CRISP-DM Twenty Years Later: From Data Mining Processes to Data Science Trajectories

Fernando Martínez-Plumed, Lidia Contreras-Ochando, Cèsar Ferri, José Hernández-Orallo, Meelis Kull, Nicolas Lachiche, María José Ramírez-Quintana, and Peter Flach

Abstract—CRISP-DM(CRoss-Industry Standard Process for Data Mining) has its origins in the second half of the nineties and is thus about two decades old. According to many surveys and user polls it is still the *de facto* standard for developing data mining and knowledge discovery projects. However, undoubtedly the field has moved on considerably in twenty years, with *data science* now the leading term being favoured over *data mining*. In this paper we investigate whether, and in what contexts, CRISP-DM is still fit for purpose for data science projects. We argue that if the project is goal-directed and process-driven the process model view still largely holds. On the other hand, when data science projects become more exploratory the paths that the project can take become more varied, and a more flexible model is called for. We suggest what the outlines of such a trajectory-based model might look like and how it can be used to categorise data science projects (goal-directed, exploratory or data management). We examine seven real-life exemplars where exploratory activities play an important role and compare them against 51 use cases extracted from the NIST Big Data Public Working Group. We anticipate this categorisation can help project planning in terms of time and cost characteristics.

Index Terms—Data science trajectories, data mining, knowledge discovery process, data-driven methodologies