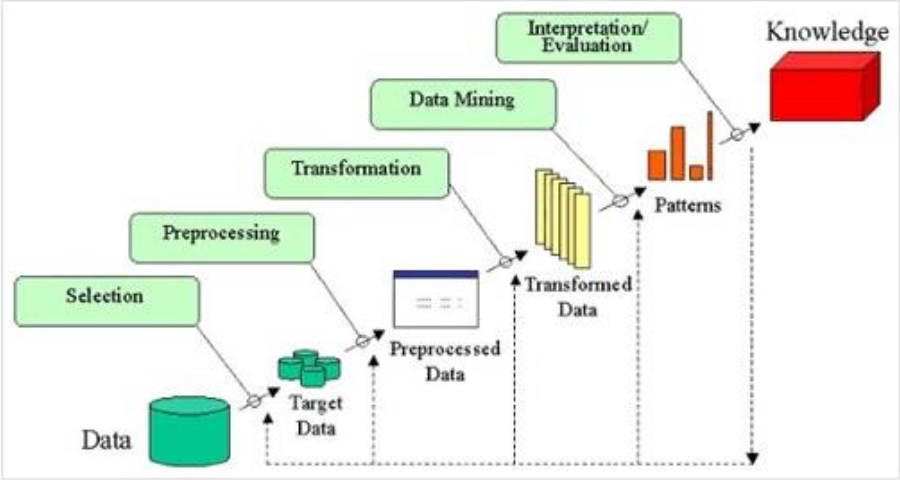
METODOLOGIAS PARA PROYECTOS DE DATA SCIENCE

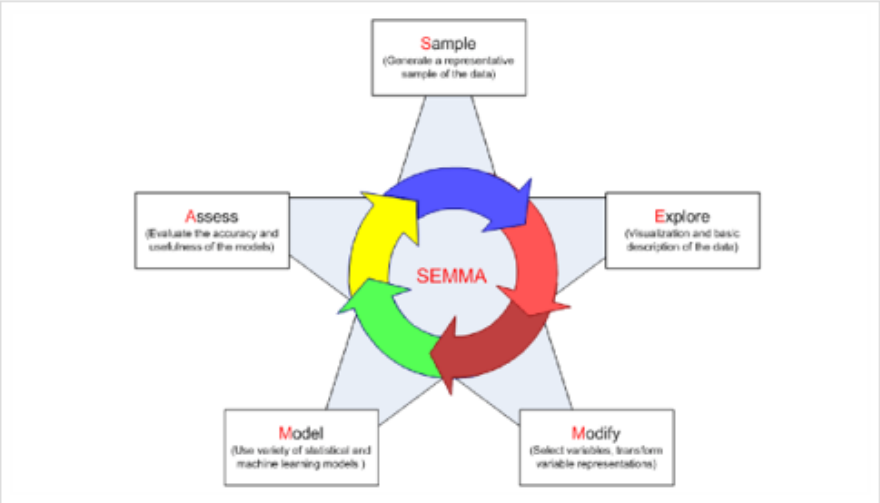
1) KDD (Knowledge Discovery in Databases)



Metodología de 5 pasos. Inicia con la selección donde de un data set principal hay que selecciónar un subconjunto de variables que nos pueden apoyar en la exploración del fenómeno que estamos estudianto. En el pre-procesamiento realizamos la limpieza y balanceo de datos. En la transformación, el método sugiere que reduzcamos dimensiones con técnicas estadísticas para manejar la menor cantidad de variables necesarias. En minería de datos buscamos patrones de interés o representativos en relación al objetivo de la minería de datos. Finalmente para colarnos al conocimiento pasamos por el proceso de intepretación y evaluación de modelo. Al final de la iteración se le otorga una calificación al modelo y si no se cumplieron satisfactoriamente los objetivos se repite hasta que sean logrados.

by in Noelia Ferrero

2) SEMMA (Sample, Explore, Modify, Model and Access)



En esta metodología iniciamos con «sample» o un muestro de la base de datos principal (que asumimos que es muy pesada y lenta de procesar) para poder hacer manipulaciones sobre este pequeño set de una manera ágil. Después exploramos los datos para ganar entendimiento e ideas, así como refinir nuestro proceso de búsqueda de anomalías, patrones y tendencias. Llegamos entonces al paso de modificar donde nos enfocamos en crear, seleccionar y transformar variables para enfocarnos en un proceso de selección. En esta etapa también se buscan anomalías y reducir el número de variables. Luego sigue la etapa de modelaje en donde debemos aplicar distintos métodos estadísticos evaluando sus fortalezas y cumplimiento de objetivos. Finalmente, la etapa de «access» que significa evaluar la confiabilidad y utilidad de los hallazgos. Se evalúa particularmente el «performance».

De la misma manera del modelo anterior, si no se logran los objetivos en una primera iteración tendremos que repetir el proceso.

3) CRISP-DM (Cross-Industry Standard Process for Data Mining)

Seguimos con el «famosisimo» CRISP-DM, el método más usado en la industria y es que IBM, la compañía dueña de Watson que antes desarrollaba poderosas computadoras, es quien desarrolló este modelo. La diferencia clave es que cualquier etapa del modele puede tener retorno o iniciar una reversa al método. Si durante la etapa en particular el especialista encontró que los datos no son suficientes para resolver su objetivo, puede regresar a cualquiera de la otras etapas.

En la etapa de «Entendimiento de negocio» primero se determinan los objetivos de negocio: Antecedentes, objetivos estratégicos de impacto y criterios de éxito. Después revisamos la situación, inventariamos recursos, realizamos un análisis de costo-beneficio, determinamos objetivos y producimos un plan de proyecto.

En «Data Understanding» es donde recolectamos los datos iniciales, describimos cada uno de estos datos, exploramos y verificamos la calidad de la información.

En «Data preparation» seleccionamos la información más razonable, la limpiamos, construimos variables de ser necesario, integramos datos y finalmente formateamos. El entregable de esta etapa sería un dataset listo para trabajar.

Para la etapa de «Modeling», similar a los otros modelos, experimentamos con distintas técnicas, consideramos supuestos, hacemos pruebas, definimos parámetros y revisamos funcionalidad general de los modelos.

En «Evaluación» es donde considerando los criterios de éxito definidos consideramos como positiva y/o negativa la evaluación. Aqui mismo definimos los siguientes pasos y tomamos las decisiones necesarias.

Finalmente, en «Deployment», esta etapa sólo se activa si el proyecto tuvo evaluación positiva. Se genera entonces un plan de desarrollo, un plan de mantenimiento, se genera un reporte final y presentación para socializar el caso de estudio.

Comparando Métodos…

KDD	SEMMA	CRISP-DM
---	---	Business Understanding
Selection	Sample	Data Understanding
Preprocessing	Explore	
Transformation	Modify	Data Preparation
Data Mining	Model	Modeling
Interpretation/Evaluation	Assess	Evaluation
---	---	Deployment