

Aplicação de Algoritmos de Mineração de Dados para Traçar um Panorama da Situação das Unidades Prisionais no Brasil

¹Odolir Daniel dos Santos Junior, ¹Anita Maria da Rocha Fernandes, ²Rodrigo de Souza Oliveira

¹Curso de Ciência da Computação – Escola do Mar, Ciência e Tecnologia - Universidade do Vale do Itajaí (UNIVALI) – Campus Kobrasol – SC – Brasil - Campus Kobrasol – SC – Brasil

²MBA em Big Data – FATEC SENAI. Cuiabá - MT

odolirdaniel@gmail.com, anita.fernandes@univali.br, rddsouzaoliveira@gmail.com

***Abstract.** This paper presents a proposal of applying data mining algorithms in a database composed by data from INFOPEN, IBGE and DATASUS, seeking to draw an overview of prisons in Brazil. Data from 2000 to 2018, categorized by Brazilian cities are being crossed on a single basis and from there will be treated and prepared for analysis by mining algorithms.*

***Resumo.** Este artigo apresenta a proposta de aplicação de algoritmos de Mineração de Dados em uma base de dados composta por dados oriundos do INFOPEN, IBGE e DATASUS, buscando traçar um panorama das unidades prisionais no Brasil. Dados de 2000 a 2018, categorizados por cidades brasileiras estão sendo cruzados em uma base única e a partir daí serão tratados e preparados para serem analisados pelos algoritmos de mineração.*

1. Introdução

A população prisional brasileira segue a tendência mundial e vive em uma situação crescente de encarceramento. De acordo com o Levantamento Nacional de Informações Penitenciárias, publicado em 2016 [Infopen 2016], o Brasil teve um aumento na população carcerária de 267,32% nos últimos 14 anos, apresentando 622.202 mil presos, sendo o quarto país que mais prende no mundo. Lermen et al (2015), apresentam que o aumento da população carcerária não foi acompanhado de melhorias das condições físicas e estruturais das prisões do país ,na proporção em que eram necessárias.

No Brasil o Ministério da Justiça e Segurança Pública, através do Departamento Penitenciário Nacional possui o INFOPEN, que é um sistema que gera estatísticas básicas sobre o sistema carcerário brasileiro. O sistema, atualizado pelos gestores dos estabelecimentos, apresenta dados sobre as unidades penais do país, disponibilizados em formato de uma planilha Excel, e referem-se à descrição das unidades prisionais, infraestrutura, capacidade, recursos humanos, população, perfil da população, dentre outros atributos. Porém, estes dados não estão analisados no sistema do INFOPEN de maneira a apresentar um diagnóstico sobre os municípios que recebem uma unidade prisional no que tange aos dados sócio-econômicos, geográficos e de saúde. Estes são

alguns exemplos de levantamentos que podem ser feitos e explorados e a partir deles pode-se gerar informações para os gestores, que a princípio passaram imperceptíveis.

Neste contexto, foi proposta a aplicação de técnicas de mineração de dados em um conjunto de dados formados pelos dados do INFOPEN, complementados pelos dados do IBGE e DATASUS, a fim de gerar subsídios para os gestores da área, buscando contribuir com a melhoria do sistema carcerário. Vale lembrar que todas estas bases são de domínio público. Para aplicação dos algoritmos estão sendo utilizadas as bibliotecas do Python e R. Após a aplicação dos algoritmos, os resultados obtidos serão analisados por especialista da área e em seguida serão apresentados em uma ferramenta online de visualização de dados.

2. Metodologia

Para a execução da proposta, cinco etapas foram estabelecidas, começando por um levantamento bibliográfico sobre os algoritmos possíveis de serem utilizados.

Na sequência, foi necessário criar uma base com dados oriundos do INFOPEN, IBGE, DATASUS e realizar o tratamento e limpeza dos dados. O tratamento dos dados coletados para composição da base integrada foi feito em termos de valores ausentes, valores fora do padrão, avaliação de dados inconsistentes, integração de esquemas internos, redundância de dados, detecção e resolução de valores conflitantes. Em alguns casos foi feita a transformação dos dados quando havia necessidade de remoção de ruídos, agregação ou sumarização de dados, bem como a generalização ou normatização de algum atributo.

Após estas etapas serem executadas, será necessário escolher os algoritmos de mineração de dados a serem utilizados, aplicá-los e avaliar os resultados obtidos. Após esta análise, será criada uma ferramenta online para visualização dos resultados obtidos. Antes da escolha dos algoritmos, será feita uma análise estatística prévia utilizando a ferramenta R. Os algoritmos que vão ser aplicados serão implementados em Python.

3. Resultados preliminares

Dentre os objetivos propostos para a execução do trabalho, criou-se uma base de dados única, com informações do INFOPEN, IBGE e DATASUS e foi realizada a limpeza dos dados. Para a limpeza dos dados utilizou-se a ferramenta Pandas [McKinney e Kinoshita 2018].

Uma das dificuldades apresentadas para a junção das tabelas provenientes de diferentes fontes está na disponibilidade das mesmas, tendo em vista que em alguns anos, as informações das 3 fontes de dados, não foram atualizadas, mesmo assim, foi possível criar uma base com dados de 2000 a 2018.

A pesquisa nas bases do DATASUS e do IBGE baseou-se nos dados que já estavam nas tabelas oriundas do INFOPEN. A partir destes dados, buscou-se nas duas bases os dados complementares. Os dados do INFOPEN são apresentados em forma de uma planilha com dados sobre infraestrutura, recursos humanos, capacidade, gestão, assistências, população prisional, perfil das pessoas presas, entre outros.

Além de temas e subtemas, tais como quantidade de pessoas presas, estabelecimento prisional, sistema prisional, vagas, assistência laboral, assistência educacional, saúde, tipo penal, tempo de pena, visitas. Cada planilha anual é composta

de 1422 linhas (unidades prisionais) e cerca de 800 colunas (variáveis) que descrevem todas as unidades carcerárias nacionais. Os dados nelas apresentados são sumarizados na forma de tabelas de frequência e médias em um relatório anual. Porém nenhuma outra análise é feita, e os relacionamentos de tais dados com a estrutura dos municípios que recebem as unidades não são analisados.

Para organizar os dados de forma efetiva, optou-se pelo banco não relacional, MongoDB facilitando a manipulação de tais informações. Também foram utilizadas técnicas de limpeza e tratamento de dados, para garantir a qualidade das informações.

O próximo passo refere-se a escolha e implementação dos algoritmos de mineração para obter uma visão sobre essas informações.

Referências

Brasil, (2016) *Levantamento Nacional de Informações Penitenciárias*. Disponível em Acessado em 20 de janeiro de 2019. <https://www.conjur.com.br/dl/infopen-levantamento.pdf>.

INFOPEN, (2014) *Levantamento Nacional de Informações Penitenciárias*. Acessado em 03 de março de 2019. Disponível em http://www.justica.gov.br/seus-direitos/politica-penal/documentos/infopen_dez14.pdf.

Lermen, H. S.; Gil B. L. Cúnico, S.D; Jesus, L.O. *Saúde no: análise das políticas sociais de saúde voltadas à população prisional brasileira*. Physis Revista de Saúde Coletiva, Rio de Janeiro, 25 [3]: 905-924, 2015.

McKinney, W.; Kinoshita, L. A. (2018). *Python Para Análise de Dados: Tratamento de Dados com Pandas, NumPy e IPython*. Novatec; Edição: 1ª (9 de janeiro de 2018).