CENTRO UNIVERSITÁRIO DE ANÁPOLIS – UniEVANGÉLICA
BACHARELADO EM ENGENHARIA DE COMPUTAÇÃO

DESCOBERTA DE CONHECIMENTO NA RELAÇÃO ENTRE ACIDENTES DE TRÂNSITO RODOVIÁRIO E FATORES CLIMÁTICOS, NO EIXO GOIÂNIA-DISTRITO FEDERAL

RAPHAEL DOS SANTOS GUEDES VIEIRA

ANÁPOLIS
2018

RAPHAEL DOS SANTOS GUEDES VIEIRA

DESCOBERTA DE CONHECIMENTO NA RELAÇÃO ENTRE ACIDENTES DE TRÂNSITO RODOVIÁRIO E FATORES CLIMÁTICOS, NO EIXO GOIÂNIA-DISTRITO FEDERAL

Trabalho de Conclusão de Curso II apresentado como requisito parcial para a conclusão de grau do curso de Bacharelado em Engenharia de Computação do Centro Universitário de Anápolis – UniEVANGÉLICA.

Orientador(a): Profª. Esp. Aline Dayany de Lemos.

Anápolis
2018

RAPHAEL DOS SANTOS GUEDES VIEIRA

DESCOBERTA DE CONHECIMENTO NA RELAÇÃO ENTRE ACIDENTES DE TRÂNSITO RODOVIÁRIO E FATORES CLIMÁTICOS, NO EIXO GOIÂNIA-DISTRITO FEDERAL

Trabalho de Conclusão de Curso II apresentado como requisito parcial para a obtenção de grau do curso de Bacharelado em Engenharia de Computação do Centro Universitário de Anápolis – UniEVANGÉLICA.

Aprovado(a) pela banca examinadora em **11** de **dezembro** de 2018, composta por:

_____
Profª. Esp. Aline Dayany de Lemos
Orientador(a)

_____
Profª. Ma. Luciana Nishi

_____
Prof. Me. Millys Fabrielle Araujo Carvalhaes

À minha família (minha mãe, meu pai e meu irmão),
que me apoiam.

Agradecimentos

Começo meus agradecimentos ao dizer que são poucas linhas para expressar tantos sentimentos, mas dentre tais um toma o primeiro plano: A Gratidão.

Primeiramente, agradeço a quem ouve todas as minhas queixas, observa-me todos os dias, instrui-me sobre o que é certo ou errado e agracia-me diariamente, com um presente diferente, pois assim como as águas correntes de um rio são distintas a cada momento, cada dia o fôlego de vida que possuo, distingue-se dos passados e dos vindouros. Assim agradeço ao meu *Abba*, Amigo, Confidente e Porto Seguro: Deus!

Em seguida agradeço a minha família. Meu irmão, por ter me incentivado o ingresso no Ensino Superior. Agradeço ao meu pai, por me admirar tanto, e fazer de mim motivo de seu orgulho, ao dizer para "quase todo mundo" que seu filho era acadêmico de Engenharia de Computação. E especialmente, agradeço com muito amor a minha mãe. Se não fosse por Deus e por ela, eu não teria chegado até aqui. Ela sempre esteve ao meu lado, sempre se preocupou comigo, encorajou-me quando precisei, deu-me afago quando chorei, acalmou-me quando quis desistir de tudo e sempre cuidou de mim. Mãe, como eu te amo! Agradeço a Deus por ter me dado uma mãe como a senhora!

Agradeço a minha família agregada. Todos aqueles que conviveram comigo durante esse tempo, observando-me, de perto ou de longe, mas que me admiravam e torciam pelo meu sucesso.

Agradeço aos meus professores da educação básica, que um dia esperavam ver-me no ensino superior. Digo que eles contribuíram, cada um de modo diferente, para minha jornada acadêmica.

Agradeço também aos meus amigos, uma lista extremamente restrita, todavia significativa. Entretanto, preciso destacar um amigo com o qual compartilhei diversos momentos (não apenas de alegria, mas também de reflexão e aprendizado), a mesma sala de aula, o mesmo ambiente de trabalho e até, quase as mesmas lamúrias. Foram mais de oito períodos de muito aprendizado, de ajuda mútua, "um ajuda ao outro que a gente consegue". Não esqueço de quando trabalhávamos juntos e dos episódios cômicos pelo qual passamos. Mas também foi uma pessoa que me deu palavras de encorajamento quando, do 2º ao 8º período, pensei em desistir. Muito obrigado Thiago Silva!

Ainda sobre a academia, agradeço a Karen Amorim, ex-secretária dos Cursos Superiores em Computação, além de uma pessoa excepcional foi uma grande conselheira. Agradeço também à Leia Silva, atual secretária, por me aturar enchendo sua paciência quase sempre!

Agradeço também a professora Viviane, por ser uma pessoa tão prestativa que sempre me recebeu muito bem quando precisei de sua ajuda. Agradeço a todo o corpo docente do curso de Engenharia de Computação que contribuiu com minha formação. Não posso esquecer do meu grupo de convivência em sala de aula, valeu pessoal!

Mas ainda não acabou! Agradeço do fundo do meu coração a professora Luciana Nishi, comumente conhecida por seu sobrenome. Mais do que uma mestra, ela foi uma mãe! Muito obrigado pelas nossas conversas, pelos valiosos conselhos, pelo partilhar de experiências e como ela mesma diz: - Pela paixão de ensinar. Levarei muito de você comigo! Não é para tanto que fiquei conhecido como seu aprendiz, nas monitorias de Projeto Interdisciplinar I e III, especialmente nesta última, como eu gostava!

Entretanto, ainda há uma outra pessoa que é motivo de muita admiração! Ela é um exemplo de profissionalismo, dedicação, força de vontade, e acima de tudo de superação. Antes de agradecer, parabenizo por ser uma mestra e pessoa excepcional. E uso de sinceridade ao escrever tudo isso. Obrigado por ministrar disciplinas tão complexas de modo tão entendível, só muito amor pelo que faz, justifica um trabalho como o seu. Obrigado pelos conselhos valiosos, os puxões de orelha na hora certa e por ser minha orientadora! Eu não imaginava que minha escolha teria sido tão acertada. Lembro como estava perdido, no oitavo período, em procurar um tema para desenvolver futuramente, como Trabalho de Conclusão de Curso e cá está o resultado de muita orientação. Obrigado por dispender seu tempo e por me ouvir quando eu disparava a falar. Agradeço também por me ajudar enxergar o que eu era capaz de atingir e pelos momentos de apoio, quando me sentia inseguro nas decisões a serem tomadas, no desenrolar desta pesquisa. Muito obrigado Aline de Lemos!

"*Àquele que está assentado no trono e ao Cordeiro sejam o louvor, a honra a glória e poder, para todo o sempre!*"
Ap. 5.13b NVI

Resumo

No mundo, as mortes por acidentes no trânsito preocupam as autoridades governamentais que buscam formas de reduzir o índice e a gravidade desses fenômenos que ceifam a vida de muitas pessoas. Medidas têm sido elaboradas para esse fim, entretanto a situação ainda preocupa. Assim, objetivou-se aplicar a descoberta de conhecimento em bases de dados para investigação de novos padrões em dados de acidentes de trânsito no eixo Goiânia-Distrito Federal entre os anos de 2012 a 2017. Para isso, foram utilizados dados das ocorrências de acidentes, disponíveis no Portal de Dados Abertos do Departamento de Polícia Rodoviária Federal. O procedimento foi precedido pela avaliação de tarefas, técnicas, algoritmos e ferramentas para mineração de dados, que resultaram na criação de um modelo de classificação empregando árvores de decisão com a implementação do algoritmo *C4.5*; auxiliado pela ferramenta *Weka 3*. A descoberta de conhecimento foi guiada pela aplicação do processo *CRISP-DM*. Ao final desta pesquisa, o modelo gerado permitiu a avaliação de novas ocorrências de acidentes para a área escolhida, com 47,3% de precisão, e a interpretação do conhecimento descoberto.

Palavras-chave: Descoberta de Conhecimento em Bases de Dados. Acidentes de Trânsito. BR-060. Acidentes e Fatores Climáticos. Árvores de Decisão. Classificação. *C4.5*.

Abstract

In the world, deaths by road traffic accidents worry government authorities, who are looking for ways to reduce the rate and severity of these events that take off the life of several people. Ways have been elaborated for this purpose, however the situation is still worrying. Thus, the objective was to apply the knowledge discovery in databases to investigate new patterns in road traffic accident data in the Goiânia-Distrito Federal region, between the years of 2012 to 2017. For this goal, it was used accident data occurrences, available at the Open Data Portal of the Federal Highway Police Department. The procedure was preceded by the evaluation of tasks, techniques, algorithms and tools for data mining, which resulted in the creation of a classification model using decision trees with the implementation of algorithm C4.5; aided by the Weka 3 tool. The discovery of knowledge was guided by the application of the CRISP-DM process. At the end of this research, the generated model allowed the evaluation of new occurrences of accidents for the chosen area, with 47.3% accuracy, and the interpretation of the discovered knowledge.

Keywords: Knowledge Discovery in Databases. Road Traffic-accidents. BR-060. Accidents and Climatic Factors. Decision Trees. Classification. *C4.5*.

Lista de Figuras

Lista de Quadros

Lista de Tabelas

Lista de Abreviaturas e Siglas

ABNT	Associação Brasileira de Normas Técnicas
AGNU	Assembleia Geral das Nações Unidas
AT	Acidente de Trânsito
CNM	Confederação Nacional de Municípios
CRISP-DM	*Cross-Industry Standard Process for Data Mining*
CTB	Código Brasileiro de Trânsito
DCBD	Descoberta de Conhecimento em Bases de Dados
DENATRAN	Departamento Nacional de Trânsito
DF	Distrito Federal
DPRF	Departamento de Polícia Rodoviária Federal
e-SIC	Sistema Eletrônico do Serviço de Informação ao Cidadão
FN	Falso Negativo
FP	Falso Positivo
GO	Goiás
HUGOL	Hospital de Urgências Governador Otávio Lage de Siqueira
IBM	*International Business Machines*
IDC	*International Data Corporation*
IPEA	Instituto de Pesquisa Econômica Aplicada
IPI	Imposto sobre Produtos Industrializados
KDD	*Knowledge Discovery in Databases*
MD	Mineração de Dados
ODS	Objetivos de Desenvolvimento Sustentável
OMS	Organização Mundial da Saúde
ONSV	Observatório Nacional de Segurança Viária
PARE	Prevenção de Acidentes e Reeducação no Trânsito
PDA	Portal de Dados Abertos
RNA	Rede Neural Artificial
SGBD	Sistema de Gerenciamento de Banco de Dados
SIM	Sistema de Informações sobre Mortalidade
SQL	*Structured Query Language*
VN	Verdadeiro Negativo
VP	Verdadeiro Positivo

Sumário

1 INTRODUÇÃO

Segundo a Organização Mundial da Saúde (OMS) (2015), acidentes de trânsito (ATs) ocupam a nona posição dentre todas as causas de mortes no mundo. O Sistema de Informações sobre Mortalidade (SIM) (2017), informou que de 2012 a 2016, em torno de 212 mil pessoas vieram a óbito em vias terrestres brasileiras, o que resulta em uma média anual com cerca de 42.430 perdas.

Embora haja ações no sentido de reduzir os eventos e suas consequências, ainda existem deficiências quanto à investigação e sistematização das informações sobre as ocorrências de dos ATs, como revelam a Confederação Nacional de Municípios (CNM) (2013) e Lima et al (2008). Eles ressaltam que é dado mais enfoque nos tipos de acidentes e comportamento dos condutores, se comparado aos fatores que contribuem para a sua realização. Entre os fatores pouco observados, estão as condições climáticas, que aumentam o índice de fatalidade dos acidentes quando presentes de forma adversa (REIS, 2014).

A partir dessa problemática, a pesquisa propõe empregar descoberta de conhecimento em bases de dados, relacionando acidentes de trânsito e fenômenos climáticos, no eixo Goiânia-Distrito Federal, entre 2012 a 2017. Para o cenário apresentado, pretende-se analisar a aplicação de tarefa, técnica, algoritmo e ferramenta de mineração de dados, avaliar o emprego de um processo de descoberta de conhecimento e identificação de padrões preditivos entre as ocorrências de acidentes e fenômenos climáticos.

Analisados isoladamente, os dados não detêm valor em si. Sua importância está centrada na habilidade de extração de conhecimento a partir deles (GOLDSCHMIDT; PASSOS; BEZERRA, 2015). Utilizando informações disponíveis sobre acidentes de trânsito no Brasil, constata-se que as análises que enfatizam os fatores influenciadores para a realização desses eventos, ainda são poucas, em vista daquelas que enfocam seus tipos e causas (BRASIL, 2010; INSTITUTO DE PESQUISA ECONÔMICA APLICADA – IPEA, 2015a, 2015b; LIMA et al., 2008).

Com o intuito de tornar evidente o conhecimento implícito em bases de dados de acidentes de trânsito, estudos que aplicam a descoberta de conhecimento em bases de dados (DCBD), têm sido realizados para auxiliar a tomada de decisão pelas organizações responsáveis pelo trânsito brasileiro. Isso contribui para a diminuição do déficit, presente nos relatórios nacionais quanto à exatidão das análises dos conjuntos de fatores que incidem num AT (GALVÃO, 2009; IPEA; DEPARTAMENTO NACIONAL DE TRÂNSITO – DENATRAN, 2006; REIS, 2014).

O uso da computação aplicada para delimitação do escopo desta pesquisa, por meio da DCBD, deu-se principalmente por conta dos resultados obtidos nos estudos acima mencionados (GALVÃO, 2009; REIS, 2014). E também, pela quantidade reduzida de trabalhos similares para o estado de Goiás – eixo Goiânia-Distrito Federal, constatada por meio de pesquisas em bases de artigos científicos, como Scielo[1] e Capes[2].

A seguir será apresentada a fundamentação teórica que norteou o desenvolvimento desta pesquisa. Ele está dividido em duas partes principais: uma contextualização dos ATs, e os aspectos relacionados a DCBD e Mineração de Dados. Na seção seguinte verificam-se os resultados alcançados. Há ainda considerações do autor sobre o que foi alcançado e aprendido, seguidas da sugestão de trabalhos futuros. Por fim, a seção de anexos e apêndices agregam informações que contribuem para o detalhamento dos resultados obtidos.

[1] Acessar: http://www.scielo.org/php/index.php
[2] Acessar: http://www.periodicos.capes.gov.br/

2 FUNDAMENTAÇÃO TEÓRICA

Para alcançar a devida compreensão da pesquisa, é importante que sejam esclarecidos os conceitos inerentes ao seu desenvolvimento, o que contribuirá para a delimitação do estudo, bem como para a delineação dos resultados obtidos.

O embasamento teórico abrange o ponto de vista da literatura sobre os acidentes de trânsito; os aspectos relacionados a dados, informação e conhecimento; o processo de descoberta de conhecimento em bases de dados, detalhando-se a etapa de mineração de dados com suas particularidades.

2.1 Acidentes de Trânsito

Acidente de trânsito é entendido como um evento não premeditado, (por vezes) passível de ser evitado, ocorrido parcial ou inteiramente em via pública, sendo que pelo menos uma das partes envolvidas deva estar em movimento. Além disso, conta com o envolvimento do ser humano, via, ambiente e demais elementos, em dimensões variáveis; podendo causar danos de diferentes níveis de gravidade a todos fatores envolvidos (ASSOCIAÇÃO BRASILEIRA DE NORMAS TÉCNICAS – ABNT, 1989; BRASIL, 2010; IPEA; DENATRAN, 2006; OMS, 2015).

Definido um AT, ABNT (1989) evidencia os seus agentes ou fatores geradores, entidades que de algum modo influenciam na realização do evento. Os quatro tipos de fatores são listados no quadro 1, a seguir:

Quadro 1 – Classificação dos Fatores Geradores de Acidentes

Fator	Descrição
Humano	Quando o comportamento do homem como pedestre, condutor ou qualquer outra condição, contribui para ocorrência do acidente.
Via	Quando uma deficiência na via ou sua sinalização contribui para a ocorrência do acidente.
Meio Ambiente	Quando fatores do meio ambiente ou da natureza prejudicam a segurança do trânsito, contribuindo para a ocorrência do acidente.
Veículo	Quando falha mecânica no veículo contribui para a ocorrência do acidente, sem que tenha havido negligência na manutenção ou fabricação.

Fonte: ABNT (1989, p. 4-5)

ATs são ações passivas com causa e efeito, isto é, carecem de agentes ativos (fatores geradores de acidentes) para concretizar a ação, possuindo uma categorização (tipos de acidentes) para a adequada sistematização (IPEA; DENATRAN, 2006; OLIVEIRA, 2012).

A gravidade de um acidente está associada ao fator humano e seu estado físico após o acontecimento. Este último pode ser classificado em três categorias, de acordo com sua

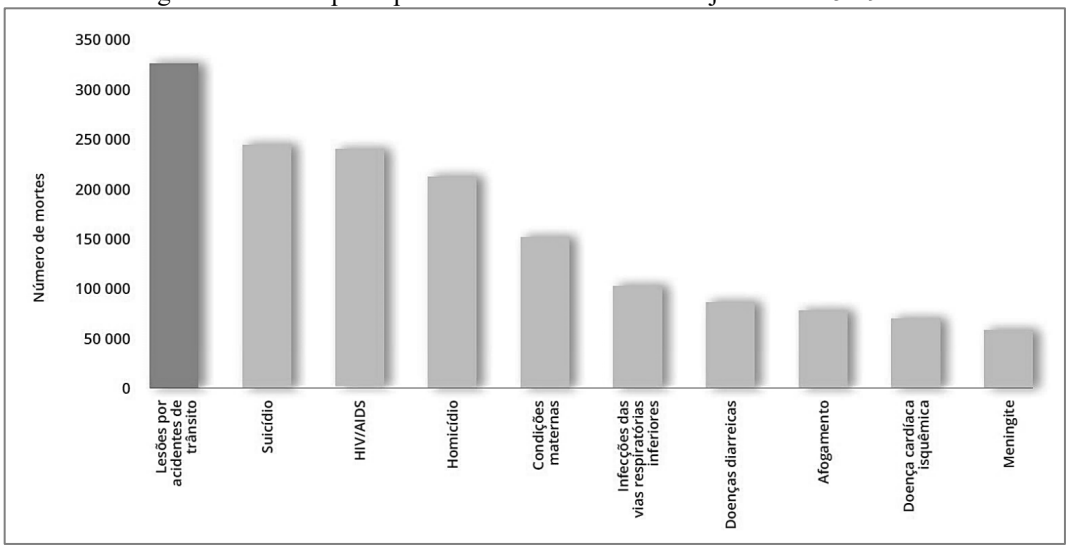
gravidade: sem vítima – quando todos os indivíduos saem ilesos; com vítima – quando pelo menos um envolvido sofre uma lesão (essa categoria se subdivide em vítimas leves e vítimas graves); e fatal – quando pelo menos um indivíduo vem a óbito no local (IPEA, 2015b).

Porém, os efeitos de um AT não figuram somente nos prejuízos causados à saúde física e mental do indivíduo. Os danos materiais, econômicos e sociais fazem parte das preocupações dos órgãos governamentais estaduais, nacionais e internacionais (IPEA, 2015a; OMS, 2015).

De acordo com a OMS (2015), o crescimento industrial, aliado à expansão populacional, contribui para o aumento da produção e comercialização de veículos motorizados em todo o mundo, principalmente nos países com renda média. Contudo, esse progresso tem colaborado com o incremento do número de ATs em muitos países.

Até 2013, 1,25 milhões de pessoas morriam, anualmente, vítimas de algum tipo de AT; considerado a principal causa de morte na faixa etária entre 15 a 29 anos (figura 1), e com maior índice de ocorrência em países de rendas média e baixa (OMS, 2015).

Figura 1 – As dez principais causas de morte entre os jovens de 15-29 anos



Fonte: OMS (2015, p. 1)

Apesar dos altos números, a taxa mundial apresenta estabilidade desde 2007. Todavia, faz-se necessário verificar como é o cenário em território nacional, já que em 2013 o índice brasileiro era de 23,4 mortes no trânsito por 100 mil habitantes (OMS, 2013).

A frota nacional de veículos também se expandiu. De 2004 até 2015, aumentou em 131%, alcançando 90 milhões de unidades em circulação (AMBEV S.A, 2017). A situação foi intensificada por volta de 2007-2008, quando o governo brasileiro decidiu reduzir os impostos sobre produtos industrializados (IPI) para combater o desemprego e reduzir os efeitos da crise mundial de 2008. Entretanto, o crescimento da frota provocou elevação das ocorrências de ATs tanto nas rodovias federais – foco deste trabalho de pesquisa, como nos centros urbanos.

(BRASIL, 2010; CNM, 2013; IPEA; DENATRAN, 2006).

Adentrando ao campo das rodovias federais, os condutores mais atingidos, em ordem, são os de veículos motorizados, passageiros e motociclistas. Os tipos de acidentes mais comuns, tanto nas rodovias como nos centros urbanos são colisão traseira e saída de pista, sendo as causas mais prováveis, a falta de atenção e o excesso de velocidade. Já os ATs fatais costumam ser do tipo colisão frontal e atropelamento de pedestre, evocando a imprudência como fator condicionante (IPEA, 2015a).

Quanto aos danos econômicos, em 2010 foram gastos R$ 6,5 bilhões de reais com ATs realizados em rodovias federais. Em 2014, os custos atingiram R$12,8 bilhões; com a maior parcela ligada ao fator humano (BRASIL, 2017b; IPEA, 2015b).

Restringindo mais o cenário de análise, chega-se ao estado de Goiás (GO). Em 2010, foram registrados 8.006 ATs em rodovias goianas. Na região entre Goiânia e Brasília, 1.407 ocorrências foram registradas (BRASIL, 2010). Com relação à mortalidade, os acidentes de trânsito ocupavam o segundo lugar nas causas de morte (GOIÁS, 2015).

As estatísticas apontaram que os acidentes que mais provocaram vítimas fatais foram do tipo colisão frontal, o qual também liderava o *ranking* nacional em ATs. Em relação às consequências econômicas, ATs em rodovias federais presentes em Goiás, custaram cerca de R$ 270 milhões de reais em 2006 (BRASIL, 2010; IPEA; DENATRAN, 2006).

O contexto de elevados índices de ATs, provocou em 1997 a criação do novo Código Brasileiro de Trânsito[3] (CTB), com propósito de conter o número de mortes no trânsito e a gravidade dos acidentes. A primeira década após a implantação do CTB foi marcada pela redução dos ATs e de seus efeitos. Porém, após o ano de 2007 as autoridades precisaram encontrar novos meios para garantir a segurança da população no trânsito, por conta dos índices altos (AMBEV S.A, 2017; BRASIL, 2010).

2.1.1 Medidas Preventivas

Com foco nos ATs, a Assembleia Geral das Nações Unidas (AGNU) declarou o período de 2011-2020 como a "Década de Ação pela Segurança no Trânsito". Essa iniciativa tem o objetivo de alcançar a "estabilização e redução do número de óbitos causados pelos acidentes de trânsito em todo o mundo" (OMS, 2011, p.5, tradução nossa).

Com base nesse movimento, a AGNU realizada em 2015 estabeleceu uma redução de 50% da quantidade de mortes no trânsito até 2020, como meta primordial dos Objetivos de Desenvolvimento Sustentável (ODS) (OMS, 2015).

[3] Lei 9.503 de 23 de setembro de 1997.

A “Década” originou movimentos nacionais e estaduais para apoiar no propósito da redução de acidentes e mortes causadas por ATs. Dentre essas iniciativas merecem destaque: a Semana Nacional de Trânsito e o Maio Amarelo.

A Semana Nacional de Trânsito tem o objetivo de alertar sobre os danos da violência no trânsito, na tentativa de conscientizar a população brasileira. Ocorre, geralmente na segunda quinzena do mês de setembro de cada ano. (BRASIL, 2017a)

O Maio Amarelo “busca chamar a atenção da sociedade para o alto índice de mortes e feridos no trânsito em todo o mundo” (MAIO AMARELO, 2014). O movimento conta os esforços da união do poder público, representado pelas entidades gestoras de trânsito e transporte e a sociedade civil.

Em Goiás, desde 2016 é realizado o programa de Prevenção de Acidentes e Reeducação no Trânsito (PARE), promovido pelo Hospital de Urgências Governador Otávio Lage de Siqueira (HUGOL). Esse programa também procura meios para reduzir o número de ATs. (GOIÁS, 2015)

2.2 Clima e Trânsito

Caleffi et al. (2016) declaram que o clima é capaz de incidir diretamente na capacidade e velocidade média do fluxo de uma rodovia. Agarwal, Maze e Souleyrette (2005), analisaram essa interferência e concluíram que as condições que mais influenciam são, em ordem: a chuva e a presença de neblina. Constatou-se que sob o cenário de precipitação leve ou intensa, a velocidade média dos veículos é reduzida em até 17%, se comparada ao tráfego comum, nas rodovias estudadas. Já Smith et al. (2003) concluem que a redução pode chegar a 30%.

Paula e Duarte (1996) analisaram, durante dois anos não contínuos, o comportamento dos ATs sob a presença de chuva na cidade de São Paulo. Eles compararam os resultados desses períodos contra épocas em que o clima se apresentou favorável e com o fluxo de trânsito igual para as duas situações. A conclusão foi que, na sazonalidade com precipitação, os acidentes tiveram um aumento de 107% (figura 2).

Figura 2 – Comparação de acidentes com clima seco e chuvoso

Tipo do acidente de trânsito	**Nº de acidentes de trânsito na área de estudos (*)**		
	938 horas de tempo seco	938 horas de tempo chuvoso	Acréscimo (Decorrência da chuva)
Com vítimas	42	74	76,1%
Atropelamento	21	34	61,9%
Sem vítimas	208	454	118,2%
Todos	271	562	107%

Fonte: Paula e Duarte (1996, p. 3)

Sobre a condição climática neblina, o Observatório Nacional de Segurança Viária (ONSV) (2014) explica a formação das neblinas como o excesso de umidade relativa do ar em épocas de geada ou, também, em pontos onde existe uma variação de temperatura devido locais onde a altitude é maior ou em regiões de depressão com aglomerações de água, como rios. A presença de acidentes sob essas circunstâncias "em geral [possuem] [...] consequências gravíssimas" (ONSV, 2016). Quando se analisa esse fenômeno, as reduções no tráfego são reduzidas em até 12%, revela Agarwal, Maze e Souleyrette (2005).

2.3 Dado, Informação e Conhecimento

Os dados constituem a pedra fundamental para o embasamento técnico desta pesquisa. É partir deles que a informação e, posteriormente, o conhecimento são gerados. Dados são entendidos como fatos, valores, resultados de medições, itens elementares ou sequências de símbolos mensuráveis que são armazenados em coleções, conhecidas como bases de dados (ELMASRI; NAVATHE, 2011; GOLDSCHMIDT; PASSOS; BEZERRA, 2015; SETZER, 2015; SILVA; PERES; BOSCARIOLI, 2016). O crescimento expressivo dos dados é realidade desde a década de 2000. E verifica-se que o aumento é inevitável (INTERNATIONAL DATA CORPORATION – IDC, 2014).

É constatado na literatura um intercambiamento que gera dúvidas, quanto a termos usados na área de dados. Mesmo que próximos, ou usualmente utilizados num mesmo contexto de aplicação: dado, informação e conhecimento apresentam definições distintas. As confusões residem no emprego de dado e informação e entre informação e conhecimento (SETZER, 2015).

Dados são informações sintáticas e não possuem um significado. Entretanto, quando uma semântica lhes é atribuída, eles adquirem valor e o transmite para quem os possui – uma organização, por exemplo. A partir daí, passam a ser identificados como informação. O conhecimento, então, é concebido quando as informações permitem tomada de decisões, a partir de análises feitas por agentes humanos (GOLDSCHMIDT; PASSOS; BEZERRA, 2015; SETZER, 2015; SILVA; PERES; BOSCARIOLI, 2016).

Sozinho, o conhecimento é apenas informação. Goldschmidt, Passos e Bezerra (2015) e Silva, Peres e Boscarioli (2016), declaram que para ele ser tornar um elemento auxiliador no processo de tomada de decisões, ainda é necessária análise humana que o valide como entendível, pertinente, novo e útil.

Goldschmidt, Passos e Bezerra (2015), ilustra que em um processo de descoberta de conhecimento voltado para a venda de veículos, pode ser gerado um padrão que associe a idade

de condutores à probabilidade de envolvimento em um AT. Embora esse "conhecimento" seja novo e válido, de modo algum é útil para a área na qual o processo foi aplicado.

2.4 Descoberta de Conhecimento em Base de Dados

Acidentes de trânsito são fenômenos diferentes de acidentes de transporte, entretanto são empregados indistintamente. Situação semelhante acontece com a Mineração de Dados (MD), ou *Data Mining*, um conceito relacionado, porém diferente de Descoberta de Conhecimento em Bases de Dados (DCBD), em inglês, *Knowledge Discovery in Databases* (KDD).

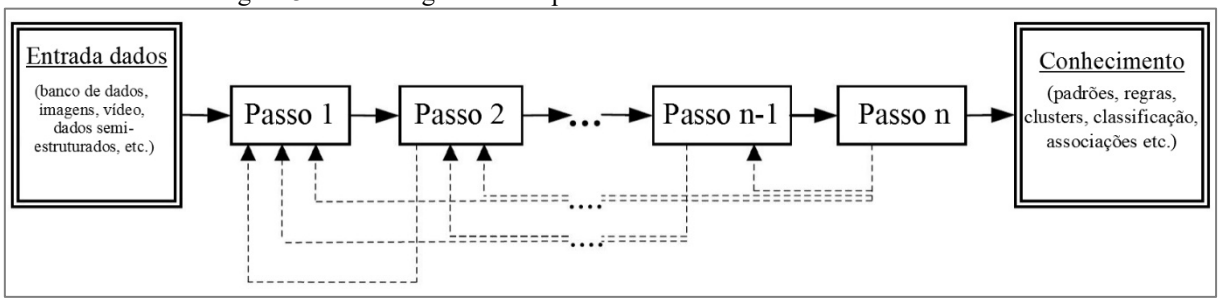
Para Fayyad et al. (1996, p. 40, tradução nossa), Descoberta de Conhecimento em Bases de Dados, é definida como: "o processo não trivial de identificar nos dados, padrões válidos, novos, potencialmente utilizáveis e compreensíveis".

Pesquisadores na área de descoberta de conhecimento, atentam ao fato da mistura realizada entre os termos. Silva, Peres e Boscarioli (2016) enfatizam que eles não podem ser compreendidos como sinônimos, dado que a MD não consegue possibilitar, sozinha, a geração de conhecimento.

Cios et al. (2007) confirma Silva, Peres e Boscarioli (2016) ao dizer que a MD é apenas uma parte do processo. Eles sinonimizam os termos descoberta de conhecimento e processo de DCBD. Porém usam uma abordagem diferente de Silva, Peres e Boscarioli (2016).

Eles discorrem sobre o processo de descoberta de conhecimento utilizando o modelo proposto por Fayyad et al. (1996), e definem um modelo geral (figura 3) que obtém os dados de uma base de dados e os submetem a diferentes etapas, definidas pelo processo em uso, incluindo a MD. Esses passos procuram completar as etapas de descoberta, apoiadas por tarefas de MD e, apresentar padrões ocultos, ao final (CIOS et al., 2007).

Figura 3 – Modelo genérico de processo de Descoberta de Conhecimento

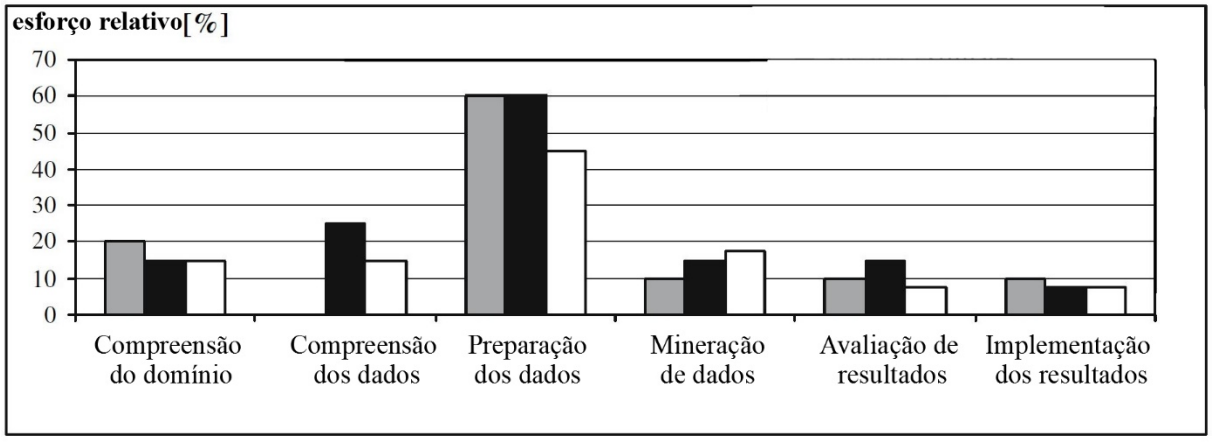


Fonte: Cios et al. (2007, p. 11, tradução nossa)

Pal e Jain (2005 apud CIOS et al., 2007) analisaram a execução do processo de DCBD e concluíram que cada etapa consume tempo e esforço variados (figura 4). A análise deles leva a concluir que as etapas inicias do processo possuem maior dispêndio de esforço, com atenção

especial à “Preparação dos Dados”. Estas fases serão melhor explicitadas na seção Processos de DCBD.

Figura 4 – Esforço relativo no processo de DCBD



Fonte: Cios et al. (2007, p. 19, tradução nossa)

Ainda nesse tópico, faz-se importante destacar que o processo de DCBD é uma atividade dependente do ser humano. Um problema de geração de conhecimento depende, no mínimo de dois papeis, um entendedor do negócio e um especialista em DCBD. Eles são dependentes entre si, pois os objetivos de descoberta a serem alcançados são definidos pelo entendedor do negócio, enquanto que as decisões no âmbito técnico são tomadas pelo especialista de DCBD.

Ao final de um processo de DCBD, ambos se reúnem para avaliar o conhecimento. O especialista técnico é responsável por prover uma visualização inteligível das informações e o de negócio, valida o conhecimento gerado (GOLDSCHMIDT; PASSOS; BEZERRA, 2015).

2.4.1 Processos de DCBD

Conforme Cios et al. (2007) os processos de DCBD, começaram a surgir a partir da década de 1990, embora Silva, Peres e Boscarioli (2016) revelem que os estudos sobre análises de dados, para descoberta de padrões, datem da década de 1960.

Azevedo e Santos (2008) e Cios et al. (2007) classificam dois ambientes onde os processos, relacionados adiante, surgiram. A academia é indicada como o primeiro ambiente. O objetivo era criar um modelo de DCBD geral, independente de ferramentas, que pudesse ser aplicado em diferentes cenários e que resultasse numa padronização, tal como existe nos bancos de dados relacionais. O segundo ambiente, o industrial, visava a aplicação prática da descoberta de conhecimento no cenário profissional, onde auxiliaria empresas no processo de tomada de decisão.

Esses processos possuem semelhanças entre si, permitindo realizar equivalências entre as etapas que os compõem. A diferença elementar entre eles, reside na quantidade de etapas e o propósito delas (CIOS et al., 2007).

A análise dos processos neste estudo consiste na caracterização de cada um deles. Sua escolha foi dada de acordo com a popularidade que eles possuem na literatura (AZEVEDO; SANTOS, 2008; SHAFIQUE; QAISER, 2014).

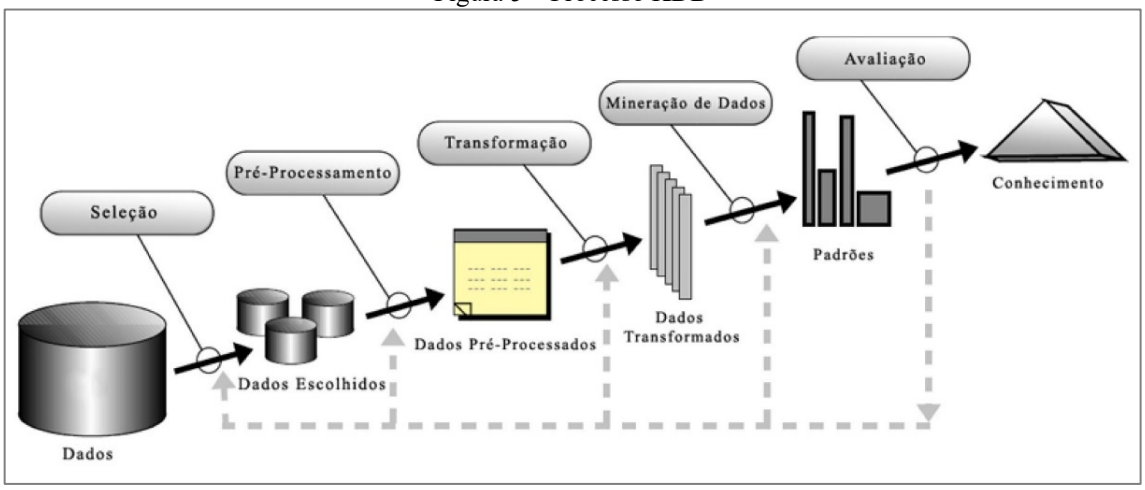
a) KDD

A princípio, deve-se ressaltar que o KDD além de dar nome a atividade de DCBD também é uma metodologia para obtenção de conhecimento, proposta por Fayyad et al. (1996). Esse é o processo criado na academia que incentivou a criação de outros, como o CRISP-DM, explicado posteriormente.

Sobre sua popularidade e análise de pontos fracos Cios et al. (2007, p. 18, tradução nossa) declaram que ele é "o processo mais popular e mais citado; provê detalhadas descrições técnicas com relação a análise de dados, porém não é ideal para aplicações em problemas comerciais".

Ele é composto por nove fases, sumarizadas na figura 5, conforme Fayyad et al. (1996), Cios et al. (2007) e Shafique e Qaiser (2014):

Figura 5 – Processo KDD



Fonte: Fayyad et al. (1996, p. 41)

Entendimento do Negócio: A primeira etapa envolve o aprendizado sobre o domínio da aplicação do processo de DCBD, também define os objetivos da descoberta de conhecimento segundo a visão dos usuários.

Integração: Consiste em selecionar um conjunto de dados, ou uma amostra do todo, com a finalidade de ser submetido como entrada para o processo de descoberta de conhecimento.

Limpeza e Pré-processamento: Esta etapa trata da remoção de dados inúteis ou anômalos, conhecidos por *outliers*, como por exemplo: data de nascimento igual a *NULL* ou idade com valores negativos.

Transformação: Nesta etapa são realizadas as modificações quanto ao modo como os dados devem se encontrar para alimentarem a etapa de mineração. Se necessário for, métodos de redução e normalização podem ser empregados afim de padronizar os dados selecionados na segunda etapa.

Escolha da Tarefa de Mineração de Dados: Essa etapa determina o resultado da sexta etapa, pois as tarefas empregadas no processo de mineração possuem um ou mais algoritmos relacionados. A escolha da tarefa relaciona-se com os objetivos que foram definidos na etapa inicial.

Escolha do Algoritmo de Mineração de Dados: Com base na tarefa escolhida no passo anterior, é possível determinar qual ou quais algoritmos poderão ser aplicados na próxima etapa. Escolhido o algoritmo, os seus parâmetros podem ser configurados para a correta execução.

Mineração de Dados: A sétima etapa do KDD, busca gerar padrões nos dados analisados, com base nas escolhas feitas nas duas etapas anteriores. Nela, caso haja algum problema na geração dos padrões, pelo aspecto interativo e iterativo do processo, é possível retornar às fases anteriores para realização de correções.

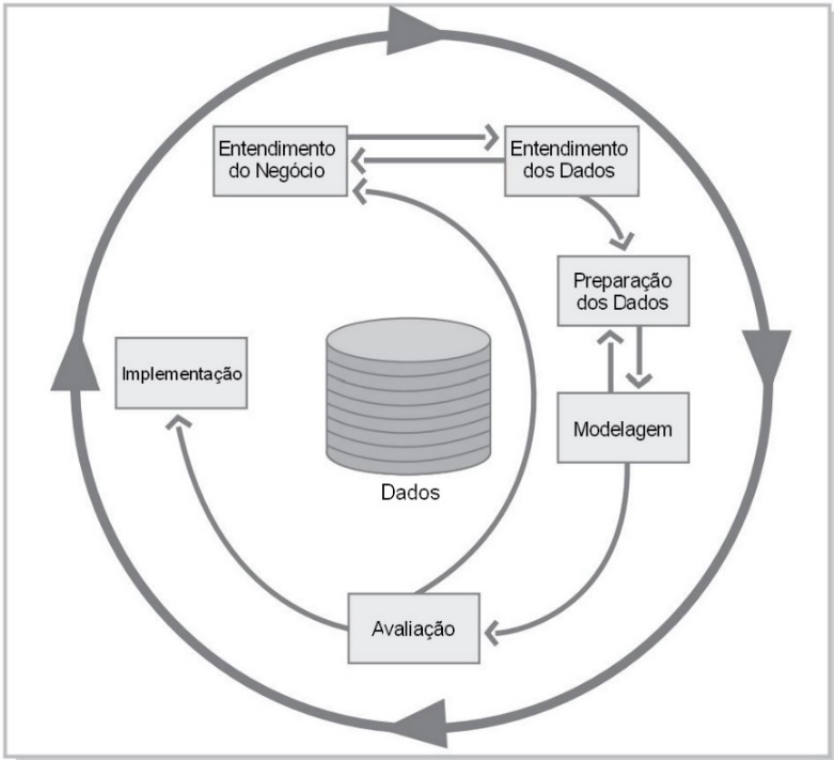
Interpretação/Avaliação: Aqui, os padrões extraídos são visualizados, analisados e interpretados. Caso necessário, o processo poderá ser retrocedido para qualquer umas das etapas precedentes.

Uso do Conhecimento Descoberto: O último passo do KDD busca avaliar o conhecimento descoberto, aplicando-o ao propósito de sua geração ou disponibilizando os resultados para quem se fizer necessário.

b) CRISP-DM

O *Cross-Industry Standard Process for Data Mining* (CRISP-DM) ou Processo Padrão de Vários Segmentos de Mercados para Mineração de Dados é composto por seis fases que podem interagir entre si, ressaltando o caráter interativo de um processo de descoberta de conhecimento. As fases (figura 6), estão dispostas de forma cíclica, o que enfatiza a iteratividade (CAMILO; SILVA, 2009; CHAPMAN et al, 2000; SHAFIQUE; QAISER, 2014).

Figura 6 – Processo CRISP-DM



Fonte: Chapman et al. (2000, p. 10, tradução nossa)

Lançado no final da década de 1990, a partir da união de esforços de empresas europeias: *Integral Solutions LTD*, *NCR Systems Engineering Copenhagen*, *DaimlerChrysler AG*, *SPSS Incorporation* e *OHRA Verzekeringen en Bank Groep B.V*, ele possui destaque por ser flexível e customizável. Isso permite sua aplicação em diferentes cenários que empregam a mineração de dados. O processo ainda é utilizado como metodologia padrão da ferramenta SPSS, pertencente à IBM (*International Business Machines*) (CIOS et al., 2007; IBM KNOWLEDGE CENTER, 2018).

Camilo e Silva (2009) e Cios et al. (2007) concordam que o CRISP-DM é o modelo de processo (ou metodologia) mais utilizado na área de DCBD com bons níveis de aceitação e aplicações bem-sucedidas. Piatetsky-Shapiro (2014) deixa claro que ele continua sendo a escolha mais popular dentre as demais metodologias de MD existentes, com maior utilização na América do Norte e Europa.

No que diz respeito à documentação, o CRISP-DM possui um guia robusto dividido em: uma visão geral da metodologia, um modelo de referência e um guia do usuário. O modelo de referência é um resumo das fases, tarefas e saídas de cada fase. O guia do usuário, apresenta o detalhamento do processo, ao especificar as tarefas e suas atividades, como também boas práticas de aplicação (CHAPMAN et al, 2000).

A seguir, tem-se a descrição de cada uma das fases do CRISP-DM segundo Camilo e Silva (2009), Shafique e Qaiser (2014) e Chapman et al. (2000):

Entendimento do Negócio: Na primeira etapa devem ser entendidas as particularidades do negócio que se beneficiará do processo de DCBD, e definidos os objetivos a serem alcançados nele.

Entendimento dos Dados: Na segunda etapa devem ser avaliados os dados necessários ao processo. Essa análise envolve a quantidade disponível, a estruturação desses dados, sua qualidade e relevância para o problema em questão.

Preparação dos Dados: A terceira fase tem o objetivo de preparar o conjunto de dados para que a etapa de modelagem possa ser aplicada. Nesta fase é realizada a limpeza dos dados, remoção de dados discrepantes, bem como organização deles consoante ao método de MD da etapa de modelagem.

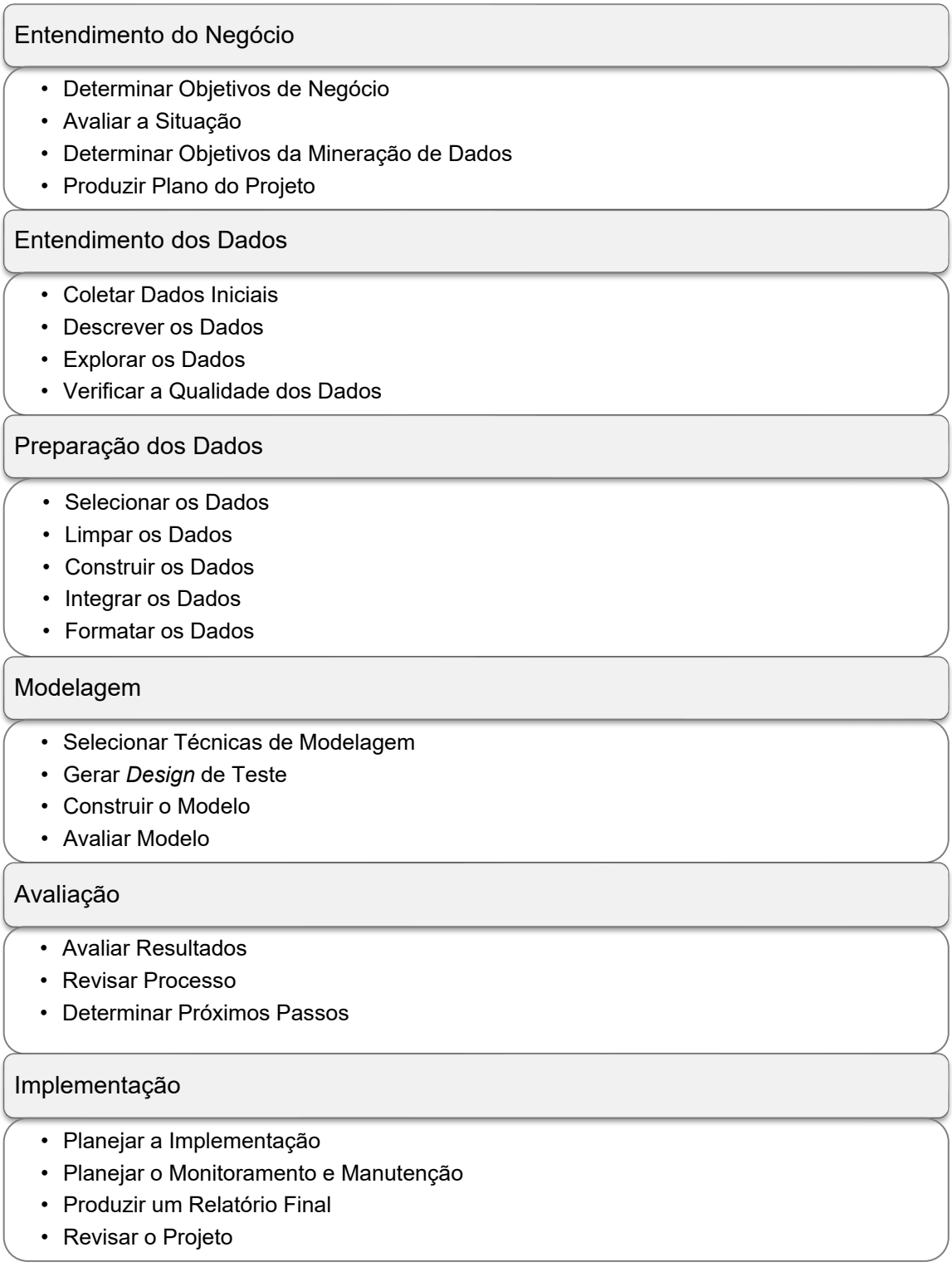
Modelagem: A etapa de modelagem é onde ocorre a MD, de fato. Nesse momento diferentes modelos podem ser gerados com base nos dados sob análise. Caso esses dados não estejam em conformidade com o método e algoritmo de MD escolhidos, eles são retrocedidos para a etapa anterior com o propósito de melhor refinamento.

Avaliação: A quinta fase do processo, consiste em avaliar, através de visualização auxiliada por ferramentas gráficas, se os modelos gerados na fase anterior atendem aos objetivos definidos no Entendimento do Negócio. Camilo e Silva (2009, p. 5) destaca que essa é uma "fase crítica do processo de mineração, [pois necessita da] [...] participação de especialistas nos dados, conhecedores do negócio e tomadores de decisão".

Implementação: A sexta e última etapa consiste na apresentação do conhecimento gerado a quem for preciso.

A figura 7, fornece um detalhamento das fases supracitadas, apresentando suas respectivas tarefas, utilizadas de acordo com a aplicação em um problema de DCBD:

Figura 7 – Fases e Tarefas CRISP-DM



Fonte: Chapman et al. (2000, tradução nossa)

2.4.1.1 Comparação dos Processos de DCBD

Estudos foram feitos, comparando os dois processos aqui descritos. O consenso entre os pesquisadores é que não existe um processo único que é o melhor para a descoberta de conhecimento. A escolha depende de quão adequado ele é em relação ao problema em questão

(CIOS et al., 2007). O quadro 2, ilustra essa comparação:

Quadro 2 – Comparação entre Processos de DCBD

Processo	KDD	CRISP-DM
Fases	1. Entendimento do Negócio	1. Entendimento do Negócio
	2. Integração	2. Entendimento dos Dados
	3. Limpeza e Pré-processamento	
	4. Transformação	3. Preparação dos Dados
	5. Escolha da Tarefa de Mineração de Dados	4. Modelagem
	6. Escolha do Algoritmo de Mineração de Dados	
	7. Mineração de Dados	
	8. Interpretação/Avaliação	5. Avaliação
	9. Uso do Conhecimento Descoberto	6. Implementação

Fonte: Shafique e Qaiser (2014) e CIOS et al. (2007)

2.4.2 Mineração de Dados

Para Silva, Peres e Boscarioli (2016, p. 10):

> a mineração de dados pode ser definida como um processo automático ou semiautomático de explorar analiticamente grandes bases de dados, com a finalidade de descobrir padrões relevantes que ocorrem nos dados e que sejam importantes para embasar a assimilação de informações importantes, suportando a geração de conhecimento.

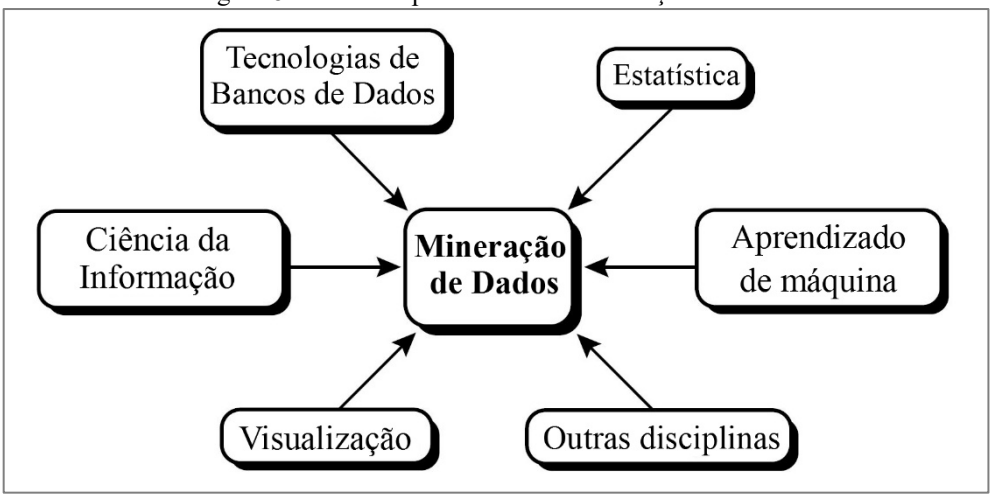
Galvão (2009) e Goldschmidt, Passos e Bezerra (2015) concordam ao dizer que a mineração de dados é a etapa mais importante do processo de DCBD. Fayyad et al. (1996), por sua vez dizem que a MD é importante, tanto quanto as outras etapas anteriores do processo de DCBD, pois elas são essenciais para que o conhecimento gerado possua conformidade. Segundo eles, um salto direto para e etapa de MD, pode provocar a descoberta de padrões sem validade, o que resultaria em conhecimento irrelevante.

Os dados submetidos à MD, necessitam estar livres de ruídos e irregularidades para que o objetivo dessa etapa seja concretizado adequadamente. Esse tratamento é executado nas etapas anteriores do processo de DCBD, principalmente na parte onde os conjuntos de dados são pré-processados. Ele é importante pois, a estrutura dos dados influencia na escolha de qual tarefa de mineração será usada para a descoberta dos padrões (SILVA; PERES; BOSCARIOLI, 2016).

Os dados podem ser de dois tipos: estruturados – aqueles, organizados em tabelas bidimensionais e, geralmente, armazenados em banco de dados – e não estruturados – dados de texto, por exemplo, que não possuem uma segmentação de suas informações, sistematizadas em variáveis e registros (AMARAL, 2016; SILVA; PERES; BOSCARIOLI, 2016).

A mineração de dados é conhecida sob outras nomenclaturas, como extração de conhecimento, descoberta de informação, processamento de padrões de dados e arqueologia de dados (CIOS et al, 2007). Vale salientar também o caráter interdisciplinar e a capacidade que a MD possui de se relacionar com outras áreas. Goldschmidt, Passos e Bezerra (2015) listam a Estatística, Aprendizado de Máquina[4], Inteligência Computacional e Reconhecimento de padrões. Han e Kamber (2006), também realizaram uma comparação conforme a figura 8, abaixo:

Figura 8 – Interdisciplinaridade em Mineração de Dados



Fonte: Han e Kamber (2006, p. 29, tradução nossa)

Finalmente, constata-se que a MD é uma atividade com possibilidade de aplicação a diversas áreas ou atividades que lidem com grandes quantidades de dados como: Astronomia, Bancos, Bioinformática, Campanhas Eleitorais, Contabilidade, Detecção de Crimes, Educação, Gerência de Relacionamento com o Cliente, Mecanismos de Buscas Virtuais, Medicina, Segurança e Telemarketing (CAMILO; SILVA, 2009; PIATETSKY-SHAPIRO; PARKER, 2006).

2.4.2.1 Tarefas

As tarefas do processo de DCBD, de acordo com Goldschmidt, Passos e Bezerra (2015) são operações inerentes à fase de MD, que guiam o modo como os padrões podem ser identificados nos conjuntos de dados. Elas são classificadas em: tarefas preditivas – que consistem em prever novos comportamentos nos dados, a partir da criação de um modelo baseado num conjunto de dados históricos e por meio de algoritmos especializados; e tarefas descritivas – que buscam revelar novas informações, sem a necessidade de um modelo anterior (GOLDSCHMIDT; PASSOS; BEZERRA, 2015; SILVA; PERES; BOSCARIOLI, 2016).

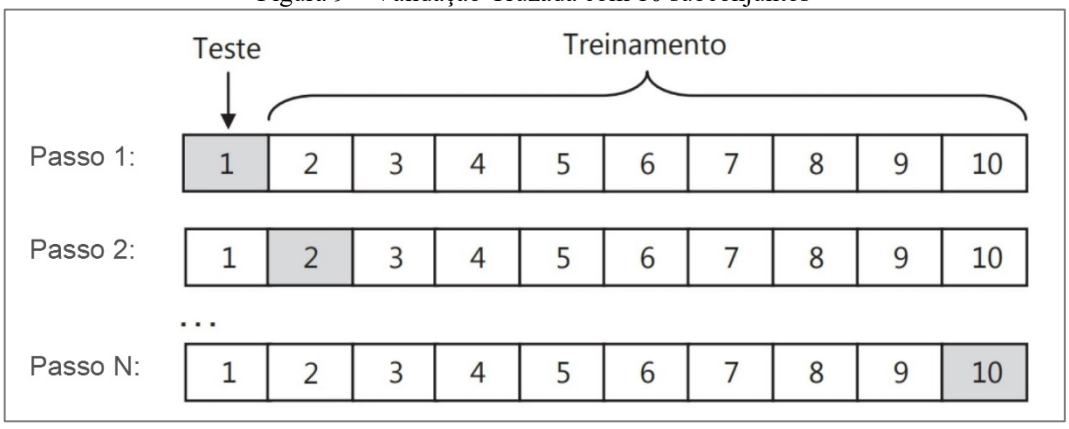
[4] *Machine Learning*

Nas tarefas preditivas, a base de dados é comumente dividida em dois subconjuntos, o primeiro é empregado na criação do modelo e o segundo, para teste. Ou seja, o conjunto de teste, avalia a qualidade do modelo criado, pois os registros ali encontrados, não participaram do treinamento (AMARAL, 2016; SILVA; PERES; BOSCARIOLI, 2016).

Dentre os métodos de partição da base de dados, dois se destacam: *Hold-out* e Validação Cruzada, ou *Cross Validation*. No *Hold-out*, os dados são divididos aleatoriamente, de modo que a porção direcionada a criação do modelo seja maior do que aquela voltada para o teste. Goldschmidt, Passos e Bezerra (2015) aponta uma razão de 2/3 para treinamento e 1/3 para teste. Esse método, todavia, pode falhar caso existam padrões na base de teste que estiveram ausentes na base de treinamento e, portanto, não foram "conhecidos" pelo modelo (AMARAL, 2016; SILVA; PERES; BOSCARIOLI, 2016).

A validação cruzada, ao contrário do *Hold-out*, executa o treinamento e teste em todo o conjunto de dados durante determinado número de vezes. Isso aumenta a qualidade e confiabilidade do modelo criado. O funcionamento desse método consiste em dividir a base original em *K* conjuntos, fazendo com que, enquanto um dos *K* conjuntos é tido como teste, o restante serve como treinamento. Esse mecanismo finaliza quando toda a base tenha sido, respectivamente, treinada e testada. Análogo ao *Hold-out*, existe uma convenção do número de divisões na validação cruzada; por padrão parametriza-se $K = 10$ (AMARAL, 2016; GOLDSCHMIDT; PASSOS; BEZERRA, 2015). A figura 9 demonstra esse método:

Figura 9 – Validação Cruzada com 10 subconjuntos



Fonte: Castro e Ferrari (2016, p. 158, adaptado)

As principais tarefas de mineração de acordo com Amaral (2016), Camilo e Silva (2009), Dias (2001), Goldschmidt, Passos e Bezerra (2015) e Silva, Peres e Boscarioli (2016) são:

a) Classificação:

Como outras tarefas citadas adiante, a classificação também faz parte do grupo de tarefas preditivas. Ela é considerada um dos métodos mais comuns, populares e importantes da MD. Conta com diversos algoritmos que a implementam (AMARAL, 2016; CAMILO; SILVA, 2009; GOLDSCHMIDT; PASSOS; BEZERRA, 2015).

Silva, Peres e Boscarioli (2016, p. 79) explicam que a classificação é "o processo pelo qual se determina um mapeamento capaz de indicar a qual classe pertence qualquer exemplar de um domínio sob análise, com base em um conjunto de dados já classificado".

Fundamentado nisso, os dados são organizados em dois grupos onde um deles consiste num vetor de atributos (registro), e o outro agrupa rótulos que classificam, isto é, categorizam cada registro. Os atributos que integram o registro são chamados de previsores e os rótulos são os atributos-alvo ou classes (GOLDSCHMIDT; PASSOS; BEZERRA, 2015; SILVA; PERES; BOSCARIOLI, 2016).

No ponto de vista de Silva, Peres e Boscarioli (2016) a tarefa de classificação é dividida em binária e de múltiplas classes. Classificação binária ocorre quando a quantidade de classes de um conjunto de dados é igual a 2. Classificação de múltiplas classes acontece quando a quantidade de classes é maior que 2. A título de exemplo, é um problema de classificação binária aquele, no qual pretende-se verificar se um indivíduo, com determinado perfil econômico, será um bom ou mau pagador. Um problema de classificação de múltiplas classes dá-se quando objetiva-se classificar um ser vivo como sendo: inseto, mamífero, ave, réptil ou peixe (GRANATYR, 2016).

Já que a classificação é uma tarefa preditiva, é esperado que um modelo de predição seja criado quando esse método for executado. Todavia é necessário que a qualidade do resultado seja julgada para determinar se ele apresenta confiabilidade e acurácia relevantes. Tornando factível a tomada de decisões com base na categorização realizada. Essa avaliação é realizada através da matriz de confusão (SILVA; PERES; BOSCARIOLI, 2016).

O objetivo da matriz de confusão é avaliar a acurácia do modelo de classificação – ou modelo preditivo – criado, evidenciando os erros e acertos dele em relação aos atributos-alvo. Essa avaliação e feita através de quatro tipos de resultados possíveis: Verdadeiro Positivo (VP), Falso Positivo (FP), Verdadeiro Negativo (VN) e Falso Negativo (FN) (AMARAL, 2016; SILVA; PERES; BOSCARIOLI, 2016).

Amaral (2016), explica que os Verdadeiros Positivos e Verdadeiros Negativos são aqueles registros que tiveram a correta classificação, enquanto que os Falsos Positivos e Falsos Negativos informam os erros do classificador.

A matriz de confusão costuma ser entendida a partir da classificação binária, em que uma matriz quadrada de tamanho 2 representa as classes reais dos registros e as classes que o modelo, através de um classificador, foi capaz de prever. As linhas, representam as classes verdadeiras enquanto que as colunas, as classes preditas. A figura 10, demonstra a estrutura de uma matriz de confusão e o quadro 3 especifica o que cada tipo de resultado deve retornar, com base na organização da ilustração:

Figura 10 – Matriz de Confusão: Classificação Binária

		Classe Predita	
		Classe A	*Classe B*
Classe Real	*Classe A*	Verdadeiro Positivo (VP)	Falso Negativo (FN)
	Classe B	Falso Positivo (FP)	Verdadeiro Negativo (VN)

Fonte: Silva, Peres e Boscarioli (2016, p. 131, adaptado)

Quadro 3 – Tipos de retornos da Matriz de Confusão

Resultado	**Descrição**
VP	Registros classificados como classe A e foram corretamente classificados (preditos) pelo classificador.
FP	Registros classificados como classe B, mas foram preditos como classe A pelo classificador.
VN	Registros classificados como classe B e foram corretamente preditos pelo classificador.
FN	Registros classificados como classe A, mas foram preditos como classe B pelo classificador.

Fonte: Vieira (2018)

Deve-se destacar que uma matriz de confusão é aplicável, de igual modo, a problemas de classificação de múltiplas classes. Neste caso, aumenta-se igualmente a quantidade de linhas e colunas, para que haja tantas linhas – e colunas – quanto a quantidade de classes (SILVA; PERES; BOSCARIOLI, 2016).

Relacionados à matriz de confusão e seus resultados, alguns avaliadores são aplicados. Dentre eles estão: a avaliação de verdadeiros positivos que é dada pela razão entre a taxa de VP e a soma das taxas de VP e FN; a avaliação de falsos positivos que é dada pela razão da taxa de FP e a soma das taxas de VN e FP; o índice de precisão que é calculado pela razão entre a taxa de VP e a soma das taxas de VP e FP (SILVA; PERES; BOSCARIOLI, 2016).

Outros dois avaliadores podem ser usados para medir a acurácia e a qualidade de um modelo de classificação são: a taxa de característica do receptor, que define quão perfeito ou quão ineficaz é um classificador. Quanto mais próximo de 1, mais perfeição existe, entretanto quanto mais próximo de 0, pior será o classificador. Outro avaliador a ser citado é o coeficiente de Kappa, que mede o nível de concordância entre os resultados obtidos por um classificador

(BALTAR; OKANO, 2017; SILVA; PERES; BOSCARIOLI, 2016).

b) Regressão:

Enquanto a classificação trabalha com dados conhecidos como categóricos, a regressão utiliza dados numéricos, desempenhando o mesmo propósito da tarefa anterior. As entradas para uma tarefa de regressão consistem num intervalo de valores numéricos, que com base em uma função, determina-se um novo valor. Essa tarefa é comumente utilizada em previsões de limite de cartões de crédito ou em avaliações de elevação do fluxo de tráfego numa rede, dada as variações de velocidade.

c) Associação:

A tarefa de associação pertence às tarefas descritivas e não exige um modelo de dados para identificar padrões. Ela fundamenta-se na análise de um conjunto de dados e estabelece relações desconhecidas entre eles, conhecidas como regras de associação. Os autores supracitados, com destaque para Camilo e Silva (2009), Goldschmidt, Passos e Bezerra (2015) e a adição de Silberchatz, Korth e Sudarshan (2012), ilustram que a associação gera regras do tipo **SE <condição(ões)> ENTÃO <resultado(s)>**.

d) Agrupamento (Clusterização)

O propósito do agrupamento reside no aspecto, de que dado um conjunto de dados onde todos eles não possuem uma segmentação, serão agrupados com base nas suas semelhanças. Quanto mais semelhante um dado for em relação ao outro, mais próximo ele estará dos seus semelhantes. Isso permite que um grupo possua apenas elementos com características comuns, criando outros grupos para aqueles que forem diferentes dos primeiros, porém semelhantes entre si. Amaral (2016) destaca que, caso um uma informação não apresente nenhuma semelhança aos agrupamentos criados, ela é identificada como *outlier*. Essa tarefa possibilita agrupar as informações de maneira não linear.

2.4.2.2 Técnicas

As técnicas, conforme citado nos processos de DCBD, ligam-se diretamente às tarefas. Isso demonstra que cada tarefa, possui uma ou mais técnicas associadas que realizam o processo de MD por meio de algoritmos específicos.

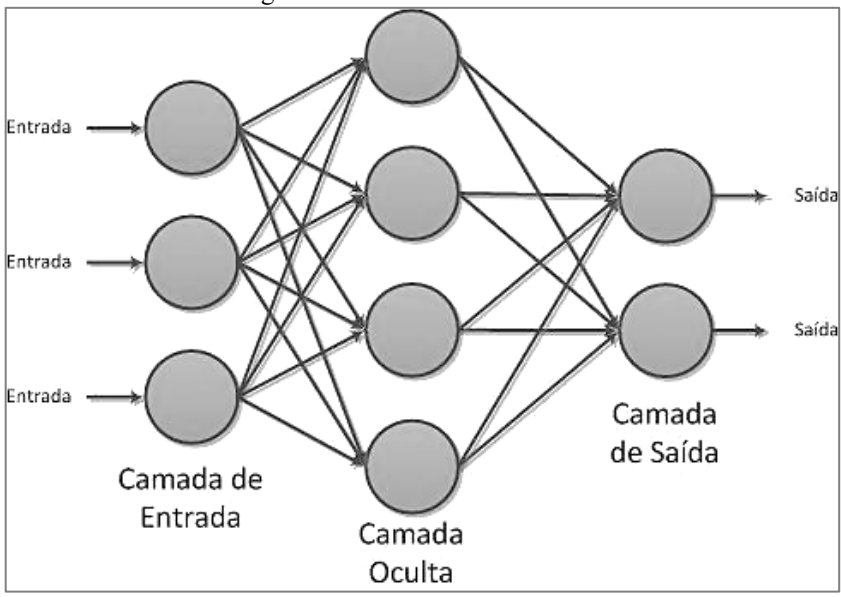
a) Redes Neurais Artificiais (RNAs):

Esta técnica é constituída de algoritmos matemáticos que simulam o funcionamento de neurônios biológicos e como Goldschmidt, Passos e Bezerra (2015, p. 30) indicam: “[podem] aprender padrões diretamente a partir dos dados por meio de um processo de repetidas apresentações dos dados à rede”. Amaral (2016) comenta que RNAs possuem como estrutura

básica, uma camada de entrada para os dados, uma camada de processamento oculta, onde as conexões e o aprendizado do algoritmo são realizados, e uma camada de saída, que extrai o resultado do processamento.

Por fim, Camilo e Silva (2009) e Goldschmidt, Passos e Bezerra (2015) ressaltam a característica das RNAs lidarem bem com valores errôneos, a partir da habilidade de melhoramento dos algoritmos.

Figura 11 – Rede Neural Artificial



Fonte: Amaral (2016, p. 43)

b) Algoritmos Genéticos:

Para Camilo e Silva (2009), Dias (2001) e Goldschmidt, Passos e Bezerra (2015) esta técnica é fundamentada nos aspectos de seleção natural e evolução. Ela está voltada para questões de busca e otimização onde uma "população" de dados passa por um processo, no qual aqueles com maior poder de predição serão selecionados para compor o modelo de classificação. Os dados mais frágeis, pela lei do mais forte, serão "extintos".

Goldschmidt, Passos e Bezerra (2015) no entanto, chamam a atenção ao afirmarem que esta técnica demonstra uma solução aproximada ou aceitável da ideal, mas apesar disso, ainda constitui uma alternativa adequada para problemas complexos.

c) Descoberta de Regras de Associação:

Segundo Dias (2001) apoiada por Camilo e Silva (2009), esta técnica implementa a tarefa de Associação ao empregar as regras de associação com o formalismo **SE <condição$_1$, ..., condição$_n$> ENTÃO <resultado$_1$, ..., resultado$_n$>** e determinar o grau de confiança da regra, que define quão exata será a associação estabelecida.

d) Árvores de Decisão:

Amaral (2016) observa que árvores de decisão é uma técnica popular na tarefa de classificação. Ela é estruturada a partir de uma árvore, na qual a raiz é ramificada em nós que resultam em folhas, as quais representam o atributo-alvo que determina a qual categoria, o registro em análise pertence. Dias (2001) acrescenta que o propósito das árvores consiste em segmentar as classes de dados.

Ainda para Amaral (2016), a maneira pela qual a árvore é percorrida é dependente do algoritmo empregado e dos dados analisados.

2.4.2.3 Algoritmos

Os algoritmos que implementam as técnicas de MD, são divididos quanto ao tipo de aprendizado que podem exercer. Algoritmos supervisionados funcionam a partir do treinamento de um algoritmo que relaciona um conjunto de atributos a uma classe através de uma função encontrada por ele. O treinamento é realizado com base num conjunto histórico de dados, particionados em treino e teste (SILVA; PERES; BOSCARIOLI, 2016).

O resultado desse processo é um modelo de classificação, que posteriormente é validado numa atividade onde os dados da porção de teste, desconhecidos pelo algoritmo, lhes são apresentados a fim de que ele possa realizar uma predição. Portanto, aprendizado supervisionado exige a existência de um modelo preditivo para que o conhecimento seja gerado. (GOLDSCHMIDT; PASSOS; BEZERRA, 2015; SILVA; PERES; BOSCARIOLI, 2016).

Algoritmos não-supervisionados não requerem uma saída determinada, isto é, um modelo; pelo contrário, eles buscam identificar padrões consoante ao agrupamento e/ou semelhança das informações. Goldschmidt, Passos e Bezerra (2015) ainda acrescentam que este tipo de aprendizado dispensa a partição dos dados em treinamento e teste.

O quadro 4, a seguir, agrupa um conjunto de algoritmos indicados para cada técnica, baseado em Camilo e Silva (2009), Dias (2001) e Goldschmidt, Passos e Bezerra (2015):

Quadro 4 – Técnicas e algoritmos de Mineração de Dados

Técnica	**Exemplos de Algoritmos**
Redes Neurais Artificiais	*Back-Propagation* e *Multilayer Perceptron*
Algoritmos Genéticos	*Rule Envolver*, Algoritmo de *Hillis* e GA-*Nuggets*.
Árvores de Decisão	*Classification and Regression Trees* (CART), *C4.5* e *ID-3*
Descoberta de Regras de Associação	*Apriori* e *FP-Growth*.

Fonte: Camilo e Silva (2009), Dias (2001) e Goldschmidt, Passos e Bezerra (2015)

2.4.2.4 Ferramentas

Na década de 2010, o mercado oferta diversas ferramentas destinadas à MD e descoberta

de conhecimento. Importante salientar que existem muitas opções disponíveis para utilização. O *site* mantido por Piatesky-Shapiro (*KDnuggets*) realizou uma listagem com mais de 44 tipos específicos de ferramentas para MD. Uma análise exaustiva dessa listagem exigiria um estudo somente para este fim (PIATETSKY-SHAPIRO, 2018).

Com o intuito de evidenciar as ferramentas e suas características, foram escolhidas duas comerciais e duas *open source*. Elas conseguem abranger a maioria das técnicas de MD disponíveis até o ano de 2018 e sua menção é comum na literatura (AMARAL; 2016; CAMILO; SILVA, 2009; GOLDSCHMIDT; PASSOS; BEZERRA, 2015; PIATETSKY-SHAPIRO, 2018). O detalhamento é demonstrado no quadro 5, adiante:

Quadro 5 – Ferramentas de Mineração de Dados

Ferramenta	Tipo	Principais Características
Enterprise Miner	Comercial	• Possui estabilidade e aceitação no mercado; • Oferece ferramentas voltadas à descoberta de conhecimento que auxiliam no processo de tomada de decisão; • Suporta árvores de decisão, redes neurais e modelos de regressão, por exemplo.
Oracle Data Mining	Comercial	• Possui integração com o Sistema de Gerenciamento de Banco de Dados (SGBD) *Oracle*; • Oferece suporte a análises preditivas, mineração de texto e análises estatísticas; • Suporta utilização da linguagem de programação R.
Orange	*Open Source*	• Oferece suporte à linguagem Python; • Pode ser empregada tanto em MD como em mineração de texto; • Suporta as tarefas de classificação, regressão, agrupamento e associação; • Disponibiliza treinamento para operacionalização da ferramenta; • Baseada no conceito de arrastar e soltar.
Weka 3	*Open Source*	• Desenvolvida em plataforma Java; • Definida como um conjunto de aplicações; • A aplicação dos algoritmos pode ser feita na própria ferramenta ou em programas aplicativos, escritos em Java; • Oferece suporte ao pré-processamento de dados e às tarefas de associação, classificação, regressão e agrupamento; • Oferece diversidade de algoritmos que implementam técnicas de MD como, kNN, Redes Neurais, *Support Vector Machines* e Árvores de Decisão; • Oferece também utilitários integrados para etapa de visualização; • Possui treinamento gratuito oficial.

Fonte: 1&1 Company (2017), Camilo e Silva (2009), Oracle (2018), SAS Institute (2018) e University of Waikato (2018)

3 RESULTADOS

Esta seção evidencia os resultados obtidos ao longo dessa pesquisa. Para isso ela foi segmentada para prover uma linearização das etapas executadas, devido ao fato de que o processo de DCBD ter sido executado de forma iterativa.

A respeito da revisão bibliográfica, foi atingido embasamento teórico necessário para compreensão dos ATs. Pôde-se tomar conhecimento de como eles são analisados pelas entidades responsáveis e identificar que o objetivo fundamental desses órgãos, centraliza-se na preservação da vida e redução das consequências desses eventos.

Semelhantemente, foi alcançado conhecimento sobre DCBD e MD. Buscou-se compreender os principais métodos de MD, detalhando-se a tarefa de classificação e discorrendo-se sobre a técnica de árvores de decisão, aplicadas no processo de descoberta de conhecimento do trabalho.

Por fim, o conhecimento adquirido foi posto em prática, através da aplicação de um processo específico de DCBD, auxiliado por uma ferramenta computacional para MD com suporte à tarefa, técnica e algoritmo escolhidos.

3.1 Trabalhos Correlatos

O detalhamento dos resultados começou ainda na primeira parte da pesquisa (TCC I), momento no qual foram investigados trabalhos que correlacionavam acidentes de trânsito, clima e descoberta de conhecimento em bases de dados. Essas produções, nortearam o desenvolvimento deste estudo e confirmaram sua relevância. As principais são de: Galvão, (2009); Agarwal, Maze e Souleyrette (2005); Reis (2014) e Oliveira (2012).

3.2 Local de Estudo

A análise da literatura e exame da região entre Goiânia e o Distrito Federal, possibilitou determinar, com mais precisão, o local estudado nesta pesquisa.

Pela razão da BR-060 passar por dentro da cidade de Goiânia (GO), foi determinado um ponto, no qual as ocorrências de ATs preservassem sua identidade de acidentes tipicamente rodoviários. Inicialmente definiu-se que, em Goiânia, o quilômetro inicial seria o 140 e no Distrito-Federal (DF), o quilômetro 0; obedecendo a natureza rodoviária dos ATs.

Após mais investigações, verificou-se que esse intervalo poderia ser refinado. Dessa forma, o km 0 foi trocado pelo 8, no DF, e o km 140, substituído pelo 137, em GO; definindo portanto, o local específico para ser analisado durante a descoberta de conhecimento.

3.3 Avaliação das Ferramentas de Mineração de Dados

Como parte dos objetivos deste trabalho, foi preciso avaliar qual ferramenta de MD, melhor atenderia a execução do processo de DCBD. Assim, uma comparação foi realizada considerando, entre outros quesitos: a gratuidade da ferramenta, detalhamento e disponibilidade da documentação, nível de popularidade e utilização, usabilidade, interface, arcabouço de funcionalidades oferecidas, liberdade de parametrização dessas funcionalidades e desempenho.

Ao considerar a gratuidade, das quatro ferramentas pesquisadas, duas foram aprovadas: *Orange Data Mining*[5] (versão 3.14) e *Weka 3*[6] (versão 3.8.2). Logo após, foi posto em pauta o fator documentação. Ambas proveem documentação detalhadas que explicam suas características e funcionalidades. Verificou-se, que a documentação da *Orange Data Ming* é visualmente mais organizada, facilitando a leitura. Por outro lado, a *Weka 3* conta com um manual para cada uma de suas versões e um apêndice gratuito, disponível no site oficial da ferramenta.

Ainda sobre a documentação, a *Weka 3* possui variabilidade de material não oficial como: artigos de *sites* eletrônicos e vídeos tutoriais, que explicam como utilizá-la. Encontrou-se também trabalhos científicos descrevendo a aplicação de MD e de DCBD e a avaliação de algoritmos de aprendizado de máquina, com a *Weka 3* como mecanismo computacional principal para a realização das atividades; observado no trabalho correlato de Reis (2014).

Quanto à popularidade, verificou-se que a *Weka 3* se destaca em relação a *Orange Data Mining*. Houve certa dificuldade na busca por exemplos de uso desta última. Importante enfatizar, que ambas as ferramentas possuem interface gráfica e documentação oficial em inglês, no entanto é possível encontrar alguns materiais em português.

A respeito da usabilidade, a *Orange Data Mining* apesar de intuitiva, não dispensa um aprendizado mínimo na operação de suas funcionalidades. A *Weka 3*, perde nesse ponto, mesmo assim, não apresenta alta complexidade para utilização, porém, assim como a outra ferramenta, é necessário um aprendizado de como operá-la, ainda que introduza apenas os conceitos introdutórios de utilização.

Sobre a interface, a *Orange Data Mining* possui melhor apresentação que a *Weka 3*, embora as duas tenham curva de aprendizado semelhantes. A seguir pode-se acompanhar uma comparação das duas ferramentas. Para melhor organização primeiro será discorrido sobre *Orange Data Mining*, em seguida sobre a *Weka 3*.

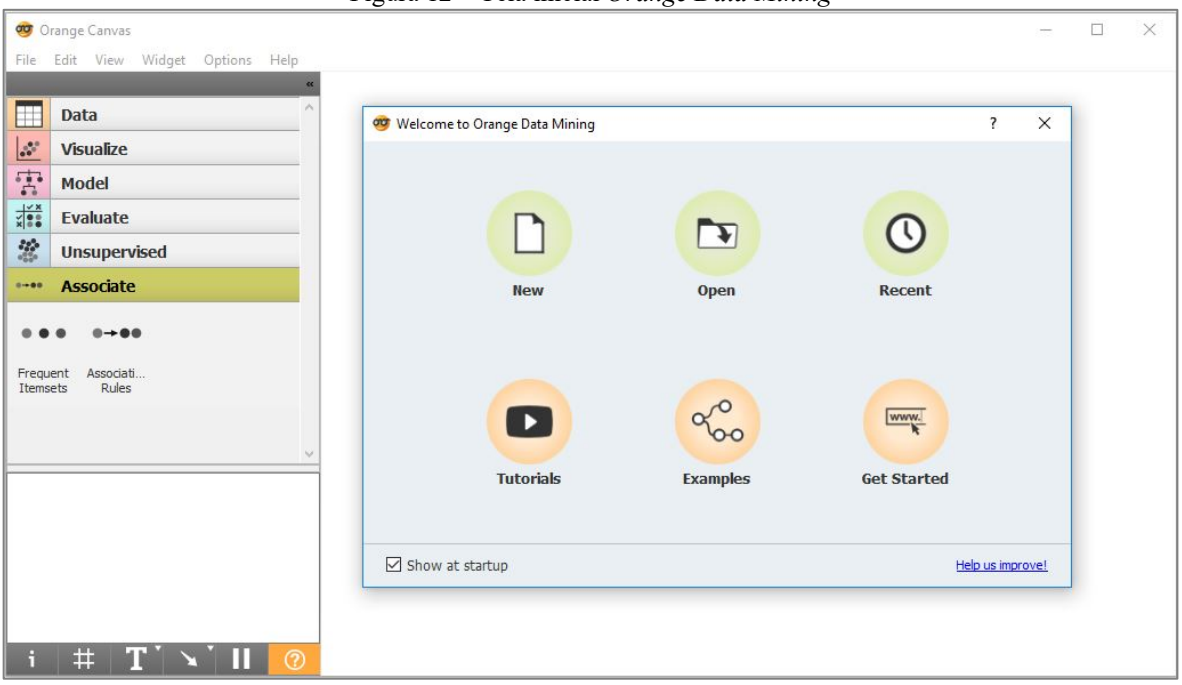
---

[5] Disponível em: https://orange.biolab.si/download/
[6] Disponível em: https://www.cs.waikato.ac.nz/ml/weka/downloading.html

a) *Orange Data Mining*

A figura 12 exibe a tela inicial da ferramenta e uma tela de bem-vindo autoexplicativa, na qual o usuário pode criar um novo projeto, abrir um existente ou recentemente aberto, acessar tutoriais em vídeo, visualizar exemplos de *workflow* (fluxo de trabalho) ou acessar a documentação. À esquerda é possível ver as categorias que agrupam as funcionalidades da ferramenta, os *widgets*.

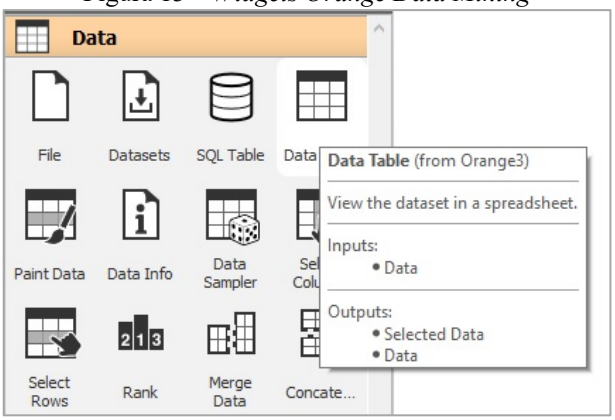
Figura 12 – Tela inicial *Orange Data Mining*



Fonte: Vieira (2018)

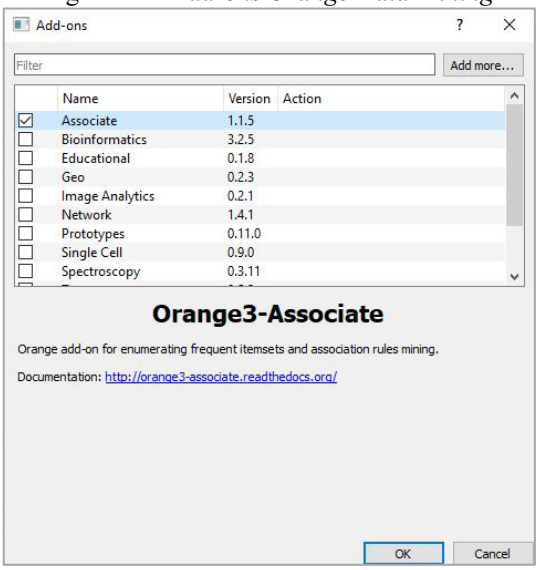
A figura 13 exibe as funcionalidades da categoria *Data* (Dados). Ao passar o *mouse* sobre qualquer uma dessas funcionalidades é mostrado ao usuário: o nome, breve descrição e as entradas e saídas do *widget.* A *Orange Data Mining* oferece outras categorias de funcionalidades, não estão instaladas por padrão. Existe uma opção no menu *Options*, chamada *Add-ons*, por onde elas podem ser obtidas (figura 14).

Figura 13 – *Widgets Orange Data Mining*



Fonte: Vieira (2018)

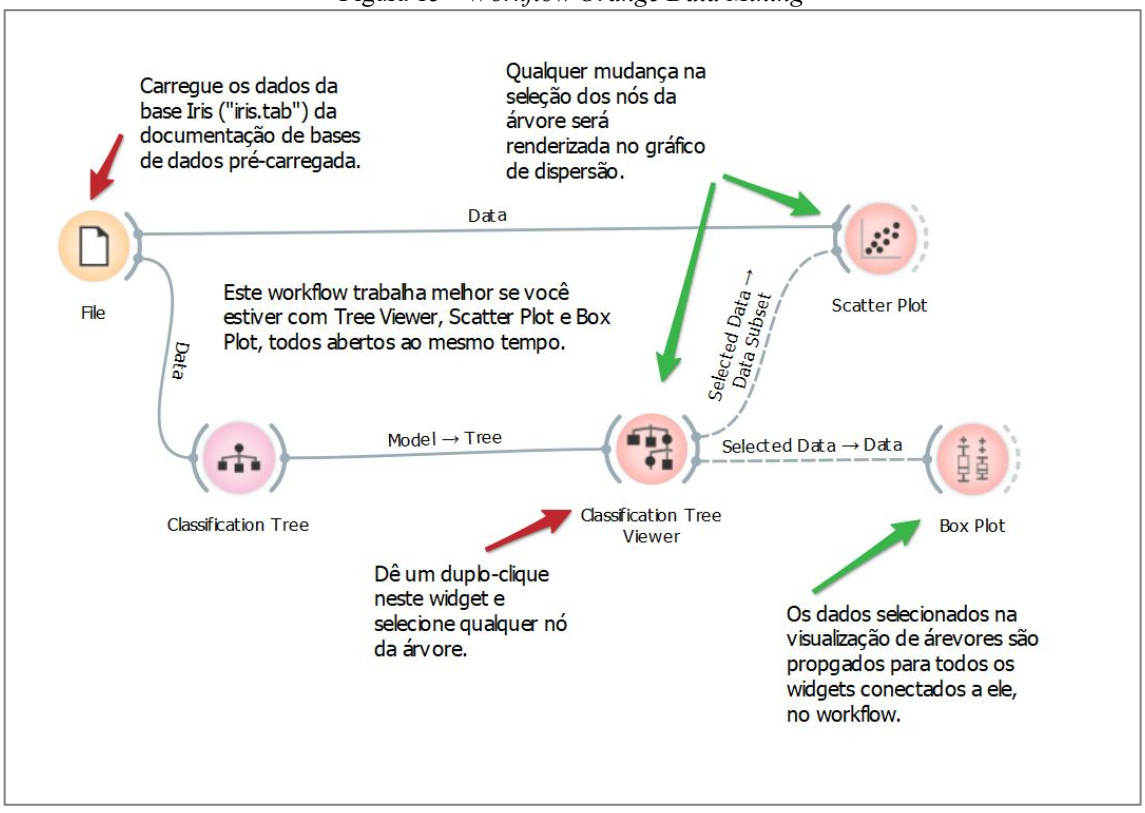
Figura 14 – *Add-ons Orange Data Mining*



Fonte: Vieira (2018)

A figura 15, ilustra o exemplo de um *workflow* obtido no item *Examples*, localizado na tela de bem-vindo.

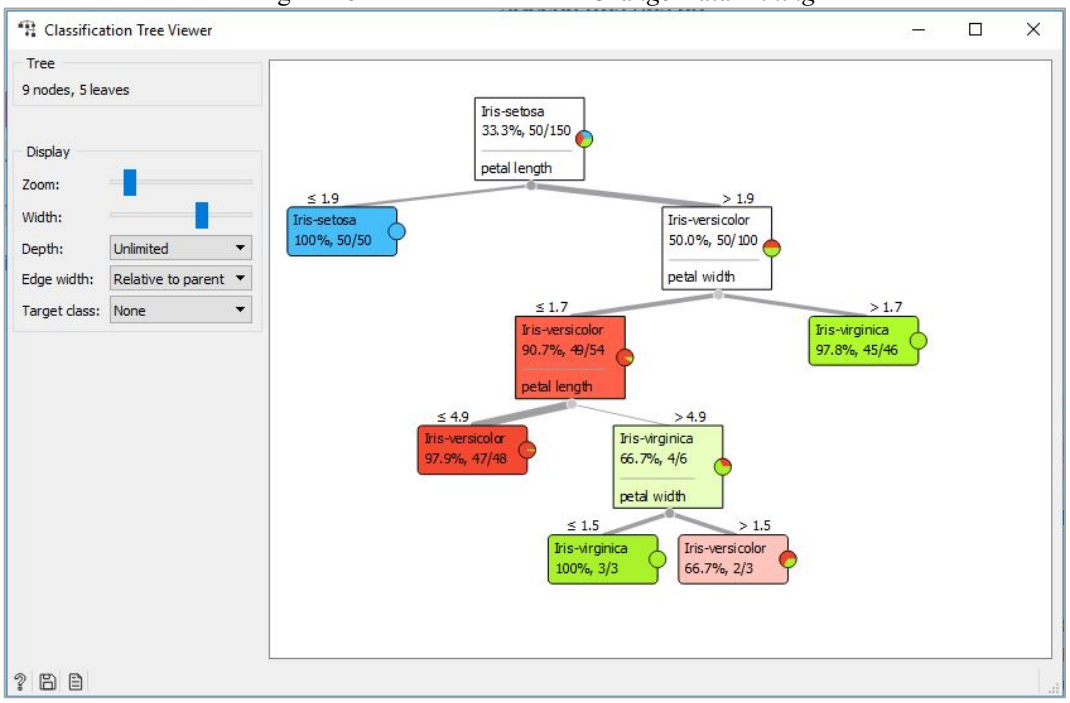
Figura 15 – *Workflow Orange Data Mining*



Fonte: Vieira (2018)

Considerando que a figura 15 faz referência à classificação com árvores de decisão, optou-se por mostrar como a *Orange Data Mining* apresenta os resultados dessa técnica de MD. A figura 16, exibe uma árvore de decisão ao centro e no lado esquerdo, as configurações de visualização.

Figura 16 – Árvore de Decisão *Orange Data Mining*



Fonte: Vieira (2018)

b) *Weka 3*

Enquanto a *Orange Data Mining* se compromete a oferecer uma experiência de usabilidade diferenciada, a *Weka 3*, centra seus esforços em oferecer ampla gama de funcionalidades que, em sua maioria, utilizam algoritmos de aprendizado de máquina. A interface dessa ferramenta é inferior à de sua concorrente, contudo é limpa e objetiva, com intuito de atender as necessidades do usuário.

A figura 17 exibe a tela inicial da *Weka 3*. Na parte superior é possível encontrar um menu que abriga opções de configuração e visualização. As funcionalidades principais, rotuladas de aplicações, estão localizadas no lado direito, e serão detalhadas adiante.

Figura 17 – Tela Inicial *Weka 3*

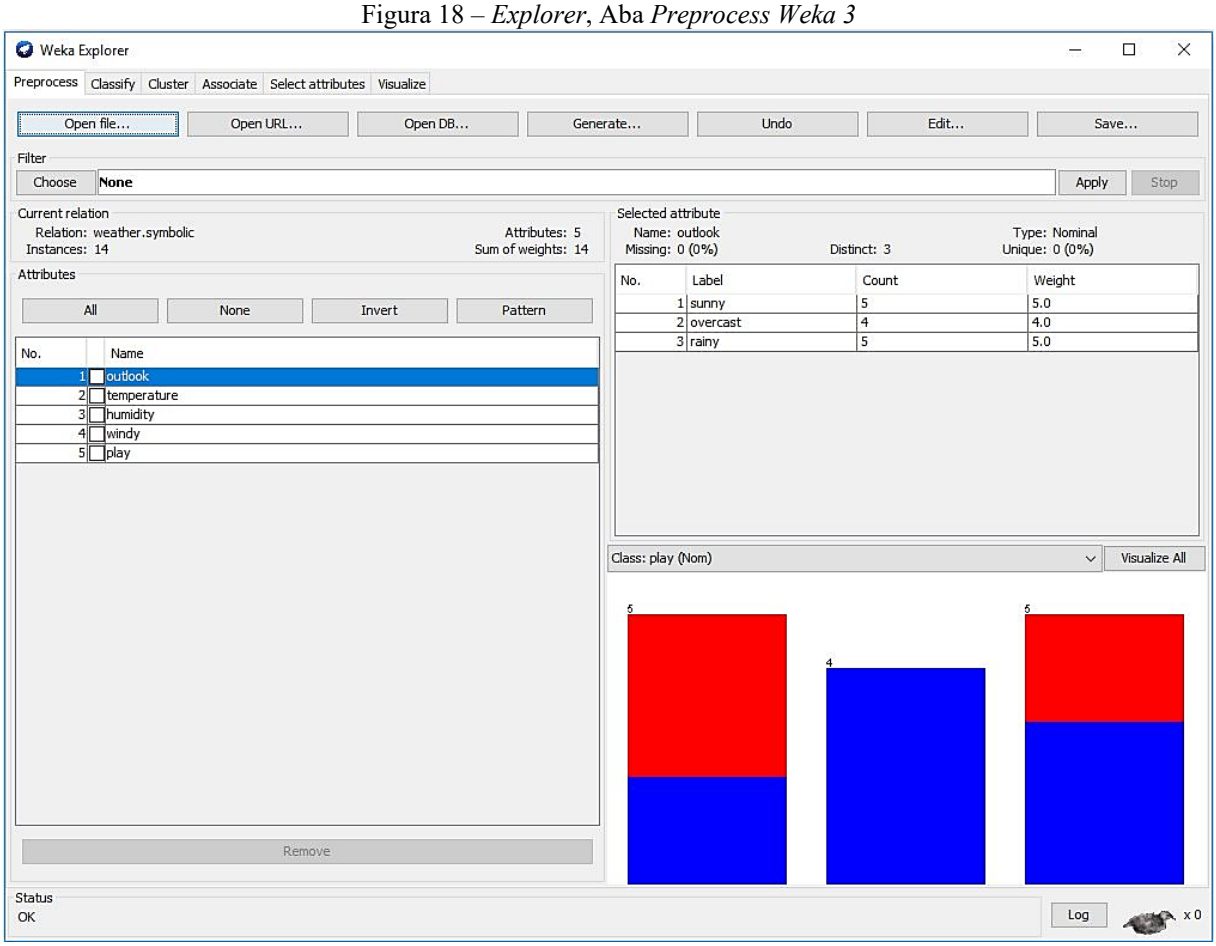
Fonte: Vieira (2018)

O item *Explorer* agrupa as funcionalidades necessárias para o pré-processamento, mineração e aspectos relacionados a visualização de dados (figura 18).

Essa seção está dividida em sete regiões. Na parte superior é possível visualizar as abas referentes ao pré-processamento, tarefas de MD e visualização de dados. A seguir existem botões utilizados para carregamento, edição e salvamento de um conjunto de dados, bem como para desfazer ações realizadas nos dados. Ainda na parte horizontal superior, há uma seção de filtros (*Filter*), com diversas escolhas para o tratamento dos dados.

Logo abaixo, no canto esquerdo, visualiza-se as informações do conjunto de dados. Deve-se esclarecer que a ferramenta lê dados no formato *.csv*, mas possui seu próprio tipo de arquivo, o .ARFF. Sob essa região estão listados todos os atributos da base de dados (denominada *relation* – relação). Ao clicar em um atributo qualquer, seus valores são exibidos ao lado, na seção *Selected atribute* (atributo selecionado).

Figura 18 – *Explorer*, Aba *Preprocess Weka 3*

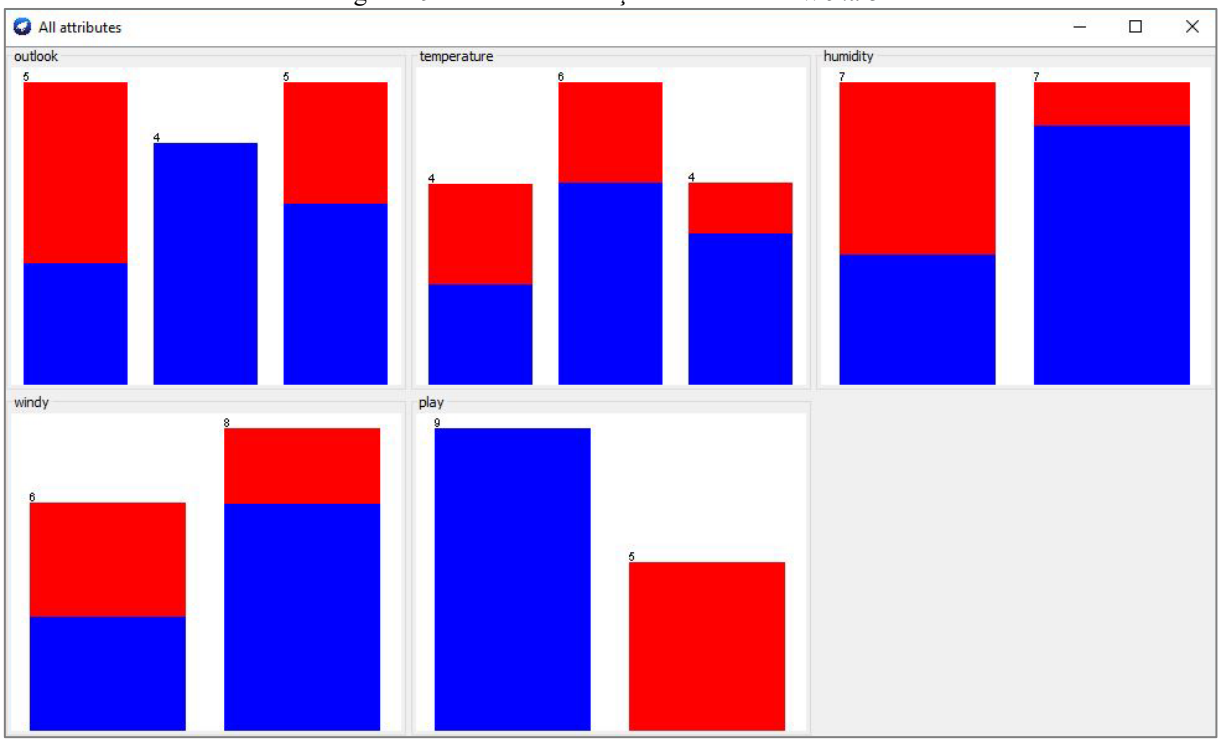


Fonte: Vieira (2018)

Por fim, ainda é possível observar um gráfico que fornece uma visualização dos dados, considerando o atributo classe do conjunto. Ao clicar em *Vizualize all*, visualiza-se a

representatividade da classe em relação aos outros atributos da relação (figura 19). Na figura em questão a classe é o último quadro, intitulado por *play*

Figura 19 – Classe em relação aos atributos *Weka 3*



Fonte: Vieira (2018)

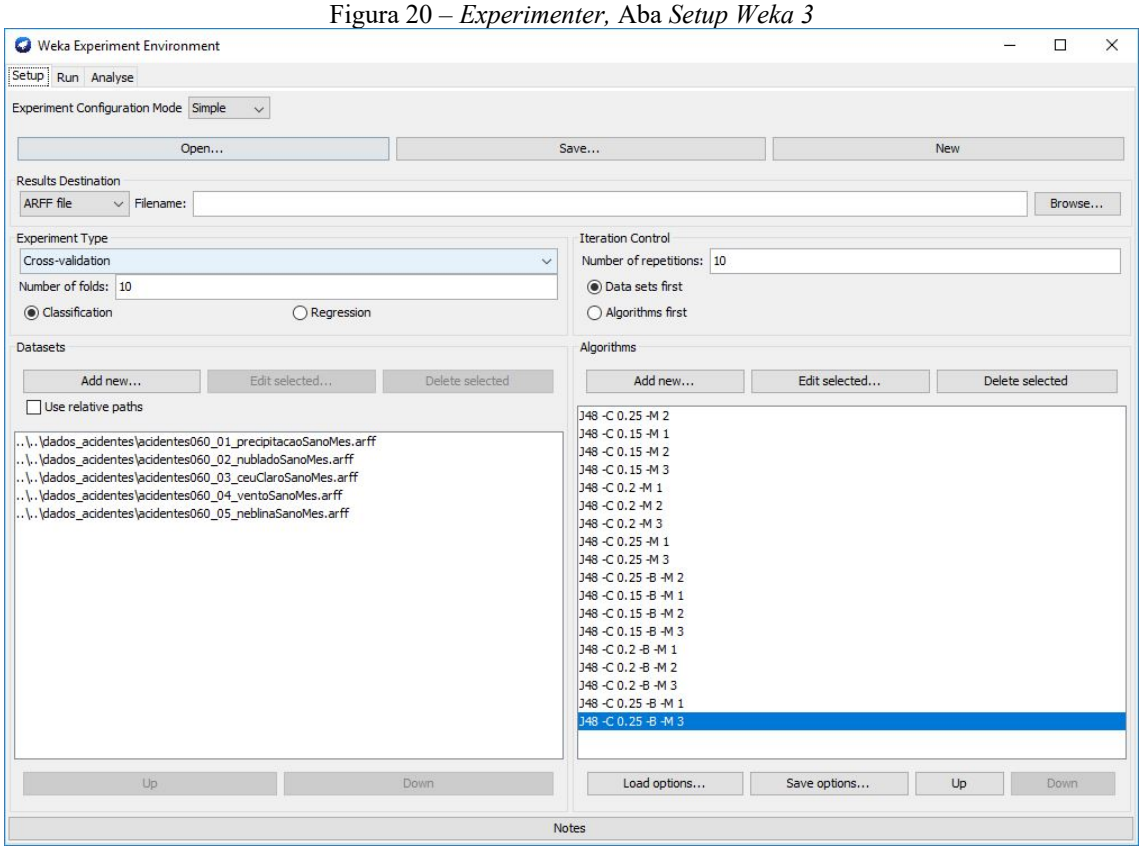
De retorno às aplicações exibidas na figura 17, tem-se o item *Experimenter*, voltado para o teste de desempenho de algoritmos de classificação, a fim de eleger a melhor solução para determinado problema ou testar parametrizações de um algoritmo específico.

Na tela inicial (figura 20) do *Experimenter* (aba *Setup*), os conjuntos de dados a serem avaliados são definidos na região *Datasets*. O tipo de divisão da base, em treinamento e teste, é feito em *Experiment Type*, onde também pode-se definir o número de dobras (*folds*), caso a validação cruzada esteja em uso.

Na parte direita da interface existem outras duas seções, uma define quantas vezes cada parametrização será testada (*Iteration Control*) e a outra permite selecionar os algoritmos com suas respectivas parametrizações. Essa área de trabalho pode ser salva e carregada posteriormente.

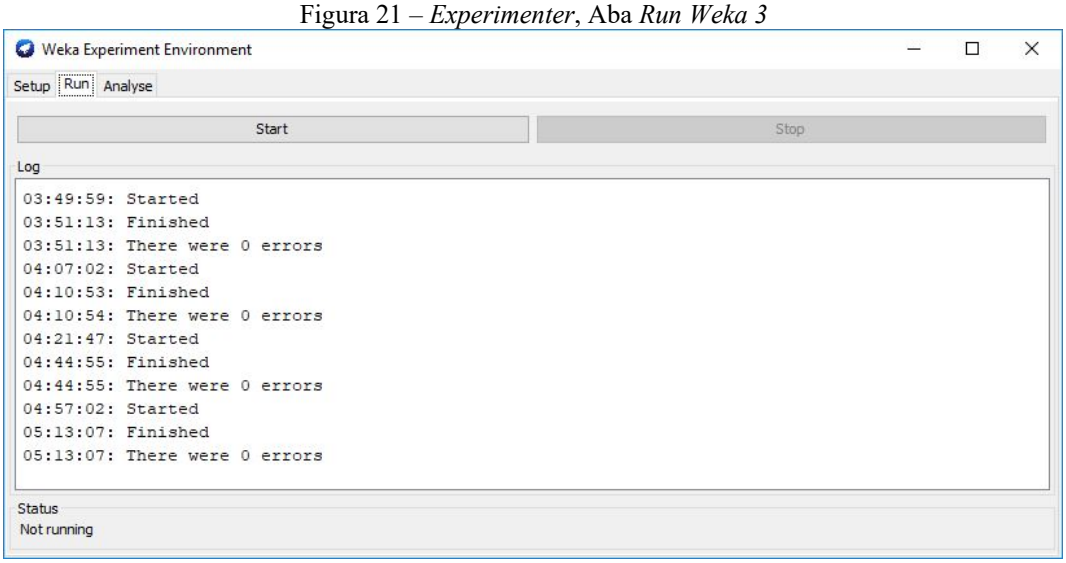
A aba *Run* (figura 21) registra informações sobre o início e término da avaliação, além de indicar se houve algum erro durante. Os erros decorrem de parametrizações inadequadas dos algoritmos.

Figura 20 – *Experimenter,* Aba *Setup Weka 3*



Fonte: Vieira (2018)

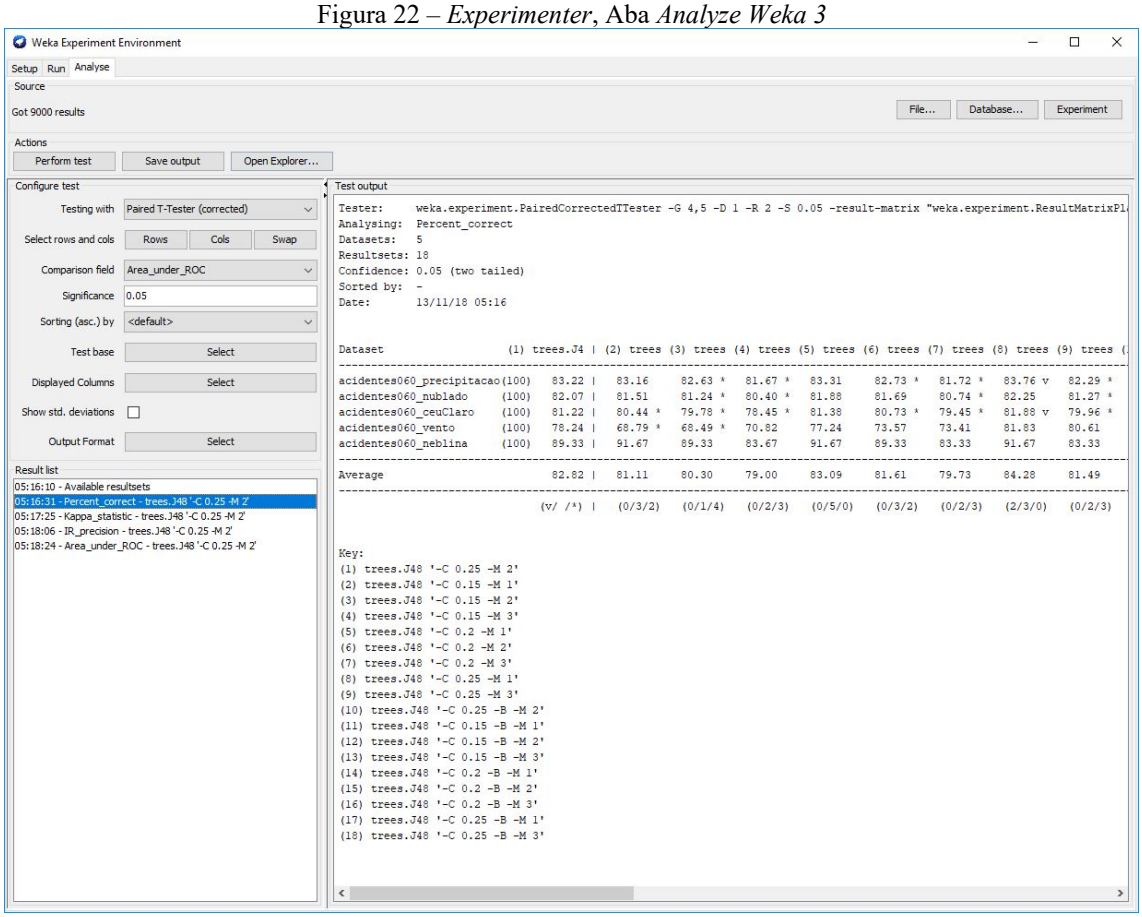
Figura 21 – *Experimenter*, Aba *Run Weka 3*



Fonte: Vieira (2018)

A figura 22 permite visualizar a área de avaliação. Após seu término do procedimento, clica-se em *Experiment* para que os dados da avaliação sejam carregados. Em *Configure Test* é possível ajustar como os resultados serão exibidos na seção *Test Output*. Feitos os ajustes de visualização, clica-se em *Perform test*, para que a avaliação seja impressa na tela e o especialista de DCBD possa comparar cada modelo, criado por um conjunto de algoritmos ou pelas parametrizações de um único.

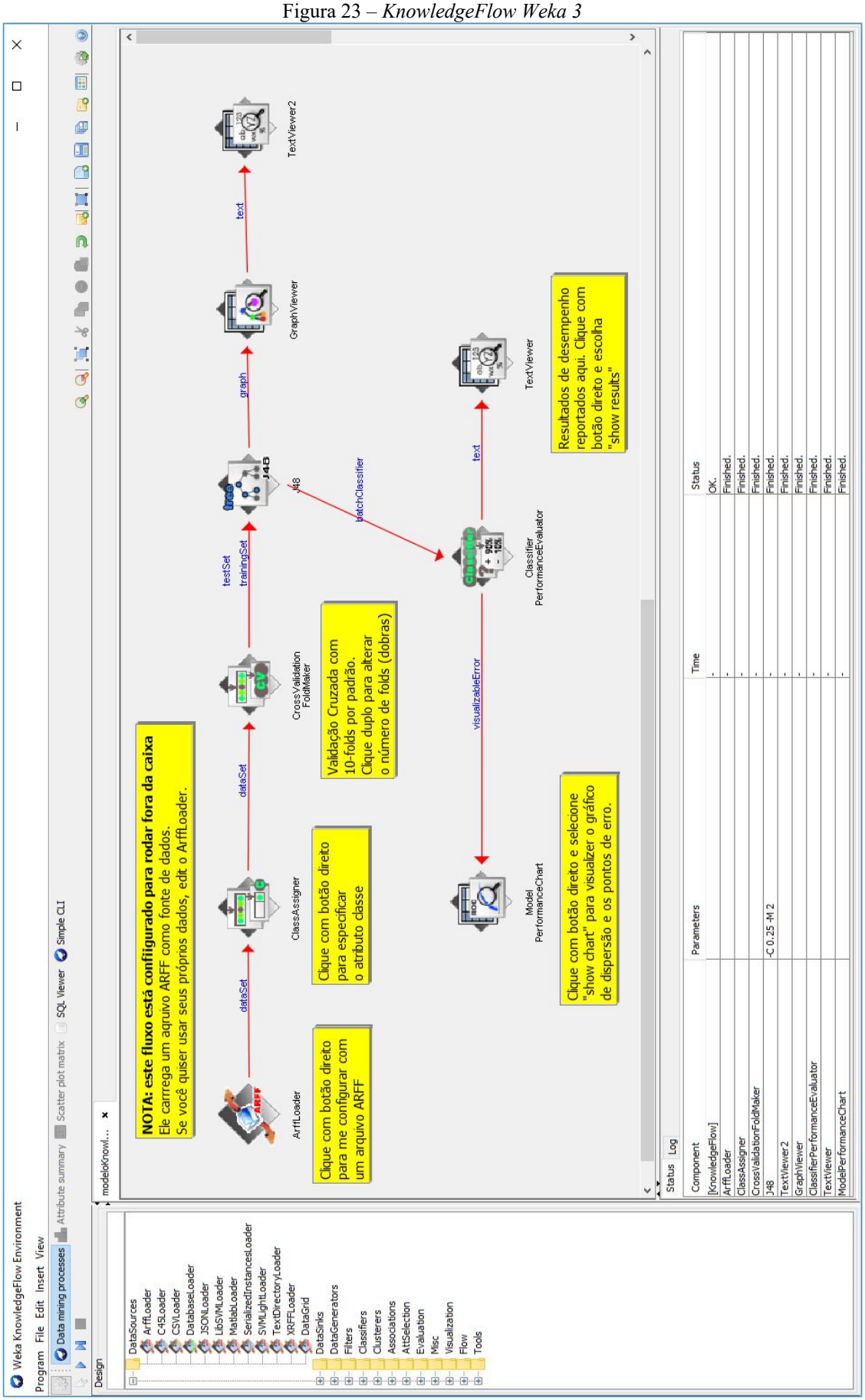
Figura 22 – *Experimenter*, Aba *Analyze Weka 3*



Fonte: Vieira (2018)

O item *KnowledgeFlow* possui funcionamento similar à ferramenta *Orange Data Mining*. Nele o usuário executa os mesmos passos que executaria na aba *Preprocess* (figura 18), porém de modo visual. Como dito, a *Weka 3*, perde em usabilidade para a *Orange*; isso fica claro, nesta seção (figura 23).

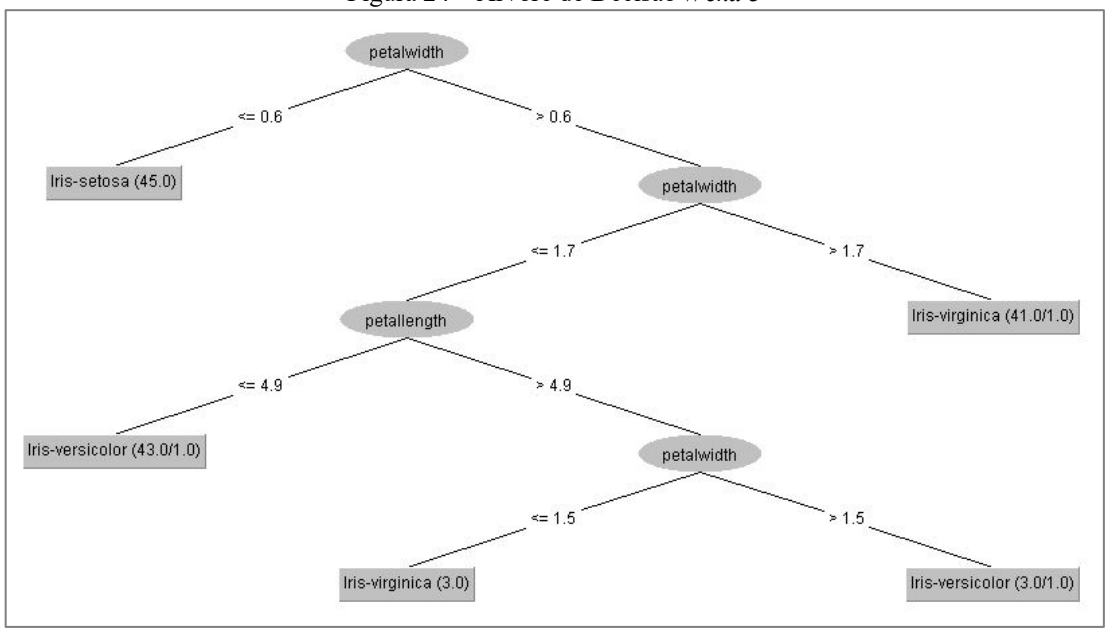
Figura 23 – *KnowledgeFlow Weka 3*



Fonte: Vieira (2018)

Assim como o *workflow* da figura 15 tratava um problema de classificação, a figura 23 exibe um problema de mesmo tipo. Como naquele exemplo, houve aqui a geração de uma árvore de decisão (figura 24), para comparar a exibição de resultados entre as ferramentas.

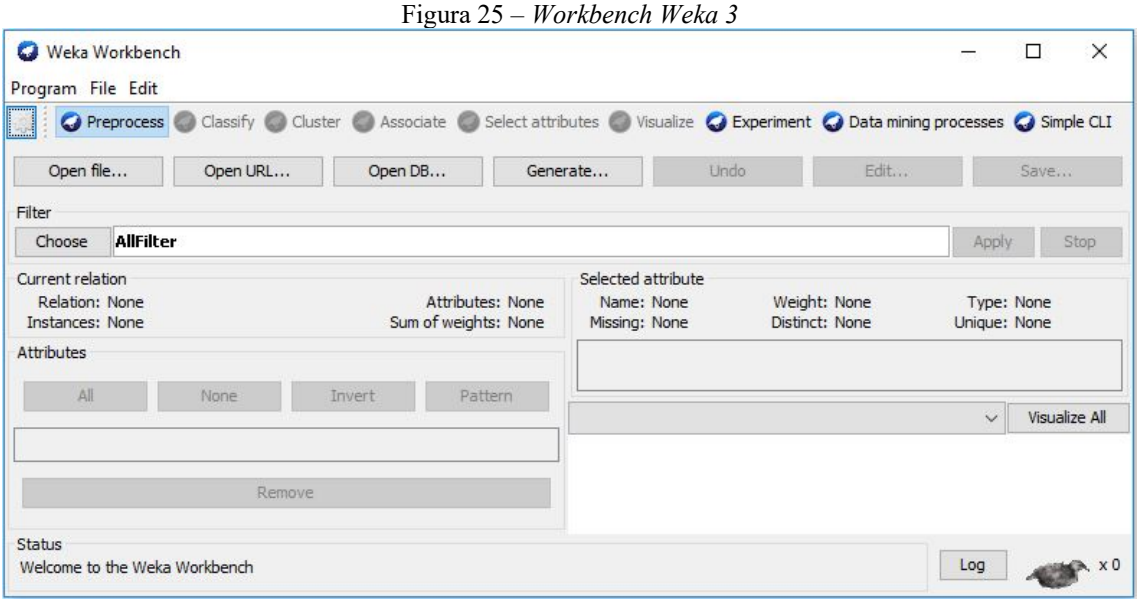
Figura 24 – Árvore de Decisão *Weka 3*



Fonte: Vieira (2018)

A aplicação *Workbench* possui visualização quase idêntica à *Explorer*, a diferença consiste em disponibilizar as aplicações da figura 17, na parte superior da tela (figura 25).

Figura 25 – *Workbench Weka 3*



Fonte: Vieira (2018)

Por fim, o item *SimpleCLI* constitui uma terceira alternativa para se trabalhar com a *Weka 3*. Nele os procedimentos são realizados através de linha de comando.

As métricas decisivas na avaliação das ferramentas foram o arcabouço de funcionalidades, a liberdade de parametrização e o desempenho geral. Assim sendo, constata-

se que a ferramenta *Orange Data Mining* consegue esconder do usuário, detalhes maiores de parametrizações, facilitando o seu uso. O que a torna ideal para aplicações em problemas comerciais, pois não demandam tanta parametrização para obtenção de resultados. Contudo, acompanhando essa facilidade surge um problema de limitação das parametrizações possíveis de suas funcionalidades, o que não acontece na *Weka 3*; lá os parâmetros são modificáveis segundo a intenção do usuário.

Feita a comparação e definidos quais os aspectos mais relevantes na avaliação, chegou-se à conclusão que a Weka 3, seria a melhor opção de ferramenta em relação a *Orange Data Mining*, para a execução do processo de DCBD nesse estudo.

3.4 Avaliação dos Processos de DCBD

Dos processos de DCBD estudados, verificou-se que ambos eram factíveis de aplicação a saber: o KDD e o CRISP-DM (seção 2.4.1).

Examinando as características, vantagens e índices de aplicabilidade, optou-se pelo CRISP-DM, pois o produto gerado ao final do processo, segue os moldes da tomada de decisão comumente verificada em empresas. Portanto, o conhecimento gerado será organizado para auxiliar as entidades gestoras de trânsito e transporte brasileiras, no que diz respeito a identificação de padrões que relacionem ATs ao clima.

Embora equivalente ao KDD, o CRISP-DM consegue condensar as nove etapas do primeiro em apenas seis, todavia as divide em itens menores (as tarefas).

3.5 Avaliação da Tarefa e Técnica e Algoritmo

As escolhas desta etapa possuíram importante relevância sobre os resultados obtidos nesta pesquisa. Embora o processo de DCBD escolhido, defina que somente na sua quinta etapa (Modelagem) devam ser escolhidos a tarefa, técnica e algoritmo para o procedimento de MD, optou-se por tomar essa decisão antes mesmo do início da avaliação dos dados. Isso tornou o procedimento mais objetivo, reduzindo possíveis iterações do processo. Desse modo, a base de dados foi pré-processada para se adequar à tarefa, técnica e algoritmo aqui escolhidos.

Outro fator que justifica a realização dessa avaliação antecipadamente e em uma seção específica, deve-se ao motivo de que a avaliação da tarefa, técnica e algoritmo integra os objetivos do trabalho.

Como mostrado no item 3.7.2, os conjuntos de dados adquiridos são constituídos, em sua maioria, por dados categóricos. A identificação de padrões preditivos a qual essa pesquisa se propõe realizar, necessita do emprego de uma tarefa preditiva que se adeque aos dados disponíveis, dado que o retorno da criação dos padrões configura-se num atributo categórico.

Das tarefas abordadas no item 2.4.2.1, aquela que consegue criar modelos preditivos com atributos categóricos e retornar uma classe de mesmo tipo é a classificação. Portanto, ela foi eleita como a tarefa de MD aplicada na execução do processo de DCBD, determinado no item anterior.

As próximas escolhas foram dependentes da tarefa de MD. Como técnica, procurou-se, dentre as apresentadas no item 2.4.2.2, aquela que além de gerar modelos de classificação, fosse capaz, também de gerar padrões que pudessem ser avaliados e interpretados em um processo de tomada de decisão. As outras técnicas preditivas de MD descritas, não geram padrões passíveis de análise, apenas um modelo de classificação.

Com base nesse pressuposto, a técnica escolhida, em concordância com a tarefa de classificação, foi Árvores de Decisão. A escolha dessa técnica também foi motivada pela taxa de presença na literatura, averiguada durante a construção do referencial teórico da pesquisa.

Por último, foi escolhido o algoritmo de aprendizado de máquina que implementasse a técnica selecionada. Também com base na literatura consultada e características do algoritmo, foi decidido que seria usado o *C4.5*. Na ferramenta *Weka 3*, ele está implementado sob o nome de *J48*. Lá existem algumas variações, porém são implementações modificadas do *C4.5*.

3.6 Ambiente de Execução do Processo

Durante esse trabalho, algumas ferramentas adicionais foram usadas, para tratamento e organização dos dados analisados e geração de resultados visuais, ao final. Este item descreve o ambiente computacional utilizado para a execução do processo de DCBD.

a) Configuração de *Hardware*:

Notebook Dell Inspiron 14-3442, com processador *Intel Core* i3-4005U de 1,70GHz. Memória *RAM* 4,00 GB com frequência de 1.666 MHz.

b) Configuração de *Software*

O quadro 6, agrupa e descreve as outras ferramentas utilizadas e sua respectiva serventia.

Quadro 6 – Configuração de *software*

Ferramenta	Objetivo
Windows 10 Pro, compilação 17134.345	• Plataforma de sistema operacional onde as demais ferramentas foram instaladas.
Microsoft Office Excel 2016	• Filtragem inicial dos dados; • Derivação de atributos; • Conversão de atributos; • Geração de comandos para inserção em banco de dados através de linguagem de consulta;

	• Auxilio na catalogação de atributos.
MySQL Server 5.7.23	• Armazenamento e agrupamento dos dados; • Pré-processamento dos dados; • Derivação, transformação e seleção de atributos; • Limpeza de dados; • Avaliação dos dados.
Graphviz 2.38.0	• Organização visual das árvores de decisão geradas.

Fonte: Vieira (2018)

3.7 Aplicação do Processo CRISP-DM

Como discorrido no embasamento teórico, o processo CRISP-DM é caracterizado por ser iterativo e interativo, com algumas etapas podendo avançar e retroceder, ou mesmo se comunicar com outras que não lhes são adjacentes (figura 6).

Dada essa premissa, convém observar que sua execução nesse estudo fez jus à sua natureza, isto é, não obedeceu a um ritmo linear. Todavia, para fins de objetividade e melhor compreensão do texto, tentou-se obedecer a uma certa linearidade na sua apresentação.

Os passos listados a seguir estão de acordo com a figura 7. Pode-se constatar que foram omitidas algumas tarefas das fases do processo, visto que o CRISP-DM é customizável e pode ser adaptado à realidade a qual é aplicado.

3.7.1 Entendimento do Negócio

O início do processo é marcado pela definição dos objetivos do negócio. Em concordância com os objetivos da pesquisa, o objetivo a nível de negócio foi decidido como: Contribuir com a diminuição de acidentes de trânsito e sua gravidade, no eixo Goiânia-Distrito Federal.

Em seguida procurou-se avaliar a situação vigente do negócio (cenário dos ATs), evidenciando possíveis pontos frágeis. Foram pontuados:

- Falta de estudos específicos sobre ATs rodoviários, no trecho escolhido, em função do clima, ou analisando condições climáticas adversas;
- Relatórios estatísticos com avaliações limitadas; e
- Baixo enfoque nos fatores que incidem em ATs.

Em seguida definiu-se os objetivos para a Mineração de Dados. Estes buscaram: (1) identificar padrões ocultos nos dados por meio de árvores de decisão; (2) utilizar uma base de dados históricos (de 2012 a 2017), para criação de um modelo de classificação que conseguisse classificar novas ocorrências de ATs, segundo o resultado do acidente (sem vítimas, com vítimas feridas ou fatais), em função do clima; (3) obter acurácia mínima de 75% de acertos,

nos modelos criados; e (4) avaliar a aplicação do modelo em registros de novas ocorrências de ATs, também em função do clima, na região em estudo.

3.7.2 Compreensão dos Dados

Nesta segunda fase do processo, buscou-se descrever o processo de obtenção e exploração dos dados, a fim de conhecer os possíveis problemas que deveriam ser tratados na fase seguinte.

a) Coleta de Dados Iniciais

Primeiramente, foi necessário adquirir os dados. Essa atividade de coleta foi realizada mais de uma vez e em locais diferentes. Inicialmente a pesquisa visava a análise de dados oriundos de duas fontes. A primeira era o Portal de Dados Abertos do Departamento de Polícia Rodoviária Federal (DPRF) e a segunda, a concessionária de rodovias Triunfo CONCEBRA.

Assim, foi realizada uma solicitação formal (Anexo D) à empresa Triunfo CONCEBRA solicitando os registros de ocorrências de acidentes e fluxo de veículos nos pedágios, entre Goiânia e o Distrito-Federal, no período de 2012 a 2017. Tais informações não foram disponibilizadas. Contudo, isso não configurou um obstáculo para o desenvolvimento da pesquisa, apenas uma redução na riqueza das informações disponíveis.

Quanto aos dados provenientes do Portal de Dados Abertos do DPRF, o trabalho de Reis (2014) revelou que nem todas as informações sobre ocorrências de ATs são disponibilizadas nesse meio, para a população. Então, foi realizada uma solicitação ao Ministério da Justiça (Anexo A), por meio do Sistema Eletrônico do Serviço de Informação ao Cidadão (e-SIC), das outras variáveis que não estavam disponíveis para consulta *online* e que foram consideradas relevantes (Anexo B).

Os dados requisitados foram fornecidos, entretanto quando sua análise foi iniciada constatou-se que as informações estavam incompletas, o que culminou em nova solicitação (Anexo F). Após esta última, a análise foi possível.

Restringida de qual fonte as informações foram fornecidas, algumas questões precisam ser esclarecidas. No Portal de Dados Abertos (PDA), os registros de acidentes são organizados por ano de ocorrência e divididos em duas categorias: dados de ocorrências e dados de pessoas. Foi feito o *download* dos arquivos referentes ao período de 2012 a 2017, das duas categorias. Por sua vez, a solicitação realizada no sistema e-SIC, forneceu os dados filtrados para a BR-060, do quilômetro 0 até o 140, para o período supramencionado.

O requerimento dos registros entre os quilômetrso 0 e 140, se justifica pelo fato que a segunda solicitação foi um pedido de correção frente às informações fornecidas anteriormente

(Anexos A).

b) Descrição dos Dados

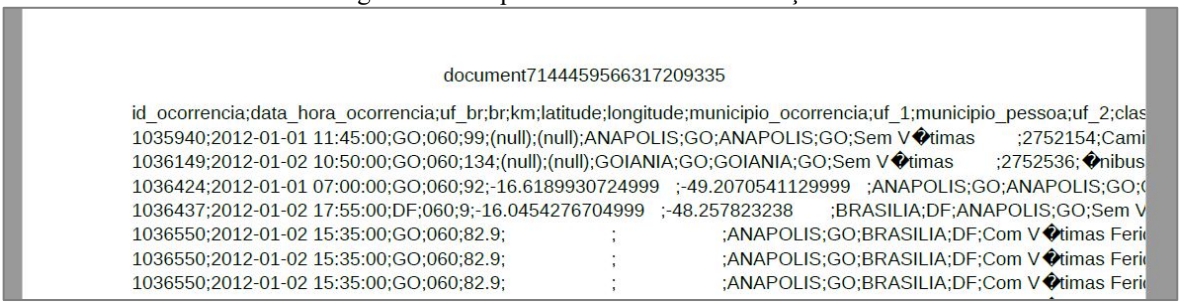
Continuando à aplicação do processo, os dados precisaram ser descritos. Buscou-se mostrar: como eles estavam organizados ao serem coletados, o formato sob os quais foram fornecidos, quais sistemas os geraram, assim como a quantidade de registros e atributos.

Os dados obtidos do PDA estavam armazenados em arquivos do tipo *.csv*. Aqueles da primeira solicitação, no sistema e-SIC (Anexo A), foram fornecidos em *.pdf*. Isso causou problemas de perda de registros e mistura de atributos, pois o arquivo cortou as informações, inviabilizando a extração das informações (figuras 26 e 27). Por esse motivo, a segunda solicitação foi realizada, requerendo os dados no formato de arquivo adequado (Anexo F).

Figura 26 – Arquivo com corte de informação 1

Fonte: Vieira (2018)

Figura 27 – Arquivo com corte de informação 2



Fonte: Vieira (2018)

O segundo pedido resultou no fornecimento das informações armazenadas em arquivos do tipo *.xlsx* e *.csv*. Verificou-se também que os dados, tanto do PDA quanto do sistema e-SIC foram gerados por sistemas diferentes: O BRBrasil, utilizado pelo DPRF até 2016, e o NovoBat, implantado a partir de 2017.

Os conjuntos adquiridos estavam organizados de acordo com a estrutura do PDA (por ocorrência e por pessoa). Nos arquivos “Por ocorrência” estavam registrados aspectos gerais das ocorrências de AT e nos “Por Pessoa”, aspectos gerais das ocorrências e específicos dos indivíduos e veículos, envolvidos em ATs rodoviários.

O quadro 7 especifica todos os arquivos obtidos para esse estudo:

Quadro 7 – Arquivos obtidos

Arquivo	Descrição	Origem
datatatran2012.csv	Dados gerais de acidentes, separados por ano e agrupados por ocorrência.	Portal de Dados Abertos
datatatran2013.csv		
datatatran2014.csv		
datatatran2015.csv		
datatatran2016.csv		
datatatran2017.csv		
acidentes2012.csv	Dados gerais de acidentes e específicos de pessoas e veículos, separados por ano e agrupados por pessoa	
acidentes2013.csv		
acidentes2014.csv		
acidentes2015.csv		
acidentes2016.csv		
acidentes2017.csv		
BAT___Por_Ocorrencia.csv	Dados gerais de acidentes, agrupados por ano e ocorrência	Sistema e-SIC
BRBRASIL___Por_Ocorrencia.csv		
BAT___Por_Pessoa.csv	Dados específicos de pessoas e veículos agrupados por ano e ocorrência	
BRBRASIL___Por_Pessoa.csv		

Fonte: Vieira (2018)

Na descrição dos dados ainda é preciso conhecer a dimensão da base. Foram somados todos os registros contidos nos arquivos adquiridos. A adição foi realizada tal como as informações foram recebidos (tabela 1), devido ao fato de que nesta parte do processo nenhuma modificação é realizada sobre os dados.

Tabela 1 – Quantidade de registros da base de dados

	PDA		Sistema e-SIC	
Ano	**Por Ocorrência**	**Por Pessoa**	**Por Ocorrência**	**Por Pessoa**
2012	184.568	396.916	1.278	2.647
2013	186.748	405.820	1.450	2.974
2014	169.201	368.506	1.374	2.851
2015	122.161	269.052	1.184	2.450
2016	96.363	216.261	1.134	2.411
2017	89.518	204.289	1.054	3.669
TOTAIS	**848.559**	**1.860.844**	**7.474**	**17.002**
	2.709.403		**24.476**	
	2.733.879			

Fonte: Vieira (2018)

Como os dados foram originados de dois sistemas de coleta foi preciso identificar um meio para relacioná-los e montar uma base única. Concluiu-se que o melhor meio de realizar a união era através dos identificadores de ocorrências e pessoas.

Os arquivos estavam organizados em estruturas tabulares, onde as linhas identificavam os registros e as colunas, os atributos. Calculado todos os atributos, obteve-se um total de 202. É importante destacar que as contagens dos registros e atributos incluíram a sua totalidade, itens duplicados foram adicionados à soma.

Na fase Modelagem, algoritmos são empregados para criação de um modelo de classificação. Todavia, eles podem exigir tipos de dados específicos, identificados como atributos. Quatro tipos possíveis podem ser definidos, segundo o processo aplicado: Atributos Categóricos, Atributos Numéricos, Atributos Booleanos e Atributos do tipo Data. Verificou-se, na base em análise, que aproximadamente 53% (107) dos atributos eram categóricos, 32,7% (66) eram numéricos, 7,4% (15) eram booleanos e 6,9% (14) eram do tipo data.

c) Exploração dos Dados

Neste passo, foi realizada outra avaliação, dessa vez a fim de identificar quais atributos seriam mais promissores no processo de descoberta de conhecimento. Foi detectado que os atributos que tratavam da condição meteorológica, data e hora do acidente, classificação, tipo e causa do acidente, detalhes do veículo, sexo e idade do indivíduo envolvido, e estado pós-acidente poderiam ser promissores.

d) Verificação da Qualidade dos Dados

Avaliar a qualidade dos dados constituiu uma tarefa importante, pois possibilitou identificar ruídos e inconsistências nos valores dos atributos. Além de ruídos, também foram localizados erros de medição. É o caso de registros de acidentes ocorridos na cidade de Santo Antônio do Descoberto, registrados como ocorridos em Santo Antônio de Goiás. O quadro 8 apresenta as inconsistências encontradas nos dados:

Quadro 8 – Atributos inconsistentes ou com ruídos

Atributo	Tipo de Ruído / Problema
ano_fabricacao_veiculo	Ano de fabricação sem valor ou com valores “NA”, “(null)” e anos inválidos.
br	BR com valores “NA”.
data_nascimento	Data de nascimento com valores “(null)”.
dia_semana	Dias da semana com valores completos e reduzidos. Ex.: Segunda e segunda-feira.
id_veiculo	Veículos com valor “(null)”.
idade	Idades com valores inválidos, como: valor negativo, valores acima de

Atributo	Tipo de Ruído / Problema
	100, 200 e 2000; ou valores “NA”.
km	Quilometragem zerada em trechos que a numeração deveria ser acima de 0 ou com valores “NA”. Acidentes ocorridos na BR-060, classificados em trechos inexistentes da BR-153.
latitude	Latitude com valor “(null)”.
lbrlatitude	Latitude com valor “(null)”.
lbrlongitude	Longitude com valor “(null)”.
longitude	Longitude com valor “(null)”.
marca	Marca de veículo sem valor, com valores: “I”, “********”, “(null)” ou armazenando dois valores (Marca e Modelo).
municipio	Ocorrências registradas em uma cidade, mas que aconteceu em outra.
município_pessoa	Município de origem da pessoa com valor “(null)”.
nacionalidade	Nacionalidade com valor “NA”.
naturalidade	Naturalidade com valor “NA”.
tipo_veiculo	Tipo de veículo sem valor ou com valores “(null)”.
uf_2	Estado de origem da pessoa com valor “(null)”
uso_solo	Variável com dois tipos de valor: Categórico ou Booleano

Fonte: Vieira (2018)

3.7.3 Preparação dos Dados

De acordo com a figura 4, a fase que consome a maior quantidade de tempo, esforços e recursos é a responsável por preparar os dados para serem minerados. Isso foi confirmado durante a seleção, limpeza e tratamento dos dados desta pesquisa.

a) Seleção dos Dados

Os dados foram selecionados para que atendessem aos objetivos definidos na primeira etapa do processo (Entendimento do Negócio). A seleção foi realizada em duas dimensões: por registros e por atributos, com o propósito de trabalhar apenas com informações relevantes para o problema.

Visto que as bases possuíam registros de ATs das várias rodovias federais brasileiras, elas foram filtradas para a região definida no item 3.2. Essa ação foi realizada em um processador de planilhas comum. A escolha do processador de planilhas frente a ferramenta de banco de dados, se justifica por questões de otimização de tempo de inserção e recuperação dos dados, dada a quantidade total de registros (tabela 1).

Os arquivos listados no quadro 7 foram filtrados segundo as regras definidas no quadro 9 e divididos em arquivos, separados por ano, já que aqueles provenientes do sistema e-SIC

estavam agrupados por essa variável.

Quadro 9 – Parâmetros de filtragem inicial dos dados

Parâmetro	Faixa de Valores
Unidade Federativa (UF)	DF e GO
Rodovia Federal (BR)	BR 060 e BR 153
Quilometragem	KM 0 ao 137

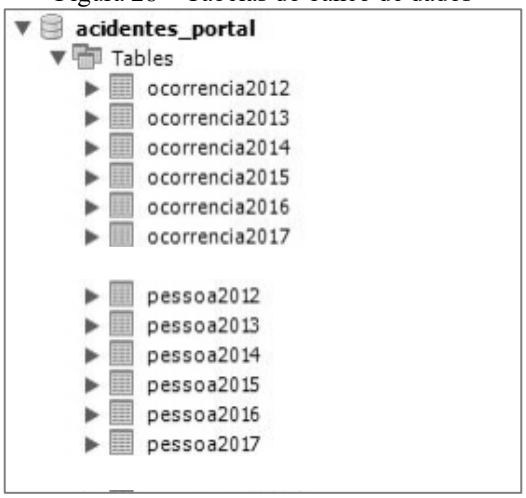
Fonte: Vieira (2018)

Embora o item 3.2 defina o quilômetro 8 como inicial, a fase Compreensão dos Dados revelou a necessidade de se considerar o ponto de partida a partir do quilômetro 0, devido aos problemas encontrados no atributo “km” (quadro 8).

A BR 153 se mescla com a 060 entre as cidades de Goiânia e Anápolis, foram buscados registros no intervalo dos quilômetros 444 a 490, da primeira rodovia, espaço que compreende a segunda. Foram encontradas apenas duas ocorrências (em 2017), registradas nos quilômetros 450,2 e 448,1.

O produto dessa filtragem, foi inserido em um banco de dados (figura 28), criado no SGBD MySQL (quadro 6).

Figura 28 – Tabelas do banco de dados



Fonte: Vieira (2018)

Entretanto, ainda houve a necessidade da remoção de registros onde o atributo “municipio” continha valores de cidades que, geograficamente, não estavam localizadas entre os quilômetros 0 e 137. Para isso, foi realizada uma consulta que recuperou todas as cidades nas quais ocorreram acidentes. Aquelas consideradas como invasoras foram removidas (figura 29). A figura 57 (Apêndice B) permite visualizar tanto as cidades invasoras quanto aquelas registradas corretamente; com atenção à cidade 08 (Santo Antônio de Goiás), cuja situação foi explicada anteriormente.

Figura 29 – Comando para remoção de cidades invasoras

```
DELETE FROM nome_tabela
WHERE municipio='ABADIA DE GOIAS' OR
      municipio='AGUA FRIA DE GOIAS' OR
      municipio='AGUAS LINDAS DE GOIAS' OR
      municipio='APARECIDA DE GOIANIA' OR
      municipio='GOIANESIA' OR
      municipio='GUAPO' OR
      municipio='HIDROLANDIA' OR
      municipio='INDIARA' OR
      municipio='LUZIANIA' OR
      municipio='PIRENOPOLIS' OR
      municipio='VARJAO';
```

Fonte: Vieira (2018)

A tabela 2 realiza uma comparação do antes e depois das duas seleções iniciais de registros. A tabela não apresenta quantidades referentes aos dados do sistema e-SIC, pois na fase anterior, análises e comparações demonstraram que as informações presentes neles, não teriam utilidade. Os motivos envolveram: redundância de registros, alta quantidade de valores de atributos incorretos, atributos presentes apenas em um ano ou quantidade de ruídos que inviabilizava um tratamento.

Tabela 2 – Quantidade de registros selecionados

Ano	Por Ocorrência	Por Pessoa
2012	1.189	2.425
2013	1.362	2.762
2014	1.282	2.631
2015	1.049	2.129
2016	976	1.988
2017	919	1.935
TOTAIS	**6.777**	**13.870**
	20.647	

Fonte: Vieira (2018)

Quanto a seleção dos atributos, foram avaliados os mais significativos para o problema. Dos 202 atributos disponíveis, 111 eram sintaticamente duplicados (tinham o mesmo nome e armazenavam a mesma informação), 19 eram semanticamente duplicados (possuíam nomes diferentes, mas armazenavam a mesma informação), 28 não estavam presentes em todos os anos (somente em 2017, devido ao uso do sistema NovoBat (item 3.7.2 b)). Restaram 44 atributos. Dentre esses, 16 foram escolhidos para serem utilizados na criação dos modelos de classificação. O quadro 10 apresenta a descrição dos selecionados.

Quadro 10 – Atributos selecionados

Atributo	Descrição
data_inversa	Data da ocorrência no formato dd/mm/aaaa.
dia_semana	Dia da semana da ocorrência. Ex.: Segunda, Terça, etc.
horario	Horário da ocorrência no formato hh:mm:ss.
municipio	Nome do município de ocorrência do acidente.
causa_acidente	Identificação da causa presumível do acidente. Ex.: Falta de atenção, Velocidade incompatível, etc.
tipo_acidente	Identificação do tipo de acidente. Ex.: Colisão frontal, Saída de pista, etc.
classificacao_acidente	Classificação quanto à gravidade do acidente: Sem Vítimas, Com Vítimas Feridas, Com Vítimas Fatais e Ignorado.
sentido_via	Sentido da via considerando o ponto de colisão: Crescente e decrescente.
condicao_meteorologica	Condição meteorológica no momento do acidente: Céu claro, chuva, vento, etc.
uso_solo	Descrição sobre as características do local do acidente: Urbano ou rural.
tipo_veiculo	Tipo do veículo conforme Art. 96 do Código de Trânsito Brasileiro. Ex.: Automóvel, Caminhão, Motocicleta, etc.
tipo_envolvido	Tipo de envolvido no acidente conforme sua participação no evento. Ex.: condutor, passageiro, pedestre, etc.
estado_fisico	Condição do envolvido conforme a gravidade das lesões. Ex.: morto, ferido leve, etc.
sexo	Sexo do envolvido. O valor "inválido" indica que não foi possível coletar tal informação.
veículos	Total de veículos envolvidos na ocorrência.
pessoas	Total de pessoas envolvidas na ocorrência.

Fonte: DPRF (2016)

Dos atributos selecionados, apenas "pessoas" e "veiculos" estavam exclusivamente nas tabelas de ocorrência (figura 28). Assim, esses atributos foram anexados aos demais registros existentes nas tabelas de pessoa. A figura 55 (Apêndice A), exibe o comando utilizado para seleção dos atributos com seus registros.

Os registros dos atributos selecionados, passaram por um processo de transformação, havendo redução da quantidade de valores por atributo (item 3.7.3 C). Após essa redução, os atributos foram novamente selecionados (figura 56 – Apêndice A). Dois novos atributos foram adicionados, "dia_mes" (item 3.1.3 C) e "id_pessoa".

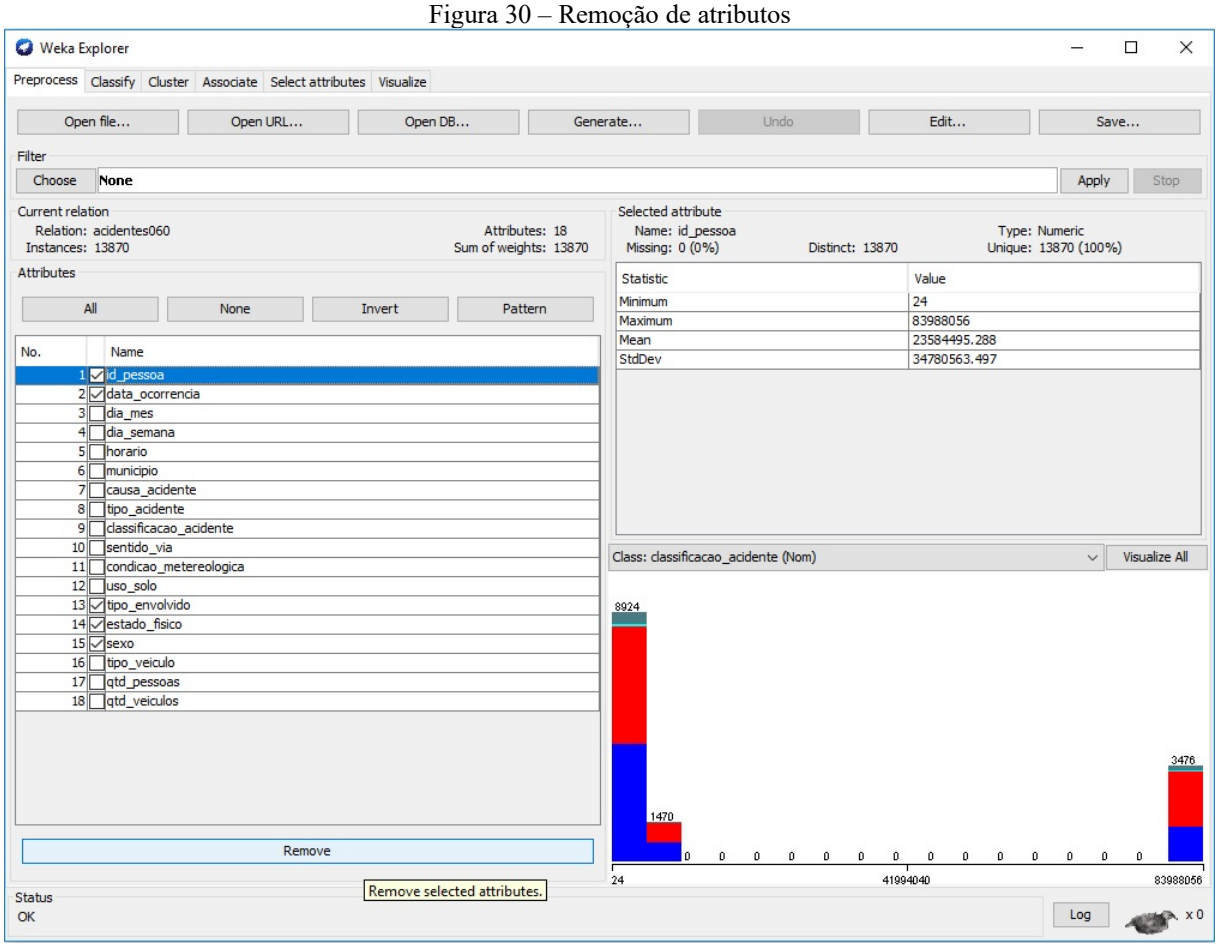
O resultado dessa seleção foi importado para a ferramenta *Weka 3*, por meio de um arquivo *.csv*, onde uma nova segmentação foi realizada considerando os objetivos da primeira fase do processo. Dessa vez foram desconsiderados os atributos: "id_pessoa", "data_inversa",

“tipo_envolvido”, “estado_fisico” e “sexo”. Os atributos foram removidos através do botão *Remove* (figura 30).

O refinamento dos dados contou com mais outra seleção (de registros) com o propósito de remover valores “Ignorados” dos atributos: “classificacao_acidente”, “condicao_meteorologica”, e “tipo_veiculo”. Essa ação removeu 503 registros, através da aplicação do filtro *RemoveWithValues*, especificando o índice do atributo (*attributeIndex*) e do valor a ser removido (*nominalIndices*) (figura 31).

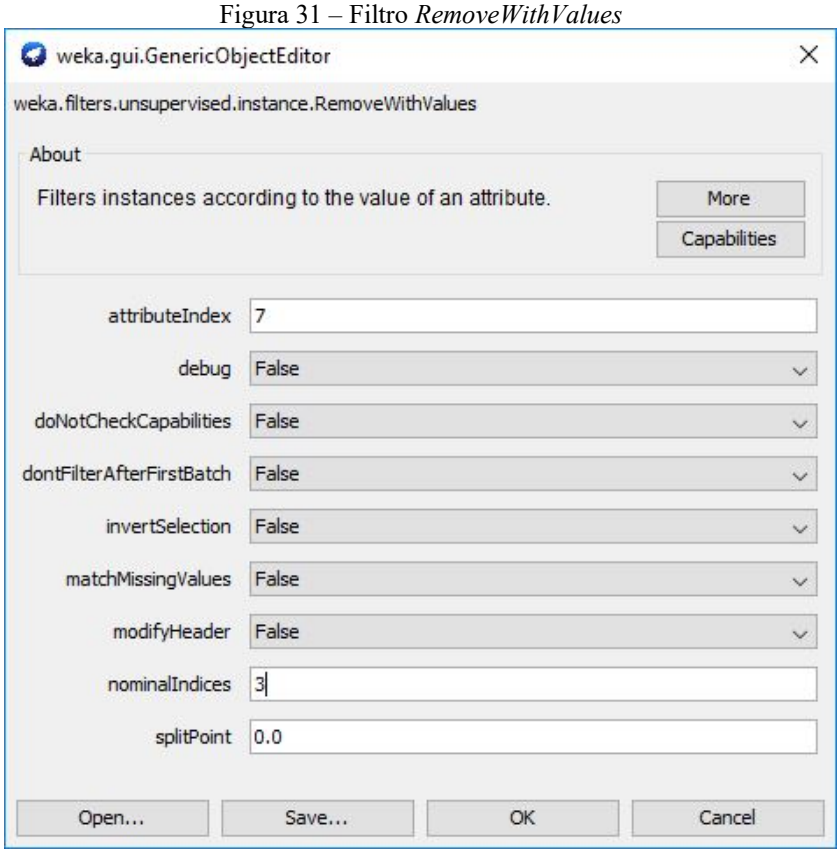
Ao fim dessas intervenções, um arquivo *.arff* foi gerado para dar continuidade às operações da atividade “Construção dos Dados”.

Figura 30 – Remoção de atributos



Fonte: Vieira (2018)

Figura 31 – Filtro *RemoveWithValues*



Fonte: Vieira (2018)

b) Limpeza dos Dados

Ainda que o passo anterior tenha demonstrado a criação do arquivo *.arff*, filtrado, a ação realizada nesta fase foi executada quando os dados ainda estavam no banco de dados. O atributo que passou por esse processo foi "tipo_veiculo", conforme quadro 11 e figura 32.

Neste passo, dos dados selecionados no passo anterior, aqueles com ruídos passaram por um processo de limpeza (quadro 8), realizada antes da inserção no banco de dados, pois devido a sujeiras presentes nos dados, houve incompatibilidades no momento da inserção.

Quadro 11 – Limpeza de ruídos

Atributo	**Ruído**	**Método**	**Resultado**
tipo_veiculo	Valor vazio "(null)".	1 - Preenchimento de valores vazios ou "(null)", com o valor NULL. 2 - Alteração do valor NULL para "Ignorado"	202 correções

Fonte: Vieira (2018)

Figura 32 – Comando para limpeza tipo_veiculo

```
UPDATE pessoa2012 SET tipo_veiculo = 'Ignorado'
WHERE tipo_veiculo is null OR tipo_veiculo = 'null';

UPDATE pessoa2013 SET tipo_veiculo = 'Ignorado'
WHERE tipo_veiculo is null OR tipo_veiculo = 'null';

UPDATE pessoa2014 SET tipo_veiculo = 'Ignorado'
WHERE tipo_veiculo is null OR tipo_veiculo = 'null';

UPDATE pessoa2015 SET tipo_veiculo = 'Ignorado'
WHERE tipo_veiculo is null OR tipo_veiculo = 'null';

UPDATE pessoa2016 SET tipo_veiculo = 'Ignorado'
WHERE tipo_veiculo is null OR tipo_veiculo = 'null';

UPDATE pessoa2017 SET tipo_veiculo = 'Ignorado'
WHERE tipo_veiculo is null OR tipo_veiculo = 'null';
```

Fonte: Vieira (2018)

c) Construção do Dados

Durante a preparação dos dados, a tarefa “Construção dos Dados” é considerada a mais demorada, pois exige análise, atenção e capacidade de resolver problemas.

Assim que os dados sofreram a primeira filtragem (quadro 9), foram criados arquivos divididos por ano e categoria (por ocorrência e por pessoa), no formato *.csv*. Após isso, todos os caracteres especiais foram retirados.

Em seguida realizou-se a derivação do atributo “data_ocorrencia” no atributo “dia_mes”. A derivação consistiu em transformar a data completa do modelo DD/MM/AAAA para DD-MMMM. Essa abordagem foi efetuada objetivando uma redução nominal, estruturada por quinzenas.

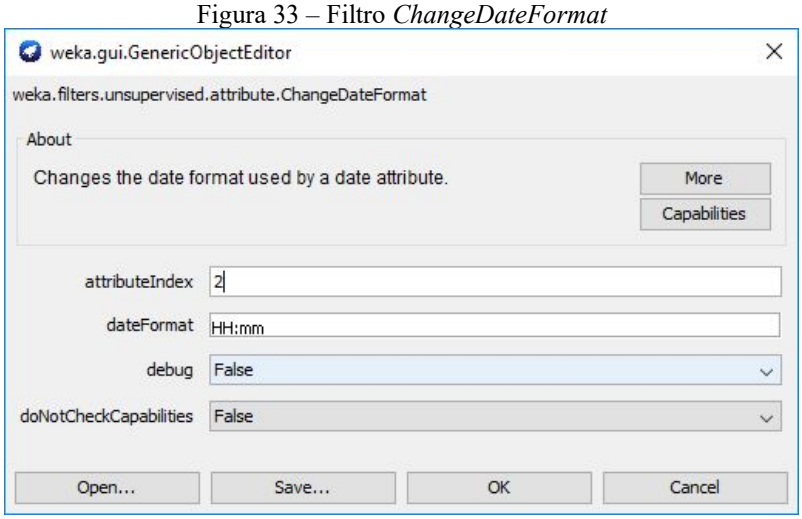
Todas as reduções nos atributos foram nominais, já que a base contava apenas com dois atributos numéricos. Elas foram efetivadas através da linguagem *SQL (Structured Query Language)* (Apêndice C) e da ferramenta *Weka 3*.

As transformações dos atributos com a *Weka 3*, foram efetuadas durante o procedimento onde foram selecionados os atributos e registros (figuras 30 e 31).

Assim foi realizada a redução do atributo “horario” e renomeação dos valores do atributo “classificacao_acidente”. Para o atributo “horario”, foi efetuada uma alteração manual no arquivo *.arff*, onde seu tipo foi alterado para *date*, com a máscara: “HH:mm:ss”.

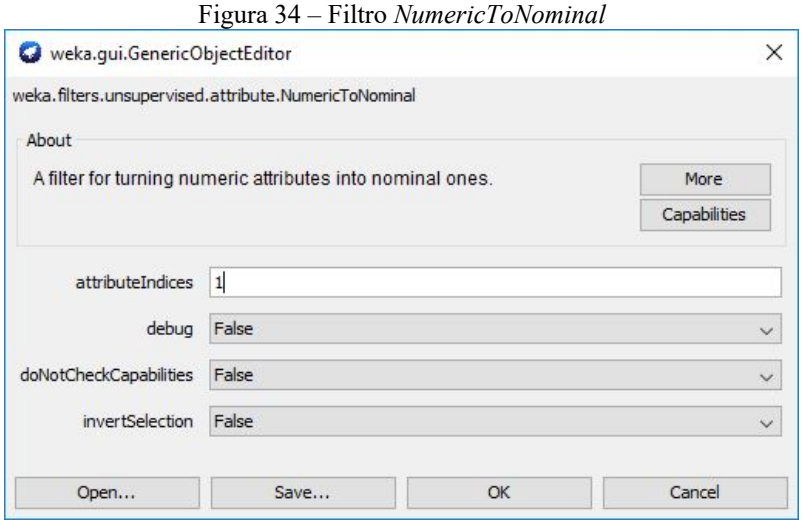
Em seguida o arquivo foi recarregado na *Weka 3* e utilizados, os filtros: *ChangeDateFormat*, para transformar a hora para “HH:mm” (figura 33); *NumericToNominal*, para converter o campo de data para nominal, de maneira ordenada (figura 34), e o filtro *MergeManyValues*, para agrupar os horários em quatro turnos (madrugada, manhã, tarde noite) (figura 35)

Figura 33 – Filtro *ChangeDateFormat*



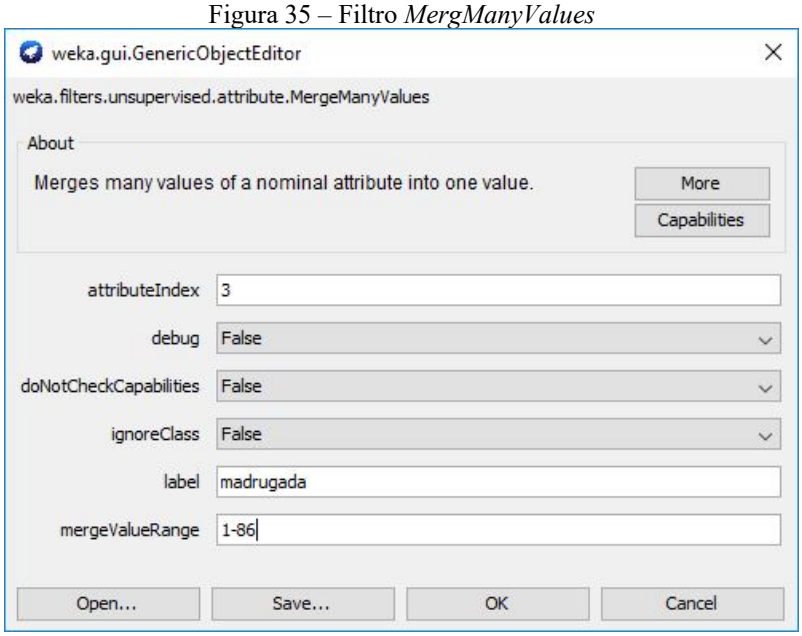
Fonte: Vieira (2018)

Figura 34 – Filtro *NumericToNominal*



Fonte: Vieira (2018)

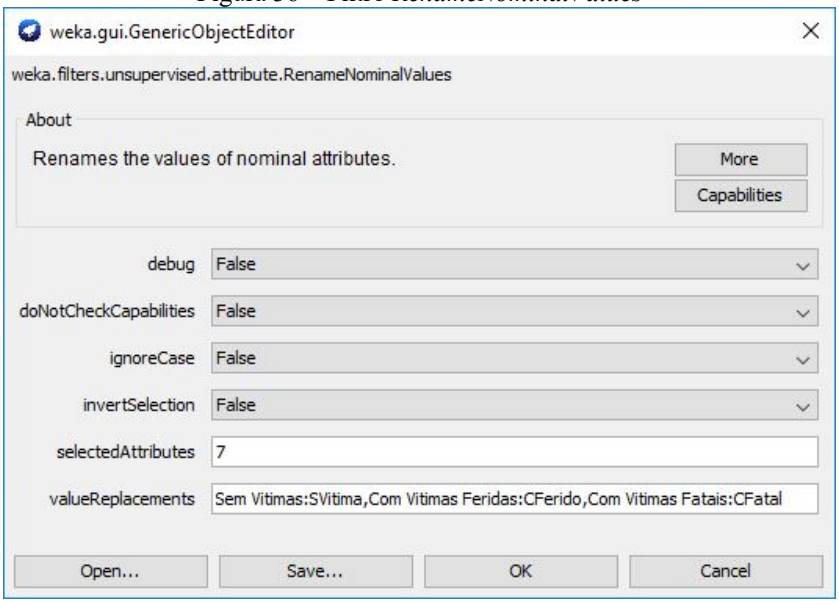
Figura 35 – Filtro *MergManyValues*



Fonte: Vieira (2018)

No atributo “classificao_acidente” foi empregado o filtro *RenameNominalValues* para alterar o nome dos valores dos atributos. “Sem Vítimas” passou a ser SVitima, “Com Vitimas Feridas” virou CFerido e “Com Vitimas Fatais”, foi transformado em CFatal (figura 36).

Figura 36 – Filtro *RenameNominalValues*



Fonte: Vieira (2018)

d) Integração dos Dados

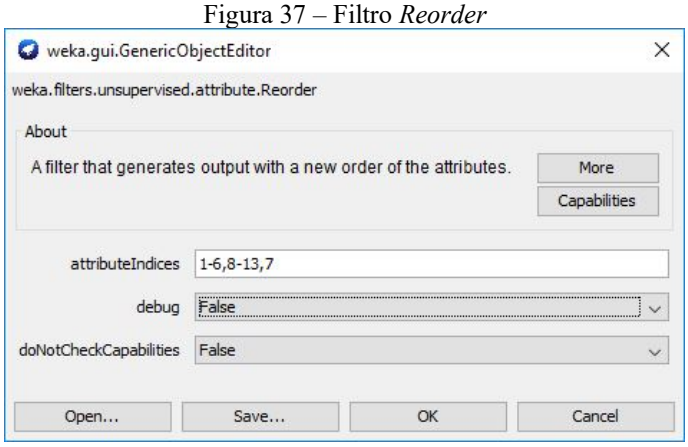
Para geração do conjunto de dados para análise, realizou-se a união das tabelas de ocorrência com as de pessoa. Essa união foi realizada no SGBD, em uma consulta com a condição *where* que igualou o identificador das ocorrências, nas tabelas de ocorrência (id), com o mesmo atributo nas tabelas de pessoa (id_ocorrencia).

Como a base também estava dividida por ano, utilizou-se o comando *SQL UNION ALL,* para unir as 6 consultas realizadas, uma para cada ano (figura 55 do Apêndice A).

e) Formatação dos Dados

A última atividade da fase de pré-processamento dos dados, consistiu em ordenar e dividir os dados. Primeiramente, houve uma ordenação nos campos do arquivo resultante dos passos B e C (item 3.7.3), onde a classe “classificacao_acidente” foi movida para última posição com a intenção de facilitar a execução do algoritmo de classificação na ferramenta *Weka 3*. Para esse procedimento, foi empregado o filtro *Reorder* (figura 37).

Figura 37 – Filtro *Reorder*



Fonte: Vieira (2018)

Os 13.367 registros finais, foram segmentados em arquivos individuais em função do clima. Essa redução da quantidade em relação a quantidade exibida na tabela 2, se deu pelas seleções realizadas na parte de seleção de dados. Foram gerados cinco arquivos *.arff*. A figura 38 ilustra a estrutura de um arquivo *.arff*.

Figura 38 – Arquivo .arff

```
@relation 'acidentes060'

@attribute ano {2012,2013,2014,2015,2016,2017}
@attribute mes {jan,fev,mar,abr,mai,jun,jul,ago,set,out,nov,dez}
@attribute dia_mes {janeiro1-15,janeiro16-31,fevereiro1-15,fevereiro16-29,marco1-15,marco16-31,
                    abril1-15,abril16-30,maio1-15,maio16-31,junho1-15,junho16-30,julho1-15,julho16-31,
                    agosto1-15,agosto16-31,setembro1-15,setembro16-30,outubro1-15,outubro16-31,novembro1-15,
                    novembro16-30,dezembro1-15,dezembro16-31}
@attribute dia_semana {Domingo,Segunda,Quarta,Sexta,Quinta,sabado,Terca}
@attribute horario {madrugada,manha,tarde,noite}
@attribute municipio {Anapolis,Goiania,Brasilia,Goianapolis,Alexania,Abadiania,'Terezopolis de Goias',
                      'Santo Antonio do Descoberto'}
@attribute causa_acidente {'Outras causas','Ingestao alcool','Distancia seguranca','Falta atencao',
                           'Desobediencia sinalizacao','Velocidade incompativel','Defeito veiculo',
                           Dormindo,'Defeito via','Animais pista',
                           'Ultrapassagem indevida','Fator ambiente'}
@attribute tipo_acidente {'Saida pista','Colisao lateral','Colisao traseira','Colisao transversal',
                          Capotagem,'Queda veiculo','Choque objeto fixo',Atropelamento,Outros,
                         'Choque objeto movel','Atropelamento animal',Tombamento,'Colisao frontal'}
@attribute sentido_via {Decrescente,Crescente,'Nao Informado'}
@attribute uso_solo {Urbano,Rural}
@attribute tipo_veiculo {Carga,Coletivo,Passeio,Motocicleta,Bicicleta,Outros}
@attribute qtd_pessoas numeric
@attribute qtd_veiculos numeric
@attribute classificacao_acidente {SVitima,CFerido,CFatal}

@data
2012,jan,janeiro1-15,Domingo,manha,Anapolis,'Outras causas','Saida pista',Decrescente,Urbano,Carga,1,1,SVitima
2012,jan,janeiro1-15,Segunda,manha,Goiania,'Outras causas','Saida pista',Decrescente,Rural,Coletivo,1,1,SVitima
2012,jan,janeiro1-15,Domingo,manha,Anapolis,'Ingestao alcool','Saida pista',Decrescente,Urbano,Passeio,1,1,CFerido
2012,jan,janeiro1-15,Segunda,manha,Goiania,'Outras causas','Saida pista',Crescente,Rural,Passeio,1,1,SVitima
2012,jan,janeiro1-15,Domingo,madrugada,Brasilia,'Outras causas','Saida pista',Crescente,Urbano,Passeio,2,1,CFerido
```

Fonte: Vieira (2018)

O quadro 12 exibe os arquivos gerados e descreve em função de qual condição climática ele foi segmentado.

Quadro 12 – Arquivos ARFF para criação dos modelos de classificação

Arquivo	Condição Climática	Quantidade dados
acidentes060_01_precipitacao.arff	Qualquer nível de precipitação	2.071
acidentes060_02_nublado.arff	Clima Nublado	1.508
acidentes060_03_ceuClaro.arff	Clima Ensolarado e/com Céu Claro	9.625
acidentes060_04_vento.arff	Existência de Vento	103

acidentes060_05_neblina.arff	Existência de Neblina	60
	Total	**13.367**

Fonte: Vieira (2018)

3.7.4 Modelagem

Após o seguimento das fases anteriores, foi possível chegar até a modelagem, na qual os dados foram minerados originando modelos de classificação.

a) Seleção da Técnica de Modelagem:

A técnica escolhida neste passo, foi definida de acordo com a tarefa avaliada no item 3.5. Aqui, apenas serão reforçadas as decisões já tomadas. 83% da base dados, incluindo a classe eram constituídos por dados categóricos, contra apenas 17% de dados numéricos.

De acordo com os objetivos definidos na etapa Entendimento do Negócio, desejou-se criar um modelo de classificação e a geração de padrões por meio de árvores de decisão.

b) Geração do *Design* de Teste:

Nesta atividade foi necessária a delimitação de alguns pontos de como os dados seriam divididos para treinamento e teste visando a reduçao de *overfitting*, e quais métricas de avaliação seriam usadas para identificar o melhor modelo criado (quadro 13).

Quadro 13 – Métricas para avaliação dos modelos

Objetivo	Métrica
Dividir os dados em teste e treinamento	Validação cruzada com 10 dobras.
Iterar resultado do classificador	Testar a configuração 10 vezes para cada parametrização.
Medir Acurácia do Modelo	Porcentagem de acertos maior ou igual a 75%.
Medir índice de concordância do Modelo	Coeficiente de Kappa (*Kappa Statistics*) maior ou igual 5.
Medir precisão classificação	Taxa de precisão das classes maior ou igual a 0.7.
Medir relação entre taxa de erros e acertos	Curva ROC maior ou igual 0,8.

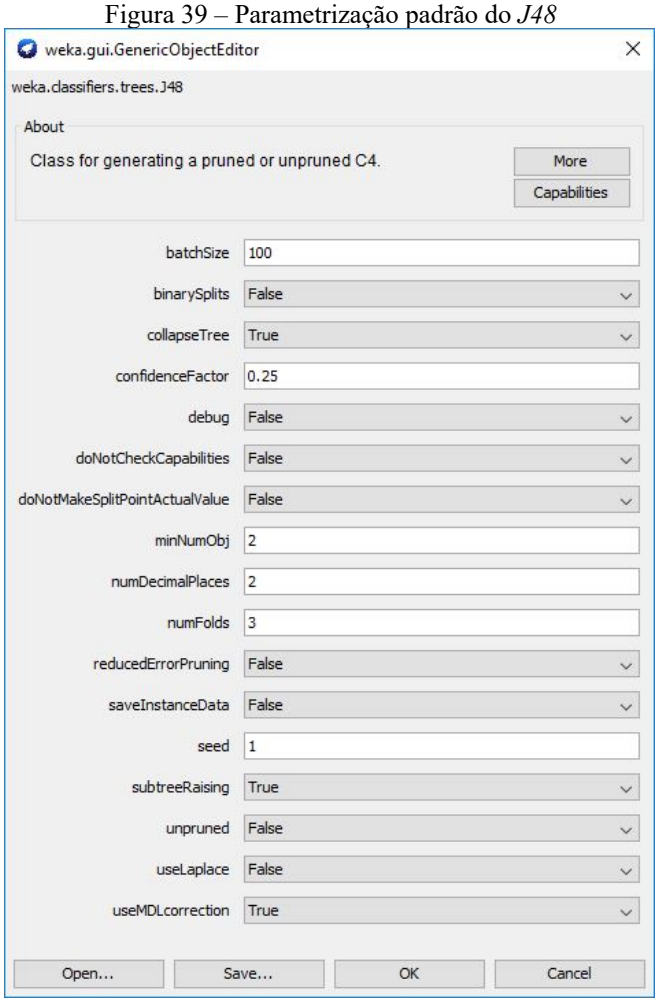
Fonte: Vieira (2018)

c) Construção do Modelo

Os modelos de classificação foram construídos e avaliados, utilizando a aplicação *Experimenter* da ferramenta *Weka 3*, dado que ela permitia uma avaliação em lote, o que facilitou o processo de escolha do melhor modelo, da melhor parametrização, assim como a exibição do resultado em forma de matriz. Dessa forma, o *Experimenter* foi utilizado para testar as combinações dos parâmetros mais relevantes do algoritmo *J48* (item 2.4.2.3).

A figura 39, exibe a tela de parametrização padrão do *J48* e o quadro 14, descreve os seus principais parâmetros.

Figura 39 – Parametrização padrão do *J48*



Fonte: Vieira (2018)

Quadro 14 – Parâmetros J48

Parâmetro	Identificador	Descrição
binarySplit	-B	Aplica divisão binária em dados nominais. Compara um valor nominal com o outro.
confidenceFactor	-C	Fator de confiança usado para poda da árvore. Varia de 0 a 1. Valores menores resultam e poda mais rigorosa.
minNumObj	-M	Determina a quantidade de instâncias por folha. Contribui para redução do *overfitting*. Valor padrão 2.
numFolds	-N	Usado em concordância com *reducedErrorPruning*, determina a quantidade de dados para realizar a poda. Valor padrão 3.
reducedErrorPruning	-R	Alternativa ao método de poda padrão do algoritmo C4.5.
subTreeRaising	-S	Realiza redução na árvore, fazendo com que sub-árvores menos significantes, sejam removidas.
unpruned	-U	Determina que a árvore gerada será podada ou não. A poda reduz o *overfitting*.
useMDLcorrection	-J	Realiza divisão de atributos numéricos e também contribui para redução do *overfitting.*

Fonte University of Waikato (2018)

Dos parâmetros listados no quadro 14, *binarySplit, confidenceFactor, minNumObj* foram selecionados para compor a parametrização que construísse o melhor classificador baseados nas métricas definidas no quadro 13. Importante destacar que a escolha desses parâmetros, foi orientada à tentativa de redução de *overfitting* nos modelos.

O quadro 15, exibe os valores possíveis que cada parâmetro recebeu. E a figura 40 exibe as 18 combinações possíveis entre esses atributos, na aplicação *Experimenter*.

Quadro 15 –Parâmetros para avaliação do algoritmo

Parâmetro	Identificador	Valores
binarySplit	-B	*True* ou *False*.
confidenceFactor	-C	0.15, 0.20 e 0.25.
minNumObj	-M	1, 2 3.

Fonte: Vieira (2018)

Figura 40 – Combinação de parâmetros usados na avaliação

```
Available resultsets

(1) trees.J48 '-C 0.25 -M 2'  ||  (10) trees.J48 '-C 0.25 -B -M 2'
(2) trees.J48 '-C 0.15 -M 1'  ||  (11) trees.J48 '-C 0.15 -B -M 1'
(3) trees.J48 '-C 0.15 -M 2'  ||  (12) trees.J48 '-C 0.15 -B -M 2'
(4) trees.J48 '-C 0.15 -M 3'  ||  (13) trees.J48 '-C 0.15 -B -M 3'
(5) trees.J48 '-C 0.2 -M 1'   ||  (14) trees.J48 '-C 0.2 -B -M 1'
(6) trees.J48 '-C 0.2 -M 2'   ||  (15) trees.J48 '-C 0.2 -B -M 2'
(7) trees.J48 '-C 0.2 -M 3'   ||  (16) trees.J48 '-C 0.2 -B -M 3'
(8) trees.J48 '-C 0.25 -M 1'  ||  (17) trees.J48 '-C 0.25 -B -M 1'
(9) trees.J48 '-C 0.25 -M 3'  ||  (18) trees.J48 '-C 0.25 -B -M 3'
```

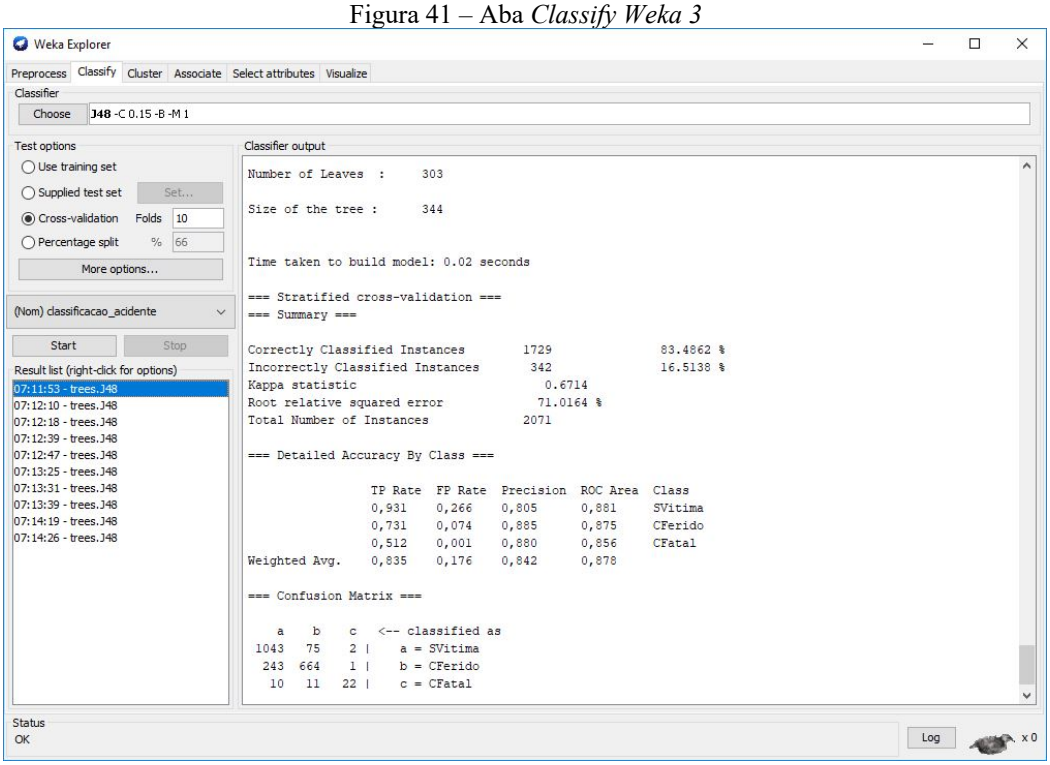
Fonte: Vieira (2018)

O resultado dessa avaliação (Apêndice D) gerou como melhores configurações os itens 8, 11, 14 e 17 (figura 40). Assim, mais uma comparação foi realizada, considerando os parâmetros *binarySplit* e *confidenceFactor*, no que resultou na escolha das configurações 8 e 11. Tais configurações foram utilizadas para criação de modelos de classificação.

Pretendeu-se ainda que os modelos pudessem gerar a árvore de decisão mais otimizada e apresentasse melhor percentual de classificação quando apresentado a dados de teste, não participantes do processo de treinamento.

Então, os conjuntos de dados apresentados no quadro 12 foram carregados na *Weka 3*, na aba *Classify*, e deu-se início a criação dos modelos (figura 41).

Figura 41 – Aba *Classify Weka 3*



Fonte: Vieira (2018)

As figuras a seguir, mostram o resultado de cada modelo de classificação criado. Primeiramente serão exibidos os modelos utilizando a parametrização 8 (figuras 42 a 46), em seguida aqueles que usaram a parametrização 11 (figuras 47 a 51).

Figura 42 – Modelo de Classificação configuração 8: Precipitação

```
Scheme:       weka.classifiers.trees.J48 -C 0.25 -M 1
Relation:     acidentes060_precipitacao

Number of Leaves  :     303
Size of the tree :  344

Time taken to build model: 0.02 seconds

=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances        1729               83.4862 %
Incorrectly Classified Instances       342               16.5138 %
Kappa statistic                          0.6714
Root relative squared error             71.0164 %
Total Number of Instances             2071

=== Detailed Accuracy By Class ===
                 TP Rate   FP Rate   Precision   ROC Area  Class
                 0,931     0,266     0,805       0,881     SVitima
                 0,731     0,074     0,885       0,875     CFerido
                 0,512     0,001     0,880       0,856     CFatal
Weighted Avg.    0,835     0,176     0,842       0,878

=== Confusion Matrix ===
    a    b    c   <-- classified as
 1043   75    2 |    a = SVitima
  243  664    1 |    b = CFerido
   10   11   22 |    c = CFatal
```

Fonte: Vieira (2018)

Figura 43 – Modelo de Classificação configuração 8: Nublado

```
Scheme:       weka.classifiers.trees.J48 -C 0.25 -M 1
Relation:     acidentes060_nublado

Number of Leaves  :     168
Size of the tree :  194

Time taken to build model: 0 seconds

=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances        1245               82.5597 %
Incorrectly Classified Instances       263               17.4403 %
Kappa statistic                          0.6722
Root relative squared error             71.6232 %
Total Number of Instances             1508

=== Detailed Accuracy By Class ===
                 TP Rate  FP Rate  Precision  ROC Area  Class
                 0,917    0,227    0,758      0,871     SVitima
                 0,768    0,098    0,897      0,870     CFerido
                 0,579    0,000    1,000      0,861     CFatal
Weighted Avg.    0,826    0,151    0,840      0,870

=== Confusion Matrix ===
   a   b   c   <-- classified as
 604  55   0 |   a = SVitima
 184 608   0 |   b = CFerido
   9  15  33 |   c = CFatal
```

Fonte: Vieira (2018)

Figura 44 – Modelo de Classificação configuração 8: Céu Claro

```
Scheme:       weka.classifiers.trees.J48 -C 0.25 -M 1
Relation:     acidentes060_ceuClaro

Number of Leaves  :     1162
Size of the tree :  1329

Time taken to build model: 0.14 seconds

=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances        7861               81.6727 %
Incorrectly Classified Instances      1764               18.3273 %
Kappa statistic                          0.6627
Root relative squared error             70.4911 %
Total Number of Instances             9625

=== Detailed Accuracy By Class ===
                 TP Rate  FP Rate  Precision  ROC Area  Class
                 0,940    0,256    0,743      0,890     SVitima
                 0,729    0,077    0,908      0,883     CFerido
                 0,623    0,002    0,939      0,933     CFatal
Weighted Avg.    0,817    0,153    0,836      0,889

=== Confusion Matrix ===
    a    b    c   <-- classified as
 3988  250    6 |    a = SVitima
 1318 3581   13 |    b = CFerido
   62  115  292 |    c = CFatal
```

Fonte: Vieira (2018)

Figura 45 – Modelo de Classificação configuração 8: Vento

```
Scheme:       weka.classifiers.trees.J48 -C 0.25 -M 1
Relation:     acidentes060_vento

Number of Leaves  :     73
Size of the tree :  81

Time taken to build model: 0 seconds

=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances          86               83.4951 %
Incorrectly Classified Instances        17               16.5049 %
Kappa statistic                          0.6907
Root relative squared error             64.8279 %
Total Number of Instances              103

=== Detailed Accuracy By Class ===
                 TP Rate  FP Rate  Precision  ROC Area  Class
                 0,714    0,107    0,714      0,901     SVitima
                 0,873    0,225    0,859      0,898     CFerido
                 0,917    0,000    1,000      0,951     CFatal
Weighted Avg.    0,835    0,167    0,836      0,905

=== Confusion Matrix ===
  a  b  c   <-- classified as
 20  8  0 |  a = SVitima
  8 55  0 |  b = CFerido
  0  1 11 |  c = CFatal
```

Fonte: Vieira (2018)

Figura 46 – Modelo de Classificação configuração 8: Neblina

```
Scheme:       weka.classifiers.trees.J48 -C 0.25 -M 1
Relation:     acidentes060_neblina
Number of Leaves  :     39
Size of the tree :  42

Time taken to build model: 0 seconds

=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances          55               91.6667 %
Incorrectly Classified Instances         5                8.3333 %
Kappa statistic                          0.8489
Root relative squared error             52.2324 %
Total Number of Instances               60

=== Detailed Accuracy By Class ===
                 TP Rate  FP Rate  Precision  ROC Area  Class
                 0,889    0,061    0,923      0,952     SVitima
                 0,933    0,067    0,933      0,972     CFerido
                 1,000    0,018    0,750      0,991     CFatal
Weighted Avg.    0,917    0,061    0,920      0,964

=== Confusion Matrix ===
  a  b  c   <-- classified as
 24  2  1 |  a = SVitima
  2 28  0 |  b = CFerido
  0  0  3 |  c = CFatal
```

Fonte: Vieira (2018)

Figura 47 – Modelo de Classificação configuração 11: Precipitação

```
Scheme:       weka.classifiers.trees.J48 -C 0.15 -B -M 1
Relation:     acidentes060_precipitacao

Number of Leaves  :     124
Size of the tree :  247

Time taken to build model: 0.06 seconds

=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances        1799               86.8662 %
Incorrectly Classified Instances       272               13.1338 %
Kappa statistic                          0.7407
Root relative squared error             65.4072 %
Total Number of Instances             2071

=== Detailed Accuracy By Class ===
                 TP Rate  FP Rate  Precision  ROC Area  Class
                 0,954    0,217    0,838      0,907     SVitima
                 0,770    0,051    0,922      0,898     CFerido
                 0,744    0,003    0,821      0,930     CFatal
Weighted Avg.    0,869    0,139    0,875      0,904

=== Confusion Matrix ===
    a    b    c   <-- classified as
 1068   52    0 |    a = SVitima
  202  699    7 |    b = CFerido
    4    7   32 |    c = CFatal
```

Fonte: Vieira (2018)

Figura 48 – Modelo de Classificação configuração 11: Nublado

```
Scheme:       weka.classifiers.trees.J48 -C 0.15 -B -M 1
Relation:     acidentes060_nublado

Number of Leaves  :     150
Size of the tree :  299

Time taken to build model: 0.08 seconds

=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances        1287               85.3448 %
Incorrectly Classified Instances       221               14.6552 %
Kappa statistic                          0.7236
Root relative squared error             69.7064 %
Total Number of Instances             1508

=== Detailed Accuracy By Class ===
                 TP Rate  FP Rate  Precision  ROC Area  Class
                 0,895    0,148    0,824      0,902     SVitima
                 0,838    0,120    0,885      0,889     CFerido
                 0,579    0,006    0,786      0,874     CFatal
Weighted Avg.    0,853    0,128    0,855      0,894

=== Confusion Matrix ===
   a   b   c   <-- classified as
 590  68   1 |   a = SVitima
 120 664   8 |   b = CFerido
   6  18  33 |   c = CFatal
```

Fonte: Vieira (2018)

Figura 49 – Modelo de Classificação configuração 11: Céu Claro

```
Scheme:       weka.classifiers.trees.J48 -C 0.15 -B -M 1
Relation:     acidentes060_ceuClaro

Number of Leaves  :     804
Size of the tree :  1607

Time taken to build model: 0.89 seconds

=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances        8153               84.7065 %
Incorrectly Classified Instances      1472               15.2935 %
Kappa statistic                          0.7174
Root relative squared error             68.9606 %
Total Number of Instances             9625

=== Detailed Accuracy By Class ===
                 TP Rate  FP Rate  Precision  ROC Area  Class
                 0,898    0,162    0,814      0,904     SVitima
                 0,820    0,116    0,881      0,891     CFerido
                 0,672    0,006    0,849      0,905     CFatal
Weighted Avg.    0,847    0,131    0,850      0,898

=== Confusion Matrix ===
    a    b    c   <-- classified as
 3812  420   12 |    a = SVitima
  842 4026   44 |    b = CFerido
   29  125  315 |    c = CFatal
```

Fonte: Vieira (2018)

Figura 50 – Modelo de Classificação configuração 11: Vento

```
Scheme:       weka.classifiers.trees.J48 -C 0.15 -B -M 1
Relation:     acidentes060_vento

Number of Leaves  :     19
Size of the tree :  37

Time taken to build model: 0 seconds

=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances          86               83.4951 %
Incorrectly Classified Instances        17               16.5049 %
Kappa statistic                          0.6982
Root relative squared error             74.7874 %
Total Number of Instances              103

=== Detailed Accuracy By Class ===
                 TP Rate  FP Rate  Precision  ROC Area  Class
                 0,786    0,133    0,688      0,886     SVitima
                 0,841    0,175    0,883      0,873     CFerido
                 0,917    0,000    1,000      0,958     CFatal
Weighted Avg.    0,835    0,143    0,844      0,887

=== Confusion Matrix ===
  a  b  c   <-- classified as
 22  6  0 |  a = SVitima
 10 53  0 |  b = CFerido
  0  1 11 |  c = CFatal
```

Fonte: Vieira (2018)

Figura 51 – Modelo de Classificação configuração 11: Neblina

```
Scheme:       weka.classifiers.trees.J48 -C 0.15 -B -M 1
Relation:     acidentes060_neblina

Number of Leaves  :     8
Size of the tree :  15

Time taken to build model: 0 seconds

=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances          57               95      %
Incorrectly Classified Instances         3                5      %
Kappa statistic                          0.9084
Root relative squared error             42.3321 %
Total Number of Instances               60

=== Detailed Accuracy By Class ===
                 TP Rate  FP Rate  Precision  ROC Area  Class
                 0,963    0,061    0,929      0,950     SVitima
                 0,933    0,033    0,966      0,951     CFerido
                 1,000    0,000    1,000      1,000     CFatal
Weighted Avg.    0,950    0,044    0,951      0,953

=== Confusion Matrix ===
  a  b  c   <-- classified as
 26  1  0 |  a = SVitima
  2 28  0 |  b = CFerido
  0  0  3 |  c = CFatal
```

Fonte: Vieira (2018)

A criação desses modelos, trouxe consigo a geração de árvores de decisão (figura 52). A ferramenta *Weka 3*, cria as árvores enquanto o modelo é gerado. Um aspecto interessante da criação de modelos de classificação em árvores de decisão usando validação cruzada, é que para cada dobra da validação, uma nova árvore é criada, de tal modo que o produto final seja a melhor árvore criada durante as iterações.

Figura 52 – Árvore de Decisão textual

```
J48 pruned tree
------------------

dia_mes = outubro1-15: CFatal (3.0)
dia_mes != outubro1-15
|   causa_acidente = Fator ambiente: CFerido (13.0)
|   causa_acidente != Fator ambiente
|   |   causa_acidente = Outras causas
|   |   |   dia_mes = dezembro16-31: SVitima (1.0)
|   |   |   dia_mes != dezembro16-31: CFerido (11.0)
|   |   causa_acidente != Outras causas
|   |   |   dia_mes = abril16-30: CFerido (2.0)
|   |   |   dia_mes != abril16-30
|   |   |   |   dia_mes = outubro16-31: CFerido (2.0)
|   |   |   |   dia_mes != outubro16-31
|   |   |   |   |   tipo_veiculo = Motocicleta: CFerido (1.0)
|   |   |   |   |   tipo_veiculo != Motocicleta: SVitima (27.0/1.0)

Number of Leaves  :     8

Size of the tree :  15
```

Fonte: Vieira (2018)

d) Avaliação do Modelo

Neste ponto os modelos criados precisaram ser avaliados, a fim de se eleger aquele do qual se extrairia o conhecimento.

A avaliação final da qualidade dos modelos de classificação, consistiu em testar dados desconhecidos durante o treinamento. O objetivo era verificar o quão adequadamente os novos registros seriam classificados. A configuração de parâmetros que melhor classificou, foi eleita como modelo de classificação final neste estudo.

Os novos registros foram coletados do PDA. Eles eram referentes às ocorrências de ATs, realizadas na região do escopo dessa pesquisa (item 3.2), mas tiveram sua ocorrência entre janeiro a agosto de 2018.

Para que esses dados pudessem ser submetidos aos modelos preditivos, foi executado sobre eles um pré-processamento semelhante àquele, no qual o conjunto de dados usados como treinamento foi submetido. Semelhantemente aos dados de treinamento, os de teste foram segmentados em função do clima (quadro 16).

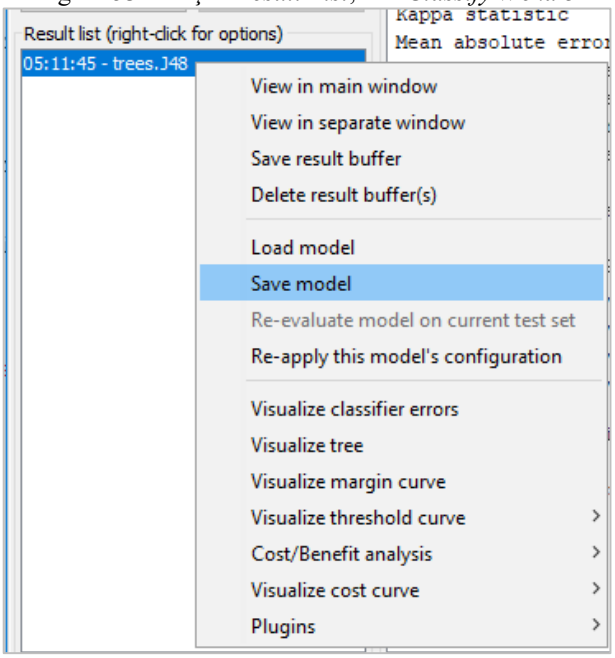
Quadro 16 – Arquivos ARFF para teste dos modelos de classificação

Arquivo	**Condição Climática**	**Quantidade Registros**
acidentes060Teste_01_precipitacao.arff	Qualquer nível de precipitação	112
acidentes060Teste_02_nublado.arff	Clima Nublado	99
acidentes060Teste_03_ceuClaro.arff	Clima Ensolarado e/com Céu Claro	659
acidentes060Teste_04_vento.arff	Existência de Vento	9
acidentes060Teste_05_neblina.arff	Existência de Neblina	1
	Total	880

Fonte: Vieira (2018)

O teste de cada classificador foi realizada através da *Weka 3*. A seção *Classify* (figura 41), possui uma região chamada *Result List* (figura 53), ao se clicar com o botão direito em cima de um resultado, é possível realizar algumas ações como: Salvar o modelo criado pelo algoritmo de aprendizado de máquina (*Save model*), Carregar um modelo previamente salvo (*Load Model*), Avaliar o modelo utilizando uma base de teste externa (*Re-evaluate model on current test set*) e, no caso específico de utilização de algoritmos de árvores de decisão, é possível visualizar em forma de grafo, a árvore gerada (*Visuzalize tree*) (figura 24).

Figura 53 – Seção *Result List*, aba *Classify Weka 3*



Fonte: Vieira (2018)

A tabela 3, permite visualizar o desempenho de cada classificador criado no item 3.7.4 C.

Tabela 3 – Avaliação dos modelos de classificação

Dados de Teste	**Configuração 8**		**Configuração 11**	
	Acertos	**Erros**	**Acertos**	**Erros**
Precipitação	60.71 % (68)	39.29 % (44)	58.03 % (65)	41.96 % (47)
Nublado	62.63% (62)	37.37% (37)	50,51% (50)	49,49% (49)
Céu Claro	62.82% (414)	37.18% (245)	58,42% (385)	41,58% (274)
Vento	0% (0)	100% (9)	22,22% (2)	77,78% (7)
Resultado	46,54%	53,46%	**47,3%**	**52,7%**

Fonte: Vieira (2018)

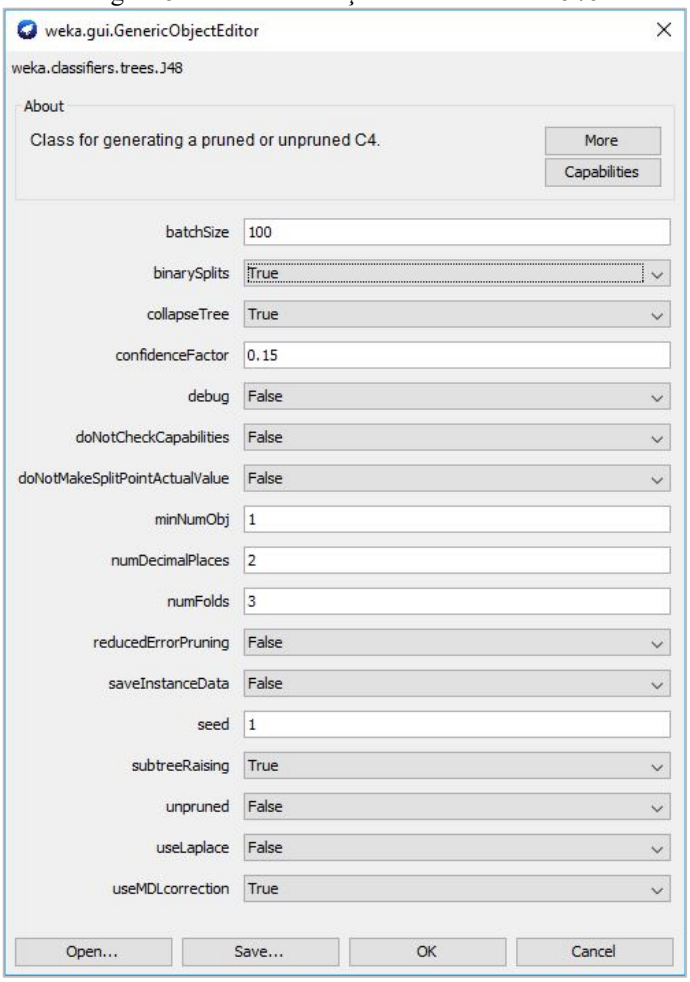
Embora os modelos criados com a configuração de parâmetros 8 (figura 41) tenham apresentado melhor taxa de acertos que a outra configuração, eles não conseguiram classificar corretamente nenhum dos registros do conjunto de teste que estavam relacionados à condição climática Vento. Desse modo, através do cálculo de uma média simples, ficou comprovado que os modelos criados a partir do resultado da segunda parametrização em avaliação, apresentaram desempenho ligeiramente melhor, em detrimento da anterior.

Todavia, cria-se a hipótese de que se houvessem mais registros no conjunto de teste relacionado a Vento, possivelmente, a configuração 8, poderia ter se sobressaído em relação a outra.

Deve-se destacar também que a condição climática Neblina não participou do teste de avaliação dos modelos, pelo motivo de que havia apenas um registro no conjunto de teste, que embora não sirva de margem para avaliação, teve aquele único registro classificado corretamente no modelo criado a partir da parametrização 11.

Conclui-se então que o melhor modelo de classificação criado neste estudo é aquele criado a partir da parametrização 11 (figura 40) visível na figura 54. Nele, a poda da árvore é mais rigorosa, eliminando as folhas menos relevantes para o modelo. As árvores de decisão dessa parametrização podem ser conferidas no Apêndice E. Os nós folhas estão organizados por cores, onde a cor verde representa classificações “Sem Vítimas”, a cor azul classificações com “Vítimas Feridas” e a cor vermelha classificações com “Vítimas Fatais”.

Figura 54 – Parametrização customizada do *J48*



Fonte: Vieira (2018)

3.7.5 Avaliação

a) Avaliação dos Resultados

Como definido no item 2.4.1 B, a penúltima etapa do CRISP-DM consiste em avaliar os resultados obtidos com os objetivos determinados na primeira fase. Dessa forma, verificou-

se neste estágio a completude de 75% dos objetivos lá definidos. O objetivo restante será avaliado na última etapa da aplicação do processo (Implementação), que no cenário aqui apresentado, será responsável por apresentar, os padrões criados por árvores de decisão.

Chama-se atenção a outro objetivo definido anteriormente. Esperava-se obter uma acurácia mínima de 75% de acertos nos modelos de classificação criados na fase Modelagem. Verificou-se que todos os modelos criados, de acordo com a melhor parametrização, apresentaram acurácia superior a 83%. (figuras 47 a 51).

b) Determinação dos Próximos Passos

Após a avaliação do modelo em relação aos objetivos definidos no início do processo, as próximas atividades a serem realizadas seriam a apresentação dos padrões identificados nos dados, como produto que possa auxiliar as entidades gestoras de trânsito e transporte na tomada de decisão em relação aos acidentes de trânsito rodoviários, perpassados no eixo Goiânia-Distrito Feral.

3.7.6 Implementação

A fase de implementação da aplicação do CRISP-DM nesta pesquisa, consistiu na execução de uma adaptação da tarefa Relatório do Produto Final, onde o conhecimento descoberto pelas árvores, criadas pelo melhor modelo de classificação foi interpretado e representado sob regras do tipo **SE** condição **(E|OU)** condição **ENTÃO** classe.

Além disso as regras deduzidas a partir das árvores foram confrontadas com as seguintes hipóteses, definidas de acordo com o referencial teórico de ATs na região do escopo da pesquisa:

1. Tipos de acidentes fatais mais comuns são colisão frontal;
2. Desobediência à sinalização (causa do acidente) está relacionada com acidentes fatais;
3. Acidentes sob neblina tendem a ser fatais;
4. Mais vítimas fatais em dezembro;
5. Falta de atenção e Excesso de velocidade estão relacionados com colisão traseira e saída de pista.

Para cada condição climática, foram escolhidas até 7 regras, para descrição textual, as quais foram construídas, prioritariamente, em função de avaliar as hipóteses supracitadas, fatalidade e acidentes com vítimas feridas, respectivamente.

a) Sob Clima com Precipitação:

SE o tipo de acidente for atropelamento E ocorrer numa quarta-feira ENTÃO há 100%

de chances de haver vítimas fatais.

SE o tipo de acidente for atropelamento E não ocorrer em Goiânia ENTÃO há 95% de chances de haver vítimas feridas.

SE o tipo de acidente for colisão frontal E envolver mais de 5 pessoas ENTÃO há 100% de chances de haver vítimas fatais

SE o tipo de veículo for motocicleta E o acidente ocorrer na primeira quinzena de junho ENTÃO há 100% de chances de haver vítimas fatais.

SE um acidente envolver mais de duas pessoas E ocorrer na primeira quinzena de janeiro E a causa do acidente for velocidade incompatível ENTÃO há 100% de chances de ocorrer vítima fatal

SE um acidente envolver mais de duas pessoas E o município for Abadiânia E o dia da semana for domingo ENTÃO há 100% de chances de haver vítimas fatais.

SE um acidente envolver mais de 2 veículos E ocorrer no sábado ENTÃO há 96% de chances de não haver vítimas.

Sobre essa condição climática, verificou-se que a hipótese de acidentes fatais em colisão frontal foi confirmada.

b) Sob clima Nublado:

SE um acidente for do tipo atropelamento E ocorrer na primeira quinzena de outubro ENTÃO há 100% de chance de haver vítimas fatais.

SE um acidente for do tipo atropelamento E ocorrer na primeira quinzena de janeiro ENTÃO há 83% de chances de haver vítimas fatais.

SE o tipo de um acidente não for atropelamento E o tipo de veículo for motocicleta e a causa do acidente for velocidade incompatível ENTÃO há 60% de chances de haver vítimas fatais.

SE o tipo de um acidente não for atropelamento E ocorrer de madrugada E a cidade for Alexânia ENTÃO há 100% de chances de haver vítimas fatais.

SE o tipo de um acidente for desobediência à sinalização ENTÃO há 100% de chances de haver vítimas feridas.

SE um acidente envolver mais que duas pessoas E a causa do acidente for dormindo ENTÃO há 100% de chances de haver vítimas fatais.

SE um acidente envolver mais que duas pessoas E ocorrer na cidade de Terezópolis de Goiás na primeira quinzena de dezembro ENTÃO há 100% de chances de haver vítimas fatais.

Sobre essa condição climática, conclui-se que nenhuma das hipóteses foram confirmadas com base nas regras descritas.

c) Sob clima com Céu Claro:

SE o tipo de veículo de um acidente for motocicleta E o tipo de acidente for colisão frontal ENTÃO há 100% de chances de haver vítimas fatais.

SE o tipo de veículo de um acidente for motocicleta E o tipo de acidente for choque com objeto fixo E a causa do acidente for velocidade incompatível ENTÃO há 100% de chances de haver vítimas fatais.

SE um acidente envolver até duas pessoas E o tipo de veículo for bicicleta ENTÃO há 76% de chances de haver vítimas feridas.

SE um acidente envolver até duas pessoas E o ocorrer durante a madrugada E na primeira quinzena de junho ENTÃO 100% de chance de haver vítimas fatais.

SE um acidente envolver acima de 4 pessoas E o tipo de acidente for colisão frontal ENTÃO há 100% de chances de haver vítimas fatais.

SE um acidente ocorrer na primeira quinzena de agosto E envolver até 7 veículos E a causa do acidente for velocidade incompatível ENTÃO há 100% de chances de haver vítimas fatais.

SE um acidente envolver até 4 pessoas E até dois veículos E o município for Abadiânia E o tipo de acidente for colisão frontal ENTÃO há 100% de chances de haver vítimas fatais.

Sobre essa condição climática, verificou-se que a hipótese de acidentes fatais em colisão frontal foi confirmada.

d) Sob clima com Vento:

SE um acidente ocorrer na primeira quinzena de maio ENTÃO há 78% de chances de haver vítimas fatais.

SE um acidente ocorrer em período diferente da primeira quinzena de maio E o município for Terezópolis de Goiás E o dia da semana for domingo ENTÃO há 100% chances de haver vítimas fatais.

SE um acidente ocorrer na primeira quinzena de julho ENTÃO há 100% de chances de haver vítimas fatais.

SE um acidente ocorrer em período diferente da primeira quinzena de maio E o tipo de veículo for carga E o dia da semana for quarta OU sábado ENTÃO há 100% de chances de haver vítimas feridas.

SE um acidente for causado por velocidade incompatível em período diferente da primeira quinzena de maio ENTÃO há 100% de chances de não haver vítimas.

SE um acidente ocorrer em período diferente da primeira quinzena de maio E o tipo de acidente for capotagem ENTÃO há 86% de chances de não haver vítimas.

Sobre essa condição climática, conclui-se que nenhuma das hipóteses foram confirmadas com base nas regras descritas.

e) Sob clima com Neblina:

SE um acidente ocorrer na primeira quinzena de outubro ENTÃO há 100% de chances de haver vítimas fatais.

SE a causa de um acidente for relacionado ao ambiente ENTÃO há 100% de chances de haver vítimas feridas.

SE um acidente ocorrer na primeira quinzena de abril ENTÃO há 100% de chances de haver vítimas feridas.

SE um acidente ocorrer na primeira quinzena de outubro ENTÃO há 100% de chances de vítimas feridas.

SE o tipo de veículo de um acidente for motocicleta ENTÃO há 100% de chances de haver vítimas feridas.

Sobre essa condição climática, conclui-se que nenhuma das hipóteses foram confirmadas com base nas regras descritas.

4 CONSIDERAÇÕES FINAIS

A Descoberta de Conhecimento em Bases de Dados com atenção à Mineração de Dados, permite analisar problemas de diversas áreas que envolvam dados, ou um mesmo problema de análise de dados sob perspectivas diferentes. Todavia, o seu resultado depende das escolhas realizadas ao longo do processo.

Esta pesquisa buscou analisar o problema de ATs rodoviários em função do clima, empregando classificação, árvores de decisão e a implementação do algoritmo *C4.5* (*J48*, na ferramenta *Weka 3*), como tarefa, técnica e algoritmos, respectivamente.

Entretanto, durante a execução do processo de DCBD, verificou-se que esse problema poderia ser tratado com o emprego de outras tarefas de MD, combinadas a técnicas e algoritmos adequados; ou também, com auxílio de técnicas adicionais para melhoria do conhecimento gerado. Isso infere a possibilidade de se descobrir padrões semelhantes ou diferentes daqueles apresentados aqui.

Foi verificado que embora o modelo tenha se apresentado eficiente no treinamento, o coeficiente de Kappa revela que os testes não tiveram bom desempenho. Foi constatado também que nem todos os dados apresentaram a qualidade necessária para avaliação. Atributos como idade, se a pessoa foi socorrida ou não após o acidente, marca e modelo de veículo apresentam erros de inserção. Isso leva a concluir que falta capacitação dos agentes que registram as informações referentes aos ATs.

4.1 Trabalhos Futuros

Como trabalhos futuros, sugere-se:

- Aplicação dos métodos desta pesquisa para outras rodovias federais, brasileiras.
- Criação de modelos de classificação com ênfase em refinamento sucessivo
- Emprego de métodos de seleção de atributos utilizando aprendizado de máquina.
- Descoberta de conhecimento, para o mesmo escopo, com utilização de algoritmos de regras de decisão.
- Descoberta de conhecimento, para o mesmo escopo, com utilização da tarefa de Associação.
- Descoberta de conhecimento, para o mesmo escopo, com uso de tarefas compostas; onde combinam-se diferentes tarefas de MD para enriquecimento do conhecimento gerado.
- Descoberta de conhecimento em função de outros fatores que incidem em ATs no eixo Goiânia-Distrito Federal, aplicando as mesmas técnicas desta pesquisa.

REFERÊNCIAS BIBLIOGRÁFICAS

1&1 COMPANY. **Data Mining Tools for Better Data Analysis**. 2017. Disponível em: <https://www.1and1.com/digitalguide/online-marketing/web-analytics/a-comparison-of-data-mining-tools/>. Acesso em: 23 maio 2018.

AGARWAL, Manish; MAZE, Tom H.; SOULEYRETTE, Reginald. **Impact of Weather on Urban Freeway Traffic Flow**: characteristics and cacility capacity. Ames: [s.n.], 2005. 15 p. Disponível em: <https://trid.trb.org/view/772850>. Acesso em: 20 abr. 2018.

AMARAL, Fernando. **Aprenda Mineração de Dados**: teoria e prática. Rio de Janeiro: Alta Books, 2016. 225 p.

AMBEV S.A (Org.). **Retrato da Segurança Viária no Brasil**: retrato da segurança viária 2017. 4. ed. Brasília: [s.n.], 2017. 101 p. Disponível em: <https://www.ambev.com.br/conteudo/uploads/2017/09/Retrato-da-Segurança-Viária_Ambev_2017.pdf>. Acesso em: 05 maio 2018.

ASSOCIAÇÃO BRASILEIRA DE NORMAS TÉCNICAS (ABNT). **NBR 10697**: pesquisa de acidentes de trânsito. [Rio de Janeiro]: ABNT Editora, 1989. 10 p.

AZEVEDO, Ana; SANTOS, Manuel Filipe. KDD, SEMMA and CRISP-DM: a parallel overview. In: IADIS EUROPEAN CONFERENCE DATA MINING, 2008. 2008, Amsterdã. **Poster.** Amsterdã: [s.n.], 2008. p. 182 - 185. Disponível em: <https://pdfs.semanticscholar.org/7dfe/3bc6035da527deaa72007a27cef94047a7f9.pdf>. Acesso em: 19 maio 2018.

BALTAR, Valéria Troncoso; OKANO, Valdir. **Análise de Concordância - Kappa.** 2017. Disponível em: <http://www.lee.dante.br/pesquisa/kappa/index.html>. Acesso em: 05 nov. 2018.

BRASIL. DEPARTAMENTO NACIONAL DE INFRAESTRUTURA DE TRANSPORTES - DNIT. **Anuário Estatístico das Rodovias Federais 2010:** acidentes de trânsito e ações de enfrentamento ao crime. [Brasília]: [s.n.], 2010. 683 p. Disponível em: <http://www.dnit.gov.br/download/rodovias/operacoes-rodoviarias/estatisticas-de-acidentes/anuario-2010.pdf>. Acesso em: 08 abr. 2018.

_____. GOVERNO DO BRASIL. **Cuidados Simples Podem Evitar Acidentes e Mortes no Trânsito**. 2017. Disponível em: <http://www.brasil.gov.br/cidadania-e-justica/2017/09/cuidados-simples-podem-evitar-acidentes-e-morte>. Acesso em: 20 maio 2018.

_____. MINISTÉRIO DOS TRANSPORTES PORTOS E AVIAÇÃO CIVIL. **Anuário Estatístico de Transportes 2010-2016**. Brasília: [s.n.], 2017. 56 p. Disponível em: <http://www.transportes.gov.br/images/2017/Sumário_Executivo_AET_-_2010_-_2016.pdf>. Acesso em: 09 abr. 2018.

CALEFFI, Felipe et al. Influência das Condições Climáticas e de Acidentes na Caracterização do Comportamento do Tráfego em Rodovias. **Transportes**, [S.l.], v. 24, n. 4, p.57-63, 1 dez. 2016. Disponível em: <https://www.revistatransportes.org.br/anpet/article/view/1104>. Acesso em: 08 jan. 2018.

CAMILO, Cássio Oliveira; SILVA, João Carlos da. **Mineração de Dados**: conceitos, tarefas, métodos e ferramentas. Goiânia: Instituto de Informática Universidade Federal de Goiás, 2009. Disponível em: <http://www.portal.inf.ufg.br/sites/default/files/uploads/relatorios-tecnicos/RT-INF_001-09.pdf>. Acesso em: 25 set. 2017.

CASTRO, Leandro Nunes de; FERRARI, Daniel Gomes. **Introdução à mineração de dados**: conceitos básicos, algoritmos e aplicações. São Paulo: Saraiva, 2016.

CHAPMAN, Pete et al. **CRISP-DM 1.0**: step-by-step data mining guide. 2000. Disponível em: <https://www.the-modeling-agency.com/crisp-dm.pdf>. Acesso em: 21 maio 2018.

CIOS, Krzysztof J. et al. **Data Mining**: a knowledge discovery approach. New York: Springer, 2007. 606 p.

CONFEDERAÇÃO NACIONAL DE MUNICÍPIOS (CNM). **Mapeamento das Mortes no Trânsito**. Brasília: CNM, 2013. 38 p. Disponível em: <http://www.cnm.org.br/cms/biblioteca/O_Mapeamento_das_mortes_no_transito.pdf>. Acesso em: 21 fev. 2018.

DEPARTAMENTO DE POLÍCIA RODOVIÁRIA FEDERAL (DPRF). Ministério da Justiça e Segurança Pública. 2016. **Acidentes**. Disponível em: < https://www.prf.gov.br/portal/dados-abertos> Acesso em: 08 jan. 2018.

DIAS, Maria Madalena. **Um Modelo de Formalização do Processo de Desenvolvimento de Sistemas de Descoberta de Conhecimento em Banco de Dados**. 2001. 2012 f. Tese (Doutorado) - Curso de Pós-graduação em Engenharia de Produção, Universidade Federal de Santa Catarina, Florianópolis, 2001. Disponível em: <https://repositorio.ufsc.br/handle/123456789/81875>. Acesso em: 21 fev. 2018. so em: 21 fev. 2018.

ELMASRI, Ramez; NAVATHE, Shamkant B. **Sistemas de Banco de Dados**. 6. ed. São Paulo: Pearson Addsion Wesley, 2011. Tradução Daniel Vieira.

FAYYAD, Usama et al. From Data Mining to Knowledge Discovery in Databases. **AI Magazine**, Palo Alto, v. 17, n. 3, p.37-54, nov. 1996. Disponível em: <https://www.aaai.org/ojs/index.php/aimagazine/article/viewFile/1230/1131>. Acesso em: 17 nov. 2017.

GALVÃO, Noemi Dreyer. **Aplicação da Mineração de Dados em Bancos da Segurança e Saúde Pública em Acidentes de Transporte.** 2009. 120 f. Tese (Doutorado) - Curso de Pós-Graduação em Enfermagem, Universidade de São Paulo, São Paulo, 2009. Disponível em: <http://repositorio.unifesp.br/handle/11600/8955>. Acesso em: 17 fev. 2018.

GOIÁS. Secretaria de Saúde. Governo de Goiás. **Acidentes de trânsito são segunda maior causa de morte em Goiás.** 2015. Disponível em: <http://www.portaldoservidor.go.gov.br/post/ver/195366/acidentes-de-transito-sao-segunda-maior-causa-de-morte-em-goias>. Acesso em: 18 fev. 2018.

GOLDSCHMIDT, Ronaldo; PASSOS, Emmanuel; BEZERRA, Eduardo. **Data mining:** conceitos, técnicas, algoritmos, orientações e aplicações. 2. ed. Rio de Janeiro: Elsevier, 2015. 276 p.

GRANATYR, Jones.; **Mineração de Regras de Associação com Weka, Apriori e Java**. 2016. Disponível em: < https://www.udemy.com/mineracao-de-regras-de-associacao-com-weka-apriori-e-java>. Acesso em: 20 ago. 2018

HAN, Jiawei; KAMBER, Micheline. **Data Mining**: concepts and techniques. San Francisco: Elsevier, 2006. 743 p.

IBM KNOWLEDGE CENTER. **Visão geral da ajuda do CRISP-DM**. [2018?]. Disponível em: <https://www.ibm.com/support/knowledgecenter/pt-br/SS3RA7_17.1.0/modeler_crispdm_ddita/clementine/crisp_help/crisp_overview.html>. Acesso em: 22 maio 2018.

INSTITUTO DE PESQUISA ECONÔMICA APLICADA (IPEA). **Acidentes de Trânsito nas Rodovias Federais Brasileiras**: caracterização, tendências e custos para a sociedade. Brasília: [s.n.], 2015. 34 p. Disponível em: <http://www.ipea.gov.br/portal/index.php?option=com_content&view=article&id=26277>. Acesso em: 08 jan. 2018.

_____. **Estimativa dos Custos dos Acidentes de Trânsito no Brasil com Base na Atualização Simplificada das Pesquisas Anteriores do Ipea.** Brasília: [s.n.], 2015. 13 p. Relatório de Pesquisa. Disponível em: <http://repositorio.ipea.gov.br/bitstream/11058/7456/1/RP_Estimativa_2015.pdf>. Acesso em: 08 abr. 2018.

_____; DEPARTAMENTO NACIONAL DE TRÂNSITO (DENATRAN). **Impactos Sociais e Econômicos dos Acidentes de Trânsito nas Rodovias Brasileiras**. Brasília: [s.n.], 2006. 79 p. Relatório Executivo. Disponível em: <http://pfdc.pgr.mpf.mp.br/informativos/edicoes-2007/janeiro/relatorio-impactos-sociais-e-economicos->. Acesso em: 08 abr. 2018.

INTERNATIONAL DATA CORPORATION (IDC). **Data Growth, Business Opportunities, and the IT Imperatives**. 2014. Disponível em: <https://www.emc.com/leadership/digital-universe/2014iview/executive-summary.htm> Acesso em: 21 nov. 2017.

LIMA, Ieda Maria de Oliveira et al. Fatores condicionantes da gravidade dos acidentes de trânsito nas rodovias brasileiras. **Texto Para Discussão**, Brasília, p.1-25, jul. 2008. Texto para discussão nº 1344. Disponível em: <http://www.ipea.gov.br/portal/images/stories/PDFs/TDs/td_1344.pdf>. Acesso em: 19 dez. 2017.

MAIO AMARELO. **O Movimento Maio Amarelo**. 2014. Disponível em: <https://www.maioamarelo.com/o-movimento/>. Acesso em: 20 maio 2018.

_____. **Observatório Alerta Motoristas para Casos de Neblina nas Estradas**. 2014. Disponível em: <http://www.onsv.org.br/observatorio-alerta-motoristas-para-casos-de-neblina-nas-estradas/>. Acesso em: 19 maio 2018.

_____. **Período é de Neblina nas Estradas e Exige Atenção Redobrada**. 2016. Disponível em: <http://www.onsv.org.br/periodo-e-de-neblina-nas-estradas-e-exige-atencao-redobrada/>. Acesso em: 19 maio 2018.

OLIVEIRA, Mariana Figueiredo de. **Relações entre os Acidentes de Trânsito e a Precipitação na Área Urbana de Rio Claro (SP)**. 2012. 67 f. TCC (Graduação) - Curso de Geografia, Universidade Estadual Paulista "Julio de Mesquita Filho" - UNESP, Rio Claro, 2012. Disponível em: <https://repositorio.unesp.br/handle/11449/120302>. Acesso em: 08 jan. 2018.

ORACLE. **Oracle Data Mining**: scalable in database predictive analytics. [2018?]. Disponível em: <http://www.oracle.com/technetwork/database/options/advanced-analytics/odm/overview/index.html>. Acesso em: 23 maio 2018.

ORGANIZAÇÃO MUNDIAL DA SAÚDE (OMS). **Classificação Estatística Internacional de Doenças e Problemas Relacionados à Saúde (CID-10)**. 15. ed. [S.l]: [s.n.], 2016. Disponível em: <http://apps.who.int/classifications/icd10/browse/2016/en>. Acesso em: 17 fev. 2018.

_____. **Relatório Global Sobre o Estado da Segurança Viária 2015**: Sumário. [Genebra]: [s.n.], 2015. 12 p. Disponível em: <http://www.who.int/violence_injury_prevention/road_safety_status/2015/en/>. Acesso em: 21 fev. 2018.

_____. **Road Safety**: Estimated road traffic death rate (per 100 000 population). 2013. Disponível em: <http://gamapserver.who.int/gho/interactive_charts/road_safety/road_traffic_deaths2/atlas.html>. Acesso em: 08 fev. 2018.

_____. **Sauver des Millions des Vies**. [S. l.]: [s.n.], 2011. 15 p. Disponível em: <http://www.who.int/violence_injury_prevention/publications/road_traffic/booklet_fr.pdf?ua=1>. Acesso em: 08 abr. 2018.

PAULA, Max Ernani Borges de; DUARTE, Augusta Maria. **Influência da chuva na ocorrência dos acidentes de trânsito.** São Paulo: [s.n.], 1996. 4 p. Nota Técnica. Disponível em: <http://www.cetsp.com.br/media/20743/nt195.pdf>. Acesso em: 20 maio 2018.

PIATETSKY-SHAPIRO, Gregory. **CRISP-DM, still the top methodology for analytics, data mining, or data science projects.** 2014. Disponível em: <https://www.kdnuggets.com/2014/10/crisp-dm-top-methodology-analytics-data-mining-data-science-projects.html>. Acesso em: 05 nov. 2018.
Citação com autor incluído no texto: Pia

_____, Gregory. **Software Suites/Platforms for Analytics, Data Mining, Data Science, and Machine Learning**. 2018. Disponível em: <https://www.kdnuggets.com/software/suites.html>. Acesso em: 20 maio 2018.

_____; PARKER, Gary. **Data Mining Course**: introduction: machine learning and data mining. 2006. Disponível em: <https://www.kdnuggets.com/data_mining_course/index.html>. Acesso em: 24 maio 2018.

REIS, Cristian Virgílio Roque. **O Uso da Descoberta de Conhecimento em Banco de Dados nos Acidentes da BR-381**. 2014. 113 f. Dissertação (Mestrado) - Curso de Mestrado Profissional em Sistemas de Informação e Gestão do Conhecimento, Universidade Fumec, Belo Horizonte, 2014. Disponível em: <http://www.fumec.br/revistas/sigc/article/view/2255>. Acesso em: 08 jan. 2018

SAS INSTITUTE. **SAS Enterprise Miner**. [2018?]. Disponível em: <https://www.sas.com/en_us/software/enterprise-miner.html>. Acesso em: 23 maio 2018.

SETZER, Valdemar W. Dado, Informação, Conhecimento e Competência. In: SETZER, Valdemar W. **Meios Eletrônicos e Educação**: uma visão alternativa. São Paulo: Escrituras, 2015. Cap. 11, p. Não paginado. Disponível em: <https://www.ime.usp.br/~vwsetzer/dado-info.html>. Acesso em: 20 maio 2018.

SHAFIQUE, Umair; QAISER, Haseeb. A Comparative Study of Data Mining Process Models (KDD, CRISP-DM and SEMMA). **International Journal of Innovation and Scientific Research**. Gurgaon, p. 217-222. nov. 2014. Disponível em: <http://www.ijisr.issr-journals.org/abstract.php?article=IJISR-14-281-04>. Acesso em: 19 maio 2018. Acesso em: 17 maio 2018.

SILBERCHATZ, Abraham; KORTH, Henry F.; SUDARCHAN, S. **Sistema de Banco de Dados**. 6. ed. Rio de Janeiro: Elsevier, 2012. 861 p.

SILVA, Leandro Augusto da; PERES, Sarajane Marques; BOSCARIOLI, Clodis. **Introdução à Mineração de Dados**: com aplicações em R. Rio de Janeiro: Elsevier, 2016. 277 p.

SISTEMA DE INFORMAÇÕES SOBRE MORTALIDADE (SIM). Departamento de Informática do Sus. **Óbitos por Causas Externas - Brasil.** 2017. Óbitos por Residência - Por Grande Grupo CID10 - Segundo Região - Período: 2012-2016. Disponível em: <http://tabnet.datasus.gov.br/cgi/tabcgi.exe?sim/cnv/ext10uf.def>. Acesso em: 20 maio 2017.

SMITH, Brian L. et al. An Investigation into the Impact of Rainfall of Freeway Traffic Flow. In: ANNUAL MEETING OF THE TRANSPORTATION RESEARCH BOARD, 83rd, 2003, Washington D.C. **Meeting…** Washington D.C.: [s.n.], 2003. 15 p. Disponível em: < https://pdfs.semanticscholar.org/4112/d4692ad05e6a18db2cf5c3659ca7906454d2.pdf >. Acesso em 20 maio 2018.

UNIVERSITY OF WAIKATO. **Weka 3**: data mining software in Java. [2018?] Disponível em: <https://www.cs.waikato.ac.nz/~ml/weka/>. Acesso em: 23 maio 2018.

ANEXO A – SOLICITAÇÃO DOS DADOS DE ACIDENTES – DPRF

Dados do Pedido

Protocolo	08850001828201813
Solicitante	Raphael dos Santos Guedes Vieira
Data de Abertura	10/04/2018 14:23
Orgão Superior Destinatário	MJ - Ministério da Justiça
Orgão Vinculado Destinatário	DPRF - Departamento de Polícia Rodoviária Federal
Prazo de Atendimento	11/05/2018
Situação	Respondido
Status da Situação	Acesso Concedido (Resposta solicitada inserida no e-SIC)
Forma de Recebimento da Resposta	Pelo sistema (com avisos por email)
Resumo	Ocorrências de Acidentes BR-060, KM 0 a KM 140
Detalhamento	Sou aluno regularmente matriculado no 9º período do curso superior em Engenharia de Computação, no CENTRO UNIVERSITÁRIO DE ANÁPOLIS - UNIEVANGÉLICA, e estou produzindo meu Trabalho de Conclusão de Curso (TCC). A pesquisa tem o objetivo de aplicar a mineração de dados nos acidentes ocorridos no eixo Goiânia-Brasília, dentro do período de 2012 a 2017. Assim, por gentileza solicito acesso aos dados das ocorrências de acidentes registradas: * Na BR-060, do KM 0 ao KM 140; * Entre janeiro de 2012 e dezembro de 2017. * Formato do arquivo: .CSV No portal de dados abertos do Departamento de Polícia Rodoviária Federal, há bases com várias informações das quais necessito, porém venho solicitar mais variáveis para que meu objetivo de pesquisa seja alcançado; baseadas nos dicionários de dados disponíveis no portal de dados abertos da PRF. Como pretendo relacionar os dados já presentes no portal com os dados anexos nessa solicitação (Arquivo em Anexo), peço que nas duas visões (por ocorrências e por pessoas), os registros sejam relacionados com os IDs respectivos. Destacados no anexo (na cor vermelha), estão variáveis que acredito ser consideradas sigilosas. Dessa forma gostaria que essas variáveis destacadas sejam analisadas e em caso afirmativo, aquelas que se enquadrem no artigo 31, parágrafo 1º e inciso I da Lei 12.527/11, podem ser desconsideradas. Caso contrário, gostaria que elas fossem juntadas ao restante dos dados solicitados. Desde já agradeço a atenção e fico no aguardo.

Dados da Resposta

Data de Resposta	11/05/2018 10:00
Tipo de Resposta	Acesso Concedido
Classificação do Tipo de Resposta	Resposta solicitada inserida no e-SIC

Resposta	Seguem as informações prestadas pela área demandada. Atenciosamente, SIC PRF
Responsável pela Resposta	Coordenador-Geral de Operações
Destinatário do Recurso de Primeira Instância:	Diretor Geral
Prazo Limite para Recurso	23/05/2018

Classificação do Pedido

Categoria do Pedido	Transportes e trânsito
Subcategoria do Pedido	Trânsito
Número de Perguntas	1

Histórico do Pedido

Data do evento	Descrição do evento	Responsável
10/04/2018 14:23	Pedido Registrado para para o Órgão DPRF - Departamento de Polícia Rodoviária Federal	SOLICITANTE
30/04/2018 09:17	Pedido Prorrogado	MJ - Ministério da Justiça/DPRF - Departamento de Polícia Rodoviária Federal
11/05/2018 10:00	Pedido Respondido	MJ - Ministério da Justiça/DPRF - Departamento de Polícia Rodoviária Federal

ANEXO B – LISTAGEM DOS DADOS SOLICITADOS – DPRF

Registros Por Ocorrência

Variável	Descrição
Informações de Localização da Ocorrência	
km	Identificação do quilômetro onde ocorreu o acidente
latitude	Latitude do local do acidente em formato geodésico decimal
longitude	Longitude do local do acidente em formato geodésico decimal
Informações da Ocorrência	
id	Identificador do acidente
intensidade_condicao_meteo	Intensidade da condição meteorológica auferida
restrição visibilidade	Tipos de restrição de visibilidade no momento da ocorrência
Informações da Rodovia/Pista	
iluminaca_rodovia	Índice de iluminação da rodovia ou km.
fluxo_movimentacao	Intensidade do fluxo de veículos no momento da ocorrência.
sinalizacao_rodovia	Índice do nível de sinalização da rodovia ou km.
relevo_pista	Tipo de relevo da pista no momento da ocorrência

Registros Por Pessoas

Variável	Descrição
id	Identificador do acidente
Informações de Localização da Ocorrência	
km	Identificação do quilômetro onde ocorreu o acidente
latitude	Latitude do local do acidente em formato geodésico decimal
longitude	Longitude do local do acidente em formato geodésico decimal.
Informações sobre o Envolvido	
id_pessoa	Identificador que identifica a vítima
data_nascimento	Data de nascimento da vítima
uf_residencia	Unidade da federação de residência da vítima
municipio_residencia	Município de residência da vítima
assistencia_medica	Vítima recebeu assistência médica no local?
Informações sobre a Situação do Envolvido	
alcoolizado/entorpecentes	Envolvido estava sob efeito de álcool e/ou entorpecentes?
Informações do Veículo	
id_veículo	Identificador do veículo
categoria_veiculo	Categoria do veículo envolvido. Ex.: Aluguel, Particular etc.
especie_veiculo	Espécie do veículo envolvido. Ex.: Carga, Especial, Passageiro etc.
cor_veiculo	Cor do veículo envolvido no acidente
condicao_pneus	Condição dos pneus do veículo envolvido na ocorrência

ANEXO C – DADOS RECEBIDOS – DPRF

Legenda BAT		
Campo	**Descrição**	**Nome do Campo**
ID_Bat	Número identificador do acidente	[id_bat]
Fase do Dia	Conforme tabela SIMPLES do banco (apenas um item por ocorrência)	[tp_fase_dia]
Iluminação Artificial	Indicador se existe iluminação artificial.	[st_existia_iluminacao]
Condição Meteorológica	Conforme tabela SIMPLES do banco (apenas um item por ocorrência)	[tp_metereologica]
Tipo de Via	Conforme tabela SIMPLES do banco (apenas um item por ocorrência)	[tp_via]
Tipo de Pista	Conforme tabela SIMPLES do banco (apenas um item por ocorrência)	[tp_pista]
Condição da Pista	Possibilidade de múltiplas escolhas e tais escolhas devem vir separadas por "-", formato: Seca- Com buraco-etc.	[tp_condicoes_pista]
Tipo de Pavimento	Conforme tabela SIMPLES do banco (apenas um item por ocorrência)	[tp_pavimento]
Estrutura Viária	Possibilidade de múltiplas escolhas, e tais escolhas devem vir separadas por "-" no formato: Reta-Em obras-Aclive-etc.	[tp_estrutura_viaria]
Urbanização	Indicador que mostra se o local é ou não urbanizado.	[st_local_urbanizado]
Acostamento	Indicador que informa a existência ou não de acostamento.	[tp_acostamento]
Canteiro Central	Indicador que informa a existência ou não de canteiro central	[tp_canteiro_central]
Sentido da Via	Se crescente ou decrescente	[tp_sentido_via]
Indicador Rodovia	Indicador que mostra se o acidente foi na rodovia ou fora dela.	[st_fora_rodovia]
ID_Envolvido	Número identificador do envolvido no acidente	[id_envolvido]
Motivo Encaminhamento Envolvido	Indica o motivo do encaminhamento do envolvido	[tp_motivo_encaminhamento]
Envolvido Não Identificado	Indicador que informa se o envolvido foi ou não identificado ou evadiu-se do local.	[st_envolvido]
Morte Após Remoção	Indicador que informa se o envolvido faleceu após a remoção do local do acidente.	[st_morreu_apos_remocao]
Estado Civil	Estado Civil do envolvido	[tp_estado_civil]
Tipo do Envolvido	Informa se o indivíduo é um condutor, passageiro, testemunha, pedestre ou cavaleiro	[tp_envolvimento]
Possibilidade Teste Alcoolemia	Indicador que informa se houve ou não possibilidade de realização do teste de alcoolemia.	[st_etilometro_indisponivel]
Recusa de Teste Alcoolemia	Indicador que informa se o usuário se recusou a realizar o teste.	[st_recusou_teste]
Sinais Embriaguez	Indicador que informa se o usuário tinha visíveis sinais de embriaguez.	[st_sinal_embriagues]
Sinais Uso Psicoativos	Indicador que informa se o usuário tinha visíveis sinais de uso de substâncias psicoativas.	[st_sinal_uso_subst]
ID_veículo	Identificador único do veículo, conforme consta no sistema.	[id_veiculo]
Ano de fabricação	Valor do ano de fabricação do veículo.	[nu_ano_fabricacao]
Marca/modelo	Identificação da marca/modelo do veículo conforme tabela do banco de dados.	[nm_marca_modelo]
Espécie	Apenas uma opção para cada veículo conforme lista simples.	[tp_especie]
Categoria	Apenas uma opção para cada veículo conforme lista simples.	[tp_categoria_veiculo]

Tipo de veículo	Apenas um tipo para cada veículo conforme lista simples.	[tp_veiculo]
Estrutura	Identificação da estrutura do veículo em "monobloco" ou "chassi", apenas se o tipo de veículo for reboque, semirreboque, caminhonete ou utilitário.	[tp_estrutura]
Articulações	Caso o tipo de veículo seja "ônibus", temos as duas possibilidades para esse atributo: "Articulado" ou "Biarticulado".	[tp_articulacao]
Data	Indica a data em que o acidente ocorreu	[dh_momento_ocorrencia]

Significado Siglas
Considerar: Pleno dia = "PD"), Plena Noite = "PN"), Amanhecer = "AM"), Anoitecer = "AN");
Considerar: Sim = "True" Não = "False"
Considerar: Céu Claro = "CC" Chuva = "CH" Garoa/Chuvisco = "GC" Granizo = "GZ" Neve = "NV" Nevoeiro/Neblina = "NN" Nublado = "NB" Sol = "SL" Vento = "VT" Ignorado = "IG"
Considerar: Marginal = "M" Principal = "P"
Considerar: Simples = "S" Dupla = "D" Múltipla = "M"
Considerar: Seca = 'A' Molhada = 'M' Com Buraco = 'B' Com Gelo = 'G' Com Lama = 'L' Com Material Granulado = 'C' Escorregadia = 'E'
Considerar: Asfalto = "A" Cascalho = "C" Concreto = "R" Paralelepípedo = "P" Terra = "T"

Considerar: Reta = 'R' Curva = 'C' Aclive = 'A' Declive = 'D' Interseção de Vias = 'I' Rotatória = 'T' Desvio Temporário = 'E' Em Obras = 'O' Ponte = 'P' Viaduto = 'V' Retorno Regulamentado = 'N' Túnel = 'L'
Considerar: Sim = "True" Não = "False"
Considerar: Sim = "True" Não = "False"
Considerar: Sim = "True" Não = "False"
Considerar: Sim = "True" Não = "False"
Considerar: "C" = Crescente "D" = Decrescente
Considerar: Sim = "True" Não = "False"
Considerar: Ausência de responsável = "A" Crime = "C" Socorro médico = "S" Outros = "O"
Considerar: Sim = "I" Não = "E" e "N"
Considerar: Sim = "True"

Não = "False"
Considerar: Solteiro(a) = "S" Casado(a) = "C" Divorciado(a) = "D" Separado(a) Judicialmente = "J" Viúvo(a) = "V" Não Informado = "N"
Condsiderar: "C" = Condutor "G" = Passageiro "E" = Testemunha "S" = Pedestre "V" = Cavaleiro
Considerar: Sim = "True" Não = "False"
Considerar: Sim = "True" Não = "False"
Considerar: Sim = "True" Não = "False"
Considerar: Sim = "True" Não = "False"
Considerar: Sim = "True" Não = "False"
Formato: aaaa.
Considerar: Passageiro = "PS" Carga = "CG" Tração = "TR" Misto = "MT" Especial = "ES" Coleção = "CL" Competição = "CP"

ANEXO D – OFÍCIO SOLICITAÇÃO DOS DADOS DE ACIDENTES – TRIUNFO CONCEBRA

Anápolis, 12 de Abril de 2018.

Of. n.º 23/2018

À Triunfo Concebra

Vimos por meio desta, solicitar dados para uma pesquisa científica, o aluno Raphael dos Santos Guedes Vieira (CPF: OCULTADO) é residente em Anápolis/GO e aluno, sob matrícula 1412248 regularmente matriculado, no 9º período do curso superior em Engenharia de Computação, no Centro Universitário de Anápolis - UniEVANGÉLICA, situado na Avenida Universitária, Km 3.5, Cidade Universitária - Anápolis/GO, CEP: 75 070-290.

Neste período o aluno está matriculado na disciplina de Trabalho de Conclusão de Curso I, ministrada pela Prof.ª Me. Luciana Nishi (Currículo Lattes: http://lattes.cnpq.br/1699054697961149) e tem como orientadora, a Prof.ª Esp. Aline Dayany de Lemos (Currículo Lattes: http://lattes.cnpq.br/9407749848661234).

O tema da pesquisa consiste na "APLICAÇÃO DE MINERAÇÃO DE DADOS NO RELACIONAMENTO ENTRE ACIDENTES RODOVIÁRIOS E FATORES CLIMÁTICOS NO EIXO GOIÂNIA-BRASÍLIA e tem como objetivo "classificar e prever padrões de acidentes relacionados a fenômenos climáticos no eixo Goiânia-Brasília no período de 2012 a 2017".

Para que o objetivo de pesquisa seja alcançado faz-se necessário o acesso às bases de dados das ocorrências existentes, de acidentes:

- Na BR-060, do KM 0 ao KM 140 (eixo Goiânia-Brasília);
- Entre janeiro de 2012 e dezembro de 2017.
- Sob formato de arquivo: .CSV

As bases que serão utilizadas e relacionadas, são as dos registros de ocorrências de acidentes do Departamento de Polícia Rodoviária Federal (DPRF), disponíveis online publicamente e as bases concedidas pela TRIUNFO CONCEBRA.

Dessa forma, vimos por meio deste, solicitar:

- O acesso às bases de dados dos registros de ocorrências de acidentes rodoviários (com exclusão de dados privados dos envolvidos, como nome e CPF); e as
- Informações relativas ao fluxo diário de veículos que passam pelos pedágios da TRIUNFO CONCEBRA.

As variáveis dos acidentes, foram elencadas com base nos dicionários de dados disponíveis no portal de dados abertos do DPRF, conforme a tabela em anexo.

Informamos que o caráter da pesquisa é puro e exclusivamente de caráter científico, excluindo toda e qualquer natureza de fim lucrativo.

Atenciosamente,

Profª Mestra Viviane C. B. Pocivi, CSM, CSPO.
UniEvangélica - Cursos de Computação
Diretora, Port. 4 de 25/01/2010

Profa. Mestra Luciana Nishi
UniEvangélica - Cursos de Computação
Coordenação de TCC

UniEVANGÉLICA
CENTRO UNIVERSITÁRIO
Associação Educativa Evangélica

Avenida Universitária, km. 3,5. Cidade Universitária – Anápolis-GO – CEP 75083-515 – Fone: (62) 3310-6600 – FAX (62) 3318-6388
"...grandes coisas fez o Senhor por nós; por isso estamos alegres." (Sl 126:3)

Registros Por Ocorrência

Variável	Descrição
Informações de Horário	
data_inversa	Data do acontecimento da ocorrência
dia_semana	Dia da semana da ocorrência
horario	Horário da ocorrência no formato hh:mm:ss
fase_dia	Fase do dia no momento do acidente
Informações de Localização da Ocorrência	
uf	Unidade da federação
br	Identificador da BR do acidente
km	Identificação do quilômetro onde ocorreu o acidente
municipio	Município de ocorrência do acidente
latitude	Latitude do local do acidente em formato geodésico decimal
longitude	Longitude do local do acidente em formato geodésico decimal
Informações da Ocorrência	
id	Identificador do acidente
causa_acidente	Identificação da causa do acidente
tipo_acidente	Identificação do tipo de acidente
classificacao_acidente	Classificação quanto à gravidade do acidente. Ex.: Com vítimas fatais, Com vítimas feridas, Sem vítimas, etc.
condicao_metereologica	Condição meteorológica no momento do acidente
intensidade_condicao_meteo	Intensidade da condição meteorológica auferida
restrição visibilidade	Tipos de restrição de visibilidade no momento da ocorrência
fluxo_movimentacao	Intensidade do fluxo de veículos no momento da ocorrência
Informações da Rodovia/Pista	
sentido_via	Sentido da via considerando o ponto de colisão
tipo_pista	Tipo da pista considerando a quantidade de faixas
tracado_via	Descrição do traçado da via

uso_solo	Descrição sobre as características do local do acidente
Informações sobre os Envolvidos	
pessoas	Total de pessoas envolvidas na ocorrência
mortos	Total de pessoas mortas envolvidas na ocorrência
feridos_leves	Total de pessoas com ferimentos leves envolvidas na ocorrência
feridos_graves	Total de pessoas com ferimentos graves envolvidas na ocorrência
ilesos	Total de pessoas ilesas envolvidas na ocorrência
ignorados	Total de pessoas envolvidas na ocorrência e que não se soube o estado físico
veiculos	Total de veículos envolvidos na ocorrência

Registros Por Pessoas

Variável	**Descrição**
id	Identificador do acidente
Informações de Horário	
data_inversa	Data do acontecimento da ocorrência
dia_semana	Dia da semana da ocorrência
horario	Horário da ocorrência no formato hh:mm:ss
Informações de Localização da Ocorrência	
uf	Unidade da federação
br	Identificador da BR do acidente
km	Identificação do quilômetro onde ocorreu o acidente
municipio	Município de ocorrência do acidente
latitude	Latitude do local do acidente em formato geodésico decimal
longitude	Longitude do local do acidente em formato geodésico decimal.
Informações sobre o Envolvido	
id_pessoa	Identificador do envolvido no acidente
data_nascimento	Data de nascimento do envolvido.

idade	Idade do envolvido
uf_residencia	Unidade da federação de residência do envolvido.
municipio_residencia	Município de residência do envolvido.
assistencia_medica	Envolvido recebeu assistência médica no local?
tipo_envolvido	Tipo de envolvido no acidente conforme sua participação no evento. Ex.: Condutor, Passageiro, Pedestre, Ciclista, Cavaleiro, etc.
sexo	Sexo do envolvido
nacionalidade	Indica a nacionalidade do envolvido
naturalidade	Indica a naturalidade do envolvido.
Informações sobre a Situação do Envolvido	
estado_fisico	Condição do envolvido conforme a gravidade das lesões. Ex.: Morto, Ferido Leve, Ferido Grave, Ileso, etc.
alcoolizado/entorpecentes	Envolvido estava sob efeito de álcool e/ou entorpecentes?
ilesos	Envolvido classificado como ileso
feridos_leves	Envolvido classificado como ferido leve
feridos_graves	Envolvido classificado como ferido grave
mortos	Envolvido classificado como morto
Informações do Veículo	
id_veiculo	Identificador do veículo
tipo_veiculo	Tipo do veículo conforme Art. 96 do Código de Trânsito Brasileiro
marca	Descrição da marca do veículo
ano_fabricacao_veiculo	Ano de fabricação do veículo
categoria_veiculo	Categoria do veículo envolvido. Ex.: Aluguel, Particular etc.
especie_veiculo	Espécie do veículo envolvido. Ex.: Carga, Especial, Passageiro etc.
cor_veiculo	Cor do veículo envolvido no acidente
condicao_pneus	Condição dos pneus do veículo envolvido na ocorrência

Quanto à Rodovia

Variável	Descrição
br	Identificador da BR
km	Identificação do quilômetro
data_manutencao	Data das manutenções realizadas na pista
tipo_manutenção	Tipo da manutenção feita na pista
iluminaca_rodovia	Índice de iluminação da rodovia
sinalizacao_rodovia	Índice do nível de sinalização da rodovia
relevo_pista	Tipo de relevo da pista

ANEXO E – SOLICITAÇÃO DOS DADOS DE ACIDENTES – TRIUNFO CONCEBRA

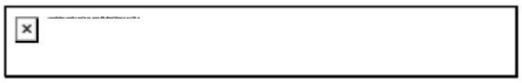
De: Prof. Luciana Nishi <OCULTADO>
Enviado em: sexta-feira, 4 de maio de 2018 08:42
Para: Protocolo Concebra
Cc: Luciana Nishi - Apoio Pedagógico, Coordenadora de TCC e Coordenadora de Estágio; Raphael Guedes; Leia Vitória dos Santos Silva - Jovem Aprendiz; OCULTADO; Aline Dayany (OCULTADO)
Assunto: Re: Oficio Engenharia de Computação - Unievangélica - Dados para Trabalho de Conclusão de Curso.
Anexos: image001.jpg

Obrigada.

Ficaremos aguardando.

Att,
Luciana

Em sex, 4 de mai de 2018 08:40, Protocolo Concebra <concebra.protocolo@triunfoconcebra.com.br> escreveu:

> Sra. Luciana,
>
> Já acionamos a área responsável, solicitamos que aguarde o retorno.
>
> Obrigada.
>
> 
>
> 
>
> **De:** Prof. Luciana Nishi [mailto:OCULTADO] **Enviada em:** sexta-feira, 4 de maio de 2018 08:36

Para: Protocolo Concebra <concebra.protocolo@triunfoconcebra.com.br>
Cc: Luciana Nishi - Apoio Pedagógico, Coordenadora de TCC e Coordenadora de Estágio <OCULTADO>; Raphael Guedes <OCULTADO>; Leia Vitória dos Santos Silva - Jovem Aprendiz <OCULTADO>; OCULTADO; Aline Dayany (OCULTADO) <OCULTADO>
Assunto: Re: Oficio Engenharia de Computação - Unievangélica - Dados para Trabalho de Conclusão de Curso.

Muito Agradecida pelo retorno.

Continuaremos aguardando resposta quanto ao ofício encaminhado. Mas Existe a possibilidade de me encaminhar o contato do setor responsável que recebeu o ofício?

Att,

Luciana

Em sex, 4 de mai de 2018 08:28, Protocolo Concebra <concebra.protocolo@triunfoconcebra.com.br> escreveu:

> Bom dia, Sra. Luciana.
>
> Informamos que ofício foi encaminhado a área responsável na data do envio e ainda não obtivemos retorno.
>
> A disposição.
>
> ---
>
> **De:** Luciana Nishi - Coordenações: TCC e Estágio [mailto:OCULTADO] **Enviada em:** quinta-feira, 3 de maio de 2018 21:04
> **Para:** Protocolo Concebra <concebra.protocolo@triunfoconcebra.com.br>
> **Cc:** Raphael Guedes <OCULTADO>; Leia Vitória dos Santos Silva - Secretária

<OCULTADO>; OCULTADO; OCULTADO; Aline Dayany
(OCULTADO) <OCULTADO>
Assunto: Re: Oficio Engenharia de Computação - Unievangélica - Dados para Trabalho de Conclusão de Curso.

Boa noite,

Reencaminho novamente o e-mail quanto ao oficio encaminhado há 20 dias.

Aguardo retorno e agradeço desde já a atenção.

Att,

Luciana

De: Luciana Nishi - Coordenações: TCC e Estágio
Enviado: sexta-feira, 13 de abril de 2018 15:55
Para: concebra.protocolo@triunfoconcebra.com.br
Cc: Viviane Carla Batista Pocivi - Diretora do Curso de Engenharia de Computação; OCULTADO; Raphael Guedes; Leia Vitória dos Santos Silva - Secretária; OCULTADO
Assunto: Oficio Engenharia de Computação - Unievangélica - Dados para Trabalho de Conclusão de Curso.

Boa tarde,

Conforme orientação da ouvidoria (e-mail abaixo), encaminho o oficio referente a coleta de dados para um trabalho de conclusão de curso do curso de engenharia de computação - Unievangélica - Anápolis.

O oficio consiste quanto a solicitação de dados para uma pesquisa científica, o aluno Raphael dos Santos Guedes Vieira (CPF: OCULTADO), que consiste na "APLICAÇÃO DE MINERAÇÃO DE DADOS NO RELACIONAMENTO ENTRE ACIDENTES RODOVIÁRIOS E FATORES CLIMÁTICOS NO EIXO GOIÂNIA-BRASÍLIA e tem como objetivo "classificar e prever padrões de acidentes relacionados a fenômenos climáticos no eixo Goiânia-Brasília no período de 2012 a 2017".

As bases que serão utilizadas e relacionadas, são as dos registros de ocorrências de acidentes do Departamento de Polícia Rodoviária Federal (DPRF), disponíveis online publicamente e as bases concedidas pela TRIUNFO CONCEBRA.

Melhor detalhamento da solicitação se encontra no oficio em anexo.

Copio o conteúdo deste e-mail para a direção do curso, assim como para a professora-orientadora e aluno-orientando do trabalho de conclusão de curso para conhecimento quanto ao encaminhamento do oficio conforme solicitado pela Triunfo-Concebra.

Atenciosamente.

Luciana Nishi - Coord. de TCC

Engenharia de Computação - Unievangélica

De: Aline Dayany <OCULTADO>

Enviada em: sexta-feira, 6 de abril de 2018 00:08
Para: Raphael Guedes <OCULTADO> **Assunto:**

Sou professora universitária e estou trabalhando um trabalho de conclusão de curso que trata sobre mineração de dados, que verifica a incidência de acidentes em BR's brasileiras. Gostariamos de limitar este escopo de pesquisa e pensamos na BR 153, de Brasilia - Goiânia. Mas para verificar essas informações de acidentes precisariamos de informações reais (par anão trabalhar com suposições).
Neste intuito, gostaria de saber se é possivel a disponibilização de alguns dados (a pesquisa não tem nenhum fim lucrativo, ou que precise de informações financeiras) sobre este treco, no período de 2016 e 2017.

* Dados caso queiram verificar a veracidade da informação
Instituição: Uni-Evangelica
Curso : Engenharia de Computação
Docente: Aline Dayany de Lemos
Matricula: 6446
Posteriormente posso encaminhar o projeto de Pesquisa.

No Aguardo

RESPOSTA

Agradecemos o contato. A Triunfo Concebra administra as rodovias BR-060/153/262, com extensão de 1.176,5 km, compreendendo o Distrito Federal e os estados de Goiás e Minas Gerais.

Por gentileza, solicitamos formalizar o pedido via ofício para o e-mail concebra.protocolo@triunfoconcebra.com.br.

Atenciosamente,
Ouvidoria

--

Aline Dayany de Lemos

"Nossa maior fraqueza está em desistir.

O caminho mais certo de vencer é tentar mais de uma vez"
Thomas Edson

Luciana Nishi

OCULTADO

Uni EVANGÉLICA
CENTRO UNIVERSITÁRIO
www.unievangelica.edu.br

ANEXO F – SEGUNDA SOLICITAÇÃO DOS DADOS DE ACIDENTES – DPRF

Dados do Pedido	
Protocolo	08850003898201814
Solicitante	Raphael dos Santos Guedes Vieira
Data de Abertura	07/08/2018 20:40
Orgão Superior Destinatário	MJ – Ministério da Justiça
Orgão Vinculado Destinatário	DPRF – Departamento de Polícia Rodoviária Federal
Prazo de Atendimento	28/08/2018
Situação	Respondido
Status da Situação	Acesso Concedido (Resposta solicitada inserida no e-SIC)
Forma de Recebimento da Resposta	Pelo sistema (com avisos por email)
Resumo	Ocorrências de Acidentes BR-060, KM 0 a KM 140 entre 2012 a 2017
Detalhamento	Olá, Essa solicitação se refere ao pedido e-SIC nº 08850001828201813 que trata da disponibilização de dados de acidentes ocorridos na BR-060, KM 0 a 140. Gostaria de agradecer o envio das informações extraídas, a saber das planilhas: * SEI 11916888; * SEI 11916929; * SEI 11916970; * SEI 11916994; * SEI 11917177. Todavia, embora tenha solicitado os dados no formato .CSV, os mesmos foram fornecidos em PDF e no modo retrato. Isso fez com que os registros fossem cortados, deixando o arquivo fornecido incompleto para o processo de análise. Por exemplo, não são exibidas na parte recebida da planilha SEI 11916970, todas as variáveis que foram informadas na planilha SEI 11917177. Desse modo, venho solicitar novamente o envio das mesmas informações fornecidas no pedido e-SIC nº 08850001828201813. Porém, por gentileza peço, que os dados: + Sejam salvos no FORMATO .CSV ou .TXT (separados por ponto e vírgula); + Sejam exportados em arquivos diferentes, de acordo com sua respectiva planilha. Ressalto que a forma como foi enviada na primeira vez (EM PDF), provocou a perda de informações relevantes para o propósito da pesquisa. Em caso de dúvidas, estou disponível para contato. Mais uma vez agradeço a atenção e fico no aguardo.

Dados da Resposta

Data de Resposta	27/08/2018 10:22
Tipo de Resposta	Acesso Concedido
Classificação do Tipo de Resposta	Resposta solicitada inserida no e-SIC
Resposta	Bom dia! Em anexo os documentos solicitados. Atenciosamente, SIC PRF
Responsável pela Resposta	PRF Márcio Corrêa - SIC/PRF
Destinatário do Recurso de Primeira Instância:	Chefe de Gabinete
Prazo Limite para Recurso	06/09/2018

Classificação do Pedido

Categoria do Pedido	Transportes e trânsito
Subcategoria do Pedido	Trânsito
Número de Perguntas	1

Histórico do Pedido

Data do evento	Descrição do evento	Responsável
07/08/2018 20:40	Pedido Registrado para para o Órgão DPRF – Departamento de Polícia Rodoviária Federal	SOLICITANTE
27/08/2018 10:22	Pedido Respondido	MJ – Ministério da Justiça/DPRF – Departamento de Polícia Rodoviária Federal

APÊNDICE A – SELEÇÃO DE ATRIBUTOS

Figura 55 – Comando (*view*) para primeira seleção de atributos

```
CREATE VIEW selecao_sem_idade as

SELECT  ps.id_pessoa, ps.data_ocorrencia, ps.dia_mes, ps.dia_semana,
        ps.horario, ps.municipio, ps.causa_acidente, ps.tipo_acidente,
        ps.classificacao_acidente, ps.sentido_via, ps.condicao_meteorologica, ps.uso_solo,
        ps.tipo_envolvido, ps.estado_fisico, ps.sexo, ps.tipo_veiculo,
        oc.pessoas AS qtd_pessoas, oc.veiculos as qtd_veiculos
FROM ocorrencia2012 oc, pessoa2012 ps WHERE oc.id = ps.id_ocorrencia

UNION ALL

SELECT  ps.id_pessoa, ps.data_ocorrencia, ps.dia_mes, ps.dia_semana,
        ps.horario, ps.municipio, ps.causa_acidente, ps.tipo_acidente,
        ps.classificacao_acidente, ps.sentido_via,ps.condicao_meteorologica, ps.uso_solo,
        ps.tipo_envolvido, ps.estado_fisico, ps.sexo,ps.tipo_veiculo,
        oc.pessoas AS qtd_pessoas, oc.veiculos as qtd_veiculos
FROM ocorrencia2013 oc, pessoa2013 ps WHERE oc.id = ps.id_ocorrencia

UNION ALL

SELECT  ps.id_pessoa, ps.data_ocorrencia, ps.dia_mes, ps.dia_semana,
        ps.horario, ps.municipio, ps.causa_acidente, ps.tipo_acidente,
        ps.classificacao_acidente, ps.sentido_via, ps.condicao_meteorologica, ps.uso_solo,
        ps.tipo_envolvido, ps.estado_fisico, ps.sexo, ps.tipo_veiculo,
        oc.pessoas AS qtd_pessoas, oc.veiculos as qtd_veiculos
FROM ocorrencia2014 oc, pessoa2014 ps WHERE oc.id = ps.id_ocorrencia

UNION ALL

SELECT  ps.id_pessoa, ps.data_ocorrencia, ps.dia_mes, ps.dia_semana,
        ps.horario, ps.municipio, ps.causa_acidente, ps.tipo_acidente,
        ps.classificacao_acidente, ps.sentido_via,ps.condicao_meteorologica, ps.uso_solo,
        ps.tipo_envolvido, ps.estado_fisico, ps.sexo, ps.tipo_veiculo,
        oc.pessoas AS qtd_pessoas, oc.veiculos as qtd_veiculos
FROM ocorrencia2015 oc, pessoa2015 ps WHERE oc.id = ps.id_ocorrencia

UNION ALL

SELECT  ps.id_pessoa, ps.data_ocorrencia, ps.dia_mes, ps.dia_semana,
        ps.horario, ps.municipio, ps.causa_acidente, ps.tipo_acidente,
        ps.classificacao_acidente, ps.sentido_via, ps.condicao_meteorologica, ps.uso_solo,
        ps.tipo_envolvido, ps.estado_fisico, ps.sexo, ps.tipo_veiculo,
        oc.pessoas AS qtd_pessoas, oc.veiculos as qtd_veiculos
FROM ocorrencia2016 oc, pessoa2016 ps WHERE oc.id = ps.id_ocorrencia

UNION ALL

SELECT  ps.id_pessoa,ps.data_ocorrencia, ps.dia_mes,ps.dia_semana,
        ps.horario, ps.municipio, ps.causa_acidente,ps.tipo_acidente,
        ps.classificacao_acidente,ps.sentido_via, ps.condicao_meteorologica, ps.uso_solo,
        ps.tipo_envolvido, ps.estado_fisico, ps.sexo,ps.tipo_veiculo,
        oc.pessoas AS qtd_pessoas, oc.veiculos as qtd_veiculos
FROM ocorrencia2017 oc, pessoa2017 ps WHERE oc.id = ps.id_ocorrencia;
```

Fonte: Vieira (2018)

Figura 56 – Comando (*view)* para segunda seleção de atributos

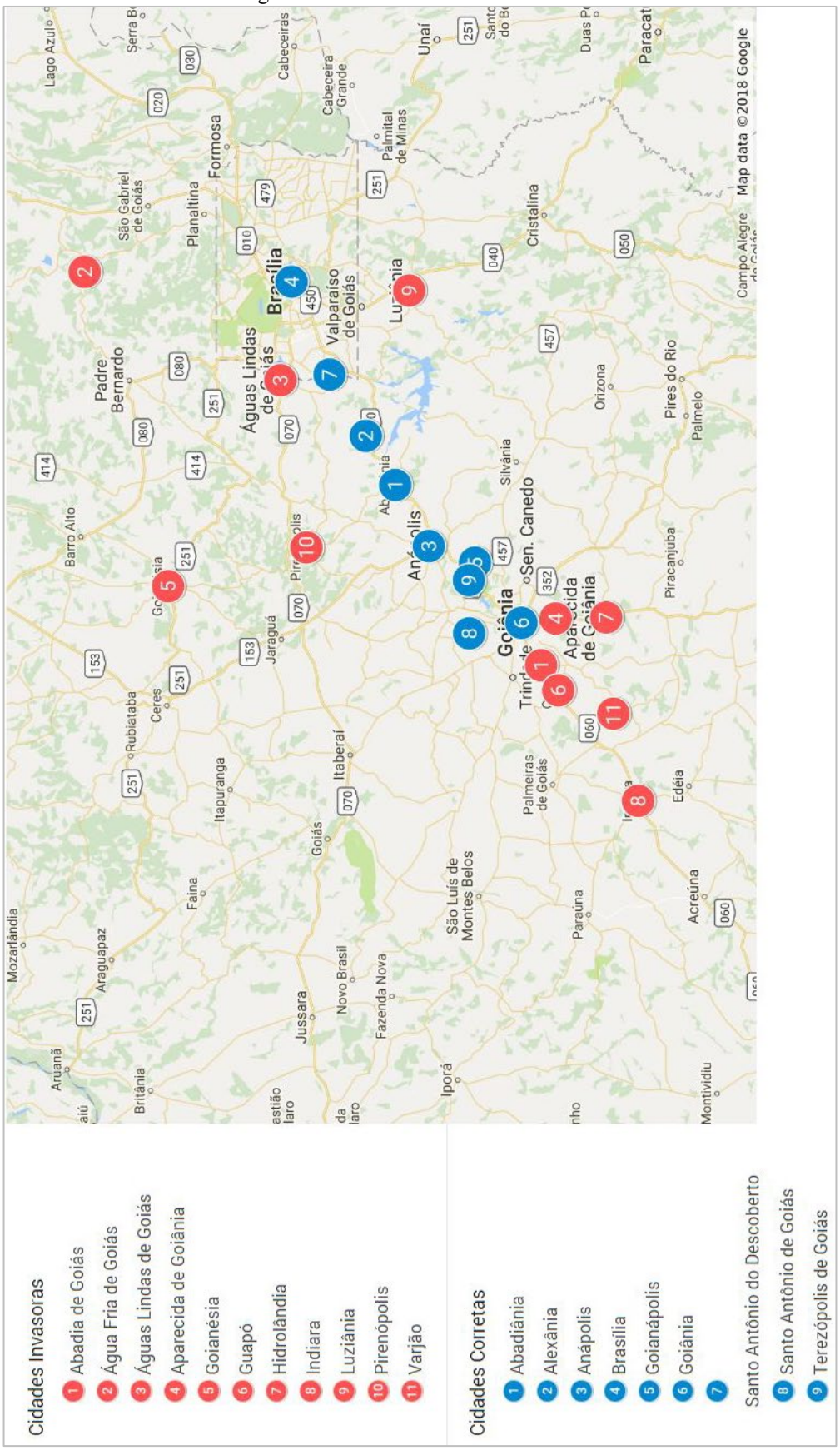
```
CREATE view selecao_sem_idade2 AS

SELECT  id_pessoa,
        data_ocorrencia,
        GetDiaMes(dia_mes) AS dia_mes,
        GetDiaSemana(dia_semana) AS dia_semana,
        horario,
        GetMunicipio(municipio) AS municipio,
        GetCausaAcidente(causa_acidente) AS causa_acidente,
        GetTipoAcidente(tipo_acidente) AS tipo_acidente,
        classificacao_acidente,
        sentido_via,
        GetCondicaoMeteo(condicao_meteorologica) AS condicao_meteorologica,
        GetUsoSolo(uso_solo) AS uso_solo,
        tipo_envolvido,
        GetEstadoFisico(estado_fisico) AS estado_fisico,
        GetSexo(sexo) AS sexo,
        GetFinalidadeVeiculo(tipo_veiculo) AS tipo_veiculo,
        qtd_pessoas,
        qtd_veiculos
FROM selecao_sem_idade;
```

Fonte: Vieira (2018)

APÊNDICE B – CIDADES REMOVIDAS

Figura 57 – Cidades invasoras e corretas



Fonte: Vieira (2018)

APÊNDICE C – REDUÇÕES DE ATRIBUTOS

Figura 58 – Redução causas de acidentes

```
delimiter $$
create function GetCausaAcidente (causa_acidente varchar(100)) returns varchar(100)
-- READS SQL DATA
-- DETERMINISTIC
begin
    declare causa_acidente_novo varchar(100);

    if (causa_acidente = 'Animais na Pista') then
        set causa_acidente_novo = 'Animais pista';

    elseif (causa_acidente = 'Avarias e/ou desgaste excessivo no pneu' or
        causa_acidente = 'Defeito mecanico em veiculo' or
        causa_acidente = 'Defeito Mecanico no Veiculo' or
        causa_acidente = 'Deficiencia ou nao Acionamento do Sistema de
        Iluminacao/Sinalizacao do Veiculo') then
        set causa_acidente_novo = 'Defeito veiculo';

    elseif (causa_acidente = 'Defeito na via' or
        causa_acidente = 'Defeito na Via' or
        causa_acidente = 'Sinalizacao da via insuficiente ou inadequada') then
        set causa_acidente_novo = 'Defeito via';

    elseif (causa_acidente = 'Desobediencia a sinalizacao' or
        causa_acidente = 'Desobediencia as normas de transito pelo condutor') then
        set causa_acidente_novo = 'Desobediencia sinalizacao';

    elseif (causa_acidente = 'Dormindo' or
        causa_acidente = 'Condutor Dormindo') then
        set causa_acidente_novo = 'Dormindo';

    elseif (causa_acidente = 'Falta de atencao' or
        causa_acidente = 'Falta de Atencao a Conducao' or
        causa_acidente = 'Falta de Atencao do Pedestre') then
        set causa_acidente_novo = 'Falta atencao';

    elseif (causa_acidente = 'Fenomenos da Natureza' or
        causa_acidente = 'Pista Escorregadia' or
        causa_acidente = 'Restricao de Visibilidade') then
        set causa_acidente_novo = 'Fator ambiente';

    elseif (causa_acidente = 'Ingestao de alcool' or
        causa_acidente = 'Ingestao de Alcool') then
        set causa_acidente_novo = 'Ingestao alcool';

    elseif (causa_acidente = 'Nao guardar distancia de seguranca') then
        set causa_acidente_novo = 'Distancia seguranca';

    elseif (causa_acidente = 'Ingestao de Substancias Psicoativas' or
        causa_acidente = 'Mal Subito' or
        causa_acidente = 'Objeto estatico sobre o leito carrocavel' or
        causa_acidente = 'Outras' or
        causa_acidente = 'Carga excessiva e/ou mal acondicionada') then
        set causa_acidente_novo = 'Outras causas';

    elseif (causa_acidente = 'Ultrapassagem indevida' or
        causa_acidente = 'Ultrapassagem Indevida') then
        set causa_acidente_novo = 'Ultrapassagem indevida';

    elseif (causa_acidente = 'Velocidade incompativel' or
        causa_acidente = 'Velocidade Incompativel') then
        set causa_acidente_novo = 'Velocidade incompativel';

    end if;
    return causa_acidente_novo;
end $$
delimiter ;
```

Fonte: Vieira (2018)

Figura 59 – Redução condições meteorológicas

```
delimiter $$
create function GetCondicaoMeteo (condicao_meteorologica varchar(25)) returns varchar(25)
-- READS SQL DATA
-- DETERMINISTIC
begin
    declare condicao_meteorologica_novo varchar(25);

    if (condicao_meteorologica = 'Ceu Claro' or
    condicao_meteorologica = 'Sol') then
    set condicao_meteorologica_novo = 'Ceu claro';

    elseif (condicao_meteorologica = 'Ignorada') then
    set condicao_meteorologica_novo = 'Ignorada';

    elseif (condicao_meteorologica = 'Nevoeiro/neblina' or
    condicao_meteorologica = 'Nevoeiro/Neblina') then
    set condicao_meteorologica_novo = 'Neblina';

    elseif (condicao_meteorologica = 'Nublado') then
    set condicao_meteorologica_novo = 'Nublado';

    elseif (condicao_meteorologica = 'Chuva' or
    condicao_meteorologica = 'Neve' or
    condicao_meteorologica = 'Garoa/Chuvisco') then
    set condicao_meteorologica_novo = 'Precipitacao';

    elseif (condicao_meteorologica = 'Vento') then
    set condicao_meteorologica_novo = 'Vento';

    end if;
    return condicao_meteorologica_novo;
end $$
delimiter ;
```

Fonte: Vieira (2018)

Figura 60 – Redução dias da semana

```
delimiter $$
create function GetDiaSemana(dia_semana varchar(15)) returns varchar(15)
-- READS SQL DATA
-- DETERMINISTIC
begin
    declare dia_semana_novo varchar(100);

    if (dia_semana = 'Domingo' or dia_semana = 'domingo') then
    set dia_semana_novo = 'Domingo';

    elseif (dia_semana = 'Segunda' or dia_semana = 'segunda-feira') then
    set dia_semana_novo = 'Segunda';

    elseif (dia_semana = 'Terca' or dia_semana = 'terca-feira') then
    set dia_semana_novo = 'Terca';

    elseif (dia_semana = 'Quarta' or dia_semana = 'quarta-feira') then
    set dia_semana_novo = 'Quarta';

    elseif (dia_semana = 'Quinta' or dia_semana = 'quinta-feira') then
    set dia_semana_novo = 'Quinta';

    elseif (dia_semana = 'Sexta' or dia_semana = 'sexta-feira') then
    set dia_semana_novo = 'Sexta';

    elseif (dia_semana = 'Sabado' or dia_semana = 'sabado') then
    set dia_semana_novo = 'sabado';

    end if;
    return dia_semana_novo;
end $$
delimiter ;
```

Fonte: Vieira (2018)

Figura 61 – Redução meses do ano 1

```
delimiter $$
create function GetDiaMes(dia_mes varchar(20)) returns varchar(20)
-- READS SQL DATA
-- DETERMINISTIC
begin
    declare dia_mes_novo varchar(20);

    if (dia_mes = '01-janeiro' or dia_mes = '02-janeiro' or dia_mes = '03-janeiro' or
        dia_mes = '04-janeiro' or dia_mes = '05-janeiro' or dia_mes = '06-janeiro' or
        dia_mes = '07-janeiro' or dia_mes = '08-janeiro' or dia_mes = '09-janeiro' or
        dia_mes = '10-janeiro' or dia_mes = '11-janeiro' or dia_mes = '12-janeiro' or
        dia_mes = '13-janeiro' or dia_mes = '14-janeiro' or dia_mes = '15-janeiro') then
        set dia_mes_novo = 'janeiro1-15';

    elseif (dia_mes = '16-janeiro' or
        dia_mes = '17-janeiro' or dia_mes = '18-janeiro' or dia_mes = '19-janeiro' or
        dia_mes = '20-janeiro' or dia_mes = '21-janeiro' or dia_mes = '22-janeiro' or
        dia_mes = '23-janeiro' or dia_mes = '24-janeiro' or dia_mes = '25-janeiro' or
        dia_mes = '26-janeiro' or dia_mes = '27-janeiro' or dia_mes = '28-janeiro' or
        dia_mes = '29-janeiro' or dia_mes = '30-janeiro' or dia_mes = '31-janeiro') then
        set dia_mes_novo = 'janeiro16-31';

    elseif (dia_mes = '01-fevereiro' or
        dia_mes = '02-fevereiro' or dia_mes = '03-fevereiro' or dia_mes = '04-fevereiro' or
        dia_mes = '05-fevereiro' or dia_mes = '06-fevereiro' or dia_mes = '07-fevereiro' or
        dia_mes = '08-fevereiro' or dia_mes = '09-fevereiro' or dia_mes = '10-fevereiro' or
        dia_mes = '11-fevereiro' or dia_mes = '12-fevereiro' or dia_mes = '13-fevereiro' or
        dia_mes = '14-fevereiro' or dia_mes = '15-fevereiro') then
        set dia_mes_novo = 'fevereiro1-15';

    elseif (dia_mes = '16-fevereiro' or
        dia_mes = '17-fevereiro' or dia_mes = '18-fevereiro' or dia_mes = '19-fevereiro' or
        dia_mes = '20-fevereiro' or dia_mes = '21-fevereiro' or dia_mes = '22-fevereiro' or
        dia_mes = '23-fevereiro' or dia_mes = '24-fevereiro' or dia_mes = '25-fevereiro' or
        dia_mes = '26-fevereiro' or dia_mes = '27-fevereiro' or dia_mes = '28-fevereiro' or
        dia_mes = '29-fevereiro') then
        set dia_mes_novo = 'fevereiro16-29';

    elseif (dia_mes = '01-março' or
        dia_mes = '02-março' or dia_mes = '03-março' or dia_mes = '04-março' or
        dia_mes = '05-março' or dia_mes = '06-março' or dia_mes = '07-março' or
        dia_mes = '08-março' or dia_mes = '09-março' or dia_mes = '10-março' or
        dia_mes = '11-março' or dia_mes = '12-março' or dia_mes = '13-março' or
        dia_mes = '14-março' or dia_mes = '15-março') then
        set dia_mes_novo = 'marco1-15';

    elseif (dia_mes = '16-março' or
        dia_mes = '17-março' or dia_mes = '18-março' or dia_mes = '19-março' or
        dia_mes = '20-março' or dia_mes = '21-março' or dia_mes = '22-março' or
        dia_mes = '23-março' or dia_mes = '24-março' or dia_mes = '25-março' or
        dia_mes = '26-março' or dia_mes = '27-março' or dia_mes = '28-março' or
        dia_mes = '29-março' or dia_mes = '30-março' or dia_mes = '31-março') then
        set dia_mes_novo = 'marco16-31';
```

Fonte: Vieira (2018)

Figura 62 – Redução meses do ano 2

```
elseif (dia_mes = '01-abril' or
    dia_mes = '02-abril' or dia_mes = '03-abril' or dia_mes = '04-abril' or
    dia_mes = '05-abril' or dia_mes = '06-abril' or dia_mes = '07-abril' or
    dia_mes = '08-abril' or dia_mes = '09-abril' or dia_mes = '10-abril' or
    dia_mes = '11-abril' or dia_mes = '12-abril' or dia_mes = '13-abril' or
    dia_mes = '14-abril' or dia_mes = '15-abril') then
    set dia_mes_novo = 'abril1-15';

elseif (dia_mes = '16-abril' or
    dia_mes = '17-abril' or dia_mes = '18-abril' or dia_mes = '19-abril' or
    dia_mes = '20-abril' or dia_mes = '21-abril' or dia_mes = '22-abril' or
    dia_mes = '23-abril' or dia_mes = '24-abril' or dia_mes = '25-abril' or
    dia_mes = '26-abril' or dia_mes = '27-abril' or dia_mes = '28-abril' or
    dia_mes = '29-abril' or dia_mes = '30-abril') then
    set dia_mes_novo = 'abril16-30';

elseif (dia_mes = '01-maio' or
    dia_mes = '02-maio' or dia_mes = '03-maio' or dia_mes = '04-maio' or
    dia_mes = '05-maio' or dia_mes = '06-maio' or dia_mes = '07-maio' or
    dia_mes = '08-maio' or dia_mes = '09-maio' or dia_mes = '10-maio' or
    dia_mes = '11-maio' or dia_mes = '12-maio' or dia_mes = '13-maio' or
    dia_mes = '14-maio' or dia_mes = '15-maio') then
    set dia_mes_novo = 'maio1-15';

elseif (dia_mes = '16-maio' or
    dia_mes = '17-maio' or dia_mes = '18-maio' or dia_mes = '19-maio' or
    dia_mes = '20-maio' or dia_mes = '21-maio' or dia_mes = '22-maio' or
    dia_mes = '23-maio' or dia_mes = '24-maio' or dia_mes = '25-maio' or
    dia_mes = '26-maio' or dia_mes = '27-maio' or dia_mes = '28-maio' or
    dia_mes = '29-maio' or dia_mes = '30-maio' or dia_mes = '31-maio') then
    set dia_mes_novo = 'maio16-31';

elseif (dia_mes = '01-junho' or
    dia_mes = '02-junho' or dia_mes = '03-junho' or dia_mes = '04-junho' or
    dia_mes = '05-junho' or dia_mes = '06-junho' or dia_mes = '07-junho' or
    dia_mes = '08-junho' or dia_mes = '09-junho' or dia_mes = '10-junho' or
    dia_mes = '11-junho' or dia_mes = '12-junho' or dia_mes = '13-junho' or
    dia_mes = '14-junho' or dia_mes = '15-junho') then
    set dia_mes_novo = 'junho1-15';

elseif (dia_mes = '16-junho' or
    dia_mes = '17-junho' or dia_mes = '18-junho' or dia_mes = '19-junho' or
    dia_mes = '20-junho' or dia_mes = '21-junho' or dia_mes = '22-junho' or
    dia_mes = '23-junho' or dia_mes = '24-junho' or dia_mes = '25-junho' or
    dia_mes = '26-junho' or dia_mes = '27-junho' or dia_mes = '28-junho' or
    dia_mes = '29-junho' or dia_mes = '30-junho') then
    set dia_mes_novo = 'junho16-30';
```

Fonte: Vieira (2018)

Figura 63 – Redução meses do ano 3

```
elseif (dia_mes = '01-julho' or
    dia_mes = '02-julho' or dia_mes = '03-julho' or dia_mes = '04-julho' or
    dia_mes = '05-julho' or dia_mes = '06-julho' or dia_mes = '07-julho' or
    dia_mes = '08-julho' or dia_mes = '09-julho' or dia_mes = '10-julho' or
    dia_mes = '11-julho' or dia_mes = '12-julho' or dia_mes = '13-julho' or
    dia_mes = '14-julho' or dia_mes = '15-julho') then
    set dia_mes_novo = 'julho1-15';

elseif (dia_mes = '16-julho' or
    dia_mes = '17-julho' or dia_mes = '18-julho' or dia_mes = '19-julho' or
    dia_mes = '20-julho' or dia_mes = '21-julho' or dia_mes = '22-julho' or
    dia_mes = '23-julho' or dia_mes = '24-julho' or dia_mes = '25-julho' or
    dia_mes = '26-julho' or dia_mes = '27-julho' or dia_mes = '28-julho' or
    dia_mes = '29-julho' or dia_mes = '30-julho' or dia_mes = '31-julho') then
    set dia_mes_novo = 'julho16-31';

elseif (dia_mes = '01-agosto' or
    dia_mes = '02-agosto' or dia_mes = '03-agosto' or dia_mes = '04-agosto' or
    dia_mes = '05-agosto' or dia_mes = '06-agosto' or dia_mes = '07-agosto' or
    dia_mes = '08-agosto' or dia_mes = '09-agosto' or dia_mes = '10-agosto' or
    dia_mes = '11-agosto' or dia_mes = '12-agosto' or dia_mes = '13-agosto' or
    dia_mes = '14-agosto' or dia_mes = '15-agosto') then
    set dia_mes_novo = 'agosto1-15';

elseif (dia_mes = '16-agosto' or
    dia_mes = '17-agosto' or dia_mes = '18-agosto' or dia_mes = '19-agosto' or
    dia_mes = '20-agosto' or dia_mes = '21-agosto' or dia_mes = '22-agosto' or
    dia_mes = '23-agosto' or dia_mes = '24-agosto' or dia_mes = '25-agosto' or
    dia_mes = '26-agosto' or dia_mes = '27-agosto' or dia_mes = '28-agosto' or
    dia_mes = '29-agosto' or dia_mes = '30-agosto' or dia_mes = '31-agosto') then
    set dia_mes_novo = 'agosto16-31';

elseif (dia_mes = '01-setembro' or
    dia_mes = '02-setembro' or dia_mes = '03-setembro' or dia_mes = '04-setembro' or
    dia_mes = '05-setembro' or dia_mes = '06-setembro' or dia_mes = '07-setembro' or
    dia_mes = '08-setembro' or dia_mes = '09-setembro' or dia_mes = '10-setembro' or
    dia_mes = '11-setembro' or dia_mes = '12-setembro' or dia_mes = '13-setembro' or
    dia_mes = '14-setembro' or dia_mes = '15-setembro') then
    set dia_mes_novo = 'setembro1-15';

elseif (dia_mes = '16-setembro' or
    dia_mes = '17-setembro' or dia_mes = '18-setembro' or dia_mes = '19-setembro' or
    dia_mes = '20-setembro' or dia_mes = '21-setembro' or dia_mes = '22-setembro' or
    dia_mes = '23-setembro' or dia_mes = '24-setembro' or dia_mes = '25-setembro' or
    dia_mes = '26-setembro' or dia_mes = '27-setembro' or dia_mes = '28-setembro' or
    dia_mes = '29-setembro' or dia_mes = '30-setembro') then
    set dia_mes_novo = 'setembro16-30';
```

Fonte: Vieira (2018)

Figura 64 – Redução meses do ano 4

```
    elseif (dia_mes = '01-outubro' or
        dia_mes = '02-outubro' or dia_mes = '03-outubro' or dia_mes = '04-outubro' or
        dia_mes = '05-outubro' or dia_mes = '06-outubro' or dia_mes = '07-outubro' or
        dia_mes = '08-outubro' or dia_mes = '09-outubro' or dia_mes = '10-outubro' or
        dia_mes = '11-outubro' or dia_mes = '12-outubro' or dia_mes = '13-outubro' or
        dia_mes = '14-outubro' or dia_mes = '15-outubro') then
        set dia_mes_novo = 'outubro1-15';

    elseif (dia_mes = '16-outubro' or
        dia_mes = '17-outubro' or dia_mes = '18-outubro' or dia_mes = '19-outubro' or
        dia_mes = '20-outubro' or dia_mes = '21-outubro' or dia_mes = '22-outubro' or
        dia_mes = '23-outubro' or dia_mes = '24-outubro' or dia_mes = '25-outubro' or
        dia_mes = '26-outubro' or dia_mes = '27-outubro' or dia_mes = '28-outubro' or
        dia_mes = '29-outubro' or dia_mes = '30-outubro' or dia_mes = '31-outubro') then
        set dia_mes_novo = 'outubro16-31';

    elseif (dia_mes = '01-novembro' or
        dia_mes = '02-novembro' or dia_mes = '03-novembro' or dia_mes = '04-novembro' or
        dia_mes = '05-novembro' or dia_mes = '06-novembro' or dia_mes = '07-novembro' or
        dia_mes = '08-novembro' or dia_mes = '09-novembro' or dia_mes = '10-novembro' or
        dia_mes = '11-novembro' or dia_mes = '12-novembro' or dia_mes = '13-novembro' or
        dia_mes = '14-novembro' or dia_mes = '15-novembro') then
        set dia_mes_novo = 'novembro1-15';

    elseif (dia_mes = '16-novembro' or
        dia_mes = '17-novembro' ordia_mes = '18-novembro' or dia_mes = '19-novembro' or
        dia_mes = '20-novembro' or dia_mes = '21-novembro' or dia_mes = '22-novembro' or
        dia_mes = '23-novembro' or dia_mes = '24-novembro' or dia_mes = '25-novembro' or
        dia_mes = '26-novembro' or dia_mes = '27-novembro' or dia_mes = '28-novembro' or
        dia_mes = '29-novembro' or dia_mes = '30-novembro') then
        set dia_mes_novo = 'novembro16-30';

    elseif (dia_mes = '01-dezembro' or
        dia_mes = '02-dezembro' or dia_mes = '03-dezembro' or dia_mes = '04-dezembro' or
        dia_mes = '05-dezembro' or dia_mes = '06-dezembro' or dia_mes = '07-dezembro' or
        dia_mes = '08-dezembro' or dia_mes = '09-dezembro' or dia_mes = '10-dezembro' or
        dia_mes = '11-dezembro' or dia_mes = '12-dezembro' or dia_mes = '13-dezembro' or
        dia_mes = '14-dezembro' or dia_mes = '15-dezembro') then
        set dia_mes_novo = 'dezembro1-15';

    elseif (dia_mes = '16-dezembro' or
        dia_mes = '17-dezembro' or dia_mes = '18-dezembro' or dia_mes = '19-dezembro' or
        dia_mes = '20-dezembro' or dia_mes = '21-dezembro' or dia_mes = '22-dezembro' or
        dia_mes = '23-dezembro' or dia_mes = '24-dezembro' or dia_mes = '25-dezembro' or
        dia_mes = '26-dezembro' or dia_mes = '27-dezembro' or dia_mes = '28-dezembro' or
        dia_mes = '29-dezembro' or dia_mes = '30-dezembro' or dia_mes = '31-dezembro') then
        set dia_mes_novo = 'dezembro16-31';

    end if;
    return dia_mes_novo;
end $$
delimiter ;
```

Fonte: Vieira (2018)

Figura 65 – Redução estado físico

```
delimiter $$
create function GetEstadoFisico (estado_fisico varchar(15)) returns varchar(15)
-- READS SQL DATA
-- DETERMINISTIC
begin
    declare estado_fisico_novo varchar(15);

    if (estado_fisico = 'Ferido Grave' or estado_fisico = 'Lesoes Graves') then
    set estado_fisico_novo = 'Ferido grave';

    elseif (estado_fisico = 'Ferido Leve' or estado_fisico = 'Lesoes Leves') then
    set estado_fisico_novo = 'Ferido leve';

    elseif (estado_fisico = 'Morto' or estado_fisico = 'Obito') then
    set estado_fisico_novo = 'Morto';

    elseif (estado_fisico = 'Ileso') then
    set estado_fisico_novo = 'Ileso';

    elseif (estado_fisico = 'Ignorado' or estado_fisico = 'Nao Informado') then
    set estado_fisico_novo = 'Ignorado';

    end if;
    return estado_fisico_novo;
end $$
delimiter ;
```

Fonte: Vieira (2018)

Figura 66 – Redução município

```
delimiter $$
create function GetMunicipio (municipio varchar(50)) returns varchar(50)
-- READS SQL DATA
-- DETERMINISTIC
begin
    declare municipio_novo varchar(50);

    if (municipio = 'ABADIANIA') then
    set municipio_novo = 'Abadiania';

    elseif (municipio = 'ALEXANIA') then
    set municipio_novo = 'Alexania';

    elseif (municipio = 'ANAPOLIS') then
    set municipio_novo = 'Anapolis';

    elseif (municipio = 'BRASILIA') then
    set municipio_novo = 'Brasilia';

    elseif (municipio = 'GOIANAPOLIS') then
    set municipio_novo = 'Goianapolis';

    elseif (municipio = 'GOIANIA') then
    set municipio_novo = 'Goiania';

    elseif (municipio = 'SANTO ANTONIO DE GOIAS' or
    municipio = 'SANTO ANTONIO DO DESCOBERTO') then
    set municipio_novo = 'Santo Antonio do Descoberto';

    elseif (municipio = 'TEREZOPOLIS DE GOIAS') then
    set municipio_novo = 'Terezopolis de Goias';

    end if;
    return municipio_novo;
end $$
delimiter ;
```

Fonte: Vieira (2018)

Figura 67 – Redução sexo

```
delimiter $$
create function GetSexo(sexo varchar(15)) returns varchar(15)
-- READS SQL DATA
-- DETERMINISTIC
begin
    declare sexo_novo varchar(15);

    if (sexo = 'Feminino' or sexo = 'F') then
    set sexo_novo = 'Feminino';

    elseif (sexo = 'Masculino' or sexo = 'M') then
    set sexo_novo = 'Masculino';

    elseif (sexo = 'Invalido' or sexo = 'I' or
    sexo = 'Ignorado' or sexo = 'Nao Informado') then
    set sexo_novo = 'Ignorado';

    end if;
    return sexo_novo;
end $$
delimiter ;
```

Fonte: Vieira (2018)

Figura 68 – Redução uso do solo

```
delimiter $$
create function GetUsoSolo(uso_solo varchar(15)) returns varchar(15)
-- READS SQL DATA
-- DETERMINISTIC
begin
    declare uso_solo_novo varchar(15);

    if (uso_solo = 'Rural' or uso_solo = 'Nao') then
    set uso_solo_novo = 'Rural';

    elseif (uso_solo = 'Urbano' or uso_solo = 'Sim') then
    set uso_solo_novo = 'Urbano';

    end if;
    return uso_solo_novo;
end $$
delimiter ;
```

Fonte: Vieira (2018)

Figura 69 – Redução tipos de acidentes

```
delimiter $$
create function GetTipoAcidente (tipo_acidente varchar(100)) returns varchar(100)
-- READS SQL DATA
-- DETERMINISTIC
begin
    declare tipo_acidente_novo varchar(100);

    if (tipo_acidente = 'Atropelamento de Pedestre' or
        tipo_acidente = 'Atropelamento de pessoa') then
        set tipo_acidente_novo = 'Atropelamento';

    elseif (tipo_acidente = 'Atropelamento de animal' or
        tipo_acidente = 'Atropelamento de Animal') then
        set tipo_acidente_novo = 'Atropelamento animal';

    elseif (tipo_acidente = 'Capotamento') then
        set tipo_acidente_novo = 'Capotagem';

    elseif (tipo_acidente = 'Colisao com objeto estatico' or
        tipo_acidente = 'Colisao com objeto fixo') then
        set tipo_acidente_novo = 'Choque objeto fixo';

    elseif (tipo_acidente = 'Colisao com bicicleta' or
        tipo_acidente = 'Colisao com objeto em movimento' or
        tipo_acidente = 'Colisao com objeto movel') then
        set tipo_acidente_novo = 'Choque objeto movel';

    elseif (tipo_acidente = 'Colisao frontal') then
        set tipo_acidente_novo = 'Colisao frontal';

    elseif (tipo_acidente = 'Colisao lateral') then
        set tipo_acidente_novo = 'Colisao lateral';

    elseif (tipo_acidente = 'Colisao Transversal' or
        tipo_acidente = 'Colisao transversal') then
        set tipo_acidente_novo = 'Colisao transversal';

    elseif (tipo_acidente = 'Colisao traseira') then
        set tipo_acidente_novo = 'Colisao traseira';

    elseif (tipo_acidente = 'Danos Eventuais' or
        tipo_acidente = 'Danos eventuais' or
        tipo_acidente = 'Derramamento de Carga' or
        tipo_acidente = 'Derramamento de carga' or
        tipo_acidente = 'Engavetamento' or
        tipo_acidente = 'Incendio') then
        set tipo_acidente_novo = 'Outros';

    elseif (tipo_acidente = 'Queda de motocicleta / bicicleta / veiculo' or
        tipo_acidente = 'Queda de motocicleta/bicicleta/veiculo' or
        tipo_acidente = 'Queda de ocupante de veiculo') then
        set tipo_acidente_novo = 'Queda veiculo';

    elseif (tipo_acidente = 'Saida de leito carrocavel' or
        tipo_acidente = 'Saida de Pista') then
        set tipo_acidente_novo = 'Saida pista';

    elseif (tipo_acidente = 'Tombamento') then
        set tipo_acidente_novo = 'Tombamento';

    end if;
    return tipo_acidente_novo;
end $$
delimiter ;
```

Fonte: Vieira (2018)

Figura 70 – Redução tipos de veículos

```
delimiter $$
create function GetFinalidadeVeiculo (tipo_veiculo varchar(25)) returns varchar(25)
-- READS SQL DATA
-- DETERMINISTIC
begin
    declare tipo_veiculo_novo varchar(25);

    if (tipo_veiculo = 'Bicicleta') then
        set tipo_veiculo_novo = 'Bicicleta';

    elseif (tipo_veiculo = 'Caminhao' or
        tipo_veiculo = 'Caminhao-Tanque' or
        tipo_veiculo = 'Caminhao-Trator' or
        tipo_veiculo = 'Caminhao-trator' or
        tipo_veiculo = 'Reboque' or
        tipo_veiculo = 'Semi-Reboque' or
        tipo_veiculo = 'Semireboque') then
        set tipo_veiculo_novo = 'Carga';

    elseif (tipo_veiculo = 'Micro-onibus' or
        tipo_veiculo = 'Microonibus' or
        tipo_veiculo = 'Onibus') then
        set tipo_veiculo_novo = 'Coletivo';

    elseif (tipo_veiculo = 'Ignorado' or
        tipo_veiculo = 'Nao identificado') then
        set tipo_veiculo_novo = 'Ignorado';

    elseif (tipo_veiculo = 'Ciclomotor' or
        tipo_veiculo = 'Motocicleta' or
        tipo_veiculo = 'Motocicletas' or
        tipo_veiculo = 'Motoneta') then
        set tipo_veiculo_novo = 'Motocicleta';

    elseif (tipo_veiculo = 'Carroca' or
        tipo_veiculo = 'Carroca-charrete' or
        tipo_veiculo = 'Outros' or
        tipo_veiculo = 'Trator de rodas' or
        tipo_veiculo = 'Trator misto') then
        set tipo_veiculo_novo = 'Outros';

    elseif (tipo_veiculo = 'Automovel' or
        tipo_veiculo = 'Caminhonete' or
        tipo_veiculo = 'Camioneta' or
        tipo_veiculo = 'Utilitario') then
        set tipo_veiculo_novo = 'Passeio';

    end if;
    return tipo_veiculo_novo;
end $$
delimiter ;
```

Fonte: Vieira (2018)

APÊNDICE D – AVALIAÇÃO DE RESULTADOS DE TESTE DE PARAMETRIZAÇÃO

Figura 71 – Avaliação das parametrizações *J48*

Analysing: Percent_correct

Dataset	(1) trees.J4	(2) trees	(3) trees	(4) trees	(5) trees	(6) trees	(7) trees	(8) trees	(9) trees	(10) tree	(11) tree	(12) tree	(13) tree	(14) tree	(15) tree	(16) tree	(17) tree	(18) tree
acidentes060_precipitacao	(100) 83.22	83.16	82.63 *	81.67 *	83.31	82.73 *	81.72 *	83.76 v	82.29 *	84.48	86.90 v	84.34	81.81	86.94 v	84.46	81.52 *	86.82 v	81.66
acidentes060_nublado	(100) 82.07	81.51	81.24 *	80.40 *	81.88	81.69	80.74 *	82.25	81.27 *	83.47	84.98 v	82.27	79.23 *	85.86 v	82.62	79.34 *	86.74 v	79.64 *
acidentes060_ceuClaro	(100) 81.22	80.44 *	79.78 *	78.45 *	81.38	80.73 *	79.45 *	81.88 v	79.96 *	83.19 v	84.80 v	82.54 v	80.52	85.37 v	82.80 v	80.54	85.90 v	80.57
acidentes060_vento	(100) 78.24	68.79 *	68.49 *	70.82	77.24	73.57	73.41	81.83	80.61	79.11	82.32	79.09	77.64	82.51	79.10	78.34	82.31	78.91
acidentes060_neblina	(100) 89.33	91.67	89.33	83.67	91.67	89.33	83.33	91.67	83.33	83.83	94.50	84.00	81.17	94.50	83.83	80.67	94.50	83.50
Average	82.82	81.11	80.30	79.00	83.09	81.61	79.73	84.28	81.49	82.82	86.70	82.45	80.07	87.04	82.56	80.08	87.25	80.86
(v/ /*)		(0/3/2)	(0/1/4)	(0/2/3)	(0/5/0)	(0/3/2)	(0/2/3)	(2/3/0)	(0/2/3)	(1/4/0)	(3/2/0)	(1/4/0)	(0/4/1)	(3/2/0)	(1/4/0)	(0/3/2)	(3/2/0)	(0/4/1)

Analysing: Kappa_statistic

Dataset	(1) trees.J	(2) tree	(3) tree	(4) tree	(5) tree	(6) tree	(7) tree	(8) tree	(9) tree	(10) tre	(11) tre	(12) tre	(13) tre	(14) tre	(15) tre	(16) tre	(17) tre	(18) tre
acidentes060_precipitacao	(100) 0.67	0.66	0.65 *	0.63 *	0.67	0.66 *	0.64 *	0.68 v	0.65 *	0.69	0.74 v	0.69	0.64	0.74 v	0.69	0.63	0.74 v	0.64
acidentes060_nublado	(100) 0.66	0.65	0.65 *	0.63 *	0.66	0.66	0.64 *	0.67	0.65 *	0.69	0.72 v	0.67	0.61 *	0.73 v	0.67	0.61 *	0.75 v	0.62 *
acidentes060_ceuClaro	(100) 0.65	0.64 *	0.63 *	0.60 *	0.66	0.64 *	0.62 *	0.67 v	0.63 *	0.69 v	0.72 v	0.68 v	0.64	0.73 v	0.68 v	0.64	0.74 v	0.64
acidentes060_vento	(100) 0.57	0.33 *	0.32 *	0.34 *	0.53	0.46	0.44	0.64	0.62	0.62	0.67	0.61	0.58	0.68	0.61	0.60	0.67	0.61
acidentes060_neblina	(100) 0.80	0.85	0.80	0.70	0.85	0.80	0.69	0.85	0.69	0.69	0.90	0.70	0.65	0.90	0.69	0.64	0.90	0.69
Average	0.67	0.63	0.61	0.58	0.67	0.64	0.61	0.70	0.65	0.68	0.75	0.67	0.62	0.76	0.67	0.62	0.76	0.64
(v/ /*)		(0/3/2)	(0/1/4)	(0/1/4)	(0/5/0)	(0/3/2)	(0/2/3)	(2/3/0)	(0/2/3)	(1/4/0)	(3/2/0)	(1/4/0)	(0/4/1)	(3/2/0)	(1/4/0)	(0/4/1)	(3/2/0)	(0/4/1)

Analysing: IR_precision

Dataset	(1) trees.J	(2) tree	(3) tree	(4) tree	(5) tree	(6) tree	(7) tree	(8) tree	(9) tree	(10) tre	(11) tre	(12) tre	(13) tre	(14) tre	(15) tre	(16) tre	(17) tre	(18) tre
acidentes060_precipitacao	(100) 0.81	0.81	0.80	0.80 *	0.81	0.80	0.80 *	0.81	0.80	0.83 v	0.84 v	0.82 v	0.80	0.84 v	0.82 v	0.81	0.85 v	0.81
acidentes060_nublado	(100) 0.76	0.75	0.75	0.74 *	0.75	0.75	0.74 *	0.76	0.75 *	0.82 v	0.82 v	0.79 v	0.76	0.83 v	0.80 v	0.77	0.85 v	0.78
acidentes060_ceuClaro	(100) 0.74	0.74	0.73 *	0.73 *	0.74 v	0.74 *	0.73 *	0.74 v	0.73 *	0.81 v	0.82 v	0.79 v	0.77 v	0.83 v	0.80 v	0.78 v	0.83 v	0.78 v
acidentes060_vento	(89) 0.64	0.23	0.20 *	0.26	0.59	0.48	0.56	0.70	0.68	0.71	0.74	0.71	0.68	0.75	0.71	0.69	0.74	0.70
acidentes060_neblina	(98) 0.90	0.94	0.90	0.86	0.94	0.90	0.86	0.94	0.86	0.84	0.95	0.84	0.84	0.95	0.84	0.83	0.95	0.89
Average	0.77	0.69	0.68	0.68	0.76	0.73	0.74	0.79	0.76	0.80	0.83	0.79	0.77	0.84	0.79	0.78	0.84	0.79
(v/ /*)		(0/5/0)	(0/3/2)	(0/2/3)	(1/4/0)	(0/4/1)	(0/2/3)	(1/4/0)	(0/3/2)	(3/2/0)	(3/2/0)	(3/2/0)	(1/4/0)	(3/2/0)	(3/2/0)	(1/4/0)	(3/2/0)	(1/4/0)

Analysing: Area_under_ROC

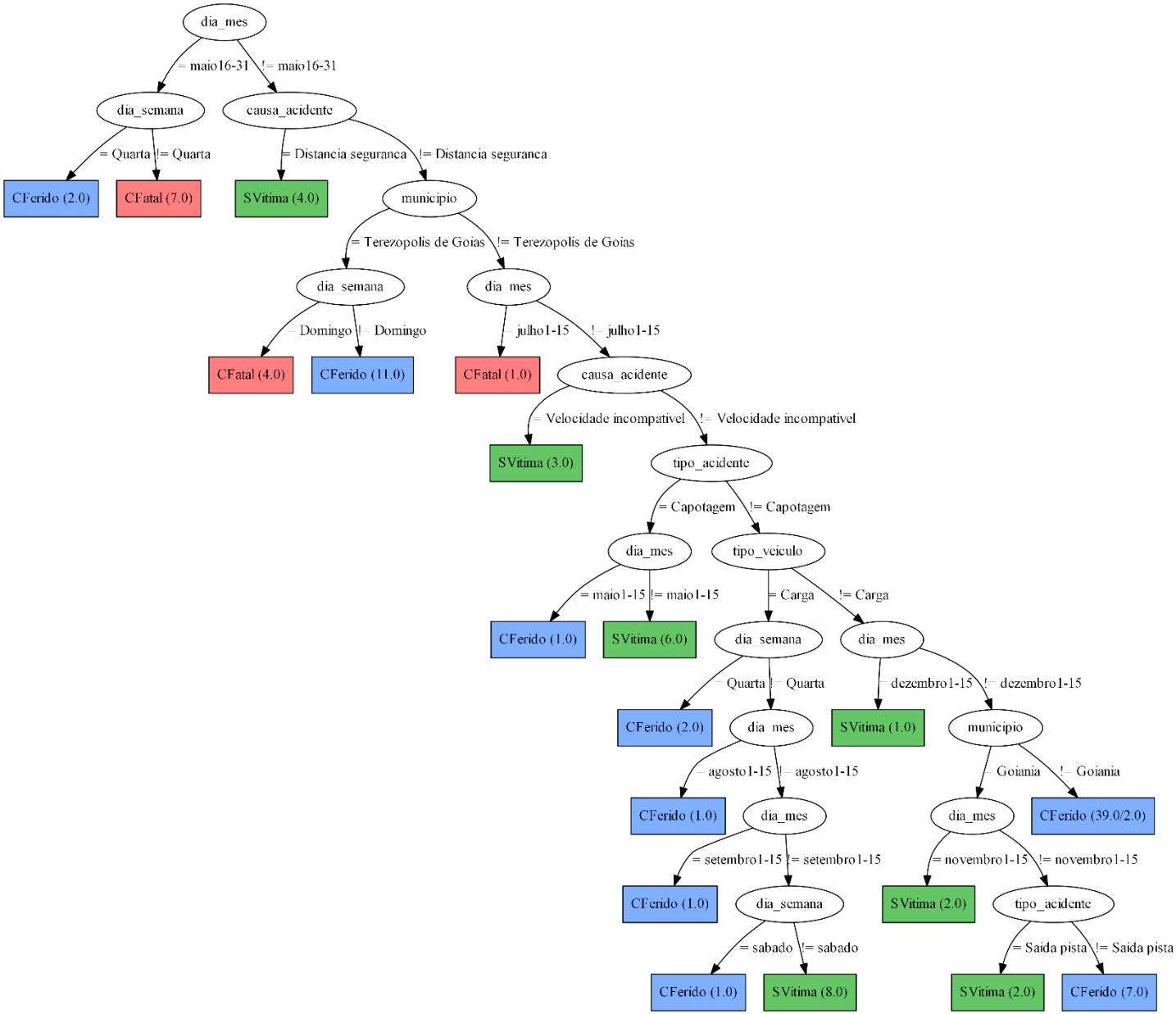
Dataset	(1) trees.J	(2) tree	(3) tree	(4) tree	(5) tree	(6) tree	(7) tree	(8) tree	(9) tree	(10) tre	(11) tre	(12) tre	(13) tre	(14) tre	(15) tre	(16) tre	(17) tre	(18) tre
acidentes060_precipitacao	(100) 0.89	0.88	0.88 *	0.87 *	0.88	0.88 *	0.87 *	0.89	0.88	0.90	0.91 v	0.89	0.88	0.91 v	0.90	0.88	0.91 v	0.89
acidentes060_nublado	(100) 0.88	0.87	0.87	0.87 *	0.88	0.88	0.87 *	0.88	0.88	0.91 v	0.90	0.89	0.88	0.90 v	0.90 v	0.88	0.91 v	0.89
acidentes060_ceuClaro	(100) 0.89	0.88 *	0.88 *	0.87 *	0.89	0.89 *	0.88 *	0.89 v	0.88 *	0.91 v	0.91 v	0.90 v	0.89	0.90 v	0.91 v	0.90	0.91 v	0.90
acidentes060_vento	(100) 0.82	0.65 *	0.65 *	0.63 *	0.79	0.76	0.74	0.85	0.83	0.82	0.84	0.82	0.82	0.84	0.82	0.83	0.83	0.83
acidentes060_neblina	(100) 0.92	0.93	0.92	0.92	0.93	0.92	0.92	0.93	0.92	0.85	0.94	0.85	0.82	0.94	0.85	0.82	0.94	0.85
Average	0.88	0.84	0.84	0.83	0.87	0.86	0.86	0.89	0.88	0.88	0.90	0.87	0.86	0.90	0.87	0.86	0.90	0.87
(v/ /*)		(0/3/2)	(0/2/3)	(0/1/4)	(0/5/0)	(0/3/2)	(0/2/3)	(1/4/0)	(0/4/1)	(2/3/0)	(2/3/0)	(1/4/0)	(0/5/0)	(3/2/0)	(2/3/0)	(0/5/0)	(3/2/0)	(0/5/0)

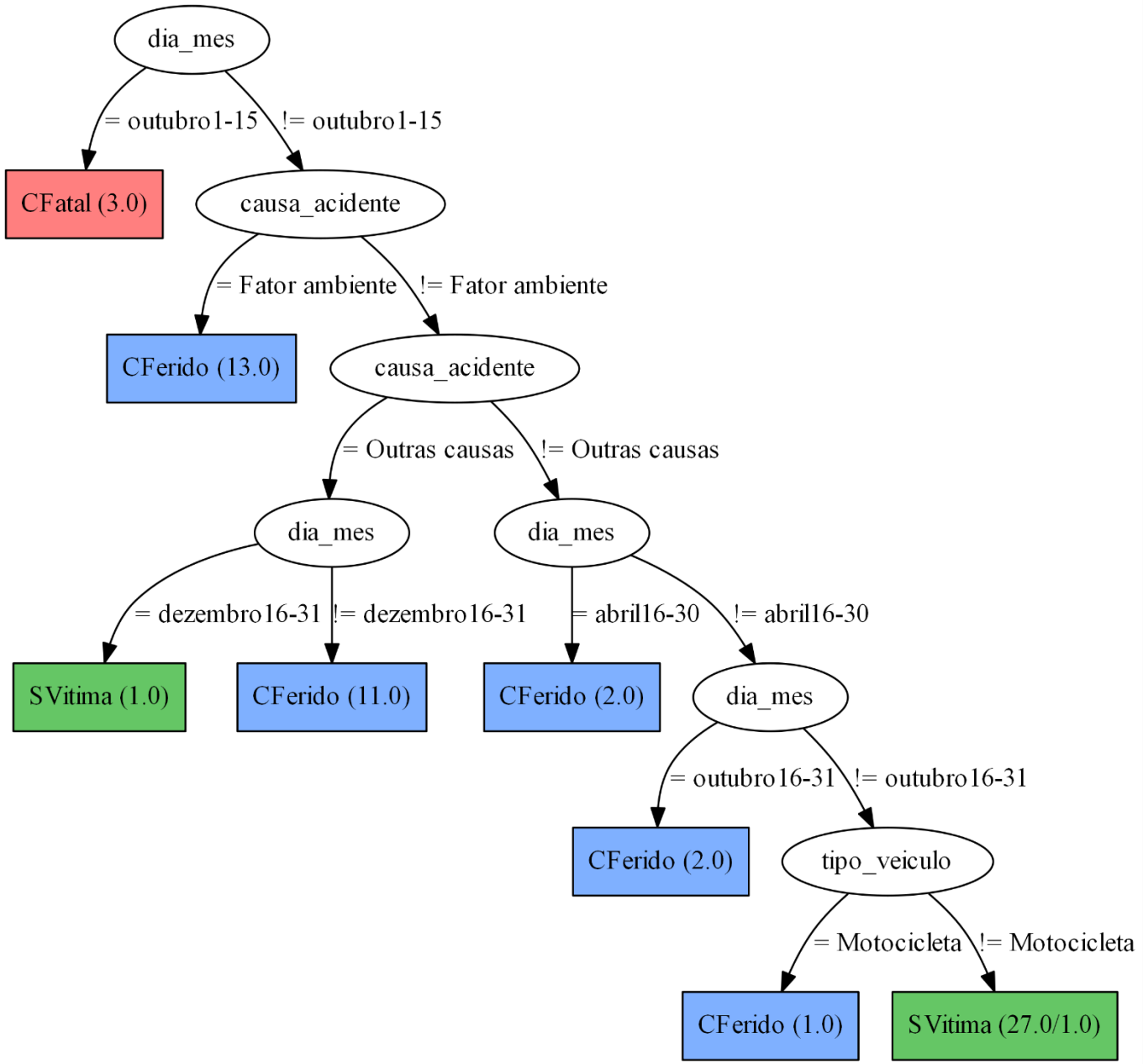
Fonte: Vieira (2018)

APÊNDICE E – ÁRVORES DE DECISÃO

Observação:

- Rótulos Vermelhos correspondem a acidentes com vítimas fatais;
- Rótulos Azuis correspondem a acidentes com vítimas feridas;
- Rótulos Verdes correspondem a acidentes sem vítimas.


dia_mes
= maio16-31
!= maio16-31
dia_semana
causa_acidente
= Quarta
!= Quarta
= Distancia seguranca
!= Distancia seguranca
CFerido (2.0)
CFatal (7.0)
SVitima (4.0)
municipio
= Terezopolis de Goias
!= Terezopolis de Goias
dia_semana
dia_mes
= Domingo
!= Domingo
julho1-15
!= julho1-15
CFatal (4.0)
CFerido (11.0)
CFatal (1.0)
causa_acidente
= Velocidade incompativel
!= Velocidade incompativel
SVitima (3.0)
tipo_acidente
= Capotagem
!= Capotagem
dia_mes
tipo_veiculo
= maio1-15
!= maio1-15
= Carga
!= Carga
CFerido (1.0)
SVitima (6.0)
dia_semana
dia_mes
= Quarta
!= Quarta
= dezembro1-15
!= dezembro1-15
CFerido (2.0)
dia_mes
SVitima (1.0)
municipio
= agosto1-15
!= agosto1-15
= Goiania
!= Goiania
CFerido (1.0)
dia_mes
dia_mes
CFerido (39.0/2.0)
= setembro1-15
!= setembro1-15
= novembro1-15
!= novembro1-15
CFerido (1.0)
dia_semana
SVitima (2.0)
tipo_acidente
= sabado
!= sabado
= Saida pista
!= Saida pista
CFerido (1.0)
SVitima (8.0)
SVitima (2.0)
CFerido (7.0)


dia_mes
= outubro1-15
!= outubro1-15
CFatal (3.0)
causa_acidente
= Fator ambiente
!= Fator ambiente
CFerido (13.0)
causa_acidente
= Outras causas
!= Outras causas
dia_mes
dia_mes
= dezembro16-31
!= dezembro16-31
= abril16-30
!= abril16-30
SVitima (1.0)
CFerido (11.0)
CFerido (2.0)
dia_mes
= outubro16-31
!= outubro16-31
CFerido (2.0)
tipo_veiculo
= Motocicleta
!= Motocicleta
CFerido (1.0)
SVitima (27.0/1.0)