PUC Minas

pista: periódico interdisciplinar

[sociedade tecnologia ambiente]

UTILIZAÇÃO DE TECNOLOGIAS PARA DESCOBERTA DE CONHECIMENTO NA BASE DE DADOS PRISIONAL DO ESTADO DE MINAS GERAIS

USE OF TECHNOLOGIES FOR KNOWLEDGE DISCOVERY IN THE PRISIONAL DATABASE OF THE STATE OF MINAS GERAIS

Submetido em: 21/01/2021
Aprovado em: 19/022021

Edson Luiz Ferreira dos Santos[1]

Magali Rezende Gouvêa Meireles[2]

RESUMO

O Brasil se tornou a terceira nação do mundo que mais possui pessoas encarceradas, ficando atrás apenas dos Estados Unidos e da China. Só em Minas Gerais, mais de 800 mil pessoas já passaram pelo Sistema Prisional. Conforme informações do Banco Nacional de Monitoramento de Prisões do Conselho Nacional de Justiça (2018), a população carcerária brasileira era de 602 mil presos, sendo Minas o terceiro estado do Brasil que mais possuía presos. Do total de encarcerados brasileiros, 24% estão presos provisoriamente e 40% sem condenação, totalizando mais de 64% de pessoas sem execução da pena definitiva. Isto faz com que o sistema prisional se torne uma escola do crime. Vários presos com perfis totalmente diferentes convivem juntos.O objetivo deste trabalho é utilizar tecnologia de descoberta de conhecimento na base de dados dos presos que receberam o livramento condicional ou cumpriram a pena, para descobrir o perfil do preso reincidente em Minas Gerais. Percebeu-se pelas regras extraídas do algoritmo árvore de decisão que, resumidamente, os presos do sexo masculino e que saem da unidade prisional ainda jovem são os que reincidem. Subentende-se que o preso que cumpre sua pena ou recebe o livramento condicional ainda jovem tem começado mais cedo sua vida criminal.

Palavras-chave: Reincidência Criminal. Sistema Penitênciário. Segurança Pública. Mineração de dados. Algoritmos de Classificação.

---

[1]Graduado em Sistemas de Informação pela PUC Minas, unidade São Gabriel. E-mail: elfsantos@sga.pucminas.br

[2]Doutora em Ciência da Informação pela Universidade Federal de Minas Gerais (UFMG), mestre em Tecnologia pelo CEFET-MG e graduação em Engenharia Elétrica pela UFMG. Professora Adjunta do Instituto de Ciências

Exatas e Informática da PUC Minas. Professora do Programa de Pós-Graduação em Modelagem Matemática e Computacional do CEFET-MG. E-mail: magali@pucminas.br

ABSTRACT

Brazil has become the third nation in the world with the greatest number of incarcerated people, being third to the United States and China. In Minas Gerais alone, more than 800,000 people have already gone through the Prison System. According to information from the National Prison Monitoring Bank of the National Council of Justice (2018), the Brazilian penitentiary population was a total of 602,000 inmates, with Minas Gerais being the third state in Brazil with the most inmates. Out of all the Brazilian prisoners, 24% are provisionally arrested and 40% are without conviction, totaling more than 64% of people without the execution of the final sentence. This makes the penitentiary system a crime school. Several prisoners with completely different profiles live together. From this information, we tried to use a classification algorithm in the database of prisoners who received conditional release or have served their sentence, to find out the profile of a repeating offender in Minas Gerais. By the rules extracted from the decision tree, it was noticed that, in short, male prisoners who leave the prison unit at a young age are those who commit repeated felonies. A possible conclusion is that the prisoner serving his sentence or receiving the parole at an early age has begun his criminal life earlier.

Keywords: *Criminal recidivism. Penitentiary system. Public security. Data mining. Classification Algorithms.*

1 INTRODUÇÃO

O Brasil se tornou a terceira nação do mundo que mais possui pessoas encarceradas, ficando atrás, apenas, dos Estados Unidos e da China, conforme afirma a revista Isto é (2017). Só em Minas Gerais, mais de 800 mil pessoas já passaram pelo sistema prisional desde 1994. Conforme informações do Banco Nacional de Monitoramento de Prisões (BNMP2.0) do Conselho Nacional de Justiça (CNJ), em oito de agosto de 2018, a população carcerária brasileira era de 602 mil presos, sendo Minas o terceiro estado do Brasil que mais possuía presos. Em Minas Gerais, na época, eram quase 60 mil presos e, atualmente, são mais de 70 mil presos. Algumas políticas públicas têm sido adotadas para tentar diminuir esse crescimento. A assistência à família do preso realizada pela Diretoria de Assistência à Família (DAF) tem como objetivo fornecer todo o suporte relacionado ao cadastramento dos familiares e ao credenciamento para visitação nas unidades prisionais. É realizada também a assistência religiosa, que tem o objetivo de incentivar a reintegração social e proporcionar a mudança de comportamento dos indivíduos privados de liberdade, por meio de atividades que valorizam aspectos éticos, morais e religiosos. Existe também o atendimento jurídico, o incentivo ao ensino e à profissionalização e a remissão da pena por meio do trabalho do preso em empresas parceiras do Sistema Prisional.

A Secretaria de Estado e Administração Prisional (SEAP) de Minas Gerais tem como missão "promover sua gestão eficiente, criando condições ideais de segurança nas unidades prisionais e atuando na ressocialização dos indivíduos privados de liberdade" (SEAP, 2019).

Administrar o crescimento de presos nas unidades prisionais sem ter em contrapartida um aumento das vagas faz com que se tenha mais unidades com superlotação, aumentando o risco de motins e rebeliões. A Lei de Execução Penal prevê o cumprimento da pena em celas individuais com, no mínimo, 6m2 de área. Atualmente, existem centros de remanejamentos do sistema prisional (CERESP) com celas com mais de 20 presos. A capacidade original era de 8 presos e o excesso dificulta a missão de atuar na ressocialização do indivíduo privado de liberdade.

Segundo o BNMP2.0 (2018), do total de encarcerados brasileiros, 24% estão presos provisoriamente e 40% sem condenação, totalizando mais de 64% de pessoas sem execução da pena definitiva. Vários presos com perfis totalmente diferentes convivem juntos. Isto faz com que tais pessoas sejam "recrutadas" para o crime dentro das unidades prisionais. O tráfico de drogas e o roubo, segundo o BNMP, são os dois crimes mais cometidos e somam mais 51% dos casos. O uso de tecnologias para descoberta do conhecimento é uma alternativa para subsidiar tomadas de decisão bem como para identificar o perfil dos presos de Minas. Descobrir em qual perfil o preso se encaixa, antes mesmo de admiti-lo em uma unidade prisional, pode auxiliar a alocação do mesmo numa cela ou em unidades que concentrem presos de um mesmo perfil. Isso já acontece com o presídio Inspetor José Martinho Drummond, que só acolhe presos que cometeram crimes sexuais e com o Centro de Referência à Gestante Privada de Liberdade, que recebe apenas presas gestantes e com filhos recém nascidos. Outro exemplo é o Complexo Penitênciário Estevão Pinto, que acolhe apenas presas do sexo feminino. Isso facilitaria a ressocialização dos presos com perfis menos agressivos, viabilizando a análise dos perfis para um futuro contrato de trabalho, realizado por meio de parcerias com a SEAP, para reduzir sua pena e auxiliar na reinserção no mercado de trabalho e na sociedade.

O objetivo deste trabalho é identificar o perfil dos presos que receberam livramento condicional ou terminaram sua pena, mas reincidiram em uma unidade prisional. A expectativa é de que estas informações possam ser utilizadas na adoção de políticas públicas específicas para o perfil identificado, contribuindo, assim, com a diminuição da reincidência criminal. Este trabalho está dividido em mais quatro seções. A Seção 2 apresenta o referencial teórico utilizado na pesquisa, bem como alguns exemplos de aplicações em outros trabalhos.

A Seção 3 descreve a metodologia do trabalho. As Seções 4 e 5 apresentam os resultados, as análises e as considerações finais.

2 REFERENCIAL TEÓRICO

As cinco próximas subseções discutem conceitos relacionados à reincidência social, à seleção de dados, ao pré-processamento da base de dados, aos algoritmos para descoberta do conhecimento e apresentam alguns trabalhos relacionados.

2.1 Conceitos Sociais

Ressocializar o preso é fazer com que o mesmo consiga se inserir novamente na sociedade. Cabral e Silva (2010) afirmam que ressocializar é “propiciar a sua (preso) reintegração social". Para Machado e Sloniak (2015), a ressocialização tem o mesmo significado que reintegração social e pode ser considerado como reabilitação, ressocialização e reintegração. Para Barcinski e outros (2017), o tratamento ressocializador busca recuperar o indivíduo que cometeu alguma infração. A lei de Execução Penal Brasileira, no artigo 10, fala sobre o dever do Estado de orientar o retorno à convivência em sociedade. Logo, o conceito de ressocializar surge do próprio conceito de socializar, que é a capacidade de se tornar social, se reunir em sociedade, assimilar os hábitos de seu grupo social.

O conceito de reincidência vindo da Lei de Execução Penal (LEP) significa que o indivíduo reincidente é aquele que possui condenação penal no período de 5 anos. Ou seja, a pessoa que ainda não teve seu processo judicial transitado julgado não pode ser considerada reincidente pela prática de um novo crime. Para Sapori e outros (2017), reincidência prisional “ocorre quando há um segundo ingresso na prisão do mesmo indivíduo por nova prática criminal”. Considera-se reincidente prisional/penitenciário o indivíduo que independente de ter condenação no âmbito jurídico já foi preso, no mínimo, 2 vezes e passou pelo sistema prisional na cadeia de custódia do estado.

2.2 Seleção e Redução de Dados

Seleção de dados, também chamada de redução de dados, é o processo de selecionar um sub-conjunto de atributos relevantes para a base de dados que será utilizada nos experimentos. O obje- tivo é remover os atributos com baixa variância, alta correlação, redundantes e informações irrelevantes para o processo de descoberta de conhecimento. É “a identificação

de quais infor- mações, dentre as bases de dados existentes, devem ser efetivamente consideradas" (PASSOS; GOLDSCHMIDT, 2005). Para Castro e Ferrari (2016), "a redução dos dados tem como um dos objetivos a construção de bases de dados menores para o aumento do desempenho computaci- onal, ao mesmo tempo em que tenta manter características originais dos dados".

Pode-se reduzir os dados de forma horizontal e vertical. A redução horizontal elimina instâncias (linhas) da base de dados e, geralmente, é feita por intermédio de amostragem aleatória, ou eliminação direta de instâncias. Pode ser utilizada quando uma linha possuir vários valores nulos ou ausentes. A redução vertical está relacionada à eliminação de uma coluna inteira da base de dados. Por exemplo, pode-se eliminar toda a coluna de nome do preso ao selecionar os dados para serem usados por um algoritmo de classificação.

2.3 Pré-processamento da Base de Dados

Antes de se utilizar algoritmos de Aprendizado de Máquina (AM), faz-se necessário o tratamento da base de dados ou dos dados brutos. Existem algoritmos que aceitam como entrada apenas dados numéricos, bem como registros não vazios ou não nulos. Para Faceli e outros (2011), existe uma "dificuldade dos algoritmos de AM para lidar com os dados no seu formato original" e o pré-processamento refina os dados para melhorar o desempenho desses algoritmos. Castro e Ferrari (2016) afirmam que "conhecer e preparar de forma adequada os dados para análise pode tornar todo o processo de mineração muito mais eficiente e eficaz. Por outro lado, dados mal processados podem inviabilizar uma análise ou invalidar um resultado".

Para Quintella e Soares Junior (2003), a atividade de pré-processamento tem como objetivo "gerar uma representação conveniente para os algoritmos de mineração". Passos e Goldschmidt (2005) ressaltam que o pré-processamento possui "fundamental relevância no processo de descoberta de conhecimento". As intercorrências mais comuns que, geralmente, acontecem numa base de dados são a incompletude, a inconsistência e os ruídos. Muitas dessas inconsistências ocorrem, por exemplo, por falta de validação na entrada dos dados ou mesmo um erro na cadeia de custódia do Estado.

2.3.1 Limpeza da Base de Dados

A limpeza da base aumenta as chances de se ter um bom desempenho dos algoritmos a serem utilizados. Para Passos e Goldschmidt (2005), a limpeza dos dados é “qualquer tratamento realizado sobre os dados selecionados de forma a assegurar a qualidade". Para limpar a base, é realizada uma análise sobre dados incompletos, dados inconsistentes e dados com ruídos.

Dados incompletos são instâncias da base de dados que possuem ausência ou nulidade em um ou vários de seus atributos. São presos sem nome, sem data de entrada ou saída no sistema penitenciário. Geralmente, podem aparecer por diversos fatores, como falha no sistema, falta de identificação do indivíduo, erro/falha do agente de segurança ao preencher os dados, ou, até mesmo, por serem dados ausentes. O tratamento da incompletude proposto por Faceli e outros (2011) envolve a imputação de um valor que demonstra a inconsistência, a utilização de média ou moda em campos numéricos e a utilização de um algoritmo para predição da instância. Já Castro e Ferrari (2016) sugerem ignorar o objeto, usar dados da mesma classe para imputar o valor e imputar o valor manualmente.

Dados inconsistentes são aqueles que não possuem uma lógica racional, que estejam fora do domínio de valores aceitos, por exemplo, a idade de 400 anos. Faceli e outros (2011) afirmam que “dados inconsistentes são aqueles que possuem valores conflitantes em seus atributos". Já Castro e Ferrari (2016) relatam que “a consistência de um dado está relacionada à sua discrepância em relação a outros dados ou a um atributo, e tal consistência influencia na validade, na utilidade e na integridade [...]". Essa discrepância pode acontecer em relação ao atributo de classificação. A forma de se tratar dados inconsistentes pode ser manual, o que não é recomendado quando a base de dados é muito grande. O tratamento pode ser realizado por meio de gráficos, ou de algoritmos que verificam as instâncias. Contudo, a melhor forma é prevenir as inconsistências por meio da aplicação, tratando da inserção dos valores na base de dados.

Castro e Ferrari (2016) consideram que “um dado ruidoso é aquele que apresenta alguma variação em relação ao seu valor sem ruído”. Quintella e Junior (2003) acreditam que a avaliação estatística é essencial para a limpeza de ruídos. Todavia, é muito difícil saber se o dado de fato apresenta ou não ruído. Geralmente, esses dados são chamados de *outliers*, que são valores bastantes diferentes dos valores da base. Uma das formas de se tratar um dado

ruidoso é por intermédio de técnicas ou de algoritmos que suavizam esse dado ou, ainda, usando técnicas de agrupamento de dados.

2.3.2 Transformação de Dados

Para Passos e Goldschmidt (2005), a transformação dos dados (codificação) serve para apresentar os dados na forma exigida pela entrada dos algoritmos. Uma das operações básicas de transformação de dados é a padronização. Para Castro e Ferrari (2016), o objetivo principal da padronização é lidar com as unidades e escalas de dados. Essa padronização é realizada por meio da capitalização (trabalhar com maiúsculas e minúsculas), remoção de caracteres especiais, padronização de formatos como CPF e data, e conversão das unidades de medida.

Para um bom resultado com algoritmos de Redes Neurais Artificiais ou de agrupamento que usam apenas dados numéricos, é essencial que se faça uma boa conversão simbólica-numérica. Quando se tem algum atributo que indica ausência ou presença, pode-se transformá-los em 1 e 0, denotando-se verdadeiro ou falso para tal atributo. Quando, porém, existem dados nominais que assumem valores multivalorados, a transformação em dados numéricos é mais difícil. Se o número de possibilidades é baixo, a binarização do atributo pode ter um bom resultado. Quando são muitas as possibilidades, podem ser utilizados bits.

A transformação de dados numéricos para simbólicos é chamada de discretização. Tem como objetivo preparar os dados para os algoritmos de classificação, como a árvore de decisão. Faceli e outros (2011) afirmam que “quando um atributo é quantitativo e discretizado, o conjunto de possíveis valores é dividido em intervalos, e cada intervalo de valores quantitativos é convertido em um valor qualitativo". Dias e outros (2016) discretizaram sua base numérica e contínua em 5 intervalor diferentes, para que pudessem utilizar algoritmos de classificação. Basicamente, são usados, no processo de discretização, o histograma, agrupamento e a entropia.

2.3.3 Balanceamento e Enriquecimento dos Dados

Ao utilizar modelos de classificação, é necessário estar atento à quantidade de dados de cada classe. Se existir uma quantidade de registros de uma classe muito maior que a de outra, tem-se uma base de dados desbalanceada. É muito comum ter dados desbalanceados numa base de dados, que podem causar problemas no processo de aprendizado de máquina. Os

algoritmos de classificação tenderão a aprender com mais facilidade o que está relacionado à classe majoritária. Dias e outros (2016) tiveram problemas ao utilizar os algoritmos de classificação em uma base de amostragem sem balanceamento. Tan e outros (2009) afirmam que, quando um algoritmo de aprendizado de máquina quer classificar classes minoritárias, age como se quisesse encontrar uma agulha em meio ao palheiro.

Conforme citam Faceli e outros (2011), umas das alternativas para o tratar o desbalanceamento seria a redefinição do tamanho do conjunto de dados, por meio do *oversampling*, que é o aumento da classe minoritária por meio da replicação de instâncias já existentes na base de dados ou da criação de instâncias artificiais. Outra alternativa seria o *undersampling*, que é a diminuição da classe majoritária pela eliminação de instâncias. Para Faceli e outros (2011), ambas as abordagens possuem riscos, a do *oversampling* envolve o risco das novas instâncias representarem situações que nunca ocorrerão, e a do *undersampling* é de se perder instâncias importantes, prejudicando a geração de um modelo satisfatório. Gramani e Santoro (2006) citam métodos heurísticos, que geram um grande número de soluções factíveis e selecionam as melhores por meio de equações matemáticas. Outra alternativa é manter a base de dados e usar um algoritmo que dê pesos aos registros. Uma base que tem 100 registros de uma classe e 1.000 de outra passa a associar um peso maior para a classe de 100 registros, em detrimento de um peso menor para a classe de 1.000 registros.

O processo de enriquecimento dos dados consiste em adicionar mais informações à base de dados e complementar os dados com informações que a base não possui, por meio de outras bases externas ou pesquisas para complementação. Passos e Goldschmidt (2005) citam que "a função de enriquecimento é conseguir informação que possa ser agregada aos registros existentes".

2.4 Algoritmos utilizados na Mineração de Dados

Um dos algoritmos que utiliza aprendizado supervisionado, que é aquele que possue um atributo de classificação que o rotula. é a árvore de decisão. "A árvore de decisão é uma técnica que, a a partir de uma massa de dados, cria e organiza regras de classificação e decisão em formato de diagramas de árvores, que vão classificar suas observações ou predizer resultados futuros" (BARBIERI, 2011). Para Castro e Ferrari (2016), é "uma estrutura em forma de árvore, na qual cada nó interno corresponde a um teste de um atributo, cada ramo

representa um resultado do teste e os nós folhas representam classes ou distribuições de classes". O nó mais elevado da árvore é conhecido como nó raiz e cada caminho da raiz até um nó folha corresponde a uma regra de classificação. A árvore também é conhecida como modelo de caixa branca, pois com ela se adquire a regra de classificação, o padrão das classes, revelando o atributo mais importante e a regra que classifica mais instâncias.

Existem algoritmos de balanceamento que aumentam ou diminuem o conjunto de dados. É o caso do algoritmo *Resample*. O aumento se dá pela cópia dos registros da classe minoritária. A diminuição da base majoritária acontece com a eliminação da quantidade de registros, de modo que ambos fiquem com a mesma quantidade. Também existem algoritmos que dão pesos diferentes a classe dos dados, como o *ClassBalancer*. Ele soma a quantidade de registros das classes e dá um peso para cada classe de modo que todas as classes tenham o mesmo peso. Assim, todas as classes passam a ter peso igual, independentemente da quantidade de registros.

2.5 Trabalhos Relacionados

Sapori e outros (2017) discutiram os fatores sociais determinantes da reincidência criminal no Brasil, tratando em especial o caso de Minas Gerais. Eles selecionaram os presos que receberam livramento condicional ou cumpriram a pena no ano de 2008. Do total de 2.116 presos, selecionaram 800 presos aleatoriamente. A trajetória dos presos reincidentes foi acompanhada num período de 5 anos para saber se reincidiram no período. Do total de 800 presos, 51,4% reincidiram no período. Por meio da análise, fica evidenciado que o preso do sexo masculino e jovem tem uma chance de reincidência maior.

O Instituto de Pesquisa Econômica Aplicada (IPEA) e o Conselho Nacional de Justiça (CNJ) fizeram um relatório sobre a reincidência criminal em 2015. Eles analisaram apenas os presos que tiveram uma reincidência jurídica, ou seja, quando o preso reincidiu na concepção legal, preconizada nos artigos 63 e 64 da Lei de Execução Penal. Só é considerado reincidente aquele que teve mais de uma condenação, em ações penais distintas, transitada em julgado. Do total de 817 apenados, 199 reincidiram, o equivalente a 24,4%. Do total de presos, cerca de 57% dos reincidentes têm faixa etária entre 18 a 29 anos.

Machado e Sloniak (2015) discutiram a ressocialização prisional. Eles afirmam que se o trabalho fosse usado como uma ferramenta de ressocialização, o índice de reincidência

certamente reduziria. Para os autores “o sistema penitenciário participa do processo de degradação humana do condenado e frustra o ideal reintegrador do modelo clínico da LEP". Apesar de o departamento penitenciário federal considerar que o trabalho prisional é fundamental para a socialização do preso, o mesmo não adota políticas públicas e não disponibiliza verba suficiente.

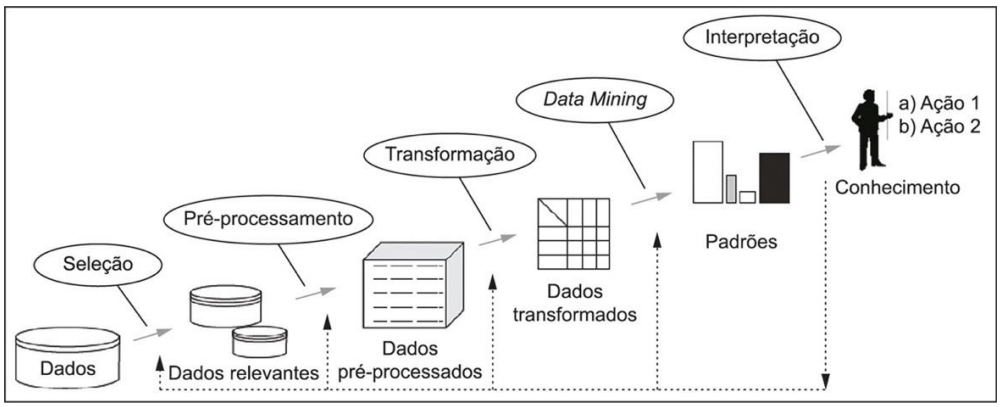
Barcinski e outros (2017) tratam o mesmo tema, porém sobre a ótica do agente de segurança penitenciário. O mesmo possui uma função ambígua no que tange a ressocialização do preso, pois ao mesmo tempo que tem o papel de agente ressocializador e facilitador da reinser- ção social do preso, tem como principal função a custódia do preso, o que inclui a disciplina e segurança. Para os autores, a “dupla missão da prisão, punir e educar, torna as relações estabe- lecidas entre presos e agentes penitenciários contraditórias e ambivalentes” (BARCINSKI et al., 2017). Eles definem que a ressocialização das presas se dá pela soma de 3 fatores primordiais: o esforço pessoal das presas, o esforço das agentes penitenciárias em realizar sua função e, por último, a sociedade e sua forma de receber as egressas do sistema penitenciário.

3 METODOLOGIA

A base de dados do Sistema Prisional possui mais de 800 mil registros de presos. Nesta base, se encontram todos os dados referentes ao preso, desde nome, identificador, filiação, data de nascimento, até a informação que afirma se o preso possui certidão de casamento, CPF e identidade. Quando uma pessoa é presa, são registrados o motivo da prisão, a data, a unidade prisional na qual ela será mantida e a qual instituição essa unidade pertence. Da mesma maneira, quando um preso é solto, são registrados o motivo do desligamento e a data.

Para identidicar o perfil do preso reincidente, foi utilizado um algoritmo de classificação na base de dados do Sistema de Informações Penitenciárias (INFOPEN) / Sistema de Gestão Prisional (SIGPRI) do Sistema Prisional de Minas Gerais. A metodologia seguiu o fluxograma apresentado na Figura 1.

Figura 1 – Descrição da Metodologia Proposta



Fonte: Steiner e outros (2006).

A metodologia é dividida em cinco etapas a saber: seleção de dados, pré-processamento, transformação dos dados, mineração de dados e interpretação dos resultados. Na primeira etapa, foram selecionados os indivíduos que foram presos em unidades do Sistema Prisional e que foram desligados por término da pena ou livramento condicional. A próxima etapa foi o pré-processamento da base de dados. A base possuía dados desde 1994, salvos por meio de uma plataforma baixa, em que, na maioria dos campos, não havia nenhum tipo de validação. São encontrados, então, registros de presos sem nome, com nome “desconhecido", “a apurar". Por existirem muitos registros, foram removidos da tabela aqueles dados que possuíam características chaves vazias ou nulas. Após esta etapa, foram tratados os dados inconsistentes, presos com sexo fora do domínio da aplicação, datas que não existiam, unidade de admissão ou desligamento que não existia, ou que não havia sido informada.

A terceira etapa foi a de transformar os dados numéricos em dados nominais, criar uma faixa para idade do preso e elaborar uma faixa significativa para a escolaridade. Foi realizada uma pesquisa na base para verificar se o preso possuía admissão após o livramento condicional ou cumprimento de pena, para classificá-lo como reincidente ou não. Após essa etapa, foi realizado o balanceamento da base de dados com o auxílio do algoritmo de balanceamento *ClassBalancer*.

A quarta etapa foi a de descoberta de conhecimento por meio da mineração de dados. Foi utilizado o algoritmo *j48*, para se obter o perfil do preso reincidente no Sistema Prisional de Minas Gerais. O algoritmo *j48* é utilizado para classificação e tem a finalidade de gerar

uma árvore de decisão. Segundo Vieira e outros (2018), ele surgiu da implementação em java do algoritmo *c4.5*. Ele é um algoritmo recursivo. Em cada rodada, ele verifica qual atributo possui mais entropia em relação aos outros, mantendo-o no nó. Esse processo é repetido várias vezes, até se chegar às folhas da árvore. Para Vieira e outros (2018), uma de suas vantagens é a de se poder utilizar dados contínuos, como a idade, e dados discretos, como a ocupação. Foram utilizados a ferramenta *Weka* para a mineração dos dados e o algoritmo *Cross-validation* com 10 *folds* para a classificação. A quinta etapa foi a de interpretação das regras geradas no processo.

4 APRESENTAÇÃO E ANÁLISE DE RESULTADOS

Na primeira subseção, foram discutidas a seleção, o pré-processamento dos dados e sua transformação. Na segunda subseção, foram apresentadas as métricas utilizadas no processo de mineração de dados. Na terceira subseção, são analisados as regras identificadas.

4.1 Seleção, Pré-processamento e Transformação dos Dados

Foram selecionados, para os experimentos, todos os presos que passaram no sistema prisional que receberam o livramento condicional ou cumpriram sua pena, também sendo verificado se os mesmos reincidiram no sistema prisional posteriormente. O próximo filtro foi a seleção da instituição onde o preso estava encarcerado, a SEAP. Foram selecionados apenas os presos do sexo masculino e feminino, os presos que tinham uma data de nascimento válida, e que possuíam mais de 18 anos de idade. Algumas das inconsistências encontradas encontram-se no Quadro 1.

Quadro 1 – Inconsistências encontradas na base de dados

Atributo	**Escopo**	**Inconsistências**
Sexo	M ou F	0, 1, 5 e 9
Data de nascimento	DD/MM/AAAA	YYYY, MM/YYYY, 99/99/9999, etc...
Idade	Maiores de 18 anos	15, 17

Fonte: Elaborado pelos autores.

Após a conclusão da etapa de pré-processamento, foi executada a etapa de transformação numérico-nominal e nominal-numérico. Foi realizada a discretização da idade em 9 faixas etárias. A escolaridade foi convertida de nominal para numérica, sendo 1 para analfabetos até

9, para aqueles que possuíam pós-graduação. O estado civil foi mantido como registrado. Um problema identificado foi relativo às profissões/ocupações dos 50.011 presos. Haviam 450 ocupações distintas, o que foi prejudicial para o processo de aprendizado do algoritmo de classificação numa primeira análise realizada. Devido a isso, foram realizadas conversões das ocupações. As novas ocupações foram baseadas nas seis ocupações do Relatório de Reincidência Criminal realizado pelo IPEA e CNJ (2015).

A próxima transformação realizada foi a da cútis do preso. Na base de dados, existem apenas 4 tipos de cútis, a branca, a amarela, a parda e a negra. Elas foram convertidas em dados numéricos de 1 a 4, de acordo com o nível de pigmentação da melanina, ficando 1 para branca, 2 para amarela, 3 para parda, 4 para negra e 99 para aqueles que não possuem cútis registrada. A última transformação realizada foi relativa ao enquadramento dos presos. No INFOPEN/SIGPRI, os enquadramentos são cadastrados contendo o número da lei e o artigo. Por exemplo, um cidadão que é preso por praticar um roubo tem o registro de enquadramento 'DL 2848 ART 157', ou seja, o mesmo infringiu o artigo 157 do decreto-lei 2848 (Código Penal Brasileiro). Existem em torno de 370 tipos de enquadramentos distintos, o que gerou um trabalho manual significativo. Foram realizadas duas conversões, a primeira baseada no trabalho realizado pelo Sapori e outros (2017) e a outra em 4 grandes grupos de crimes mostrados no trabalho do *Bureau of Justice Statistics* (2018) conforme apresentado no Quadro 2.

Quadro 2 – Transformação relativa ao Enquadramento

Lei	Breve Descrição	E. Codebook	E. Sapori
L 9605	Atividades lesivas ao meio ambiente	Property	Outros
L 9504	Lei Eleitoral	Property	Outros
L 9503	Trânsito	Property	Outros
L 9472	Telecomunicações	Property	Outros
L 9455	Crimes de ortura	Violent	Outros
L 9437/10826	Armas de fogo	Public Order	Arma de fogo
L 8176	Ordem econômica	Property	Outros
L 8137	Ordem tributária	Property	Outros
L 8072	Crimes hediondos	Violent	Outros
L 8069	ECA	Violent	Outros

L 11343 / 6368	Tráfico de drogas	Drug	Tráfico
L 5869 / 13105	Direito Civil	Property	Outros
L 12850	Organização criminosa	Violent	Outros
L 3688	Contravenção Penal	Property	Outros
L 2252	Corrupção de menor	Violent	Corrupção de Menor
L 7661	Plano costeiro	Property	Outros
L 8072	Crimes hediondos	Violent	Outros
L 10741	Estatuto do idoso	Violent	Outros
L 11101	Empresarial	Property	Outros
L 11340	Maria da Penha	Violent	Outros
DL 2848 art 121	Homicídio	Violent	Homicídio
DL 2848 art 129	Lesão corporal	Violent	Lesão corporal
DL 2848 art 155	Furto	Property	Furto
DL 2848 art 157	Roubo	Violent	Roubo
DL 2848 art 171	Estelionato	Property	Estelionato
DL 2848 art 171	Receptação	Property	Receptação
DL 2848 art 213	Estupro	Violent	Estupro

Fonte: Elaborado pelos autores.

O Quadro 3 descreve, na segunda coluna, cada uma das variáveis, da primeira coluna, selecionadas para o processo, e apresenta suas respectivas categorias na terceira coluna.

Quadro 3 – Descrição das variáveis e das categorias

Variável	Definição	Categorias
Faixa etária	Foi considerada a idade atual do preso	Menor de 25 anos 25 a 29 anos 30 a 34 anos 35 a 39 anos 40 a 44 anos 45 a 49 anos

		50 a 54 anos 55 a 59 anos A partir de 60 anos
Possui pai	Informação disponível	Sim Não
Sexo	Sexo do preso	Masculino Feminino
Escolaridade	Nível de escolaridade informado pelo preso	1 - Analfabeto 2 - Semi-alfabetizado 3 – Fundamental Incompleto 4 – Fundamental Completo 5 – EM Incompleto 6 – EM Completo 7 - Superior Incompleto 8 - Superior Completo 9 - Pós-graduado 99 - Não Informado
Estado civil	Estado civil informado pelo preso	Solteiro Casado União estável Concubinato Divorciado Separado/Desquitado Viúvo Não Informado
Ocupação	Ocupação informada pelo preso	Aposentado Desempregado Estudante Ocupado Sem profissão
Cútis	Cor da pele	1 - Branca

		2 - Amarela 3 - Parda 4 - Negra 99 - Não Informado
Idade que saiu	Idade que o preso recebeu o livramento condicional ou termino da pena	
Enquadramento Codebook	Tipo de crime relacionado à pena do preso	Violent Property Public Order Drug
Enquadramento Sapori	Tipo de crime relacionado à pena do preso	Tráfico Roubo Furto Homicídio Arma de fogo Estupro Lesão corporal Estelionato Receptação Corrupção de menor Outros Sem informação
Reincidente	Se voltou a ser preso após a saída da unidade prisional	Não Sim

Fonte: Elaborado pelos autores.

4.2 Métricas Utilizadas

As métricas utilizadas têm como objetivo avaliar os resultados encontrados pelo algoritmo *j-48*. Para Castro e Ferrari (2016), o desempenho de um algoritmo de classificação pode ser avaliado por meio de uma matriz de confusão, conforme mostra a Figura 2.

Figura 2 - Exemplo de Matriz de Confusão

		Classe predita	
		Positiva	Negativa
Classe original	Positiva	*VP*	*FN*
	Negativa	*FP*	*VN*

Fonte: Castro e Ferrari (2016).

Nesta matriz, VP é verdadeiro positivo e significa um objeto da classe positiva sendo classificado como positivo. VN é verdadeiro negativo e significa um objeto da classe negativa sendo classificado como negativo. FP é falso positivo e significa um objeto da classe negativa sendo classificado com positivo. E FN é falso negativo e significa um objeto da classe positiva sendo classificado como negativo. A partir da matriz de confusão, consegue-se extrair as seguintes métricas: acurácia, precisão, *recall*, medida-F dentre outras. A acurácia é a porcentagem de instâncias corretamente classificadas. Para Shalev-Shwartz e Ben-David (2014), a medida de acurácia não é suficiente para considerar o classificador como eficiente. Devido a isso, neste trabalho, foram utilizadas as métricas precisão e *recall*. A precisão, descrita pela Equação 1, é o percentual de instâncias classificadas corretamente como positivas dentre todas as instâncias que foram classificadas como positivas. O *recall*, descrito pela Equação 2, é o percentual de instâncias classificadas corretamente como positivas dentre todas as instâncias daquela classe.

$$Pr = VP/(FP + VP) \quad (1)$$

$$Re = VP/(FN + VP) \quad (2)$$

O primeiro experimento foi realizado para avaliar a codificação do enquadramento e identificar a codificação que fez com que o algoritmo tivesse uma melhor taxa de acerto. Utilizando a codi- ficação proposta pelo trabalho do *Bureau of Justice Statistics* (2018), obteve-se uma acurácia de 67,38%. A matriz de confusão encontra-se no Quadro 4, com 69,3% de precisão e 62,4% de *recall* para os não reincidentes, e 65,8% de precisão e 72,4%

de *recall* para os reincidentes. O modelo teve médias de precisão e *recall* de 67,6% e 67,4% respectivamente.

Utilizando a codificação proposta pelo trabalho de Sapori e outros (2017), obteve-se uma acurácia de 67,71%. Com 69,8% de precisão e 62,5% de *recall* para os não reincidentes, e 66,1% de precisão e 72,9% de *recall* para os reincidentes. O modelo teve médias de precisão e *recall* de 67,9% e 67,7% respectivamente.

Quadro 4 – Comparativo

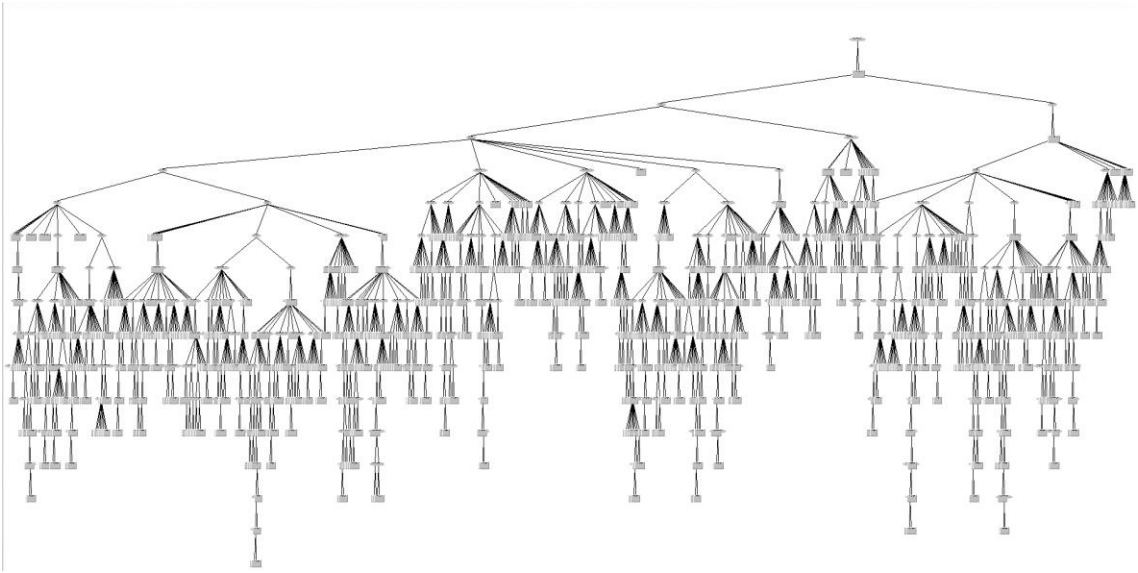
Matriz de Confusão							
Codebook	**Não**	**Sim**		**Sapori**	**Não**	**Sim**	
	15600,17	9405,33	**Não**		15638,31	9367,19	**Não**
	6908,27	18097,23	**Sim**		6781,45	18224,05	**Sim**
Métricas							
Codebook	**Precisão**	***Recall***	**Classe**	**Sapori**	**Precisão**	***Recall***	**Classe**
	0,693	0,624	**Não**		0,698	0,625	**Não**
	0,658	0,724	**Sim**		0,661	0,729	**Sim**
	0,676	0,674	**Média**		0,679	0,677	**Média**

Fonte: Elaborado pelos autores.

4.3 Análise dos Resultados

Ambas as codificações alcançaram resultados semelhantes quando avaliadas pelas métricas descritas. Optou-se por ficar com a codificação feita por Sapori e outros (2017) por ser uma codificação brasileira e por também ser bastante semelhante à usada pelo IPEA e pelo CNJ. A árvore criada pelo algoritmo possui 1.268 folhas. Isto significa que foram criadas 1.268 regras para classificar um preso, conforme apresentado na Figura 3.

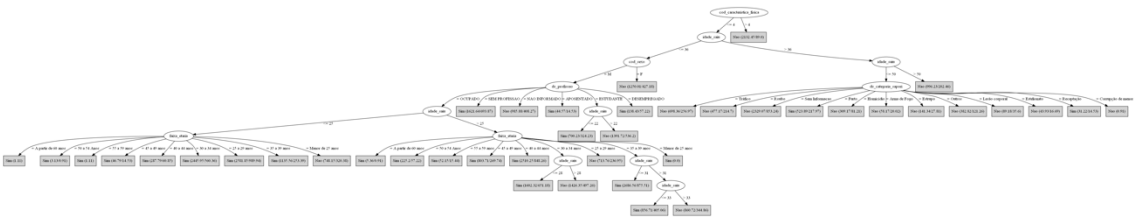
Figura 3 - Árvore gerada sem poda



Fonte: Elaborada pelos autores.

Foi preciso fazer o procedimento de poda da árvore para que seja reduzido o número de folhas, de modo que não se perca muito a qualidade gerada. O procedimento foi realizado no *Weka*, alterando os seguintes atributos do algoritmo *J48*: *minNumObj* - Número mínimo de instâncias por folha de 2 para 500 e *reducedErrorPruning* - reduzir a quantidade de erro na poda de falso para verdadeiro. Um percentual de 66,64% instâncias foram classificadas corretamente, cerca de 3% a menos. A imagem da nova árvore encontra-se na Figura 4 e a nova matriz de confusão encontra-se no Quadro 5. A nova precisão foi de 68,1% e 62,5% de *recall* para os não reincidentes, e 65,4% de precisão e 70,7% de *recall* para os reincidentes. O modelo teve médias de precisão e sensibilidade próximas, sendo 66,8% e 66,6% respectivamente.

Figura 4 - Árvore gerada com poda



Fonte: Elaborada pelos autores.

Quadro 5 – Métricas

Matriz de Confusão		
Não	**Sim**	
15639,22	**9366,28**	**Não**
7314,31	**17691,19**	**Sim**
Métricas		
Precisão	***Recall***	**Classe**
0,681	**0,625**	**Não**
0,654	**0,707**	**Sim**
0,668	**0,666**	**Média**

Fonte: Elaborado pelos autores.

A partir da árvore, foram extraídas as regras associadas aos presos do Sistema Prisional. Foram apresentadas, no Quadro 6, as 7 regras mais relevantes sobre os presos reincidentes.

Quadro 6 – Regras

Número	**Regra**	**Cobertura**
1	SE cútis <= 4 E sexo = Masculino E ocupação = Ocupado E idade que saiu <= 25 E faixa etária = 25 a 29 anos ENTÃO Reincidente	6,366%
2	SE cútis <= 4 E sexo = Masculino E ocupação = Ocupado E idade que saiu <= 25 E faixa etária = 30 a 34 anos ENTÃO Reincidente	7,541%
3	SE cútis <= 4 E sexo = Masculino E ocupação = Ocupado E idade que saiu <= 25 E faixa etária = 35 a 39 anos ENTÃO Reincidente	3,528%
4	SE cútis <= 4 E sexo = Masculino E ocupação = Ocupado E 25 < idade que saiu <= 31 E faixa etária = 35 a 39 anos ENTÃO Reincidente	7,243%
5	SE cútis <= 4 E sexo = Masculino E ocupação = Ocupado E 25 < idade que saiu <= 36 E faixa etária = 40 a 44 anos ENTÃO Reincidente	6,646%
6	SE cútis <= 4 E sexo = Masculino E ocupação = Ocupado E 25 < idade	3,804%

	que saiu <= 28 E faixa etária = 30 a 34 anos ENTÃO Reincidente	
7	SE cútis <= 4 E sexo = Masculino E ocupação = Sem profissão E idade que saiu <= 36 ENTÃO Reincidente	3,710%

Fonte: Elaborado pelos autores.

Analisando as regras extraídas, percebe-se que o preso que reincide é basicamente do sexo masculino. Nas regras de número 1 a 3, todos os presos saíram da unidade prisional com menos de 26 anos, eram do sexo masculino e tinham alguma ocupação, diferenciando-se pelas faixas etárias. Desde a faixa etária de 25 a 29 até a de 35 a 39 anos, têm-se uma alta taxa de reincidência. Na regra 4, percebe-se o mesmo padrão de sexo e ocupação, porém a idade que o preso saiu foi maior que 25 anos e menor que 31 anos, sendo que, atualmente, está na faixa etária de 35 a 39 anos. As regras 5 e 6 também possuem a semelhança do sexo e ocupação. Em ambas as regras, os presos saíram com mais de 25 anos. Na regra 5, eles saíram com menos de 36 anos e, na regra 6, um pouco mais jovem, com menos de 28 anos. Estes reincidem na faixa etária subsequente à sua saída. Isso mostra que eles cometem um novo crime num curto período de tempo. Na regra 7, percebe-se que os presos que não tinham profissão e saíram da unidade com menos de 37 anos reincidiram. A cobertura das regras chega a quase 40%. Essas não foram as únicas regras retornadas pela árvore de decisão, mas foram as mais relevantes.

5 CONSIDERAÇÕES FINAIS E TRABALHOS FUTUROS

Percebe-se pelas regras extraídas da árvore de decisão que, resumidamente, os presos do sexo masculino e que saem da unidade prisional ainda jovem são os que reincidem. Subentende- se que o preso que cumpre sua pena ou recebe o livramento condicional ainda jovem, tem começado mais cedo sua vida criminal. Sapori e outros (2017), em seu resultado da análise de regressão, chegaram à conclusão que “no que se refere aos atributos sociodemográficos dos presos liberados em 2008 em Minas Gerais, os resultados apontam que a probabilidade de reincidência é, de fato, maior para homens e quanto menor for a idade do indivíduo” (SAPORI et al., 2017, p.13).

O IPEA e o CNJ (2015), em seu relatório de pesquisa sobre a reincidência criminal no Brasil, afirmam que, “em síntese, a parcela de reincidentes da amostra é composta basicamente de homens jovens, brancos, de baixa escolaridade e com uma ocupação”. Os resultados obtidos neste trabalho confirmam os resultados de outros trabalhos a respeito do assunto. A única discordância diz respeito à cor da pele e à baixa escolaridade.

Deve-se considerar que, das regras extraídas, a probabilidade da presa mulher reincidir é muito baixa. Todavia, nesse trabalho, não houve a preocupação de balancear a quantidade de homens e mulheres, ou separá-las em outra base de dados. Os resultados obtidos a partir de uma nova base podem trazer uma regra mais objetiva quanto ao assunto.

Deve-se considerar que esse trabalho não levou em consideração a região onde o preso morava, nem se ele recebia visitas de seus parentes próximos, como pai, mãe, mulher e filhos. Estes fatores podem contribuir para que o preso não regresse ao Sistema Prisional.

Uma proposta de continuidade dos estudos para tratar questões identificadas nesse trabalho seria a de realizar a separação da base de dados em duas, uma para o sexo masculino e outra para o feminino, com o intuito de descobrir o padrão para mulheres que reincidem, que não foi encontrado nesse trabalho. Outra proposta seria a do enriquecimento da base de dados com outras variáveis sociodemográficas como as citadas anteriormente. Além disso, existem outros algoritmos de aprendizado de máquina, como é o caso da Rede Neurais Artificiais, que podem encontrar uma taxa de acerto maior.

Tendo em vista os aspectos relativos ao perfil do preso reincidente, confirma-se que uma política pública que trabalhe com o jovem infrator, mesmo após sua saída da unidade prisional, pode ser uma boa alternativa para redução da reincidência prisional. As análises sugerem que a continuidade do acompanhamento dos presos pelos assistentes sociais e psicólogos poderia auxiliá-lo em sua reinserção na sociedade e evitar a reincidência.

REFERÊNCIAS

BARBIERI, Carlos. BI2 - Business Intelligence: modelagem e qualidade. Rio de Janeiro: Elsevier, 2011. ISBN 9788535247220.

BARCINSKI, Mariana; CUNICO, Sabrina Daiana; BRASIL, Marina Valentim. Significa- dos da Ressocialização para Agentes Penitenciárias em uma Prisão Feminina: Entre o Cuidado e o Controle. Trends in Psychology, v. 25, p. 1257 – 1269, 09 2017. ISSN 2358-1883. BUREAU OF JUSTICE STATISTICS. Survey of Inmates in State and Federal Correctional Facilities. United States: Inter-university Consortium for Political and Social Research, 2018.

CABRAL, Luisa Rocha; SILVA, Juliana Leite. O trabalho penitenciário e a ressocialização do preso no brasil. Revista do CAAP, v. 2010, n. 1, 2010.

CASTRO, Leandro Nunes de; FERRARI, Daniel Gomes. Introdução à Mineração de Dados: Conceitos básicos, algoritmos e aplicações. São Paulo: Editora Saraiva, 2016. ISBN 978-85-472-0100-5.

CNJ, Conselho Nacional de Justiça. Cadastro nacional de presos. In: Banco Nacional de Monitoramento de Prisões - BNMP 2.0. Brasília: [s.n.], 2018.
FACELI, Katti et al. Inteligência artificial: uma abordagem de aprendizado de máquina. [S.l.]: LTC, 2011.

GRAMANI, Maria Cristina N; SANTORO, Miguel Cezar. Problema de balanceamento de linhas: Modelagem e abordagens de resolução. 2006.

IPEA, Instituto Pesquisa Econômica Aplicada; CNJ, Conselho Nacional de Justiça. Relatório final de atividades da pesquisa sobre reincidência criminal, conforme acordo de cooperação técnica entre o conselho nacional de justiça e o ipea. In: Reincidência criminal no Brasil. Brasília: [s.n.], 2015.

ISTOé. População carcerária no Brasil já é a terceira maior do mundo. Disponível em: <https://istoe.com.br/populacao-carceraria-no-brasil-ja-e-terceira-maior-do-mundo/>. Acesso em: 07 jun. 2019.

MACHADO, Bruno Amaral; SLONIAK, Marcos Aurélio. Disciplina ou ressocialização? racionalidades punitivas, trabalho prisional e política penitenciária. Rev. direito GV, São Paulo, v. 11, n. 1, p. 189–222, jun. 2015. PASSOS, Emmanuel; GOLDSCHMIDT, Ronaldo. Data mining: um guia prático: conceitos, técnicas, ferramentas, orientações e aplicações. Rio de Janeiro: Elsevier, 2005. ISBN 8535218777.

QUINTELLA, Rogério Hermida; JUNIOR, Jair Sampaio Soares. Sistemas de apoio à decisão e descoberta de conhecimento em bases de dados: uma aplicação potencial em políticas públicas . Organização Sociedade, v. 10, p. 83 – 98, 12 2003. ISSN 1984-9230.

SAPORI, Luis Flávio; SANTOS, Roberta Fernandes; MAAS, Lucas Wan Der. FATORES SOCIAIS DETERMINANTES DA REINCIDÊNCIA CRIMINAL NO BRASIL: O CASO DE MINAS GERAIS. Revista Brasileira de Ciências Sociais, v. 32, 00 2017. ISSN 0102-6909.

SEAP. Secretaria de Estado e Administração Prisional. Disponível em: <http://www.seap.mg.gov.br/index.php/a-secretaria/sobre>. Acesso em: 07 jun. 2019.

SHALEV-SHWARTZ, Shai; BEN-DAVID, Shai. Understanding Machine Learning: From Theory to Algorithms. New York, NY, USA: Cambridge University Press, 2014. ISBN 1107057132, 9781107057135.

STEINER, Maria Teresinha Arns et al. Abordagem de um problema médico por meio do processo de KDD com ênfase à análise exploratória dos dados. Gestão Produção, v. 13, p. 325 – 337, 05 2006. ISSN 0104-530X.

TAN, Pang Ning; STEINBACH, Michael; KUMAR, Vipin. Introdução ao DATAMINING Mineração de Dados. Rio de Janeiro: Ciência Moderna, 2009. ISBN 9788573937619.

VIEIRA, Elamara Marama de Araujo et al. Avaliação da performance do algoritmo j48 para construção de modelos baseados em árvores de decisão. Revista Brasileira de Computação Aplicada, v. 10, n. 2, p. 80–90, jul. 2018.