1 Método

O capítulo todo é um resumo do CRISP-DM 1.0... fico dando
cite em todo parágrafo? o mesmo para o capítulo sobre aprendizado de máquina

Por que um processo?

(DOMINGOS, 2012) (DRUMMOND, 2008) (SCULLEY et al., 2014) (DRUMMOND, 2009)

O processo de aplicação de técnicas de Mineração de Dados para atender a um problema real demanda não apenas conhecimentos e execução de atividades relacionadas, estritamente, a aprendizado de máquina, mas também conhecimentos e execução de atividades relacionadas ao entendimento do problema em questão. A fim de organizar as atividades envolvidas nesse processo, foram desenvolvidas especificações de processos que orientam quais os passos devem ser seguidos para partir do entendimento de um problema e chegar a uma solução baseada em Mineração de Dados, como o CRISP-DM, o KDD e o SEMMA(AZEVEDO, 2008).

(AZEVEDO, 2008)

por que crisp-dm

1.1 Processo CRISP-DM

CRISP-DM(CRoss-Industry Standard Process for Data Mining) é um processo de aplicação de Mineração de Dados, desenvolvido pelo CRISP-DM Special Interest Group e publicado em 2000. Foi concebido em 1996 por três empresas que utilizavam Mineração de Dados: DaimlerChrysler(à época Daimler-Benz), SPSS(à época ISL) e NCR; motivadas pela incerteza com relação à qualidade de seus trabalhos, pelo questionamento de se toda nova empresa que quisesse aplicar Mineração de Dados teria que passar pelo aprendizado pelo qual passaram, baseado em tentativa e erro, e como garantirem, para seus clientes, que Mineração de Dados era uma área suficientemente madura para ser incorporada a seus processos de negócio. Em 1999 foi publicado um draft do CRISP-DM versão 1.0, sendo aplicado pela DaimlerChrysler, SPSS e NCR a vários tipos de aplicações, indústrias e problemas de negócio, sendo considerado, então, validado suficientemente para ser publicado e distribuído(CHAPMAN et al., 2000).

falar sobre mineração de dados como super área no capítulo sobre aprendiza de máquina

CRISP-DM segue uma estrutura hierárquica, composta por quatro níveis de abstração(do mais genérico ao mais específico):fase, tarefa genérica, tarefa especializada e instância de processo. Os dois primeiros níveis, fase e tarefa genérica, foram modelados a fim de serem: genéricos o suficiente para atenderem às todas aplicações de Mineração de Dados; completos, abrangendo todo o processo de Mineração de Dados; e estáveis, sendo aplicáveis tanto para as técnicas de Mineração de Dados existentes, quanto às que venham a ser desenvolvidas. O terceiro nível, tarefa especializada, é composto pelas tarefas a serem executadas em situações específicas para alcançar os objetivos das tarefas genéricas. Exemplificando, seja a tarefa genérica Limpar dados: a ela relacionadas estão as tarefas especializadas Limpar dados numéricos e Limpar dados categóricos. O quarto nível, instância de processo, é composto pelos registros de ações, decisões e resultados de uma execução do processo(CHAPMAN et al., 2000).

Apesar de a representação do processo sugerir que ele é composto por uma sequência fixa de fases, na prática as tarefas podem ser executadas seguindo outras ordens: é o caso de, por exemplo, na tarefa Avaliação do modelo ser verificado que são necessários mais dados, a serem adquiridos através de tarefas anteriores, de acordo com o diagrama do processo.

Para o mapeamento do modelo em uma instância do processo, a especificação do CRISP-DM identifica como relevantes quatro dimensões do contexto de Mineração de Dados: domínio de aplicação, tipo de problema de Mineração de Dados, aspectos técnicos e ferramentas e técnicas. Os valores dessas dimensões são utilizados nas decisões sobre que tarefas específicas podem ou devem ser executadas.

O processo de mapeamento do CRISP-DM a uma instância do processo é, de acordo com o CRISP-DM, composto pelas etapas:

1. Analisar o contexto, identificando os valores para as dimensões domínio de aplicação, tipo de problema de Mineração de Dados, aspectos técnicos e ferramentas e técnicas
2. Remover do modelo CRISP-DM os detalhes não aplicáveis ao contexto analisado
3. Adicionar detalhes específicos do contexto analisado
4. Especializar(ou instanciar) elementos genéricos do modelo de acordo com características concretas do contexto
5. Possivelmente renomear elementos genéricos do modelo a fim de tornar mais explícito seu significado, de acordo com o contexto

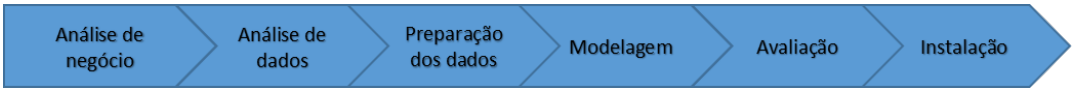

Figura 1 – Fases do CRISP-DM

A seguir segue uma descrição breve de cada uma das fases:

1. **Análise do negócio**: O objetivo desta fase é entender os requisitos e objetivos do projeto sob uma perspectiva de negócio, traduzí-los para requisitos e objetivos sob uma perspectiva de Mineração de Dados e então traçar um plano preliminar para alcançá-los.
2. **Análise dos dados**: Esta fase inicia com uma coleta inicial de dados e segue para o estudo dos dados a fim de identificar problemas de qualidade, obter insights e detectar possíveis subconjuntos de dados que permitam desenvolver hipóteses sobre informações que não estejam presentes.
3. **Preparação dos dados**: Esta fase é composta por atividades necessárias para gerar, a partir dos dados inicialmente coletados, um conjunto de dados a ser utilizado pelas ferramentas de modelagem. Inclui atividades como seleção de tabelas, de registros, de atributos e transformação de dados.

4. **Modelagem**: Nesta fase são desenvolvidos e otimizados modelos. Normalmente aplicam-se ao problema mais de uma técnicas de modelagem. Como algumas técnicas de modelagem podem possuir pré requisitos sobre os dados, pode ser necessário voltar para a fase Preparação dos dados.

5. **Avaliação**: Esta fase é iniciada quando já foi desenvolvido um modelo com alta qualidade, do ponto de vista da Mineração de Dados. Nela são avaliados a adequação do modelo como ferramenta para alcançar o objetivo de negócio que motivou o projeto e a qualidade da instância do processo. Ela termina com a decisão pela utilização ou não dos resultados obtidos.

6. **Instalação**: Após o desenvolvimento de um modelo, faz-se necessário que ele seja disponibilizado para os usuários finais, seja na forma de relatórios, seja na forma de sistemas de apoio à tomada de decisão, para que seja efetivamente utilizado, auxiliando no alcance dos objetivos de negócio que motivaram a criação do projeto.

A seguir segue o detalhamento de cada fase, especificando suas tarefas genéricas e os documentos que são gerados.

1 - Análise do negócio

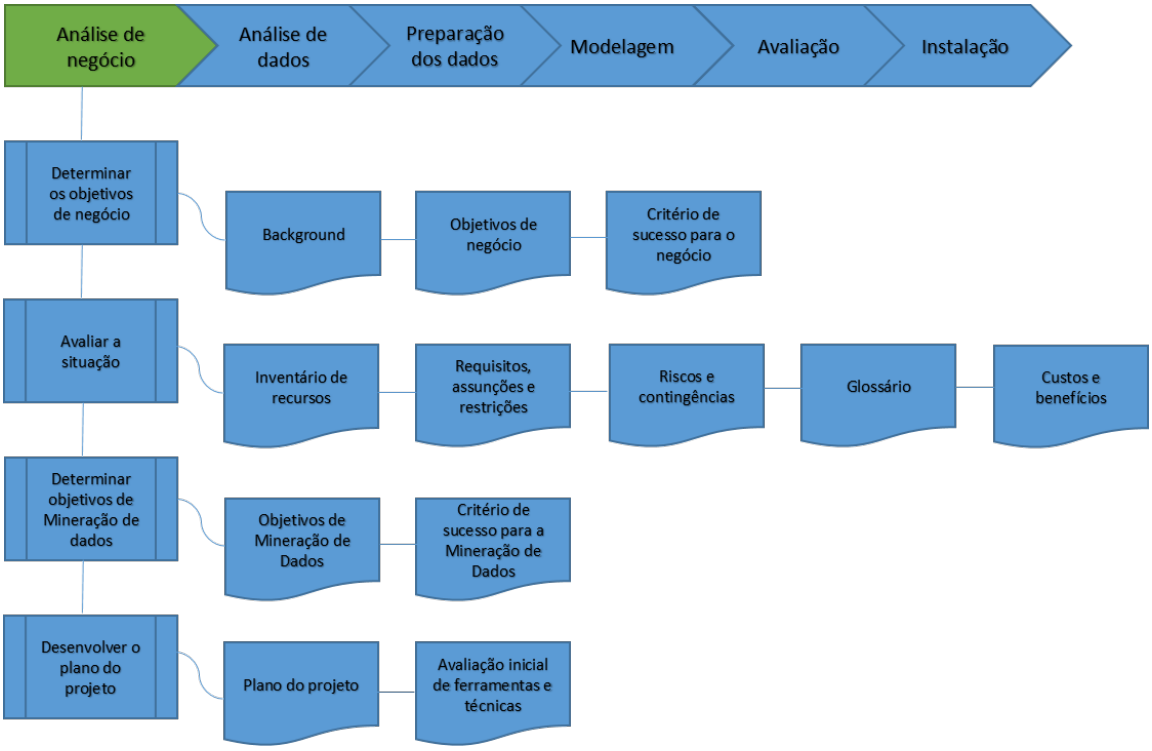

Figura 2 – Detalhamento da fase Análise de negócio

1.1 - Determinar os objetivos de negócio

O primeiro passo em um projeto de Mineração de Dados é analisar, sob uma perspectiva de negócio, o que o cliente deseja alcançar. Normalmente o cliente possui várias restrições e objetivos concorrentes, que devem ser balanceados. O objetivo desta tarefa é descobrir fatores importantes do negócio que possam influenciar no resultado do projeto. Uma das consequências de se negligenciar esse passo é o projeto finalizar respondendo corretamente a perguntas erradas.

Background

esqueci de traduzir Background...

Registro das informações levantadas sobre o estado do negócio no início do projeto.

Objetivos do negócio

Registra os objetivos de negócio que motivaram a criação do projeto, além de questões a eles relacionadas que o cliente deseje responder.

Critério de sucesso para o negócio

Registra os critérios para que o projeto seja considerado um sucesso, sob uma perspectiva de negócio.

1.2 - Avaliar a situação

O objetivo desta tarefa é analisar mais detalhadamente informações importantes para determinar os objetivos da Mineração de Dados e desenvolver um plano para o projeto. São analisadas informações como quais os recursos estão disponíveis, quais as restrições impostas, quais as assunções e outros fatores que forem relevantes para a especificação dos objetivos de Mineração de Dados e para o desenvolvimento do plano do projeto.

Inventário de recursos

Registra os recursos disponíveis para o projeto, como recursos humanos(especialistas do negócio, analistas de dados, técnicos de suporte), dados(arquivos, bases de dados operacionais, data warehouses), hardwares e softwares.

Requisitos, assunções e restrições

Registra os requisitos do projeto, incluindo prazos, níveis de qualidade, segurança e aspectos legais; as assunções do projeto, sejam assunções que poderão ser verificadas a partir dos dados utilizados pelo projeto, sejam assunções que não poderão ser verificadas, que devem ser registradas, visto que podem afetar a validade dos resultados do projeto; e as restrições do projeto, sejam restrições na disponibilidade de recursos, sejam restrições tecnológicas.

Riscos e contingências

Registra os eventos que, caso ocorram, poderão afetar os prazos ou a qualidade do projeto, bem como os planos de contingência, detalhando que ações devem ser executadas caso esses eventos ocorram.

Glossário

Registra o conjunto de termos e seus significados que são relevantes para o projeto. Inclui tanto termos pertencentes à terminologia do negócio, quanto termos pertencentes à terminologia de Mineração de Dados.

Custos e benefícios

Registra uma análise dos custos do projeto comparados com os potenciais benefícios para o negócio, caso o projeto alcance sucesso. Essa comparação deve ser o mais específico possível. Por exemplo, pode-se utilizar o custo estimado do projeto e a economia esperada, em termos monetários.

1.3 - Determinar os objetivos de Mineração de Dados

O objetivo desta tarefa é traduzir para objetivos de Mineração de Dados os objetivos de negócio analisados na tarefa Determinar os objetivos de negócio.

Objetivos de Mineração de Dados

Registra os objetivos a serem alcançados pelo projeto para auxiliar no alcance dos objetivos de negócio.

Critério de sucesso para a Mineração de Dados

Registra, em termos técnicos, os critérios para determinar se o projeto alcançou sucesso, sob uma perspectiva de Mineração de Dados.

1.4 - Desenvolver o plano do projeto

O objetivo desta tarefa é desenvolver um plano para alcançar os objetivos de Mineração de Dados e então os objetivos de negócio, analisando que atividades serão executadas e que ferramentas e técnicas serão utilizadas.

Plano do projeto

Registra as atividades a serem desenvolvidas, incluindo duração, recursos necessários, entradas, saídas e dependências. É importante que sejam registradas as dependências e riscos das atividades e como podem impactar nos prazos. Dado o aspecto iterativo de um projeto de Mineração de Dados, o plano de projeto é um documento dinâmico, sendo recomendado que ao fim de cada fase seja revisado e atualizado.

Avaliação inicial de ferramentas e técnicas

Registra a avaliação de um conjunto de ferramentas e técnicas que poderão ser utilizadas no projeto. É importante que essa análise seja realizada no início do projeto, dado que a escolha das ferramentas e técnicas podem influenciar todo o resto do projeto.

2 - Análise de dados

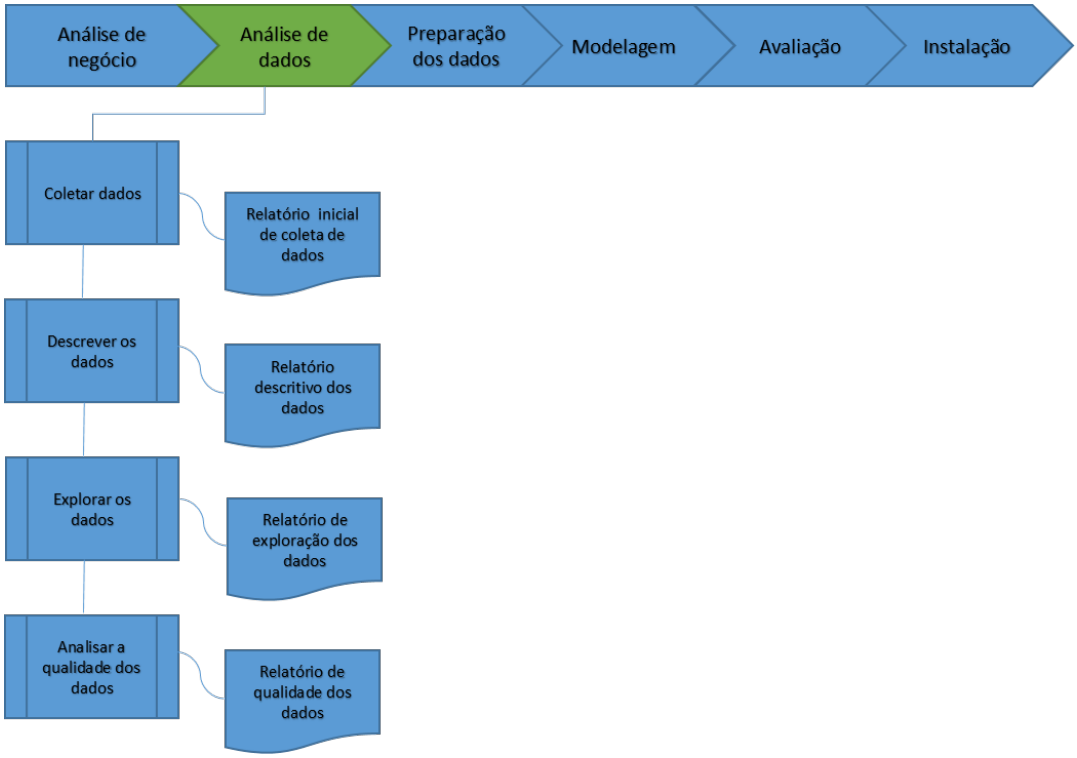

Figura 3 – Detalhamento da fase Análise de dados

2.1 - Coletar dados

O objetivo desta tarefa é realizar a coleta dos dados indicados nos recursos do projeto. Nesta tarefa estão inclusos tanto o trabalho de extração quanto de integração dos dados, caso provenham de fontes diferentes, e carregamento dos dados em ferramenta específica, caso necessário.

Relatório inicial de coleta de dados

Registra os conjuntos de dados coletados, suas localizações, os métodos utilizados na coleta e problemas, com respectivas soluções adotadas, que nela tenham ocorrido.

2.2 - Descrever os dados

O objetivo desta tarefa é realizar uma análise estrutural dos dados, avaliando se eles satisfazem os requisitos do projeto.

Relatório descritivo dos dados

Registra informações estruturais sobre os dados coletados, como formato, quantidade de registros e nomes de atributos.

2.3 - Explorar os dados

O objetivo desta tarefa é realizar uma análise da distribuição dos dados, através de consultas, visualizações e técnicas de report. Nela estão inclusas análise da distribuição

qual a tradução para report

de atributos dos dados, análise do relacionamento entre pares de atributos, análise de subpopulações e análise estatística. Essa análise serve tanto para suportar diretamente os objetivos de Mineração de Dados, quanto para refinar as informações sobre a estrutura e a qualidade dos dados.

Relatório de exploração dos dados

Registra as informações descobertas na tarefa Explorar os dados e o impacto que poderão causar no projeto.

2.4 - Analisar a qualidade dos dados

O objetivo desta tarefa é analisar a qualidade dos dados, verificando, por exemplo, se são completos(há registros para todos os casos necessários), se são corretos(frequência de erros), se há dados ausentes; e analisar soluções para os problemas de qualidade encontrados.

Relatório de qualidade dos dados

Registra os resultados da análise de qualidade dos dados, indicando os problemas de qualidade e as possíveis soluções.

3 - Preparação dos dados

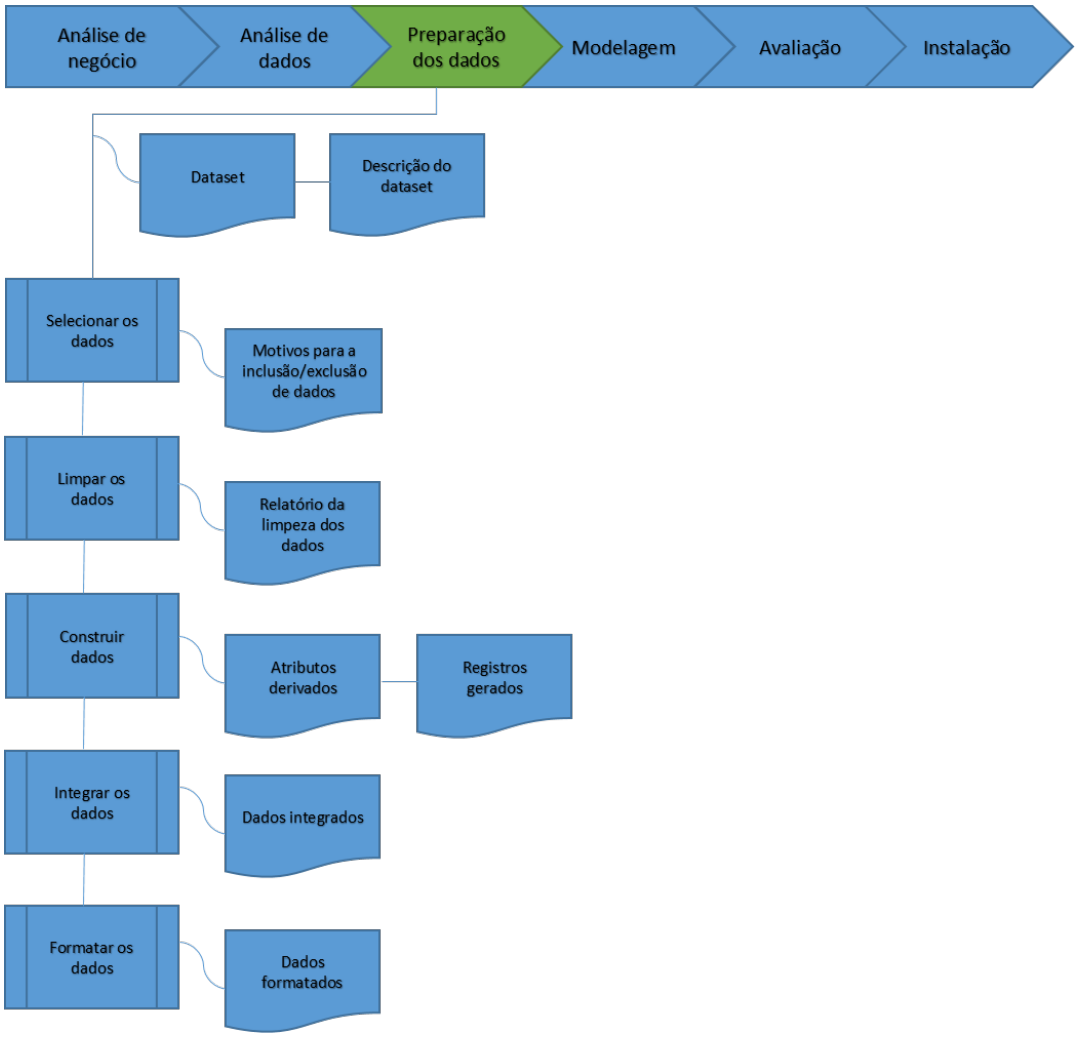

Figura 4 – Detalhamento da fase Preparação dos dados

Datasets

Datasets produzidos nesta fase, a serem utilizados no desenvolvimento de modelos ou em análises.

Descrição dos datasets

Registra informações sobre os datasets produzido nesta fase.

3.1 - Selecionar dados

O objetivo desta tarefa é selecionar datasets a serem utilizados em análises posteriores. Essa seleção envolve tanto a seleção de registros quanto a seleção de atributos. A lista de critérios para essa seleção inclui relevância dos dados para os objetivos de Mineração de Dados, qualidade e restrições técnicas, como limite no volume dos dados.

Motivos para inclusão/exclusão de dados

Registra os motivos para inclusão e exclusão de dados realizadas na tarefa Selecionar dados.

3.2 - Limpar os dados

O objetivo desta tarefa é produzir um dataset com nível de qualidade adequado para a aplicação das técnicas e modelos selecionados pelo projeto, resolvendo os problemas de qualidade analisados na tarefa Analisar a qualidade dos dados. Para tanto, atividades como seleção de subconjunto dos dados, inserção de valores padrão e estimação de valores ausentes poderão ser necessárias.

Relatório da limpeza dos dados

Registra as alterações realizadas nos dados para resolver problemas de qualidade, indicando os motivos e possíveis consequências.

3.3 - Construir dados

O objetivo desta tarefa é a criação de novos dados, através da derivação, a partir dos dados disponíveis, de novos registros ou atributos.

Atributos derivados

Registra os atributos que foram construídos a partir de outros já existentes. Por exemplo, $área = altura \times largura$.

Registros gerados

Registra a geração de novos registros.

3.4 - Integrar os dados

O objetivo desta tarefa é criar novos dados através da integração de dados de fontes diversas.

Dados integrados

Esta saída é composta tanto pelos dados que foram gerados a partir da integração de dados de fontes diversas, quanto dados agregados.

3.5 - Formatar os dados

O objetivo desta tarefa é realizar transformações nos dados que não alterem seus significados, necessárias para que os dados possam ser utilizados pelas ferramentas. Exemplos de transformações são mudança do formato do arquivo onde estão os dados, alteração na ordem das colunas ou alteração na ordem dos registros.

Dados formatados

Registra as transformações realizadas nos dados, indicando motivos e possíveis consequências.

4 - Modelagem

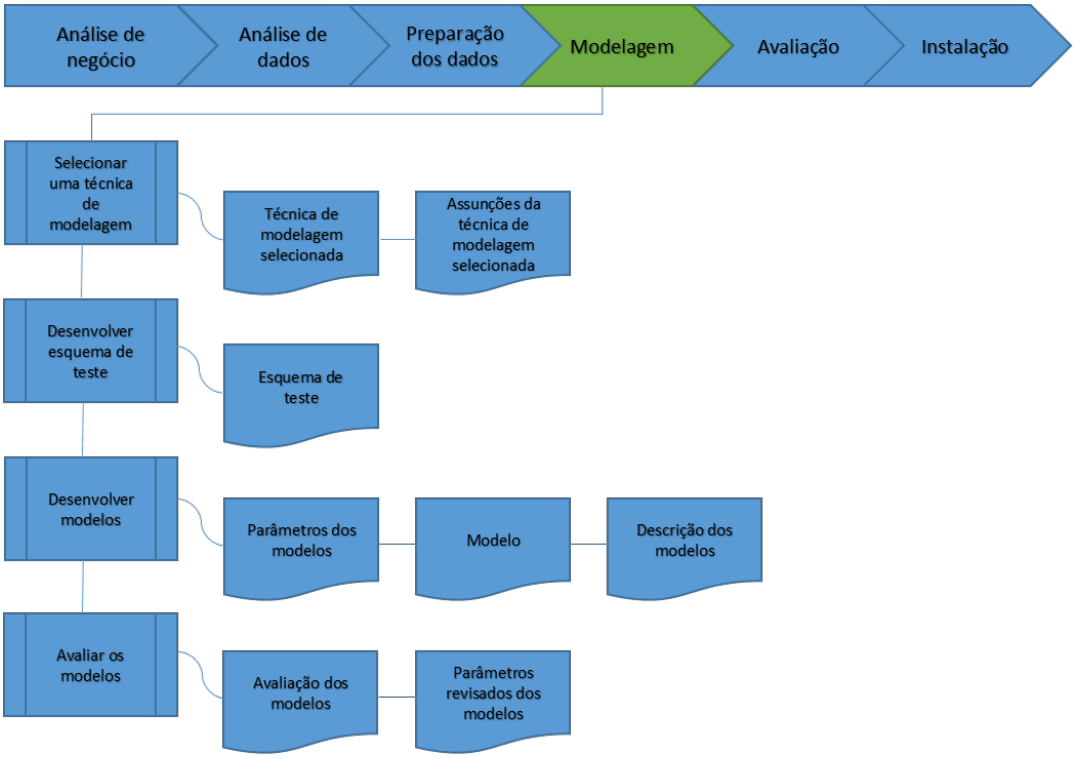

Figura 5 – Detalhamento da fase Modelagem

4.1 - Selecionar uma técnica de modelagem

O objetivo desta tarefa é selecionar uma técnica de modelagem a ser aplicada a um dataset gerado. No caso de várias técnicas de modelagem terem sido escolhidas para serem aplicadas, esta tarefa deve ser executada para cada uma delas.

Técnica de modelagem selecionada

Registra informações sobre a técnica de modelagem selecionada.

Assunções da técnica de modelagem selecionada

Registra as assunções feitas pela técnica de modelagem selecionada. Por exemplo, que todos os registros são independentes ou que todos os atributos possuem distribuição uniforme.

4.2 - Desenvolver esquema de teste

O objetivo desta tarefa é desenvolver um procedimento ou mecanismo para testar a qualidade e validade do modelo a ser desenvolvido. Para tanto, deve-se decidir, por exemplo, sobre como os dados serão particionados em subconjuntos de treinamento e de teste e quais métricas serão utilizadas para avaliar o desempenho.

Esquema de teste

Registra um plano para treinamento, teste e avaliação do modelo a ser desenvolvido.

4.3 - Desenvolver modelos

O objetivo desta tarefa é aplicar a técnica de modelagem escolhida ao dataset desenvolvido na fase anterior.

Parâmetros dos modelos

Registra os parâmetros utilizados pelos modelos desenvolvidos, bem como os motivos para suas escolhas.

Modelos

Modelos desenvolvidos.

Descrição dos modelos

Registra informações sobre os modelos desenvolvidos, como, por exemplo, como interpretá-los.

4.4 - Avaliar os modelos

O objetivo desta tarefa é avaliar os modelos desenvolvidos sob uma perspectiva de Mineração de Dados, verificando se os critérios de sucesso de Mineração de Dados foram satisfeitos, se os resultados do testes foram satisfatórios. Os modelos desenvolvidos devem ser então comparados e ordenados de acordo com critérios de avaliação.

Avaliação dos modelos

Registra os resultados da tarefa Avaliar o modelo, como a performance dos modelos desenvolvidos e uma ordem dos modelos de acordo com critérios de qualidade.

Parâmetros revisados dos modelos

Registra alterações propostas em parâmetros dos modelos desenvolvidos de acordo com a avaliação dos modelos. Os parâmetros revisados servem para serem utilizados no desenvolvimento, em uma nova iteração, de novos modelos.

5 - Avaliação

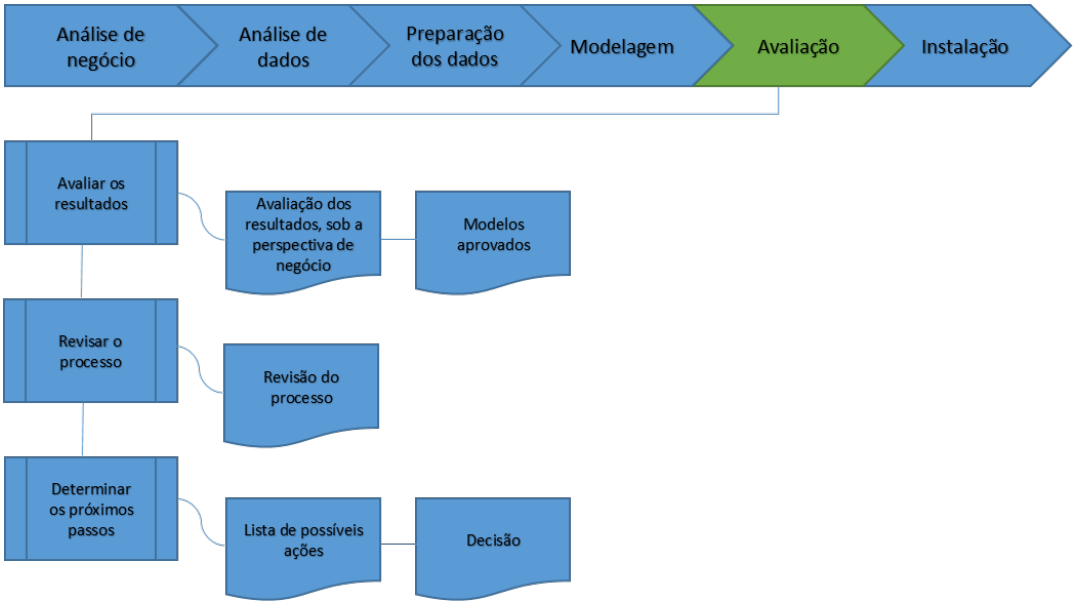

Figura 6 – Detalhamento da fase Avaliação

5.1 - Avaliar os resultados

O objetivo desta tarefa é avaliar em que medida os resultados alcançados com a Mineração de Dados, sejam modelos desenvolvidos, sejam informações extraídas, auxiliam no alcance dos objetivos de negócio e, se possível, testar os modelos desenvolvidos em aplicações reais.

Avaliação dos resultados, sob a perspectiva de negócio

Registra a avaliação dos resultados, indicando se o projeto obteve sucesso em suportar os objetivos de negócio.

Modelos aprovados

Conjunto de modelos que na avaliação dos resultados apresentaram resultados satisfatórios.

5.2 - Revisar o processo

A partir deste ponto do processo os modelos desenvolvidos já apresentam resultados satisfatórios e torna-se apropriada a realização de uma revisão do processo, a fim de verificar a qualidade das atividades até então desenvolvidas.

Revisão do processo

Registra os resultados da revisão do processo, indicando atividades que não foram desenvolvidas com a qualidade esperada e que deverão ser repetidas.

5.3 - Determinar os próximos passos

O objetivo desta tarefa é definir quais as próximas atividades a serem desenvolvidas, de acordo com os resultados da avaliação dos resultados, da revisão do processo e dos recursos disponíveis para o projeto. Pode-se decidir pela implantação dos modelos desenvolvidos, pela realização de uma nova iteração ou a finalização do projeto.

Lista de possíveis ações

Registra as potenciais ações a serem executadas e os respectivos motivos para execução ou não.

Decisão

Registra a decisão sobre quais os próximos passos a serem seguidos e a respectiva motivação.

6 - Instalação

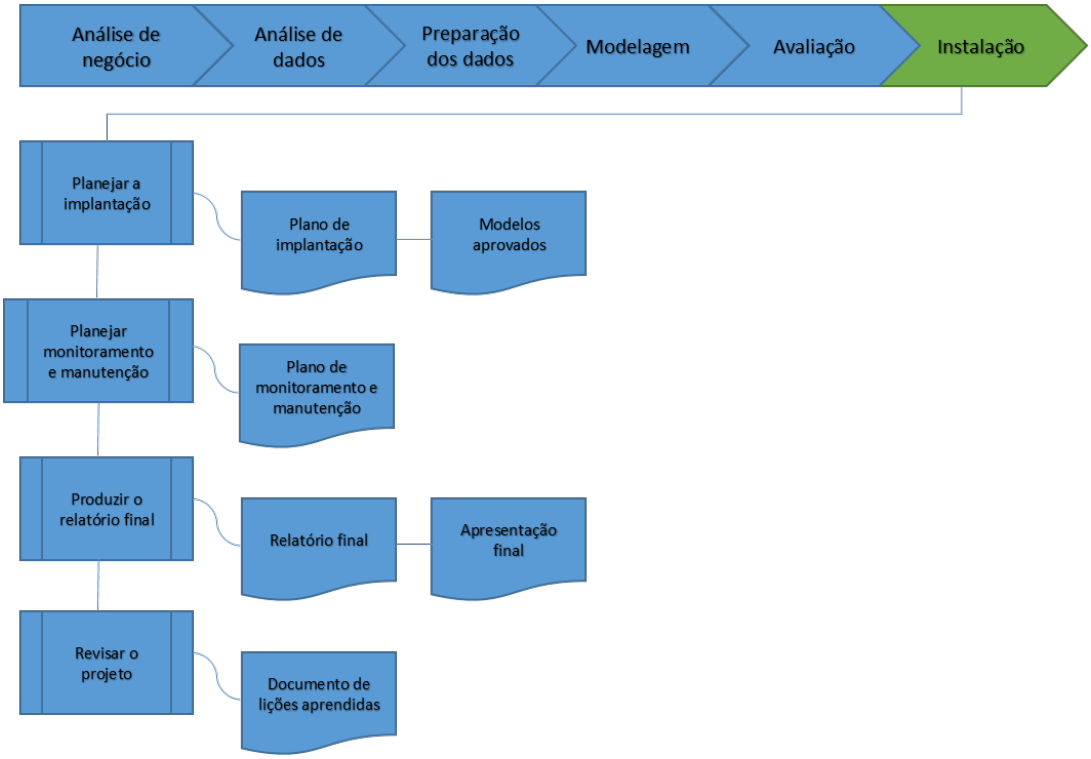

Figura 7 – Detalhamento da fase Implantação

6.1 - Planejar a implantação

O objetivo desta tarefa é produzir um plano de implantação dos modelos aprovados.

Plano de implantação

Registra as atividades necessárias para a implantação dos modelos aprovados.

6.2 - Planejar monitoramento e manutenção

O objetivo desta tarefa é desenvolver planos de monitoramento e manutenção, objetivando verificar os resultados dos modelos implantados e evitar que sejam utilizados indevidamente ou tornem-se obsoletos.

Plano de monitoramento e manutenção

Registra as estratégias de monitoramento e de manutenção.

6.3 - Produzir relatório final

O objetivo desta tarefa é produzir um relatório registrando um histórico do projeto.

Relatório final

Registra um histórico do projeto.

Apresentação final

Apresentação dos resultados alcançados pelo projeto.

6.4 - Revisar o projeto

O objetivo desta tarefa é analisar o que foi feito correta e incorretamente no projeto.

Documento de lições aprendidas

Registra as lições aprendidas no projeto, indicando que ações foram executadas corretamente, para que sejam reforçadas, e que ações foram executadas incorretamente, para que sejam corrigidas em futuros projetos.

1.2 Ferramentas utilizadas

(BUITINCK et al., 2013) (MCKINNEY, 2010)

Referências

AZEVEDO, A. I. R. L. Kdd, semma and crisp-dm: a parallel overview. 2008.

BUITINCK, L. et al. Api design for machine learning software: experiences from the scikit-learn project. *arXiv preprint arXiv:1309.0238*, 2013.

CHAPMAN, P. et al. Crisp-dm 1.0 step-by-step data mining guide. 2000.

DOMINGOS, P. A few useful things to know about machine learning. *Communications of the ACM*, ACM, v. 55, n. 10, p. 78–87, 2012.

DRUMMOND, C. Finding a balance between anarchy and orthodoxy. In: *Proceedings of the Twenty-Fifth International Conference on Machine Learning: Workshop on Evaluation Methods for Machine Learning III (4 pages)*. [S.l.: s.n.], 2008.

DRUMMOND, C. Replicability is not reproducibility: nor is it good science. 2009.

MCKINNEY, W. Data structures for statistical computing in python. In: *Proceedings of the 9th*. [S.l.: s.n.], 2010. v. 445, p. 51–56.

SCULLEY, D. et al. Machine learning: The high interest credit card of technical debt. In: *SE4ML: Software Engineering for Machine Learning (NIPS 2014 Workshop)*. [S.l.: s.n.], 2014.