UNIVERSIDADE FEDERAL DO ESPÍRITO SANTO

CENTRO TECNOLÓGICO

PROGRAMA DE PÓS-GRADUAÇÃO EM ENGENHARIA ELÉTRICA

EDUARDO RIBEIRO VARGAS DUARTE

UTILIZANDO CIÊNCIA DE DADOS PARA UM MELHOR ENTEDIMENTO DO CENÁRIO DA VIOLÊNCIA NO ESPÍRITO SANTO

VITÓRIA

2020

EDUARDO RIBEIRO VARGAS DUARTE

UTILIZANDO CIÊNCIA DE DADOS PARA UM MELHOR ENTEDIMENTO DO CENÁRIO DA VIOLÊNCIA NO ESPÍRITO SANTO

Dissertação de mestrado apresentada ao Programa de Pós-Graduação em Engenharia Elétrica da Universidade Federal do Espírito Santo, como requisito para a obtenção do título de Mestre em Engenharia Elétrica, na área de Telecomunicações.
Orientador: Prof. Dr. Anilton Salles Garcia

VITÓRIA

2020

Ficha catalográfica disponibilizada pelo Sistema Integrado de Bibliotecas - SIBI/UFES e elaborada pelo autor

---

Ribeiro Vargas Duarte, Eduardo, 1991-
R484u Utilizando ciência de dados para um melhor entendimento da situação da violência no Espírito Santo / Eduardo Ribeiro Vargas Duarte. - 2020.
102 f. : il.

Orientador: Anilton Salles Garcia.
Coorientador: Jair Adriano Lima Silva.
Dissertação (Mestrado em Engenharia Elétrica) - Universidade Federal do Espírito Santo, Centro Tecnológico.

1. Aprendizado do computador. 2. Banco de dados. 3. Violência. I. Salles Garcia, Anilton. II. Adriano Lima Silva, Jair. III. Universidade Federal do Espírito Santo. Centro Tecnológico. IV. Título.

CDU: 621.3

EDUARDO RIBEIRO VARGAS DUARTE

UTILIZANDO CIÊNCIA DE DADOS PARA UM MELHOR ENTEDIMENTO DO CENÁRIO DA VIOLÊNCIA NO ESPÍRITO SANTO

Dissertação de mestrado apresentada ao Programa de Pós-Graduação em Engenharia Elétrica da Universidade Federal do Espírito Santo, como requisito para a obtenção do título de Mestre em Engenharia Elétrica - Área de Telecomunicações.

Aprovado em 08 de junho de 2020.

COMISSÃO EXAMINADORA

_____

Prof. Dr. Anilton Salles Garcia
Universidade Federal do Espírito Santo
Orientador

_____

Prof. Dr. Jair Adriano Lima Silva
Universidade Federal do Espírito Santo
Tutor/Membro Titular Interno

_____

Prof. Dr. Helder Roberto de O. Rocha
Universidade Federal do Espírito Santo
Membro Titular Interno

_____

Prof.ª Dr.ª Maria Carmen Viana
Universidade Federal do Espírito Santo
Membro Titular Externo

AGRADECIMENTOS

Gostaria de, primeiramente, agradecer a Deus por essa oportunidade de realizar mais um grande sonho, que é a obtenção do título de mestre.

Um agradecimento especial aos meus pais Nelson e Beth, que me ofereceram apoio incondicional durante todo esse processo.

À Anilton, meu orientador, que me permitiu fazer parte desse incrível trabalho junto à Secretaria de Saúde do Espírito Santo e que me ofereceu todo o apoio necessário.

À Edleusa Cupertino e toda a equipe da Secretaria de Saúde do Espírito Santo, que nos auxiliou com o acesso as informações utilizadas nesse trabalho, além de esclarecimento de eventuais dúvidas.

Aos amigos e colegas de trabalho da $(QI)^3$, que me incentivaram a ir atrás dessa conquista pessoal e também ofereceram a todo momento o apoio necessário.

Um agradecimento também aos professores que aceitaram fazer parte da minha banca de mestrado e separaram um tempo de suas rotinas para poder agregar qualidade a este trabalho, mesmo nesse período difícil de pandemia que estamos vivendo.

RESUMO

Este trabalho tem como objetivo auxiliar a Secretaria de Saúde do Espírito Santo na melhor compreensão da situação da morbimortalidade por causas externas (violência e acidentes), que representa a terceira causa de morte na população em geral e a primeira na população de 1 a 49 anos no Brasil.

Para atingir tal objetivo, foi proposto a construção de um banco de dados, semanticamente estruturado, onde os dados dos registros das vítimas, disponibilizados em fichas textuais sejam organizados com a finalidade de possibilitar o uso de ferramentas para geração de inteligência em relação ao problema da violência. A partir da base de dados, o objetivo é desenvolver um conjunto de ferramentas com o uso de técnicas de Ciência de Dados de modo a possibilitar a identificação, após a integração das informações, das possíveis causas e consequências desses casos de violência e, com base em técnicas de raciocínio inteligente, propor ações e medidas que possam prevenir os casos mais comuns de violência e/ou amenizar suas consequências.

Um banco de dados estruturado foi desenvolvido com base nos registros de violência por causas externas no estado do Espírito Santo de 2011 a 2018. Com essa base de dados, foram desenvolvidos painéis que ajudaram a analisar as cidades com maior ocorrência de incidentes, como as características pessoais da vítima (cor, idade, estado civil, sexo) afetam os casos e como a relação entre vítima e agressor influencia os incidentes. O algoritmo de aprendizado de máquina teve resultados consistentes na previsão dos casos de violência que ocorreram várias vezes. O classificador *Support Vector Machine* (SVM) teve o melhor desempenho, com precisão de 0.7827, recall de 0.7009 e F1 score de 0.7235.

Com este estudo, foi possível ter um conhecimento mais consistente em relação à violência por causas externas no estado do Espírito Santo. O modelo de *machine learning* provou ser uma ótima ferramenta para prever os casos de violência que ocorreram repetidas vezes.

Palavras-chave: Ciência de Dados, Violência, *Analytics*, Modelagem de Banco de Dados, *Machine Learning*

ABSTRACT

This work aims to support the Health's Department of Espírito Santo to understand better the situation of violence from external cause (violence and accidents), which represents the third cause of death in the general population and the first in the population aged from 1 to 49 years in Brazil.

To achieve this goal, it was proposed the creation of a Semantic data model where the data of the violence victims, made available in textual records, are organized with the purpose of enabling the use of tools to generate intelligence related to the problem of violence. From the database, the objective is to develop a set of tools using Data Science techniques in order to identify, after integration of the information, the possible causes and consequences of these cases of violence and, based on intelligent reasoning techniques, propose actions and measures that can prevent the most common cases of violence and / or mitigate its consequences.

A structured database was developed using the records of violence from external causes in the state of Espírito Santo from 2011 to 2018. With this database it was developed dashboards that helped to analyze the cities with the highest occurrence of incidents, how personal characteristics of the victim (color, age, marital status, gender) affects the cases and how the relationship between victim and aggressor influences the incidents. The machine-learning algorithm had consistent results in predicting the cases of violence that happened several times. The Support Vector Machine (SVM) classifier had the best performance, with precision 0.7827, recall 0.7009 and F1 score 0.7235.

With this study, it was possible to have a more consistent knowledge related to the violence from external causes in the state of Espírito Santo. The machine-learning model proved to be a great tool to predict the cases of violence that happened several times.

Keywords: Data Science, Violence, Analytics, Database Design, Machine Learning

LISTA DE ABREVIATURAS

MS – Ministério da Saúde

SUS – Sistema Único de Saúde

OMS – Organização Mundial da Saúde

IDH – Índice de Desenvolvimento Humano

VIVA – Sistema de Vigilância de Acidentes e Violências

SINAN – Sistema de Informações sobre Agravos de Notificação

SDM – *Semantic Database Model*

DBMS – *Database Management System*

SQL – Structured Query Language

I/O – Input/Output

VPI – Violência por Parceiro Íntimo

NLP – *Natural Language Processing*

ES – Espírito Santo

SVM – *Support Vector Machine*

UF – Unidade da Federação

KNN – *K-Nearest Neighbor*

CBR – *Case-Based Reasoning*

SESA – Secretaria de Saúde do Espírito Santo

ETL – *Extract, Transform, Load*

IPEA – Instituto de Pesquisa Aplicad

SUMÁRIO

1 INTRODUÇÃO **10**

1.1 CONTEXTUALIZAÇÃO.....10

1.2 JUSTIFICATIVA 13

1.3 OBJETIVOS.....15

1.3.1 Objetivos Gerais.....15

1.3.2.Objetivos Específicos.....15

1.4 TRABALHOS RELACIONADOS.....15

1.5 METOLOGIA.....17

1.6 RESULTADOS E CONTRIBUIÇÕES.....17

1.7 ORGANIZAÇÃO DA DISSERTAÇÃO.....18

2 REFERENCIAL TEÓRICO 20

2.1 MODELAGEM CONCEITUAL BASEADA EM ONTOLOGIA.....20

2.2 BANCO DE DADOS 22

2.3 ONTOLOGIA X BANCO DE DADOS 23

2.4 MODELAGEM DE DADOS COM ESTRUTURAÇÃO SEMÂNTICA.....24

2.5 CIÊNCIA DE DADOS.....26

2.6 PROCESSAMENTO DE LINGUAGEM NATURAL 27

2.7 CLASSIFICADORES DE *MACHINE LEARNING*.....29

3 CARACTERIZAÇÃO DO PROBLEMA E MÉTODOS DE SOLUÇÃO.....**33**

3.1 PROBLEMA EM QUESTÃO.....33

3.2 MÉTODOS DE SOLUÇÃO.....33

3.2.1 Modelagem conceitual baseada em ontologia.....33

3.2.2 Modelagem de dados.....34

3.2.3 Uso de NLP para extração de informações do campo de 'Observações Adicionais'.....37

3.2.4 Uso de *machine learning* para prever casos de violência que ocorreram múltiplas vezes ...39

4 RESULTADOS OBTIDOS.....**41**

4.1 ONTOLOGIA DE DOMÍNIO.....41
4.2 DASHBOARDS.....43
4.3 ALGORITMO DE MACHINE LEARNING.....47
5 CONCLUSÃO E TRABALHOS FUTUROS.....50
REFERÊNCIAS.....52
ANEXOS.....54

1 INTRODUÇÃO

1.1 CONTEXTUALIZAÇÃO

No Brasil e no mundo, o impacto da morbimortalidade por causas externas (violências e acidentes) constitui uma das maiores preocupações para chefes de Estado e dirigentes do setor Saúde. No Brasil, as violências e os acidentes representam a terceira causa de morte na população geral e a primeira na população de 1 a 49 anos (BRASIL, 2016).

Dados do Ministério da Saúde (MS) registraram, no período de 2000 a 2013, um total de 1.874.508 óbitos por causas externas. Entre o início e o fim desse período, houve um aumento de 28,1%, passando de 118.397 óbitos por causas externas, em 2000, para 151.683, em 2013. Em 2013, as causas externas representaram 12,5% do total de óbitos no País. Segundo o MS, em 2013, nos hospitais que integram o Sistema Único de Saúde (SUS), ocorreram 1.056.372 internações por causas externas, perfazendo 9,5% do total de internações, enquanto em 2014, esse percentual foi de 9,9% (BRASIL, 2016).

As Figuras 1 e 2 mostram a evolução ao longo dos anos (de 2000 a 2018) das quatro principais causas de óbito no Brasil: Doenças do aparelho circulatório, Neoplasias (tumores), Doenças do aparelho respiratório e Causas externas morbidade e mortalidade. Para a Figura 2 o Universo de Análise é reduzido para as vítimas de 1 a 49 anos. Enquanto na Figura 1 é possível notar que as causas externas oscilam entre a 3ª e 4ª causa de óbitos ao longo dos anos, na Figura 2 as causas externas são, disparadamente, a maior causa de óbitos em todo o período registrado. Esses dados foram retirados do DATASUS e o *software Tableau Desktop* foi utilizado para realizar a plotagem dos gráficos.

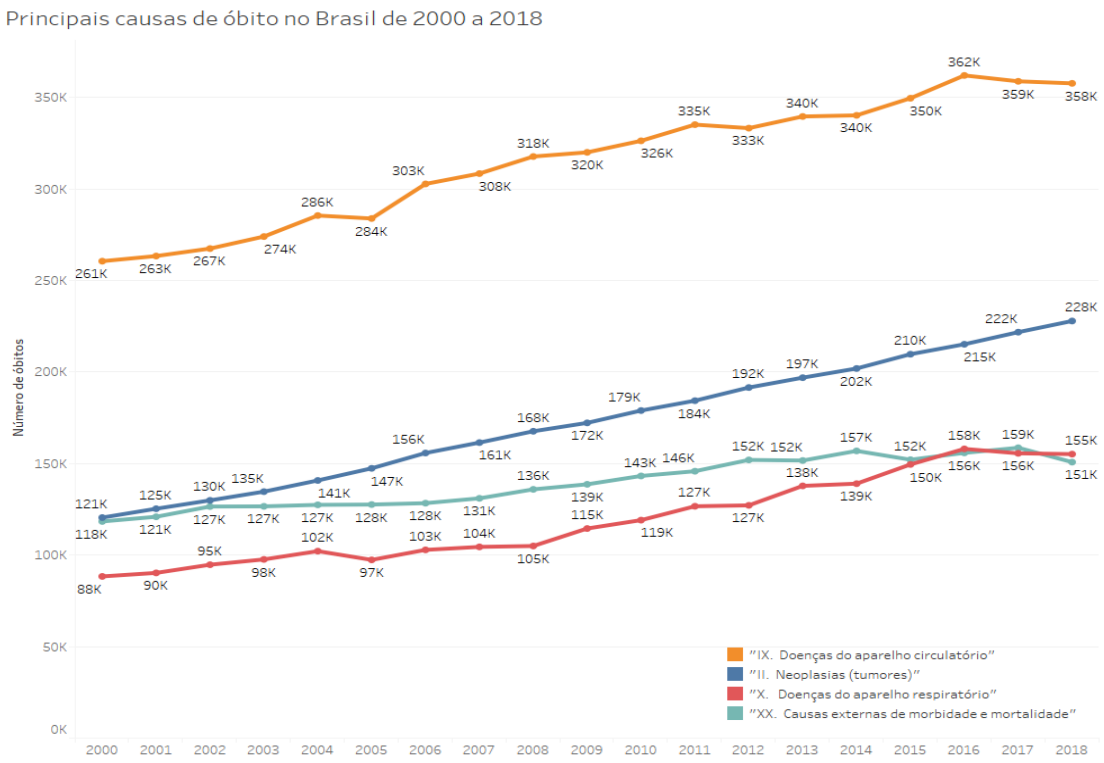


Figura 1: Principais causas de óbito no Brasil de 2000 a 2018

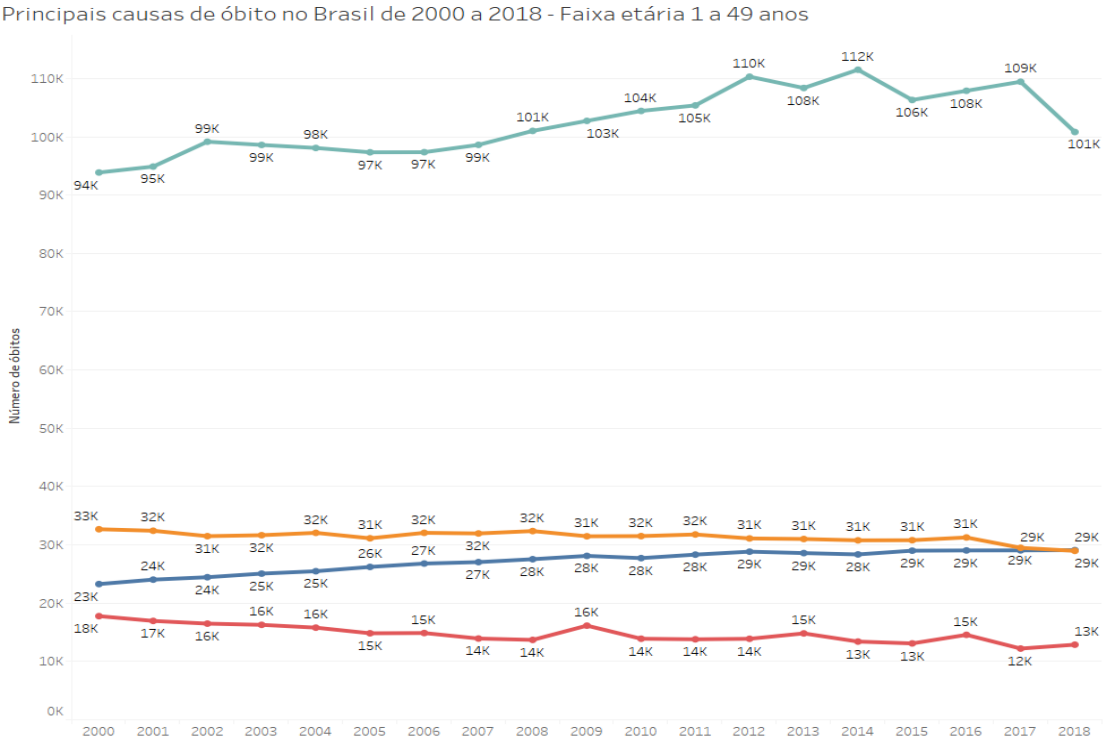
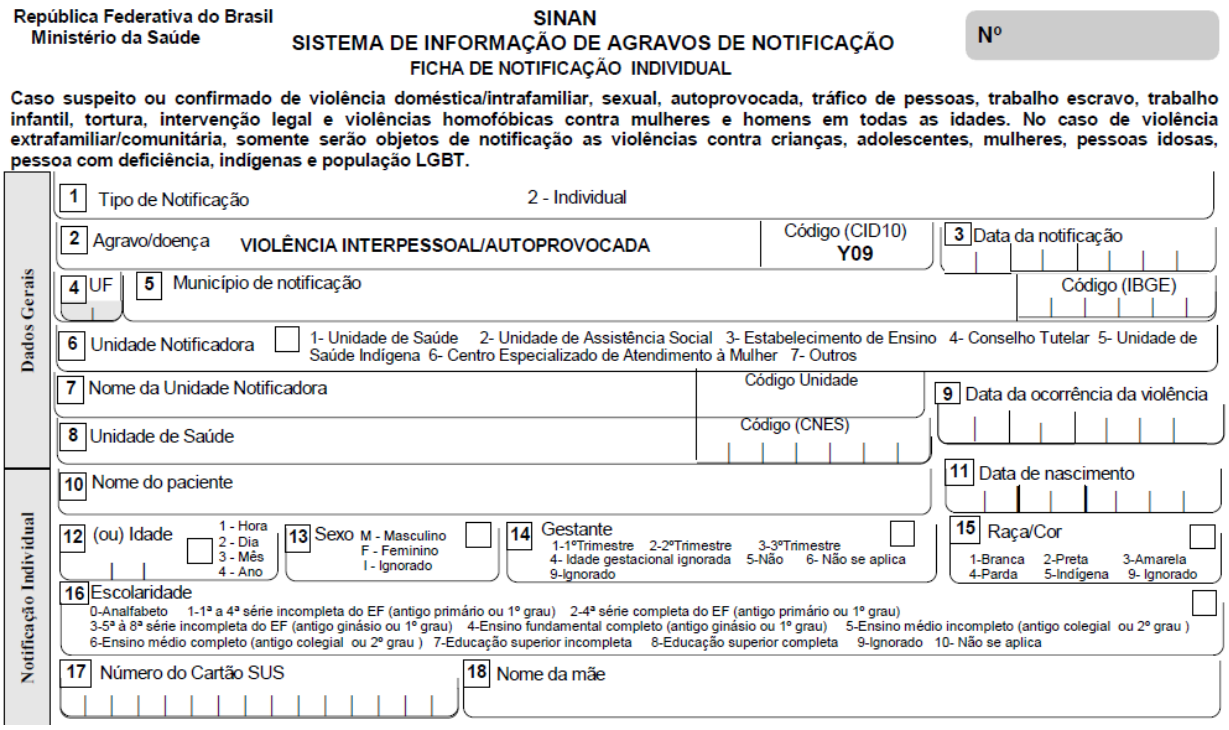


Figura 2: Principais causas de óbito no Brasil de 2000 a 2018 de vítimas de 1 a 49 anos

Preocupado com esses dados alarmantes, o Ministério da Saúde lançou em 2001 a Política Nacional de Redução da Morbimortalidade por Acidentes e Violências. Faz parte dessa política a Ficha de Notificação de Violência Interpessoal e Autoprovocada (vide Figura 3), que é a ficha de notificação compulsória de todos os casos suspeitos e ou confirmados de violência atendidos em todos os serviços de saúde do país, de forma a padronizar a coleta de dados, a partir de um conjunto de variáveis e categorias, que retratam as violências perpetradas contra grupos populacionais. Essa ficha é preenchida nos serviços de saúde, quando há suspeita e/ou confirmação de ocorrência da violência, tanto interpessoais quanto autoprovocadas. Depois de preenchida, as informações de cada uma dessas fichas são digitalizadas pelos agentes de saúde.


República Federativa do Brasil
Ministério da Saúde

SINAN
SISTEMA DE INFORMAÇÃO DE AGRAVOS DE NOTIFICAÇÃO
FICHA DE NOTIFICAÇÃO INDIVIDUAL

Nº

Caso suspeito ou confirmado de violência doméstica/intrafamiliar, sexual, autoprovocada, tráfico de pessoas, trabalho escravo, trabalho infantil, tortura, intervenção legal e violências homofóbicas contra mulheres e homens em todas as idades. No caso de violência extrafamiliar/comunitária, somente serão objetos de notificação as violências contra crianças, adolescentes, mulheres, pessoas idosas, pessoa com deficiência, indígenas e população LGBT.

Dados Gerais

1 Tipo de Notificação 2 - Individual

2 Agravo/doença VIOLÊNCIA INTERPESSOAL/AUTOPROVOCADA Código (CID10) Y09

3 Data da notificação

4 UF

5 Município de notificação Código (IBGE)

6 Unidade Notificadora 1- Unidade de Saúde 2- Unidade de Assistência Social 3- Estabelecimento de Ensino 4- Conselho Tutelar 5- Unidade de Saúde Indígena 6- Centro Especializado de Atendimento à Mulher 7- Outros

7 Nome da Unidade Notificadora Código Unidade

8 Unidade de Saúde Código (CNES)

9 Data da ocorrência da violência

Notificação Individual

10 Nome do paciente

11 Data de nascimento

12 (ou) Idade 1 - Hora 2 - Dia 3 - Mês 4 - Ano

13 Sexo M - Masculino F - Feminino I - Ignorado

14 Gestante 1-1ºTrimestre 2-2ºTrimestre 3-3ºTrimestre 4- Idade gestacional ignorada 5-Não 6- Não se aplica 9-Ignorado

15 Raça/Cor 1-Branca 2-Preta 3-Amarela 4-Parda 5-Indígena 9- Ignorado

16 Escolaridade 0-Analfabeto 1-1ª a 4ª série incompleta do EF (antigo primário ou 1º grau) 2-4ª série completa do EF (antigo primário ou 1º grau) 3-5ª à 8ª série incompleta do EF (antigo ginásio ou 1º grau) 4-Ensino fundamental completo (antigo ginásio ou 1º grau) 5-Ensino médio incompleto (antigo colegial ou 2º grau) 6-Ensino médio completo (antigo colegial ou 2º grau) 7-Educação superior incompleta 8-Educação superior completa 9-Ignorado 10- Não se aplica

17 Número do Cartão SUS

18 Nome da mãe

Figura 3: Primeira parte da Ficha de Notificação de Violência Interpessoal

A ficha de notificação é estruturada em dez blocos: Dados Gerais, Notificação Individual, Dados de Residência, Dados da Pessoa Atendida, Dados da Ocorrência, Violência, Violência Sexual, Dados do Provável Autor da Violência, Encaminhamento e Dados Finais, bem como campos para anotação de informações complementares. Essa notificação deve ser preenchida pelo

profissional de saúde em duas vias, sendo uma delas enviada à vigilância epidemiológica municipal. Às Secretarias Municipais de Saúde, compete a digitação, consolidação e análise dos dados, disseminação das informações e envio da base de dados para as Secretarias de Estado de Saúde, e posteriormente para o Ministério da Saúde (BRASIL, 2016).

Apesar de haver instruções específicas para o preenchimento dessa ficha, muitas vezes esse preenchimento não é feito de maneira adequada. Existe a possibilidade de a vítima preencher informações antagônicas em campos que se complementam ou mesmo inserir informações importantes para o entendimento do tipo de violência em um campo de 'Observações Adicionais', que pode ficar esquecido na hora de analisar os dados.

Com este trabalho busca-se desenvolver uma modelagem conceitual das informações e estruturar semanticamente a base de dados, a partir das fichas e documentos de referência, para que seja possível integrar essas informações preenchidas nas fichas de notificação em meios digitais e alimentar os bancos secundários, de forma a auxiliar a análise dos gestores e, por consequência, às vítimas de violências.

1.2 JUSTIFICATIVA

O elevado índice de violência faz com que o Brasil seja considerado um dos países mais perigosos do mundo. Isso traz inúmeros problemas para o país, incluindo altas taxas de homicídio, evasão escolar, superlotação da população carcerária, baixo aproveitamento da capacidade turística, além de muitas outras desvantagens.

A Figura 4 mostra o 'Atlas de violência' do Brasil em 2017, produzido pelo Instituto de Pesquisa Aplicada (IPEA), que destrincha o cenário da violência ao longo do território brasileiro. Um dado que se destaca neste infográfico é a elevada proporção de mortes violentas não esclarecidas em relação ao total de mortes violentas. Altas proporções de mortes violentas não esclarecidas quando comparadas às mortes por causas externas indicam problemas na qualidade do sistema de informação da saúde (CERQUEIRA et al., 2016).

Figura 4: Atlas da violência no Brasil no ano de 2017 (CERQUEIRA et al., 2016)

Enquanto, em países com alto Índice de Desenvolvimento Humano (IDH), as mortes violentas indeterminadas representam um resíduo inferior a 1% do total de mortes violentas, em alguns estados do Brasil esse número foi superior a 10%. Isto ocorre, pois nesses países se reconhece a importância de se descobrir as causas que levaram o indivíduo ao óbito, como elemento fundamental para evitar mortes futuras (CERQUEIRA et al., 2016).

Esse conjunto de fatores ajuda a reforçar o fato que se deve fazer um estudo mais aprofundado do cenário da violência, identificando suas possíveis causas de modo a tentar agir de maneira preventiva para evitar possíveis novos casos.

1.3 OBJETIVOS

1.3.1 Objetivo Geral

O objetivo geral deste trabalho é propor um modelo conceitual baseado em ontologias para a gestão da informação e do conhecimento na área de vigilância epidemiológica, com foco inicial nas questões relativas à violência.

1.3.2 Objetivos Específicos

Para atingir o objetivo geral, os seguintes objetivos específicos foram delineados:

I. Estruturar semanticamente o banco de dados existente, modelado com base na ontologia definida para o domínio;
II. Uma vez estruturado o banco, apontar possíveis incorreções nos registros, buscando excelência de informação;
III. Utilizar esse banco de dados semanticamente estruturado para o desenvolvimento de diversos painéis (*dashboards*), atualizados à medida em que os novos dados são introduzidos, de forma a auxiliar na melhor compreensão dos cenários de violência presentes nas Fichas de Notificação;
IV. Auxiliar no processo de localização e exploração de novos conhecimentos que o Sistema de Saúde, apesar de possuir, não consegue identificar claramente, facilitando na melhor compressão dos perfis das vítimas de violência e, dessa forma, embasando o processo de tomada de decisão;
V. Desenvolver ferramentas de Ciência de Dados com o objetivo de apontar a possibilidade de novos casos mais graves de violência (ex: tentativa de suicídio ou feminicídio) baseado em informações já disponíveis nos bancos de dados referentes às vítimas ou aos agressores;

1.4 TRABALHOS RELACIONADOS

Em 2018 Corassa, em sua dissertação de mestrado, descreveu o perfil das violências contra mulheres perpetradas por parceiro íntimo e identificou fatores de risco para o homicídio entre mulheres vítimas de violência por parceiro íntimo (VPI) notificada ao Sistema de Vigilância de Violências e Acidentes (Viva-contínuo/Sinan), no Brasil, período de 2011 a 2015.

Com esse trabalho, foi possível identificar que existem diferenciais importantes no perfil da VPI notificada em relação à idade, que precisam ser considerados na elaboração de intervenções. Mulheres negras, com baixa escolaridade, separadas ou divorciadas, vivendo em zonas rurais e vítimas de violência durante a gestação apresentaram maior risco de serem vítimas de homicídio (CORASSA, 2018).

Em 2019, Modesto e outros pesquisadores da UFBA, fizeram uma revisão da literatura dos fatores que influenciam na mortalidade de jovens por causas externas no Brasil. Com o estudo demonstrou que as vítimas mais ocorrentes são: homens, pretos, com baixos níveis socioeconômicos e escolares e que possuem o local do acidente como principal agravante. Os fatores que influenciam nos óbitos envolvem questões sociopolíticas, que são reflexos da violência estrutural somada à violência institucional (Modesto et al., 2019).

Em 2018, Stone e outros pesquisadores americanos fizeram uma análise dos casos de suicídio nos estados americanos do período de 1999 a 2016. Nesse estudo, os pesquisadores obtiveram como resultados que, no período analisado, as taxas de suicídio aumentaram significativamente em 44 estados, com 25 estados apresentando aumentos superiores a 30%. As taxas aumentaram significativamente entre homens e mulheres em 34 e 43 estados, respectivamente. Os falecidos por suicídio sem condições conhecidas de saúde foram comparados com aqueles com condições conhecidas de saúde mental em 27 estados. Enquanto os falecidos eram predominantemente do sexo masculino e brancos não hispânicos, aqueles sem condições de saúde mental conhecidas, em comparação com aqueles com problemas de saúde mental, eram mais propensos a ser do sexo masculino e pertencente a uma minoria racial / étnica. Os falecidos por suicídio sem condições conhecidas de saúde mental também obtiveram chances significativamente maiores de perpetrar homicídio seguido de suicídio. Entre os falecidos com idade ≥ 18 anos, 20,1% daqueles sem condições de saúde mental conhecidas e 15,3% daqueles com problemas de saúde mental já haviam servido anteriormente nas forças armadas dos EUA ou estavam servindo no momento da morte (Stone et al., 2018).

Esse conjunto de estudos demonstra a importância e a relevância do tema da violência no cenário nacional e internacional e que ainda há muito a ser compreendido. Com este trabalho espera-se obter uma compreensão melhor do cenário da violência no estado do Espírito Santo e também entender melhor quais os principais fatores que estão relacionados com o considerável aumento no número de casos de violência nos últimos anos.

1.5 METODOLOGIA

O trabalho desenvolvido é uma pesquisa descritiva utilizando dados secundários. Dessa forma, um estudo minucioso do tema foi realizado e os dados fornecidos pela SESA foram utilizados com o intuito de analisar mais a fundo a situação da violência no Espírito Santo. As etapas de desenvolvimento deste trabalho estão divididas da seguinte forma:

I. Pesquisa bibliográfica sobre o tema incluindo-se nessa etapa as bases de divulgação científica, a documentação existente na Organização Mundial da Saúde (OMS) e no MS;
II. Modelagem Conceitual baseada em ontologias das informações referentes aos diversos tipos de notificações;
III. Utilização dos dados do VIVA/SINAN para realizar a estruturação de uma base de dados semanticamente estruturada com base na ontologia definida para o domínio.
IV. Desenvolvimento de painéis (*dashboards*) que possam ajudar a compreender melhor o cenário de violência aprendido atráves nas Fichas de Notificação;
V. Utilização de técnicas de *Data Science/Machine Learning* para identificar correlações entre os dados e tentar traçar perfis de possíveis novos casos de agressão baseado nas informações já disponíveis das Fichas de Notificação.

1.6 RESULTADOS E CONTRIBUIÇÕES

Contribuições significativas foram obtidas com este trabalho. A primeira delas, e talvez a mais importante, foi que com este projeto foi possível ter uma visão mais ampla e completa da situação da violência no estado do Espírito Santo. Anteriormente já era feito um trabalho de

exploração e visualização dos dados presentes nas fichas de notificação pela Secretaria de Saúde do Espírito Santo (SESA), mas esse trabalho era realizado de forma limitada, sem explorar todos os recursos presentes na ficha de notificação. Informações importantes como por exemplo, dados geográficos, correlações entre diferentes dados da ficha e utilização das informações adicionais oferecidas pelas vítimas de violência que não se encaixavam nos demais campos da ficha de notificação, mas que possuem informações crucias para o entendimento de alguns casos, foram utilizados para ter uma compreensão dos motivos e circunstâncias por trás dos casos de violência no estado do Espírito Santo.

Para obter tais resultados, primeiramente foi desenvolvida uma ontologia de domínio, descrita nas seções 3.2.1 e 4.1 deste trabalho, que permitiu categorizar e definir os relacionamentos dos diversos campos presentes na ficha de notificação em um modelo único, sem a necessidade de um artefato adicional (dicionário de dados) para a consulta das características e relacionamentos entre cada um dos campos.

Em seguida, uma base de dados estruturada (seção 3.2.2) foi desenvolvida para permitir que os profissionais de saúde possam fazer suas próprias análises de maneira fácil e intuitiva, por meio de ações de arrastar e soltar (*drag and drop)* e sem a necessidade de conhecimentos profundos de linguagens de programação. Essa mesma fonte de dados foi utilizada para a criação dos *dashboards* descritos na seção 4.2 deste projeto, que permitiram a observação de maneira fácil e visual das notificações descritas nas fichas.

Por fim, um modelo de *Machine Learning* foi desenvolvido com o intuito de tentar prever os casos de violências que ocorreram múltiplas vezes e que não foi possível obter essa informação por meio da ficha de notificação. Esse modelo está descrito nas seções 3.2.4 e 4.2 deste trabalho, o qual obteve resultados próximos a 70% de acerto na previsão para alguns classificadores utilizados.

1.7 ORGANIZAÇÃO DA DISSERTAÇÃO

A organização deste trabalho foi feita da seguinte forma: no capítulo 2 é detalhado todo o referencial teórico da pesquisa que foi utilizada como base para a construção deste trabalho.

No capítulo 3 é feita a caracterização do problema que este trabalho se propõe a resolver e os métodos de solução que foram utilizados com o objetivo de solucionar o problema em questão.

No capítulo 4 são apresentados os resultados obtidos com este trabalho, que incluem a ontologia de domínio, os *dashboards* que foram desenvolvidos para melhor compreensão do cenário da violência e o modelo de *Machine Learning*.

Por fim, no capítulo 5, são apresentadas as considerações finais do trabalho e os possíveis trabalhos futuros que podem ser desenvolvidos para aprimorar ainda mais a qualidade dos resultados já obtidos com esse trabalho.

2 REFERENCIAL TEÓRICO

2.1 MODELAGEM CONCEITUAL BASEADA EM ONTOLOGIA

“ Ontologia é uma especificação explícita da conceituação. O termo é emprestado da filosofia, onde Ontologia é um relato sistemático da Existência ” (Tom Gruber e o Laboratório de Sistemas de Conhecimento da Universidade de Stanford).

Na ciência da computação, a ontologia é uma representação formal do conhecimento por um conjunto de conceitos dentro de um domínio e os relacionamentos entre esses conceitos. É usado para raciocinar sobre as propriedades desse domínio e pode ser usado para descrever o domínio. Em teoria, ontologia é uma “especificação formal e explícita de uma conceituação compartilhada ”. A ontologia fornece um vocabulário compartilhado, o qual pode ser usado para modelar um domínio que compreende os tipos de objetos e / ou conceitos que existem, e suas propriedades e relações (MAN, 2013).

Ontologias são usadas em Inteligência Artificial, Web Semântica, Engenharia de Sistemas, Engenharia de Software, Informática Biomédica, Biblioteconomia, *Bookmarking* Corporativo e Arquitetura da Informação como forma de conhecimento da representação sobre o mundo ou parte dele. A criação de ontologias de domínio também é fundamental para a definição e uso de um *framework* de arquitetura corporativa. Existem quatro categorias de ontologia: estática, dinâmica, intencional e social. A ontologia estática descreve as coisas que existem, seus atributos e relacionamento. A ontologia dinâmica descreve o mundo em termos de estados, transições e processos de estado. A ontologia intencional abrange o mundo dos agentes, as coisas que se acredita e que se quer provar ou refutar e argumentar sobre. A ontologia social abrange contextos sociais, estruturas organizacionais permanentes ou mudança de redes de alianças e independências (GRUBER, 1993).

Um conjunto preliminar de critérios de design para ontologias cujo objetivo é compartilhar e interoperar conhecimento entre programas baseados em uma conceituação compartilhada são:

a) Clareza. Uma ontologia deve comunicar o significado proposto de termos definidos. As definições devem ser objetivas e independentes do contexto social e computacional. Uma definição completa (um predicado definido pelas condições necessárias e suficientes) é preferida a uma definição parcial (definida apenas pelas condições necessárias ou suficientes). Todas as definições devem ser documentadas em linguagem natural.

b) Coerência. A ontologia deve permitir inferências que sejam consistentes com as definições. No mínimo, os axiomas definidores devem ser logicamente consistentes. A coerência também deve se aplicar aos conceitos definidos informalmente, como os descritos na documentação e nos exemplos da linguagem natural. Se uma frase que pode ser inferida a partir dos axiomas contradiz uma definição ou exemplo dado informalmente, a ontologia é incoerente.

c) Extensibilidade. A ontologia deve ser projetada para antecipar os usos do vocabulário compartilhado. Ele deve oferecer uma base conceitual para uma série de tarefas previstas e a representação deve ser criada para que se possa estender e especializar a ontologia monotonicamente. Em outras palavras, deve-se definir novos termos para usos especiais, com base no vocabulário existente, de uma maneira que não exija a revisão das definições existentes.

d) Viés mínimo de codificação. A conceitualização deve ser especificada no nível do conhecimento sem depender de uma codificação específica no nível do símbolo. Um viés de codificação ocorre quando as escolhas de representação são feitas exclusivamente para a conveniência da notação ou implementação. O viés de codificação deve ser minimizado porque os agentes de compartilhamento de conhecimento podem ser implementados em diferentes sistemas de representação e estilos de representação.

e) Compromisso Ontológico mínimo. A ontologia deve exigir o mínimo comprometimento ontológico suficiente para apoiar as atividades de compartilhamento de conhecimento pretendidas. A ontologia deve fazer o mínimo de afirmações possíveis sobre o mundo que está sendo modelado, permitindo que as partes comprometidas com a liberdade da ontologia se especializem e instanciem a ontologia conforme necessário (GRUBER, 1995).

2.2 BANCO DE DADOS

Um banco de dados é basicamente um sistema informatizado de manutenção de registros. Em outras palavras, é um sistema informatizado o qual o propósito geral é armazenar dados e permitir aos usuários a recuperação e atualização desses dados sob demanda. A informação em questão pode ser qualquer coisa que seja de importância para o indivíduo ou organização interessada, que auxilie no processo geral de execução do negócio daquele indivíduo ou organização.

A Figura 5 mostra um esquema simplificado de um banco de dados. Esse tipo de sistema pode ser resumido em quatro componentes principais: dado, *hardware*, *software* e usuários. É importante diferenciar os termos "dados" e "informação". Dado se refere a aquilo que é de fato armazenado no banco de dados e informação se refere ao significado que aquele dado possui para um usuário ou negócio (DATE, 2004).

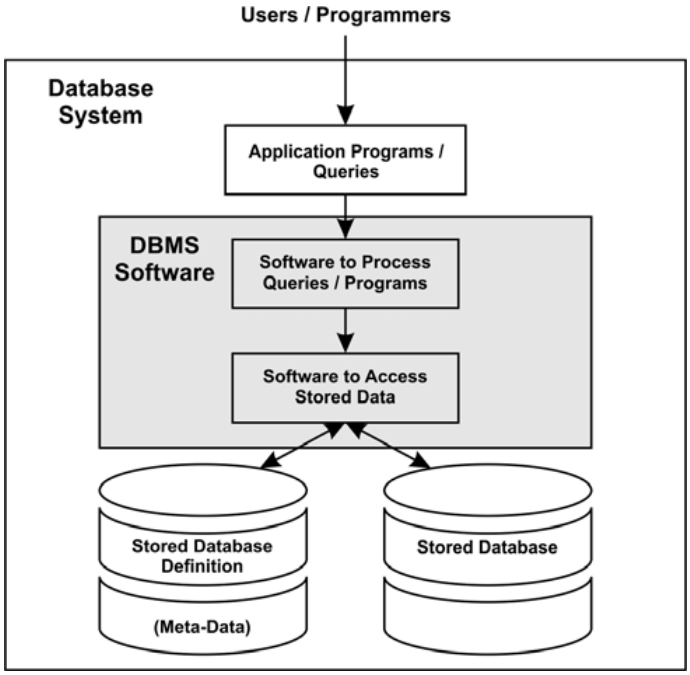


Figura 5: Esquema simplificado de um banco de dados tradicional (KRSTIC, 2018)

Os componentes de *hardware* de um sistema de banco de dados consistem de: o processador de *hardware* e a memória principal associada que são utilizados como suporte para a execução do *software* do banco de dados e um volume secundário de armazenamento (geralmente

um disco magnético), que é usado para manter os dados armazenados, em conjunto com os dispositivos de I/O (*Input/Output),* controladores e outros dispositivos.

Entre o banco de dados fisicamente armazenado e os usuários do sistema, está uma camada de *software* conhecida por nomes como: gerenciamento de banco de dados, servidor de banco de dados ou mais comumente de Sistema de Gerenciamento de Banco de Dados *(Database Management System* – DBMS). Todas as requisições de acesso ao banco de dados são controladas pelo DBMS, como por exemplo: adicionar ou remover arquivos (ou tabelas), recuperar e atualizar dados de tabelas, além de outras funcionalidades.

Por fim, os programadores são os usuários responsáveis pela escrita dos programas em alguma linguagem de programação suportada pelo DBMS (C, C++, Java ou alguma outra linguagem de alto nível) para a estruturação do *software*. Já os usuários finais acessarão as informações necessárias que estão armazenadas no banco de dados por meio de instruções apropriadas, geralmente por Linguagem de Consulta Estruturada (*Structured Query Language* – SQL) (DATE, 2004).

2.3 ONTOLOGIA X BANCO DE DADOS

Apesar de Ontologia e Banco de Dados serem formas de representação de conhecimento, existem diversas diferenças na maneira que cada uma delas realiza essa representação. Para começar, o foco principal de um esquema de Banco de Dados é o dado em si. Toda a estrutura de um banco de dados é pensada para o armazenamento e consulta de dados de maneira eficiente. Já em uma Ontologia, o foco é o significado que aquele determinado dado ou informação possui. A estrutura da Ontologia é elaborada para comunicação humana, interoperabilidade e pesquisa.

Uma outra importante diferença na estrutura é que em um Banco de Dados, o dado em si, é armazenado em forma de registros, em que cada linha de uma tabela representa um registro armazenado nesse banco. Já na Ontologia, a estrutura é baseada na hierarquia '*is-a*' (ou 'é um' em português'), onde o relacionamento entre cada uma das classes forma o *backbone* dessa estrutura. Também há diferenças nas formas de relacionamento entre cada uma delas. Enquanto no Banco de Dados definem-se tabelas, colunas e chaves para realizar o armazenamento de registros e o

relacionamento entre tabelas, na Ontologia definem-se classes, indivíduos e relacionamentos entre as classes.

Por fim, um Banco de Dados necessita de um artefato separado, geralmente um dicionário de dados, para explicar e exemplificar como é realizado o relacionamento entre cada uma das colunas do banco, as características de cada uma dessas colunas e também o tipo de dado (número, *string*, dado geográfico e etc.) que aquele campo possui. Na ontologia esse tipo de documento não é necessário, pois todas essas informações já estão representadas e contidas na própria estrutura ontológica, dispensando assim um documento separado para detalhar essas especificações (USCHOLD, 2011).

2.4 MODELAGEM DE DADOS COM ESTRUTURAÇÃO SEMÂNTICA

A Modelagem de Dados com Estruturação Semântica (*Semantic Database Model* – SDM) foi projetada com vários tipos específicos de usos em mente. Inicialmente, o SDM deve servir como um mecanismo formal de especificação para descrever o significado de um banco de dados. Um esquema SDM fornece documentação precisa e um meio de comunicação para usuários de banco de dados. Em particular, um novo usuário de um grande banco de dados complexo deve encontrar seu esquema de uso do SDM para determinar quais informações estão contidas na base de dados.

Segundo, o SDM fornece a base para uma variedade de interfaces de usuário baseadas em semântica de alto nível para um banco de dados. Esses recursos de interface podem ser construídos como *front-ends* para os Sistemas de Gerenciamento de Banco de Dados *(Database Management System* – DBMS) existentes ou como a linguagem de consulta de um novo sistema de Gerenciamento de Banco de Dados. Essas interfaces aprimoram o processo de identificação e recuperação de informações relevantes do banco de dados (HAMMER; MC LEOD, 1981).

SDM foi projetado para atender a vários critérios que não são atendidos pelos modelos de banco de dados contemporâneos, mas que se acredita serem essenciais na descrição eficaz da base de dados e formalismo de estruturação. Esses critérios são:

a) As construções de um modelo de banco de dados devem fornecer a especificação explícita de uma grande parte do significado de uma base de dados. Muitos bancos de dados contemporâneos exibem compromissos entre o desejo de fornecer uma organização de um conjunto de dados orientada ao usuário e a necessidade de oferecer suporte a recursos eficientes de armazenamento e manipulação de banco de dados. No entanto, a expressividade semântica dos modelos hierárquico, de rede e relacional é limitada; eles não fornecem mecanismo suficiente para permitir que um esquema de banco de dados descreva o significado de uma base de dados. Ao fazer isso, eles inevitavelmente perdem informações sobre os dados; eles fornecem a expressão de apenas uma faixa limitada do conhecimento do designer sobre o ambiente do aplicativo. Isso é uma consequência do fato de que suas estruturas são essencialmente todas orientadas a registros. É necessário romper com a tradição da modelagem baseada em registros e basear um modelo de banco de dados em construções estruturais altamente orientadas ao usuário e expressivas do ambiente de aplicações.

b) Um modelo de banco de dados deve suportar uma visão relativista do significado de uma base de dados e permitir que sua estrutura suporte formas alternativas de visualizar as mesmas informações. Para acomodar várias visualizações dos mesmos dados e permitir a evolução de novas perspectivas sobre os dados, um modelo de banco de dados deve suportar esquemas flexíveis, potencialmente logicamente redundantes e integrados. A flexibilidade é essencial para permitir múltiplas e iguais visualizações dos dados.

c) Um modelo de banco de dados deve suportar a definição de esquemas baseados em entidades abstratas. Especificamente, isso significa que um modelo de banco de dados deve facilitar a descrição de entidades relevantes no ambiente de aplicação, coleções de tais entidades, relacionamentos (associações) entre entidades e interconexões estruturais entre as coleções. Além disso, as próprias entidades devem ser distinguir-se de seus identificadores sintáticos (nomes); a visualização no nível de usuário de um banco de dados deve ser baseada em entidades reais, e não em nomes de entidades artificiais (HAMMER; MC LEOD, 1981).

2.5 CIÊNCIA DE DADOS

O termo "ciência" implica conhecimento adquirido através de estudo sistemático. Em uma definição, é uma organização sistemática que constrói e organiza conhecimento sob a forma de explicações testáveis e previsões. A ciência de dados pode, portanto, implicar um enfoque envolvendo dados e, por extensão, estatísticas, ou o estudo sistemático da organização, propriedades, e análise de dados e seu papel na inferência, incluindo nossa confiança na inferência. Por qual motivo então precisamos de um novo termo como ciência de dados quando tivemos estatística por séculos? O fato de termos agora grandes quantidades de dados não devem por si só justificar a necessidade de um novo termo.

A resposta curta é que a ciência de dados é diferente de estatística e outras disciplinas existentes em várias maneiras importantes. Para começar, a matéria-prima, a parte “dados” da ciência de dados, está cada vez mais heterogênea e não estruturada (texto, imagens, vídeo) geralmente emanando de redes com relacionamentos complexos entre suas entidades (DAHR, 2013).

Cientista de Dados é um profissional que usa de métodos científicos para expor e criar significado e análise a partir de dados brutos. Estatística significa a prática ou ciência de coletar e analisar dados numéricos em grandes quantidades. Para um estatístico, essa definição de estatística já parece abranger qualquer coisa que a definição de cientista de dados possa abranger, mas a definição de estatístico parece limitada, uma vez que muito trabalho estatístico é explicitamente sobre inferências para ser feito a partir de amostras muito pequenas (DONOHO, 2017).

A Figura 6 mostra o crescimento do volume de dados estruturados e não estruturados desde a década de setenta até os dias de hoje. Nota-se que os dados não estruturados apresentaram o crescimento bem mais elevado do que os estruturados, representando atualmente mais de 80% de todo o volume de dados existente.

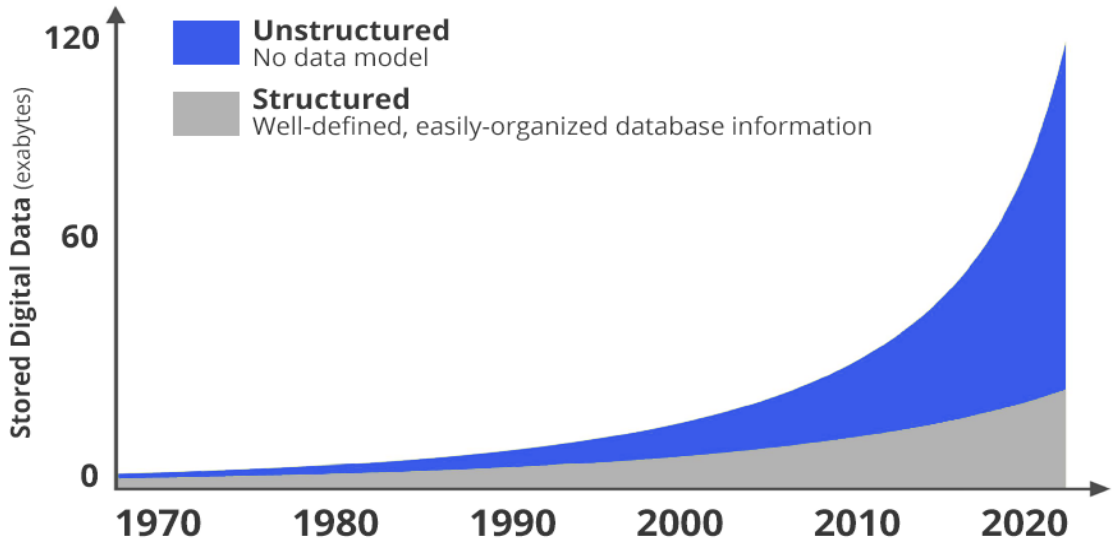


Figura 6: Comparação entre o crescimento do volume de dados estruturados e não estruturados no mundo (DURANT, 2020)

Análises, incluindo a combinação dos dois tipos de dados, requer integração, interpretação e sentido que é cada vez mais derivado através de ferramentas de ciência da computação, linguística, econometria, sociologia e outras disciplinas. A proliferação de linguagens e *tags* de marcação é projetada para permitir que os computadores interpretem os dados automaticamente, tornando-os agentes ativos do processo de tomada de decisão.

Diferentemente das linguagens de marcação iniciais (como HTML) que enfatizava a exibição de informações para consumo humano, a maioria dos dados gerados por seres humanos e computadores hoje são para consumo por computadores; isto é, computadores cada vez mais fazem o trabalho em segundo plano para eles mesmos tomarem decisões automaticamente. Essa escalabilidade na tomada de decisão se tornou possível porque o *big data* serve como matéria-prima para a criação de novos conhecimentos (DAHR, 2013).

2.6 PROCESSAMENTO DE LINGUAGEM NATURAL

Tudo que os seres humanos expressam (seja verbalmente ou por escrito) carrega uma grande quantidade de informação. Os assuntos escolhidos, o tom de voz, a escolha de palavras, tudo adiciona algum tipo de informação que pode ser interpretada e extraída algum valor. Em tese,

é possível entender e mesmo prever o comportamento humano usando essa informação. Mas há um problema: uma pessoa pode gerar centenas ou milhares de palavras em uma declaração, cada sentença com sua complexidade correspondente. Se deseja-se dimensionar e analisar centenas, milhares ou milhões de declarações de pessoas em uma determinada região, a situação fica incontrolável.

Dados gerados a partir de conversas, declarações ou até *tweets* são exemplos de dados não estruturados. Os dados não estruturados não se encaixam perfeitamente na estrutura tradicional de linhas e colunas de bancos de dados relacionais e representam a grande maioria dos dados disponíveis no mundo real. É confuso e difícil de manipular. No entanto, graças aos avanços em disciplinas como aprendizado de máquina, uma grande revolução está acontecendo em relação a esse tópico. Atualmente, não se trata mais de tentar interpretar um texto ou discurso com base em suas palavras-chave (a maneira mecânica antiquada), mas de entender o significado por trás dessas palavras (a maneira cognitiva). Dessa forma, é possível detectar figuras de linguagem como ironia, ou mesmo realizar análises de sentimentos (YZE, 2019).

O Processamento de Linguagem natural (Natural Language Processing - NLP) é um campo da inteligência artificial que oferece às máquinas a capacidade de ler, entender e derivar significado das linguagens humanas. É uma disciplina que se concentra na interação entre ciência de dados e linguagem humana e está sendo dimensionada para vários setores. Hoje, a NLP está crescendo graças às enormes melhorias no acesso aos dados e ao aumento do poder computacional, que estão permitindo que os profissionais obtenham resultados significativos em áreas como assistência médica, mídia, finanças e recursos humanos, entre outras.

Os principais desafios que se enfrenta atualmente com a NLP estão relacionadas ao fato de que a linguagem em si é muito complicada. O processo de compreensão e manipulação da linguagem é extremamente complexo e, por esse motivo, é comum usar técnicas diferentes para lidar com diferentes desafios antes de vincular tudo. Alguns dos algoritmos mais utilizados atualmente são:

Bag of Words *(*Saco de Palavras): É um modelo comumente usado que permite contar todas as palavras em um pedaço de texto. Basicamente, cria uma matriz de ocorrência para a sentença ou documento, desconsiderando a gramática e a ordem das palavras. Essas frequências ou ocorrências de palavras são usadas como recursos para o treinamento de um classificador.

Tokenization (Tokenização): É o processo de segmentar o texto corrente em frases e palavras. Em essência, é a tarefa de cortar um texto em pedaços chamados *tokens* e, ao mesmo tempo, jogar fora alguns caracteres com pouco valor semântico, como por exemplo a pontuação. O processo de tokenização pode ser particularmente problemático ao lidar com domínios de texto biomédicos que contêm muitos hifens, parênteses e outros sinais de pontuação.

Stop Words Removal (Remoção de Palavras Vazias): Inclui a eliminação de artigos, pronomes e preposições em idiomas comuns, como "e", "a" ou "para" em português. Nesse processo, algumas palavras muito comuns que parecem fornecer pouco ou nenhum valor ao objetivo da NLP são filtradas e excluídas do texto a ser processado, removendo, portanto, termos comuns e frequentes que não são informativos sobre o texto correspondente. As palavras vazias podem ser ignoradas com segurança, realizando uma pesquisa em uma lista predefinida de palavras-chave, liberando espaço no banco de dados e melhorando o tempo de processamento.

Stemming (Stemização): Refere-se ao processo de fatiar o final ou o início das palavras com a intenção de remover afixos (adições lexicais à raiz da palavra). Os afixos anexados no início da palavra são chamados prefixos (por exemplo, "astro" na palavra "astrobiologia") e os afixados no final da palavra são chamados sufixos (por exemplo, "mente" na palavra "sutilmente").

Lemmatization (Lematização): Tem como objetivo reduzir uma palavra à sua forma base e agrupar diferentes formas da mesma palavra. Embora pareça intimamente relacionado ao processo de stemização, a lematização usa uma abordagem diferente para alcançar as formas raiz das palavras. A lematização determina as palavras em sua forma de dicionário (conhecida como lema), para a qual requer dicionários detalhados nos quais o algoritmo pode pesquisar e vincular palavras aos lemas correspondentes. Por exemplo, as palavras "corria", "corre" e "correndo" são todas formas da palavra "correr", portanto "correr" é o lema de todas as palavras anteriores (YZE, 2019).

2.7 CLASSIFICADORES DE *MACHINE LEARNING*

Classificação é o processo de prever a classe de determinado conjunto de dados. Às vezes, as classes são chamadas de alvo / rótulo ou categorias. A modelagem preditiva de classificação é a tarefa de aproximar uma função de mapeamento (f) de variáveis de entrada (X) para variáveis de

saída discretas (*y*). Por exemplo, a detecção de *spam* nos provedores de serviços de e-mail pode ser identificada como um problema de classificação. Esta é uma classificação binária, uma vez que existem apenas duas classes como *spam* e não *spam*. Um classificador utiliza alguns dados de treinamento para entender como determinadas variáveis de entrada se relacionam com a classe. Nesse caso, os e-mails conhecidos de *spam* e não-*spam* devem ser usados como dados de treinamento. Quando o classificador é treinado com precisão, ele pode ser usado para detectar um e-mail desconhecido.

A classificação pertence à categoria de aprendizado supervisionado, onde os objetivos também são fornecidos com os dados de entrada. Existem muitas aplicações na classificação em muitos domínios, como aprovação de crédito, diagnóstico médico, mercado alvo etc. Existem dois tipos de aprendizagem na classificação que são a aprendizagem preguiçosa (*lazy learning*) e aprendizagem ansiosa (*eager learning*).

A *lazy learning* simplesmente armazena os dados do treinamento e espera até que os dados do teste apareçam. Quando isso ocorre, a classificação é realizada com base nos dados mais relacionados nos dados de treinamento armazenados. Comparado a *eager learning*, a *lazy learning* demanda menos tempo de treinamento, mas mais tempo na previsão. Como exemplos de algoritmos de *lazy learning* pode-se citar o *K-Nearest Neighbor* (KNN) e o *Case-Based Reasoning* (CBR)

Já o *eager learning* constrói um modelo de classificação com base nos dados de treinamento fornecidos antes de receber os dados para classificação. Ele deve ser capaz de se empenhar com uma única hipótese que cubra todo o espaço da instância. Devido à construção do modelo, o *eager learning* leva muito tempo para treinar e menos tempo para prever. Como exemplos de algoritmos de *eager learning* tem-se *a* Árvore de Decisão, *Naive Bayes* e Redes Neurais Artificiais (ASIRI, 2018).

Árvore de Decisão: A árvore de decisão constrói modelos de classificação ou regressão na forma de uma estrutura em árvore. Ele utiliza um conjunto de regras se-então que é mutuamente exclusivo e exaustivo para classificação. As regras são aprendidas sequencialmente usando os dados de treinamento, um de cada vez. Cada vez que uma regra é aprendida, as tuplas (conjunto com nomes e valores) cobertas pelas regras são removidas. Esse processo continua no conjunto de treinamento até atender a uma condição de término.

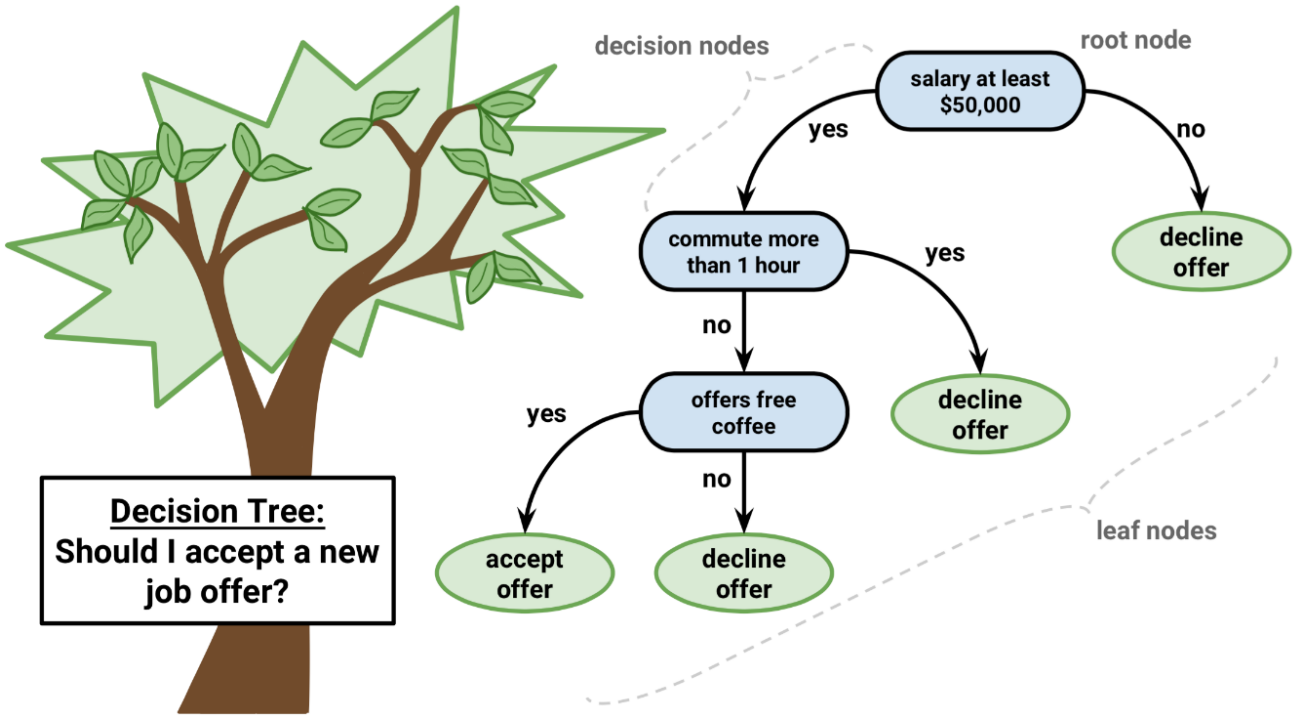


Figura 7: Exemplo de modelo de Árvore de Decisão (ASIRI, 2018)

A Figura 7 demonstra um exemplo de um modelo de Árvore de Decisão, no caso um modelo que tenta prever se uma determinada pessoa deve ou não aceitar uma proposta de trabalho. O modelo se inicia no nó raiz, que nesse exemplo é uma condição de salário mínimo anual para o aceite da vaga de emprego. Se a oferta não atinge um valor mínimo de salário, a oferta é prontamente recusada. Se essa condição é aceita, o modelo continua a avaliar as demais regras e, se todas as regras previamente definidas forem atendidas, a oferta é aceita. Os algoritmos *Random Forest, Gradient Boosting Classifier, Extra Tree Classifier e XGBoost Classifier* utilizados neste trabalho se baseiam nos conceitos do modelo de Árvore de Descisão.

Redes Neurais Artificiais: Conjunto de unidades de entrada / saída conectadas, nas quais cada conexão tem um peso associado. Seu funcionamento é baseado nos neurônios humanos. Durante a fase de aprendizado, a rede aprende ajustando os pesos para poder prever as classes corretas de entrada.

Podem haver várias camadas ocultas no modelo, dependendo da complexidade da função que será mapeada pelo modelo. Ter mais camadas ocultas permitirá modelar relacionamentos complexos, como redes neurais profundas. No entanto, quando existem muitas camadas ocultas,

leva muito tempo para treinar e ajustar os pesos. Uma outra desvantagem é a baixa interpretabilidade do modelo em comparação com outros modelos, como as Árvores de Decisão, devido ao significado simbólico desconhecido por trás dos pesos aprendidos.

As redes neurais artificiais tiveram um desempenho impressionante na maioria das aplicações do mundo real. É alta tolerância a dados ruidosos e é capaz de classificar padrões não treinados. Normalmente, as redes neurais artificiais têm melhor desempenho com entradas e saídas de valor contínuo (ASIRI, 2018). Neste trabalho, o algoritmo utilizado que se baseia nessa técnica é o *Perceptron*, que é o modelo mais simples de rede neural (apenas um neurônio).

3 CARATERIZAÇÃO DO PROBLEMA E MÉTODOS DE SOLUÇÃO

3.1 PROBLEMA EM QUESTÃO

Desde 1980, as causas externas (violência e acidentes) assumiram, tanto em âmbito mundial quanto no Brasil, uma posição de destaque no ranking da morbimortalidade, especialmente em grandes áreas urbanas, tendo como suas principais vítimas os homens jovens. Portanto, a redução da magnitude e da gravidade da violência na nossa sociedade, incluindo, para tanto, a ampliação do conhecimento sobre seus fatores de risco, configura-se como um dos maiores desafios para as políticas públicas de saúde na atualidade; e a ascensão destes eventos indesejáveis tem motivado estudos de grande importância no cenário nacional e internacional (Brasil, 2014).

A partir de 2015, a mortalidade por causas externas passou a ocupar o terceiro lugar no ranking da mortalidade geral no Espírito Santo, atrás das doenças cardiovasculares e neoplasias. No ano de 2008, 19,80% dos óbitos ocorridos entre os residentes no Espírito Santo, 18,11% em 2013 e em 2018, 14,39 % foram por causas externas.

Os óbitos por causas externas ocorrem em todas as faixas etárias, porém ao estratificar por faixa etária, a população de 05 a 49 anos se mostra ainda mais vulnerável, posto que há alguns anos, a primeira causa de mortalidade para essa faixa etária foram as causas externas. Destacam-se as faixas etárias de 15 a 19 e de 20 a 29 anos, quando, respectivamente, 81,6% e 73,2% da população para essas faixas etárias foram a óbito por estes agravos (Espírito Santo, 2019).

Os dados citados, tanto em âmbito regional quanto nacional, mostram que a mortalidade por causas externas representa uma importante preocupação para os chefes de estado do Espirito Santo e do Brasil. Com este trabalho busca-se, com o uso de diversas técnicas de Ciência de Dados, ter uma melhor compreensão do cenário atual e auxiliar no processo de tomada de decisão no que envolve medidas que amenizem e reduzam esses elevados índices da mortalidade.

3.2 MÉTODOS DE SOLUÇÃO

3.2.1 Modelagem Conceitual baseada em Ontologia

A primeira parte deste projeto consistiu em desenvolver uma ontologia com base nas informações apresentadas nos formulários de notificação e em um dicionário de dados fornecido pela Secretaria de Saúde do Espírito Santo. Este dicionário contém informações sobre os dados apresentados nos formulários, como nome, tipo, categoria do campo, uma pequena descrição e algumas características sobre o campo. Esse campo pode ser um campo obrigatório (a ausência de dados impossibilita a inclusão da notificação), um campo essencial (embora não obrigatório, registra os dados necessários para investigar o caso) ou regular. O dicionário de dados também contém informações sobre a relação entre os campos, por exemplo, se o sexo do paciente for preenchido com Masculino, o campo "Grávida" deve ser preenchido obrigatoriamente com “Não se aplica”.

Todas as informações e regras apresentadas no dicionário de dados foram usadas para desenvolver a ontologia do paciente. Essa ontologia, cuja visão geral está representada na Figura 8, foi desenvolvida no *software* Protégé e contém todas as classes, atributos, relações, regras e axiomas relacionados ao formulário de notificação.

Uma grande vantagem do uso de uma ontologia de domínio é sua capacidade de definir um modelo semântico dos dados combinado com o conhecimento do domínio associado (MUNIR; ANJUM, 2018). Além disso, é uma representação de conhecimento compartilhável e reutilizável que permite dimensionar o que está sendo feito a Secretaria de Saúde do Espírito Santo para os demais estados brasileiros.

3.2.2 Modelagem de Dados

Outra parte do trabalho desenvolvido foi a elaboração de um banco de dados contendo os dados das vítimas de agressão de maneira estruturada. Usando o software Tableau Desktop, essa formulação de banco de dados foi realizada. Para esta pesquisa, foram utilizadas as informações fornecidas pela Secretaria de Saúde do Espírito Santo, com dados da ficha de notificação de violência interpessoal e autoprovocada no Espírito Santo no período de 2011 a 2018, que possuem aproximadamente trinta e sete mil registros de ocorrências de violência.

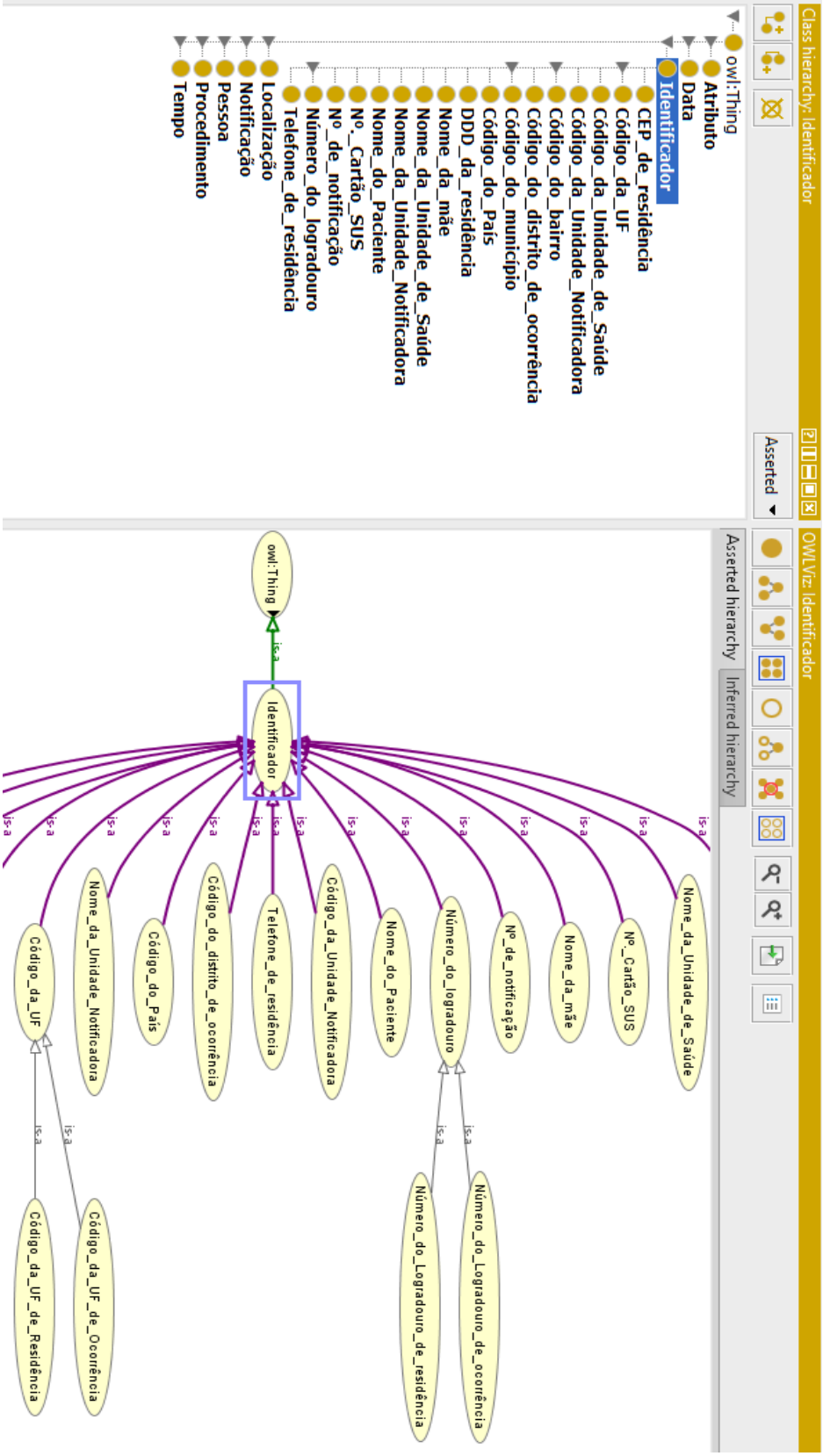
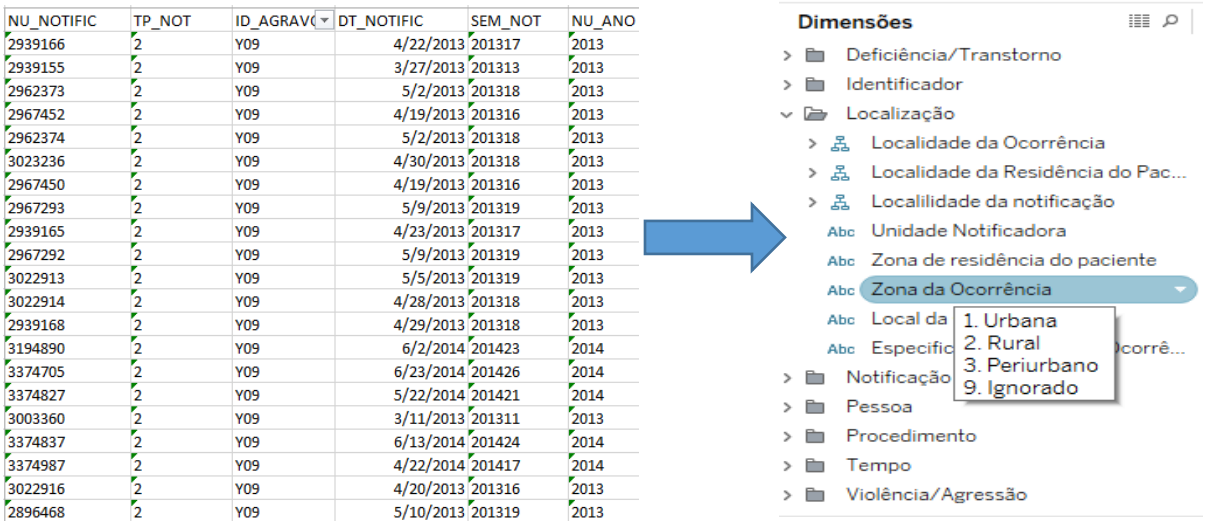


Figura 8: Parte da ontologia desenvolvida mostrando a classe Identificador

Nessa nova estrutura, em vez de uma tabela tradicional que normalmente apenas pessoas com um conhecimento considerável de computação e programação podem gerenciar, foi criada uma modelagem intuitiva que, com algumas ações de arrastar e soltar, um profissional de saúde pode criar sua própria análise e gráficos (Figura 9). Essa estrutura também foi usada para entender melhor a situação das vítimas usando técnicas de Ciência de Dados, Processamento de Linguagem Natural (NLP – Natural Language Processing) e *Machine Learning*.



NU_NOTIFIC	TP_NOT	ID_AGRAVO	DT_NOTIFIC	SEM_NOT	NU_ANO
2939166	2	Y09	4/22/2013	201317	2013
2939155	2	Y09	3/27/2013	201313	2013
2962373	2	Y09	5/2/2013	201318	2013
2967452	2	Y09	4/19/2013	201316	2013
2962374	2	Y09	5/2/2013	201318	2013
3023236	2	Y09	4/30/2013	201318	2013
2967450	2	Y09	4/19/2013	201316	2013
2967293	2	Y09	5/9/2013	201319	2013
2939165	2	Y09	4/23/2013	201317	2013
2967292	2	Y09	5/9/2013	201319	2013
3022913	2	Y09	5/5/2013	201319	2013
3022914	2	Y09	4/28/2013	201318	2013
2939168	2	Y09	4/29/2013	201318	2013
3194890	2	Y09	6/2/2014	201423	2014
3374705	2	Y09	6/23/2014	201426	2014
3374827	2	Y09	5/22/2014	201421	2014
3003360	2	Y09	3/11/2013	201311	2013
3374837	2	Y09	6/13/2014	201424	2014
3374987	2	Y09	4/22/2014	201417	2014
3022916	2	Y09	4/20/2013	201316	2013
2896468	2	Y09	5/10/2013	201319	2013

Figura 9: Comparativo entre a estrutura de dados antiga e a nova estrutura proposta

Na Figura 9 é possível observar na parte esquerda uma porção da estrutura antiga dos dados armazenados das fichas de notificação, baseados no modelo tradicionais de sistemas DBMS. Geralmente, nesses modelos, os nomes das colunas são organizados por códigos abreviados (NU = número, DT = data, ID = identificador), que facilitam o armazenamento dessas informações em banco de dados, mas dificultam o entendimento de usuários de perfis não técnicos, que muitas vezes necessitam consultar documentos como dicionário de dados para entender o que cada uma dessas colunas significa.

Na nova estrutura, que pode ser parcialmente vista na parte direita da Figura 9, cada uma das colunas foram divididas por pastas, que representam as classes principais da ontologia citada neste trabalho. Ao posicionar o cursor do *mouse* em cima de cada um desses campos, é possível obter uma descrição sobre o que cada um dos códigos representa. No exemplo da Figura 6, para o campo “Zona de Ocorrência” da violência, o código ‘1’ representa zona urbana, o código ‘2’ zona

rural, o ‘3’ representa periurbano (espaço onde as zonas urbanas e rurais se misturam), e o ‘9’ é o Ignorado, quando não foi possível coletar essa informação para determinado caso.

3.2.3 Uso de NLP para extração de informações do campo de ‘Observações Adicionais’

No formulário de notificação, existe um campo chamado "Observações adicionais", onde é possível escrever algumas observações sobre os casos que não puderam ser explicados apenas com os outros campos. Na maioria das vezes, há informações relevantes escritas nesses campos, mas é difícil extrair as informações desses dados para auxiliar no processo de tomada de decisão.

Para contornar esse obstáculo, foi proposto a utilização um algoritmo de NLP para extrair informações do campo 'Observações Adicionais' e agrupá-las por similaridade. Durante o processo de inserção das informações dos formulários no sistema pela equipe de saúde, havia uma quantidade considerável de palavras escritas incorretamente ou de forma abreviada no banco de dados. O algoritmo conseguiu agrupá-los por similaridade para permitir que essas informações fossem usadas posteriormente na análise (Figura 10).

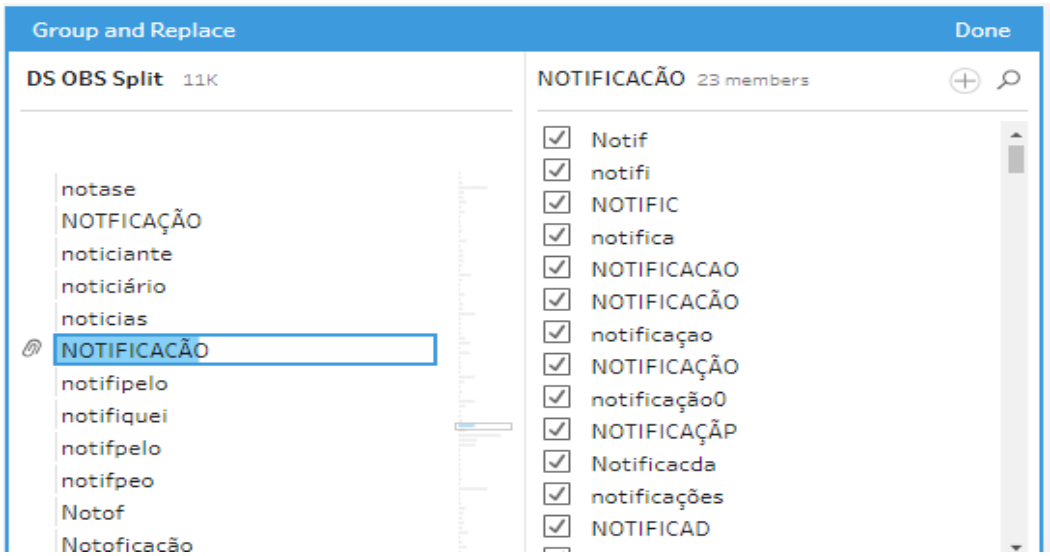
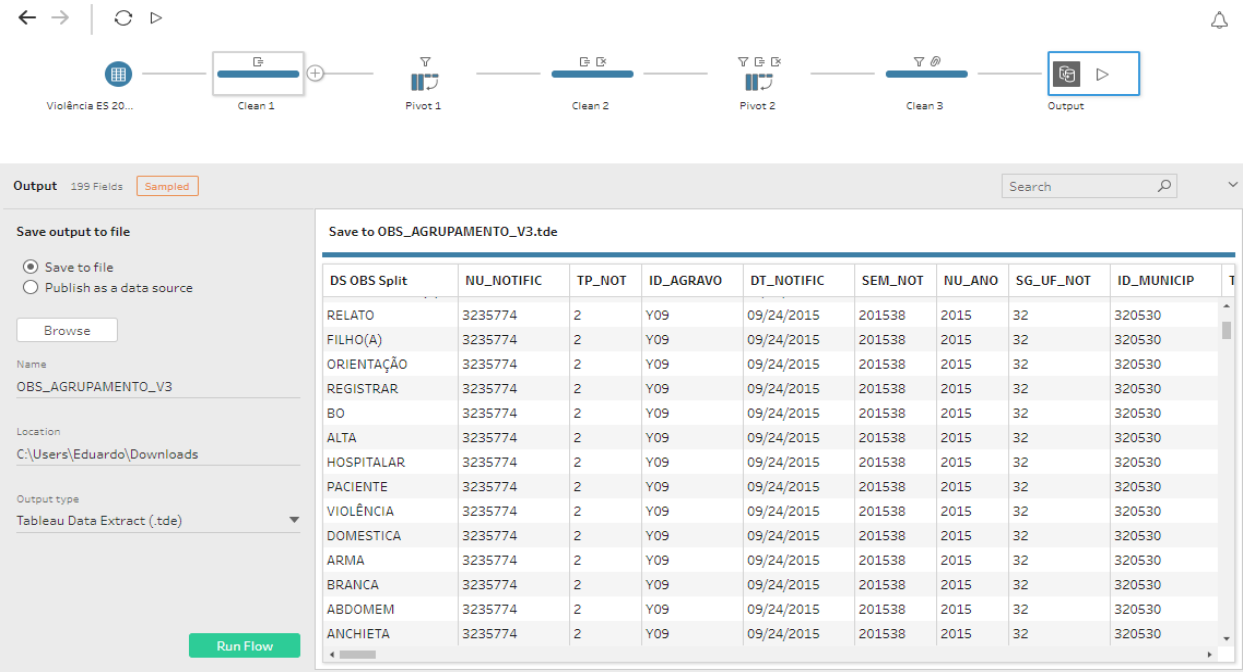


Figura 10: Exemplo do algoritmo de NLP agrupando a palavra “Notificação” com similares

Para realizar tal análise, foi criado um mapa de ETL (do inglês, *Extract*, *Transform*, *Load*) no programa *Tableau Prep* que extrai as informações presentes nas fichas de notificação, faz uma série de transformações nos dados como por exemplo filtragens de palavras e pontuações com pouco significado para a compreensão do contexto (como por exemplo artigos: a, o, de, da, na, além de outras palavras com pouco valor semântico), faz o 'pivoteamento' (transformar linhas em colunas) das informações presentes no campo de observações adicionais, agrupa as demais palavras por similaridade utilizando a técnica de lematização citada na seção 2.6 deste trabalho e retorna como saída do processo um novo arquivo onde as informações de observações adicionais podem ser usadas de maneira a criar a nuvem de palavras (*wordcloud)* apresentada na seção 4.2 deste trabalho.

A Figura 11 mostra o mapa com o processo de ETL desenvolvido no *Tableau Prep*, além de algumas linhas de exemplo do novo formato dos dados depois deste processo. Na Figura 8 é possível observar as etapas de limpeza de dados (representados pelas etapas de *Clean)*, os pivoteamentos (representados pelas etapas de *Pivot)* e a etapa final que gera o arquivo de saída (representado pelo *Output*). O processo de pivoteamento é necessário para que o *software* possa contar corretamente cada ocorrência de uma determinada palavra. Sem esse processo, a contagem seria por frases e não por palavras.



DS OBS Split	NU_NOTIFIC	TP_NOT	ID_AGRAVO	DT_NOTIFIC	SEM_NOT	NU_ANO	SG_UF_NOT	ID_MUNICIP
RELATO	3235774	2	Y09	09/24/2015	201538	2015	32	320530
FILHO(A)	3235774	2	Y09	09/24/2015	201538	2015	32	320530
ORIENTAÇÃO	3235774	2	Y09	09/24/2015	201538	2015	32	320530
REGISTRAR	3235774	2	Y09	09/24/2015	201538	2015	32	320530
BO	3235774	2	Y09	09/24/2015	201538	2015	32	320530
ALTA	3235774	2	Y09	09/24/2015	201538	2015	32	320530
HOSPITALAR	3235774	2	Y09	09/24/2015	201538	2015	32	320530
PACIENTE	3235774	2	Y09	09/24/2015	201538	2015	32	320530
VIOLÊNCIA	3235774	2	Y09	09/24/2015	201538	2015	32	320530
DOMESTICA	3235774	2	Y09	09/24/2015	201538	2015	32	320530
ARMA	3235774	2	Y09	09/24/2015	201538	2015	32	320530
BRANCA	3235774	2	Y09	09/24/2015	201538	2015	32	320530
ABDOMEM	3235774	2	Y09	09/24/2015	201538	2015	32	320530
ANCHIETA	3235774	2	Y09	09/24/2015	201538	2015	32	320530

Figura 11: Mapa de ETL desenvolvido no *Tableau Prep* e alguns registros de saída

3.2.4 Uso de *machine learning* para prever casos de violência que ocorreram múltiplas vezes

Existe um campo no formulário de notificação chamado "Ocorreu outras vezes?", que indica se um caso de violência ocorreu repetidamente. Esse campo é de extrema importância para a compreensão dos casos e é usado como indicativo para casos que podem evoluir para uma violência mais extrema, como homicídios ou suicídios.

Mesmo sendo um campo importante para análise, nem sempre é possível coletar essas informações sobre o caso. Cerca de 25% dos casos registrados foram preenchidos com "Nulo" ou "Ignorado" para este campo (os outros 75% são para os registros de “Sim” ou “Não”). Para contornar esse problema, foi proposto o uso de técnicas de *machine learning* para tentar prever a violência que ocorreu de maneira repetida nos casos em que não era possível registrar esses dados.

Para realizar essa predição, diversos algoritmos foram utilizados, entre eles a regressão logística, *Perceptron, Random Forest, Support Vector Machine e* outros algoritmos que são comumente utilizados para realizar previsões de valores tomados por uma variável categórica, frequentemente binária, que é o caso da variável utilizada neste trabalho, que indica se uma violência ocorreu repetidamente, e é igual a 1 se sim e igual a 0 se não ocorreu mais de uma vez.

Apesar de cada um desses algoritmos possuírem suas particularidades, o passo-a-passo utilizado para o treinamento de cada um desse algoritmos é bem similar. Inicialmente é feita a leitura do arquivo com os dados que são utilizados para o treinamento desse modelo, que nesse caso são os dados das fichas de notificação. Em seguida é feita uma etapa de limpeza desses dados, já que há muitas colunas que possuem boa parte de seus dados nulos, além de outras colunas que não são tão relevantes para esse trabalho de predição.

Em seguida, são identificados os campos que possuem maior correlação com a variável que se deseja prever, que no caso é a variável que indica se a violência ocorreu mais de uma vez. Encontrando esses campos (que são citados na seção 4.3 deste trabalho), é feita uma etapa de normalização e codificação desses dados, para deixá-los em uma mesma escala e para transformar variáveis categóricas em numéricas, já que a maioria dos algoritmos necessitam de variáveis numéricas para realizar a previsão corretamente.

Com esses dados no formato correto, é feita a divisão de conjunto de dados de teste e de treinamento (pode também ser incluído também um conjunto de dados de validação). O conjunto

de dados de treinamento (geralmente a maior parte dos dados disponíveis, de 65 a 75%) é utilizado para treinar o modelo. Nessa etapa o modelo “vê” e aprende com os dados disponíveis. Com o conjunto de dados de validação é feito o ajuste dos parâmetros dos modelos, para tentar obter melhores resultados de previsão. Já o conjunto de testes é usado para, de fato, avaliar o modelo. Nele são utilizados os diversos algoritmos citados neste capítulo e são comparados os resultados obtidos com cada um deles.

Os indicadores de resultado mais comuns são o de precisão (razão entre as observações positivas previstas corretamente e o total de observações positivas previstas), *recall* (a proporção de observações positivas previstas corretamente e todas as observações reais) e *F1 Score* (a média harmônica da precisão e *recall)*, os quais foram utilizados nesse trabalho para a avaliação dos algoritmos.

4 RESULTADOS OBTIDOS

4.1 ONTOLOGIA DE DOMÍNIO

Conforme citado na sessão 3.2.1 deste trabalho, uma ontologia de domínio foi desenvolvida para descrever os diversos campos de dados (classes) presentes na ficha de notificação e os relacionamentos (relações) entre cada uma dessas classes.

Nessa ontologia foram criadas 8 classes principais, que são: **Atributo** (campos da fonte de dados que descrevem as características da vítima, do autor da violência ou a própria violência), **Data** (datas relacionadas com a notificação, com a ocorrência da violência ou datas de nascimento), **Identificador** (códigos e nomes que identificam um local ou que identificam a vítima), **Localização** (detalha o local de ocorrência da violência, da residência da vítima ou a localidade da Unidade Notificadora), **Pesso**a (descreve as pessoas envolvidas na notificação de violência), **Notificação** (descreve o tipo de notificação), **Procedimento** (procedimentos que podem ser realizados se a violência ocorrida for do tipo sexual, além de encaminhamentos para órgãos responsáveis) e **Tempo** (relação temporal de ocorrência da violência).

Dentro de cada uma dessas classes principais, há diversas subclasses que caracterizam mais detalhadamente os campos presentes na Ficha de Notificação de Violência Interpessoal/Autoprovocada. Dessa maneira foi possível representar todos os campos presentes na ficha em forma de modelo ontológico.

Além das classes e subclasses, também foram definidos os indivíduos. Os indivíduos são os exemplares que representam cada classe. Por exemplo, no caso de referirmos ao campo da ficha "UF de ocorrência da violência", um indivíduo dessa classe é código "32" que representa o código do estado do Espírito Santo padronizado pelo IBGE. O conjunto dos 28 códigos da UF (27 estados e Distrito Federal) representam a classe "UF de ocorrência da violência".

Os relacionamentos também constituem uma importante parte da Ontologia. Toda a estrutura da Ontologia é construída em um modelo de relação de inclusão (também conhecido como *is-a* ou em português "é um"). Nesse modelo, todas as subclasses são conectadas a 'classe pai' por meio de uma relação do tipo *'is-a'*. A Figura 9 mostra um exemplo desse relacionamento utilizando a classe "Unidade Notificadora".

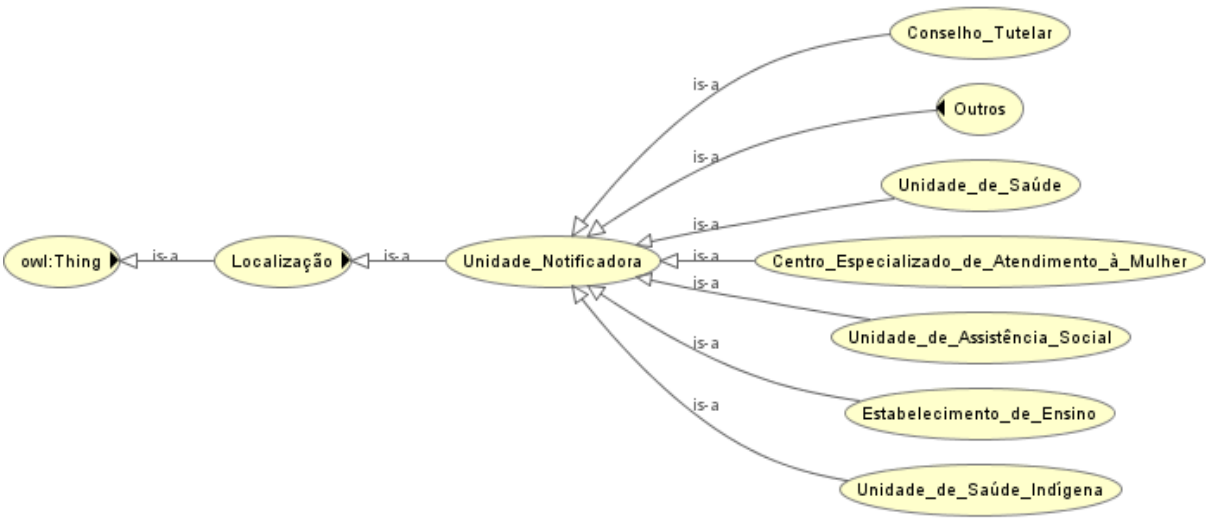


Figura 12: Exemplo de relação 'is-a' utilizando a classe Unidade Notificadora

Na Figura 12 é possível observar as diversas subclasses da classe "Unidade Notificadora", que são conectadas à classe pai por meio de uma relação do tipo *'is-a'*. Já a classe "Unidade Notificadora" está conectada a classe "Localização" e essa está conectada a superclasse *owl: Thing* (do inglês "coisa"), que é a classe pai de todas as classes.

Além das relações diretas do tipo 'é um" também foram definidos outros tipos de relacionamentos como, por exemplo, a relação "faz parte de". Nesse tipo de relacionamento é possível demonstrar como objetos se combinam para formar objetos compostos. No caso das hierarquias de localidade, podemos representar de seguinte maneira: Logradouro "faz parte de" um Bairro, Bairro "faz parte de" um Município, Município "faz parte de" uma Unidade da Federação, Unidade da Federação "faz parte de" um País.

Definições de restrições entre o relacionamento de classes também foram definidos quando necessário. Tomemos por exemplo a classe 'Gestante', um dos campos presentes na ficha de notificação. No dicionário de dados é definido que " Se Sexo=Masculino ou idade menor ou igual a 10 anos, o campo é preenchido automaticamente com Não se Aplica". Esse tipo de relacionamento também pode ser definido na ontologia por meio de relações de disjunção. Esse tipo de relação significa que esses eventos são mutualmente exclusivos, ou seja, não podem ocorrer simultaneamente. Ao definir regras como essa é possível encontrar, por exemplo, inconsistências nos registros da fonte de dados, onde dois campos que deveriam ser disjuntos foram preenchidos com informações conflitantes.

4.2 *DASHBOARDS*

Diversos painéis foram desenvolvidos para entender melhor a situação da violência por causas externas no estado do Espírito Santo. O primeiro painel (Figura 13) faz uma representação geral da situação da Violência no estado do ES. Esses painéis são interativos, portanto, é possível, por exemplo, clicar em uma cidade e filtrar os outros gráficos para entender melhor a situação nessa região específica.

No canto superior esquerdo é mostrado o mapa do Espírito Santo, onde as cidades com cores mais intensas representam as cidades com maior quantidade de incidentes registrados. A maioria das cidades de cor mais escuras estão localizadas na região da Grande Vitória, onde vive a maioria da população e onde se concentra as principais atividades econômicas da região. Alterando o filtro de Indicador para “% de casos” (total de casos registrados/população total), o perfil de cidades com cores mais escuras se altera drasticamente, e outros munícipios como Marechal Floriano (4,81 casos a cada mil habitantes) e Itaguaçu (3,91 casos a cada mil habitantes) aparecem de maneira proeminente.

No canto superior direito, é possível ver a taxa de incidentes pela cor da vítima. As vítimas que se definiram como "Pardas" representam mais da metade de todos os casos registrados (55,20%), seguido das que se definiram brancas (30,6%) e negras (13,0%). Fazendo um comparativo entre o tipo de violência sofrida (Interpessoal ou Autoprovocada), temos um aumento do percentual para as raças pardas e negras (56,3% e 14,1%, respectivamente) para a violência interpessoal, enquanto para a raça branca há um decréscimo de mais de 2 pontos percentuais (28,41%). Fazendo a análise para a violência do tipo autoprovocada, o cenário é contrário. Há um aumento considerável para as vítimas que se definiram como brancas (36,0%) e uma nítida diminuição percentual para aquelas que se definiram como pardas ou negras (52,5 % e 10,4%, respectivamente).

No canto inferior esquerdo, há a divisão dos incidentes por sexo da vítima. De cada quatro casos registrados, três são violência contra mulheres (75,1%). Por fim, no canto inferior direito, é possível analisar a pirâmide etária dos incidentes. Uma faixa etária que chama a atenção são os "10 a 19 anos", em que um número considerável de incidentes foi registrado em ambos os sexos. Essa informação se torna mais alarmante, considerando que estamos falando de crianças e

adolescentes, que ainda estão no começo de suas vidas e poderiam, de alguma forma, adquirir algum tipo de trauma devido a esses incidentes (Finkelhor e. al., 2013).

É importante ressaltar que a Ficha de Notificação de Violência Interpessoal e Autoprovacada não se aplica à violência extrafamiliar cujas vítimas sejam adultos (20 a 59 anos) do sexo masculino, como, por exemplo, brigas entre gangues, brigas nos estádios de futebol e outras. Essa modalidade de violência pode ser monitorada por meio de outros sistemas de informação e do componente do Viva Sentinela (BRASIL, 2016).

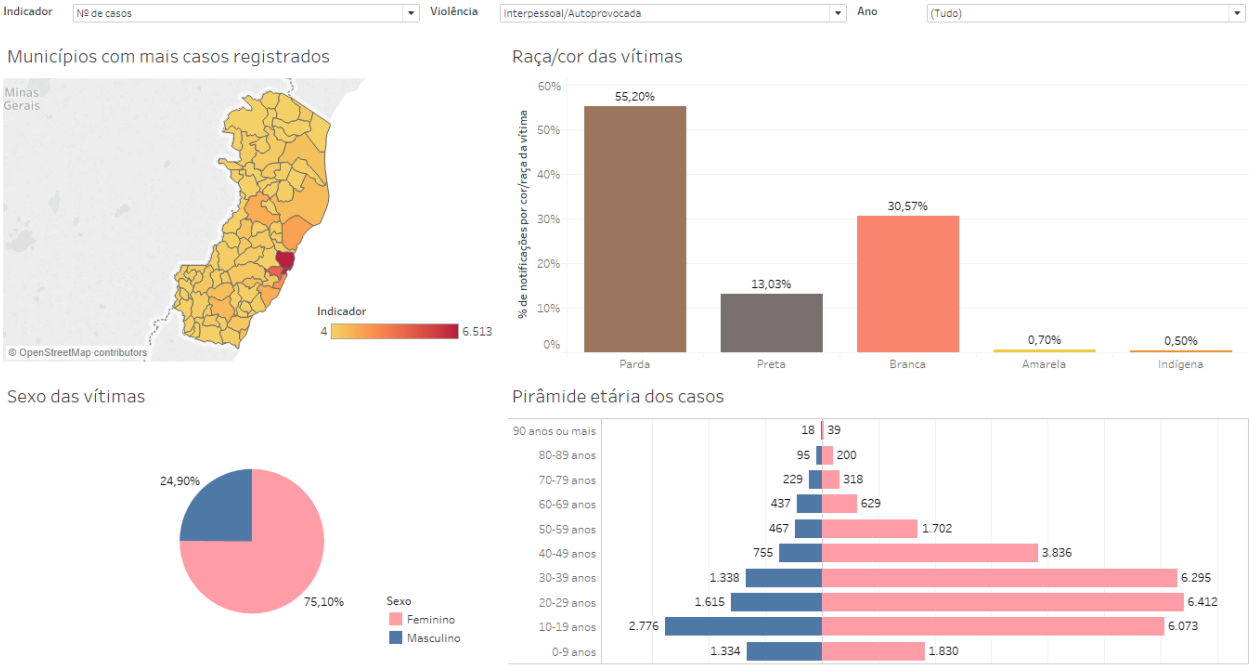


Figura 13: *Dashboard* que exibe a situação da violência Interpessoal e Autoprovocada no ES

Um outro painel desenvolvido diz respeito a relação entre a data de ocorrência da violência e a data em que essa violência foi notificada, conforme mostrado na Figura 14. Nesse painel é possível identificar visualmente os dias do ano com maiores incidências de notificações e ocorrências de violência, representados por cores mais intensas no calendário.

Por esse painel é possível identificar que há um número grande de notificações nos primeiros dias da semana (geralmente na segunda-feira) e um baixo índice de registro de notificações nos finais de semana. Já os registros de ocorrência não se concentram em um dia específico da semana, mas costumam ser mais incidentes em datas festivas e feriados.

Também é possível, por esse painel, visualizar casos atípicos como por exemplo o que está em destaque na Figura (dia 06/02/2017), que foi o dia que se intensificou a crise de segurança pública no Espírito Santo devido a paralisação da Polícia Militar. É possível notar também pela imagem, que alguns casos só foram notificados meses depois do acontecimento. Esse fato acontece também para vários outros dias de ocorrência, mostrando que algumas vítimas costumam procurar ajuda bem depois do ocorrido (em alguns casos depois de anos).

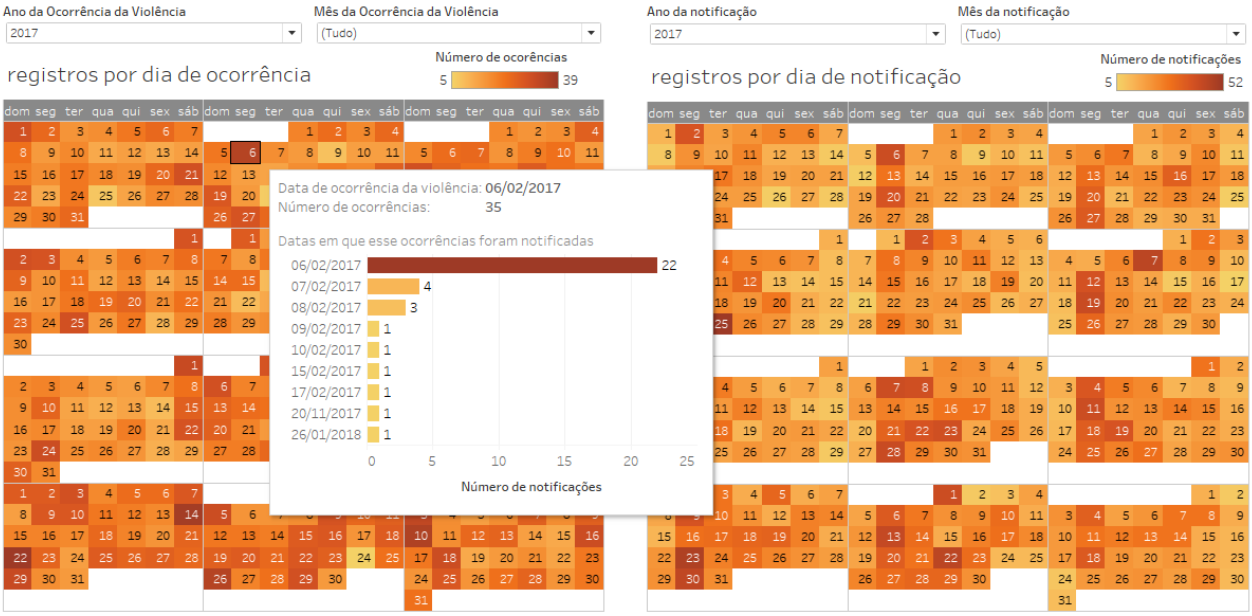


Figura 14: *Dashboard* que exibe a relação entre os registros de violência por dia de ocorrência e notificação

Outro painel desenvolvido está demonstrado na Figura 15, cujo canto superior esquerdo exibe um gráfico do tipo *treemap* que mostra a porcentagem de incidentes pela relação entre agressor e vítima. O que chama muita atenção é que a agressão causada pela própria pessoa (autoagressão) é a que possui mais registros (26,0% do total de incidentes).

No canto superior direito há um gráfico de área que mostra a evolução dos incidentes por ano, onde é possível escolher um parâmetro de análise. Este parâmetro pode ser definido como cor, sexo, faixa etária ou estado civil da vítima.

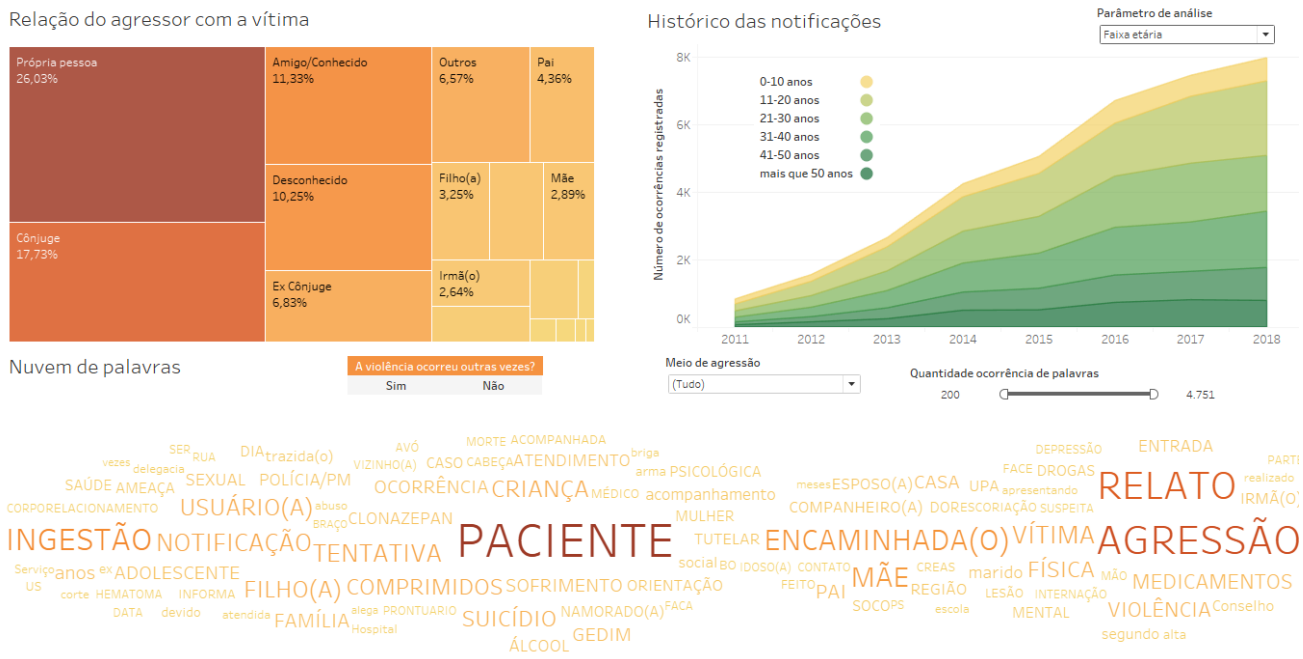


Figura 15: *Dashboard* que exibe mais a fundo a situação da violência Interpessoal e Autoprovocada no ES

Por fim, na parte inferior, há uma nuvem de palavras feitas com os comentários preenchidos no campo 'Observações adicionais'. Essas palavras foram agrupadas conforme mencionado na seção 3.2.3 desta dissertação e o número de ocorrências representa o tamanho da palavra na nuvem e quanto mais ocorrências, mais escura é a cor da palavra no painel. É possível analisar que algumas palavras como 'Paciente', 'Agressão', 'Ingestão', 'Relato' têm um número maior de registros. Também é possível filtrar o painel para ver apenas as ocorrências que ocorreram mais de uma vez (ou não), limitar a quantidade de ocorrência de palavras na nuvem de palavras e escolher o meio de agressão.

Esse também é um painel interativo e algumas informações importantes que podem ser extraídas deste painel. Por exemplo, ao clicar na caixa relacionada à autoagressão, é possível observar um crescimento quase exponencial dos casos entre 2011 e 2018 e o padrão da nuvem de palavras muda consideravelmente e outros palavras como “comprimidos”, “suicídio”, “tentativa” e "medicamentos" aparecem com mais destaque.

Resultados significativos foram obtidos alterando o parâmetro de análise e interagindo com os filtros. Por exemplo, ao definir o parâmetro para 'faixa etária' e analisar cada uma das faixas, observou-se que, na faixa de 0 a 10 anos, o pai e a mãe da vítima são os maiores agressores (26,90% e 19,00% das ocorrências, respectivamente). Nas faixas de 11 a 20 anos e de 21 a 30 anos, a

violência autoprovocada é a que possui mais registros (30,2% e 29,8%, respectivamente) e nas faixas de 31 a 40 anos e de 41 a 50 anos a agressão causada pelo cônjuge é a que possui mais ocorrências junto com a lesão autoprovocada (29,1% e 28,9%, respectivamente). Por fim, na faixa de 'Mais de 50 anos', o filho (a) das vítimas é quem tem mais registros de agressão (23,3% dos casos).

Alterando o valor do parâmetro de análise para “Situação conjugal” e analisando individualmente cada uma das categorias, temos que para as vítimas Casadas ou em União consensual, o cônjuge é responsável por 46,10% dos incidentes registrados. Considerando somente os casos de violência que ocorreram mais de uma vez, esse valor percentual aumenta para 58,2%. Para as vítimas que se encontram separadas, o ex-cônjuge é o principal autor dos registros de violência com 35,1% dos casos (43,8% para as violências que ocorreram repetidamente). Para os que se declararam solteiros, a violência autoprovocada é que apresenta mais casos (26,9%), para os viúvos o (a) filho (a) é o maior autor de violências (41,2%) e para a categoria “Não se aplica”, que representa as crianças com até 10 anos de idade, o pai é o maior agressor (23,8%).

4.3 MODELO DE *MACHINE LEARNING*

Como mencionado na seção 3.2.4, diversos algoritmos de *machine learning* foram desenvolvidos para tentar prever se a violência ocorreu mais de uma vez nos casos em que não foi possível registrar esses dados. Este algoritmo foi desenvolvido usando a linguagem *Python* e o *Jupyter* Notebook e pode ser consultado no Anexo B desse trabalho.

Inicialmente, como uma boa prática de códigos de *Python* (e de outras linguagens), é feita a importação dos pacotes e bibliotecas necessários para a utilização de funções já prontas que facilitam e reduzem a complexidade do código. Em seguida é feita a leitura do arquivo com as informações dos casos notificados. No arquivo, há alguns campos com a grande maioria de seus registros em branco ou nulos pois nem sempre é possível coletar todas as informações relacionadas ao incidente. Esses campos são descartados, pois sem essas informações não é possível obter *insights* significativos.

Fazendo uma análise de correlação, foi possível descobrir os campos com a maior correlação com o campo "Ocorreu outras vezes?" apresentado no formulário de notificação. Os

campos VIOL_PSICO (indica se houve violência psicológica), REL_CONJ (indica se o cônjuge é o autor da violência), CS_SEXO (o sexo da vítima), REL_EXCON (indica se o ex-cônjuge é o autor da violência), AUTOR_ALCO (indica se o autor da violência tinha suspeita de ter ingerido álcool) apresentam a maior correlação positiva e o campo REL_DESCO (indica se o autor da violência é uma pessoa desconhecida) tem a maior correlação negativa com o campo "Ocorreu outras vezes? ".

Após uma etapa de limpeza, normalização e análise exploratória de dados (*Exploratory Data Analysis*, EDA), esses campos foram utilizados para treinar o algoritmo de aprendizado de máquina. Alguns dos classificadores mais comuns na literatura e conhecidos por conseguirem melhores resultados foram testados neste trabalho. A Tabela 1 mostra os resultados da precisão, *recall* e F1 *score* de alguns classificadores utilizados. Os classificadores tiveram, em geral, resultado semelhante. O SVM (*Support Vector Machine*) foi um pouco melhor que os demais em precisão, *recall* e F1 *score.*

Tabela 1: Performance dos algoritmos classificadores de *machine learning*

Algoritmo	**Precisão**	**Recall**	**F1 score**
Logistic Regression	0.7352	0.6926	0.7055
Perceptron	0.7526	0.6882	0.7070
Random Forest	0.7509	0.6906	0.7082
Gradient Boosting Classifier	0.7662	0.6994	0.7183
Support Vector Machine	0.7827	0.7009	0.7235
Extra Tree Classifier	0.7458	0.6903	0.7066
XGBoost Classifier	0.7680	0.7004	0.7195

A Figura 16 mostra a matriz de Confusão do algoritmo de Regressão Logística. A matriz de Confusão permite visualizar de maneira intuitiva os resultados do modelo. No eixo Y da figura temos as classes verdadeiras do modelo (1 se a violência ocorreu mais de uma vez e 2 se não ocorreu). Já no eixo X temos as classes previstas pelo modelo. O valor que aparece em cada um dos quadrantes é o total de registros que o modelo obteve para a determinada classe verdadeira/prevista. É possível notar pela figura que para a classe 2, os resultados de previsão são muito bons, acertando a grande maioria das previsões. Já para a classe 1, os resultados não são tão precisos. O ajuste de alguns parâmetros no modelo permite que se obtenha resultados melhores.

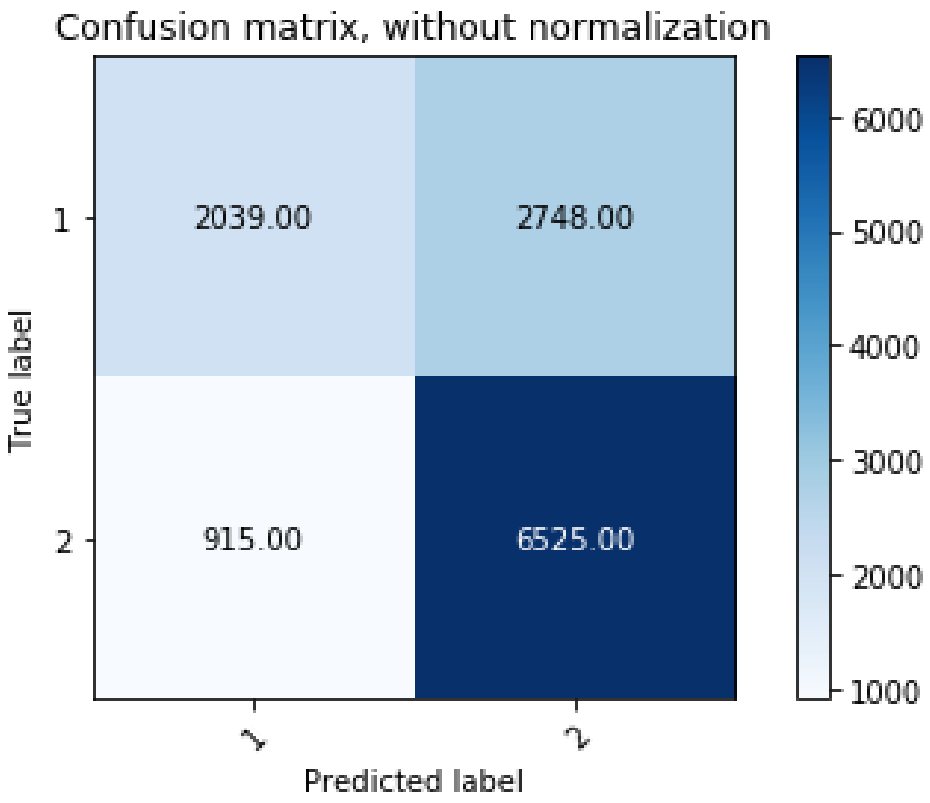


Figura 16: Matriz de Confusão para o modelo de Regressão Logística

5 CONCLUSÃO E TRABALHOS FUTUROS

Uma modelagem baseada em ontologia foi desenvolvida com os dados das fichas de notificação das vítimas de violência. Nesse modelo ontológico foi possível representar os diversos campos presentes nas fichas de notificação e também os tipos de relacionamentos que existem entre eles.

Também foi desenvolvida uma fonte de dados estruturada no *software Tableau Desktop*, que permite pessoas sem conhecimentos profundos de programação possam desenvolver suas próprias análises e relatórios com ações simples de *drag and drop (*arrastar e soltar). Para análises mais complexas, o programa permite a criação de "campos calculados", que são operações que permitem manipular os diversos campos presentes na base de dados.

A partir da fonte de dados estruturada, foram desenvolvidos *dashboards* que abordam mais a fundo a situação da violência no estado do Espírito Santo, de modo que foi possível compreender mais claramente os fatores que desencadeiam os diversos casos de violência no estado. Dentre os diversos resultados obtidos, pode-se citar alguns mais importantes como: vítimas que se definiram como "Pardas" representam mais da metade de todos os casos de violência registrados (55,20%), de cada quatro casos registrados, três são violência contra mulheres (75,10%) e um grande número de casos na faixa etária de 10 a 19 anos tanto para homens quanto para mulheres (2776 e 6073 registros de violência, respectivamente), mais de 20% de todos os casos registrados.

Também foram desenvolvidos modelos de Machine Learning com o objetivo de prever os casos de violência que ocorreram múltiplas vezes, no caso em que essa informação não foi possível de ser coletada. A maioria dos algoritmos obtiveram bons resultados, próximos a 70% para os indicadores utilizados, sendo que *Support Vector Machine (SVM)* teve o melhor desempenho, com precisão de 0.7827, recall de 0.7009 e F1 score de 0.7235.

Este estudo demonstrou que alguns *insights* importantes sobre a situação de violência foram obtidos com este projeto e, além disso, quase todo o trabalho desenvolvido para a Secretaria de Saúde do Espírito Santo pode ser dimensionado para outros estados brasileiros, já que a Ficha de Notificação de Violência Interpessoal/Autoprovocada é estruturada e preenchida da mesma maneira em todos os estados brasileiros.

Os painéis desenvolvidos nos ajudaram a obter informações importantes do cenário atual no Espírito Santo e fornecer mais conhecimento aos gerentes de saúde para apoio ao processo de tomada de decisão. O algoritmo de *Machine Learning* teve um desempenho satisfatório para os classificadores testados e poderia ser usado como um indicativo para casos em que não se puderam obter informações mais detalhadas sobre a violência.

Como trabalho futuro, pretende-se escalar o que foi feito para o Espírito Santo para os demais estados do Brasil, usando os dados dos outros estados brasileiros para treinar os modelos de *Machine Learning* desenvolvidos neste trabalho e obter melhores resultados de previsão. Além dos algoritmos já utilizados neste trabalho, também deseja-se futuramente utilizar outros tipos de algoritmos, como algoritmos mais complexos de Rede Neural Profunda.

Com os dados dos demais estados do Brasil também seria possível a criação de *dashboards* a nível nacional, permitindo realizar comparações entre Unidades da Federação para saber como um estado está situado em relação aos demais e ao Brasil em relação a Violência Interpessoal e Autoprovocada.

Um outro ponto importante que poderia ser explorado como trabalho futuro seria a atualização dos dados das fichas de notificação até a data mais recente, para que seja possível realizar análises de violência com a realidade atual que estamos enfrentando com a pandemia do covid-19 e poder saber como a situação de isolamento social que está sendo enfrentada está relacionada com o aumento dos casos de violência doméstica.

REFERÊNCIAS

Brasil, da Saúde, M. (2016) **"VIVA: Instrutivo – Notificação de Violência Interpessoal e Autoprovocada"**. Online, http://bvsms.saude.gov.br/bvs/publicacoes/viva_instrutivo_violencia_interpessoal_autoprovocada_2ed.pdf.

Cerqueira, D., Lima, S., Bueno, S., Valencia, L., Hanashiro, O., Machado, P., Lima, A. (2017) **"Atlas da Violência".** Online, http://www.ipea.gov.br/portal/images/170602_atlas_da_violencia_2017.pdf.

Man D., Didactica Mathematica, Vol. 31(2013), No 1, pp. 43–46 **Ontologies in Computer Science.**

Date, Christopher John. An introduction to database systems. **Pearson Education India**, 1981.

Gruber, T., A translation approach to portable ontologies, **Knowledge Acquisition**, 5 (1993), 199–220

Gruber, T., Toward principles for the design of ontologies used for knowledge sharing, **International Journal of Human-Computer Studies**, 43 (1995), 907–928.

Uschold, M. (2011) **Ontologies and Database Schema: Whats's the difference?** Online, https://pt.slideshare.net/UscholdM/ontologies-and-db-schema-whats-the-difference.

Hammer, M., & Mc Leod, D. (1981). Database description with SDM: a semantic database model. **ACM Transactions on Database Systems**, 6(3), 351–386. doi:10.1145/319587.319588

Dhar, V. (2013). Data science and prediction. **Communications of the ACM**, 56(12), 64–*73.* doi:10.1145/2500499

David Donoho (2017) **50 Years of Data Science, Journal of Computational and Graphical Statistics**, 26:4, 745-766, DOI: 10.1080/10618600.2017.1384734

Yze, Diego Lopez. (2019). **"Your Guide to Natural Language Processing (NLP)".** Online, https://towardsdatascience.com/your-guide-to-natural-language-processing-nlp-48ea2511f6e1

Asiri, Sidath. (2018). **"Machine Learning Classifiers".** Online, https://towardsdatascience.com/machine-learning-classifiers-a5cc4e1b0623

Corassa, R. (2018). **Violência por parceiro íntimo e homicídio de mulheres no Brasil, de 2011 a 2016**, Dissertação de Mestrado**.**

Modesto, J. G., Alves, A. Y. M., Santos, L. V., Archanjo, C. C. C., & Araújo, G. S. (2019). Fatores que influenciam na mortalidade de jovens por causas externas no Brasil: uma revisão da literatura**. Multidebates**, *3*(2), 137-155.

Stone D. M., Simon T. R., Fowler K. A., et al. Vital Signs: Trends in State Suicide Rates — United States, 1999–2016 and Circumstances Contributing to Suicide — 27 States, 2015. **MMWR Morb Mortal Wkly Rep** (2018); 67:617–624. DOI: http://dx.doi.org/10.15585/mmwr.mm6722a1.

Brasil. Ministério da Saúde. Secretaria de Vigilância em Saúde. **Saúde Brasil 2014: uma análise da situação de saúde e das causas externas**; 2014. Brasília, 2015

Espírito Santo, **Boletim Epidemiológico: uma análise da situação de saúde**, 2019, in https://saude.es.gov.br/Media/sesa/Documentos/Boletim%20Epidemiol%C3%B3gico%202%C2%AA%20Edi%C3%A7%C3%A2o.pdf

K. Munir, M. Anjum, The use of ontologies for effective knowledge modelling and information retrieval, in: **Applied Computing and Informatics 14** (2018), pp. 116-126. https://doi.org/10.1016/j.aci.2017.07.003

D. Finkelhor, H.A. Turner, A. Shattuck, S.L. Hamby. Violence, Crime, and Abuse Exposure in a National Sample of Children and Youth: An Update**. *JAMA Pediatr.*** 2013;167(7):614–621. https://doi.org/10.1001/jamapediatrics.2013.42

ANEXO A – Ficha Individual de Notificação/Investigação de Violência Interpessoal/ Autoprovocada

República Federativa do Brasil
Ministério da Saúde

SINAN
SISTEMA DE INFORMAÇÃO DE AGRAVOS DE NOTIFICAÇÃO
FICHA DE NOTIFICAÇÃO INDIVIDUAL

Nº

Caso suspeito ou confirmado de violência doméstica/intrafamiliar, sexual, autoprovocada, tráfico de pessoas, trabalho escravo, trabalho infantil, tortura, intervenção legal e violências homofóbicas contra mulheres e homens em todas as idades. No caso de violência extrafamiliar/comunitária, somente serão objetos de notificação as violências contra crianças, adolescentes, mulheres, pessoas idosas, pessoa com deficiência, indígenas e população LGBT.

Dados Gerais

1 Tipo de Notificação — 2 - Individual

2 Agravo/doença — **VIOLÊNCIA INTERPESSOAL/AUTOPROVOCADA** — Código (CID10) **Y09**

3 Data da notificação

4 UF

5 Município de notificação — Código (IBGE)

6 Unidade Notificadora — 1- Unidade de Saúde 2- Unidade de Assistência Social 3- Estabelecimento de Ensino 4- Conselho Tutelar 5- Unidade de Saúde Indígena 6- Centro Especializado de Atendimento à Mulher 7- Outros

7 Nome da Unidade Notificadora — Código Unidade

8 Unidade de Saúde — Código (CNES)

9 Data da ocorrência da violência

Notificação Individual

10 Nome do paciente

11 Data de nascimento

12 (ou) Idade — 1 - Hora 2 - Dia 3 - Mês 4 - Ano

13 Sexo — M - Masculino F - Feminino I - Ignorado

14 Gestante — 1-1ºTrimestre 2-2ºTrimestre 3-3ºTrimestre 4- Idade gestacional ignorada 5-Não 6- Não se aplica 9-Ignorado

15 Raça/Cor — 1-Branca 2-Preta 3-Amarela 4-Parda 5-Indígena 9- Ignorado

16 Escolaridade — 0-Analfabeto 1-1ª a 4ª série incompleta do EF (antigo primário ou 1º grau) 2-4ª série completa do EF (antigo primário ou 1º grau) 3-5ª à 8ª série incompleta do EF (antigo ginásio ou 1º grau) 4-Ensino fundamental completo (antigo ginásio ou 1º grau) 5-Ensino médio incompleto (antigo colegial ou 2º grau) 6-Ensino médio completo (antigo colegial ou 2º grau) 7-Educação superior incompleta 8-Educação superior completa 9-Ignorado 10- Não se aplica

17 Número do Cartão SUS

18 Nome da mãe

Dados de Residência

19 UF

20 Município de Residência — Código (IBGE)

21 Distrito

22 Bairro

23 Logradouro (rua, avenida,...) — Código

24 Número

25 Complemento (apto., casa, ...)

26 Geo campo 1

27 Geo campo 2

28 Ponto de Referência

29 CEP

30 (DDD) Telefone

31 Zona — 1 - Urbana 2 - Rural 3 - Periurbana 9 - Ignorado

32 País (se residente fora do Brasil)

Dados Complementares

Dados da Pessoa Atendida

33 Nome Social

34 Ocupação

35 Situação conjugal / Estado civil — 1 - Solteiro 2 - Casado/união consensual 3 - Viúvo 4 - Separado 8 - Não se aplica 9 - Ignorado

36 Orientação Sexual — 1-Heterossexual 2-Homossexual (gay/lésbica) 3-Bissexual 8-Não se aplica 9-Ignorado

37 Identidade de gênero: — 1-Travesti 2-Mulher Transexual 3-Homem Transexual 8-Não se aplica 9-Ignorado

38 Possui algum tipo de deficiência/ transtorno? — 1- Sim 2- Não 9- Ignorado

39 Se sim, qual tipo de deficiência /transtorno? — 1- Sim 2- Não 8-Não se aplica 9- Ignorado
Deficiência Física; Deficiência visual; Transtorno mental; Outras_____
Deficiência Intelectual; Deficiência auditiva; Transtorno de comportamento

Dados da Ocorrência

40 UF

41 Município de ocorrência — Código (IBGE)

42 Distrito

43 Bairro

44 Logradouro (rua, avenida,...) — Código

45 Número

46 Complemento (apto., casa, ...)

47 Geo campo 3

48 Geo campo 4

49 Ponto de Referência

50 Zona — 1 - Urbana 2 - Rural 3 - Periurbana 9 - Ignorado

51 Hora da ocorrência (00:00 - 23:59 horas)

52 Local de ocorrência
01 - Residência
02 - Habitação coletiva
03 - Escola
04 - Local de prática esportiva
05 - Bar ou similar
06 - Via pública
07 - Comércio/serviços
08 - Indústrias/construção
09 - Outro _____
99 - Ignorado

53 Ocorreu outras vezes? — 1 - Sim 2 - Não 9 - Ignorado

54 A lesão foi autoprovocada? — 1 - Sim 2 - Não 9 - Ignorado

SVS 15.06.2015

Violência

55 Essa violência foi motivada por: 01-Sexismo 02-Homofobia/Lesbofobia/Bifobia/Transfobia 03-Racismo 04-Intolerância religiosa 05-Xenofobia 06-Conflito geracional 07-Situação de rua 08-Deficiência 09-Outros_____ 88-Não se aplica 99-Ignorado

56 Tipo de violência 1- Sim 2- Não 9- Ignorado
- Física
- Psicológica/Moral
- Tortura
- Sexual
- Tráfico de seres humanos
- Financeira/Econômica
- Negligência/Abandono
- Trabalho infantil
- Intervenção legal
- Outros _____

57 Meio de agressão 1- Sim 2- Não 9- Ignorado
- Força corporal/ espancamento
- Enforcamento
- Obj. contundente
- Obj. pérfuro-cortante
- Substância/ Obj. quente
- Envenenamento, Intoxicação
- Arma de fogo
- Ameaça
- Outro_____

Violência Sexual

58 Se ocorreu violência sexual, qual o tipo? 1- Sim 2 - Não 8 - Não se aplica 9- Ignorado
- Assédio sexual
- Estupro
- Pornografia infantil
- Exploração sexual
- Outros _____

59 Procedimento realizado 1- Sim 2 - Não 8 - Não se aplica 9- Ignorado
- Profilaxia DST
- Profilaxia HIV
- Profilaxia Hepatite B
- Coleta de sangue
- Coleta de sêmen
- Coleta de secreção vaginal
- Contracepção de emergência
- Aborto previsto em lei

Dados do provável autor da violência

60 Número de envolvidos
1 - Um
2 - Dois ou mais
9 - Ignorado

61 Vínculo/grau de parentesco com a pessoa atendida 1-Sim 2-Não 9-Ignorado
- Pai
- Mãe
- Padrasto
- Madrasta
- Cônjuge
- Ex-Cônjuge
- Namorado(a)
- Ex-Namorado(a)
- Filho(a)
- Irmão(ã)
- Amigos/conhecidos
- Desconhecido(a)
- Cuidador(a)
- Patrão/chefe
- Pessoa com relação institucional
- Policial/agente da lei
- Própria pessoa
- Outros_____

62 Sexo do provável autor da violência
1 - Masculino
2 - Feminino
3 - Ambos os sexos
9 - Ignorado

63 Suspeita de uso de álcool
1- Sim
2 - Não
9- Ignorado

64 Ciclo de vida do provável autor da violência:
1-Criança (0 a 9 anos)
2-Adolescente (10 a 19 anos)
3-Jovem (20 a 24 anos)
4-Pessoa adulta (25 a 59 anos)
5-Pessoa idosa (60 anos ou mais)
9-Ignorado

Encaminhamento

65 Encaminhamento: 1-Sim 2-Não 9-Ignorado
- Rede da Saúde (Unidade Básica de Saúde,hospital,outras)
- Rede da Assistência Social (CRAS, CREAS, outras)
- Rede da Educação (Creche, escola, outras)
- Rede de Atendimento à Mulher (Centro Especializado de Atendimento à Mulher, Casa da Mulher Brasileira, outras)
- Conselho Tutelar
- Conselho do Idoso
- Delegacia de Atendimento ao Idoso
- Centro de Referência dos Direitos Humanos
- Ministério Público
- Delegacia Especializada de Proteção à Criança e Adolescente
- Delegacia de Atendimento à Mulher
- Outras delegacias
- Justiça da Infância e da Juventude
- Defensoria Pública

Dados finais

66 Violência Relacionada ao Trabalho
1 - Sim 2 - Não 9 - Ignorado

67 Se sim, foi emitida a Comunicação de Acidente do Trabalho (CAT)
1- Sim 2 - Não 8 - Não se aplica 9- Ignorado

68 Circunstância da lesão
CID 10 - Cap XX

69 Data de encerramento

Informações complementares e observações

Nome do acompanhante | Vínculo/grau de parentesco | (DDD) Telefone

Observações Adicionais:

TELEFONES ÚTEIS

Disque Saúde - Ouvidoria Geral do SUS 136

Central de Atendimento à Mulher 180

Disque Direitos Humanos 100

Notificador

Município/Unidade de Saúde | Cód. da Unid. de Saúde/CNES

Nome | Função | Assinatura

ANEXO B – Jupyter Notebook utilizado para a obtenção dos resultados de Machine Learning apresentados neste artigo

```
import numpy as np # linear algebra
import pandas as pd # data processing, CSV file I/O (e.g. pd.read_csv)
import matplotlib.pyplot as plt
import matplotlib.ticker as ticker


data = pd.ExcelFile("Violência_ES_2011_a_2018_v2.xlsx")
data.sheet_names
df = data.parse("Violência ES 2011 a 2018")
```

In [3]:

```
thresh = len(df) * .2
df.dropna(thresh = thresh, axis = 1, inplace = True)
t = pd.DataFrame(data={"col": df.dtypes.index, "type": df.dtypes}).reset_inde
x(drop=True)
col_names = t["col"][t.type != "object"]
col_names
```

Out[3]:

```
0      NU_NOTIFIC
1          TP_NOT
3      DT_NOTIFIC
4         SEM_NOT
5          NU_ANO
6       SG_UF_NOT
7      ID_MUNICIP
8      ID_UNIDADE
9      ID_REGIONA
10     ID_RG_RESI
11        DT_OCOR
12        SEM_PRI
14        DT_NASC
15     NU_IDADE_N
17     CS_GESTANT
18        CS_RACA
19     CS_ESCOL_N
20     ID_CNS_SUS
22          SG_UF
23     ID_MN_RESI
24      ID_BAIRRO
31            DDD
```

```
33           ZONA
34        ID_PAIS
36     SIT_CONJUG
37      DEF_TRANS
38     DEF_FISICA
39     DEF_MENTAL
40     DEF_VISUAL
41     DEF_AUDITI
          ...
119      REL_TRAB
120       REL_CAT
122    CLASSI_FIN
123      EVOLUCAO
124     DT_DIGITA
125    DT_TRANSSM
126    DT_TRANSSE
127     NU_LOTE_V
128    IDENT_MICR
129       REL_MAD
130      TPUNINOT
131    ORIENT_SEX
132     IDENT_GEN
133    VIOL_MOTIV
134      CICL_VID
135      REDE_SAU
136    ASSIST_SOC
137    REDE_EDUCA
138    ATEND_MULH
139    CONS_TUTEL
140      CONS_IDO
141    DELEG_IDOS
142     DIR_HUMAN
143           MPU
144    DELEG_CRIA
145    DELEG_MULH
146         DELEG
147     INFAN_JUV
148    DEFEN_PUBL
149    DT_ENCERRA
Name: col, Length: 131, dtype: object
```

In [4]:

```
#df = df[df != 9]
#df.dropna(subset=['OUT_VEZES'])
pd.DataFrame(abs(df.corr()['OUT_VEZES']).sort_values(ascending = False))
```

Out[4]:

	OUT_VEZES
OUT_VEZES	1.000000
LOCAL_OCOR	0.409799
AUTOR_ALCO	0.402655
ORIENT_SEX	0.350727
DEF_TRANS	0.349237
SIT_CONJUG	0.346805
CS_ESCOL_N	0.343554
AUTOR_SEXO	0.331427
REL_SEXUAL	0.329042
NUM_ENVOLV	0.319692
REL_CONHEC	0.283268
CICL_VID	0.280130
REL_DESCO	0.275837
REL_CONJ	0.271609
REL_EXNAM	0.265988
REL_EXCON	0.264599
REL_NAMO	0.258290

	OUT_VEZES
REL_IRMAO	0.254699
REL_PAD	0.250677
REL_MAD	0.248532
REL_POL	0.246738
REL_PAI	0.245831
REL_FILHO	0.244600
REL_OUTROS	0.243605
REL_CUIDA	0.243136
REL_INST	0.240983
REL_MAE	0.236329
REL_PATRAO	0.236131
CS_RACA	0.223375
LES_AUTOP	0.213272
...	...
SEX_OUTRO	0.048334
VIOL_SEXU	0.045673
VIOL_FINAN	0.044686

	OUT_VEZES
ID_CNS_SUS	0.043299
PEN_ANAL	0.042785
VIOL_INFAN	0.037494
VIOL_TRAF	0.037433
VIOL_LEGAL	0.036939
VIOL_FISIC	0.036531
NU_ANO	0.034834
NU_NOTIFIC	0.033638
TPUNINOT	0.030455
REL_CAT	0.029875
DDD	0.027084
VIOL_OUTR	0.024914
VIOL_NEGLI	0.024303
ID_BA_OCOR	0.016576
SEM_PRI	0.005304
NU_LOTE_V	0.005042
SEM_NOT	0.004615

	OUT_VEZES
SG_UF_NOT	0.004053
ID_MN_OCOR	0.003352
ID_MN_RESI	0.002865
SG_UF	0.002763
ID_MUNICIP	0.002204
ID_PAIS	0.002077
IDENT_MICR	0.001119
ZONA	0.000991
SG_UF_OCOR	0.000780
TP_NOT	NaN

124 rows × 1 columns

In [ ]:

```
fig = plt.figure(figsize=(50,50))

ax = fig.add_subplot(1,1,1)
cax = ax.matshow(df.corr(), interpolation = 'nearest')

fig.colorbar(cax)

ax.xaxis.set_major_locator(ticker.MultipleLocator(1))
ax.yaxis.set_major_locator(ticker.MultipleLocator(1))

ax.set_xticklabels(col_names)
ax.set_yticklabels(col_names);
```

In [ ]:

```
df = df[df.OUT_VEZES != 9]
fig = plt.figure(figsize=(50,50))
```

```
ax = fig.add_subplot(1,1,1)
cax = ax.matshow(df.corr())

fig.colorbar(cax)

ax.xaxis.set_major_locator(ticker.MultipleLocator(1))
ax.yaxis.set_major_locator(ticker.MultipleLocator(1))

ax.set_xticklabels(col_names)
ax.set_yticklabels(col_names);
```

In [5]:

```
#df = df[df.OUT_VEZES != 9]
df1 = df[['REL_DESCO','REL_CONJ','OUT_VEZES','VIOL_PSICO','CS_SEXO','AUTOR_AL
CO','REL_CONHEC','REL_EXCON','AUTOR_SEXO','NU_IDADE_N']]
#df1.dropna(subset=['OUT_VEZES'])
#df1=df1.fillna(df1.mean())
#df1['OUT_VEZES'] = df1['OUT_VEZES'].astype(str).str.replace("9.0","2.0").ast
ype(float)
#df1['REL_DESCO'] = df1['REL_DESCO'].astype(str).str.replace("9.0","2.0").ast
ype(float)
#df1['REL_CONJ'] = df1['REL_CONJ'].astype(str).str.replace("9.0","2.0").astyp
e(float)
#df1['REL_CONHEC'] = df1['REL_CONHEC'].astype(str).str.replace("9.0","2.0").a
stype(float)
#df1['VIOL_PSICO'] = df1['VIOL_PSICO'].astype(str).str.replace("9.0","2.0").a
stype(float)

#df1=df1.fillna(2)
df1 = df1[df1 != 9]
df1=df1.fillna(2)
#df1['REL_SEXUAL'] = df1['REL_SEXUAL'].astype(str).str.replace("8.0","2.0").a
stype(float)
df1.head(50)
#print(df1.dtypes)
```

Out[5]:

	REL_DESCO	REL_CONJ	OUT_VEZES	VIOL_PSICO	CS_SEXO	AUTOR_ALCO	REL_CONHEC	REL_EXCON	AUTOR_SEXO	NU_IDADE_N
0	2.0	2.0	1.0	1.0	F	2.0	2.0	2.0	1.0	4086
1	2.0	2.0	2.0	2.0	F	2.0	2.0	2.0	1.0	4022

	REL_DESCO	REL_CONJ	OUT_VEZES	VIOL_PSICO	CS_SEXO	AUTOR_ALCO	REL_CONHEC	REL_EXCON	AUTOR_SEXO	NU_IDADE_N
2	2.0	1.0	1.0	1.0	F	2.0	2.0	2.0	1.0	4029
3	2.0	2.0	2.0	2.0	F	2.0	2.0	2.0	2.0	4041
4	2.0	1.0	1.0	1.0	F	1.0	2.0	2.0	1.0	4045
5	2.0	1.0	1.0	1.0	F	2.0	2.0	2.0	1.0	4037
6	2.0	2.0	2.0	2.0	F	2.0	2.0	2.0	2.0	4030
7	2.0	2.0	2.0	2.0	F	2.0	2.0	2.0	2.0	4020
8	2.0	2.0	2.0	2.0	F	2.0	2.0	2.0	1.0	3010
9	2.0	2.0	2.0	2.0	F	2.0	2.0	2.0	2.0	4033
10	2.0	2.0	2.0	2.0	F	2.0	2.0	2.0	2.0	4045
11	2.0	1.0	1.0	2.0	F	1.0	2.0	2.0	1.0	4024
12	2.0	1.0	1.0	2.0	F	1.0	2.0	2.0	1.0	4048
13	2.0	2.0	2.0	2.0	M	2.0	2.0	2.0	2.0	4033
14	2.0	2.0	2.0	2.0	F	2.0	2.0	2.0	2.0	4034
15	2.0	2.0	2.0	2.0	M	2.0	2.0	2.0	2.0	4056
16	2.0	1.0	1.0	2.0	F	2.0	2.0	2.0	1.0	4029

	REL_DESCO	REL_CONJ	OUT_VEZES	VIOL_PSICO	CS_SEXO	AUTOR_ALCO	REL_CONHEC	REL_EXCON	AUTOR_SEXO	NU_IDADE_N
17	2.0	2.0	2.0	2.0	F	2.0	2.0	2.0	2.0	4032
18	2.0	2.0	2.0	2.0	F	2.0	2.0	2.0	2.0	4105
19	2.0	2.0	2.0	2.0	F	2.0	1.0	2.0	1.0	4040
20	2.0	1.0	1.0	1.0	F	1.0	2.0	2.0	1.0	4025
21	2.0	2.0	2.0	1.0	M	2.0	1.0	2.0	1.0	4016
22	2.0	2.0	2.0	2.0	M	2.0	1.0	2.0	1.0	4007
23	2.0	2.0	1.0	1.0	F	2.0	2.0	2.0	2.0	4022
24	2.0	2.0	2.0	2.0	F	2.0	2.0	2.0	2.0	4042
25	2.0	2.0	2.0	1.0	F	2.0	1.0	2.0	1.0	4012
26	2.0	2.0	2.0	2.0	M	2.0	2.0	2.0	2.0	4039
27	2.0	2.0	2.0	2.0	M	2.0	2.0	2.0	2.0	4027
28	2.0	2.0	2.0	2.0	M	2.0	1.0	2.0	1.0	4011
29	2.0	2.0	2.0	2.0	F	2.0	2.0	2.0	2.0	4025

	REL_DESCO	REL_CONJ	OUT_VEZES	VIOL_PSICO	CS_SEXO	AUTOR_ALCO	REL_CONHEC	REL_EXCON	AUTOR_SEXO	NU_IDADE_N
30	2.0	2.0	1.0	2.0	M	2.0	2.0	2.0	1.0	4071
31	2.0	2.0	2.0	2.0	M	2.0	2.0	2.0	2.0	4034
32	2.0	1.0	2.0	2.0	F	2.0	2.0	2.0	2.0	4043
33	2.0	2.0	2.0	2.0	F	2.0	1.0	2.0	1.0	4006
34	2.0	1.0	1.0	2.0	M	2.0	2.0	2.0	3.0	4058
35	1.0	2.0	2.0	1.0	F	2.0	1.0	2.0	1.0	4009
36	2.0	2.0	1.0	2.0	M	2.0	1.0	2.0	1.0	4012
37	1.0	2.0	2.0	2.0	F	1.0	2.0	2.0	1.0	4022
38	2.0	2.0	2.0	2.0	F	2.0	2.0	2.0	3.0	4001
39	2.0	2.0	2.0	1.0	F	2.0	2.0	1.0	2.0	4039
40	1.0	2.0	2.0	2.0	F	2.0	2.0	2.0	1.0	4014
41	2.0	2.0	2.0	1.0	F	2.0	2.0	2.0	1.0	4012
42	2.0	2.0	2.0	2.0	F	2.0	2.0	2.0	1.0	4005

	REL_DESCO	REL_CONJ	OUT_VEZES	VIOL_PSICO	CS_SEXO	AUTOR_ALCO	REL_CONHEC	REL_EXCON	AUTOR_SEXO	NU_IDADE_N
43	2.0	2.0	2.0	2.0	F	2.0	1.0	2.0	1.0	4034
44	2.0	2.0	2.0	2.0	F	1.0	2.0	2.0	1.0	4045
45	2.0	2.0	1.0	1.0	F	2.0	2.0	1.0	1.0	4033
46	2.0	1.0	2.0	1.0	F	1.0	2.0	2.0	1.0	4026
47	2.0	2.0	2.0	2.0	F	2.0	1.0	2.0	2.0	4088
48	2.0	2.0	2.0	2.0	M	2.0	2.0	2.0	1.0	4026
49	2.0	2.0	1.0	2.0	F	1.0	2.0	2.0	1.0	4007

In [6]:

```
df1['NU_IDADE_N'] = df1['NU_IDADE_N'].apply(str)
df1['NU_IDADE_N'] =df1['NU_IDADE_N'].apply(lambda x: x[2:4] if x>"4000" else
"0")
df1['NU_IDADE_N'] = df1['NU_IDADE_N'].apply(float)
df1['FAIXA_ETARIA']=pd.cut(df1['NU_IDADE_N'], bins=[-1,10, 20, 30, 40,50,100]
)
#df1['FAIXA_ETARIA'] =df1['NU_IDADE_N'].apply(lambda x: "0-10 anos" if x<=10
else if "0")
#NU_IDADE_N=IF(NU_IDADE[:1]=4:NU_IDADE[2:4]:ELSE:0)
df1.head(100)
```

Out[6]:

	REL_DESCO	REL_CONJ	OUT_VEZES	VIOL_PSICO	CS_SEXO	AUTOR_ALCO	REL_CONHEC	REL_EXCON	AUTOR_SEXO	NU_IDADE_N	FAIXA_ETARIA
0	2.0	2.0	1.0	1.0	F	2.0	2.0	2.0	1.0	86.0	(50, 100]

	REL_DESCO	REL_CONJ	OUT_VEZES	VIOL_PSICO	CS_SEXO	AUTOR_ALCO	REL_CONHEC	REL_EXCON	AUTOR_SEXO	NU_IDADE_N	FAIXA_ETARIA
1	2.0	2.0	2.0	2.0	F	2.0	2.0	2.0	1.0	22.0	(20, 30]
2	2.0	1.0	1.0	1.0	F	2.0	2.0	2.0	1.0	29.0	(20, 30]
3	2.0	2.0	2.0	2.0	F	2.0	2.0	2.0	2.0	41.0	(40, 50]
4	2.0	1.0	1.0	1.0	F	1.0	2.0	2.0	1.0	45.0	(40, 50]
5	2.0	1.0	1.0	1.0	F	2.0	2.0	2.0	1.0	37.0	(30, 40]
6	2.0	2.0	2.0	2.0	F	2.0	2.0	2.0	2.0	30.0	(20, 30]
7	2.0	2.0	2.0	2.0	F	2.0	2.0	2.0	2.0	20.0	(10, 20]
8	2.0	2.0	2.0	2.0	F	2.0	2.0	2.0	1.0	0.0	(-1, 10]
9	2.0	2.0	2.0	2.0	F	2.0	2.0	2.0	2.0	33.0	(30, 40]
10	2.0	2.0	2.0	2.0	F	2.0	2.0	2.0	2.0	45.0	(40, 50]
11	2.0	1.0	1.0	2.0	F	1.0	2.0	2.0	1.0	24.0	(20, 30]
12	2.0	1.0	1.0	2.0	F	1.0	2.0	2.0	1.0	48.0	(40, 50]
13	2.0	2.0	2.0	2.0	M	2.0	2.0	2.0	2.0	33.0	(30, 40]
14	2.0	2.0	2.0	2.0	F	2.0	2.0	2.0	2.0	34.0	(30, 40]
15	2.0	2.0	2.0	2.0	M	2.0	2.0	2.0	2.0	56.0	(50, 100]

	REL_DESCO	REL_CONJ	OUT_VEZES	VIOL_PSICO	CS_SEXO	AUTOR_ALCO	REL_CONHEC	REL_EXCON	AUTOR_SEXO	NU_IDADE_N	FAIXA_ETARIA
16	2.0	1.0	1.0	2.0	F	2.0	2.0	2.0	1.0	29.0	(20, 30]
17	2.0	2.0	2.0	2.0	F	2.0	2.0	2.0	2.0	32.0	(30, 40]
18	2.0	2.0	2.0	2.0	F	2.0	2.0	2.0	2.0	5.0	(-1, 10]
19	2.0	2.0	2.0	2.0	F	2.0	1.0	2.0	1.0	40.0	(30, 40]
20	2.0	1.0	1.0	1.0	F	1.0	2.0	2.0	1.0	25.0	(20, 30]
21	2.0	2.0	2.0	1.0	M	2.0	1.0	2.0	1.0	16.0	(10, 20]
22	2.0	2.0	2.0	2.0	M	2.0	1.0	2.0	1.0	7.0	(-1, 10]
23	2.0	2.0	1.0	1.0	F	2.0	2.0	2.0	2.0	22.0	(20, 30]
24	2.0	2.0	2.0	2.0	F	2.0	2.0	2.0	2.0	42.0	(40, 50]
25	2.0	2.0	2.0	1.0	F	2.0	1.0	2.0	1.0	12.0	(10, 20]
26	2.0	2.0	2.0	2.0	M	2.0	2.0	2.0	2.0	39.0	(30, 40]
27	2.0	2.0	2.0	2.0	M	2.0	2.0	2.0	2.0	27.0	(20, 30]
28	2.0	2.0	2.0	2.0	M	2.0	1.0	2.0	1.0	11.0	(10, 20]

	REL_DESCO	REL_CONJ	OUT_VEZES	VIOL_PSICO	CS_SEXO	AUTOR_ALCO	REL_CONHEC	REL_EXCON	AUTOR_SEXO	NU_IDADE_N	FAIXA_ETARIA
29	2.0	2.0	2.0	2.0	F	2.0	2.0	2.0	2.0	25.0	(20, 30]
...	...	...	...	...	...	...	...	...	...	...	...
70	2.0	2.0	1.0	2.0	M	2.0	2.0	2.0	2.0	16.0	(10, 20]
71	2.0	2.0	2.0	2.0	F	2.0	2.0	2.0	2.0	14.0	(10, 20]
72	2.0	2.0	2.0	2.0	M	2.0	2.0	2.0	2.0	26.0	(20, 30]
73	2.0	1.0	1.0	1.0	F	2.0	2.0	2.0	1.0	30.0	(20, 30]
74	2.0	2.0	2.0	1.0	F	1.0	2.0	2.0	1.0	15.0	(10, 20]
75	2.0	2.0	2.0	2.0	M	2.0	2.0	2.0	2.0	29.0	(20, 30]
76	2.0	2.0	2.0	1.0	M	2.0	2.0	2.0	1.0	12.0	(10, 20]
77	2.0	1.0	1.0	1.0	F	1.0	2.0	2.0	1.0	25.0	(20, 30]
78	2.0	1.0	1.0	1.0	F	2.0	2.0	2.0	1.0	24.0	(20, 30]
79	2.0	2.0	2.0	2.0	M	2.0	2.0	2.0	1.0	10.0	(-1, 10]
80	2.0	2.0	2.0	2.0	M	2.0	2.0	2.0	2.0	23.0	(20, 30]

	REL_DESCO	REL_CONJ	OUT_VEZES	VIOL_PSICO	CS_SEXO	AUTOR_ALCO	REL_CONHEC	REL_EXCON	AUTOR_SEXO	NU_IDADE_N	FAIXA_ETARIA
81	2.0	2.0	2.0	2.0	F	2.0	2.0	2.0	2.0	40.0	(30, 40]
82	2.0	2.0	2.0	1.0	F	2.0	2.0	2.0	1.0	36.0	(30, 40]
83	2.0	1.0	1.0	2.0	F	1.0	2.0	2.0	1.0	48.0	(40, 50]
84	2.0	2.0	1.0	2.0	F	2.0	2.0	2.0	1.0	2.0	(-1, 10]
85	1.0	2.0	2.0	2.0	F	1.0	2.0	2.0	1.0	52.0	(50, 100]
86	2.0	2.0	2.0	2.0	F	2.0	2.0	2.0	2.0	45.0	(40, 50]
87	2.0	2.0	2.0	2.0	M	2.0	2.0	2.0	2.0	17.0	(10, 20]
88	2.0	2.0	2.0	2.0	M	2.0	2.0	2.0	2.0	32.0	(30, 40]
89	2.0	2.0	2.0	2.0	F	2.0	2.0	2.0	2.0	24.0	(20, 30]
90	2.0	2.0	2.0	2.0	F	2.0	2.0	2.0	2.0	22.0	(20, 30]
91	2.0	1.0	2.0	1.0	F	2.0	2.0	2.0	1.0	36.0	(30, 40]
92	2.0	2.0	2.0	2.0	F	2.0	1.0	2.0	2.0	26.0	(20, 30]
93	2.0	2.0	1.0	2.0	M	2.0	2.0	2.0	1.0	14.0	(10, 20]

	REL_DESCO	REL_CONJ	OUT_VEZES	VIOL_PSICO	CS_SEXO	AUTOR_ALCO	REL_CONHEC	REL_EXCON	AUTOR_SEXO	NU_IDADE_N	FAIXA_ETARIA
94	2.0	2.0	1.0	1.0	F	1.0	2.0	1.0	1.0	18.0	(10, 20]
95	1.0	2.0	2.0	2.0	F	2.0	2.0	2.0	1.0	37.0	(30, 40]
96	2.0	1.0	1.0	1.0	F	1.0	2.0	2.0	1.0	20.0	(10, 20]
97	2.0	2.0	2.0	2.0	F	2.0	2.0	1.0	1.0	33.0	(30, 40]
98	2.0	1.0	1.0	2.0	F	1.0	2.0	2.0	1.0	33.0	(30, 40]
99	2.0	2.0	2.0	2.0	M	2.0	2.0	2.0	2.0	30.0	(20, 30]

100 rows × 11 columns

In [7]:

```
# we can use label encoder for conversion of numeric data
#df1 = df1[df1.CS_SEXO != "I"]
from sklearn.preprocessing import LabelEncoder
def encoding(feature):
    if (feature.dtype == 'object'):
        return LabelEncoder().fit_transform(feature)
    else:
        return feature
```

In [8]:

```
df1['FAIXA_ETARIA'] = df1['FAIXA_ETARIA'].cat.codes
df1=df1.drop(columns=['NU_IDADE_N'])
df1.head(50)
```

Out[8]:

	REL_DESCO	REL_CONJ	OUT_VEZES	VIOL_PSICO	CS_SEXO	AUTOR_ALCO	REL_CONHEC	REL_EXCON	AUTOR_SEXO	FAIXA_ETARIA
0	2.0	2.0	1.0	1.0	F	2.0	2.0	2.0	1.0	5
1	2.0	2.0	2.0	2.0	F	2.0	2.0	2.0	1.0	2
2	2.0	1.0	1.0	1.0	F	2.0	2.0	2.0	1.0	2
3	2.0	2.0	2.0	2.0	F	2.0	2.0	2.0	2.0	4
4	2.0	1.0	1.0	1.0	F	1.0	2.0	2.0	1.0	4
5	2.0	1.0	1.0	1.0	F	2.0	2.0	2.0	1.0	3
6	2.0	2.0	2.0	2.0	F	2.0	2.0	2.0	2.0	2
7	2.0	2.0	2.0	2.0	F	2.0	2.0	2.0	2.0	1
8	2.0	2.0	2.0	2.0	F	2.0	2.0	2.0	1.0	0
9	2.0	2.0	2.0	2.0	F	2.0	2.0	2.0	2.0	3
10	2.0	2.0	2.0	2.0	F	2.0	2.0	2.0	2.0	4
11	2.0	1.0	1.0	2.0	F	1.0	2.0	2.0	1.0	2
12	2.0	1.0	1.0	2.0	F	1.0	2.0	2.0	1.0	4
13	2.0	2.0	2.0	2.0	M	2.0	2.0	2.0	2.0	3
14	2.0	2.0	2.0	2.0	F	2.0	2.0	2.0	2.0	3

	REL_DESCO	REL_CONJ	OUT_VEZES	VIOL_PSICO	CS_SEXO	AUTOR_ALCO	REL_CONHEC	REL_EXCON	AUTOR_SEXO	FAIXA_ETARIA
15	2.0	2.0	2.0	2.0	M	2.0	2.0	2.0	2.0	5
16	2.0	1.0	1.0	2.0	F	2.0	2.0	2.0	1.0	2
17	2.0	2.0	2.0	2.0	F	2.0	2.0	2.0	2.0	3
18	2.0	2.0	2.0	2.0	F	2.0	2.0	2.0	2.0	0
19	2.0	2.0	2.0	2.0	F	2.0	1.0	2.0	1.0	3
20	2.0	1.0	1.0	1.0	F	1.0	2.0	2.0	1.0	2
21	2.0	2.0	2.0	1.0	M	2.0	1.0	2.0	1.0	1
22	2.0	2.0	2.0	2.0	M	2.0	1.0	2.0	1.0	0
23	2.0	2.0	1.0	1.0	F	2.0	2.0	2.0	2.0	2
24	2.0	2.0	2.0	2.0	F	2.0	2.0	2.0	2.0	4
25	2.0	2.0	2.0	1.0	F	2.0	1.0	2.0	1.0	1
26	2.0	2.0	2.0	2.0	M	2.0	2.0	2.0	2.0	3
27	2.0	2.0	2.0	2.0	M	2.0	2.0	2.0	2.0	2

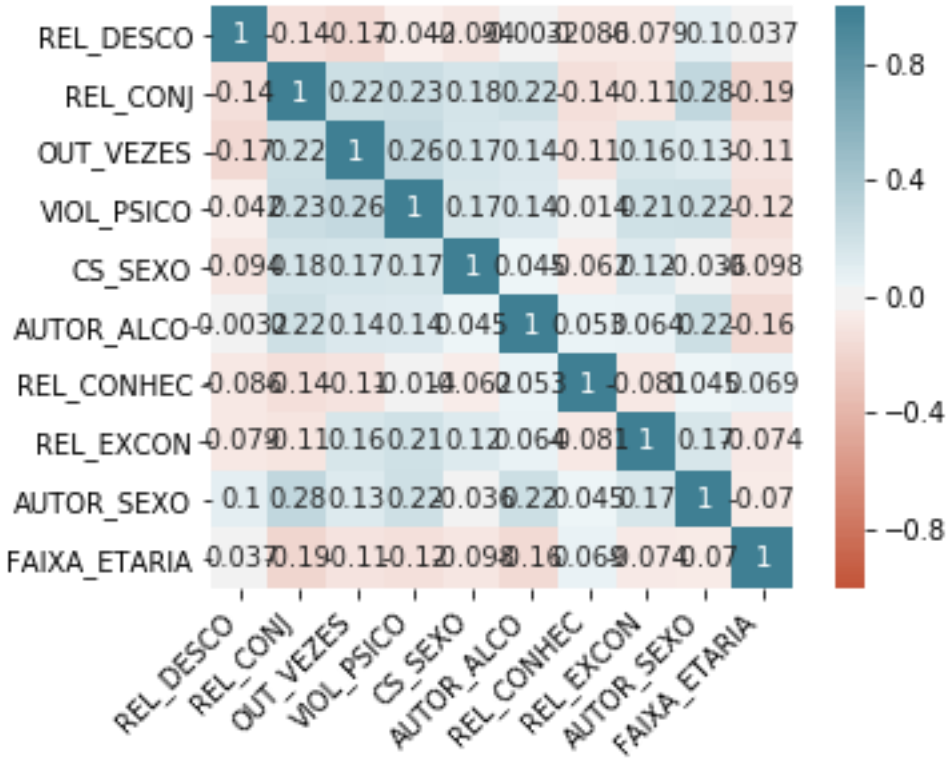
	REL_DESCO	REL_CONJ	OUT_VEZES	VIOL_PSICO	CS_SEXO	AUTOR_ALCO	REL_CONHEC	REL_EXCON	AUTOR_SEXO	FAIXA_ETARIA
28	2.0	2.0	2.0	2.0	M	2.0	1.0	2.0	1.0	1
29	2.0	2.0	2.0	2.0	F	2.0	2.0	2.0	2.0	2
30	2.0	2.0	1.0	2.0	M	2.0	2.0	2.0	1.0	5
31	2.0	2.0	2.0	2.0	M	2.0	2.0	2.0	2.0	3
32	2.0	1.0	2.0	2.0	F	2.0	2.0	2.0	2.0	4
33	2.0	2.0	2.0	2.0	F	2.0	1.0	2.0	1.0	0
34	2.0	1.0	1.0	2.0	M	2.0	2.0	2.0	3.0	5
35	1.0	2.0	2.0	1.0	F	2.0	1.0	2.0	1.0	0
36	2.0	2.0	1.0	2.0	M	2.0	1.0	2.0	1.0	1
37	1.0	2.0	2.0	2.0	F	1.0	2.0	2.0	1.0	2
38	2.0	2.0	2.0	2.0	F	2.0	2.0	2.0	3.0	0
39	2.0	2.0	2.0	1.0	F	2.0	2.0	1.0	2.0	3
40	1.0	2.0	2.0	2.0	F	2.0	2.0	2.0	1.0	1

	REL_DESCO	REL_CONJ	OUT_VEZES	VIOL_PSICO	CS_SEXO	AUTOR_ALCO	REL_CONHEC	REL_EXCON	AUTOR_SEXO	FAIXA_ETARIA
41	2.0	2.0	2.0	1.0	F	2.0	2.0	2.0	1.0	1
42	2.0	2.0	2.0	2.0	F	2.0	2.0	2.0	1.0	0
43	2.0	2.0	2.0	2.0	F	2.0	1.0	2.0	1.0	3
44	2.0	2.0	2.0	2.0	F	1.0	2.0	2.0	1.0	4
45	2.0	2.0	1.0	1.0	F	2.0	2.0	1.0	1.0	3
46	2.0	1.0	2.0	1.0	F	1.0	2.0	2.0	1.0	2
47	2.0	2.0	2.0	2.0	F	2.0	1.0	2.0	2.0	5
48	2.0	2.0	2.0	2.0	M	2.0	2.0	2.0	1.0	2
49	2.0	2.0	1.0	2.0	F	1.0	2.0	2.0	1.0	0

In [9]:

```
df1=df1.apply(encoding)
df1=df1.dropna()
#df1=df1.fillna(df1.mean())
#df1 = df1[df1.REL_CONJ != 9]
#df1 = df1[df1.REL_EXCON != 9]

#import pandas.rpy.common as com
import seaborn as sns

corr = df1.corr()
ax = sns.heatmap(
    corr,
```

```
    vmin=-1, vmax=1, center=0,
    cmap=sns.diverging_palette(20, 220, n=200),
    square=True, annot=True
)
ax.set_xticklabels(
    ax.get_xticklabels(),
    rotation=45,
    horizontalalignment='right'
);
```



In [10]:

```
df1["OUT_VEZES"]=df1["OUT_VEZES"].astype(int)
X = df1.drop(['OUT_VEZES'], axis = 1)
y = df1["OUT_VEZES"]
df1.head(100)
```

Out[10]:

	REL_DESCO	REL_CONJ	OUT_VEZES	VIOL_PSICO	CS_SEXO	AUTOR_ALCO	REL_CONHEC	REL_EXCON	AUTOR_SEXO	FAIXA_ETARIA
0	2.0	2.0	1	1.0	0	2.0	2.0	2.0	1.0	5
1	2.0	2.0	2	2.0	0	2.0	2.0	2.0	1.0	2

	REL_DESCO	REL_CONJ	OUT_VEZES	VIOL_PSICO	CS_SEXO	AUTOR_ALCO	REL_CONHEC	REL_EXCON	AUTOR_SEXO	FAIXA_ETARIA
2	2.0	1.0	1	1.0	0	2.0	2.0	2.0	1.0	2
3	2.0	2.0	2	2.0	0	2.0	2.0	2.0	2.0	4
4	2.0	1.0	1	1.0	0	1.0	2.0	2.0	1.0	4
5	2.0	1.0	1	1.0	0	2.0	2.0	2.0	1.0	3
6	2.0	2.0	2	2.0	0	2.0	2.0	2.0	2.0	2
7	2.0	2.0	2	2.0	0	2.0	2.0	2.0	2.0	1
8	2.0	2.0	2	2.0	0	2.0	2.0	2.0	1.0	0
9	2.0	2.0	2	2.0	0	2.0	2.0	2.0	2.0	3
10	2.0	2.0	2	2.0	0	2.0	2.0	2.0	2.0	4
11	2.0	1.0	1	2.0	0	1.0	2.0	2.0	1.0	2
12	2.0	1.0	1	2.0	0	1.0	2.0	2.0	1.0	4
13	2.0	2.0	2	2.0	2	2.0	2.0	2.0	2.0	3
14	2.0	2.0	2	2.0	0	2.0	2.0	2.0	2.0	3
15	2.0	2.0	2	2.0	2	2.0	2.0	2.0	2.0	5
16	2.0	1.0	1	2.0	0	2.0	2.0	2.0	1.0	2

	REL_DESCO	REL_CONJ	OUT_VEZES	VIOL_PSICO	CS_SEXO	AUTOR_ALCO	REL_CONHEC	REL_EXCON	AUTOR_SEXO	FAIXA_ETARIA
17	2.0	2.0	2	2.0	0	2.0	2.0	2.0	2.0	3
18	2.0	2.0	2	2.0	0	2.0	2.0	2.0	2.0	0
19	2.0	2.0	2	2.0	0	2.0	1.0	2.0	1.0	3
20	2.0	1.0	1	1.0	0	1.0	2.0	2.0	1.0	2
21	2.0	2.0	2	1.0	2	2.0	1.0	2.0	1.0	1
22	2.0	2.0	2	2.0	2	2.0	1.0	2.0	1.0	0
23	2.0	2.0	1	1.0	0	2.0	2.0	2.0	2.0	2
24	2.0	2.0	2	2.0	0	2.0	2.0	2.0	2.0	4
25	2.0	2.0	2	1.0	0	2.0	1.0	2.0	1.0	1
26	2.0	2.0	2	2.0	2	2.0	2.0	2.0	2.0	3
27	2.0	2.0	2	2.0	2	2.0	2.0	2.0	2.0	2
28	2.0	2.0	2	2.0	2	2.0	1.0	2.0	1.0	1
29	2.0	2.0	2	2.0	0	2.0	2.0	2.0	2.0	2

	REL_DESCO	REL_CONJ	OUT_VEZES	VIOL_PSICO	CS_SEXO	AUTOR_ALCO	REL_CONHEC	REL_EXCON	AUTOR_SEXO	FAIXA_ETARIA
...	...	...	...	...	...	...	...	...	...	...
70	2.0	2.0	1	2.0	2	2.0	2.0	2.0	2.0	1
71	2.0	2.0	2	2.0	0	2.0	2.0	2.0	2.0	1
72	2.0	2.0	2	2.0	2	2.0	2.0	2.0	2.0	2
73	2.0	1.0	1	1.0	0	2.0	2.0	2.0	1.0	2
74	2.0	2.0	2	1.0	0	1.0	2.0	2.0	1.0	1
75	2.0	2.0	2	2.0	2	2.0	2.0	2.0	2.0	2
76	2.0	2.0	2	1.0	2	2.0	2.0	2.0	1.0	1
77	2.0	1.0	1	1.0	0	1.0	2.0	2.0	1.0	2
78	2.0	1.0	1	1.0	0	2.0	2.0	2.0	1.0	2
79	2.0	2.0	2	2.0	2	2.0	2.0	2.0	1.0	0
80	2.0	2.0	2	2.0	2	2.0	2.0	2.0	2.0	2
81	2.0	2.0	2	2.0	0	2.0	2.0	2.0	2.0	3

	REL_DESCO	REL_CONJ	OUT_VEZES	VIOL_PSICO	CS_SEXO	AUTOR_ALCO	REL_CONHEC	REL_EXCON	AUTOR_SEXO	FAIXA_ETARIA
82	2.0	2.0	2	1.0	0	2.0	2.0	2.0	1.0	3
83	2.0	1.0	1	2.0	0	1.0	2.0	2.0	1.0	4
84	2.0	2.0	1	2.0	0	2.0	2.0	2.0	1.0	0
85	1.0	2.0	2	2.0	0	1.0	2.0	2.0	1.0	5
86	2.0	2.0	2	2.0	0	2.0	2.0	2.0	2.0	4
87	2.0	2.0	2	2.0	2	2.0	2.0	2.0	2.0	1
88	2.0	2.0	2	2.0	2	2.0	2.0	2.0	2.0	3
89	2.0	2.0	2	2.0	0	2.0	2.0	2.0	2.0	2
90	2.0	2.0	2	2.0	0	2.0	2.0	2.0	2.0	2
91	2.0	1.0	2	1.0	0	2.0	2.0	2.0	1.0	3
92	2.0	2.0	2	2.0	0	2.0	1.0	2.0	2.0	2
93	2.0	2.0	1	2.0	2	2.0	2.0	2.0	1.0	1
94	2.0	2.0	1	1.0	0	1.0	2.0	1.0	1.0	1

	REL_DESCO	REL_CONJ	OUT_VEZES	VIOL_PSICO	CS_SEXO	AUTOR_ALCO	REL_CONHEC	REL_EXCON	AUTOR_SEXO	FAIXA_ETARIA
95	1.0	2.0	2	2.0	0	2.0	2.0	2.0	1.0	3
96	2.0	1.0	1	1.0	0	1.0	2.0	2.0	1.0	1
97	2.0	2.0	2	2.0	0	2.0	2.0	1.0	1.0	3
98	2.0	1.0	1	2.0	0	1.0	2.0	2.0	1.0	3
99	2.0	2.0	2	2.0	2	2.0	2.0	2.0	2.0	2

100 rows × 10 columns

In [11]:

```
from sklearn.model_selection import train_test_split
X_train, X_test, y_train, y_test = train_test_split(X, y,test_size = .33, ran
dom_state=0)
```

In [12]:

```
y.sample(5)
```

Out[12]:

```
5109     1
17388    2
20799    1
18262    2
25203    1
Name: OUT_VEZES, dtype: int32
```

In [13]:

```
headers = X_train.columns

X_train.head()
```

Out[13]:

	REL_DESCO	REL_CONJ	VIOL_PSICO	CS_SEXO	AUTOR_ALCO	REL_CONHEC	REL_EXCON	AUTOR_SEXO	FAIXA_ETARIA
19646	2.0	2.0	2.0	0	2.0	2.0	2.0	2.0	1
6871	2.0	2.0	1.0	0	1.0	1.0	2.0	1.0	2
33275	2.0	2.0	2.0	0	2.0	2.0	2.0	2.0	2
17003	2.0	2.0	2.0	2	2.0	2.0	2.0	2.0	2
7342	2.0	2.0	1.0	0	1.0	2.0	1.0	1.0	5

In [14]:

```
from sklearn.preprocessing import StandardScaler
sc = StandardScaler()

## transforming "train_x"
X_train = sc.fit_transform(X_train)
## transforming "test_x"
X_test = sc.transform(X_test)

## transforming "The testset"
#test = sc.transform(test)
```

In [15]:

```
pd.DataFrame(X_train, columns=headers).head(40)
```

Out[15]:

	REL_DESCO	REL_CONJ	VIOL_PSICO	CS_SEXO	AUTOR_ALCO	REL_CONHEC	REL_EXCON	AUTOR_SEXO	FAIXA_ETARIA
0	0.320247	0.444606	0.524406	-0.577474	0.557140	0.336732	0.259489	0.981732	-0.886793
1	0.320247	0.444606	-1.906921	-0.577474	-1.794881	-2.969718	0.259489	-0.806333	-0.208723

	REL_DESCO	REL_CONJ	VIOL_PSICO	CS_SEXO	AUTOR_ALCO	REL_CONHEC	REL_EXCON	AUTOR_SEXO	FAIXA_ETARIA
2	0.320247	0.444606	0.524406	-0.577474	0.557140	0.336732	0.259489	0.981732	-0.208723
3	0.320247	0.444606	0.524406	1.732516	0.557140	0.336732	0.259489	0.981732	-0.208723
4	0.320247	0.444606	-1.906921	-0.577474	-1.794881	0.336732	-3.853727	-0.806333	1.825488
5	0.320247	-2.249183	0.524406	1.732516	-1.794881	0.336732	0.259489	0.981732	1.147417
6	0.320247	0.444606	0.524406	1.732516	-1.794881	-2.969718	0.259489	-0.806333	0.469347
7	0.320247	0.444606	0.524406	-0.577474	0.557140	-2.969718	0.259489	-0.806333	-1.564864
8	0.320247	0.444606	0.524406	-0.577474	0.557140	0.336732	0.259489	0.981732	1.147417
9	0.320247	-2.249183	0.524406	-0.577474	0.557140	0.336732	0.259489	-0.806333	1.147417
10	-3.122586	0.444606	-1.906921	-0.577474	0.557140	0.336732	0.259489	-0.806333	-0.208723
11	0.320247	0.444606	0.524406	-0.577474	0.557140	0.336732	0.259489	-0.806333	-1.564864
12	0.320247	0.444606	0.524406	1.732516	0.557140	0.336732	0.259489	0.981732	-0.208723

	REL_DESCO	REL_CONJ	VIOL_PSICO	CS_SEXO	AUTOR_ALCO	REL_CONHEC	REL_EXCON	AUTOR_SEXO	FAIXA_ETARIA
13	0.320247	0.444606	0.524406	-0.577474	0.557140	0.336732	0.259489	0.981732	1.147417
14	0.320247	0.444606	0.524406	-0.577474	0.557140	0.336732	0.259489	-0.806333	-0.208723
15	0.320247	0.444606	0.524406	1.732516	0.557140	-2.969718	0.259489	-0.806333	1.147417
16	0.320247	-2.249183	0.524406	-0.577474	0.557140	0.336732	0.259489	-0.806333	0.469347
17	0.320247	-2.249183	0.524406	-0.577474	0.557140	0.336732	0.259489	-0.806333	-0.208723
18	0.320247	0.444606	-1.906921	-0.577474	0.557140	0.336732	0.259489	0.981732	1.825488
19	0.320247	0.444606	0.524406	-0.577474	0.557140	0.336732	0.259489	0.981732	-0.208723
20	-3.122586	0.444606	0.524406	-0.577474	0.557140	0.336732	0.259489	0.981732	1.147417
21	0.320247	-2.249183	0.524406	-0.577474	-1.794881	0.336732	0.259489	-0.806333	-0.208723
22	0.320247	0.444606	0.524406	1.732516	0.557140	0.336732	0.259489	0.981732	-1.564864
23	0.320247	-2.249183	0.524406	-0.577474	-1.794881	0.336732	0.259489	-0.806333	0.469347

	REL_DESCO	REL_CONJ	VIOL_PSICO	CS_SEXO	AUTOR_ALCO	REL_CONHEC	REL_EXCON	AUTOR_SEXO	FAIXA_ETARIA
24	0.320247	0.444606	-1.906921	1.732516	0.557140	0.336732	0.259489	-0.806333	1.825488
25	0.320247	-2.249183	-1.906921	-0.577474	-1.794881	0.336732	0.259489	-0.806333	0.469347
26	0.320247	0.444606	0.524406	-0.577474	0.557140	0.336732	0.259489	2.769798	-1.564864
27	0.320247	-2.249183	0.524406	1.732516	-1.794881	0.336732	0.259489	0.981732	0.469347
28	0.320247	0.444606	-1.906921	-0.577474	-1.794881	0.336732	-3.853727	-0.806333	0.469347
29	0.320247	0.444606	0.524406	-0.577474	0.557140	-2.969718	0.259489	0.981732	-0.886793
30	0.320247	0.444606	0.524406	-0.577474	0.557140	0.336732	0.259489	0.981732	0.469347
31	0.320247	0.444606	0.524406	-0.577474	-1.794881	-2.969718	0.259489	0.981732	1.825488
32	0.320247	-2.249183	0.524406	-0.577474	-1.794881	0.336732	0.259489	-0.806333	1.147417
33	0.320247	0.444606	-1.906921	-0.577474	0.557140	0.336732	-3.853727	-0.806333	-0.208723
34	0.320247	0.444606	-1.906921	-0.577474	0.557140	-2.969718	0.259489	-0.806333	-0.886793

	REL_DESCO	REL_CONJ	VIOL_PSICO	CS_SEXO	AUTOR_ALCO	REL_CONHEC	REL_EXCON	AUTOR_SEXO	FAIXA_ETARIA
35	0.320247	0.444606	0.524406	-0.577474	0.557140	0.336732	0.259489	0.981732	-1.564864
36	0.320247	0.444606	0.524406	-0.577474	-1.794881	0.336732	0.259489	0.981732	-0.208723
37	0.320247	-2.249183	-1.906921	-0.577474	-1.794881	0.336732	0.259489	-0.806333	1.147417
38	-3.122586	0.444606	0.524406	-0.577474	0.557140	0.336732	0.259489	-0.806333	-0.208723
39	0.320247	-2.249183	0.524406	-0.577474	-1.794881	0.336732	0.259489	-0.806333	-0.208723

In [17]:

```
from sklearn.linear_model import LogisticRegression
from sklearn.metrics import mean_absolute_error, accuracy_score,classificatio
n_report, confusion_matrix

## call on the model object
logreg = LogisticRegression(solver='liblinear')

## fit the model with "train_x" and "train_y"
logreg.fit(X_train,y_train)

## Once the model is trained we want to find out how well the model is perfor
ming, so we test the model.
## we use "test_x" portion of the data(this data was not used to fit the mode
l) to predict model outcome.
y_pred = logreg.predict(X_test)

## Once predicted we save that outcome in "y_pred" variable.
## Then we compare the predicted value( "y_pred") and actual value("test_y")
to see how well our model is performing.
print(confusion_matrix(y_pred, y_test))
print(classification_report(y_pred, y_test,digits=4))
```

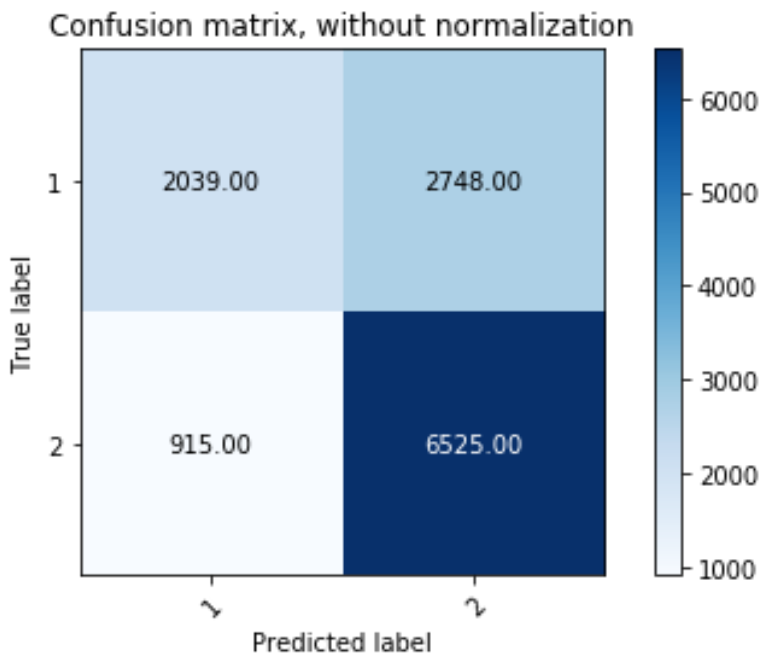
```
print ("So, Our accuracy Score is: {}".format(round(accuracy_score(y_pred, y_
test),4)))
```

```
[[2218 1189]
 [2569 6251]]
              precision    recall  f1-score   support

           1     0.4633    0.6510    0.5414      3407
           2     0.8402    0.7087    0.7689      8820

    accuracy                         0.6926     12227
   macro avg     0.6518    0.6799    0.6551     12227
weighted avg     0.7352    0.6926    0.7055     12227

So, Our accuracy Score is: 0.6926
```

In [19]:

```
from sklearn.utils.multiclass import unique_labels
from sklearn.metrics import confusion_matrix
import numpy as np


def plot_confusion_matrix(y_true, y_pred, classes,
                          normalize=True,
                          title=None,
                          cmap=plt.cm.Blues):
    """
    This function prints and plots the confusion matrix.
    Normalization can be applied by setting `normalize=True`.
    """
    if not title:
        if normalize:
            title = 'Normalized confusion matrix'
        else:
            title = 'Confusion matrix, without normalization'

    # Compute confusion matrix
    cm = confusion_matrix(y_true, y_pred)


    fig, ax = plt.subplots()
    im = ax.imshow(cm, interpolation='nearest', cmap=cmap)
    ax.figure.colorbar(im, ax=ax)
    # We want to show all ticks...
    ax.set(xticks=np.arange(cm.shape[1]),
```

```
           yticks=np.arange(cm.shape[0]),
           # ... and label them with the respective list entries
           xticklabels=classes, yticklabels=classes,
           title=title,
           ylabel='True label',
           xlabel='Predicted label')

    # Rotate the tick labels and set their alignment.
    plt.setp(ax.get_xticklabels(), rotation=45, ha="right",
             rotation_mode="anchor")

    # Loop over data dimensions and create text annotations.
    fmt = '.2f' if normalize else 'd'
    thresh = cm.max() / 2.
    for i in range(cm.shape[0]):
        for j in range(cm.shape[1]):
            ax.text(j, i, format(cm[i, j], fmt),
                    ha="center", va="center",
                    color="white" if cm[i, j] > thresh else "black")
    fig.tight_layout()
    return ax


np.set_printoptions(precision=2)


class_names = np.array(['1','2'])

# Plot non-normalized confusion matrix
plot_confusion_matrix(y_test, y_pred, classes=class_names,
                      title='Confusion matrix, without normalization')

# Plot normalized confusion matrix
plot_confusion_matrix(y_test, y_pred, classes=class_names, normalize=True,
                      title='Normalized confusion matrix')

plt.show()
```



In [50]:

```
# Random Forest
from sklearn.ensemble import RandomForestClassifier

randomforest = RandomForestClassifier()
randomforest.fit(X_train, y_train)
y_pred = randomforest.predict(X_test)
acc_randomforest = round(accuracy_score(y_pred, y_test) * 100, 2)
print(confusion_matrix(y_pred, y_test))
print(classification_report(y_pred, y_test,digits=4))
print(acc_randomforest)
```

```
[[2044 1040]
 [2743 6400]]
             precision    recall  f1-score   support

          1     0.4270    0.6628    0.5194      3084
          2     0.8602    0.7000    0.7719      9143

avg / total     0.7509    0.6906    0.7082     12227

69.06
```

In [130]:

```
pd.DataFrame(y_train, columns=headers).head()
```

Out[130]:

REL_DESCO	REL_CONJ	VIOL_PSICO	CS_SEXO	AUTOR_ALCO	REL_CONHEC	REL_EXCON	AUTOR_SEXO	FAIXA_ETARIA

In [51]:

```
# Gradient Boosting Classifier
from sklearn.ensemble import GradientBoostingClassifier

gbk = GradientBoostingClassifier()
gbk.fit(X_train, y_train)
y_pred = gbk.predict(X_test)
acc_gbk = round(accuracy_score(y_pred, y_test) * 100, 2)
print(confusion_matrix(y_pred, y_test))
print(classification_report(y_pred, y_test,digits=4))
print(acc_gbk)
```

```
[[2039  928]
 [2748 6512]]
             precision    recall  f1-score   support

          1     0.4259    0.6872    0.5259      2967
          2     0.8753    0.7032    0.7799      9260

avg / total     0.7662    0.6994    0.7183     12227

[[2039  928]
 [2748 6512]]
             precision    recall  f1-score   support

          1     0.4259    0.6872    0.5259      2967
          2     0.8753    0.7032    0.7799      9260

avg / total     0.7662    0.6994    0.7183     12227

69.94
```

In [60]:

```
# Support Vector Machines
from sklearn.svm import SVC

svc = SVC()
svc.fit(X_train, y_train)
```

```
y_pred = svc.predict(X_test)
acc_svc = round(accuracy_score(y_pred, y_test) * 100, 2)
print(confusion_matrix(y_pred, y_test))
print(classification_report(y_pred, y_test,digits=4))
print(acc_svc)
```

```
[[1934  804]
 [2853 6636]]
             precision    recall  f1-score   support

          1     0.4040    0.7064    0.5140      2738
          2     0.8919    0.6993    0.7840      9489

avg / total     0.7827    0.7009    0.7235     12227

70.09
```

In [260]:

```
# Stochastic Gradient Descent
from sklearn.linear_model import SGDClassifier

sgd = SGDClassifier()
sgd.fit(X_train, y_train)
y_pred = sgd.predict(X_test)
acc_sgd = round(accuracy_score(y_pred, y_test) * 100, 2)
print(acc_sgd)
```

```
60.74
C:\Users\Eduardo Duate\Anaconda3\lib\site-packages\sklearn\linear_model\stoch
astic_gradient.py:128: FutureWarning: max_iter and tol parameters have been a
dded in <class 'sklearn.linear_model.stochastic_gradient.SGDClassifier'> in 0
.19. If both are left unset, they default to max_iter=5 and tol=None. If tol
is not None, max_iter defaults to max_iter=1000. From 0.21, default max_iter
will be 1000, and default tol will be 1e-3.
  "and default tol will be 1e-3." % type(self), FutureWarning)
```

In [49]:

```
# Perceptron
from sklearn.linear_model import Perceptron

perceptron = Perceptron()
perceptron.fit(X_train, y_train)
y_pred = perceptron.predict(X_test)
acc_perceptron = round(accuracy_score(y_pred, y_test) * 100, 2)
print(confusion_matrix(y_pred, y_test))
print(classification_report(y_pred, y_test,digits=4))
print(acc_perceptron)
```

```
[[1997 1022]
 [2790 6418]]
             precision    recall  f1-score   support

          1     0.4172    0.6615    0.5117      3019
          2     0.8626    0.6970    0.7710      9208

avg / total     0.7526    0.6882    0.7070     12227

68.82
C:\Users\Eduardo Duate\Anaconda3\lib\site-packages\sklearn\linear_model\stoch
astic_gradient.py:128: FutureWarning: max_iter and tol parameters have been a
dded in <class 'sklearn.linear_model.perceptron.Perceptron'> in 0.19. If both
are left unset, they default to max_iter=5 and tol=None. If tol is not None,
max_iter defaults to max_iter=1000. From 0.21, default max_iter will be 1000,
and default tol will be 1e-3.
  "and default tol will be 1e-3." % type(self), FutureWarning)
```

In [58]:

```
from sklearn.ensemble import ExtraTreesClassifier
ExtraTreesClassifier = ExtraTreesClassifier()
ExtraTreesClassifier.fit(X_train, y_train)
y_pred = ExtraTreesClassifier.predict(X_test)
extraTree_accy = round(accuracy_score(y_pred, y_test), 3)
print(confusion_matrix(y_pred, y_test))
print(classification_report(y_pred, y_test,digits=4))
print(extraTree_accy)
```

```
[[2084 1086]
 [2703 6354]]
             precision    recall  f1-score   support

          1     0.4353    0.6574    0.5238      3170
          2     0.8540    0.7016    0.7703      9057

avg / total     0.7455    0.6901    0.7064     12227

0.69
```

In [195]:

```
from sklearn.gaussian_process import GaussianProcessClassifier
GaussianProcessClassifier = GaussianProcessClassifier()
GaussianProcessClassifier.fit(X_train, y_train)
y_pred = GaussianProcessClassifier.predict(X_test)
gau_pro_accy = round(accuracy_score(y_pred, y_test), 3)
print(gau_pro_accy)
```

```
---------------------------------------------------------------------------
MemoryError                               Traceback (most recent call last)
<ipython-input-195-6722f851e430> in <module>()
      1 from sklearn.gaussian_process import GaussianProcessClassifier
      2 GaussianProcessClassifier = GaussianProcessClassifier()
----> 3 GaussianProcessClassifier.fit(X_train, y_train)
      4 y_pred = GaussianProcessClassifier.predict(X_test)
      5 gau_pro_accy = round(accuracy_score(y_pred, y_test), 3)

~\Anaconda3\lib\site-packages\sklearn\gaussian_process\gpc.py in fit(self, X,
y)
    611                                 % self.multi_class)
    612
--> 613         self.base_estimator_.fit(X, y)
    614
    615         if self.n_classes_ > 2:

~\Anaconda3\lib\site-packages\sklearn\gaussian_process\gpc.py in fit(self, X,
y)
    230         else:
    231             self.log_marginal_likelihood_value_ = \
--> 232                 self.log_marginal_likelihood(self.kernel_.theta)
    233
    234         # Precompute quantities required for predictions which are in
dependent

~\Anaconda3\lib\site-packages\sklearn\gaussian_process\gpc.py in log_marginal
_likelihood(self, theta, eval_gradient)
    338             K, K_gradient = kernel(self.X_train_, eval_gradient=True)
    339         else:
--> 340             K = kernel(self.X_train_)
    341
    342         # Compute log-marginal-likelihood Z and also store some tempo
raries

~\Anaconda3\lib\site-packages\sklearn\gaussian_process\kernels.py in __call__
(self, X, Y, eval_gradient)
    756                                     K2_gradient * K1[:, :, np.newa
xis]))
    757         else:
--> 758             return self.k1(X, Y) * self.k2(X, Y)
    759
    760     def diag(self, X):
```

```
~\Anaconda3\lib\site-packages\sklearn\gaussian_process\kernels.py in __call__
(self, X, Y, eval_gradient)
   1204         if Y is None:
   1205             dists = pdist(X / length_scale, metric='sqeuclidean')
-> 1206             K = np.exp(-.5 * dists)
   1207             # convert from upper-triangular matrix to square matrix
   1208             K = squareform(K)

MemoryError:
```

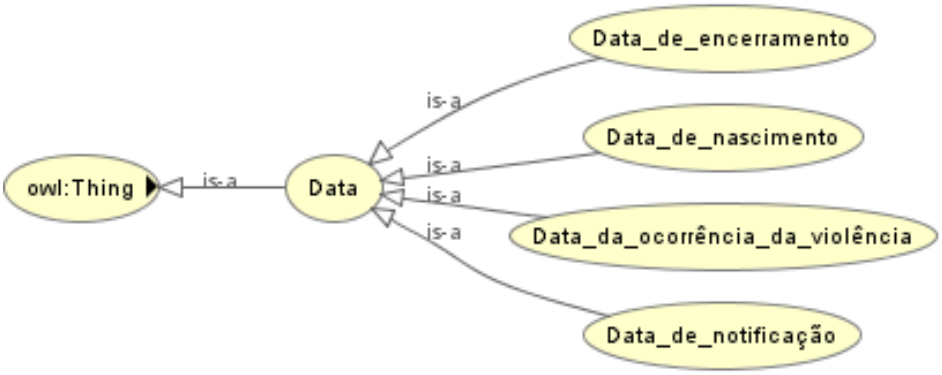
In [18]:

```
##!pip install xgboost
from xgboost import XGBClassifier
XGBClassifier = XGBClassifier()
XGBClassifier.fit(X_train, y_train)
y_pred = XGBClassifier.predict(X_test)
print(confusion_matrix(y_pred, y_test))
print(classification_report(y_pred, y_test,digits=4))
XGBClassifier_accy = round(accuracy_score(y_pred, y_test), 3)
print(XGBClassifier_accy)
```

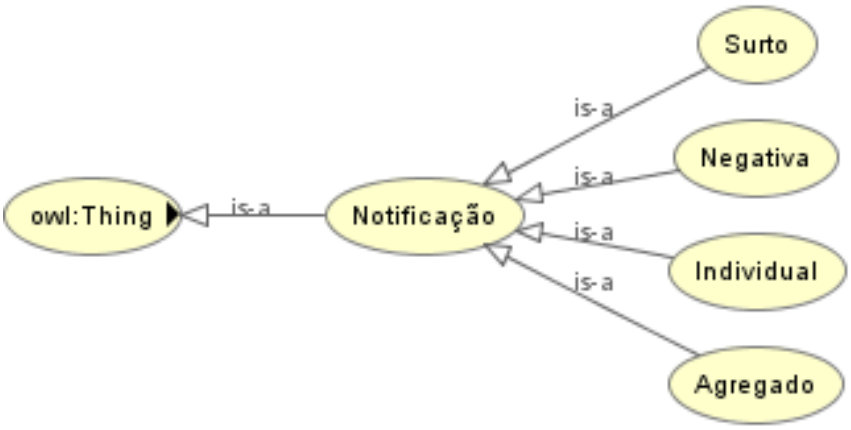
```
[[2039  915]
 [2748 6525]]
              precision    recall  f1-score   support

           1     0.4259    0.6903    0.5268      2954
           2     0.8770    0.7037    0.7808      9273

    accuracy                         0.7004     12227
   macro avg     0.6515    0.6970    0.6538     12227
weighted avg     0.7680    0.7004    0.7195     12227

0.7
```

ANEXO C – Classes principais da Ontologia desenvolvida no Protégé

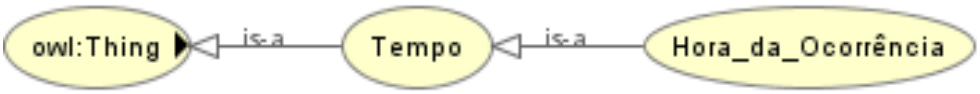
Classe Data



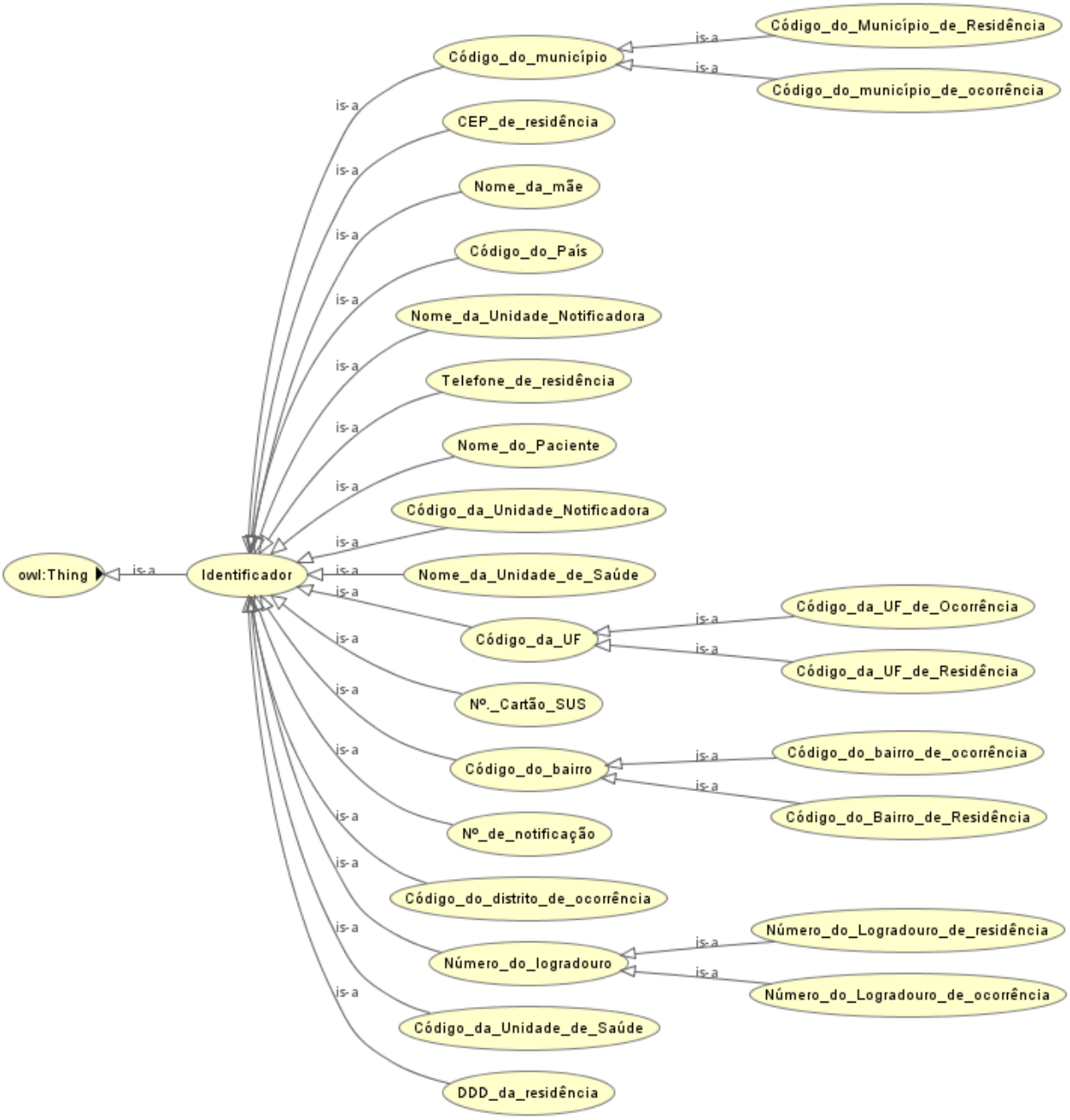
Classe Notificação



Classe Tempo

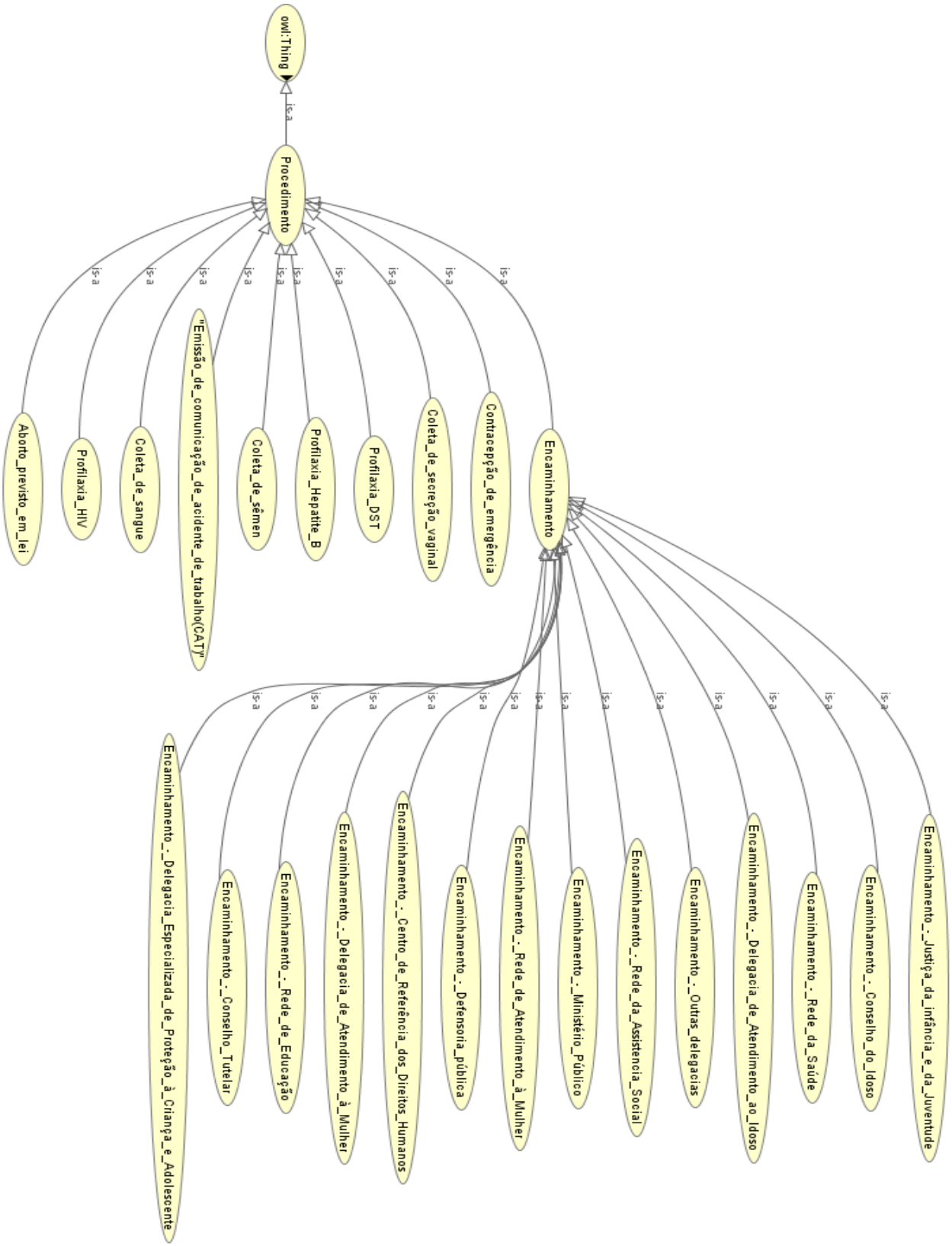


Classe Identificador

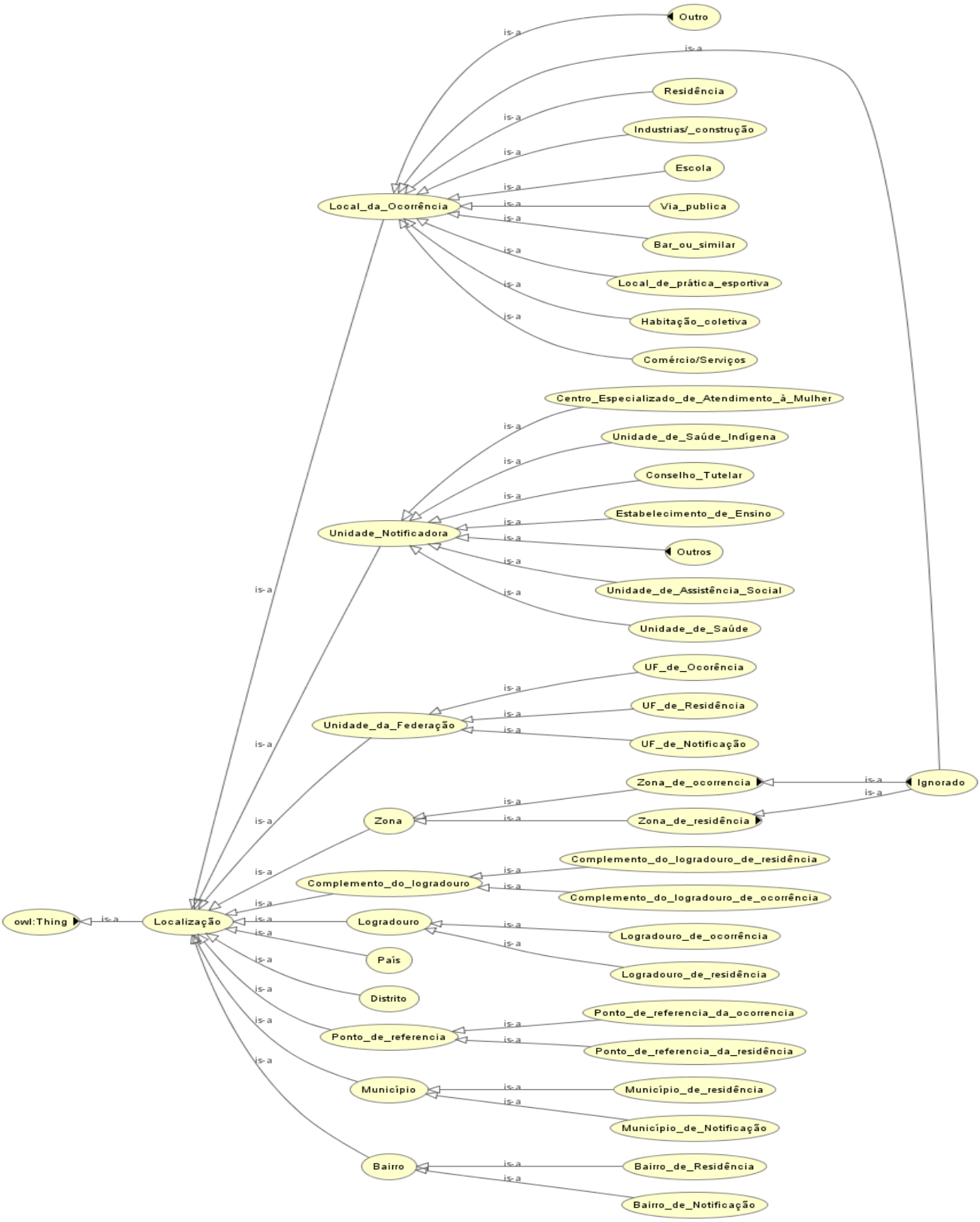


Classe Pessoa

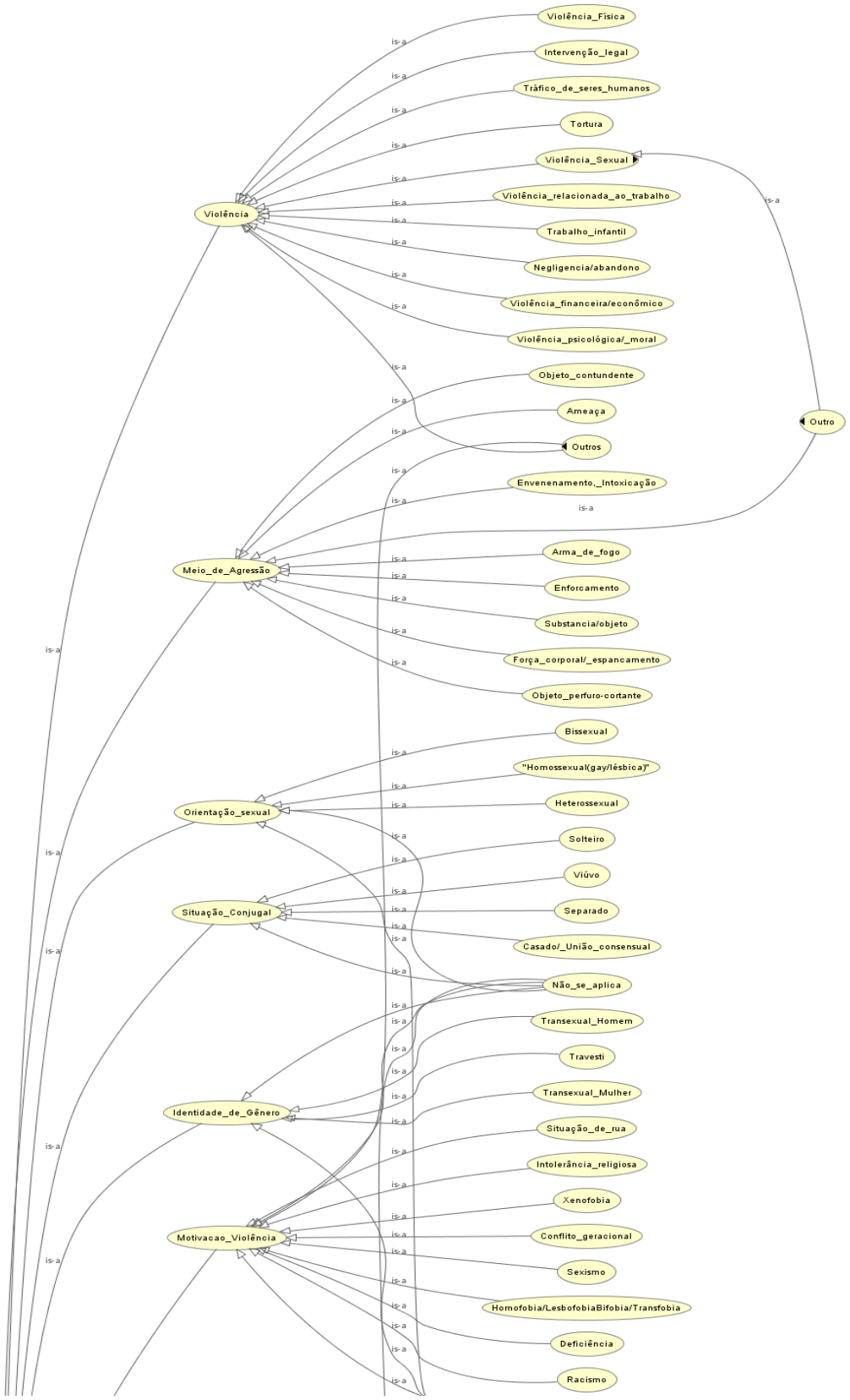
Classe Procedimento

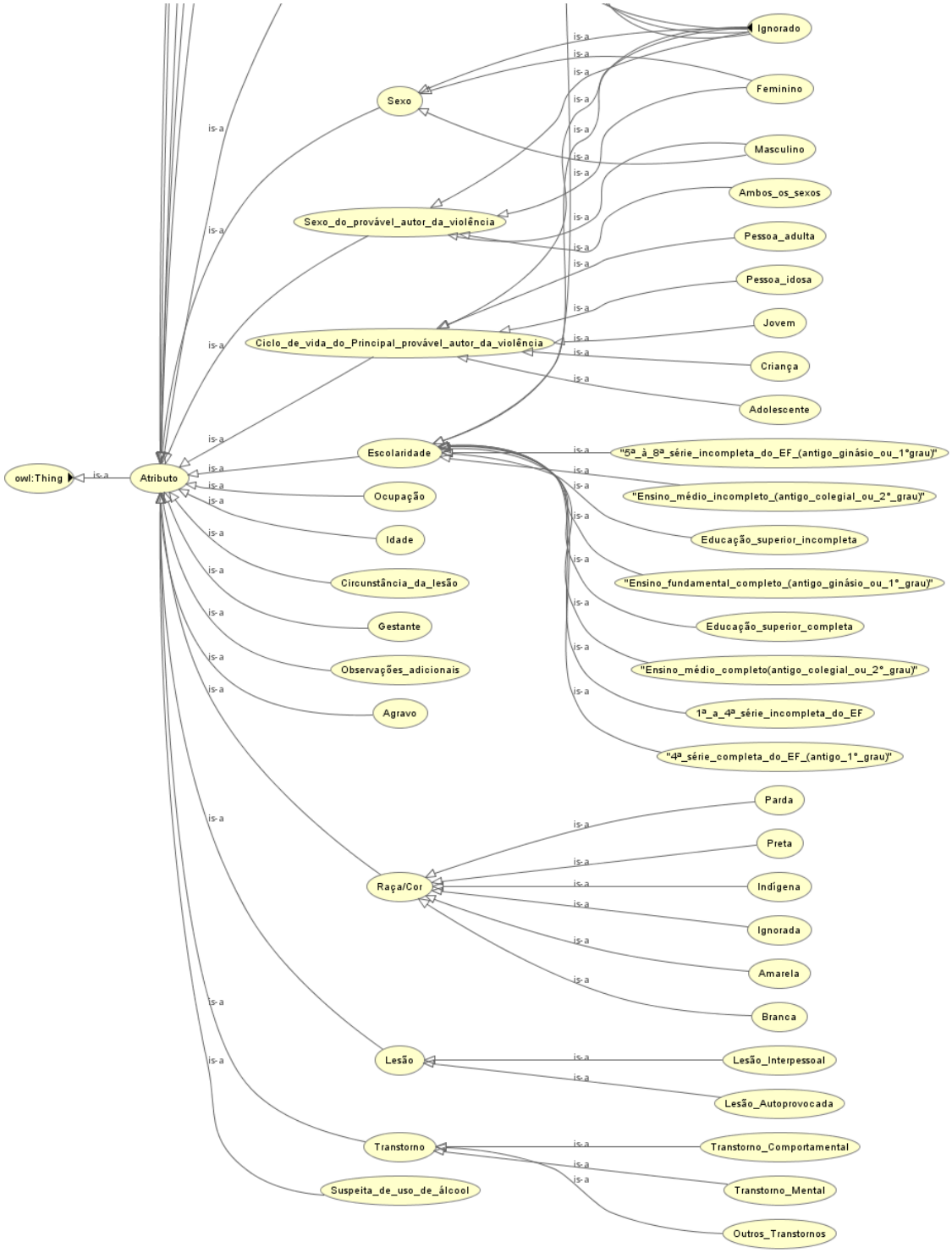


Classe Localização



Classe Atributo




owl:Thing
Atributo
Sexo
Ignorado
Feminino
Masculino
Ambos_os_sexos
Sexo_do_provável_autor_da_violência
Pessoa_adulta
Pessoa_idosa
Jovem
Ciclo_de_vida_do_Principal_provável_autor_da_violência
Criança
Adolescente
Escolaridade
"5ª_à_8ª_série_incompleta_do_EF_(antigo_ginásio_ou_1°grau)"
"Ensino_médio_incompleto_(antigo_colegial_ou_2°_grau)"
Educação_superior_incompleta
"Ensino_fundamental_completo_(antigo_ginásio_ou_1°_grau)"
Educação_superior_completa
"Ensino_médio_completo(antigo_colegial_ou_2°_grau)"
1ª_a_4ª_série_incompleta_do_EF
"4ª_série_completa_do_EF_(antigo_1°_grau)"
Ocupação
Idade
Circunstância_da_lesão
Gestante
Observações_adicionais
Agravo
Raça/Cor
Parda
Preta
Indígena
Ignorada
Amarela
Branca
Lesão
Lesão_Interpessoal
Lesão_Autoprovocada
Transtorno
Transtorno_Comportamental
Transtorno_Mental
Outros_Transtornos
Suspeita_de_uso_de_álcool
is-a