Universidade do Estado do Amazonas
Escola Superior de Tecnologia
Data: 20 de setembro de 2020
Disciplina: Fundamentos da Ciência de Dados
Professor: Carlos Maurício S. Figueiredo
Aluno: Marcos Wenneton Vieira de Araújo

Atividade Avaliativa II
Metodologias para projetos de Ciência de Dados

1 Cross Industry Standard Process for Data Mining – Crisp-DM

Essa técnica é uma das mais utilizadas em Mineração de Dados. Suas principal vantagem é que ela pode ser aplicada a qualquer tipo de negócio, e não tem dependência de ferramenta para ser executada. Abaixo estão listadas as fases desta metodologia:

- **Business Understanding**: Na primeira fase da CRISP-DM, deve-se identificar o problema negocial que quer resolver. Como insumo dessa fase, você deve entregar três coisas: um background, o objetivo do projeto e o critério de sucesso;
- **Data Understanding**: Coletar e tratar o dado é uma tarefa responsável por mais de 70% do tempo gasto em um projeto por cientistas de dados, e é exatamente sobre isso que essa fase e a próxima dizem respeito. Aqui, deverá acontecer a coleta, descrição — usando estatísticas — , exploração e verificação da qualidade do dado.
- **Data Preparation**: Nessa fase os dados serão preparados para a modelagem, que ocorre na próxima etapa. Esse processo consiste, principalmente, de quatro tarefas:
  - **Data Selection**: seleção dos dados que serão utilizados no modelo. Por exemplo, talvez você não seja necessário utilizar *outliers*, ou todas as colunas da tabela. Aqui é feita a escolha de tudo que será relevante para o modelo, e deve existir uma documentação com o motivo de escolhê-los;
  - **Data Cleaning**: pode ser que o dado não venha formatado corretamente. Datas em formato incorreto e números inteiros sendo interpretados como string, etc. É nesse passo que tudo isso irá ser tratado;
  - **Construct Data**: talvez nem todos os dados disponíveis esteja a disposição. É possível que seja necessária a criação de novos dados para o futuro modelo;
  - **Integrating Data**: essa tarefa é necessária quando é preciso juntar duas fontes de dados diferente;

- **Modeling**: É nesse momento que será realizada a construção do modelo. Essa fase consiste em escolher um algoritmo de Aprendizado de Máquina, criar o modelo em si e realizar um *fine-tuning* nos seus parâmetros. Vários modelos diferentes podem ser criados e comparados na próxima fase;
- **Evaluation**: Hora de avaliar os resultados dos modelos gerados e checar se os critérios de sucesso que foram definidos lá na primeira fase foram atingidos. Se não, é necessário voltar a primeira fase e entender o que deu de errado, determinar um novo escopo e tentar novamente.
- **Deployment**: Nesta fase final é hora de colocar o modelo em produção, para que possa ser usado. O *deployment* coloca fim ao projeto, mas é sempre bom monitorar os resultados e adaptar o modelo sempre que necessário.

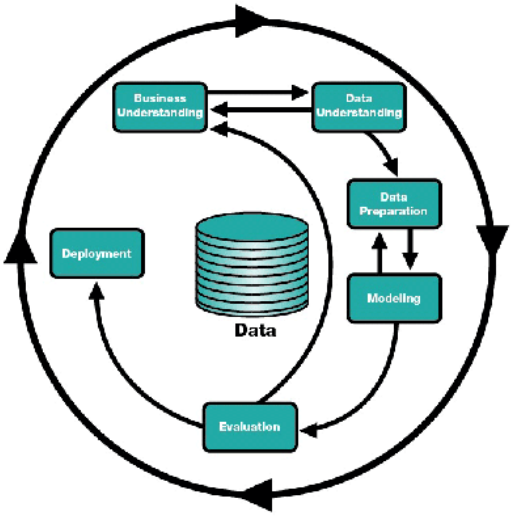

Figura 1: O fluxo de funcionamento da metodologia CRISP-DM

2 Knowledge Discovery in Databases – KDD

Esse é um dos métodos mais antigos existentes, tendo sua criação feita em 1980. Essa técnica é muito utilizada em Data Mining e, diferente de outras metodologias, não foca em questões de negócio ou geração de modelos, mas sim na descoberta de conhecimentos a partir dos dados.

As fases do KDD, mostradas na Figura 2, são resumidas a coleta dos dados que serão usados para análise; seu pré-processamento, que inclui limpeza e integração; transformação dos dados; e a interpretação e avaliação de padrões encontrados a partir de técnicas de Data Mining e Análise Exploratória.

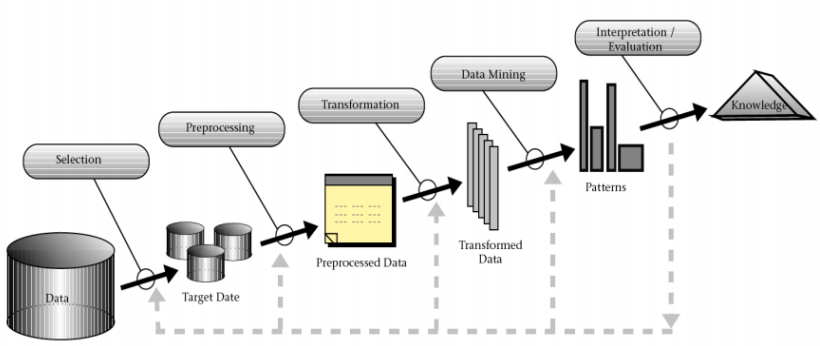

Figura 2: Descoberta de Conhecimento em Bancos de Dados utilizando KDD

3 Sample, Explore, Modify, Model e Assess – SEMMA

Uma outra metodologia muito utilizada é a SEMMA, criada pela SAS Institute. Ela é parecida com a CRISP-DM em muitos aspectos, mas, ela foca principalmente nas tarefas de criação do modelo, deixando as questões de negócio de fora.

Suas tarefas são muito parecidas com as da CRISP, como: explorar informações básicas dos dados, modificar e transformar variáveis, gerar o modelo e validá-lo. Um bom cenário para a aplicação desta metodologia pode ser: um projeto pequeno, que não tenha um forte impacto a nível de negócio. A Figura 3 mostra o fluxo de funcionamento da metodologia.

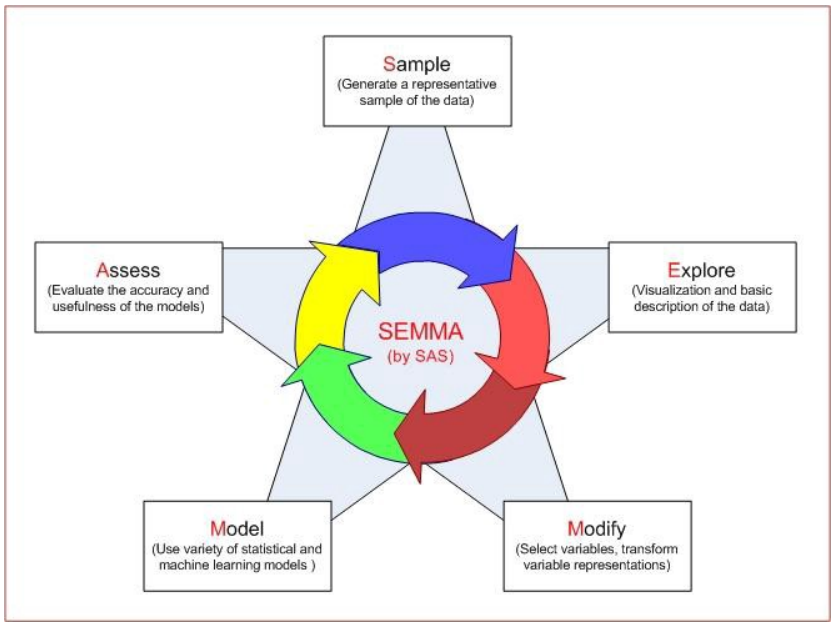

Figura 3: A metodologia SEMMA

4 Analytics Solutions Unified Method for Data Mining – ASUM-DM

O ASUM-DM é um guia passo-a-passo para conduzir um ciclo de vida completo de implementação em análise de dados desenvolvido pela IBM. O método pode ser utilizado por clientes e parceiros de negócio da empresa para implementar soluções desenvolvidas no software chamado IBM Cognos Analytics.

Alguns de seus benefícios são o risco reduzido, a escalabilidade, a facilidade de compreensão de todas as etapas de desenvolvimento e roteiros de implementação específicos do produto. Abaixo encontram-se as fases abordadas durante esta técnica:

- **Análise**: Define o que a solução busca resolver, considerando os atributos não-funcionais (desempenho, usabilidade, etc). Nesta fase busca-se o consenso entre todas as partes sobre os requisitos;
- **Design**: Define todos os componentes da solução e suas dependências, identificação de recursos e instalação de ambientes de desenvolvimento;
- **Configuração e construção**: Configuração, construção e integração de componentes baseado numa abordagem incremental. Utiliza-se de um multi-ambiente de teste e um plano de validação baseado na metodologia ágil clássica conhecida como V-Model;
- **Deploy**: Nesta fase é criado um plano rodar e manter a solução, incluindo um cronograma de suporte;
- **Operação e Otimização**: Aqui é aplicado o uso da solução em IBM Cognos Analytics. Esta operação inclui a manutenção dos modelos criados e garante a saúde da aplicação;
- **Gerenciamento do projeto**: Esta fase ocorre durante todo o processo, e consiste na monitoração e gerenciamento do progresso e manutenção do projeto.

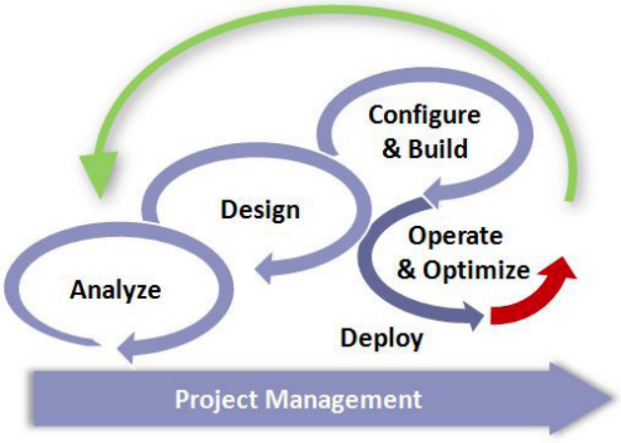

Figura 4: O método ASUM criado pela IBM