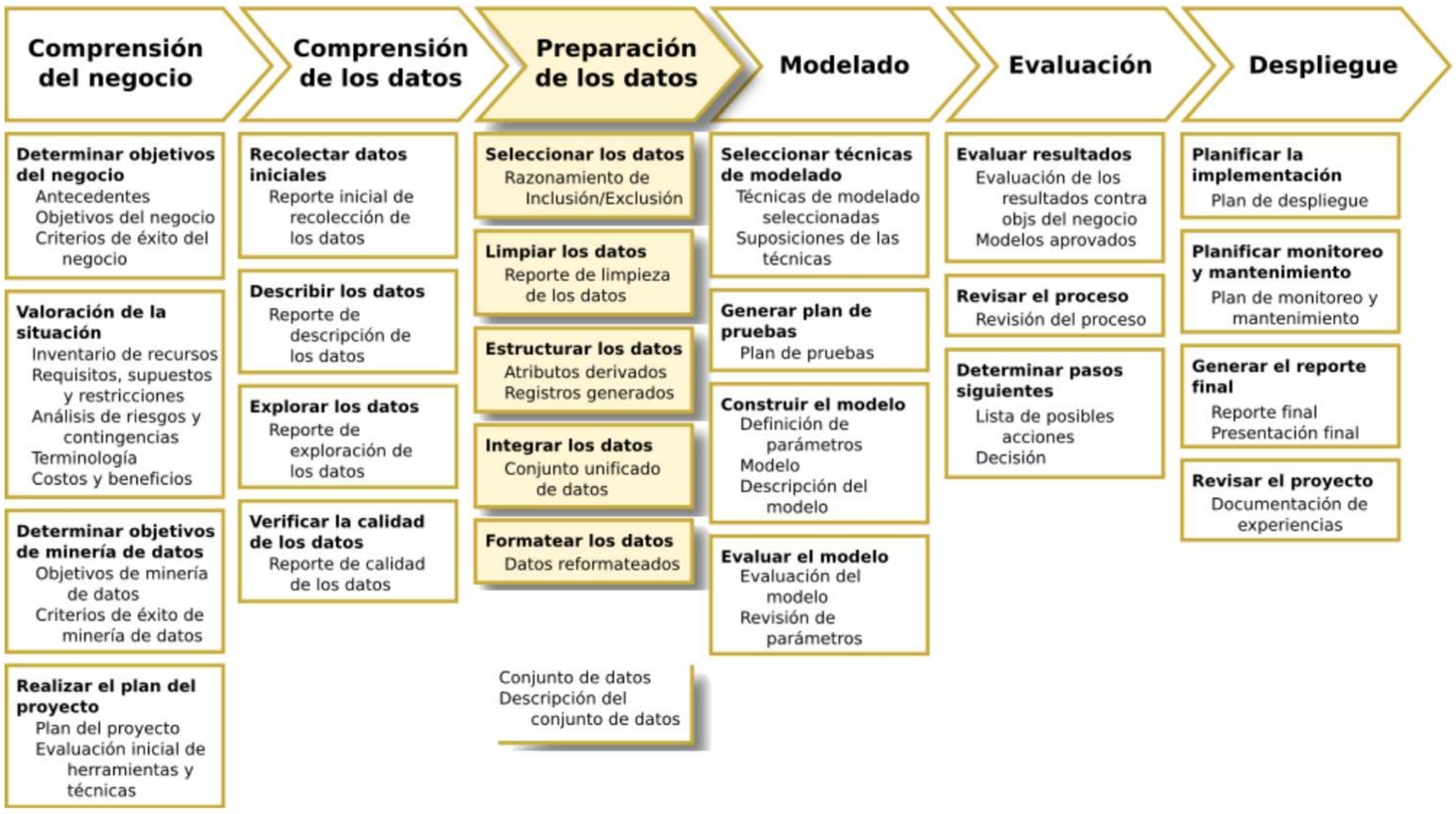
- **Comprensión de los datos**
  La comprensión de los datos se refiere a familiarizarse con las características de los datos disponibles para el proyecto y de los requerimientos adicionales de datos. Es una actividad fundamental para el desarrollo del proyecto dado que ésta es la base de todas las actividades que se realizarán a continuación. Es por ello que, en ocasiones, será necesario regresar a analizar el negocio con el fin de comprender mejor los datos.

  Las actividades en esta fase, con sus respectivos resultados esperados son:
  - *Recolección inicial de datos*. En este primer paso, se toman muestras de los datos disponibles y se identifica el alcance, así como posibles dificultades para su recolección y uso. Se distinguen los datos propios de la empresa de otros conjuntos de datos complementarios adquiridos y se identifican las fuentes adicionales de datos.
  - *Descripción de los datos*. A continuación, se identifican las características generales de los datos, como son el número de variables disponibles, la cantidad de registros, la frecuencia de generación de cada variable, su identificación, el significado de cada campo y el formato inicial. El resultado más importante de este análisis preliminar es una determinación de si los datos disponibles son suficientes para alcanzar los objetivos de la minería de datos.
  - *Exploración de los datos*. El objetivo de esta actividad es identificar la distribución general de los datos a través de pruebas estadísticas básicas y establecer hipótesis preliminares. Este análisis permite identificar la complejidad del problema y realizar una selección preliminar de técnicas a utilizar.
  - *Verificación de la calidad de los datos*. En este paso se verifica la completitud de los datos. Se buscan los porcentajes de datos incompletos, valores fuera de rango o no típicos y variables equivalentes. Se definen estrategias generales para resolver los problemas identificados.



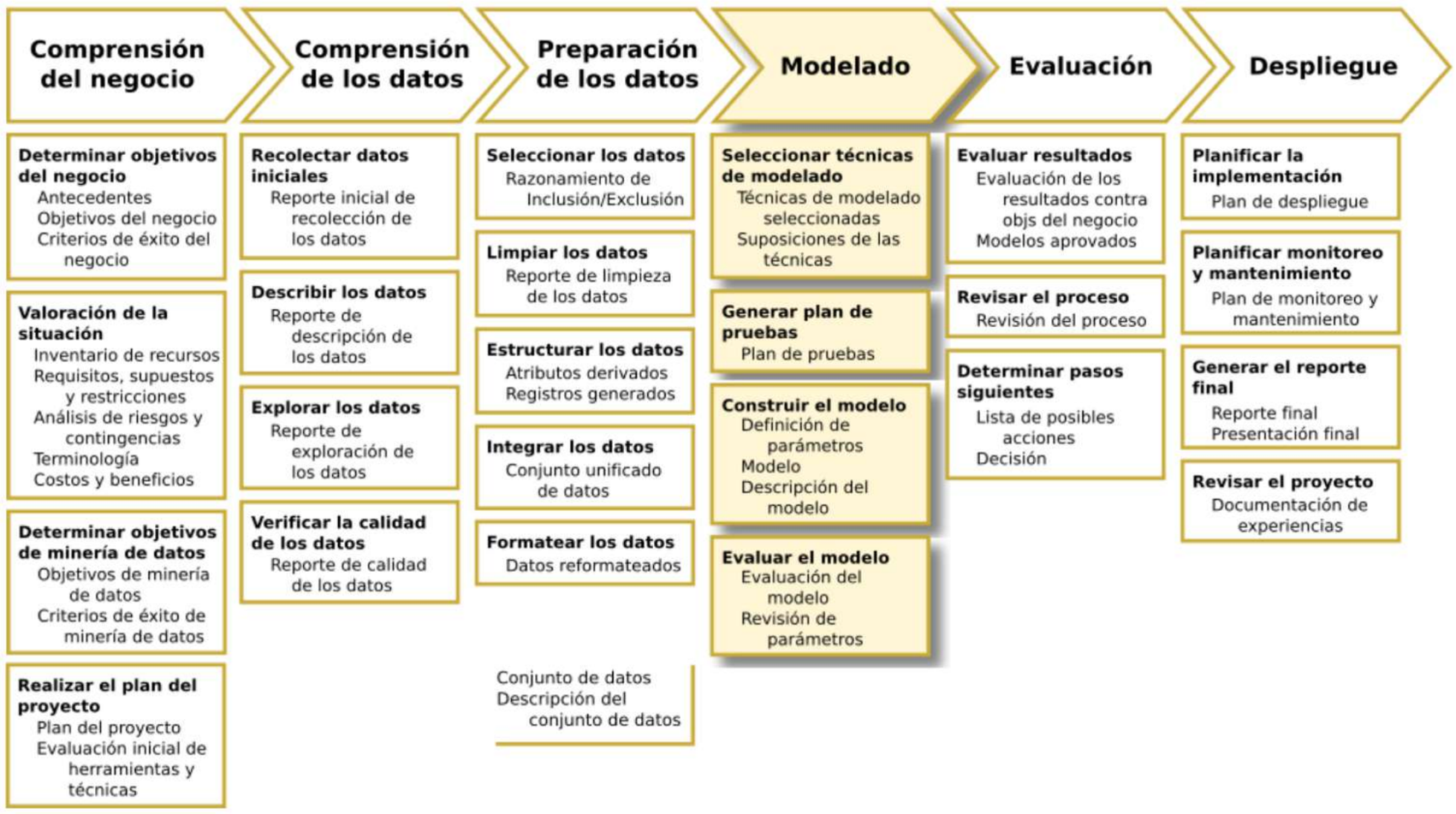
- **Preparación de los datos**
  La fase de preparación de los datos incluye todas las actividades necesarias para generar el conjunto de datos final que se utilizará para alimentar los modelos, partiendo de los datos originales. Está ampliamente reconocido que esta suele ser la fase que consume más tiempo en un proyecto de ciencia de los datos.

  En esta fase se realizan las siguientes actividades:
  - *Selección de los datos*. Aquí se busca escoger una porción del volumen total de datos preseleccionados que parezca representativa del problema de minería de datos. Por una parte, se realiza una selección de registros suficientemente amplia para cubrir todo el universo de objetos a analizar y, por otra parte, se seleccionan las características (variables) que mejor describen los diferentes objetos, tratando de tener la representación más rica posible y evitar, al mismo tiempo, variables que sean básicamente equivalentes. Es importante justificar y documentar las razones por las que diferentes subconjuntos de datos se van a incluir o excluir.
  - *Limpieza de los datos*. Esta actividad, que es la que más tiempo y recursos consume, tiene el objeto de subsanar las deficiencias de los datos identificadas en la fase previa. Entre las principales tareas a realizar sobresale el tratamiento a datos con valores faltantes y el manejo de datos atípicos y/o inconsistentes.
  - *Estructuración de los datos*. Esta actividad consiste en generar la estructura de los registros que se emplearán en el análisis, principalmente mediante la generación de nuevas variables que resulten más descriptivas de los datos y que ayuden a reducir la complejidad del espacio de representación.
  - *Integración de los datos*. La integración de datos consiste en unir datos de diferentes fuentes en un sólo conjunto de datos. Puede tratarse de crear una tabla unificada a partir de diferentes tablas o de generar registros o columnas nuevas a partir de la agregación de datos de diferentes fuentes.
  - *Formateo de los datos*. Esta actividad tiene el objeto de poner los datos en la forma en que serán procesados, típicamente mediante transformaciones que no alteran su significado. Entre las tareas más tareas más comunes de formateo de datos se encuentran el cambio de escala, la eliminación de caracteres especiales y el reordenamiento de columnas y renglones en datos tabulares.



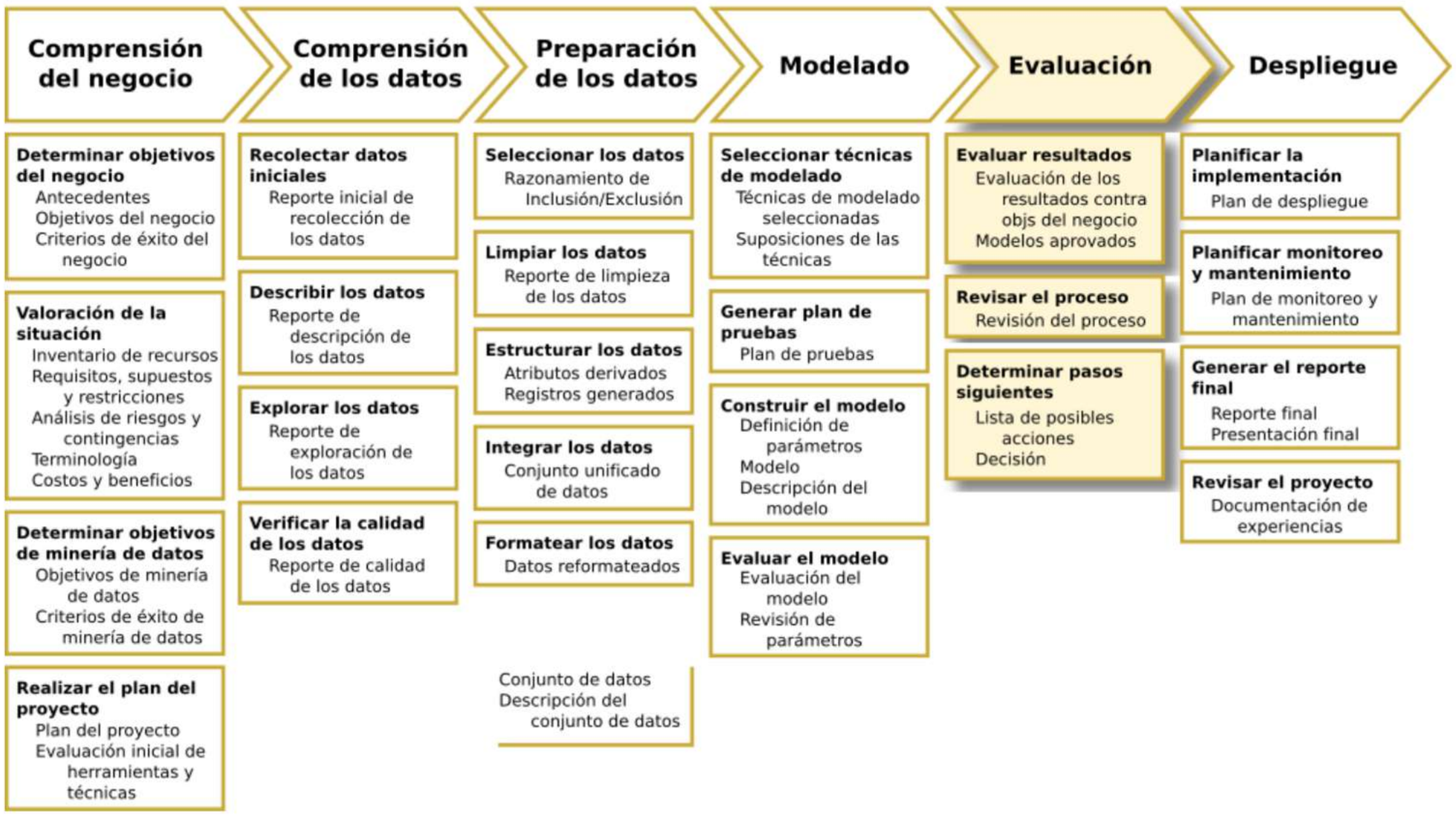
- **Modelado**
  En esta fase se eligen y se prueban diversas técnicas de modelado, afinando sus parámetros para ajustarse a la dinámica representada por los datos. En el proceso, suele ocurrir que una técnica requiera datos no contemplados en las fases previas y sea necesario dar marcha atrás para rectificar la construcción del conjunto de datos. La elección de las técnicas a utilizar se realiza utilizando criterios técnicos (como la pertinencia de la técnica para el problema específico) y prácticos (como la disponibilidad de datos adecuados, el tiempo disponible para obtener un modelo o el conocimiento de la técnica por parte del equipo de desarrollo).

  Las actividades en esta fase, con sus respectivos resultados esperados son:
  - *Selección de las técnicas de modelado*. En este paso, se eligen las técnicas de modelado que se emplearán. La selección depende de una serie de factores relativos al proyecto, entre los que destacan el grado de estructuración del problema y de los datos (¿existen modelos formales del sistema?, ¿Existen relaciones bien identificadas entre variables?, ¿se dispone de conocimiento experto?, ¿existen suficientes datos de observación?, ¿de qué tipo son los datos existentes?), los objetivos de la minería de datos, el dominio de técnicas por parte del equipo de desarrollo, restricciones legales y las preferencias del cliente. En cualquier caso, es muy común que se requiera el uso de diversas técnicas para la solución de un problema, particularmente si es un problema complejo.
  - *Generación del plan de pruebas*. Una vez seleccionadas las técnicas de modelado, se debe crear un plan de cómo realizar la implementación del prototipo (lo que suele llamarse la "**prueba de concepto**"). Esta actividad incluye tareas como la selección de bibliotecas y herramientas, la implementación de las técnicas, definición de una estrategia de segmentación de datos para creación del modelo y para realización de pruebas y selección de medidas de evaluación.
  - *Construcción del modelo*. A continuación, se construye el modelo (o conjunto de modelos, colaborativos o competidores). Se definen los parámetros de cada modelo, se hacen pruebas preliminares y se realizan ajustes al modelo. Aquí es importante destacar el carácter incierto típico en todo proyecto de minería de datos que obliga, frecuentemente a regresar a pasos previos, en este caso a la selección de modelos, definición de parámetros e incluso, a la fase de preparación de los datos. El resultado debe ser un modelo afinado, adecuado al problema.
  - *Evaluación del modelo*. Finalmente, se evalúa el modelo, haciendo pruebas con los datos reservados para ello, se realiza un reporte de los niveles de precisión/error, tiempos de respuesta, potenciales puntos críticos y cualquier otra información que sea relevante para la implementación final del sistema.



- **Evaluación**
  El resultado esperado de la fase de modelado es un modelo o un conjunto de modelos con un buen desempeño desde un punto de vista de los datos; es decir, un conjunto de modelos capaces de "explicar" el comportamiento de los datos. En la fase de evaluación se analiza la pertinencia de los modelos desarrollados en relación con los objetivos del negocio.

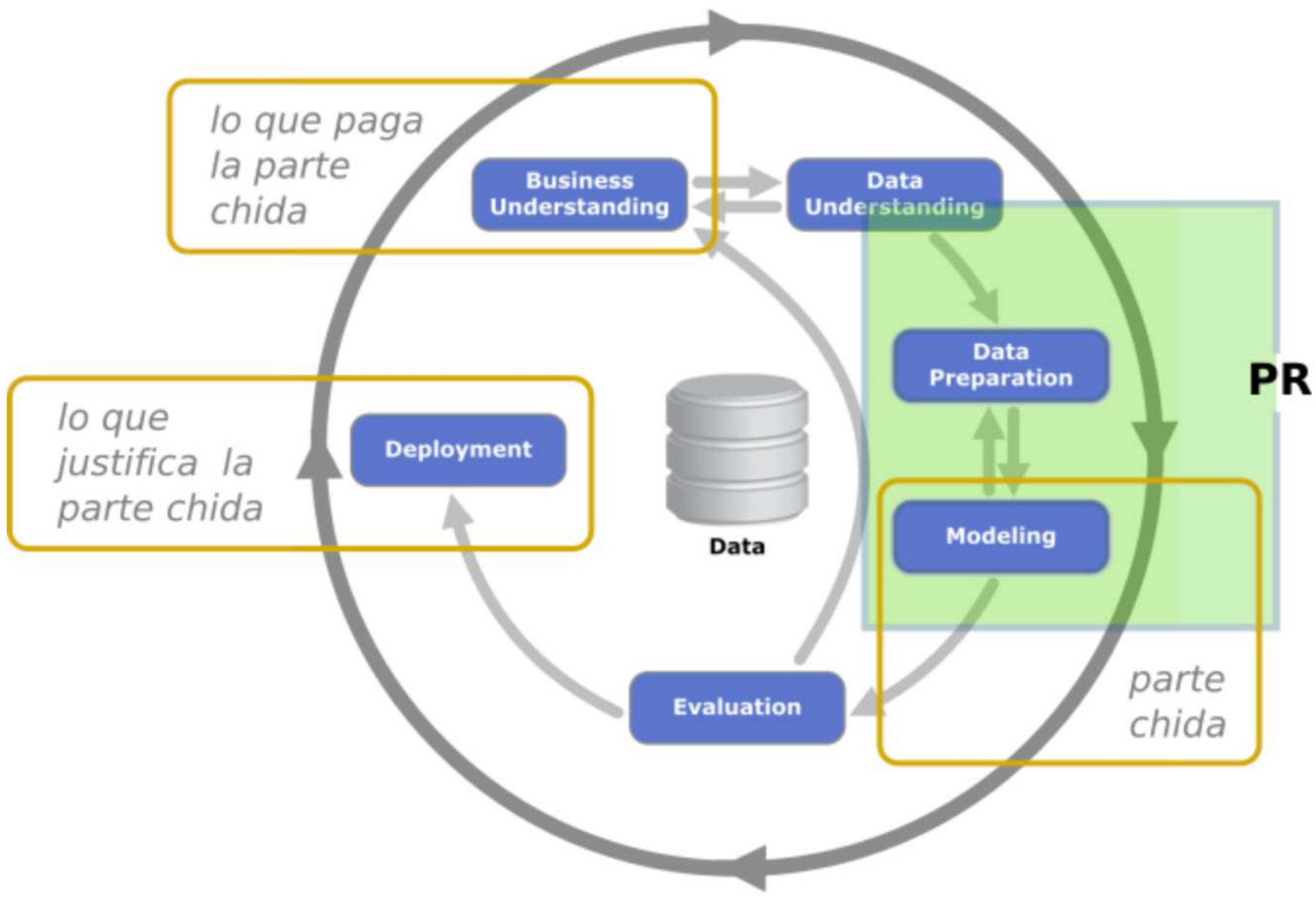
  Las actividades por realizar en esta fase son:
  - *Evaluación de los resultados*. En esta fase se evalúan los resultados que arrojan los modelos desarrollados y se comparan tales resultados con los objetivos de negocio. Se identifican objetivos del negocio que pudieran no estar resueltos y que pudieran requerir incluir nuevas herramientas e incluso, se analiza la posibilidad de ampliar los objetivos de negocio con resultados emergentes del modelado.
  - *Revisión del proceso*. Aquí se realiza una revisión de todo el proceso seguido hasta el momento, desde la comprensión del negocio, se realizan los ajustes necesarios en cada etapa yt se presentan propuestas de como mejorar todo el proceso.
  - *Determinación de los pasos siguientes*. En este paso se plantean las opciones a seguir, que pueden ir desde abandonar el proyecto (si los resultados obtenidos hasta el momento prevén un impacto no rentable en el negocio), regresar a la fase inicial, replantear y corregir los pasos necesarios o preceder a la implementación.



- **Despliegue**
  Una vez realizado el modelado del problema y obtenidos resultados satisfactorios, es necesario transformar el modelo obtenido en un producto. Este producto puede ser un nuevo sistema de información para la toma de decisiones, una herramienta para detección de determinados comportamientos de interés en los datos que genera la empresa o una ampliación del conocimiento de la empresa que conduzca a nuevos procedimientos.

  Esta fase contempla las siguientes actividades:
  - *Planificación de la implementación*. Este paso es determinante para lograr que la implementación se integre adecuadamente y con la menor perturbación posible al sistema actual. Deben definirse como se extenderán los sistemas actuales para incorporar los nuevos resultados, como se modificarán los procedimientos de acuerdo a la nueva información de negocios, cómo se desarrollarán los recursos humanos necesarios para la implementación de los cambios.
  - *Planificación del control y del mantenimiento*. Además de planificar la implementación, hay que planificar los procedimientos de control y mantenimiento que permitan darle seguimiento a los resultados que se espera obtener, particularmente en lo que se refiere a rentabilidad y estabilidad.
  - *Generación de un informe final*. El paso final del proyecto es elaborar un reporte final, que incluya la documentación técnica del proyecto, manuales de usuario y casos de uso. Adicionalmente, deben generarse un reporte ejecutivo que resuma los resultados del proyecto y cualquier otro apoyo para la presentación final ante el cliente.
  - *Revisión del proyecto*. Adicionalmente, en esta fase se elabora también un reporte anecdótico acerca del desarrollo del proyecto, que facilite el desarrollo de posteriores proyectos de innovación.

La metodología CRISP-DM permite mantener la atención puesta a todos los aspectos importantes para asegurar el éxito de un proyecto de minería de datos, por lo qué, de manera explicita o implicita, es conveniente tenerlo como guia.



Conclusiones

- El interés de la industria por las metodologías de aprendizaje automático (y reconocimiento de patrones, por lo tanto), se debe a la promesa de mejorar la productividad de las empresas. Este interés genera oportunidades de negocio, opciones de empleo bien pagadas, vinculación industria-universidad, desarrollo regional.
- Para poder aprovechar estas oportunidades, es importante establecer un plan de acción sistemático, basado en una buena comprensión de lo que el cliente busca.
- CRISP-DM es la metodología *de facto* en la industria para el desarrollo de proyectos de minería de datos. Esta metodología hace énfasis en garantizar que las soluciones planteadas se integren adecuadamente al negocio y le generen valor.

Tarea 2

Presente el análisis de un problema de reconocimiento de patrones utilizando la metodología CRISP-DM.

Fecha de entrega: 31 de agosto.