Metodologias de Gerenciamento de Projetos de Ciência de Dados

Diego Machado de Assis*

15 de maio de 2021

Resumo

Metodologias de gerenciamento de projetos de Ciência de Dados podem ser grande aliadas no desenvolvimento de projetos robustos, eficazes e eficientes na área. Diversos padrões existem e atendem a uma grande heterogeneidade de necessidades e requisitos. Alguns destes padrões são mais amplamente utilizados e consolidados, tornando-se padrões *de facto* na indústria. Este trabalho apresenta um breve resumo e traça um paralelo entre três das principais metodologias existentes hoje: *KDD*, *CRISP-DM* e *SEMMA*.

Palavras-chave: Ciência de dados. Gerenciamento de projetos. Mineração de dados. KDD. CRISP-DM. SEMMA.

Data de submissão: 06 de junho de 2021

1 Introdução

Métodos sistemáticos para extração de conhecimento a partir de dados ajudam a organizar de forma estruturada o desenvolvimento e avaliação de respostas para problemas de análise de dados (PROVOST; FAWCETT, 2013). Conhecer as principais técnicas e etapas em um projeto de Ciência de Dados é fundamental para maior acurácia dos resultados, ao mesmo tempo que o desenvolvimento seja o mais eficiente e ágil possível.

Cada projeto possui suas particularidades e não existe metodologia ubíqua que abstraia por completo as necessidades concretas de cada caso. Dessa forma, conhecer os principais métodos e ferramentas disponíveis é fundamental para composição do arsenal de um cientista de dados. Mais importante do que ranquear os processos é conseguir entender as semelhanças e diferenças entre eles, assim como identificar que tipo de problemas cada um se propõe a atacar.

Neste trabalho procuramos definir algumas das metodologias mais utilizadas hoje e traçar um breve comparativo entre elas. Na seção 2 apresentamos as metodologias *KDD*, *CRISP-DM* e *SEMMA*. Na seção 3 fazemos um paralelo entre as principais diferenças entre elas e na seção 4 tentamos consolidar as ideias apresentadas de cada uma e olhar qual nos seria mais adequada em um projeto de Ciência de Dados.

* CP301343X diego.assis@aluno.ifsp.edu.br

2 Metodologias de Gerenciamento de Projetos

2.1 KDD

O processo de Extrair de Conhecimento a partir de Dados (ou *Knowledge Discovery from Data - KDD*) muitas vezes se confunde com o próprio conceito de Mineração de Dados (ou *Data Mining*). O processo de KDD consiste em uma sequência iterativa das seguintes etapas (HAN; KAMBER, 2006):

1. Limpeza dos dados
2. Integração dos dados
3. Seleção dos dados
4. Transformação dos dados
5. Mineração dos dados
6. Avaliação dos padrões
7. Apresentação dos resultados

A **Limpeza dos dados** consiste na remoção de ruídos e dados inconsistentes (*outliers*). Nesta fase deve ser definida a estratégia para tratamento de dados faltantes, assim como é feito o mapeamento dos atributos com os seus respectivos tipos.

A **Integração dos dados** consiste na combinação de mútlipas fontes de dados em um formato de armazenamento coerente e o salvamento dos mesmos em uma fonte de fácil acesso (por exemplo um *data warehouse*).

Na etapa de **Seleção de dados**, é realizada a seleção apenas dos dados relevantes para os objetivos de análise pretendidos.

A **Transformação dos dados** objetiva consolidar os dados em um formato apropriado para a etapa de mineração. Para tanto, podem ser realizadas sumarizações, agregações, generalizações e normalizações do conjuto de dados. Em alguns casos, esta etapa pode vir acompanhada de **Redução dos dados**, para que se tenha uma representação mais suscinta dos dados originais, com objetivo de ganhar eficiência sem comprometer a integridade da informação.

Estas 4 primeiras etapas podem ser entendidas em conjunto como uma etapa de **Pré-processamento dos dados**, em que os dados são preparados para a etapa de mineração.

A **Mineração dos dados** é a principal etapa do processo, em que métodos de *inteligência* são aplicados para extração de padrões. Em outras palavras, podemos dizer que trata-se do processo de gerar informação útil a partir de um grande volume de dados, utilizando diferentes técnicas tais como *classificação*, *regressão*, *clusterização*, *modelagem sequencial*, *análise linear* (CHEHAB, 2020). A mineração de dados consiste em produzir modelos, ajustá-los e determinar os padrões observados nos dados.

A **Avaliação dos padrões** é o processo de interpretar os resultados obtidos com a aplicações das técnicas de mineração de dados nos modelos produzidos. Esta etapa identifica quais dos padrões extraídos na etapa anterior são de real interesse e podem

representar uma forma de *conhecimento*, de acordo com as medidas objetivadas. Neste ponto, o foco é está no quão útil e compreensível são os modelos produzidos (CHEHAB, 2020).

A **Apresentação dos resultados** é a etapa final do processo de KDD. Neste ponto, técnicas de visualização de dados, *storytelling* e representação de dados podem ser usadas para apresentar as descobertas para os usuários. O conhecimento adquirido pode também ser utilizado como entrada de outros sistemas para processamentos e ações adicionais.

2.2 CRISP-DM

Cross-industry standard process for data mining (CRISP-DM) é um padrão que define abordagens para problemas de análise de dados. Ele foi concebido em 1996, e a primeira versão foi apresentada no 4^{th} CRISPDM SIG Workshop em Bruxelas, na Bélgica, em março de 1999 (WIKIPEDIA, 2021a).

Segundo Chapman (1999), os objetivos do CRISP-DM são:

- Garantir a qualidade dos resultados de projetos de mineração de dados
- Reduzir as habilidades necessárias para a extração de conhecimento
- Reduzir custos e tempo

E seus benefícios podem ser listados:

- De propósito geral, ou seja, independente de aplicações
- Robusto a mudanças do ambiente
- Indendente de ferramentas e técnicas
- Suporte para ferramentas
- Suporte para documentação dos projetos
- Salvaguarda experiência para resumo
- Suporte para transferência de conhecimento e treinamento

O processo CRISP-DM pode ser esquematizado de acordo com o diagrama na figura 1. O diagrama deixa claro que o processo é bastante iterativo, e cada etapa é descrita a seguir.

O **Entendimento do negócio** é a primeira etapa e trata-se de conhecer o problema o qual se quer resolver. Deve-se determinar os objetivos do negócio, os objetivos da mineração de dados, os critérios de avaliação de sucesso, as limitações, os riscos e contigências, os custos e benefícios e os recursos disponíveis. Nesta etapa é produzido um plano do projeto e uma avaliação inicial de ferramentas e técnicas necessárias.

O **Entendimento dos dados** consiste na avaliação dos *dados crus* a partir dos quais será feita a análise. É importante que seja bem compreendido o potencial e as limitações dos dados disponíveis e se serão necessários esforços para coleta de novos dados.

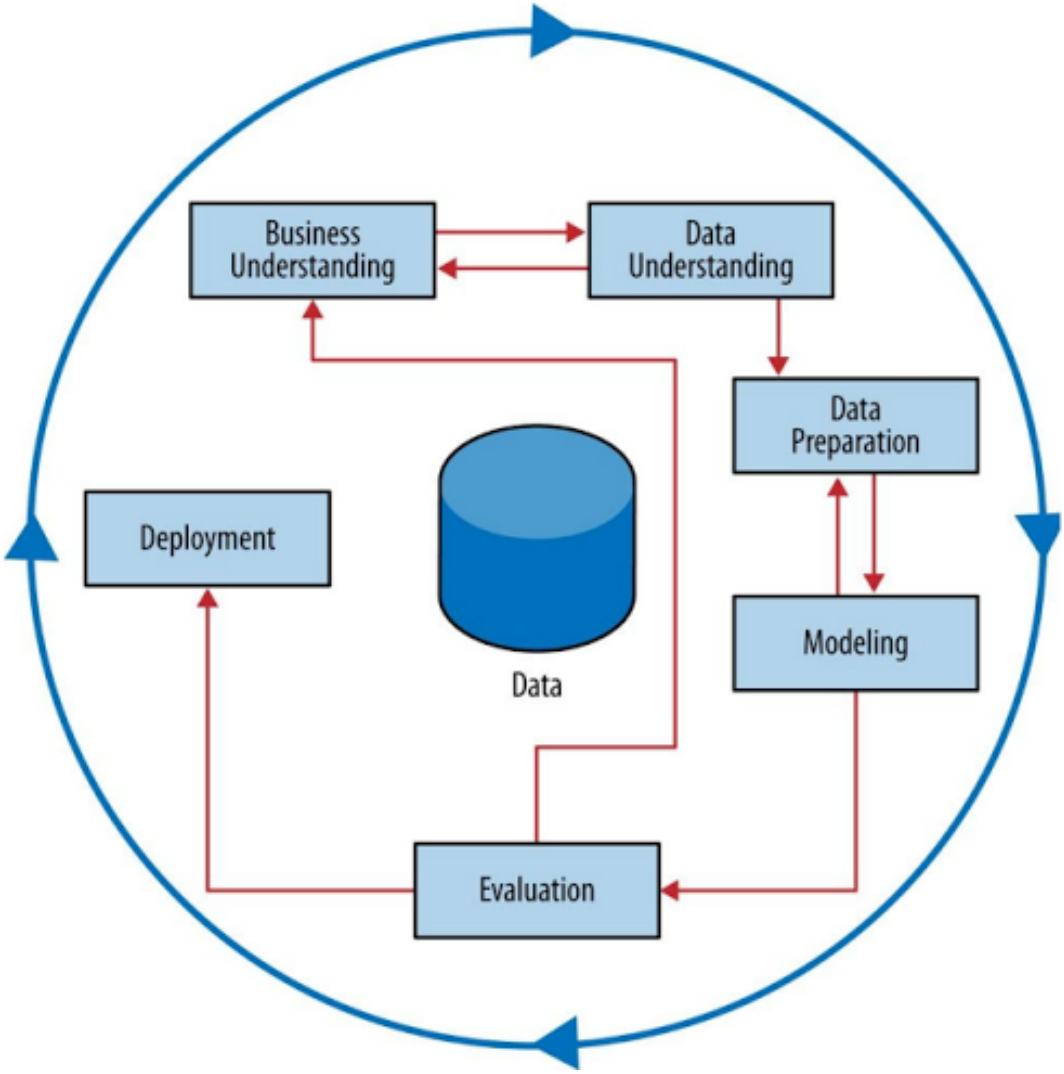

Figura 1 – Fases do CRISP-DM

Esta etapa pode ser entendida em subfases, consistindo em *Coleta inicial dos dados*, *Descrição dos dados*, *Exploração dos dados* e *Verificação da qualidade dos dados*.

Estas duas primeiras etapas são bastante iterativas entre si, implicando que um melhor entendimento de cada uma delas pode levar a um aperfeiçoamento da outra. Dessa forma, é comum que várias iterações sejam feitas entre elas antes de se passar para a próxima etapa.

A **Preparação dos dados** consiste em converter os dados disponíveis para os formatos esperados pelas ferramentas de processamento e análise que serão utilizadas. Tarefas comuns desta etapa incluem limpeza dos dados (tratamento de *outliers* e valores faltantes), conversão de tipos de dados, seleção de dados, identificação de atributos derivados, integração de diferentes conjuntos de dados e formatação dos dados.

Na fase de **Modelagem** são aplicadas as técnicas de mineração de dados para geração de modelos e identificação de padrões nos dados. Comumente, durante esta etapa é necessário voltar na fase de preparação dos dados para adequar os dados disponíveis para

os formatos esperados.

Na etapa de **Avaliação** os modelos e padrões extraídos na etapa anterior são avaliados de acordo com os objetivos do negócio, definidos na primeira fase. A partir desta etapa é possível ter uma ideia geral de todo o processo e os resultados alcançados e, sendo necessário, é o momento de realizar uma nova iteração do ciclo, voltando à etapa de entendimento do negócio. Neste ponto é possível também determinar os próximos passos, ou seja, as possívei ações e decisões que podem decorrer do processo.

A fase de **Implantação** consiste em colocar em uso os modelos gerados na mineração de dados. Deve ser elaborado um plano de implantação, que muitas vezes inclui otimizações e adequações dos modelos para o ambiente de produção. É necessário também que já se planeje como será feito o monitoramento da solução em produção, assim como sua manutenção. Por fim, pode ser salutar que o projeto como um todo seja revisto e as experiências acumuladas sejam documentadas.

2.3 SEMMA

O acrônimo para *Sample, Explore, Modify, Model, Assess - SEMMA* define um processo para desenvolvimento de projetos de mineração de dados criado pelo SAS Institute (uma tradicional empresa desenvolvedora de soluções de *software* analíticos e estatísticos). As cinco etapas definidas pelo padrão são as seguintes (WIKIPEDIA, 2021b; AZEVEDO; SANTOS, 2008):

A primeira etapa, de **Amostragem**, consiste em selecionar o conjunto de dados a ser modelado. O conjunto precisa ser grande o suficiente para conter informações relevantes a serem extraídas, porém pequeno o suficiente para ser usado de forma eficiente. Esta etapa também cuida do particionamento dos dados. De acordo com Azevedo e Santos (2008), esta etapa é opcional.

A etapa seguinte, de **Exploração**, cuida do entendimento dos dados, olhando para a relações entre as variáveis e buscando tendências e anomalias, com ajuda de mecanismos de visualização de dados.

A etapa de **Modificação** prepara os dados para modelagem através de métodos de seleção, criação e transformação de variáveis.

A **Modelagem** aplica técnicas de mineração de dados para criação de modelos que que combinem os dados para levar aos resultados desejados.

Por fim, a fase de **Avaliação** compara os resultados dos modelos de forma a mensurar a confiabilidade, utilidade e performance dos mesmos.

3 Principais diferenças

Primeiramente vale ressaltar que as três metodologias possuem objetivos semelhantes e, dessa forma, possuem diversos elementos em comum entre elas. De uma certa maneira, todas elas implementam subdivisões em três principais blocos de tarefas bem definidos: *Preparação*, *Modelagem* e *Avaliação*.

Mais do que um método bem definido, *KDD* é um conjunto de conceitos que, juntos, levam a uma extração de conhecimento a partir de dados. Não existe uma especificação oficial e formal de suas etapas, sendo suas práticas apresentadas de formas ligeiramente diferentes por cada autor.

Já *CRISP-DM* é sim uma metodologia bem definida, com especificação documentada formalmente, padronização de etapas e nomenclatura. Conta a seu favor também o fato de ser um padrão aberto, adotado pela União Europeia (WIKIPEDIA, 2021a), e com participação de diversas companhias. Outra vantagem do modelo é uma melhor especificação dos seus caminhos de iteração, que ajudam a visualizar quais ciclos podem ser traçados em cada etapa do projeto. Esta esquemática não é especificada pelos outros dois modelos.

Por sua vez, *SEMMA* também é uma especificação melhor definida, mas tem a particularidade de ser criada por uma só empresa que, de certa forma, adequa suas especificações às particularidades das ferramentas produzidas por ela própria.

Podemos dizer que *CRISP-DM* é mais ampla que as outras metodologias por especificar as etapas de *Entendimento do negócio* e *Implantação* (apesar de que na *KDD*, a etapa final de *Apresentação dos resultados* possa, em alguns aspectos, se assemelhar a esta etapa de implantação da *CRISP-DM*). Podemos dizer que as etapas da *SEMMA* seriam mais "técnicas" (focadas na mineração de dados), por não considerarem essas etapas de início e fim que estariam mais ligadas ao gerenciamento de projeto.

4 Considerações finais

As metodologias apresentadas se propõem a objetivos específicos, porém possuem suas particularidades que tornam uma ou outra mais adequada para determinadas situações.

KDD se aplica como um conjunto mais geral de técnicas e boas práticas, sendo ideal para cientistas de dados experientes, que conhecem bem as etapas de mineração de dados e necessitem de um *framework* flexível e modular que possa se adequar a necessidade de cada projeto em particular.

CRISP-DM é um conjunto bem especificado e mais rígido de regras e etapas, aplicáveis por cientistas novatos ou experientes para qualquer tamanho ou complexidade de projeto. É o *framework* mais completo, detalhado e robusto mas, por isso mesmo, as vezes mais complexo do que as necessidades específicas de um projeto.

SEMMA é um *framework* bem especificado e também bastante adequado para cientistas pouco experientes. É mais focado nas etapas técnicas do processo e, por isso, mais adequado para projetos que já tenham uma boa especificação de negócio e ambiente bem definido de homologação e produção. Apesar de servir para uso geral, se aplica melhor se combinado com a utilização das ferramentas da mesma empresa que o desenvolveu.

De uma forma geral, se eu fosse inicar um projeto de Ciência de Dados hoje, tenderia a utilizar o *CRISP-DM*, principalmente pela minha pouca experiência em projetos desse tipo. Nesta situação, um *framework* mais robusto e detalhado poderia ser um melhor guia para organização das tarefas.

Referências

AZEVEDO, A.; SANTOS, M. F. Kdd, semma and crisp-dm: a parallel overview. In: *IADIS European Conf. Data Mining*. [S.l.: s.n.], 2008. Citado na página 5.

CHAPMAN, P. *The CRISP-DM User Guide*. 1999. Disponível em: <https://s2.smu.edu/~mhd/8331f03/crisp.pdf>. Acesso em: 30 mai 2021. Citado na página 3.

CHEHAB, M. *Knowledge Discovery Data (KDD)*. 2020. Analytics Vidhya. Disponível em: <https://medium.com/analytics-vidhya/knowledge-discovery-data-kdd-a8b41509bff9>. Acesso em: 29 mai 2021. Citado 2 vezes nas páginas 2 e 3.

HAN, J.; KAMBER, M. *Data Mining: Concepts and Techniques*. San Francisco, CA, USA: Morgan Kaufmann Publishers, 2006. Citado na página 2.

PROVOST, F.; FAWCETT, T. *Data Science for Business*. Sebastopol, CA, USA: O'Reilly, 2013. Citado na página 1.

WIKIPEDIA. *Cross-industry standard process for data mining*. 2021. Wikipedia: the free encyclopedia. Disponível em: <https://en.wikipedia.org/wiki/Cross-industry_standard_process_for_data_mining>. Acesso em: 30 mai 2021. Citado 2 vezes nas páginas 3 e 6.

WIKIPEDIA. *SEMMA*. 2021. Wikipedia: the free encyclopedia. Disponível em: <https://en.wikipedia.org/wiki/SEMMA>. Acesso em: 30 mai 2021. Citado na página 5.