Association for Information Systems
AIS Electronic Library (AISeL)

Proceedings of the XI Brazilian Symposium on Information Systems (SBSI 2015)

Brazilian Symposium on Information Systems (SBIS)

5-2015

Decision Making in Public Administration Supported by Knowledge Discovery: A Case Study in Project Management

Caue Rodrigues do Prado
Universidade de São Paulo, caue.prado@usp.br

Sarajane Marques Peres
Universidade de São Paulo, sarajane@usp.br

Marcelo Fantinato
Universidade de São Paulo, m.fantinato@usp.br

Tomada de Decisão na Administração Pública Apoiada pela Descoberta de Conhecimento: Um Estudo de Caso em Gestão de Projetos

Alternative Title: Decision Making in Public Administration Supported by Knowledge Discovery: A Case Study in Project Management

Cauê R. do Prado
Universidade de São Paulo
São Paulo – SP – Brasil
caue.prado@usp.br

Sarajane M. Peres
Universidade de São Paulo
São Paulo – SP – Brasil
sarajane@usp.br

Marcelo Fantinato
Universidade de São Paulo
São Paulo – SP – Brasil
m.fantinato@usp.br

RESUMO

Este artigo relata um estudo de caso realizado para analisar o potencial da descoberta de conhecimento como estratégia de apoio à tomada de decisão na administração pública. A descoberta de conhecimento ocorreu com a modelagem de um problema de classificação binária para predizer o sucesso ou insucesso quanto ao cumprimento do planejamento de custo e prazo de projetos, a partir da análise de dados descritivos dos mesmos. Os dados históricos foram obtidos da base de dados de um sistema de gestão de projetos construído com base nas práticas do guia PMBOK. Estratégias adotadas no estudo de caso bem como as dificuldades encontradas e os resultados obtidos são aqui apresentados.

Palavras-Chave

Árvores de decisão, classificação binária, Descoberta de Conhecimento, PMBOK, gestão de projetos

ABSTRACT

In this paper, we report a case study carried out in order to analyze the potential of applying knowledge discovery as a strategy to support decision making in public administration. The discovery of knowledge was implemented by using a binary classifier to predict the success of failure concerning cost and deadline plans. The prediction was made analyzing descriptive data of the plans. The dataset was obtained from a project management system that was built based on the practices of the PMBOK guide. The strategies used in this case study, the difficultiesfaced during the classifier modeling process and the results are discussed herein.

Categories and Subject Descriptors

K.6.1 [**Management of Computing and Information Systems**]: Project and People Management—*management techniques*; I.2.6 [**Artificial Intelligence**]: Learning—*induction, knowledge acquisition*

General Terms

Management, Design, Theory

Keywords

Decision Trees, Binary Classification, Knowledge Discovery, PMBOK, Project Management

1. INTRODUÇÃO

Em um estado democrático, os governantes são eleitos com base em sua proposta de governo, cujo objetivo é solucionar os problemas públicos existentes e desenvolver a economia, a saúde e a educação pública, de modo a melhorar a qualidade de vida dos cidadãos. Com isso, cresce a preocupação com a qualidade da execução de projetos, nascendo portanto um espaço para técnicas e ferramentas computacionais capazes de ajudar no apoio à tomada de decisão. De acordo com o "Estudo de *Benchmarking* em Gestão de Projetos Brasil 2010", realizado pelo capítulo brasileiro do PMI (*Project Management Institute*), somente 10% dos projetos estão alinhados ao planejamento estratégico do governo. Para evitar o mal planejamento e à má execução de projetos, o PMI identificou que a aplicação correta de um conjunto de boas práticas para gestão de projetos – definido no PMBOK®(*Project Management Body of Knowledge*), que envolve o uso de habilidades, ferramentas e técnicas, pode aumentar as chances de sucesso na execução de projetos [1].

Na esfera da administração pública, mesmo com planejamento adequado, sem um acompanhamento detalhado da execução do projeto, um gestor tem uma chance menor de sucesso, principalmente devido a falhas na aplicação de recursos ou no cumprimento do cronograma. Buscando evitar tais situações, técnicas de gestão baseadas no PMBOK são de grande auxílio. Nesse contexto, inserem-se os Sistemas de Gestão de Projetos (SGP) que oferecem ampla cobertura às boas práticas do PMBOK, tendo um deles sido usado como base para este estudo de caso. As funcionalidades desse tipo de sistema atendem aos "grupos de processos" e às "áreas de conhecimento" do PMBOK, com visualizações por meio de gráficos como EAP (Estrutura Analítica do Projeto), gráfico

de Gantt, valor agregado, e curva S. Os SGP são capazes de monitorar projetos e suas atividades, por meio de funcionalidades que possibilitam registrar, por exemplo, datas e valores do projeto e de suas atividades. O SGP usado especificamente neste estudo de caso é usado pela administração pública de uma cidade turística brasileira.

Neste contexto, este artigo relata um estudo de caso realizado sobre o uso do SGP supracitado, apresentando uma contribuição referente à descoberta de conhecimento capaz de apoiar a tomada de decisão no cenário de planejamento de projetos na administração pública. No estudo de caso realizado, a descoberta de conhecimento ocorreu via a análise de dados existentes no banco de dados gerado a partir do ambiente de produção desse SGP. Por meio da análise de dados de custo e prazo planejados e executados, atores envolvidos, tipo de projetos, sazonalidade, entre outros, todos referentes a projetos já concluídos, foi possível construir modelos de predição que permitem apontar alguns fatores que contribuem para a chance de sucesso ou insucesso do projeto – no que diz respeito ao cumprimento de cronograma e orçamento. Para realizar esse estudo, o processo de descoberta de conhecimento em base de dados (KDD – *knowledge discovery in databases*) foi executado com o objetivo de resolver a tarefa de mineração de dados de classificação binária [5], que foi realizada por meio da aplicação de árvores de decisão [10], com o apoio da ferramenta WEKA[1] [11].

A fim de apresentar o estudo de caso realizado, este artigo está estruturado da seguinte forma: a Seção 2 apresenta brevemente os conceitos teóricos necessários ao entendimento do contexto do estudo de caso; a Seção 3 apresenta o estudo de caso, incluindo informações sobre o SGP que gera os dados sobre os quais a descoberta de conhecimento foi realizada, a definição da tarefa de classificação buscada, os procedimentos de pré-processamento de dados, e o projeto dos experimentos realizados; a Seção 4 apresenta análises sobre os resultados obtidos e as dificuldades encontradas durante o processo de descoberta de conhecimento; finalmente, a Seção 5 apresentada as considerações finais.

2. CONCEITOS FUNDAMENTAIS

O estudo de caso aqui apresentado está baseado nas grandes áreas de estudo de gestão de projetos e de descoberta de conhecimento, e também na técnica de aprendizado de máquina chamada árvores de decisão. Esta seção apresenta um breve resumo sobre os principais conceitos dessas áreas.

2.1 Gestão de Projetos

De acordo com Raynal [9], um projeto é a expressão de um desejo, de uma vontade, de uma intenção e de uma ambição. É também a expressão do necessitar de algo, de uma situação futura vivamente desejada. Um projeto também pode ser visto, segundo Kerzner [6], como um empreendimento, com objetivos bem definidos, que consome recursos, e opera sob pressões de prazos, custos e qualidade. Em uma visão similar [7], projeto é um empreendimento temporário ou uma sequência de atividades com objetivos claros, definidos em função de algum problema, oportunidade ou interesse pessoal ou organizacional, com ações conduzidas por pessoas, e seguindo indicadores de qualidade. Essas definições de projeto se complementam em uma visão abrangente, em que um projeto é apresentado como um empreendimento com início e fim determinados, e executado sob um ciclo de vida.

Do ponto de vista estratégico, um projeto é visto como um meio para as organizações atingirem seus objetivos e conseguirem eficácia e eficiência e, consequentemente, vantagem competitiva. Entretanto, para isso, as organizações precisam empregar boas práticas de gestão de projetos, para possibilitar o cumprimento de prazo, custo e qualidade propostos. Portanto, é importante buscar um modo de gerenciá-los, criando-se um cenário com objetivos claros e alcançáveis.

Para Kerzner [6], a gestão de projetos se refere ao planejamento, a programação e o controle de tarefas integradas de forma a atingir seus objetivos com êxito, para benefício dos participantes do projeto. De acordo com o PMBOK [1], a gestão de projetos é a aplicação de conhecimentos, habilidades, ferramentas e técnicas para atender os requisitos do projeto. Portanto, de forma geral e simplificada, a gestão de projetos trata do planejamento e do controle dos projetos. Se um projeto for bem planejado, de acordo com as necessidades e expectativas do cliente, e sua execução for devidamente controlada, o resultado será a satisfação do mesmo quanto à qualidade do produto ou serviço desenvolvidos, respeitando os custos e os prazos definidos [8]. Complementarmente, é benéfico o uso de práticas de gestão também na área pública, visando ao cumprimento dos projetos e obras públicas com garantia de prazo, custo e qualidade.

Via um trabalho com especialistas da área de gestão de projetos, o PMI fez um levantamento de uma série de conhecimentos, entendidos como "boas práticas" para gestão de projetos. Essas práticas foram compiladas em um guia para gestão de projetos, o PMBOK (*Project Management Body of Knowledge*) [1], apresentando-as no modo de aproximadamente 50 macro-processos de gestão, organizados em duas classificações ortogonais: grupos de processos e áreas de conhecimento. Os grupos de processo dividem os 50 processos em termos de tipo de atividade no ciclo de vida: iniciação, planejamento, execução, monitoramento e controle, e encerramento. Ortogonalmente, as áreas de conhecimento dividem os mesmos 50 processos em termos de tipo de assunto tratado na gestão do projeto: integração, escopo, tempo, custo, qualidade, recursos humanos, comunicações, riscos, aquisições e partes interessadas.

2.2 O processo KDD

A transformação de dados em conhecimento consiste em um processo onde informação especializada é extraída dos dados e apresentada em um formato adequado para interpretação e uso, de forma prática, em processos reais. Porém, a necessidade de que tal extração seja eficiente e eficaz exige, atualmente, o uso de técnicas e ferramentas automatizadas. O termo descoberta de conhecimento em bases de dados (do inglês *Knowledge Discovery in Database* – KDD) é usado para denominar a resolução do problema de automatizar a extração do conhecimento.

O processo de KDD envolve etapas que são executadas de forma iterativa e interativa. O estabelecimento de fronteiras claras entre cada etapa do processo de KDD não é consenso na literatura especializada. A Figura 1 apresenta uma proposta de divisão do processo KDD em cinco etapas. Para o estudo de caso apresentado neste artigo, tais etapas foram formatadas em três ações: *(i)* pré-processamento, que inclui a seleção e a formatação de dados; *(ii)* mineração de dados, implementada no modo de resolução da tarefa de classificação com o uso de um modelo de classificação; e *(iii)* análise,

[1]http://www.cs.waikato.ac.nz/ml/weka/

apresentada nesse estudo em termos de medidas de avaliação da classificação e de avaliação de informação útil à tomada de decisão na gestão de projetos.

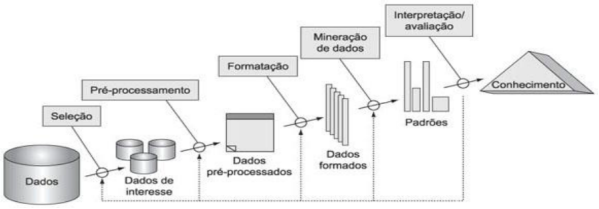

Figura 1: Etapas do processo de descoberta de conhecimento em base de dados [3].

Dentro do processo de KDD, a etapa de mineração de dados refere-se à aplicação de técnicas para processamento dos dados e extração de padrões ou modelos. Para definir uma taxonomia para os diferentes tipos de conhecimento que podem ser minerados de uma base de dados, Han and Kamber [5] sugerem duas categorias: a categoria que objetiva a extração de conhecimento preditivo e a categoria que objetiva a extração de conhecimento descritivo. A extração de conhecimento preditivo é implementada por métodos que usam atributos descritivos de dados históricos para criar um modelo capaz de predizer algum tipo de conhecimento.

A predição de conhecimento que classifica um exemplo do conjunto de dados em uma determinada classe (ou seja, uma partição) recebe o nome de "tarefa de classificação". A derivação de um modelo de classificação é baseada na análise de um conjunto de dados de treinamento em que todos os dados já se encontram classificados. Após a derivação, o modelo pode ser usado para predizer a classe de novos dados. Neste trabalho, a técnica de árvores de decisão foi usada para extração de modelos preditivos.

2.3 Árvores de Decisão

Segundo Rokach and Maimon [10], uma árvore de decisão é um modelo preditivo que pode ser usado tanto para a construção de classificadores quanto para a construção de modelos de regressão. No primeiro caso, as árvores são conhecidas como árvores de classificação.

Na resolução de uma tarefa de classificação, as árvores de decisão são usadas para classificar dados, objetos ou instâncias de um conjunto de dados, em um conjunto pré-definido de classes, a partir da análise dos valores assumidos por seus atributos descritivos. A árvore é construída de forma indutiva, ou seja, a partir da análise de um conjunto de dados de treinamento. Durante o processo de construção da árvore, os atributos que mais eficientemente discriminam os dados em suas classes são escolhidos para compor seus nós internos. Os valores que esses atributos podem assumir determinam os ramos dos respectivos nós, fazendo com que esses nós assumam o papel de particionadores do espaço de decisão. Para cada ramo, uma nova árvore é construída, recursivamente, considerando os atributos descritivos ainda não usados (na árvore ou em um caminho dela, a depender da estratégia adotada). O processo se repete até que não existam mais atributos ou até que não seja mais necessário analisar atributos para obter a resposta para a classificação dos dados. Os valores constantes nos nós folhas da árvore são as classes em que um dado pode ser classificado. A classificação de um novo dado, a partir do modelo preditivo representado na árvore, ocorre por meio do caminho que é seguido na árvore de acordo com a verificação dos valores que os atributos que compõem a árvore assumem para aquele dado.

Os algoritmos usados para construir uma árvore de decisão diferem entre si principalmente pelo modo de analisar os atributos e determinar seu poder de discriminação. Neste estudo de caso, dois algoritmos foram aplicados: J48 e CART. J48 é uma implementação derivada do clássico algoritmo C4.5, que usa o raio do ganho de informação como medida para análise do poder de discriminação de um atributo. O ganho de informação é uma medida que usa informação de entropia para analisar o quão puras[2] são as partições formadas pelo uso do atributo. CART usa o índice Gini, que também mede a impureza de uma partição, porém trabalha com partições binárias. Informações detalhadas sobre o cálculo desses índices são encontradas em [5].

3. O ESTUDO DE CASO

Para apresentar o estudo de caso a que se refere este artigo, as próximas seções estão divididas em: *(i)* uma apresentação resumida do SGP responsável por apoiar a gestão de projetos e, portanto, responsável por gerar os dados analisados pelo processo de descoberta de conhecimento; *(ii)* a definição da tarefa de mineração de dados responsável por criar os modelos que representam o conhecimento descoberto; *(iii)* os procedimentos de pré-processamento de dados que representam uma das maiores dificuldades na área de descoberta de conhecimento e mineração de dados sob dados de sistemas que originalmente não foram preparados para tal; e *(iv)* a descrição do escopo dos experimentos realizados. Na seção posterior (Seção 4) são apresentados os resultados obtidos e algumas análises deles derivadas.

3.1 O Sistema de Gestão de Projetos

O Sistema de Gestão de Projetos (SGP)[3] usado neste estudo de caso foi desenvolvido de acordo com as áreas de conhecimento e grupos de processo do PMBOK. Trata-se de um sistema especialmente projetado para apoiar o trabalho do gestor público; mais especificamente o trabalho desenvolvido na esfera executiva da gestão municipal. O SGP em questão oferece uma série de funcionalidades, permitindo a persistência e gestão de dados sobre projetos e sobre as atividades relacionadas, apoiando o acompanhamento da execução do projeto e a comunicação entre os atores que formam as respectivas equipes de execução. O SGP está otimizado para atender a requisitos da administração pública.

Especificamente para este estudo de caso, apenas os dados armazenados em relações do banco de dados pertinentes aos processos dos grupos *iniciação* e *planejamento* do PMBOK foram submetidos à mineração. A Figura 2 apresenta a parte do modelo relacional selecionada para análise. Nesse modelo parcial de dados, verificam-se relações e atributos que permitem a persistência de dados sobre: datas de início e fim para o projeto, tanto previstas quanto de fato realizadas; valor do projeto (custo); alíneas de fontes de recursos (por licitação, recursos próprios, recursos financiados, provenientes de reservas municipais, estaduais ou federais); estado atual do

[2]Nesse contexto, uma partição é pura se ela contém apenas dados de uma única classe.

[3]Trata-se de um sistema real implantado em uma prefeitura de uma cidade turística do interior do país.

projeto (em execução, suspenso, finalizado); atores (cargos e funções) que formam a equipe de execução; existência de aditivos no caso de extensão do projeto e informações pertinentes; datas de início e fim para as atividades do projeto, tanto previstas quanto de fato realizadas; valores associados a cada atividade; e dados referentes a riscos associados (impacto, probabilidade e descrição).

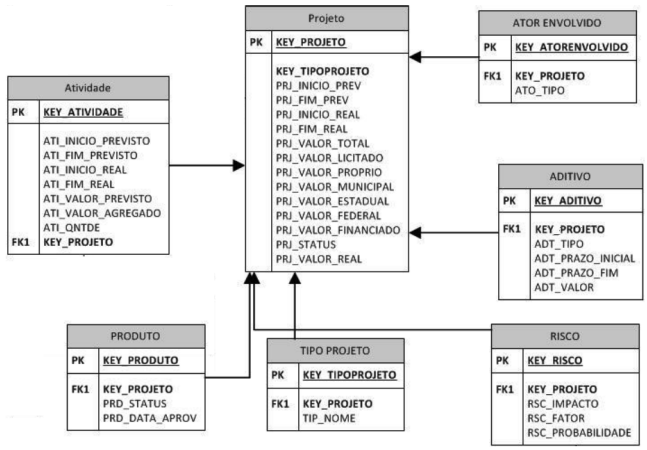

Figura 2: Modelo de dados do SGP: relações pertinentes a processos de planejamento.

A escolha das relações e dos atributos a serem usados no estudo de caso foi realizada visando a apoiar o conhecimento que se pretendia minerar (conforme descrito na Seção 3.2). Para isso, foi necessário realizar um processo de engenharia reversa para documentar a relação existente entre as relações e os atributos da base de dados e os processos do PMBOK.

A partir do detalhamento da conformidade e relação dos elementos do modelo relacional com os processos do PMBOK, verificou-se a necessidade de documentar o modo de inserção de dados no SGP. Isso foi necessário para analisar o tipo de tarefa de mineração que poderia ser útil no contexto de uso desse SGP. Saber como as informações são produzidas durante o processo real permite avaliar melhor que tipo de conhecimento pode ajudar na tomada de decisão relacionada ao processo. Assim, o modo como a informação é inserida no SGP foi mapeada usando diagrama de atividades da UML e uma visão resumida é apresentada na Figura 3.

A Figura 3 apresenta a ordem em que ocorre a iniciação, o planejamento, a execução, o monitoramento e controle, e o encerramento do projeto no escopo do SGP. Um projeto é criado, os riscos são adicionados juntamente com o planejamento de comunicação e os atores designados para o projeto. As fases de execução e de monitoramento e controle iniciam-se após a inserção das informações de acompanhamento, que se referem a informações sobre a execução das atividades relacionadas ao projeto: conforme as atividades vão sendo realizadas, percentuais de conclusão vão sendo cadastrados até que a atividade seja de fato concluída. Quando todas as atividades do projeto estão concluídas e todos produtos e solicitações do projeto são aprovados, então o projeto pode ser encerrado.

3.2 Tarefa de Mineração

Basicamente, o motivo para realizar um processo de descoberta de conhecimento no âmbito do SGP foi entender

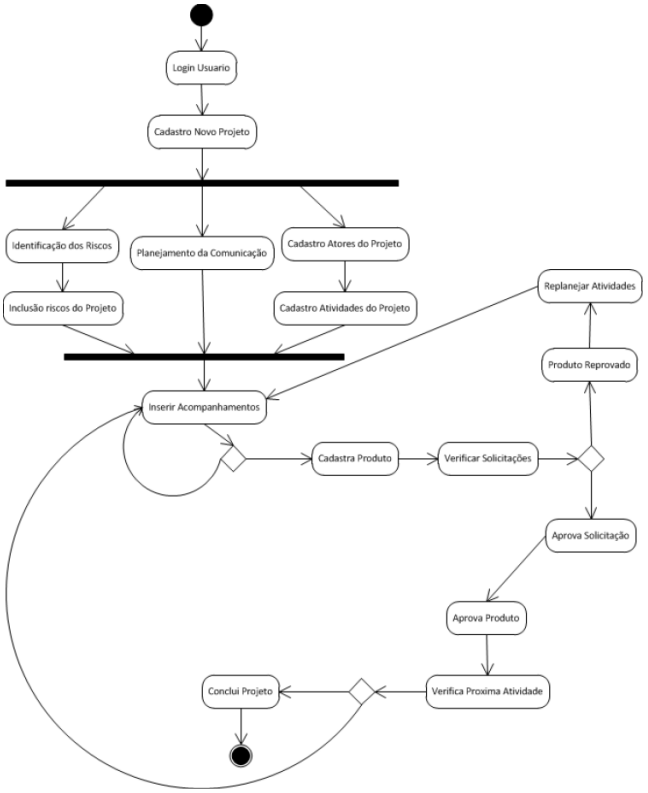

Figura 3: Fluxo de atividades no SGP: visão resumida.

quais os fatores poderiam influenciar na gestão de um projeto na administração pública, levando-o ao sucesso ou ao fracasso. O entendimento de tais fatores é visto aqui como a descoberta do conhecimento que apoia a tomada de decisões.

Os dados disponíveis para análise e descoberta de conhecimento eram resultantes da implantação do SGP na prefeitura de uma cidade. Durante o período de julho de 2012 a novembro de 2013, os projetos foram cadastrados e acompanhados pelo SGP e, assim, os dados usados nesse estudo são provenientes desse intervalo de tempo. Um intervalo de tempo totalmente inserido no passado foi escolhido para que fosse possível construir uma base de dados histórica referente a projetos já encerrados.

Então, uma análise dos atributos presentes no modelo de dados relacional foi realizada e uma lista de atributos foi selecionada para compor a descrição de um projeto na base de dados histórica. Os atributos foram escolhidos considerando: o conhecimento especialista sobre quais dados poderiam representar informação útil à obtenção de sucesso ou fracasso em um projeto; e, a presença de valores para tais atributos na base de dados legada. De fato, o conjunto de atributos que formam a base de dados legada é bem maior do que a usada. Porém, devido a políticas adotadas para uso do SGP e devido a ocorrências referentes a erros de usuários no cadastro das informações, muitos atributos (como aqueles referentes à gestão de riscos e planos de comunicação) não puderam ser aproveitados por terem uma número muito grande de valores ausentes, inconsistentes ou ruidosos. Os atributos selecionados para uso no processo de descoberta de conhecimento estão listados na Tabela 1, organizados de

acordo com a relação de onde são provenientes (Figura 2).

Tabela 1: Atributos selecionados para uso no processo de descoberta de conhecimento

Projeto	**Aditivo**	**Atividade**
key_projeto	key_aditivo	key_atividade
inicio_previsto	prazo_inicial	inicio_previsto
fim_previsto	prazo_fim	fim_previsto
valor_licitado		valor_previsto
valor_proprio		
valor_estadual		
valor_federal		
valor_real		
inicio_real		
fim_real		

Tipo Projeto	**Ator Envolvido**	**Produto**
key_projeto	key_atorenvolvido	key_produto
	tipo	

Com base nos atributos escolhidos e no tipo de informação que é possível extrair deles, a tarefa de mineração definida para ser usada no processo de descoberta de conhecimento foi a "tarefa de classificação", ou seja, uma tarefa preditiva. De forma mais específica, dois tipos de predição foram estabelecidos: a predição sobre a possível ocorrência de um estouro de orçamento; e, a predição sobre a possível ocorrência de um estouro de prazo. Na realidade, a tarefa de classificação implementada é do tipo binária e as classes de predição são: classe **positiva** para **ocorrência** de um estouro e classe **negativa** para a **não ocorrência** de um estouro.

A motivação por trás de tais objetivos é que em um processo real, a partir da inserção de informações sobre um projeto e seu planejamento, a aplicação de um modelo preditivo daria ao gestor a predição sobre a ocorrência de estouros de orçamento e estouros de prazo para aquele planejamento, de forma que medidas preventivas poderiam ser tomadas no caso de predições positivas.

Após a escolha dos atributos e definição das tarefas de classificação, um procedimento de seleção de dados foi necessário visto que ainda restaram alguns dados que não possuíam valores cadastrados para alguns dos atributos. Ao final desse processo, chegou-se ao conjunto de dados mencionado na Tabela 2.

Tabela 2: Descrição do conjunto de dados

	Para estudo de estouro de **orçamento**	Para estudo de estouro de **prazo**
Classe +	29	59
Classe −	111	96

3.3 Pré-processamento

Os dados, em seu estado original na base de dados legada apresentam incompletude, inconsistências e ruído. Muitos projetos não possuíam valores cadastrados para todos os atributos, principalmente em relação a atributos de gestão de comunicações e de riscos. Visto que esse problema ocorria para a grande maioria dos casos, atributos que se relacionavam a esses quesitos precisaram ser desconsiderados, ou não seria possível compor a base histórica. Alguns poucos casos, em que os valores para alguns atributos não estavam de acordo com o esperado e não podiam ser imputados manualmente, foram considerados como ruído e os projetos relacionados não foram incluídos da extração de dados.

Em relação a inconsistências, foram observados alguns casos referentes a dados sobre datas. Havia projetos com data de início real maior do que a data de final prevista. Essas ocorrências foram consideradas como inconsistências e corrigidas manualmente durante a extração de dados para a construção da base histórica. A justificativa para esses casos era que se tratavam de projetos que embora tenham sido planejados em um momento no tempo, não puderam ser executados por diferentes motivos, e tiveram sua execução retomada em um momento no futuro. No entanto, eles não poderiam, por motivos legais, constarem como novos projetos e, por isso, houve apenas a criação de "aditivos".

Após a limpeza da base de dados, foi necessário executar uma série de procedimentos para preparar os dados para a mineração do conhecimento. Esses procedimentos envolveram a derivação de novos atributos e a transformação de alguns atributos. A Tabela 3 apresenta a lista final de atributos, usada para a construção das árvores de decisão. Os atributos apresentados em negrito são aqueles que sofreram uma transformação (procedimento de binarização ou procedimento de categorização); os atributos em itálico são atributos derivados; e os demais atributos mantiveram-se em seu estado original.

Tabela 3: Lista de atributos usada na construção das árvores de decisão (bin − binário; cat − categórico)

Tipo de projeto	**Valor Estadual (bin)**
Fim Previsto	**Valor Federal (bin)**
Início Previsto	**Valor licitado (bin)**
Duração Prevista	**Valor Próprio (bin)**
Estouro de duração (bin)	**Valor Total (cat)**
Estouro de prazo (bin)	*Valor Outros*
Qtde de atividades	**Ator Tipo 2 (bin)**
Qtde de obras previstas	**Ator Tipo 3 (bin)**
Valor Estadual	**Ator Tipo 4 (bin)**
Valor Federal	**Ator Tipo 5 (bin)**
Valor Licitador	**Ator Tipo 6 (bin)**
Valor Próprio	**Ator Tipo 7 (bin)**
Valor Total	**Ator Tipo 8 (bin)**
Média de duração previstas para atividades	
Duração prevista máxima para atividades	
Média de valores previstos para atividades	
Valor previsto máximo para atividades	

Os procedimentos de binarização indicaram quando um atributo se aplicava ou não a um projeto, por exemplo: projetos que não receberam verba federal receberam o valor binário FALSE para o atributo *criado Valor Federal (bin)*, enquanto o atributo *Valor Federal* estava preenchido com o valor real 0. Os procedimentos de categorização dividiram os valores de um atributo em faixas. As faixas foram determinadas empiricamente pelos projetistas da descoberta de conhecimento.

Por fim, o problema de desbalanceamento, evidente no conjunto de dados usado para estudo de estouro de orçamento, foi tratado com a aplicação da estratégia SMOTE [2], implementada na ferramenta WEKA. Nessa estratégia, exemplos (dados) sintéticos são gerados a partir da classe minoritária, fazendo com que a sua região de decisão se torne mais representativa dentro do conjunto de dados. O

tratamento de dados desbalanceados é importante para minimizar os efeitos que a classe majoritária pode causar na construção de modelos classificadores. Mais detalhes sobre o problema de construção de árvores na presença de dados desbalanceados podem ser consultados em [4].

3.4 Experimentos

A escolha do algoritmo a ser usado na ferramenta WEKA foi baseada em testes preliminares com os conjuntos de dados. Após análise de alguns algoritmos que implementam árvores de decisão, optou-se por aplicar os algoritmos J48 e o SimpleCART, visto que eles apresentaram resultados interessantes em termos de complexidade de análise (as árvores de decisão foram geradas com a aplicação de procedimentos de poda, o que as tornam mais compactas). Assim, tais algoritmos foram executados a fim de construir modelos preditivos no modo de árvores de decisão capazes de predizer estouros de orçamento e de prazo, e também capazes de informar quais são os atributos descritivos do projeto que mais contribuem para chegar em tais predições.

Os algoritmos foram primeiramente testados usando a estratégia de re-substituição, em que o próprio conjunto de treinamento é usado para realização de testes do modelo [5]. Essa estratégia de teste é interessante para verificar a capacidade básica da técnica em aprender o modelo preditivo inerente aos dados. Porém, esse teste não avalia o quanto o modelo gerado é capaz de generalizar a predição para novos dados. Então, a fim de avaliar essa característica, os algoritmos foram também testados sob a estratégia de *k-fold cross-validation*, com $k = 10$. Essa estratégia separa o conjunto em k subconjuntos $(k1, ..., kn)$ e então $n - 1$ subconjuntos são usados para treino do modelo, e um deles é usado para estimação do erro (ou seja, o teste) [5].

Existem vários parâmetros que podem ser estudados na aplicação dos algoritmos J48 e SimpleCART. Para este estudo de caso, os parâmetros estudados se basearam nos valores padrão da ferramenta WEKA[4].

Árvores de decisão foram geradas considerando diferentes combinações de atributos e de parâmetros. Os resultados para discussão neste estudo de caso foram selecionados de acordo com a possibilidade de geração de conhecimento que pudesse ser usado para gestão de projetos, ou seja, resultados em que foi obtido um bom nível de generalização sob a estratégia *cross-validation*, e provenientes de árvores de decisão por meio das quais fosse possível interpretar, em algum nível, as regras de decisão geradas.

4. RESULTADOS E DISCUSSÃO

Quatro árvores de decisão foram escolhidas para ilustrar este estudo de caso. Suas capacidades de classificação foram analisadas em termos de: taxa de classificações corretas (acurácia), precisão, revocação e *f-measure* [5]. A Tabela 4 apresenta essas medidas para cada um dos experimentos. Verifica-se, nesses resultados, que os modelos gerados apresentam resultados medianos que podem ser melhorados a partir de ações de otimização no pré-processamento dos dados, na variação dos parâmetros de configuração dos algoritmos, e no enriquecimento da base de dados.

[4]Foram realizados alguns experimentos nos quais aplicou-se alterações nos valores dos parâmetros. No entanto, até o presente estado do estudo, as alterações não geraram resultados expressivamente melhores do que os resultados apresentados neste artigo.

Tabela 4: Avaliação dos modelos preditivos obtidos nos experimentos

	J48		**SimpleCART**	
Estudo de estouro de **orçamento**	Acurácia:	0,83	Acurácia:	0,71
	Precisão:	0,84	Precisão:	0,70
	Revocação:	0,83	Revocação:	0,71
	F-measure:	0,83	*F-measure*:	0,70
Estudo de estouro de **prazo**	Acurácia:	0,71	Acurácia:	0,74
	Precisão:	0,70	Precisão:	0,75
	Revocação:	0,71	Revocação:	0,75
	F-measure:	0,70	*F-measure*:	0,75

Entretanto, embora as avaliações referentes à acurácia de classificação ainda precisem ser melhoradas, as árvores de decisão geradas trouxeram conhecimento sobre informações relevantes sobre projetos, que devem ser consideradas no trabalho de gestão a fim de minimizar resultados negativos (ou seja, estouro em orçamento e estouro em prazo).

A seguir são detalhados, em duas subseções separadas, os resultados dos estudos para estouro de orçamento e de prazo. Nessas seções, as árvores de decisão são graficamente apresentadas. A interpretação dos gráficos deve ser realizada da seguinte forma:

- Cada linha da representação gráfica representa um nó folha da árvore se há informação numérica (a primeira linha da Figura 4), e um nó interno caso contrário (a terceira linha da Figura 4).
- Cada nó folha, nas árvores geradas pelo J48 (segunda linha da Figura 4) é composto pelo nome do atributo (*tipoprojeto*), valor do atributo (*Obras*), decisão sobre a classe (0 – não há estouro), número de exemplares no conjunto de dados que chegam neste nó folha (14), e número de exemplares classificados incorretamente (1); já nas árvores geradas pelo SimpleCART (primeira linha da Figura 5) é composto pelo nome do atributo (*FimPrevisto*), valores do atributo (*janeiro, dezembro, maio, junho e fevereiro*), decisão sobre a classe (1 – há estouro), número de exemplares que chegam nesse nó folha e estão corretamente classificados (38), e número de exemplares classificados incorretamente (16);
- Cada nó interno (terceira linha da Figura 4) é composto pelo nome do atributo (*valorlicitadobinario*) e valor do atributo (*construcao*).
- Uma regra simples na árvore (primeira linha da Figura 4) é "**SE** *tipo de projeto é licitaçãoCP* **ENTÃO** *não haverá estouro de orçamento*".
- Uma regra composta na árvore (terceira à quinta linhas da Figura 4) é "**SE** *tipo de projeto é construção* **E** *não há valor licitado* **ENTÃO** *não haverá estouro de orçamento* **SENÃO**, **SE** *há valor licitado* **ENTÃO** *haverá estouro de orçamento*".

4.1 Estudo para Estouro de Orçamento

A Figura 4 apresenta a árvore de decisão para o experimento que considerou a estratégia de balanceamento de dados SMOTE e o algoritmo de árvore de decisão J48, especificamente para estudo do estouro de orçamento. Trata-se

de uma árvore de profundidade 3, ou seja, apenas até três perguntas precisam ser respondidas para tomar a decisão final sobre estouro de orçamento. Os atributos do conjunto de dados considerados relevantes e necessários para predizer o estouro de orçamento são: *tipo do projeto*, *valor licitado (bin)*, *data de fim prevista (mês)*, *valor total (cat)*.

```
tipoprojeto = LicitacaoCP: 0 (32.0)
tipoprojeto = Obras: 0 (14.0/1.0)
tipoprojeto = Construcao
|   ValorlicitadoBinario = 0: 0 (13.0/1.0)
|   ValorlicitadoBinario = 1: 1 (19.0/2.0)
tipoprojeto = EquipeInterna: 0 (1.0)
tipoprojeto = Pav./Drenag./Prep.: 0 (8.0/4.0)
tipoprojeto = Reforma: 0 (11.0/2.0)
tipoprojeto = GestaoUrbana: 0 (25.0/2.0)
tipoprojeto = Saude: 0 (5.0/1.0)
tipoprojeto = Outros: 0 (16.0/3.0)
tipoprojeto = PlanodeAcao: 0 (2.0)
tipoprojeto = Educacao: 0 (3.0)
tipoprojeto = Praca: 0 (1.0)
tipoprojeto = TAC: 0 (8.0)
tipoprojeto = ConvenioDADE: 0 (2.0)
tipoprojeto = Servicos
|   FimPrevisto = Maio: 1 (4.0)
|   FimPrevisto = Junho: 1 (3.0)
|   FimPrevisto = Abril: 0 (2.0/1.0)
|   FimPrevisto = Novembro
|   |   Valortotalclasses = MuitoAlto: 1 (32.0/11.0)
|   |   Valortotalclasses = SemValor: 0 (0.0)
|   |   Valortotalclasses = Medio: 0 (26.0/1.0)
|   |   Valortotalclasses = Baixo: 0 (30.0/2.0)
|   FimPrevisto = Julho: 0 (0.0)
|   FimPrevisto = Outubro: 0 (2.0)
|   FimPrevisto = Fevereiro: 0 (8.0/1.0)
|   FimPrevisto = Marco: 1 (69.0/17.0)
|   FimPrevisto = Janeiro: 0 (0.0)
|   FimPrevisto = Dezembro: 0 (0.0)
|   FimPrevisto = Setembro: 0 (0.0)
tipoprojeto = ComissaodeObras: 0 (1.0)
tipoprojeto = Treinamento: 0 (1.0)
```

Figura 4: Árvore de decisão gerada no estudo de estouro de orçamento com o algoritmo J48.

A Figura 5 apresenta a árvore de decisão gerada pelo WEKA, para o experimento que considerou a estratégia de balanceamento de dados SMOTE e o algoritmo de árvore de decisão SimpleCART, para estudo do estouro de orçamento. Trata-se de uma árvore um pouco mais complexa do que aquela gerada pelo algoritmo J48. É uma árvore de profundidade 4, e os atributos considerados relevantes à tomada de decisão são: *data de início prevista (mês)*, *tipo de projeto*, *data de fim prevista (mês)*, *valor licitado (bin)*, *valor total do projeto* e *valor próprio*.

Os modelos preditivos representados pelas duas árvores de decisão apresentadas nesta seção fornecem apoio similar à tomada de decisão de gestores públicos. Ambos evidenciam que, com base nos dados históricos usados, o tipo do projeto, o valor do projeto e a época em que os projetos ocorrem (iniciam ou finalizam) são relevantes para a previsão sobre se haverá ou não estouro de orçamento. O primeiro modelo evidencia o risco de ocorrer estouro de orçamento em projetos relacionados a contratações com valores licitados (portanto, provavelmente valores de altas cifras), em projetos que envolvem serviços, e em projetos de alto custo. O segundo modelo reforça a evidencia de problemas relacionados a projetos de alto custo, contudo apresenta regras mais específicas e fortemente baseadas em atributos referentes ao período em que eles são realizados. Infelizmente, aparentemente não há uma explicação lógica para as decisões baseadas na informação sobre meses de início e fim dos projetos. Porém, a observação dessa informação no modelo pode ser um indicativo de que há alguma relação entre problemas ocorridos na execução dos projetos e fatores sazonais.

4.2 Estudo para Estouro de Prazo

As Figuras 6 e 7 apresentam árvores de decisão, geradas pelo WEKA, para os experimentos que consideraram os algoritmos de árvore de decisão J48 e SimpleCART, respectivamente, para estudo do estouro de prazo[5]. Ambas as árvores apresentam estrutura simples, com profundidade 2. Contudo, diferentemente dos modelos criados para estudo de estouro de orçamento, nesse caso, cada um dos modelos levou em consideração atributos diferentes como sendo relevantes à tomada de decisão. Enquanto o primeiro modelo (J48) foi construído com base nos atributos *tipo do projeto* e *valor total (cat)*, o segundo (SimpleCART) foi construído com base nos atributos *data de fim prevista (mês)* e *número de atividades do projeto*. Assim, os dois modelos indicam o uso de diferentes atributos na decisão se haverá ou não estouro de prazo, e portanto podem ser usados de forma complementar.

```
tipoprojeto = LicitacaoCP: 1 (13.0/4.0)
tipoprojeto = Obras: 1 (8.0)
tipoprojeto = Reforma: 0 (9.0/4.0)
tipoprojeto = Construcao: 1 (9.0/1.0)
tipoprojeto = EquipeInterna: 1 (1.0)
tipoprojeto = Pav./Drenag./Prep.: 0 (6.0/1.0)
tipoprojeto = Ciclovia: 1 (2.0)
tipoprojeto = GestaoUrbana
|   Valortotalclasses = MuitoAlto: 1 (1.0)
|   Valortotalclasses = SemValor: 1 (1.0)
|   Valortotalclasses = Medio: 0 (6.0/3.0)
|   Valortotalclasses = Baixo: 0 (6.0/1.0)
tipoprojeto = ConvenioDADE: 1 (3.0/1.0)
tipoprojeto = Saude: 0 (5.0/2.0)
tipoprojeto = Outros: 1 (9.0/4.0)
tipoprojeto = PlanodeAcao: 1 (3.0/1.0)
tipoprojeto = Educacao: 0 (2.0/1.0)
tipoprojeto = Praca: 1 (1.0)
tipoprojeto = TAC: 0 (3.0)
tipoprojeto = Servicos: 0 (65.0/5.0)
tipoprojeto = ComissaodeObras: 1 (1.0)
tipoprojeto = Treinamento: 1 (1.0)
```

Figura 6: Árvore de decisão gerada no estudo de estouro de prazo com o algoritmo J48.

Verificou-se que o segundo modelo indica que projetos terminando em meses próximos a período de férias sofrem um atraso de execução. Essa informação parece fazer mais sentido quando somada ao fato da cidade onde esses projetos estão sendo executadas ser uma cidade turística. De fato, durante os meses de férias, a cidade se volta ao atendimento dos serviços turísticos, o que acarreta problemas com projetos que ainda estavam por finalizar nessa época.

[5]Para o caso da predição sobre ocorrência de estouro de prazo, o balanceamento dos dados não foi realizado.

```
InicioPrevisto=(Novembro) | (Fevereiro) | (Abril)
|   tipoprojeto=(GestaoUrbana) | (Outros) | (Servicos) | (Construcao) | (LicitacaoCP) | (Obras) | (EquipeInterna) | (Pav./Drenag./Prep.) | (PlanodeAcao) | (TAC) | (ConvenioDADE) | (ComissaodeObras)
|   |   FimPrevisto=(Abril) | (Marco) | (Maio) | (Julho) | (Outubro) | (Janeiro) | (Dezembro) | (Setembro): 1(23.0/5.0)
|   |   FimPrevisto!=(Abril) | (Marco) | (Maio) | (Julho) | (Outubro) | (Janeiro) | (Dezembro) | (Setembro)
|   |   |   ValorlicitadoBinario=(1): 1(8.0/1.0)
|   |   |   ValorlicitadoBinario!=(1): 0(12.0/5.0)
|   tipoprojeto!=(GestaoUrbana) | (Outros) | (Servicos) | (Construcao) | (LicitacaoCP) | (Obras) | (EquipeInterna) | (Pav./Drenag./Prep.) | (PlanodeAcao) | (TAC) | (ConvenioDADE) | (ComissaodeObras): 0(6.0/0.0)
InicioPrevisto!=(Novembro) | (Fevereiro) | (Abril)
|   tipoprojeto=(Reforma) | (Pav./Drenag./Prep.) | (Saude) | (Servicos) | (Construcao) | (Praca) | (Treinamento)
|   |   Valortotalclasses=(MuitoAlto)
|   |   |   InicioPrevisto=(Marco) | (Junho) | (Dezembro) | (Maio) | (Outubro) | (Janeiro) | (Fevereiro) | (Abril) | (Agosto) | (Novembro): 1(10.0/3.0)
|   |   |   InicioPrevisto!=(Marco) | (Junho) | (Dezembro) | (Maio) | (Outubro) | (Janeiro) | (Fevereiro) | (Abril) | (Agosto) | (Novembro): 0(4.0/1.0)
|   |   Valortotalclasses!=(MuitoAlto): 0(35.0/8.0)
|   tipoprojeto!=(Reforma) | (Pav./Drenag./Prep.) | (Saude) | (Servicos) | (Construcao) | (Praca) | (Treinamento)
|   |   InicioPrevisto=(Dezembro)
|   |   |   ValorProprioBinario=(1): 1(2.0/0.0)
|   |   |   ValorProprioBinario!=(1): 0(2.0/0.0)
|   |   InicioPrevisto!=(Dezembro): 0(43.0/1.0)
```

Figura 5: Árvore de decisão gerada no estudo de estouro de orçamento com o algoritmo SimpleCART.

```
FimPrevisto=(Julho) | (Janeiro) | (Dezembro) | (Maio) | (Junho) | (Fevereiro): 1(38.0/16.0)
FimPrevisto!=(Julho) | (Janeiro) | (Dezembro) | (Maio) | (Junho) | (Fevereiro)
|   Quantidadedeatividades=(7) | (9) | (26) | (48) | (3): 1(10.0/1.0)
|   Quantidadedeatividades!=(7) | (9) | (26) | (48) | (3): 0(79.0/11.0)
```

Figura 7: Árvore de decisão gerada no estudo de estouro de prazo com o algoritmo SimpleCART.

5. CONCLUSÃO

Neste artigo, foi apresentado um estudo de caso referente à execução de um processo de descoberta de conhecimento no contexto de gestão de projetos da administração pública. Árvores de decisão foram geradas para predizer a ocorrência de estouro de orçamento e estouro de prazo a partir da análise de informações referentes ao planejamento de projeto.

As principais dificuldades desse estudo se referem à escassez de dados. Ainda que diante do uso de um SGP, e a prefeitura onde o sistema foi implantado tendo sido capaz de gerar um número importante de dados sobre projetos, muito do potencial dessa base de dados não pode ser explorada, pois existe uma dificuldade referente uso do SGP por seus usuários que leva à inserção de inconsistências e ruídos na base de dados. Assim, dados que inicialmente pareciam bastante promissores – aquelas referentes à análise de risco e planos de comunicação, por exemplo – infelizmente não puderam ser usados, visto que, embora o SGP apoiasse a inserção de tais dados, eles não foram inseridos por seus usuários.

Os experimentos executados são ainda iniciais e não representam um estudo exaustivo sobre as possibilidades inerentes à essa iniciativa. Porém, os resultados obtidos demonstram um grande potencial de descoberta de conhecimento útil para o contexto envolvido a partir da exploração de dados referentes à gestão de projetos apoiada pela metodologia PMBOK por meio de um sistema automatizado. Diante do potencial demonstrado, conclui-se que esta é uma área de estudo que pode levar a resultados importantes à gestão de projetos na administração pública.

6. REFERÊNCIAS

[1] *A Guide to the Project Management Body of Knowledge: PMBOK(R) Guide*. Project Management Institute, 5th edition, 2013.

[2] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer. Smote: Synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research*, 16:321–357, 2002.

[3] U. Fayyad. *From Data Mining to Knowledge Discovery in Databases*. American Association for Artificial Inteligence, 1996.

[4] C. Frizzarini and M. S. Lauretto. Proposta de um algoritmo para indução de Árvores de classificação para dados desbalanceados. In *Annais do X Simpósio Brasileiro de Sistemas de Informação*, pages 722–733, maio 2013.

[5] J. Han and M. Kamber. *Data Mining: Concepts and Techniques*. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 2000.

[6] H. Kerzner. *Gestão de Projetos: As Melhores Práticas*. Bookman Companhia, 2006.

[7] A. C. A. Maximiano. *Administração de Projetos: Como Transformar Ideias em Resultados*. Editora Atlas, 2008.

[8] D. S. Prado. *Planejamento e controle de projetos*. Nova Lima, 2004.

[9] S. Raynal. *A gestão por projecto*. Inst. Piaget, 1996.

[10] L. Rokach and O. Maimon. *Data Mining with Decision Trees: Theory and Applications*. World Scientific Publishing Company, Singapore, 2008.

[11] I. H. Witten, E. Frank, and M. A. Hall. *Data Mining: Practical Machine Learning Tools and Techniques*. Morgan Kaufmann, 3rd edition, 2011.