Ministério da Educação
Universidade Tecnológica Federal do Paraná
Pró-Reitoria de Pesquisa e Pós-Graduação

RELATÓRIO FINAL DE ATIVIDADES DE INICIAÇÃO CIENTÍFICA

Extração de regras de associação em dados de acidentes rodoviários do Paraná

vinculado ao projeto

Cidades inteligentes: utilizando agentes inteligentes em problemas de transporte e trânsito

Júlio César Guimarães Costa
Bolsista Voluntário
Ciência da Computação
Data de ingresso no programa: 08/2021
Prof(ª). Dr(ª). Gleifer Vaz Alves

Área do Conhecimento: 1.03.03.00-6 Metodologia e Técnicas da Computação

CAMPUS PONTA GROSSA, 2022

Júlio César Guimarães Costa
Gleifer Vaz Alves

EXTRAÇÃO DE REGRAS DE ASSOCIAÇÃO EM DADOS DE ACIDENTES RODOVIÁRIOS DO PARANÁ

Relatório de Pesquisa do – Programa Institucional Voluntario de Iniciação Cientifica da Universidade Tecnológica Federal do Paraná.

Ponta Grossa, 2022

SUMÁRIO

INTRODUÇÃO	**4**
TRABALHOS RELACIONADOS	**4**
ALGORITMO APRIORI	**4**
PROGRAMA WEKA	**5**
METODOLOGIA	**5**
EXPERIMENTOS E DISCUSSÕES	**8**
CONCLUSÃO	**9**
REFERÊNCIAS	**10**

INTRODUÇÃO

Segundo a Confederação Nacional do Transporte no Brasil, a matriz de transportes é predominantemente rodoviária, contendo cerca de 120.767,3 km de extensão de rodovias pavimentadas e não pavimentadas por todo o país [1]. Segundo os dados da Polícia Rodoviária Federal, durante o ano de 2021 ocorreram 64.515 acidentes documentados em rodovias brasileiras. Este relatório, portanto, propõe a aplicação de estratégias de filtragem, pré-processamento, extração e a interpretação de regras de associação pela aplicação do algoritmo Apriori em dados abertos disponíveis eletronicamente, e coletados pela Polícia Rodoviária Federal.

Este trabalho restringiu-se a somente os dados referentes ao estado do Paraná no ano de 2021, considerando dois principais fatores: a porcentagem de acidentes em rodovias do estado do Paraná representarem 11% (7.335 acidentes) dos incidentes coletados pela Polícia Rodoviária Federal no Brasil, e a existência de trabalhos e metodologias semelhantes aplicadas em outras regiões do país, contendo resultados relevantes para a descoberta de conhecimentos e padrões úteis para a criação de planos de contingência.

Para atingir os objetivos do trabalho, foi necessário identificar uma base de dados que permitisse obter informações relevantes sobre os acidentes no estado do Paraná. Para tal, se baseando na literatura conhecida, foram utilizados os dados abertos da Polícia Rodoviária Federal, disponíveis do portal eletrônico do Governo Federal [2]. Além da base de dados foi necessário identificar ferramentas que para filtrar e manipular dados (usou-se a ferramenta LibreOffice Calc) e a ferramenta Weka [3] para o pré-processamento e aplicação do algoritmo Apriori.

TRABALHOS RELACIONADOS

Existem diversos trabalhos que propõem a aplicação do algoritmo Apriori, na extração de regras de associação em dados sobre acidentes em rodovias no Brasil. Este trabalho, no entanto, se diferencia dos demais, pela aplicação de técnicas de pré-processamento, parametrização e a delimitação do estudo pelos dados referentes ao estado do Paraná.

Oliveira [4], propõe a utilização de técnicas de generalização na etapa de pré-processamento nos atributos *km*, *idade* e *horário* pois, “em determinados atributos, para a criação de classes de dados, visando tornar a descoberta de padrões mais eficiente, uma vez que uma grande variabilidade de dados dificulta tal descoberta.”.

Nogueira, Lee e Rissino [5], reduziram o escopo de busca aplicando o algoritmo Apriori em dados referentes a BR-101 entre o quilômetro 55 e 90, desta forma conseguindo identificar regras específicas mesmo que com altos níveis de suporte.

ALGORITMO APRIORI

O algoritmo Apriori proposto por Agrawal, R. and Srikant, R. [5], permite através da frequência de itens individuais, identificar ocorrências subsequentes nos dados e delas, extrair associações, utilizadas na descoberta de conhecimento. No algoritmo existem parâmetros que podem ser ajustados para a extração das regras, entre eles, os mais relevantes para a descoberta de conhecimento são o número de regras geradas, o suporte mínimo e máximo e a métrica de avaliação aos quais podem ser utilizadas a taxa de confiança, o suporte, o aproveitamento e a convicção.

Neste trabalho, foi utilizada a taxa de confiança, por se tratar de uma métrica, que indica a relação da ocorrência de um determinado conjunto, quando um outro conjunto de dados está presente, indicando assim a recorrência de um determinado cenário na base de dados. Os parâmetros de suporte mínimo e máximo ajustam a quantidade de instâncias utilizadas nos conjuntos, na aplicação do algoritmo, o ajuste desses parâmetros portanto permite atingirmos diferentes abordagens, pois faixas de valores de suporte mais baixas, permitem a identificação de regras mais específicas, quanto a utilização de faixas de valores de suporte mais altas, geram abordagens mais genéricas.

PROGRAMA WEKA

A principal ferramenta escolhida para o desenvolvimento do trabalho foi a ferramenta Weka. Desenvolvida pela Universidade de Waikato na Nova Zelândia [6], esta aplicação é capaz de facilitar a análise de dados permitindo a rápida aplicação de filtros e técnicas de pré-processamento tanto nos atributos quanto nas instâncias. As opções tanto de pré-processamento quanto de associação estão presente no menu superior da ferramenta e permitiram a aplicação das técnicas e do próprio algoritmo Apriori e sua parametrização que serão descritas na metodologia deste relatório**.** A figura 1 abaixo indica a tela principal da ferramenta.

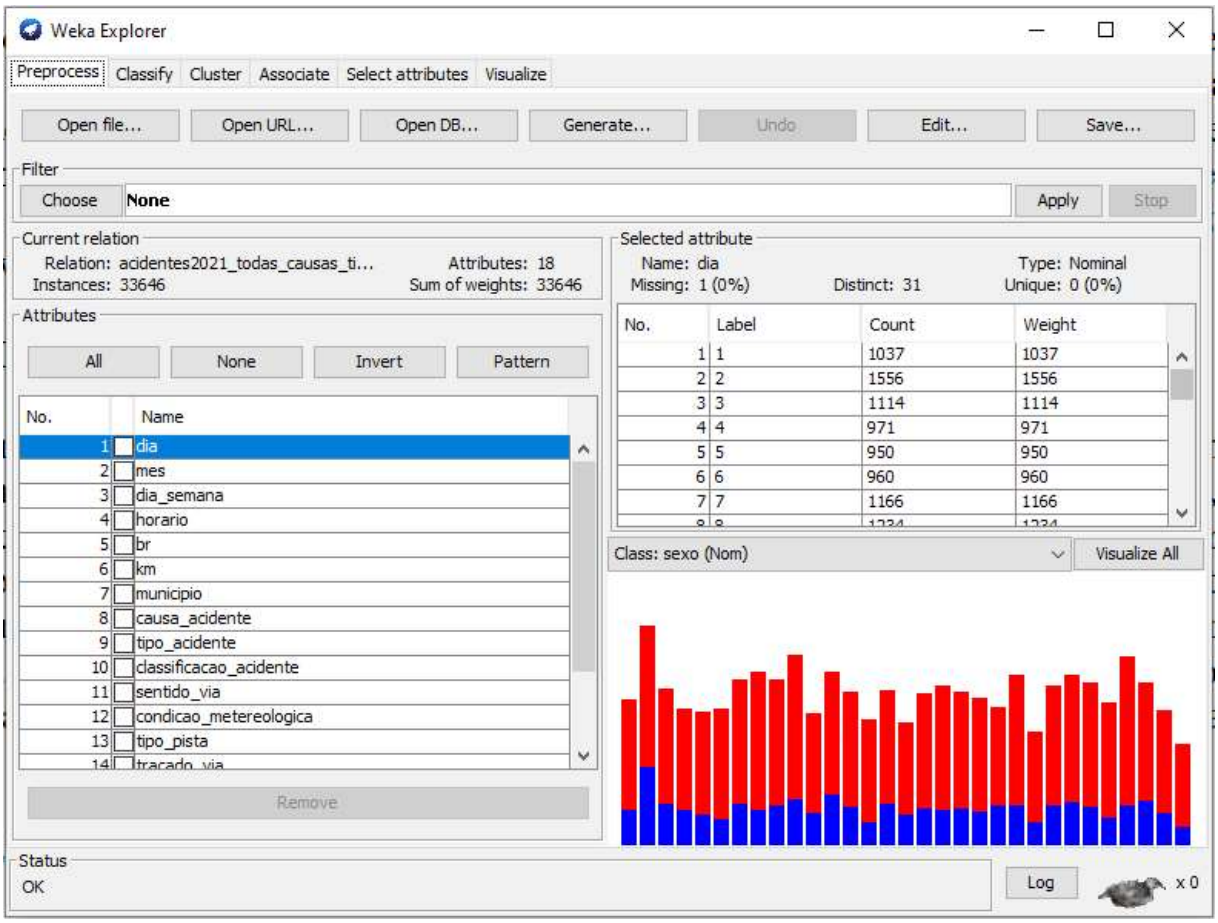


Figura 1. Tela principal da ferramenta Weka.

METODOLOGIA

As técnicas aqui apresentadas foram desenvolvidas se baseando em técnicas da literatura e buscaram se adaptar aos dados disponíveis, foi constatado então que com a

aplicação das técnicas propostas foi possível diminuir a redundância entre atributos e extrair regras contra intuitivas que indicaram comportamentos inesperados nos dados.

Obtenção dos dados

Os dados foram obtidos eletronicamente, pelo portal da Polícia Rodoviária Federal no web site do Governo Federal [2], e estavam disponíveis organizadas das seguintes maneiras:

- Por pessoas;
- Ocorrências;
- Por pessoas — Todas as causas e tipos de acidentes.

Para todos os formatos disponíveis poderiam ser obtidos os dados referentes aos anos de 2017 em diante, não havendo quaisquer restrições de acesso. Neste trabalho foram escolhidos os dados agrupados Por Pessoa — Todas as causas e tipos de acidentes, referentes somente ao ano de 2021. Com o arquivo no formato CSV importado na ferramenta LibreOffice Calc, foi possivel identificar os seguintes atributos disponíveis: *id, pesid, data_inversa, dia_semana, horario, uf, br, km, municipio, causa_principal, causa_acidente, ordem_tipo_acidente, tipo_acidente, classificacao_acidente, fase_dia, sentido_via, condicao_metereologica, tipo_pista, tracado_via, uso_solo, id_veiculo, tipo_veiculo, marca, ano_fabricacao_veiculo, tipo_envolvido, estado_fisico, idade, sexo, ilesos, feridos_leves, feridos_graves, mortos, latitude, longitude, regional, delegacia e uop*.

Filtragem e delimitação do estudo

A primeira modificação feita nos dados foi a filtragem das instâncias referentes somente ao estado do Paraná, para que assim fosse feita a delimitação do estudo, para atingir este objetivo foi aplicado o filtro automático da ferramenta, e filtradas somente as instâncias cujo atributo "uf" indicassem o estado do Paraná (PR), totalizando 42.152 instâncias referentes a 7.335 acidentes que ocorreram no estado em 2021.

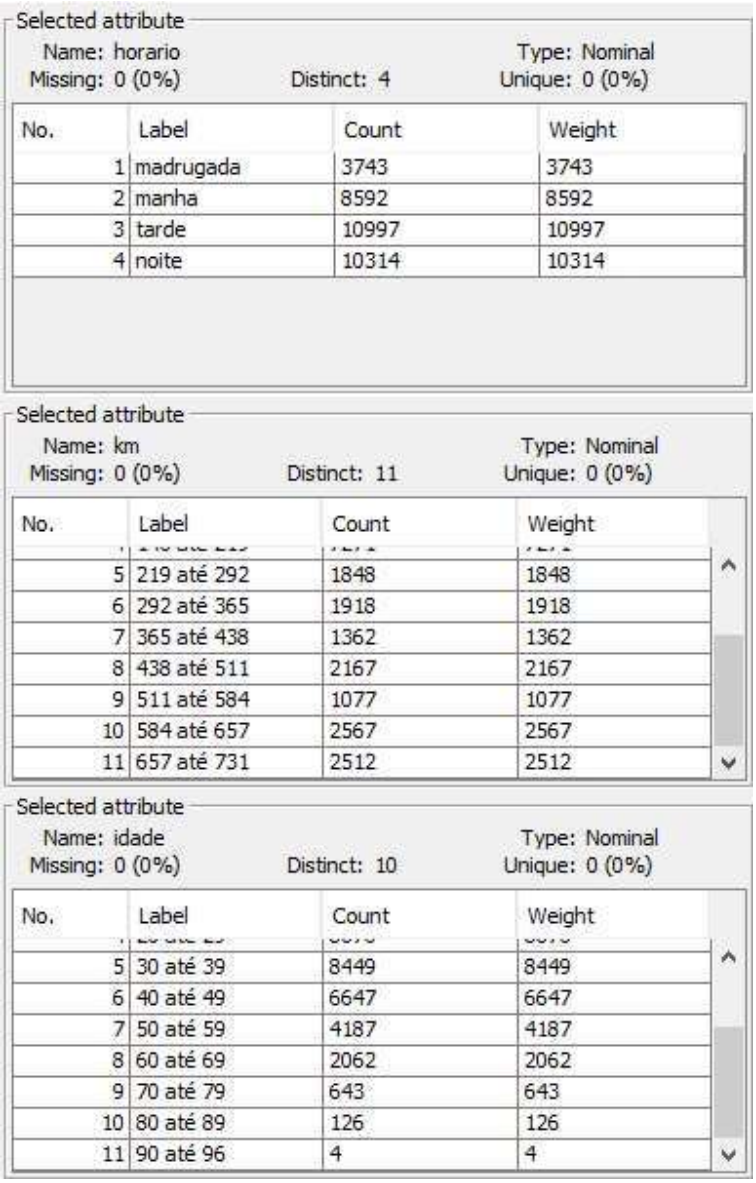
Ainda na ferramenta LibreOffice, foi feita uma modificação no atributo *data_inversa*, o dividindo em dois novos atributos, *dia e mes*, esta modificação foi feita por viabilizar regras de associação relacionadas aos dias do mês e a meses específicos do ano. Além desta modificação foi necessária a implementação de um breve programa para remoção de acentos e caracteres especiais em todos os dados, pois devido a uma limitação, estes valores poderiam não estar padronizados e portanto não seriam corretamente interpretados pelo programa Weka.

Pré-processamento dos dados

Com os dados devidamente filtrados, e em formato CSV, foi utilizada a função *"explorer"*[6] do programa Weka a fim de importar a base de dados e assim transformá-la em formato .arff, permitindo que assim seja reconhecida pelo programa Weka. Dando início portanto a etapa de pré-processamento, foram inicialmente removidos 19 dos 37 atributos disponíveis. Esta alteração foi feita através do filtro "*Remove*"[6] da ferramenta Weka e decorreu da existência de diversos atributos com valores redundantes e identificadores numéricos. Como consequência, a presença desses atributos poderiam prejudicar a extração das regras de associação favorecendo a presença de regras com confiança em 100% que não indicassem qualquer informação relevante. Por fim os atributos restantes foram: *dia, mês, dia_semana, horário, br, km, municipio, causa_acidente, tipo_acidente,*

classificação_acidente, sentido_via, condicao_metereologica, tipo_pista, tracado_via, tipo_veiculo, tipo_envolvido, idade e sexo.

Em decorrência do algoritmo Apriori permitir somente atributos com valores nominais, foi feita uma generalização nos atributos *horário*, *idade* e *km* que continham dados numéricos em grandes faixas de valores. Esta modificação foi necessária, pois, não existia volume suficiente nos conjuntos isoladamente que permitisse gerar regras que utilizassem estes atributos, portanto, foi aplicada a seguinte estratégia: para os atributos de *idade* e *km* os valores foram divididos em 10 grupos levando em consideração o maior e o menor valor encontrado, já para o atributo horário foi feita a divisão das 24 horas do dia em 4 grupos de 6 horas: madrugada, manhã, tarde e noite. Nesta etapa, foi utilizado o filtro "*MergeManyValues*"[6] (Frank et al. 2016) e os resultados estão apresentados na figura 2.



Selected attribute

Name: horario Type: Nominal
Missing: 0 (0%) Distinct: 4 Unique: 0 (0%)

No.	Label	Count	Weight
1	madrugada	3743	3743
2	manha	8592	8592
3	tarde	10997	10997
4	noite	10314	10314

Selected attribute

Name: km Type: Nominal
Missing: 0 (0%) Distinct: 11 Unique: 0 (0%)

No.	Label	Count	Weight
5	219 até 292	1848	1848
6	292 até 365	1918	1918
7	365 até 438	1362	1362
8	438 até 511	2167	2167
9	511 até 584	1077	1077
10	584 até 657	2567	2567
11	657 até 731	2512	2512

Selected attribute

Name: idade Type: Nominal
Missing: 0 (0%) Distinct: 10 Unique: 0 (0%)

No.	Label	Count	Weight
5	30 até 39	8449	8449
6	40 até 49	6647	6647
7	50 até 59	4187	4187
8	60 até 69	2062	2062
9	70 até 79	643	643
10	80 até 89	126	126
11	90 até 96	4	4

Figura 2. Generalização dos atributos *Horário, Idade e Km.*

Como resultado da generalização foi identificado que das 42.152 instâncias, 8.506 instâncias tinham os atributos de *idade* e *km* faltantes ou valores discrepantes, foi também identificado que a ausência desses valores ocorria majoritariamente nas mesmas instâncias, para então evitar regras com confiança 100% que envolvessem esses valores, foram

removidas as 8.506 instâncias, usando o filtro *"RemoveWithValues"*[6] no atributo idade, totalizando 33.646 instâncias restantes a serem utilizadas nos experimentos a seguir.

EXPERIMENTOS E DISCUSSÕES

A correta interpretação das regras que serão apresentadas é feita da seguinte maneira: quando está presente o conjunto X, este implica em Y. O exemplo a seguir: {leite, ovos} 6 ⇒ {pão} 4, pode portanto ser lido como: quando há leite e ovos em uma instância, pão também aparece. Os números que acompanham os conjuntos indicam o número de instâncias em que esta associação ocorre.

No primeiro experimento foi aplicada a parametrização padrão da ferramenta Weka mantendo o suporte mínimo em 10%, o suporte máximo em 100% e a taxa de confiança em 90%, gerando as 6 regras indicadas a seguir:

1. tipo_veiculo=Motocicleta tipo_envolvido=Condutor 4686 ⇒ sexo=Masculino 4383 <conf: (0.94)>
2. classificacao_acidente=Com Vitimas Feridas tipo_veiculo=Motocicleta sexo=Masculino 4082 ⇒ tipo_envolvido=Condutor 3806 <conf:(0.93)>
3. tipo_veiculo=Motocicleta sexo=Masculino 4719 ⇒ tipo_envolvido=Condutor 4383 <conf: (0.93)>
4. classificacao_acidente=Com Vitimas Feridas tipo_veiculo=Motocicleta tipo_envolvido=Condutor 4101 ⇒ sexo=Masculino 3806 <conf:(0.93)>
5. dia_semana=domingo tipo_envolvido=Condutor 3897 ⇒ sexo=Masculino 3533 <conf: (0.91)>
6. br=376 tipo_pista=Dupla tipo_envolvido=Condutor 4042 ⇒ sexo=Masculino 3645 <conf: (0.9)>

Dos resultados obtidos é possível observar que, todas as regras geradas contém o atributo *sexo* e *tipo_envolvido*, isto ocorre, pois, no atributo *sexo* existem somente 4 rótulos possíveis, sendo que masculino é bastante predominante em relação aos outros, já no atributo *tipo_envolvido*, é intuitivo de se pensar que em todo acidente tenha um condutor. Outros três atributos também se mostraram prejudiciais ao serem testados em valores de suporte menor, *classificacao_acidente* e *tipo_pista*, pelo mesmo motivo do atributo *sexo*, tendo um rótulo bastante predominante sobre os outros. Outro atributo identificado como prejudicial foi o *km,* pois gerou associações entre faixas de quilometragens e municípios e portanto não indicavam qualquer informação relevante.

Então, para identificar regras específicas, as seguintes estratégias foram adotadas: a remoção dos atributos *tipo_envolvido*, *sexo*, *classificação_acidente*, *km, tipo_pista* e a utilização da seguinte parametrização: suporte mínimo em 1%, suporte máximo em 10%, a taxa de confiança em 70%, o número de regras em 50 e uma modificação no parâmetro padrão “delta” que indica a taxa de decaimento do suporte em cada ciclo performado pelo algoritmo, para 1%. Como resultado, das 50 regras foram selecionadas as 3 principais regras que indicavam informações contra intuitivas, a lista final dos atributos utilizados pode ser verificada na figura 3 abaixo.

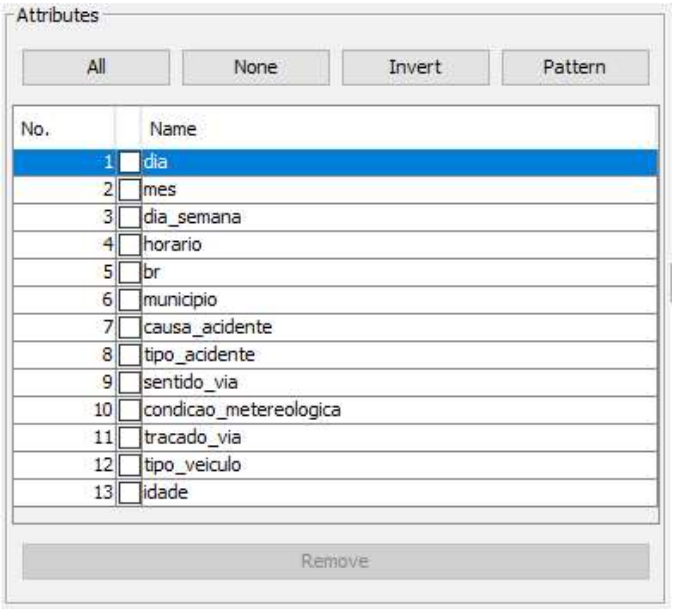


Figura 3. Atributos do segundo experimento.

24. municipio=GUARATUBA 839 ⇒ sentido_via=Crescente 730 <conf:(0.87)>
Dos 839 acidentes que ocorreram no município de Guaratuba, 730 (87%) ocorreram no sentido crescente da via.

34. br=277 municipio=MORRETES 1010 ⇒ sentido_via=Decrescente 841 <conf:(0.83)>
Dos 1.010 acidentes da BR 277 no município de Morretes, 841 (83%) acidentes ocorreram no sentido decrescente da via.

46. br=116 tracado_via=Curva 1179 ⇒ municipio=CAMPINA GRANDE DO SUL 932 <conf:(0.79)>
Dos 1.179 acidentes que ocorreram em curvas na BR 116, 932 (79%) ocorreram no município de Campina Grande do Sul.

É de se notar que nas regras 24 e 34 o sentido da via foi fator crucial na ocorrência dos acidentes nestes municípios, podendo indicar algum problema ou sobrecarga de automóveis nestes sentidos nas rodovias. Já na regra 46 pode-se inferir que há algum problema com as curvas da BR 116 na região do município de Campina Grande do Sul.

CONCLUSÃO

Este trabalho apresentou uma metodologia para descoberta de conhecimento nas bases da PRF, onde a metodologia aplicada consistiu nas etapas de pré-processamento, filtragem e parametrização do algoritmo Apriori, foi observado então, que aplicar estratégias e diferentes abordagens permite surpassar as limitações que a simples aplicação do algorítimo gera como resposta, e portanto, a importância da correta parametrização e escolha dos atributos para cada tipo de abordagem. Outro fator crucial foi entender como as modificações nos parâmetros e a escolha dos atributos alteravam os resultados obtidos.

Das dificuldades encontradas no desenvolvimento do trabalho, a mais crucial foi justamente encontrar um equilíbrio entre a remoção de atributos e parametrizações que viciavam o algoritmo e a possível perda no ganho de informação pela remoção dos mesmos.

Em trabalhos futuros, uma possível abordagem será a aplicação das estratégias descritas neste trabalho, aplicadas a diferentes regiões e estados, outra possível aplicação observada, visa utilizar os dados somente das principais rodovias do Brasil isoladamente, pois

como visto nos resultados, abordagens mais específicas, tendem a gerar resultados menos triviais.

REFERÊNCIAS

[1] CONFEDERAÇÃO NACIONAL DO TRANSPORTE. CNT lança painel com dados do transporte rodoviário no Brasil. **Painel Informativo CNT**, [*s. l.*], 16 dez. 2020. Disponível em:https://cnt.org.br/agencia-cnt/cnt-lanca-painel-com-dados-do-transporte-rodoviario-no-brasil. Acesso em: 22 abr. 2022.

[2] MINISTÉRIO DA JUSTIÇA E SEGURANÇA PÚBLICA. Polícia Rodoviária Federal (org.). **Dados Abertos Acidentes**. [*S. l.*], 5 set. 2022. Disponível em: https://www.gov.br/prf/pt-br/acesso-a-informacao/dados-abertos/dados-abertos-acidentes. Acesso em: 22 abr. 2022.

[3] Frank, E. Hall, M. A. and Witten, I. H (2016). **The WEKA Workbench**. Data Mining: Practical Machine Learning Tools and Techniques, Morgan Kaufmann, 4 ed., 2016.

[4] Oliveira, Isaac José. **Mineração de dados em padrões de acidentes de trânsito: o uso de dados abertos da Polícia Rodoviária Federal no RN**. 2020. Trabalho de Conclusão de Curso (Graduação) - UFRN, Caicó, 2020.

[5] Nogueira, F. da S., Lee, L. & Rissino, S. das D. (2018). **Descoberta de Conhecimento na base de dados aberta da Polícia Rodoviária Federal: Identificação de Pontos Críticos na Rodovia BR 101 no Município de São Mateus/ES**. Brazilian Journal of Production Engineering, 4(4): 70-90.

[6] Agrawal, R. and Srikant, R. (1994) **Fast Algorithms for Mining Association Rules in Large Databases**. Proceedings of the 20th International Conference on Very Large Data Bases, Santiago de Chile, 12-15 September 1994, pg. 487-499.

Gleifer Vaz Alves

Júlio César Guimarães Costa