Universidade do Estado do Amazonas
Data: 18 de setembro de 2020
Disciplina: Fundamentos da Ciência de Dados
Professor(a): Carlos Maurício Serodio Figueiredo
Alunos: Lucas Pereira Reis

Atividade Avaliativa #2

Nesta atividade será realizado um *overview* de diferentes metodologias direcionadas ao processo de *Data Science* e ao final, uma comparação entre cada um.

KDD

A metodologia KDD (*Knowledge Discovery in Databases*) consiste na utilização de métodos para minerar, extrair e descobrir conhecimentos à partir de dados presentes em um banco de dados. É uma das metodologias mais antigas, sua estrutura foi proposta em meados de 1980. A Figura 1 exemplifica o passo a passo deste processo.

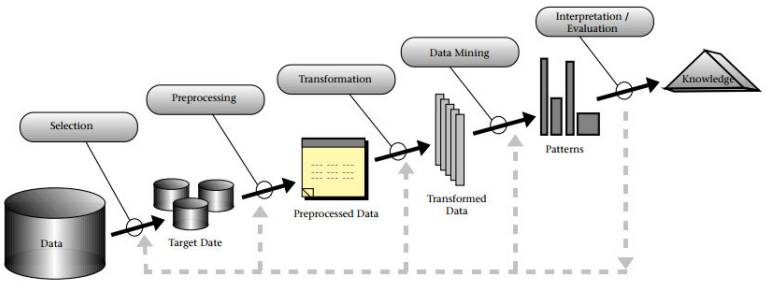

Figura 1: Os estágios do KDD.

A etapa de Seleção (*Selection*) consiste em criar um *dataset* ou dados de amostras para serem analisados nas etapas posteriores. O Pré-processamento (*Pre-processing*) realiza a limpeza, correção e o pré-processamento em si para obter um *dataset* consistente. A Transformação (*Transformation*) utiliza métodos de redução de dimensionalidade e transformações para obter dados mais densos. A etapa de Mineração de Dados (*Data Mining*) consiste em procurar e separar padrões de interesse em uma representação específica dos dados conforme o objetivo da análise. A última etapa, de Interpretação/Avaliação (*Interpretation/Evaluation*), avalia os padrões e análises obtidas para então de fato extrair o conhecimento a partir delas.

CRISP-DM

A metodologia CRISP-DM (*Cross-Industry Standard Process for Data Mining*) consiste em um ciclo com seis estágios cruciais. Esta metodologia é extremamente estruturada

e documentada, permanecendo no topo como metodologia mais utilizada. A Figura 2 representa o seu ciclo, seguido de uma breve descrição sobre cada um.

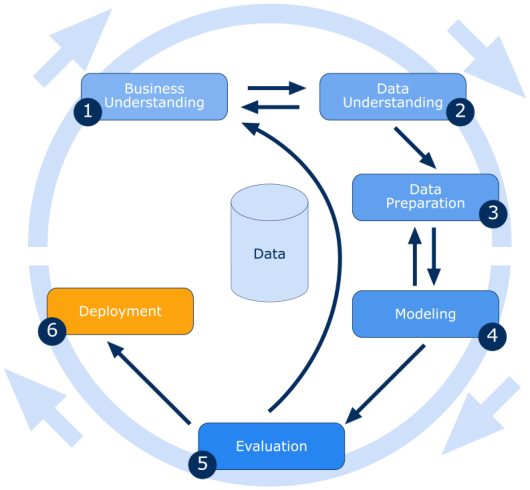

Figura 2: O ciclo de vida do CRISP-DM.

O estágio *Business Understanding* é a fase inicial focada em entender o objetivo do projeto, os requisitos por uma perspectiva de negócio e definindo qual será a métrica que ditará se seu projeto atingiu o sucesso ou não. *Data Understanding* começa com uma coleta de dados e procede em análises para descrever, explorar e verificar possíveis *insights* iniciais para seus dados. *Data Preparation* contempla as etapas de processamento de dados (como *data selection*, *data cleaning*) para produzir o *dataset* final. A próxima fase é responsável pela construção do modelo consistindo na aplicação de diferentes algoritmos para criar o modelo em si e tunar seus parâmetros, normalmente é produzido diferentes modelos para compará-los.

Após a etapa de criação dos modelos, é necessário avaliar a solução proposta. As métricas utilizadas são as mesmas especificadas na primeira etapa, caso o critério de sucesso não for atingido, é necessário voltar a primeira fase para revisar as regras de negócio levantadas. A última etapa, o *deployment*, tem como objetivo colocar o melhor modelo obtido em produção, para que possa ser utilizado e apresentado para os clientes.

SEMMA

A metodologia SEMMA (*Sample, Explore, Modify, Model, Assess*), criada pelo SAS Institute, possui um ciclo de cinco estágios, sua estrutura é semelhante da CRISP-DM mas focando principalmente nas tarefas de criação dos modelos. A Figura 3 apresenta suas etapas.

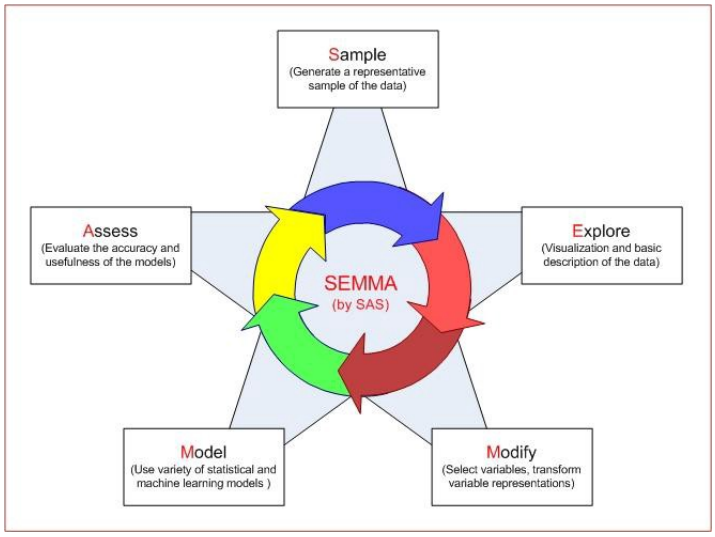

Figura 3: O ciclo de vida do SEMMA.

O estágio *Sample* consiste em obter um amostra de dados de um *dataset* maior que seja grande o suficiente para conter informações significativas, mas que ainda seja possível manipular com facilidade. O segundo estágio, *Explore*, realiza a exploração em busca procurando por tendências e anomalias para obter conhecimento e compreensão dos dados. A etapa de *Modify* consiste na modificação dos dados criando, selecionando e transformando o *dataset* preparando sua modelagem. Em *Modify*, é utilizado algoritmos para criação de modelos para gerar o resultado desejado. Por fim, o estágio de *Assess* avalia os dados obtidos observando a utilidade e a confiabilidade dos resultados obtidos.

ASUM-DM

A metodologia ASUM-DM (*Analytics Solutions Unified Method for Data Mining*) foi desenvolvida pela IBM atuando como uma versão refinada da CRISP-DM possuindo também cinco fases, além de utilizar princípios ágeis. A Figura 4 apresenta sua estrutura.

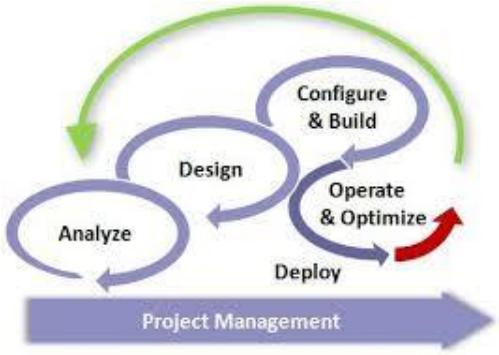

Figura 4: As fases do ASUM-DM.

Sua primeira fase, de *Analyze*, define o que a solução precisa realizar, tanto em termo de *features* quanto em atributos não funcionais (performance, usabilidade, etc), nesta etapa também é realizado a coleta e exploração dos dados para compreendê-los. Em

Design é definido todos os componentes da solução e seus dependentes e seus ambientes de desenvolvimento, teste e produção. Em *Configure Build* é realizado todo o pré-processamento de dados necessário preparando o *dataset* para a construção de diferentes modelos que serão avaliados. Devido aos princípios desta metodologias, estas três fases possuem uma natureza iterativa podendo ser reavaliadas e refeitas por diversas *sprints*.

Na fase de *Deploy*, inicialmente é criado uma estratégia para executar e manter a solução no ambiente de produção, além de realizar qualquer configuração necessária. Após isto, é iniciado a etapa de *Operate & Optimize* para monitorar a aplicação a fim de buscar otimizações e melhorias para o sistema. Além de todas as etapas descritas, é executado em paralelo o *Project Management* que consiste em processos que auxiliam com administração e monitoramento do progresso e manutenção do projeto.

Comparação entre metodologias

Todas as metodologias apresentadas possuem seus processos, estruturas e objetivos bem definidos, algumas buscam abordar um escopo menor enquanto outras tentam abranger todo o ciclo de vida.

É interessante observar que a KDD e a SEMMA apresentam fases extremamente semelhantes, pode-se até dizer que a SEMMA é uma versão mais refinada que a KDD, mas com os mesmos princípios. Estas metodologias são válidas para projetos onde não é necessário ter um entendimento profundo das regras de negócio ou enviar o modelo para a produção.

Embora a CRISP-DM apresente fases semelhantes com as duas metodologias anteriores, esta possui etapas adicionais, focando justamente nos estágios que não estão presentes nas anteriores, de *Business Understanding* e *Deployment*. Atualmente é a metodologia mais popular e a mais utilizada.

A ASUM-DM apresenta de fato uma estrutura mais refinada que as outras metodologias, incluindo também o estágio de *Operate & Optimize* para monitoramento do modelo. Porém seu uso ainda não é tão difundido, provavelmente devido a ser mais voltada para os serviços da própria IBM.

Em geral quando estamos desenvolvendo projetos pessoais de Data Science, normalmente seguimos um *pipeline* semelhante ao KDD e SEMMA pois acaba sendo um padrão difundido., dessa forma estas duas metodologias são úteis para aplicações e projetos mais simples. Quando estamos em um projeto profissional e que precisamos entregar de fato uma solução bem documentada e funcional, a CRISP-DM se torna a metodologia mais eficiente para esta situação, e se for necessário realizar um monitoramento contínuo desta solução, torna-se viável aplicar conceitos visto na ASUM-DM ou até mesmo implementar sua metodologia por completo.