Tecnologias de Descoberta de Conhecimento
Lista de Exercícios I - Parte A

Pedro Augusto Duarte de Almeida

Setembro/2016

1 Considere uma base de dados com informações sobre a frequência acadêmica de alunos. Nesse contexto, aponte o que poderiam ser um dado, uma informação e um conhecimento.

1.1 Dados

Dados são códigos que constituem a matéria prima da informação, ou seja, é a informação não tratada. Os dados representam um ou mais significados que isoladamente não podem transmitir uma mensagem ou representar algum conhecimento.

Table 1: Dados

N	S	D	idD	CH
Igor	P	2016-09-29	2	68
Gabriel	A	2016-09-29	4	34
Pedro	P	2016-09-29	3	34

A tabela acima constitui apenas um conjunto de **Dados**, pois somente com ela não se pode estabelecer um contexto nem um significado. Podemos até imaginar alguns significados, mas não há como assegurar sua validade.

1.2 Informação:

Informações são dados tratados. O resultado do processamento de dados são as informações. As informações têm significado, podem ser tomadas decisões ou fazer afirmações considerando as informações. **Informação** é o sentido que um conjunto de dados tem para alguém. Um conjunto de Dados representa uma Informação, para uma pessoa, quando ela consegue perceber suas relações com outros Dados, que lhe definem um contexto e, ainda, com outros Dados e Informações que já lhe são familiares, estabelecendo assim seu significado para ela.

Informação é, portanto, a leitura que cada indivíduo faz de um conjunto de Dados, é o significado que lhe atribui ao internalizar esses Dados.

Table 2: Informação

Nome	Situação	Data	idDisciplina	CargaHoraria
Igor	P	2016-09-29	2	68
Gabriel	A	2016-09-29	4	34
Pedro	P	2016-09-29	3	34

Na tabela acima, se for esclarecido que se trata de um registro de frequência de alunos que especifica o nome do aluno, a data da chamada, a situação do aluno naquela data (P de Presente ou A de Ausente), e id da

disciplina; e se esses termos e conceitos fazem sentido, o usuário passa a entender o contexto e o significado dos Dados, que passam a constituir uma Informação. Para um usuário não habituado ao funcionamento de frequência de alunos numa universidade, os Dados podem fazer menos sentido, não transmitindo a mesma Informação.

1.3 Conhecimento:

É a capacidade, adquirida por alguém, de interpretar e operar sobre um conjunto de Informações. Essa capacidade é criada a partir das relações que ele estabelece sobre o conjunto de Informações, e desse conjunto com outros conjuntos que já lhe são familiares, que lhe permitem compreende-lo e tirar conclusões sobre ele e a partir dele.

Com referência à mesma tabela acima, se o usuário tem conhecimento sobre o funcionamento da frequência academia dos alunos, o mesmo é capaz de interpretar as informações com maior profundidade, por exemplo, deduzindo o número de faltas que um aluno pode ter para ser reprovado numa determinada disciplina, baseando-se na carga horária total da mesma juntamente com alguma regra de negócio que estipula a porcentagem máxima de faltas permitida pela instituição de ensino.

Table 3: Dados x Informação x Conhecimento

Dado	Informação	Conhecimento
Nome do Aluno	Presente ou Ausente; Data da Presença ou Ausência; Disciplina	Estudo do perfil do ano; Previsão de frequência até o final do semestre; Cálculo de faltas restantes permitidas

2 Considerando a plataforma Data Viva, faça:

2.1 Dê uma visão geral sobre os dados que podem ser encontrados nessa plataforma

Dados sobre as exportações, atividades econômicas locais, ocupações e educação em todo o Brasil.

2.2 Identifique pelo menos dois tipos de técnicas de data mining que poderiam ser aplica das nas bases de dados da plataforma

1. Descoberta de Associações: Associar a compra de um produto x com um produto y
2. Regressão: Predição da soma da biomassa presente em uma floresta; a predição do risco de determinados investimentos;

2.3 Aponte um potencial tipo de conhecimento que um poderia ser extraído dessa base de dados

- Números das profissões com as melhores médias salariais de cada estado;
- Melhores opções de investimento em cada estado;
- Tomar decisão sobre onde instalar uma startup e como integrá-la à cadeia de suprimentos das empresas locais;
- Nome e localização das universidades que mais possuem formandos por semestre.

3 Pesquisa no site Kaggle uma competição que chame sua atenção e descreva:

3.1 Objetivo da competição

Nesta competição, é disponibilizado participantes uma tabela com dados (como nome, idade, valor da passagem, etc) de 891 passageiros (dataset de treino) que estavam a bordo do Titanic e seu respectivo destino após o acidente (faleceu/sobreviveu). O objetivo é criar modelos que corretamente prevejam o destino dos 418 passageiros restantes (dataset de teste).

3.2 Os dados disponibilizados

Para cada passageiro:

1. Nome;
2. Classe;
3. Sexo;
4. Idade;
5. Número de irmãos(as)/esposos(as) à bordo;
6. Número de pais;
7. Número do ticket;
8. Tarifa;
9. Localização da cabine;
10. Porto de embarque
11. Indicador de destino final (0 = Não sobreviveu 1 = Sobreviveu)

3.3 Como você pensaria em resolver o problema apresentado na competição

1. Importar todos os dados fornecidos pela competição
2. Excluir informações que serão irrelevantes para minha análise, como o número do ticket, tarifa, e o porto de embarque
3. "Mulheres e crianças primeiro". Verificar a proporção de sobreviventes entre homens e mulheres. Filtrar os dados pelo sexo e a idade de cada passageiro
4. Em posse do conhecimento da proporção de sobreviventes entre homens e mulheres, cruzar com os dados dos passageiros com destino (sobreviveu ou não) desconhecido.

4 Faça uma análise comparativa entre as metodologias KDD e SEMMA. Nessa análise você deverá:

4.1 Descrever cada uma das metodologias e suas etapas

4.1.1 KDD

"Processo não trivial, de extração de informações implícitas, previamente desconhecidas e potencialmente úteis, a partir dos dados armazenados em um banco de dados"

O processo é não trivial já que alguma técnica de busca ou inferência é envolvida, ou seja, não é apenas um processo de computação direta. Os padrões descobertos devem ser válidos com algum grau de certeza, novos (para o sistema e de preferência também para o usuário), potencialmente úteis e compreensíveis.

O processo KDD, como foi apresentado por Fayyad, é o processo de usar métodos de Data Mining para extrair o que é considerado conhecimento em acordo com a especificação de metricas e limiares. São considerados 5 estágios:

1. Seleção: Esse estágio consiste em criar uma coleção de dados alvo, ou focar em um sub conjunto de variáveis ou dados simples, nos quais a descoberta e feita.

2. Pré Processamento: Esse estágio consiste na limpeza dos dados alvo e o seu pre processamento visando obter dados consistentes.

3. Transformação: Esse estágio consiste na transofrmação dos dados reduzindo as suas dimensões ou métodos de transformação.

4. Data Mining: Esse estágio consiste na busca por padrões de interesse numa particular forma representacional, dependendo do objetivo do data mining.

5. Interpretação/Avaliação: Esse estágio consiste na interpretação e avaliação dos padrões mineirados.

O processo KDD é interativo e iterativo, envolvendo vários passos com várias decisões sendo feitas pelo usuário. Adicionalmente, o processo KDD deve proceder pelo desenvolvimento do conhecimento do dominio da aplicação, o conhecimento relevante prévio e os objetivos do usuário final. Também deve ser continuada pela consolidação de conhecimento encorporando esse conhecimento no sistema.

4.1.2 SEMMA

SEMMA desenvolvido pela SAS Institute, refere-se ao processo de conduzir um projeto de Data Mining. A SAS Institute considera um ciclo com 5 estágios para o processo:

1. Amostragem: Esse estágio consiste na amostragem dos dados extraindo uma porção de uma grande coleção de dados, suficientemente grande para conter a informação significante, pequeno o suficiente para manipular rapidamente. Esse estágio é apontado como sendo opcional.

2. Explorar: Esse estágio consiste na exploração dos dados, buscando por tendencias não antecipadas e anomalias, de forma a ganhar entendimento e ideias.

3. Modificar: Esse estágio consiste na modificação dos dados, criando, selecionando e transformando as variáveis para focar no processso de seleção do modelo.

4. Modelo: Esse estágio consiste na modelagem dos dados permitindo o software a buscar automaticamente por uma combinação de dados que fielmente prevê um resultado.

5. Avaliar: Esse estágio consiste na avaliação dos dados, avaliando a usabilidade e confiabilidade dos resultados do processod e Data Mining e a estimativa de quão bem é executado.

Embora o processo SEMMA seja independente da ferramente de Data Mining escolhida, o SEMMA e relacionado ao software SAS Enterprise Minere pretende guiar o usuário na implementação de aplicações de Data Mining. SEMMA oferece uma forma fácil de entender processo, permitindo o desenvolvimento e a manutenção de forma organizada e adequada em projetos de Data Mining.

4.2 Apresentar as principais vantagens de uma em relação a outra

Comparando os estágios do KDD com os do SEMMA, em primeira instância, afirmamos que são equivalentes:

- Sample (Amostra) pode ser identificada com Selection (Seleção) Explore (Explorar) pode ser identificado com Pre processing (Pré processamento)
- Modify (Modificar) pode ser identificado com Transformation (Transformação)
- Model (Modelo) pode ser identificado com Data Mining (Mineração de Dados)
- Assess (Avaliação) pode ser identificado com Interpretation/Evaluation (Interpretação/Avaliação)

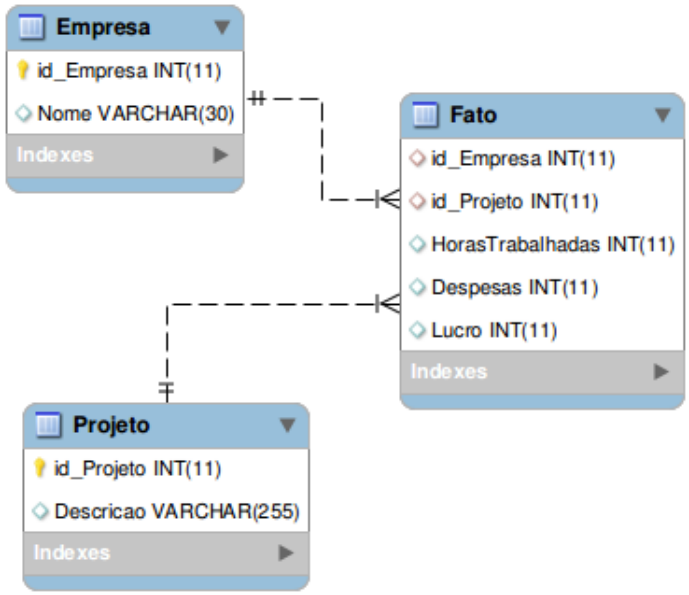
Examinando como um todo, podemos afirmar que os cinco estágios do processo SEMMA podem ser vistos como uma implementação pratica dos cinco estágios do processo KDD, desde que o SEMMA é diretamente ligado ao software SAS Enterprise Miner.

Table 4: KDD x SEMMA

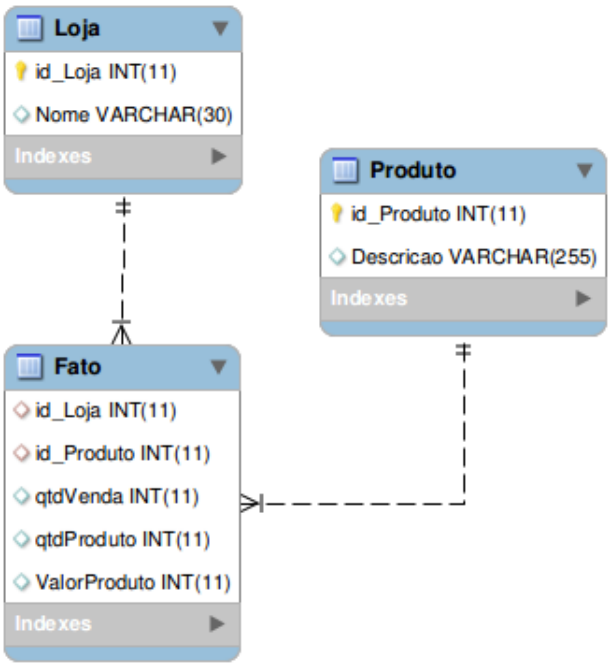
KDD	SEMMA
Selection	Sample
Pre Processing	Explore
Transformation	Modify
Data Mining	Model
Interpretation/Evaluation	Assessment

5 Resolva pelo menos 3 questões entre os exercícios propostos (1 a 5) no final capítulo 8 do livro

5.1 Questão 1:



5.2 Questão 2:



5.3 Questão 4: