Notas de importancia

Con base a las metodologías descritas en el capítulo anterior, cabe señalar, que no tienen un dominio especifico; por lo tanto queda abierto para los investigadores la selección de la metodología que se considere más adecuada para un problema determinado. Dicho lo anterior, en el estado del arte se encontraron metodologías, que, aunque abarcan el proceso necesario para el análisis de datos, no generan el valor suficiente para la solución de un problema determinado. Por esta razón, autores como \cite{Martinez2021} consideran que aunque la comunidad de investigación en ciencia de datos está creciendo día a día, está explorando nuevos dominios, creando nuevos roles especializados y adicional está realizando un gran esfuerzo de investigación para desarrollar análisis avanzados, mejorar modelos y generar nuevos algoritmos apoyados de los campos de las matemáticas, la estadística y la informática, estas habilidades no son suficientes para su aplicación en proyectos reales, puesto que la mayoría de proyectos basados en datos presentan problemas organizativos y socio-técnicos, tales como: una falta de visión y claridad en los objetivos, un énfasis sesgado en cuestiones técnicas y ambigüedad de roles. Y aunque estos problemas existen en proyectos de ciencia de datos del mundo real, la comunidad no se ha preocupado demasiado por ellos y no se ha escrito lo suficiente sobre las soluciones para abordar estos problemas.

Cabe resaltar que la metodología \textit{CRIPS-DM} fue seleccionada debido a que comprende todas las fases básicas que debe cumplir cualquier proyecto que tenga en su contexto el análisis de datos, ademas de que es bastante utilizada por una cantidad considerable de investigadores, sin embargo aunque sea la metodología más utilizada, carece de lineamientos que profundicen en la organización de los equipos de trabajo para llevar a cabo procesos de gestión que se alienen con el software y las metodologías de desarrollo ágiles. Para comprender mejor, autores como\cite{Schroer2021} consideran que aunque esta metodología sigue siendo un estándar por defecto en la minería de datos, tiene inconvenientes, ya que la mayoría de los estudios no prevén una fase de implementación.

Algo semejante sucede con las metodologías \textit{KDD} y \textit{SEMMA}, antecesoras de CRISP-DM. Por el lado de KDD, tenemos que, aunque parte del entendimiento del dominio, su ciclo de vida termina en la visualización e interpretación del conocimiento obtenido sin incluir la retroalimentación por parte del experto. Por lo tanto, aunque se evalué la precisión de los modelos utilizados, esto no es suficiente para determinar si el conocimiento obtenido realmente puede aplicarse para tomar una decisión que ayude a solucionar un problema en especifico. En el caso de metodología SEMMA aunque se enfoque en el manejo de proyectos de minería de datos a nivel empresarial en escenarios prácticos, terminan su ciclo de vida en la evaluación de datos sin tener en cuenta la retroalimentación continua del negocio.

Por otra parte, aunque muchos investigadores prefieran ejecutar proyectos de ciencia de datos con metodologías tradicionales como CRISP-DM, KDD o SEMMA, otros investigadores consideran que aspectos como la organización y el trabajo en equipo también hacen parte de las fases metodológicas para el análisis de datos. Pongamos por caso la metodología \textit{RAMSYS}, la cual permite llevar a cabo proyectos de minería de datos a distancia de manera colaborativa. Esta metodología parte de la definición del problema y de la claridad de los objetivos del proyecto para todos los participantes, pero ademas de esto permite que sin importar la localización geográfica de los integrantes del equipo, todos colaboren de forma activa en un problema común. RAMSYS no limita la innovación de los participantes para proponer soluciones y propone que se comparta todo el conocimiento obtenido de manera efectiva por medio de una herramienta web. Asimismo, esta metodología también abarca la seguridad de los datos, considerando que es información sensible, que no debe revelarse fuera del proyecto. Por consiguiente, esta metodología ademas de tratar de encontrar conocimiento en los datos, permite la organización de equipos a distancia generando soluciones innovadoras gracias a la libertad del equipo para la solucionar problemas y manteniendo la seguridad de los datos a través del control de acceso a la información.