Proposta do projeto final de Data Science - Violência Armada e suas características e incidências nos EUA.

Lerisson Florêncio de Freitas
lff3@cin.ufpe.br,
Matheus Raz de Oliveira Leandro
mrol@cin.ufpe.br

Universidade Federal de Pernambuco

Universidade Federal de Pernambuco
CIn - Centro de Informática

Recife
2018

Lerisson Florêncio de Freitas
lff3@cin.ufpe.br,
Matheus Raz de Oliveira Leandro
mrol@cin.ufpe.br

Proposta do projeto final de Data Science - Violência Armada e suas características e incidências nos EUA.

Proposta de projeto levantada com o objetivo de construir um projeto como parte dos requisitos para a conclusão da disciplina IF1015 Intro a Ciência dos Dados.

Área de concentração: Ciência da Computação

Orientador: Dr. Renato Vimieiro

Recife
2018

Capítulo 1

Proposta

1.1 Base

Nossa base de estudo foi coletada no repositório online Kaggle para aprendizagem de máquina: Dados de violência armada. Registro abrangente de mais de 260 mil incidentes de violência armada nos EUA entre 2013-2018

1.2 Informações sobre a base

Registrou-se mais de 260 mil incidentes de violência armada, com informações detalhadas sobre cada incidente, disponíveis no formato CSV. O arquivo CSV contém dados de todos os incidentes registrados de violência armada nos EUA entre janeiro de 2013 e março de 2018, inclusive.

Coleta dos dados:

- **Fase 1:** para cada data entre 1/1/2013 e 31/03/2018, um script Python consultou todos os incidentes que ocorreram naquela data em particular, depois digitalizou os dados e os escreveu em um arquivo CSV. Cada mês tem seu próprio arquivo CSV, com exceção de 2013, já que não foram registrados muitos incidentes a partir de então.

- **Fase 2:** cada entrada foi aumentada com dados adicionais que não podem ser visualizados diretamente na página de resultados da consulta, como informações do participante, dados de geolocalização etc.

- **Fase 3:** as entradas foram classificadas em ordem crescente e depois mescladas em um único arquivo CSV.

1.3 Proposta

Atualmente, faltam quantidades grandes e facilmente acessíveis de dados detalhados sobre a violência armada em geral. Para tal resolvemos focar na análise do contexto Norte Americano por ser um país que permite a regulamentação do uso de armas em alguns estados, gerando assim um índice de violência que demanda uma maior incidência de crimes cometidos com o uso delas. Dessa forma, usaremos o método Self-organizing map (som) para verificar e apoiar a validação de nossas hipóteses. Como o método faz uso de uma abordagem não supervisionada de aprendizagem, através de redes neurais e separa as labels por região de aproximação da matriz de neurônios, como dito tentaremos confirmar nossos estudos através desta.

1.4 Metodologia

Podemos identificar e levantar padrões que nos levem a entender melhor a necessidade de um reforço em tais áreas com maior índice de ocorrências desse viés avaliando correlações entre os atributos presentes no *dataset*, onde aplicamos a análise exploratória sobre os dados para corroborar hipóteses levantadas por nós, tais quais:

Hipóteses:

1. A quantidade de incidentes cometidos por menores de idade é maior em estados que possuem baixo investimento em educação.
2. Os estados com alto Índice de Pobreza possuem uma maior mortalidade em incidentes do que estados que possuam baixo índice.
3. A quantidade de mulheres que são vítimas em crimes tende a diminuir ao longo dos anos em cada estado.

No que diz respeito a análise exploratória será montadas a visualização destes dados através do que foi aprendido na disciplina com uso de gráficos que carregam consigo relações de medidas estatísticas, para verificar e entender o comportamento dos dados. Tudo isso com o objetivo de obter a validação ou a refutação das hipóteses anteriores. Podemos afirmar que nessa etapa procuraremos descobrir relações como:

1. A distribuição de crimes a mão armada por grupos de pessoas no país.
2. A evolução desses crimes nos condados no decorrer dos anos
3. A relação de Desemprego por pobreza.

4. Cruzar a relação acima pela incidência de crimes armados.

Também será construído um mapa de calor para inferir o que não foi coberto pela a análise por se só dos dados. Assim ficará clara a distribuição desses crimes no EUA.

Após as análises descritas acima, será aplicada, também, a técnica de Clusterização nos dados de saída do (SOM) com o objetivo de extrair alguns padrões nessa base de dados, e predizer grupos para *clusters* dos crimes coletados.