Rastro-DM: Mineração de Dados com Rastro

METODOLOGIA PARA DOCUMENTAÇÃO DE PROJETOS DE MINERAÇÃO DE DADOS E SUA APLICAÇÃO NA CONSTRUÇÃO DE UM CLASSIFICADOR TEXTUAL DE DOCUMENTOS ASSOCIADOS A DANOS AO ERÁRIO PÚBLICO

MARCUS VINÍCIUS BORELA DE CASTRO [1]
Auditor do TCU desde 1996. Bacharel em Informática pela Universidade Federal de Viçosa (1990) e especialista em Governança de TI pela Universidade de Brasília (2012) e em Análise de Dados pela Escola Superior do TCU (2019).

REMIS BALANIUK
Auditor do TCU desde 1989. Graduado em Ciência da Computação pela Unb(1986), mestre em Ciência da Computação pela UFRGS (1989), doutor em Informática pelo Institut National Polytechnique de Grenoble (1996) e pós-doutor em Computação pela Stanford University (2002), pelo Institut National pour la Recherche en Informatique et Automatique (2000) e pesquisador visitante da University of Oxford (2020)

RESUMO

Este trabalho propõe uma metodologia de documentação de projetos de mineração de dados (DM), Rastro-DM, com foco não no modelo gerado, mas no processo por trás de sua construção, de forma a deixar um rastro das ações planejadas, dos treinamentos realizados, dos resultados obtidos e dos aprendizados concebidos. As práticas propostas são complementares às metodologias estruturantes de DM, como o CRISP-DM, que trazem todo o arcabouço metodológico e paradigmático para o processo de DM. Ilustra-se o seu uso e seus benefícios em um projeto de classificação textual de documentos em PDF associados a danos ao Erário Público Federal Brasileiro denominado Cladop. A produção do Rastro-DM de um projeto é um pequeno passo que pode levar a um salto organizacional, a ser obtido com a partilha e o uso dos rastros de forma corporativa.

1 Trata-se de uma versão sintetizada do trabalho de conclusão de curso de pós-graduação lato sensu em Análise de Dados indicado na referência (Castro, 2019).

Palavras-chave: Mineração de dados; Análise de dados; Ciência de dados; Aprendizado de máquina; Conhecimento organizacional; Metodologia; Boas práticas; Análise de dados no Governo; Classificação textual; Documentação; Documentação de projetos de mineração de dados.

1. INTRODUÇÃO

Os dados estão mudando tudo e a capacidade de manipulá-los e entender Ciência de Dados está se tornando cada vez mais crítica para atuais e futuras descobertas e inovações (BERMAN *et al*, 2018).

Projetos de mineração de dados (DM) são desafiadores não só pelo complexo processo usado, exploratório, mas também por, em geral, serem inovadores, únicos e muitas vezes desenvolvidos por indivíduos ou pequenas equipes.

Trata-se de projetos inovadores, quer por usarem técnicas e algoritmos que podem não estar consolidadas ou na Organização ou em pesquisas acadêmicas, quer por envolverem a construção de modelos que simulam processos cognitivos, inteligência natural, por máquinas.

São projetos quase sempre únicos. As particularidades de cada contexto, dos dados envolvidos, dos requisitos de qualidade, impedem ou dificultam o reaproveitamento de código para outros projetos.

São projetos complexos pois as técnicas empregadas em geral têm concepção de difícil entendimento e envolvem conhecimento interdisciplinar de áreas como ciência da computação, matemática e estatística, além do entendimento do negócio para o qual a solução se destina.

São processos exploratórios, pois a atividade de mineração de dados pode ser definida como o processo de explorar um conjunto de dados com técnicas diversas extraindo ou ajudando a evidenciar padrões e auxiliando na descoberta de conhecimento.

E, no caso de organizações com baixo grau de maturidade em DM, são projetos que ficam sob a responsabilidade de pequenas equipes ou mesmo de um único indivíduo. Nesse caso, a partilha do conhecimento e de práticas adotadas nos projetos fica ainda mais difícil.

Infelizmente, esses trabalhos complexos, de caráter exploratório, inovadores, únicos, e, em geral, individuais, não deixam rastro do que fazem. Ao final temos a solução implementada. Pode-se ter até uma documentação do produto criado, mas não do processo seguido, das escolhas feitas e das técnicas usadas nas diversas atividades do projeto. Resta um temor no ar para o caso de ser necessário reconstruir o modelo sem a presença do seu criador. E o responsável se torna pai do produto, pois só ele o conhece e pode dar manutenção.

Uma questão intrigante na gestão do conhecimento em geral é como coletar, colher ou tornar explícita a experiência de projetos para que possam ser utilizáveis para outros (DINGSØYR *et al*, 2001). Pois a memória de uma organização não pode se basear apenas na memória de seus indivíduos (STATA, 1980).

Diante do exposto surge a desafiadora questão: como sistematizar a documentação das tarefas de um projeto de mineração de dados de forma a potencializar sua auditabilidade e a partilha dos aprendizados?

Buscando responder à questão colocada, o objetivo deste trabalho é propor uma metodologia enquanto conjunto de boas práticas de registro semi-automatizado das atividades de um projeto de DM de forma a deixar um rastro das escolhas feitas, dos processamentos realizados e dos resultados obtidos, com foco não no produto gerado, mas no processo por trás de sua construção. Boas práticas que podem ser mescladas à metodologia corporativa de DM em uso na organização. A produção de rastros em projetos de DM acelera a curva de aprendizagem e a formação de uma cultura organizacional em torno do uso da análise de dados.

São três os objetivos específicos: contextualização do referencial teórico sobre metodologias e documentação em projetos de mineração de dados bem como do potencial impacto das experiências adquiridas nos projetos para uma organização; proposição do Rastro-DM com a descrição de suas atividades; e a ilustração de sua aplicação em um projeto de classificação textual de documentos em PDF associados a danos ao Erário Público Federal Brasileiro denominado Cladop, demonstrando sua viabilidade e os benefícios de seu uso. Os objetivos específicos serão tratados nas três seções do desenvolvimento.

2. DESENVOLVIMENTO

2.1 REFERENCIAL TEÓRICO

2.1.1 METODOLOGIAS DE MINERAÇÃO DE DADOS

Segundo BERMAN *et al* (2018), a Ciência de Dados se concentra nos processos de extração de conhecimento ou *insights* a partir de dados estruturados ou não. Esse processo de descoberta de conhecimento em dados (KDD - *Knowledge-Discovery in Databases*), segundo BECKER e GHEDINI (2005), é complexo e popularmente chamado de Mineração de Dados (DM - *Data Mining)*. No escopo deste trabalho chamaremos indistintamente KDD de DM.

CHAPMAN *et al* (2000) detalham alguns objetivos de DM traduzidos em grupos de problemas tratados: descrição e sumarização de dados, segmentação (clusterização), descrição de conceitos, classificação, predição (regressão) e análise de dependência.

WIRTH e HIPP (2000) confirmam que DM é um processo complexo e associam o sucesso de um projeto a uma adequada combinação de boas ferramentas, analistas qualificados, uso de uma metodologia sólida e um eficaz gerenciamento de projetos. Quanto à metodologia, os mesmos autores afirmam que DM precisa de uma abordagem padrão que ajude a transformar problemas de negócios em tarefas de mineração de dados, que sugira transformações de dados apropriadas e técnicas a empregar, e forneça meios para avaliar a eficácia dos resultados

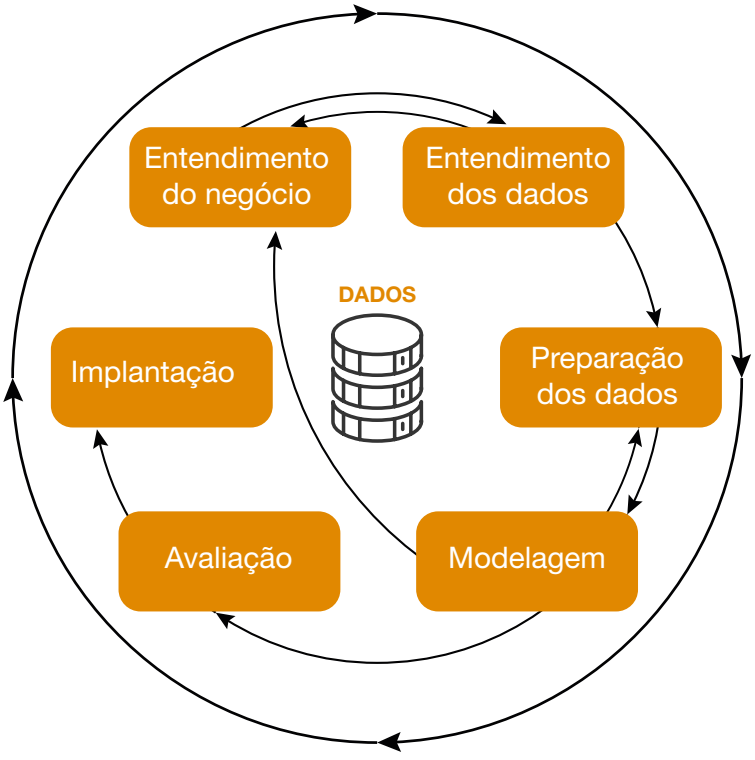
e documentar a experiência. E destacam que o uso de uma metodologia no planejamento e na apresentação de relatórios inspira confiança nos usuários e nos patrocinadores.

KURGAN e MUSILEK (2006) afirmam que metodologias acrescentam uma melhor compreensão e entendimento do processo, além de promoverem economia de tempo e de custos, com o estabelecimento de um roteiro a seguir para o planejamento e a execução dos projetos. Mas alertam que um grande número de projetos segue metodologias próprias.

Um exemplo de metodologia usada em DM é o CRISP-DM (*Cross-Industry Standard Process for Data Mining*). Ela é considerada padrão de fato e um dos fatores de seu sucesso é ser neutra em termos de indústria, ferramentas e aplicações (MARISCAL *et al*, 2010).

CHAPMAN *et al* (2000) atestam que a metodologia CRISP-DM é um modelo de processo hierárquico, consistindo em conjuntos de tarefas descritas em quatro níveis de abstração (do geral ao específico): fase, tarefa genérica, tarefa especializada e instância do processo. Afirmam que a sequência das fases não é rígida. Que, na prática, muitas das tarefas podem ser executadas em uma ordem diferente, e, muitas vezes, faz-se necessário voltar repetidamente para tarefas anteriores e repetir certas ações. Ressaltam que representar todas as rotas possíveis por meio do processo de mineração de dados exigiria um modelo de processo excessivamente complexo, por isso não o fazem. Afirmam que nunca uma fase é completamente concluída antes que a fase seguinte comece. A Figura 1 mostra a iteração entre as fases do CRISP-DM.

Figura 1 - Visão geral do CRISP-DM.



Fonte: CHAPMAN *et al* (2000).

Segundo MARBÁN *et al* (2007), o CRISP-DM não cobre muitas tarefas relacionadas à gestão, organização e qualidade de projetos. Pelo menos não da maneira exigida pela crescente complexidade dos recentes projetos de DM que envolvem não apenas grandes volumes de dados, mas também o gerenciamento e a organização de grandes times interdisciplinares. Os mesmos autores agrupam as tarefas de um processo de DM, em relação à construção de um modelo, em três estágios: pré-desenvolvimento, desenvolvimento e pós-desenvolvimento. Afirmam que todas as metodologias usadas para DM se concentram no estágio de desenvolvimento, que equivale a coleta e a análise dos dados disponíveis para o projeto, a criação de novos dados a partir dos disponíveis, a adaptação para algoritmos de DM e a criação de modelos.

Segundo CHOLLET (2017), o desenvolvimento de um projeto está centrado em experimentações de um modelo: você começa com uma ideia e a expressa como um experimento, tentando validar ou invalidar sua ideia. Depois, executa-se essa experiência e processam-se as informações geradas. Isso inspira sua próxima ideia. Quanto mais iterações desse círculo repetitivo você conseguir executar, mais refinadas e poderosas suas ideias se tornarão. Ressalta a importância de se obter o máximo possível de informações das experimentações, incluindo o desempenho dos modelos.

A experimentação, ou seja, a aplicação de algoritmos matemáticos aos dados para a extração de padrões, é chamada de treinamento (training) por NGUYEN (2018). Para simplificar, no contexto deste trabalho, serão usados os termos treinamento e modelo representando a experimentação e o produto resultante dos padrões detectados nos dados.

GREFF *et al* (2017) confirmam o número significativo de experimentos computacionais com muitas configurações diferentes de hiperparâmetros e alertam sobre o desafio prático da documentação. Devido à pressão de prazo e à inerente natureza imprevisível de um projeto de DM, constatam que há pouco incentivo para a construção de infraestruturas robustas e, como resultado, o código muitas vezes evolui rapidamente, o que compromete, entre outras coisas, a documentação do projeto.

BECKER e GHEDINI (2005) acentuam que a estrutura de um processo de DM é altamente dependente da metodologia adotada, das habilidades, da experiência e do estilo da pessoa responsável pelo processo, bem como dos recursos disponíveis na corporação. E confirmam a alta iteratividade e interatividade dos processos. Ressaltam que embora a estrutura conceitual do processo sugira uma ordem entre as fases, na prática, os analistas passam de qualquer fase para quase qualquer outra fase a qualquer momento, até porque muitos problemas relacionados a fases anteriores (por exemplo, preparação de dados) só podem ser detectados muito mais tarde, quando os padrões e os modelos são avaliados. As mesmas autoras afirmam que os projetos são desenvolvidos em geral de maneira não estruturada, ad hoc: após a análise inicial dos dados, decide-se experimentar uma determinada técnica, cujos resultados podem sugerir a reestruturação dos dados e a execução de novos tipos de análises; e assim por diante. E, à medida que o tempo passa, não se pode lembrar quais treinamentos foram realizados, os conjuntos de dados utilizados, os hiperparâmetros usados e, mais importante, os resultados que

foram derivados dos conjuntos de dados. E essa situação, segundo as autoras, leva à reexecução de treinamentos. Alertam que a situação é ainda pior se forem considerados projetos de longo prazo envolvendo várias pessoas. Constatam que, apesar da diversidade de conhecimento das pessoas, das técnicas e das ferramentas, a maioria dos projetos de DM enfrenta, na prática, as mesmas dificuldades: o desperdício de se refazer um trabalho e o gerenciamento de recursos e de resultados. Afirmam as autoras que a documentação do histórico das tarefas face a iteratividade e a interatividade do processo é um problema aberto no gerenciamento de projetos de DM.

2.2 DOCUMENTAÇÃO: CAMINHO PARA A GERAÇÃO DE CONHECIMENTO PARTILHÁVEL

GHEDINI e BECKER (2000) ressaltam que a documentação de experimentos e de todas as partes relevantes de um projeto não só evita a perda de conhecimento confinado nas mentes das pessoas como também permite o seu compartilhamento, tornando-se uma rica fonte de conhecimento para referência futura e reuso corporativo. Destacam também que essa atividade leva a um melhor gerenciamento de esforços, de recursos e de resultados de um projeto de DM.

Segundo PRAKASH *et al* (2012), informações de treinamentos e código de implementação (e seu histórico de mudança) contêm uma riqueza de informações sobre o estado, o progresso e a evolução de um projeto de software. Afirmam também que DM está se tornando uma ferramenta cada vez mais importante para transformar esses dados em informações. Por analogia, espera-se que dados de projetos de DM sejam convertidos também em informações por atividades de mineração de dados.

WIRTH e HIPP (2000) enfatizam que, talvez, o maior benefício de terem aplicado uma metodologia foi a documentação gerada. Admitem terem pulado inicialmente algumas tarefas de documentação e planejamento por serem demoradas e por considerarem desnecessárias para especialistas como eles. Mas apresentam o preço que pagaram por essa ação e se arrependem constatando que todo esforço vale a pena. Relatam alguns benefícios observados a partir da documentação produzida: evita o desperdício de esforço (por exemplo, em caminhos não frutíferos ou com trabalho repetitivo); promove um gerenciamento eficaz e uma melhor comunicação da equipe; permite a identificação de pontos críticos no processo; promove um melhor planejamento de projetos futuros, com base em uma melhor percepção de como o esforço foi gasto e dos recursos necessários; promove o uso de experiências documentadas em outros contextos.

BECKER e GHEDINI (2005) também identificam o papel da documentação na aprendizagem e na reutilização e afirmam que um benefício imediato da documentação é a efetividade no gerenciamento, no planejamento e na comunicação. Constatam que a documentação é completamente dependente da equipe do projeto, uma vez que está em sua responsabilidade a veracidade dos registros e o nível de detalhe que impacta diretamente sua utilidade. Quanto à resistência, afirmam que, à medida que os benefícios da atividade são percebidos, há um estímulo a documentar o processo com mais detalhes e de forma

concomitante com a execução das atividades e que os melhores resultados são alcançados a longo prazo, quando a equipe do projeto descobre qual estratégia melhor se adequa ao seu estilo de trabalho, bem como as melhores maneiras de obter vantagens dos recursos e das técnicas e da flexibilidade do modelo.

Contudo, WIRTH e HIPP (2000) alertam sobre a dificuldade de se documentar ao final, de se tentar reconstruir o que foi feito e suas motivações. Enfatizam que os processos de DM são vivos e, como tal, a documentação deve ser flexível e viva, e não deve ser atualizada após o final do projeto (post mortem). Defendem que a definição de uma estratégia de documentação deve ser um ponto de partida, mas a flexibilidade para a evolução e a mudança deve ser uma premissa. Enfatizam que encontrar o nível certo de detalhes para se planejar e se documentar um processo de DM é difícil e faz parte de um longo processo de aprendizado, e pode ser influenciado por diversos fatores como complexidade do projeto, duração e tamanho da equipe.

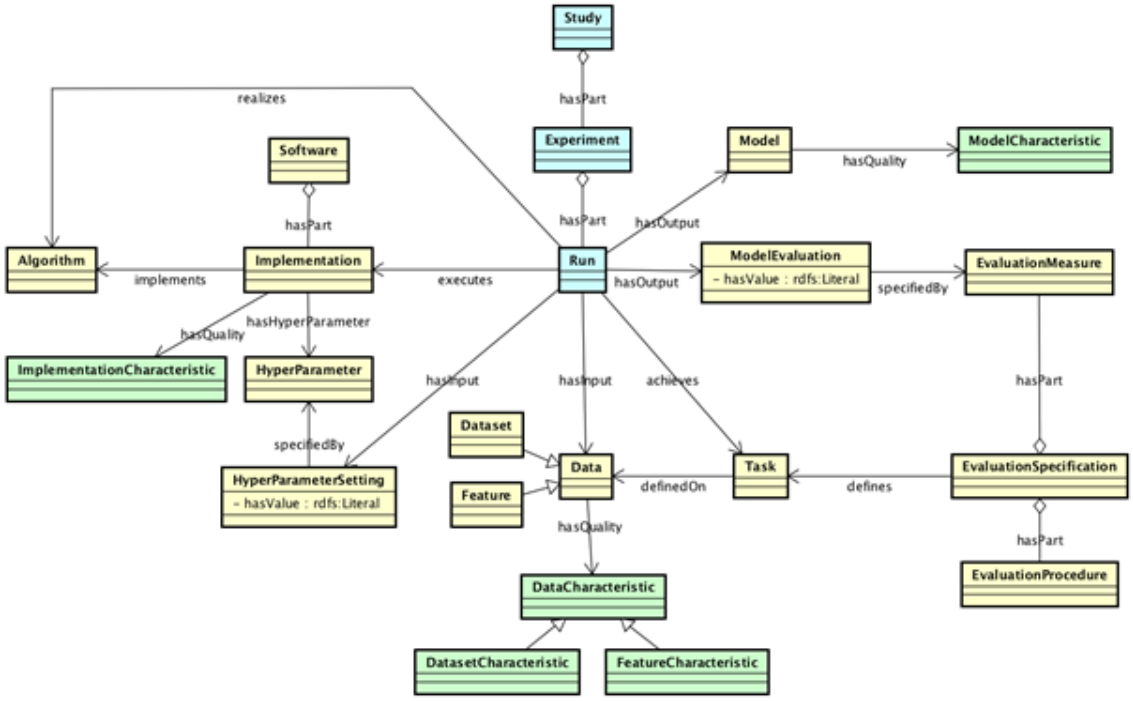
GREFF *et al* (2017) citam que uma dificuldade para o compartilhamento e a colaboração e a reprodutibilidade dos treinamentos é o uso pelas equipes de configurações particulares de área de trabalho no uso das diferentes ferramentas necessárias para tratar os aspectos distintos de um processo de DM, citando, entre outras, bancos de dados, sistemas de controle de versão, ferramentas automatizadas de otimização de hiperparâmetros, scripts e planilhas. Os mesmos autores apresentam Sacred, um framework Python de código aberto que visa fornecer infraestrutura básica para a execução de experimentos computacionais de forma independente dos métodos e bibliotecas utilizados. Concentram-se em resolver problemas como o gerenciamento de configurações, a documentação e a reprodutibilidade dos resultados. Para cada treinamento, informações relevantes, como parâmetros, dependências de pacotes, informações do host, código-fonte e resultados, são capturadas automaticamente e armazenadas em um repositório centralizado, de onde pode-se consultar também detalhes dos hiperparâmetros usados e dos resultados obtidos.

PRAKASH *et al* (2012) relatam que numerosos trabalhos estão sendo feitos no desenvolvimento de plataformas integradas para Machine Learning (ML) e para Engenharia de Software baseadas em componentes reutilizáveis, citando, entre as de código aberto mais conhecidas, WEKA e *Rapid Miner*.

PUBLIO *et al* (2018) acreditam que a visão de modelos canônicos e padronizados pode levar a uma melhor compreensão dos dados e dos algoritmos de ML empregados em DM e pode promover a interoperabilidade dos experimentos, independentemente da plataforma ou da solução de fluxo de trabalho adotada.

W3C *Machine Learning Schema Community Group* (2017) publicou uma ontologia ML *Schema* que fornece um conjunto de classes, propriedades e restrições para representar e intercambiar informações sobre algoritmos de aprendizado de máquina, conjuntos de dados e experimentos. Segundo o grupo, a ontologia pode ser facilmente estendida e mapeada para outras ontologias mais específicas de domínio desenvolvidas na área de aprendizado de máquina e de mineração de dados. A Figura 2 traz a visualização das principais abstrações do ML *Schema*.

Figura 2: Conceitos principais do ML *Schema*.



Fonte: W3C *Machine Learning Schema Community Group* (2017).

De forma resumida, o conceito central é o treinamento (Run) que, no contexto de um experimento (*Experiment, Study*), produz um modelo (*Model*), com suas características (ModelCharacteristic), e um conjunto de avaliações de qualidade (*ModelEvaluation*) que consideram métricas padronizadas (*EvaluationMeasure*, *EvaluationSpecification*, *EvauationProcedure*). O treinamento é uma execução de uma implementação (*Implementation*) de um algoritmo (*Algorithm*) em uma plataforma (*Software, ImplementationCharacteristic*) usando uma configuração específica de parâmetros (*HyperParameter, HyperParameterSetting*). O treinamento tem como entrada um conjunto de dados (*Data, Dataset, Feature*) que têm suas características próprias (*DataCharacteristic , DatasetCharacteristic, Featurecharacteristic*) e são utilizados para uma tarefa específica (*Task*).

KURGAN e MUSILEK (2006) afirmam a possibilidade de integração e de interoperabilidade dos modelos de DM com o uso de padrões industriais como PMML (*Predictive Model Markup Language*) que representa um modelo em um esquema XML (*Extensible Markup Language*). Segundo os autores, pode-se usar ferramentas diferentes para a geração, visualização e análise de um mesmo modelo.

2.3 CONHECIMENTO EM PROJETO DE DM: UM POTENCIAL SALTO ORGANIZACIONAL

WIRTH e HIPP (2000) salientam que o sucesso ou o fracasso de um projeto de mineração de dados é altamente dependente da pessoa ou da equipe e que práticas de sucesso não são necessariamente repetidas em toda a empresa.

CHAPMAN *et al* (2000) ressaltam que projetos de mineração de dados podem ser beneficiados com as experiências de projetos anteriores. As lições aprendidas durante o processo e a partir da solução implantada podem desencadear novas questões de negócios, muitas vezes mais focadas.

BECKER e GHEDINI (2005) exemplificam benefícios de se partilhar conhecimento entre projetos citando que experiências de projetos anteriores podem ser usadas para estabelecer planos de projetos de DM mais razoáveis, com estimativas mais precisas de cronogramas, orçamento, etc. A experiência torna mais fácil defender recursos mais realistas, pois há uma compreensão mais profunda de como o esforço é realmente gasto. Além disso, pode-se usar experiências anteriores para lidar com certas classes de problemas ou de técnicas. Defendem que a documentação da execução de projetos deve ser tratada como um recurso corporativo, que pode ser compartilhado pela equipe, ser usado como referência e pode estar sujeito a políticas e padrões corporativos.

Segundo BHATT (2001), à medida que os indivíduos nas organizações interagem com outros, eles tendem a entender e partilhar suas visões diferentes sobre as mesmas situações, construindo suas comunidades e compartilhando técnicas eficientes de trabalho e facilitando a integração de um corpo diversificado de conhecimentos nas organizações. O mesmo autor afirma que o conhecimento organizacional é formado por padrões únicos de interações entre tecnologias, técnicas e pessoas, que não podem ser copiados facilmente, porque essas interações são únicas da organização, moldadas pela sua história e pela sua cultura. Credita o mesmo autor a sustentação de vantagens competitivas da empresa no longo prazo ao incentivo ao crescimento desse conhecimento com a criação de um ambiente estimulante e prático (aprender-fazendo).

DINGSØYR et al (2001) tratam de forma indistinta conhecimento e experiência. Embora reconheçam que experiência em um sentido estrito é algo que reside nos seres humanos e que não pode ser transferido para os outros (que teriam que experimentar por si mesmo para ter a experiência), em uma definição menos estrita, afirmam que experiência é informação que é operacional, isto é, utilizável em alguma situação. Entendem que uma descrição de um evento acontecido em um projeto é um item de experiência.

Como comentado na introdução, uma questão interessante na gestão do conhecimento em geral é como coletar, colher ou tornar explícita a experiência de projetos para que possam ser utilizáveis para outros (DINGSØYR *et al*, 2001).

NGUYEN (2018) defende o uso de processos de gestão do conhecimento (KMP - *Knowledge Management Process*) para fazer circular o conhecimento em toda a organização para garantir que o conhecimento certo chegue à pessoa certa para entender e ter conhecimento suficiente para tomar decisões e executar bem as tarefas. Afirma o autor que KMP pode ser usado em qualquer nível, desde a organização como um todo até dentro da equipe. E complementa que seus estágios (identificação, criação, armazenamento, transferência e utilização de conhecimento) estão interligados e são iterativos, uma vez que o conhecimento é continuamente formado e alterado. O autor encontra na combinação entre KMP e DM um grande potencial na exploração e no gerenciamento do conhecimento valioso de big data: DM suporta o KMP na geração de conhecimento inestimável e KMP suporta DM na coleta e

armazenamento de conhecimento como entrada de DM. Mas ressalta que há um grande vazio de pesquisas nessa área.

STATA (1980) traz o conceito de aprendizado organizacional (OL - Organizational Learning), que ocorre por meio de insights compartilhados, conhecimento e modelos mentais e que se baseia no conhecimento e na experiência do passado, ou seja, na memória. O autor detalha que a memória organizacional depende de mecanismos institucionais (por exemplo, políticas e estratégias) usados para reter o conhecimento, não podendo depender exclusivamente da memória dos indivíduos, pois há sempre o risco de se perder lições e experiências duramente conquistadas à medida que as pessoas migram de um emprego para outro. Além de outros motivos de saída de pessoas (aposentadoria, afastamentos, transferências, falecimento, etc.), podemos acrescentar também o risco do esquecimento.

DINGSØYR *et al* (2001) defendem que seja elaborado ao final de um projeto um Relatório de Experiência para coletar o que deu certo e o que deu errado no processo adotado. A atividade Revisar Projeto do CRISP-DM produz relatório semelhante que conforme descrição em CHAPMAN *et al* (2000): Resuma a importante experiência adquirida durante o projeto. Por exemplo, armadilhas, abordagens enganosas ou dicas para selecionar as técnicas de mineração de dados mais adequadas em situações semelhantes poderiam fazer parte dessa documentação.

Segundo BECKER e GHEDINI (2005), com a documentação do processo de DM, à medida que o conhecimento é tornado explícito e gerenciado, ele aumenta o intelecto da organização, tornando-se uma base para a comunicação e para a aprendizagem, apoiando a disseminação de conhecimento e a experiência dentro da organização em vários níveis. As autoras reforçam que a ideia de capturar e armazenar todo o conhecimento informal relevante gerado e usado durante um processo de DM, de modo que esteja disponível para recuperação posterior, constitui uma abordagem interessante para lidar com a dificuldade citada de refletir na documentação a iteratividade e a interatividade do processo.

É fato que as empresas precisam educar um número maior de pessoas sobre os processos e as melhores práticas associadas de DM (MARISCAL *et al*, 2010).

Segundo NGUYEN (2018), através de processos e práticas, pode ser incorporado, em indivíduos e em organizações, o conhecimento, *commodity* cara para as organizações, com origens diversas, como, por exemplo, documentos, processos, pessoas, comunicação, cultura e aprendizagem. O autor afirma que a transferência de conhecimento estimula inovações e que o armazenamento de conhecimento é o caminho para se criar uma propriedade inestimável para as organizações. Um bem que se acumula ao longo do tempo e que não pode ser comprado por dinheiro algum.

Diante do exposto no referencial teórico, pode-se concluir que a documentação da memória de um projeto de DM é um pequeno passo, uma vez que localizado em um contexto menor de um projeto, que pode significar um salto em direção à memória organizacional e aos ganhos derivados do aprendizado partilhado e do conhecimento gerenciado para uma organização. Na próxima seção, o Rastro-DM é apresentado como um caminho em potencial para se alcançar esse salto organizacional.

2.4 RASTRO-DM

2.4.1 VISÃO GERAL

O Rastro-DM é uma metodologia que objetiva a documentação de projetos de DM com foco no processo de construção dos modelos, de forma a deixar um rastro das ações executadas, dos treinamentos realizados e os resultados obtidos e dos aprendizados concebidos. Engloba três atividades que correspondem aos conceitos documentados:

- Definição de ação;
- Registro de treinamento;
- Síntese de aprendizado.

Diferentemente do foco tradicional de se documentar o produto final (e seus artefatos), Rastro-DM foca na documentação do processo por trás da construção dos modelos. Conforme CONKLIN (1996), promove-se assim o aumento da memória organizacional, com o registro do contexto de criação dos artefatos: os pressupostos, os valores, as experiências, o motivo, as conversas e as decisões conduzidas.

Como visto no referencial teórico, todo o conhecimento informal relevante deve ser documentado para refletir na documentação a iteratividade e a interatividade do processo e estar disponível para recuperação posterior.

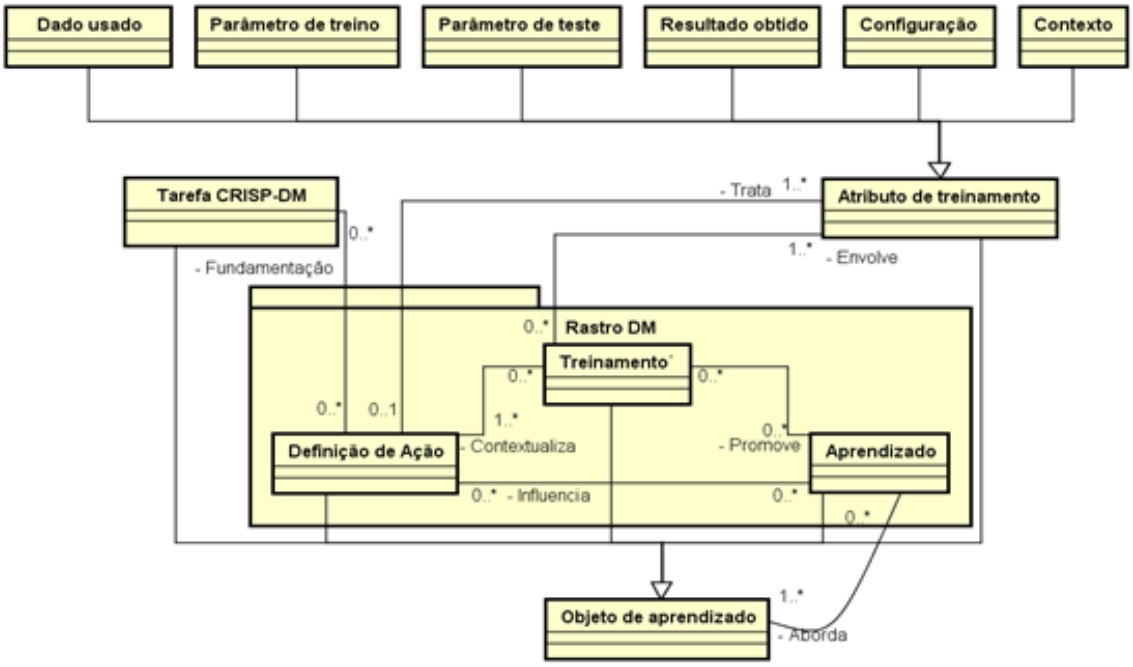
As atividades do Rastro-DM são complementares às tarefas previstas na metodologia em uso na organização, qualquer que seja ela, aqui chamada de metodologia base, que traz todo o arcabouço metodológico e paradigmático usado em um processo de DM. MINGERS e BROCKLESBY (1997) identificam várias maneiras de se combinar metodologias. Afirmam que o estabelecimento de boas práticas é uma forma de criação de uma nova metodologia, logo, não há incongruência em se referir ao Rastro-DM como uma metodologia.

Por padronização e clareza, será usado nas explicações o CRISP-DM como metodologia base, e os passos da metodologia base serão referenciados como tarefas CRISP-DM e os do Rastro-DM como atividades Rastro-DM. As atividades Rastro-DM ocorrem várias vezes durante um projeto e podem estar associadas a uma ou mais tarefas CRISP-DM.

2.4.2 VISÃO INTEGRADA DOS CONCEITOS DO RASTRO

Os conceitos do rastro se relacionam: os treinamentos acontecem no contexto de uma ação definida no projeto e esses treinamentos podem promover aprendizados que por sua vez podem influenciar, em um círculo virtuoso, novas ideias definidas em ações. Alguns relacionamentos possíveis entre os principais conceitos do Rastro-DM podem ser visualizados na Figura 3. O diagrama não tem a pretensão de ser completo, mas apenas promover uma melhor compreensão.

Figura 3: Principais conceitos associados ao Rastro-DM.



Fonte: Castro (2019).

Percebe-se que o conceito Atributo de treinamento, subdivido em seis categorias, pode, além de ser envolvido em treinamentos, ser objeto de definições de ação. As definições de ação podem ter fundamentação em tarefas da metodologia base, representada na figura pelo CRISP-DM. E todos os elementos podem ser abordados em um aprendizado.

A seguir, serão descritas as atividades do Rastro-DM.

2.4.3 DEFINIÇÃO DE AÇÃO

Atividade de registro da definição dos passos de um projeto, executados ou não. O objetivo da atividade não é o registro dos detalhes da execução de uma ação, mas as informações sobre sua definição, como a declaração de seu objetivo e as técnicas a serem usadas ou experimentadas quando da execução da ação.

Conceitualmente, uma definição de ação corresponde à descrição de uma ou mais tarefas específicas do CRISP-DM instanciadas em um projeto.

Diante das dificuldades de gerenciamento em processos de DM identificadas no referencial teórico, pode-se avaliar a possibilidade do registro de recursos usados nas ações definidas (pessoas, tempo, etc.) para apoiar estimativas de prazo e o estabelecimento de cronogramas do projeto. E o histórico desses dados pode servir de base para alocação de recursos em projetos futuros.

A atividade de definição de ação pode ocorrer a qualquer momento de um projeto de DM. É importante pelo menos a definição de cada ação que se inicia, pois, de certa forma, justifica os treinamentos que se seguirão. Além disso, saber o que já foi feito reduz a perda de esforço com execuções repetidas.

O nível de abstração e de detalhe, se mais próximo de uma tarefa CRISP-DM, como, formatação de dados, ou mais detalhada, como, formatação do nome do arquivo para possível retirada de *stopwords* e de pontuações, cabe a cada equipe se não houver um padrão corporativo.

Embora a simplicidade de um campo textual descritivo seja aceitável, uma vez que o conteúdo é o mais importante, quanto mais estruturado for o registro, melhor traduzirá a compreensão do processo e maior será o potencial de sua contribuição. As entidades do ML *Schema* constantes da Figura 2 são exemplos de atributos associados a uma Definição de Ação.

Para clarear os conceitos, usaremos um projeto hipotético de aprendizagem supervisionada que objetiva a predição do preço de uma casa (regressão) para uma imobiliária a partir de alguns atributos do imóvel.

Seguem exemplos de definições de ação do projeto hipotético associados às datas de registro de início da ação.

- 10/10/2018; Experimentar como características do imóvel no modelo o número de pavimentos e o número de banheiros;
- 8/11/2018; Avaliar se o modelo alcança uma melhor performance se os valores da característica distância do centro forem ajustados para outra escala;
- 1/2/2019; Testar os algoritmos *lightgbm* e *randomforest* para a regressão.

A compreensão dos termos técnicos usados em exemplos de registros não é relevante para o objetivo deste artigo. O foco, no caso, é a capacidade de descrição de uma ação e não nos elementos dessa descrição (*lightgbm*, *randomforest*). Contudo, para um conhecimento mais aprofundado sobre técnicas associadas a DM, sugere-se a leitura do livro *Data Mining and Analysis*: *Fundamental Concepts and Algorithms* de ZAKI e MEIRA (2014).

2.4.4 REGISTRO DE TREINAMENTO

Atividade de documentação dos treinamentos realizados, os parâmetros usados e os resultados obtidos. Conforme o referencial teórico, o treinamento é a atividade central de todo o processo de mineração de dados e é importante armazenar o máximo possível de informações relevantes deles.

O ML *Schema* constante da Figura 2 fornece um conjunto de entidades envolvidas em um treinamento e que são candidatas a terem informações documentadas no contexto da atividade de registro de treinamento.

Objetivando uma maior compreensão, sem a intenção de completude da classificação nem dos exemplos, os dados envolvidos em um treinamento podem ser agrupados em 6 categorias:

- Dados usados: dados de teste, dados de treinamento, variáveis usadas como features para o modelo;
- Parâmetros de treinamentos: algoritmo usado, hiperparâmetros considerados, implementações de técnicas aplicadas;
- Parâmetros de testes: métrica considerada, forma de apuração;
- Resultados: modelo com suas características e as métricas apuradas;
- Configuração: identificação do programa, versões das bibliotecas usadas, dados de hardware;
- Contexto: código do treinamento, data e hora, número de épocas de treinamento, mensagem de erro em caso interrupção da execução.

O nível de detalhe de documentação impacta a sua utilidade. O código de um treinamento, dado de contexto, pode ser usado como chave de identificação do modelo treinado.

Seguem exemplos de registros de treinamentos do projeto hipotético de predição do preço de uma casa:

- Código: 1; Data: 7/6/2018; Variáveis usadas: área da casa, área do lote e CEP do endereço; Algoritmo usado: *linear regression*; Erro: 0.8; Separação de dados de teste: 10%, não estratificada;
- Código: 100; Data: 7/7/2018; Variáveis usadas: área da casa, área do lote, CEP do endereço, número de quartos e data da construção; Algoritmo usado: *randomForest*; Erro: 0.7; Separação de dados de teste: 5%, dados estratificados.

2.4.5 SÍNTESE DE APRENDIZADO

Atividade de síntese e registro dos aprendizados concebidos ao longo do projeto, de forma automática ou não.

A atividade pode ocorrer a qualquer momento de um projeto de DM e pode estar associada ou não a treinamentos. Embora possam ser sintetizados no estágio de pré-desenvolvimento do modelo, nas fases CRISP-DM de entendimento do negócio e dos dados, os aprendizados, em sua maioria, são gerados a partir dos treinamentos realizados. Pode-se sintetizar, automaticamente ou não, que uma determinada seleção de variáveis ou que o uso de um determinado hiperparâmetro de uma técnica levou à geração de um modelo de melhor performance. Pode haver aprendizados que envolvem treinamentos que falharam que objetivam documentar como evitar que erros aconteçam novamente.

Um Aprendizado pode ter como atributo informações das entidades do ML *Schema* (Figura 2), mas também pode envolver outros conceitos, como Ações Definidas e Treinamentos.

A síntese de aprendizados e o uso efetivo deles no projeto ou em projetos futuros promovem o amadurecimento da equipe no processo de DM, no conhecimento sobre a alta iteratividade e interatividade de suas tarefas, sobre as várias técnicas e sobre as ferramentas usadas.

A documentação dos aprendizados impede que eles se percam nas memórias dos indivíduos ou mesmo com os indivíduos quando deixam o projeto ou a organização.

Para clarear o conceito, seguem exemplos de possíveis aprendizados concebidos no projeto hipotético de predição de preço de uma casa associados à data de registro e à tarefa CRISP-DM de origem:

- 3/10/2018; Selecionar técnica; A técnica randomforest se mostrou superior à técnica decisiontree no contexto avaliado;
- 26/5/2019; Formatar dados; É necessário que os valores dos imóveis nos dados usados para treinamento sejam atualizados para uma mesma referência monetária;
- 26/7/2019; Selecionar dados; O acréscimo das variáveis número de quartos e número de vagas para automóveis promoveu uma melhora de 10% na acurácia do modelo.

2.5 DIRECIONAMENTOS

Seguem alguns direcionamentos práticos para a efetiva e eficaz aplicação do Rastro-DM.

2.5.1 ADAPTABILIDADE

As atividades do Rastro-DM são complementares às tarefas previstas na metodologia base, que traz todo o arcabouço metodológico e paradigmático usado em um processo de DM. A documentação do rastro, fora do contexto rígido de uma metodologia base, permite tratar melhor a interatividade e iteratividade das tarefas, algo não bem mapeado pelas metodologias de DM, conforme discutido no Referencial Teórico. Como as tarefas e as fases de um processo de DM se misturam e muitas vezes são executadas de forma concomitante, fica difícil se manter uma documentação efetiva por tarefa da metodologia base. Por exemplo, se no contexto de uma tarefa CRISP-DM for produzido um relatório e depois se voltar a essa tarefa diversas vezes, a documentação precisaria ser constantemente atualizada, de forma restrita ao escopo da tarefa. O custo do retrabalho torna impeditiva a documentação tempestiva no CRISP-DM.

A documentação gerada pelo Rastro-DM pode se encaixar nos artefatos de saída das metodologias bases. Os registros podem, em um momento posterior, por exemplo, ser agrupados por tarefa CRISP-DM. Por exemplo, o documento Razões para exclusões e seleções da atividade Seleção de dados da fase Preparação de dados do CRISP-DM pode ser um

relatório construído automaticamente a partir de aprendizados e de definições de ações que tratam seleções de dados, enriquecido com um resumo dos treinamentos relacionados a cada critério de seleção experimentado.

2.5.2 TEMPESTIVIDADE

O registro deve ser tempestivo para ajudar a conduzir o projeto em andamento de forma mais eficaz. Com o adequado registro do rastro, evita-se o desperdício com execuções repetidas de trabalhos no transcorrer do projeto. Como visto, uma documentação post mortem não se adequa a projetos de DM, que são considerados vivos dada sua complexidade e sua iteratividade.

2.5.3 FLEXIBILIDADE

A definição dos atributos a serem armazenados deve ser flexível e pode variar conforme o objetivo de mineração (classificação, regressão, etc), a plataforma de treinamento e os objetivos da documentação. Exemplificando, se a documentação tiver como objetivo a reprodutibilidade, há que se gravar mais detalhes de configuração de software (versões de bibliotecas) e de sementes de números aleatórios usados.

Rastro-DM é flexível ao não definir uma relação mínima de atributos para cada conceito. Afinal cada projeto e organização tem sua complexidade particular e seu grau de amadurecimento em DM. Em última análise, cabe à equipe garantir a veracidade e a efetividade da documentação. Espera-se que, com o amadurecimento em DM, as organizações elaborem um padrão corporativo mínimo de documentação por objetivo de DM. Mas, esse padrão não pode afastar a criatividade da equipe nem se tornar apenas uma sobrecarga de trabalho.

2.5.4 AUTOMAÇÃO

É imprescindível que a atividade de registro de treinamento seja automática e esteja vinculada à realização de treinamento de modelo na plataforma em uso. Ainda que haja um custo inicial de construção do arcabouço de software para uma determinada configuração de ferramentas, esse esforço será implementado uma única vez e reaproveitado nos demais projetos. Como visto, GREFF *et al* (2017) alertam sobre o desafio prático quando não se incentiva a construção dessa estrutura.

É desejável que todas as atividades sejam realizadas de forma integrada às plataformas, para que o esforço não seja uma barreira para a documentação. Afinal, para ser útil, um modelo de documentação deve corresponder, tanto quanto possível, ao modo como as pessoas trabalham (BECKER e GHEDINI, 2005).

Castro (2019) ilustra uma implementação simples de uma infraestrutura de construção de um rastro em python, criado a um baixo custo com arquivos locais.

2.5.5 USABILIDADE

Em relação às definições de ação e aos aprendizados, deve-se priorizar o seu registro em detrimento à sua categorização, de forma que esta atividade não seja uma barreira para a documentação. O registro pode se dar inicialmente até mesmo em formato de texto livre.

A categorização, por exemplo quanto às técnicas envolvidas, deve ser realizada, de preferência, de forma automática e pode ser um requisito corporativo importante para se potencializar a utilidade do rastro através da partilha do conhecimento. Como visto, a combinação de DM para o KMP potencializa o crescimento do conhecimento organizacional (NGUYEN, 2018).

Também deve se pensar na usabilidade das aplicações de consulta ao rastro, que tende a se tornar corporativo. Consultas simples podem, por exemplo, encontrar projetos que experimentaram determinada técnica, identificar dicas de seu uso e avaliar os resultados alcançados, servindo de apoio para novos projetos e para a definição de estratégias de DM na organização.

Na próxima seção será ilustrada a aplicação do Rastro-DM em um projeto de DM.

2.6 PROJETO CLADOP – ILUSTRAÇÃO DE USO DO RASTRO-DM

O projeto Cladop[2] é um exemplo de utilização do Rastro-DM. O termo Cladop é usado tanto para indicar o processo (projeto) quanto o produto construído (Classificador de Documentos em PDF). Consistiu o projeto no desenvolvimento por aprendizagem supervisionada de um classificador automático de tipo para documentos em formato PDF (*Portable Document Format*) inseridos no sistema de gestão de Tomadas de Contas Especiais (e-TCE) do Tribunal de Contas da União (TCU).

Um processo de Tomada de Contas Especial, em última análise, objetiva o ressarcimento ao Erário Público de danos gerados por agentes públicos e a devida responsabilização destes. Com o objetivo de tornar mais célere e eficaz o trâmite do processo de apuração de danos, com a padronização e a otimização de procedimentos, foi desenvolvido pelo TCU com cooperação da CGU o sistema e-TCE, plataforma única de acesso a todas as entidades da Administração Pública que atuam em alguma fase da TCE. Castro (2019) detalha o contexto de negócio e os normativos relacionados.

O classificador desenvolvido promoveu benefícios ao negócio tanto na usabilidade do sistema, ao identificar automaticamente os tipos em apoio aos usuários, quanto na qualidade dos dados, ao garantir uma maior correção dos tipos e uma melhor qualidade do texto dos documentos gerados por OCR (*Optical Character Recognition*). Documentos com classificação correta de

2 Após a finalização do projeto Cladop para o sistema eTCE, projeto abordado neste artigo, iniciou-se a construção de um classificador semelhante para outro sistema: Protocolo Eletrônico. O novo projeto, em andamento, foi chamado de Cladop@Protocolo. Então, deve-se compreender Cladop neste artigo como Cladop@eTCE.

tipo e com um conteúdo textual de melhor qualidade são fundamentais para um rito processual assistido pelo computador.

2.6.1 ENTENDIMENTO DOS DADOS

Durante a construção do Classificador, foram considerados 118.266 documentos relativos a 4.384 danos que se distribuíam em 84 tipos de documento. O tipo Outros era o mais usado com 18,81% dos documentos. O Gráfico 1 permite a visualização parcial do acentuado desbalanceamento da quantidade de documentos entre os tipos ativos.

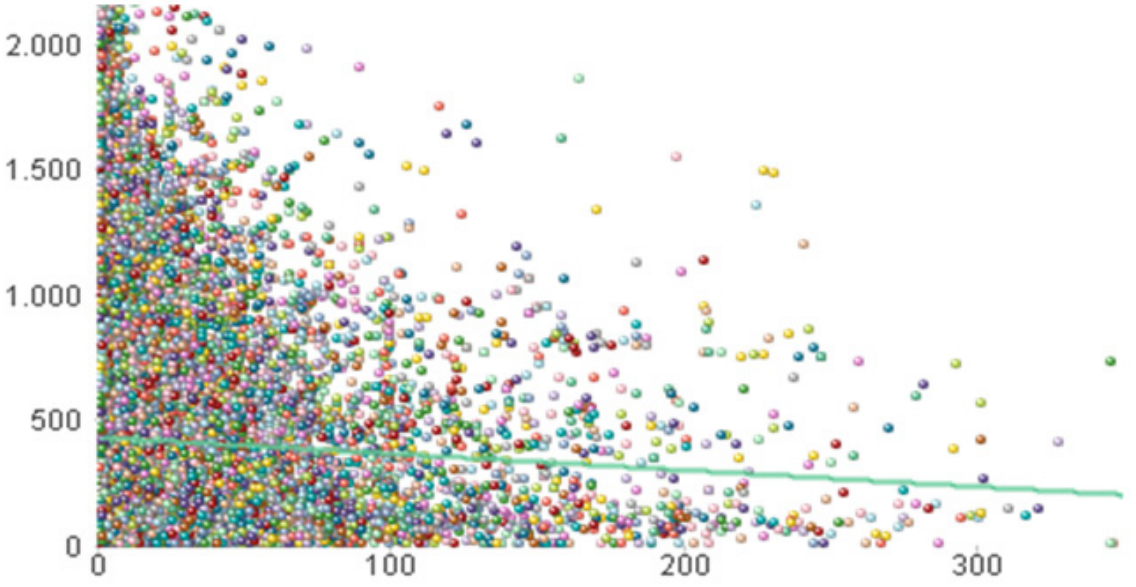
Gráfico 1: Visão parcial do desbalanceamento de documentos por tipo (referência: 17/4/2019).

Fonte: Castro (2019) - visão parcial.

BRANTING (2017) afirma que o formato PDF tem sido usado em tribunais e que o texto obtido desses documentos apresenta muitos erros e não preservam a sequência original do documento, devido ao processo usado de OCR.

O alcance dos resultados esperados pelo sistema e-TCE depende da qualidade dos documentos protocolados no sistema. A eficácia do classificador também tende a ser superior se os dados tiverem uma maior qualidade. Foi encontrada uma baixa qualidade no OCR dos documentos. Um exemplo dessa situação é o documento de protocolo 58.900.414 que tem 162 páginas, mas não se consegue via OCR identificar nem uma centena de palavras válidas, ou seja, menos de uma palavra válida por página. A baixa qualidade do OCR pode ser ilustrada no Gráfico 2 que apresenta o número de páginas e o número de palavras válidas por página para alguns documentos.

Gráfico 2: Documentos – quantidade de palavras válidas por página (eixo y) x número de páginas (eixo x).



Fonte: os autores (2020).

Foram experimentadas formas diferentes de pré-processamento de texto dos conteúdos dos documentos. Conforme Castro (2019), os métodos experimentados retrataram diferentes combinações de métodos de obtenção de texto do documento (algoritmos de OCR), de critérios para definição de palavras válidas e de classes substitutas de palavras consideradas (cpf, data, nome pessoa física, etc).

2.6.2 DESCRIÇÃO FUNCIONAL DO CLASSIFICADOR

O Cladop, em sua versão 2.1, alcança a acurácia de 91,1% com desvio padrão de 0,3%, desempenho apurado com validação cruzada de 7 partições e 2 repetições, totalizando 14 amostras. No contexto deste trabalho, o termo acurácia deve ser entendido como acurácia micro, que leva em consideração os resultados, acertos e erros, por documento independentemente do seu tipo.

O classificador foi implementado em python na forma de um webservice que recebe como parâmetro um arquivo PDF, e a partir do seu nome e do seu conteúdo retorna nove tipos mais prováveis, com suas respectivas probabilidades. A acurácia do classificador3 sobe de 91% (do primeiro tipo) para 99% quando consideradas as nove primeiras previsões. O retorno de mais de um tipo possibilita que o sistema, por exemplo, possa apresentar os demais tipos em uma segunda tela caso o primeiro tipo não seja acatado pelo usuário.

3 Após a escrita deste artigo, houve inclusão e exclusão de tipos de documentos e foi construída, considerando documentos mais recentes, a versão 3.0 do Cladop que alcançou acurácia de 93,6% com desvio padrão de 0,3%.

A Tabela 1 apresenta, para alguns tipos de documento, algumas métricas apuradas sobre os dados de validação, 5% do total, quando da geração da versão em produção do classificador. Os números gerais em dados de validação são: acurácia: 90.99%; precisão macro: 81,04% e weighted: 91,06%; recall macro: 80,29% e weighted: 90,99%; F1 macro: 80,16% e weighted: 90,87%.

Tabela 1: Métricas apuradas por tipo sobre os dados de validação (visão parcial).

Descrição	Precisão	Recall	F1	Documentos
Ação judicial - petição inicial	93,33%	93,33%	93,33%	15
Acórdão	85,71%	85,71%	85,71%	7
Análise de Prestação de Contas	77,27%	62,96%	69,39%	27
Análise defesa	83,33%	52,63%	64,52%	19
Suspensão de inadimplência	73,68%	82,35%	77,78%	17
Termo de concessão e de aceitação da bolsa e aditivos	75,00%	54,55%	63,16%	11
Termo de recebimento definitivo da obra	100,00%	100,00%	100,00%	4

Fonte: Castro (2019) – visão parcial.

Adicionalmente às predições, o classificador retorna informações derivadas do pré-processamento do texto do arquivo, que podem ser úteis para o sistema e-TCE exigir no cadastro dos documentos uma qualidade mínima de conteúdo textual, como: quantidade de palavras válidas, quantidade de valores, quantidade de nomes, etc. Assim, o sistema pode impedir que documentos críticos para o processo de TCE tenham qualidade baixa de OCR com um alto percentual de palavras inválidas ou mesmo que tenham conteúdo incompleto, como o caso de um documento de AR, Aviso de Recebimento, sem informação de CPF ou CNPJ e data.

2.6.3 APLICAÇÃO DO RASTRO-DM NO PROJETO

A utilização do Rastro-DM se mostrou uma solução efetiva para a dificuldade apresentada no referencial teórico de se refletir na documentação a iteratividade e a interatividade de um processo de DM.

As seções seguintes exemplificam a execução das ações da metodologia no projeto e os benefícios auferidos.

Detalhes do rastro do projeto Cladop (Rastro-DM@Cladop) e o código usado para sua construção estão acessíveis em https://gitlab.com/MarcusBorela/rastro-dm.git.

2.6.4 DEFINIÇÃO DE AÇÃO

As primeiras definições de ação foram registradas como comentários no próprio código de implementação dos treinamentos. Porém, com o tempo, elas passaram a ser mais complexas e amplas e perpassavam vários códigos. Então passaram a ser persistidas em banco de dados. Definitivamente não se pode confiar o registro de definições de ação a código. Devido à natureza imprevisível de um projeto DM o código muitas vezes evolui rapidamente e acaba comprometendo, entre outras coisas, sua documentação (GREFF *et al*, 2017). A Tabela 2 ilustra algumas definições de ação registradas durante o projeto.

Tabela 2: Exemplos de definições de ação registradas no Cladop.

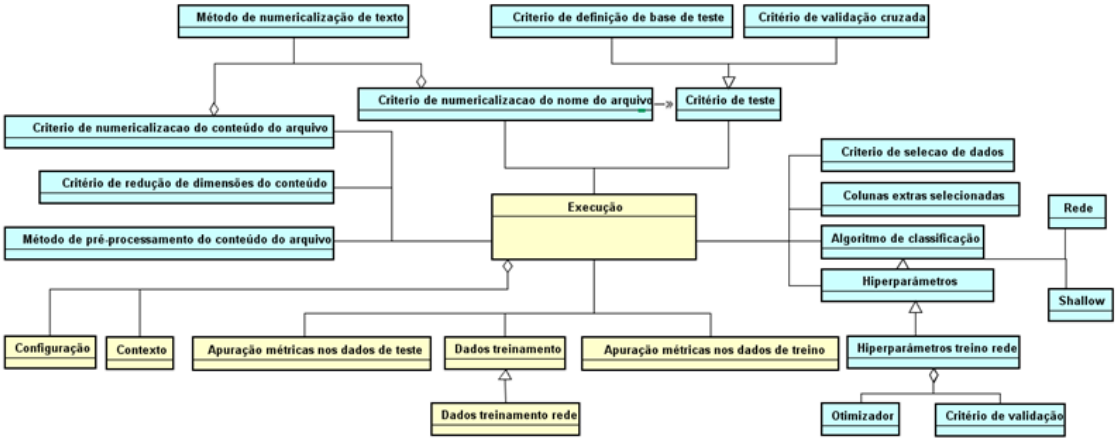
Momento	Descrição	Tarefa CRISP-DM
12/3/2019-17:04	Criando estrutura (código e dados) para tratar k-fold em shallow algorithms	Projeto de testes
14/3/2019-19:27	Iniciadas execuções para experimentar otimizadores (MLP): nadam, adadelta	Construir Modelo
18/3/2019-11:40	Experimentando colunas não mais binárias com MLP	Formatar Dados
26/3/2019-11:57	Iniciando inclusão de nomes de arquivos no modelo	Selecionar Dados
29/5/2019-19:30	Alterado programa (shallow) para gravar também recall e f1 micro	Projeto de testes
4/6/2019-19:35	Alterado programa (shallow) para não gravar recall e f1 micro, pois são equivalentes à acurácia (e à precisão)	Projeto de testes

Fonte: Castro (2019) – visão parcial.

2.6.5 REGISTRO DE TREINAMENTO

O conjunto de dados persistidos em banco de dados sobre treinamentos evoluiu no decorrer do projeto acompanhando o amadurecimento da equipe em DM. A Figura 4 apresenta os principais conceitos persistidos que podem ser relacionados aos subtipos da Figura 3. Em azul claro, constam os conceitos que são entrada para o treinamento e em amarelo os dados derivados do processo.

Figura 4 - Informações persistidas de treinamento no projeto Cladop agrupadas por conceito



Fonte: Castro (2019).

A Tabela 3, de caráter meramente ilustrativo, apresenta alguns atributos dessas entidades e seus respectivos valores para os treinamentos de teste e de geração do modelo construído na versão 2.1 do Cladop.

Tabela 3 - Rastro dos treinamentos de teste e de geração da versão 2.1 do Cladop.

Grupo	Item	Teste	Geração
Contexto	Código (chave única)	13.134	13.138
	Data e hora de registro	05/07/2019-00:44:59	05/07/2019-18:02:12
	Número de épocas treinadas	33	24
	Quantidade de documentos usados	63.468	
	Quantidade de tipos tratados	81	
	Tempo de execução (em segundos)	460	475
Configuração	Programa executado	Cladop_monitoramento.ipynb	
Parâmetros	Modo de conversão de texto para números	Tfidf	
	Número de palavras do modelo "*bag of words*" do texto	24.576	
	Número de palavras do modelo "*bag of words*" do nome do arquivo	1.000	
	Método de pré-processamento do texto	7	
	Critério de seleção dos dados	Documentos de tipo diferente de Outros e criados após 1/5/2018	
	Método de teste	Validação cruzada com 7 partições e 2 repetições	Não foram separados dados para teste
	Otimizador usado para a rede	Adam	
	Tamanho do batch de treinamento	256	
	Quantidade mínima exigida por tipo	0	
	Dimensão final do vetor de conteúdo após redução	768	
	Algoritmo de redução de tamanho	TruncatedSVD	
	Dados foram embaralhados a cada época?	Não	
	Critério para dados de validação	5% dos dados, sem estratificação e com shuffle	
Resultado	Acurácia apurada em teste	91,1% +- 0,3%	
	Acurácia apurada em treinamento	96,4% +- 0,7%	95,6%
	Acurácia apurada em validação	90,8% +- 0,6%	92,1%

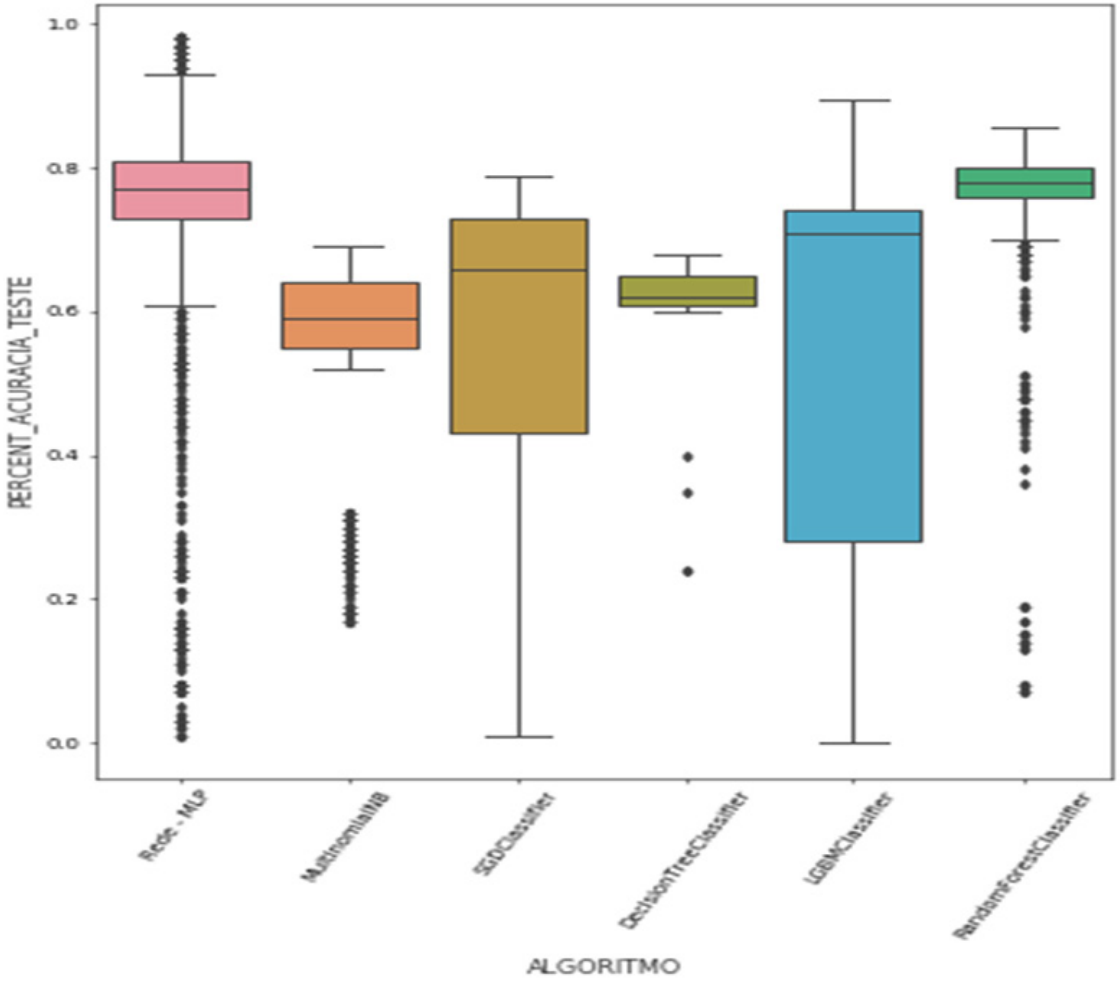
Fonte: Castro (2019) – versão adaptada.

O treinamento que gerou a versão 2.1 do modelo tinha o código 13.138 (conforme segunda linha da Tabela 3), o que demonstra que foram executados mais de treze mil treinamentos de modelos. Alguns treinamentos envolvendo apenas documentos com alta qualidade de OCR e tipos com grande quantidade de exemplares levaram a classificadores com acurácia superior

a 98% em base de teste. É o caso do treino de código 147, em 24/11/2018, que envolveu 11.110 documentos de melhor qualidade de OCR dos quatro tipos de documentos mais frequentes. Contudo, no decorrer do projeto, optou-se pela construção de um classificador que abrangesse todos os tipos de documentos e com critérios menos restritivos.

A partir dos dados de treinamento podem ser feitas diversas análises. Um exemplo é a comparação do desempenho dos modelos quanto aos algoritmos utilizados. Os modelos com maior acurácia para a seleção final de dados e de parâmetros foram implementados com redes neurais (rede MLP). Modelos implementados usando o algoritmo *LGBMClassifier* ficaram logo atrás, com uma diferença de cerca de 2%. O Gráfico 3 ilustra a acurácia em base de teste alcançada por alguns dos algoritmos experimentados. A grande variação de acurácia por algoritmo se explica pela variação de combinações de parâmetros experimentadas nos treinamentos. O entendimento desses algoritmos não está no escopo deste artigo. Para interessados, sugere-se a leitura do livro *Data Mining and Analysis: Fundamental Concepts and Algorithms* de ZAKI e MEIRA (2014).

Gráfico 3 - *Boxplot* com acurácia em base de teste dos algoritmos experimentados.



Fonte: Castro (2019) – visão parcial.

2.6.6 SÍNTESE DE APRENDIZADO

Dos aprendizados registrados, a maioria poderia ser sintetizada por mineração de dados nos registros de treinamentos. Por exemplo, a partir de critérios de igualdade de valores em algumas colunas, poderia ser detectado o aprendizado do dia 10/4/2019 da Tabela 4 que ilustra alguns aprendizados. Outros carecem da intervenção humana, como o do dia 27/2/2019. A geração de conhecimentos de forma automática se mostra um caminho promissor para a elevação do conhecimento organizacional, o que foi referendado por NGUYEN (2018), que, como visto, ressalta haver um grande vazio de pesquisas nessa área.

Tabela 4 – Exemplos de aprendizados registrados durante o projeto

Data e hora de registro	Descrição	Atividade CRISP-DM de contexto
27/2/2019-10:32	Bastam as variáveis de qualidade do PDF e de contexto para se alcançar um resultado de 44% de acurácia, com otimizador Adam. Com Adagrad:42%; SGD:28%.	Selecionar Dados
27/2/2019-10:39	Ao usar pca (*sklearn*): melhor separar o comando *fit* do comando *transform*. O comando *fit_transform* estava travando!	Formatar Dados
27/2/2019-10:53	Ao usar pca (sklearn): se o número de dimensões for bem pequeno e o *array* for grande (ver detalhes na documentação da função), melhor usar o método randomized do que o *full*. Pois é mais rápido, e os resultados (variância alcançada) são equivalentes.	Formatar Dados
29/03/2019-09:54	Percebida uma melhora de cerca de 5% na acurácia dos modelos após inclusão do nome do arquivo como característica.	Selecionar Dados
1/4/2019-15:36	Para algoritmos *shallow*, as colunas extras com valores não dummies (uma só coluna com vários valores discretos) levaram a um resultado melhor. Para Rede Neural, há uma pequena melhora usando valores *dummies*.	Formatar Dados
10/4/2019-10:49	Em classificação de multiclasses, as métricas *f1_micro*, *recall_micro*, *precision_micro* e acurácia são equivalentes	Projeto de testes
22/5/2019-20:04	O otimizador Adam (passando objeto com Amsgrad=*False*) foi o melhor otimizador até o momento. Acima do Amsgrad=*True* em 0.1%, do Rmsprop e do Adadelta em 0.2% no contexto avaliado das últimas execuções.	Construir Modelo

Fonte: Castro (2019) – visão parcial.

Passados menos de seis meses do final do projeto, Castro (2019) experimentou contabilizar, de forma amostral, quantos deles ainda se encontravam completos em sua memória e chegou a uma estimativa de apenas 27%. Os demais, infelizmente, ou não estavam completos ou haviam sido perdidos da memória do autor, mas, graças ao rastro, não foram perdidos da memória do projeto.

Vimos que o esquecimento pode levar à execução repetida de experimentos (BECKER e GHEDINI, 2005). Durante o projeto houve um esquecimento que levou a esforços

desnecessários em execução de tarefas. O aprendizado de 10/4/2019 foi ignorado com a execução desnecessária da ação para se gravar também as métricas no dia 29/5/2019 (ver Tabela 2), o que levou a um retrabalho com a alteração do programa para não gravar essas métricas em 6/4/2019. Ficou a lição aprendida de que não basta ter o rastro, é necessário fazer uso do mesmo.

2.6.7 BENEFÍCIOS ALCANÇADOS

Os benefícios auferidos com a aplicação da metodologia no projeto podem ser agrupados em 3 aspectos: gestão de projetos, capacitação técnica da equipe e automação de atividades.

Na gestão de projetos, perceberam-se ganhos na comunicação, no planejamento e na diminuição de custos.

O rastro construído viabilizou um acompanhamento gerencial contínuo da evolução do projeto e do desempenho dos modelos gerados. O uso de uma padronização na apresentação de relatórios de acompanhamento inspirou confiança nos usuários e nos patrocinadores.

O rastro se tornou subsídio para um melhor planejamento para projetos futuros semelhantes, afinal, projetos de mineração de dados podem ser beneficiados com as experiências de projetos anteriores (CHAPMAN *et al*, 2000). O relato da experiência contida no rastro torna mais fácil defender recursos mais realistas, pois há uma compreensão mais profunda de como o esforço é realmente gasto. Permite, também, a identificação de pontos críticos no processo seguido.

Constatou-se uma diminuição no esforço em trabalhos repetidos ou mesmo em caminhos não frutíferos.

Os bons impactos do uso do Rastro-DM na gestão levaram à sua inclusão como requisito técnico para a terceirização de desenvolvimento de projetos de DM que se inicia no Tribunal.

A capacitação técnica da equipe foi incrementada dado o ambiente criado, o amadurecimento alcançado e o compartilhamento dos aprendizados.

Promoveu-se um ambiente estimulante e prático (aprender-fazendo). A equipe se sentiu incentivada à documentação concomitante com a execução das atividades dada a facilidade de registro e os benefícios percebidos no decorrer do projeto, como a própria documentação gerada.

Não só a síntese de aprendizados e o uso efetivo deles no projeto como também a evolução estrutural dos registros de treinamento traduziram o amadurecimento da equipe em DM. À medida que se alcançava uma melhor compreensão e entendimento do

processo e das técnicas envolvidas, novos campos foram criados para armazenamento e os conteúdos ficaram mais ricos.

Quanto ao compartilhamento, o rastro do Cladop foi disponibilizado no contexto de um trabalho de conclusão de curso de Especialização em Análise de Dados (Castro, 2019) e a metodologia foi apresentada no 5º Seminário Internacional de Análise de Dados da Administração Pública em 2019.

Espera-se avançar na discussão interna à organização sobre políticas e padrões para se transformar os rastros em recursos corporativos, como apoio à disseminação de conhecimento em vários níveis. Consultas simples, por exemplo, poderiam encontrar projetos que aplicaram determinada técnica e identificar dicas de seu uso e avaliações dos resultados alcançados. Consultas gerenciais poderiam subsidiar estratégias na área de DM para a organização.

Outro benefício foi a automação de rotinas, que não só garante uma qualidade mínima para atividades como reduz o custo de suas realizações. A lista cronológica de aprendizados e de definições de ação enriquecida com alguns gráficos gerados automaticamente a partir dos treinamentos podem compor relatórios automáticos sobre tarefas da metodologia base, como ilustrado em Castro (2019). O mesmo autor demonstra que o rastro viabilizou o monitoramento automático do desempenho dos modelos face a novos documentos cadastrados no sistema. Esse rastreamento promoveu, inclusive, a geração automática da versão 2.1 do Classificador a partir de dados mais atualizados.

3. CONCLUSÃO

Os objetivos propostos para o trabalho foram alcançados. Uma breve contextualização com referencial teórico foi apresentada sobre metodologias e documentação em projetos de mineração de dados, bem como do potencial impacto das experiências adquiridas nos projetos para uma organização.

Foi proposto o Rastro-DM como uma metodologia de documentação de projetos de DM com foco no processo que, entre outras características, é flexível e pode ser mesclado à metodologia em uso por uma organização.

O Rastro-DM mostrou-se viável e com resultados benéficos conforme ilustrado na sua aplicação no projeto Cladop.

Há muito ainda por fazer. Trabalhos futuros podem experimentar o uso Rastro-DM em outros contextos: outros objetivos de mineração que não classificação, outras plataformas de desenvolvimento ou mesmo em outras culturas organizacionais. A metodologia pode ser revisada com o acréscimo de novas atividades ou mesmo enriquecida com técnicas e procedimentos que complementem suas atividades. Estudos podem ser feitos para o levantamento de políticas e padrões para que um rastro possa ser usado como subsídio em contratações de DM. Também é um desafio para trabalhos futuros a mineração de aprendizados a partir de rastros de treinamentos, pois preocupa o fato de a integração

entre DM e KMP possuir um grande vazio de pesquisas (NGUYEN, 2018). E, talvez o mais importante ponto a evoluir, que não foi escopo deste trabalho, seja como tornar o rastro um recurso corporativo, que permita a partilha das experiências adquiridas e o uso de táticas institucionais para o crescimento do conhecimento organizacional. Afinal, o valor dos dados está em como eles são interpretados e usados (BERMAN *et al*, 2018).

Por fim, ficou claro que, além de contribuir para a equipe do projeto, o rastro dos projetos tem potencial para promover um salto no conhecimento organizacional, se forem adotadas medidas institucionais para incentivar sua construção, sua partilha e a busca automática de aprendizados neles contidos. Mas há que se ter em mente que nenhum direcionamento corporativo deve inibir a liberdade dos analistas de dados, pois diminuiria sua criatividade. E a liberdade criativa, de evoluir o próprio rastro, é algo de que os seres humanos não podem prescindir, pois é um dos grandes diferenciais que os impedem de serem classificados como máquinas.

4. REFERÊNCIAS

BECKER, Karin; GHEDINI, Cinara. **A documentation infrastructure for the management of data mining projects**. Information and Software Technology, v. 47, n. 2, p. 95-111, 2005.

BERMAN, Francine et al . **Realizing the potential of data science**. Communications of the ACM, v. 61, n. 4, p. 67-72, 2018.

BHATT, Ganesh D. Knowledge management in organizations: examining the interaction between technologies, techniques, and people. Journal of knowledge management, v. 5, n. 1, p. 68-75, 2001.

CASTRO, Marcus Vinícius Borela. **Mineração de Dados com Rastro: Boas Práticas para Documentação de Processos e sua Aplicação em um Projeto de Classificação Textual**. Trabalho de Conclusão de Curso (Especialização em Análise de Dados para o Controle) – Escola Superior do Tribunal de Contas da União, Instituto Serzedello Corrêa, Brasília DF. Disponível em: https://portal.tcu.gov.br/biblioteca-digital/mineracao-de-dados-com-rastro-boas-praticas-para-documentacao-de-processo-e-sua-aplicacao-em-um-projeto-de-classificacao-textual.htm. Acesso em 26 jul. 2020. 2019.

CHAPMAN, Pete et al . **CRISP-DM 1.0:** Step-by-step data mining guide. SPSS inc, v. 16, 2000.

CHOLLET, Francois. **Deep Learning with Python**. Manning Publications Co., Greenwich, CT, USA. 2017.

CONKLIN, Jeffret. **Capturing Organisational Memory**. In: Groupware and Computer-Supported Cooperative Work, R.M. Barcker (Ed.), Morgan Kaufman, pp. 561-565. 1996.

DINGSØYR, Torgeir; Moe, Nils Brede; Øystein. Nytrø. **Augmenting experience reports with lightweight postmortem reviews**. Lecture Notes in Computer Science, 2188:167–181, 2001.

GHEDINI, Cinara; BECKER, Karin. **KDD application management through documentation**. Disponível em: https://www.researchgate.net/profile/Karin_Becker2/publication/268253354_ KDD_application_management_through_documentation/links/5657a5ec08ae1ef9297bf1d1/KDD-application-management-through-documentation.pdf. Acesso em 27 jul. 2019. 2000.

_____. **A documentation model for KDD application management support**. In: SCCC 2001. 21st International Conference of the Chilean Computer Science Society. IEEE. p. 105-114. 2001.

GREFF, Klaus et al . **The sacred infrastructure for computational research.** In: Proceedings of the Python in Science Conferences-SciPy Conferences. 2017.

KURGAN, Lukasz A.; MUSILEK, Petr. **A survey of Knowledge Discovery and Data Mining process models**. The Knowledge Engineering Review, v. 21, n. 1, p. 1-24, 2006.

MARBÁN, Óscar et al . **An engineering approach to data mining projects**. In: International Conference on Intelligent Data Engineering and Automated Learning. Springer, Berlin, Heidelberg, p. 578-588. 2007.

MARISCAL, Gonzalo; MARBAN, Oscar; FERNANDEZ, Covadonga. **A survey of data mining and knowledge discovery process models and methodologies**. The Knowledge Engineering Review, v. 25, n. 2, p. 137-166, 2010.

MINGERS, John; BROCKLESBY, John. **Multimethodology: Towards a framework for mixing methodologies**. Omega, v. 25, n. 5, p. 489-509, 1997.

NGUYEN, Ngoc Buu Cat. Data Mining in Knowledge Management Processes: Developing an Implementing Framework, 2018.

PRAKASH, BV Ajay; ASHOKA, D. V.; ARADHYA, VN Manjunath. **Application of data mining techniques for software reuse process**. Procedia Technology, v. 4, p. 384-389, 2012.

PUBLIO, Gustavo Correa et al . **ML-Schema: Exposing the Semantics of Machine Learning with Schemas and Ontologies**. arXiv preprint arXiv:1807.05351, 2018.

STATA, Ray. **Organizational learning: The key to management innovation**. Massachusetts Institute of Technology, 1980.

W3C (World Wide Web Consortium) **Machine Learning Schema Community Group**. W3c machine learning schema. Disponível em: https://www.w3.org/community/ml-schema. Acesso em 30 jul. 2019. 2017.

WIRTH, Rüdiger; HIPP, Jochen. **CRISP-DM: Towards a standard process model for data mining**. In: Proceedings of the 4th international conference on the practical applications of knowledge discovery and data mining. Citeseer, p. 29-39. 2000.

ZAKI, Mohammed J.; MEIRA, Wagner. **Data mining and analysis: fundamental concepts and algorithms**. Cambridge University Press, 2014.