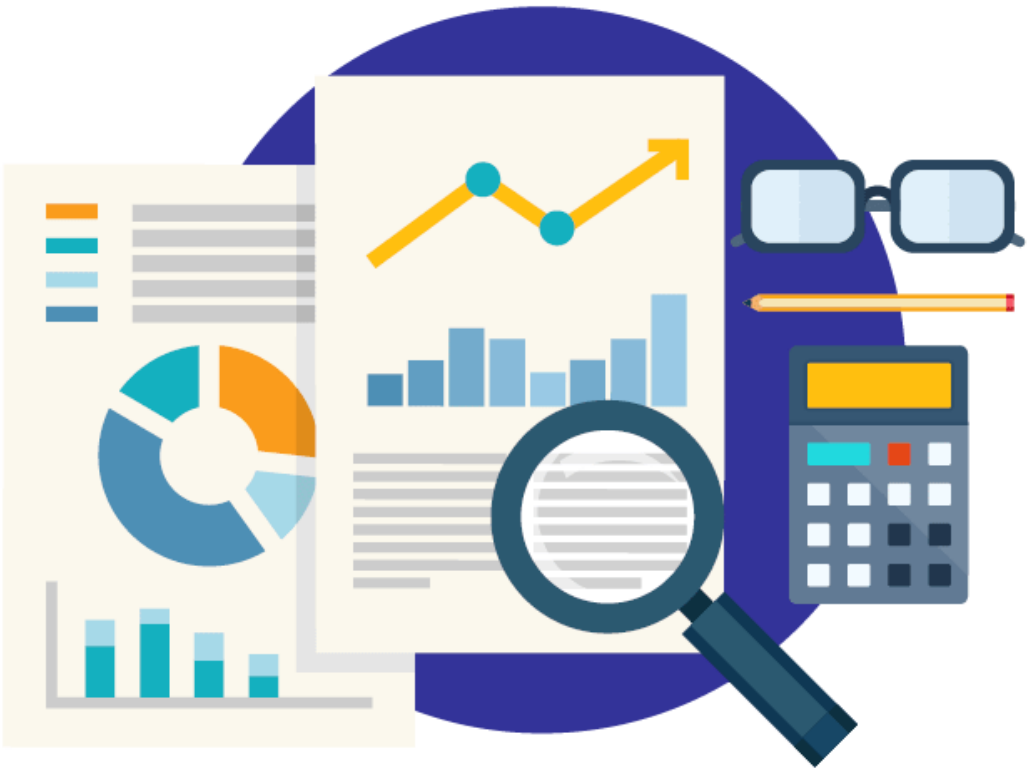

Formación en Ciencia de Datos

Una Perspectiva Estudiantil

Wilmer González / wilmer.a.gonzalez@ucv.ve

La reciente demanda generada por la Ciencia de Datos en el último año, ha tenido como consecuencia, un auge asociado a la participación en cursos ofrecidos bajo plataformas MOOC en estas áreas. Sin embargo, existen otras alternativas para el aprendizaje de esta disciplina.

Antes de llegar a la recomendación de nuevas herramientas, cursos o aplicaciones, resulta necesario entender algunos principios que contribuyen al aprendizaje adecuado de los conceptos requeridos por esta área del conocimiento. La subsanación de esta necesidad, evita caer en la mera memorización de conceptos como parte de la formación.

Principios metodológicos:

Entre ellos, se tienen:

- **Abstracción:** Es importante delimitar el contexto del problema y así cercar los conceptos que manejaremos durante el desarrollo del método. Inicialmente, este principio ayuda a concentrarse en la parte técnica de lo que requerimos aprender.

- **Rigor y Formalidad:** Al momento de conocer el antecedente teórico necesario, es imprescindible mantener durante el aprendizaje, un lenguaje propio, pragmático uso de los términos, y la menesterosa formalidad en cualquier formulación planteada.

- **Incrementalidad:** Los principios anteriores confieren un orden de ideas deseable de seguir, pero es la apropiada construcción de conceptos complejos, que se basan en fundamentos elementales o de menor nivel, lo que permitirá entender problemas más profundos y complicados.

- **DRY (Don't Repeat Yourself):** Al comenzar la implementación o puesta en práctica de los conceptos aprendidos, resulta tentador desarrollar herramientas a la medida del problema, sin revisar el estado del arte del mismo, a pesar de esto, es muy probable que el problema en cuestión posea una solución existente (al menos parcialmente) que sirva como punto de partida de una adaptación más eficiente.

Ruta de aprendizaje

Tomando en cuenta los principios antes mencionados, la ruta de aprendizaje puede resultar en:

1. Aprendizaje continuo de fundamentos teóricos:
   Se plantea la revisión de bibliografía con conceptos asociados a probabilidad y estadística (como es el caso del libro de Inferencia Estadística de Sanjeev Kulkarni y Gilbert Harman.), así como inteligencia artificial (con el texto de Peter Norvig) y tópicos relacionados con Álgebra Matricial (donde convendría Algebra and Applications in Data Mining de Elden).

2. Seguimiento de cursos prácticos: Igualmente es recomendable realizar cursos que permitan practicar las actividades que requieren los problemas reales, para ello, los siguientes cursos se han vuelto estándar en cuanto al área de ciencia de datos:
   - Machine Learning. University of Stanford.
   - Data Science. John Hopkins University.
   - Data Mining. University of Illiniois at Urbana-Champaign.

En paralelo al material antes señalado, la resolución de problemas existentes (Learn by doing), sigue siendo, una de las formas de aprendizajes más eficientes disponibles.

Aprender haciendo

Una vez contemplados los fundamentos teóricos mencionados, la resolución de estos problemas reales, puede iniciar de forma asistida, mediante problemas bien definidos en sitios como Kaggle.com. UCI-repository también permite encontrar datos abiertos dispuestos para tareas específicas de minería de datos, una extensa colección de estos set de datos está alojada en el Repositorio: awesome-public-datasets (github).

También puede generarse de forma espontánea accediendo a datos públicos y planteándose preguntas de interés. Por ejemplo: Memorias y Cuentas de las instituciones gubernamentales, Análisis en redes sociales, Web Scraping de comercio electrónico, entre muchas otras alternativas. El límite se diluye cada vez más, mientras las corrientes de digitalización de datos y datos de acceso público, cobran fuerza.

Buenas prácticas

En conjunto con los principios y herramientas antes mencionados, existen también ciertas convenciones que proveen un ecosistema adecuado para la colaboración y desarrollo de soluciones en la Ciencia de Datos, garantizando características como la reproducibilidad de las actividades, la transparencia en ejecución de modelos, y otras que permitan mantener el ciclo de desarrollo activo.

Algunas de estas convenciones(o buenas prácticas) son:

- Alojar el código usado en los proyectos en repositorios públicos que permitan su evolución (siempre que sea posible) como: Github.
- Alojar el código usado en la elaboración de artículos científicos, productos de la investigación, en estructuras que permitan su reproducibilidad como: Figshare, Overleaf, entre otras tecnologías disponibles.
- Difundir y/o promover el uso de directorios de contenido de acceso público.

Este artículo intenta resumir algunas consideraciones metodológicas resultantes de la constante interacción e investigación sobre Ciencia de Datos en Venezuela; y procura catalizar la investigación de personas que recién incursionan en la mencionada disciplina.

Finalmente, una vez que percibimos la velocidad de producción de conocimiento de esta área, se vuelve inevitable integrar distintas fuentes de información para complementar el conocimiento. ■