CENTRO UNIVERSITÁRIO DE ANÁPOLIS – UniEVANGÉLICA
BACHARELADO EM ENGENHARIA DE COMPUTAÇÃO

ANÁLISE DOS PROCESSOS DE DESCOBERTA DE CONHECIMENTOS COM AS FERRAMENTAS DE MINERAÇÃO DE DADOS: ORANGE E WEKA

KETLEN FERNANDES DA CUNHA DOS SANTOS
SIDNEY JÚNIO MOURA PEREIRA

ANÁPOLIS
2020

KETLEN FERNANDES DA CUNHA DOS SANTOS
SIDNEY JÚNIO MOURA PEREIRA

ANÁLISE DOS PROCESSOS DE DESCOBERTA DE CONHECIMENTOS COM AS FERRAMENTAS DE MINERAÇÃO DE DADOS: ORANGE E WEKA

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.

Orientadora: Profª. Ma. Aline Dayany Lemos

Anápolis
2020

Às nossas famílias e amigos que os quais sem
o seu apoio nada seria possível.

Agradecimentos

Dizer que o caminho até aqui foi de pedras, montanhas e águas profundas seria um eufemismo. Os últimos anos foram compostos de dúvidas, insegurança, ansiedade, choro e mais dúvidas, mas também foram de alegria, empolgação, companheirismo, comemoração e agora de alívio pelo dever cumprido.

Nosso muito obrigado às nossas famílias pelo apoio desde o primeiro momento, da escolha do curso, ao ingresso na faculdade, as presenças sutis nas madrugadas em claro, nas notas boas e ruins. Aos nossos parceiros de amor e de vida que nos escutaram e escutam nos momentos de incerteza sobre o futuro antes de dormir.

Às amizades verdadeiras que fizemos durante o curso e que levaremos para vida o nosso muito obrigado, sem vocês a história seria bem menos empolgante, as batalhas seriam mais difíceis e as comemorações seriam apenas detalhes. Durante os 5 anos que estivemos juntos partilhamos não só o dia-a-dia de estudos, mas também nossas vidas pessoais, apoiamos e fomos apoiados, estivemos presentes nos bons e nos maus momentos um do outro e isto nos fez vencedores desta importante etapa de nossas vidas, então muito obrigado à turma que ficava presente até o final das aulas com os professores (nem sempre, mas na maioria) e até depois só para rir mais um pouco ou reclamar junto: Cairo Mateus, Éber Lucas e Rodrigo Martins!

À academia como um todo, desde as secretárias que sempre nos ajudaram aos professores, em especial à nossa querida Luciana Nish por ser um poço de paciência e disponibilidade para nos atender e nos dar sempre o seu olhar de "por favor" sempre que escutava um "Nish, pode entregar amanhã?" e nossa orientadora e parceira desde a disciplina de SGBD, a quem admiramos imensamente, quem muitas vezes procuramos em busca de conselhos, quem se mostrou sempre prestativa até mesmo para ajudar em assuntos fora da grade curricular, àquela que acompanhou as lutas que travamos desde nossas primeiras tentativas de finalizar o curso (quando ainda não firmamos essa dupla). Muito obrigado Aline Dayany Lemos por não desistir de nós, que tenhamos te deixado orgulhosa!

No final nós somos um conjunto de partes de cada pessoa importante nas nossas vidas e se ao final desta trajetória conseguirmos honrá-los esta será a nossa maior vitória. Obrigado!

“No difícil está a Verdade.”
Provérbio Romano

Resumo

Em uma pesquisa científica é necessário a utilização de parâmetros científicos em sua elaboração, e isso não difere em pesquisas que envolvem o processo de descoberta de conhecimento em banco de dados. Entretanto, ao analisar os critérios de escolha das ferramentas de mineração de dados realizadas por pesquisadores em suas pesquisas, nota-se que não são baseadas em critérios científicos, e sim pessoais, fazendo que o trabalho não possa ser replicado com o mesmo rigor. Portanto, o presente trabalho visa avaliar o processo de descoberta de conhecimento com o auxílio das ferramentas de mineração de dados Orange e WEKA e com isto proporcionar a futuros estudos parâmetros científicos para sustentar a escolha de utilização entre as duas ferramentas. Para isto será aplicado o processo de descoberta de conhecimento com cada uma das ferramentas e analisados seus comportamentos em cada etapa deste projeto.

Palavras-chave: Mineração de Dados, Descoberta de Conhecimento, Base de Dados, KDD, *Cluster*, *K-means*, Orange, WEKA

Abstract

In scientific research, it is necessary to use scientific parameters in its elaboration, and this does not differ in research involving the process of discovering knowledge in a database. However, when analyzing the criteria for choosing data mining tools carried out by researchers in their research, it is noted that they are not based on scientific criteria, but rather personal, making the work cannot be replicated with the same rigor. Therefore, the present work aims to evaluate the knowledge discovery process with the aid of data mining tools Orange and WEKA and thereby provide future studies with scientific parameters to support the choice of use between the two tools. For this, the knowledge discovery process will be applied with each of the tools and their behavior will be analyzed in each stage of this project.

Keywords: *Data Mining, Knowledge Discovery, Database, KDD, Cluster, K-means, Orange, WEKA*

Lista de Figuras

Lista de Quadros

Lista de Abreviaturas e Siglas

ATS	*Air Traffic Service*
CENIPA	Centro de Investigação e Prevenção de Acidentes Aeronáuticos
CPU	*Central Process Unit*
ETL	*Extract, Transform and Load*
GNU	*General Public License*
GUI	*Graphical User Interface*
IAA	Inquérito de Acidente Aeronáutico
ICAO	*International Civil Aviation Organization*
IPM	Inquérito Policial Militar
ITA	Instituto de Tecnologia da Aeronáutica
KDD	*Knowledge Discovery in Databases*
KNIME	*Konstanz Information Miner*
MDS	Gráfico de Dispersão Multidimensional
NASA	*National Aeronautics and Space Administration*
ONU	Organização das Nações Unidas
PUC-GO	Pontifícia Universidade Católica de Goiás
SIPAER	Serviço de Investigação e Prevenção de Acidentes Aeronáuticos
UFMG	Universidade Federal de Minas Gerais
UFPE	Universidade Federal de Pernambuco
UFRJ	Universidade Federal do Rio de Janeiro
UnB	Universidade de Brasília
Unespe	Universidade Estadual Paulista
USP	Universidade de São Paulo
XML	*eXtensible Markup Language*
WEKA	*Waikato Environment for Knowledge Analysis*

Sumário

1 INTRODUÇÃO

Por intermédio da internet, a humanidade vem produzindo cada vez mais dados nas mais diversas plataformas digitais através de uma interação contínua de dispositivos interconectados (sensores, computadores, celulares, câmeras, entre outros) e aplicativos com a *web*. O resultado dessa interação é uma quantidade de registros, sinais, imagens, vídeos e *posts* que vêm sendo coletados e armazenados, e que no final somam cerca de 90% de todos dados produzidos desde 2017. Por consequência, os dados gerados são abundantes, produzidos de forma rápida e servem de base para geração de estratégia e tomada de decisão na esfera pública e privada (BUGNION; MANIVANNAN; NICOLAS, 2017; MAGRANI, 2019; VAN DER AALST, 2014 apud RAUTENBERG; CARMO, 2019).

Esta massa de dados tem atingido enormes proporções, o que tornou um problema para o ser humano que apresenta a incapacidade natural de assimilar tanta informação. Por consequência, essa análise de dados ficou lenta, cara e subjetiva e com o volume de dados crescendo essa análise manual se torna cada vez mais impraticável. Neste cenário, o desenvolvimento de soluções computacionais e estatísticas tornou necessário para que se possa realizar a descoberta de conhecimento em bases de dados onde as massas de dados são armazenadas (AMARAL, 2001; CAMPOS NETO, 2016; FAYYAD, et. al, 1996).

Para analisar tanta informação foram desenvolvidas ferramentas de mineração de dados que facilitam o manuseio, processamento e a análise de informações. Algumas ferramentas de mineração de dados como Orange, RapidMiner, Tanangra, WEKA (*Waikato Environment for Knowledge Analysis*), dentre outras, que consistem em buscar padrões e anomalias em bases de dados sendo compostas de métodos tais como descoberta de regras de associação, árvores de decisão, raciocínio baseado em casos, algoritmos genéticos, redes neurais artificiais, etc; e de tarefas como associação, descrição, estimação, predição, agrupamento, associação, dentre outras (CAMILO; SILVA, 2009; CAMPOS NETO, 2016; DIAS, 2002).

Portanto, dentre diversos métodos e tarefas de mineração, a escolha das ferramentas de mineração de dados muitas vezes transcende os critérios científicos e acaba sendo realizada através de escolhas pessoais. Entretanto a escolha da ferramenta de mineração de dados pode ser uma tarefa difícil podendo estar relacionada a um dos fatores como falta de conhecimento dos métodos existentes de mineração de dados; implementação e aplicação complexas de um método de mineração de dados; falta de ferramentas adequadas; custo elevado das ferramentas de mineração de dados disponíveis no mercado; falta de parâmetros de referência no momento

da escolha do método e da ferramenta mais adequada de acordo com o problema a ser resolvido (DIAS, 2002).

Ao analisar as justificativas de escolha de ferramentas de mineração de dados presentes em 24 trabalhos científicos incluindo Trabalhos de Conclusão de Curso, Dissertações e Teses que possuem o tema associado a mineração de dados foram constatados que as justificativas foram pautadas em cima de escolhas pessoais como: ferramenta gratuita, facilidade de uso, ferramenta utilizada em projetos semelhantes, conhecimento dos algoritmos da ferramenta e alguns casos não justificados.

Os trabalhos foram analisados pelos autores deste trabalho e pertenciam às Instituições de Ensino Superior como ITA (Instituto de Tecnologia da Aeronáutica), PUC-GO (Pontifícia Universidade Católica de Goiás), UFMG (Universidade Federal de Minas Gerais), UFPE (Universidade Federal de Pernambuco), UFRJ (Universidade Federal do Rio de Janeiro), UnB (Universidade de Brasília), Unespe (Universidade Estadual Paulista), USP (Universidade de São Paulo) e o Centro Universitário de Anápolis - UNIEVANGÉLICA.

Neste sentido, o presente trabalho aborda uma análise de todas as etapas do processo de KDD utilizando a base de dados de ocorrências aeronáuticas do CENIPA com o auxílio das ferramentas de mineração de dados Orange e WEKA. Através da Técnica de Clusterização e utilização do algoritmo *K-means* foi gerado um paralelo de uso de cada ferramenta, e ainda com o contraste da Representação do Conhecimento obtido após a utilização de cada ferramenta que foi possível gerar parâmetros científicos para justificar a escolha de utilização das ferramentas que possam ser utilizadas e replicadas futuramente.

Foram apresentados ao longo do trabalho conceitos teóricos que direcionaram o desenvolvimento deste estudo. Eles estão divididos em duas partes principais: aspectos relacionados ao Processo de Descoberta de Conhecimento em Bancos de Dados, e uma contextualização ao ambiente de pesquisa que foi escolhido, ocorrências aeronáuticas. Na Seção seguinte são demonstrados os resultados alcançados. Há ainda considerações dos autores sobre o que foi alcançado e aprendido, em seguida com sugestão de trabalhos futuros. Por fim, a Seção de Anexos e Apêndices adicionam informações que contribuem para o detalhamento dos resultados obtidos.

2 FUNDAMENTAÇÃO TEÓRICA

Para atingir a compreensão da pesquisa, é essencial o esclarecimento dos conceitos abordados durante o desenvolvimento deste trabalho, o que contribuiu tanto para a definição do tema abordado quanto para o alinhamento dos resultados obtidos.

A fundamentação teórica abrange o ponto de vista da literatura e pesquisadores sobre: processos de identificação de padrões em base de dados e detalhamento da etapa de mineração de dados e suas especificidades.

2.1 Dado, informação e conhecimento

Desde o surgimento dos primeiros computadores o objetivo das organizações eram armazenar e processar dados. Isto ficou mais evidente com a queda dos custos de aquisição de *hardware*, tornando possível armazenar quantidades massivas de dados. Alguns exemplos deste fato é uma empresa que cria os cadastros com os dados pessoais de seus clientes para conhecê-los melhor, e Instituições de ensino que armazenam notas, pagamentos de mensalidades e informações pessoais dos alunos. Com essas necessidades foram criadas novas tecnologias de armazenamento como banco de dados, *Data Warehouse*, Bibliotecas Virtuais, Servidores *Web* e dentre outras (CIOS et al., 2007; HAN; KAMBER; PEI, 2012).

Como exemplo de geração das massas de dados pode-se citar aplicações como os satélites de observação da NASA (*National Aeronautics and Space Administration*) que geram cerca de um *terabyte* de dados por dia; o Projeto Genoma armazena milhares de *bytes* para cada uma das bilhões de bases genéticas; Instituições que mantém repositórios com milhares de transações dos seus clientes; uso intenso de dispositivos interconectados (sensores, computadores, câmeras, celulares, *tablets*) comunicando com a internet que coletam muitos dados e de forma indireta, e dentre outras aplicações (BRAMER, 2007; RAUTENBERG; CARMO, 2019).

É importante ressaltar que o dado é a menor unidade obtida em uma observação do mundo com significados implícitos, que podem ser armazenados em uma base de dados. Informação é o uso dos dados, dentro de um contexto, que por meio de interpretações adquire significado e auxilia na tomada de decisões. Já o conhecimento, é a informação sendo aplicada com um propósito ou utilidade através de atividades humanas (BITTENCOURT, 2004; ELMASRI; NAVATHE, 2011; SOUSA, 2017).

Na figura 1, será apresentada uma metáfora sobre dado, informação, apresentação e conhecimento, num contexto mais lúdico, para que possa ser compreensível os conceitos dos termos a serem utilizados.

Figura 1. Metáfora sobre dado, informação, apresentação e conhecimento

Fonte: Johnstone (2019)

Nessa metáfora os dados são os ingredientes utilizados no bolo, as menores unidades de um todo; a informação é a combinação de todos os ingredientes para formação do bolo; a apresentação é a forma como a informação é apresentada; o bolo, no momento que é consumido representa o conhecimento.

2.2 Processo de Descoberta de Conhecimento em Bancos de Dados

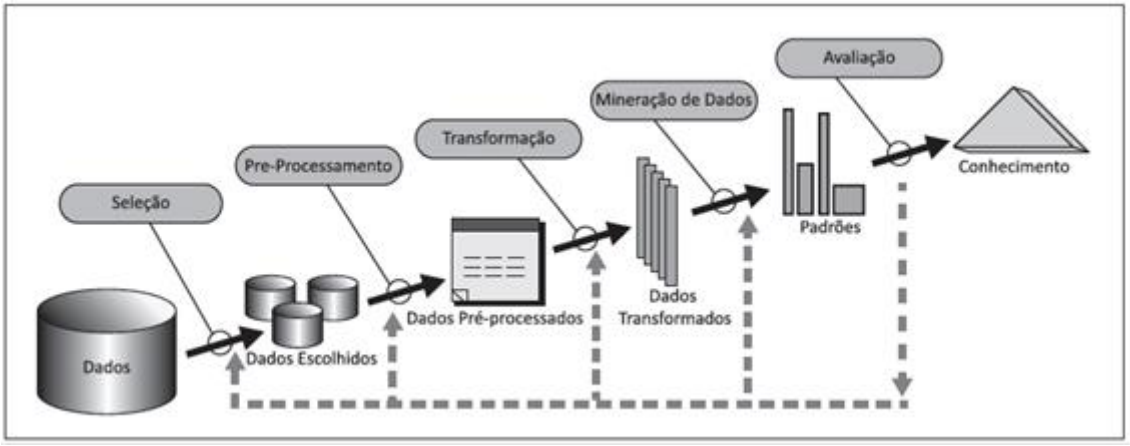
O estudo de dados com a intenção de transformá-los em conhecimento através de seus padrões é um método tradicional visto nas mais diversas áreas como Medicina, Agronomia, Geografia, Administração, dentre outras. Na área de Administração, por exemplo, o profissional realiza diversas análises a partir dos números de vendas e cancelamentos, admissão e evasão de funcionários, e receita e prejuízo anual da empresa. As análises geralmente são feitas tanto de forma manual quanto interpretativa e seus resultados auxiliam o profissional em suas projeções e tomadas de decisões. Ainda, à medida que o volume de dados da empresa aumenta este processo se torna complexo e custoso (AMARAL, 2001).

Para automatizar este processo de descoberta de conhecimento podem ser utilizadas ferramentas auxiliares e métodos mais modernos, como o *Knowledge Discovery in Database* (KDD - Descoberta de Conhecimento de Base de Dados). De acordo com Fayyad et. al (1996, p. 40, tradução nossa): “KDD é o processo não trivial de identificação de padrões válidos, novos, potencialmente úteis e finalmente compreensíveis nos dados.”

O processo de KDD, ilustrado na Figura 2, é composto por várias etapas que envolvem alguns pontos principais, sendo eles: Seleção, Pré-processamento, Transformação, Mineração

de Dados, Pós-Processamento e Aplicação do Conhecimento. Este processo é interativo e iterativo, onde o analista pode retornar à etapa anterior para aprimorar os dados coletados ou reavaliar os resultados obtidos (BERNABEU, 2004).

Figura 2. Etapas do Processo de KDD



Fonte: Camilo e Silva (2009, p. 3)

2.2.1 Etapas do KDD

O processo de KDD envolve duas fases, a preparação e a mineração dos dados. A primeira parte envolve as etapas de seleção, pré-processamento e transformação; já a segunda parte envolve as etapas mineração de dados, pós-processamento e utilização do conhecimento. A seguir serão descritas cada etapa mencionada que está envolvida no processo de KDD (AMARAL, 2001).

Na etapa de Seleção, o analista precisa dedicar grande esforço e cuidado para determinar aquilo que se deseja extrair de conhecimento e a base de dados a ser utilizada para que seja possível obter o resultado esperado ao final do projeto. Tendo em vista a escolha do problema a ser abordado, é crucial a correta definição do conjunto de dados-alvo, a qual o analista terá como foco determinados atributos ou instâncias de dados, pois esta poderá ser determinante para o sucesso do projeto (AMARAL, 2001; BERNABEU, 2004).

A seguir, a etapa de Pré-processamento possui algumas subetapas onde todas agem com um único objetivo de garantir a qualidade dos dados a serem utilizados, reduzindo possíveis atrasos e complexidades que não possuem utilidade para o processo de mineração. Para isto, inicialmente são tratadas formas de limpeza dos dados a fim de eliminar quaisquer inconsistências, ruídos ou dados incompletos que possam comprometer a mineração dos dados (AMARAL, 2001).

Além da limpeza dos dados, no Pré-processamento, ocorre a Seleção dos Dados, onde Bernabeu (2004) aponta como finalidade diminuir o tempo de processamento do algoritmo, reduzindo seu espaço de busca considerando a qualidade dos registros a serem analisados.

Em seguida, inicia-se o processo de Extração de Padrões, responsável por garantir o cumprimento dos objetivos definidos na etapa de Seleção. Conforme Rezende (2003, p. 317), "nessa etapa é realizada a escolha, a configuração e execução de um ou mais algoritmos para extração de conhecimento". Sendo este um processo iterativo, é comum ser executado diversas vezes, conforme necessário.

São comuns nos processos de KDD a utilização de mais de uma fonte de dados. Sendo assim, para que seja possível aplicar o processo de mineração, os dados são acoplados e transformados com o intuito de que apresentem o mesmo formato e estrutura. De acordo com Fayyad et al. (1996), após concluir as etapas anteriores os dados necessitam de uma transformação para a aplicação dos algoritmos (que trabalham apenas com um tipo de valor).

Um exemplo da etapa de Transformação, seria da formatação de duas bases diferentes, onde uma delas apresenta apenas dados em formatos numéricos e uma segunda base onde há valores categóricos. Neste caso é necessário transformar os valores numéricos em categóricos ou vice-versa. Para a transformação dos dados são usados diferentes critérios conforme o objetivo do projeto, existindo apenas um modelo padrão a ser seguido (CAMILO; SILVA, 2009).

Enfim, a etapa de Mineração de Dados consiste na aplicação de algoritmos de análise e descoberta para abstrair conhecimento a partir dos dados. Considerada como a etapa principal do KDD, é comum ser referenciada como um sinônimo para todo o processo. Um problema a ser enfrentado nesta etapa é a capacidade de processamento necessário para realizar as buscas em grandes bases de dados (FAYYAD et al., 1996; REIS, 2014).

Com a conclusão da etapa de mineração, ingressa-se na etapa de Pós-Processamento, destinada à avaliação e estudo dos resultados obtidos na mineração de dados. Para total aproveitamento do conhecimento extraído, é necessário o auxílio de um profissional especialista no assunto para realizar as interpretações e validar as informações como coerentes. Para complementar ainda mais os resultados, o analista pode utilizar algoritmos de análise de padrões para buscarem conhecimentos específicos. O resultado final é compilado e documentado em forma de relatórios em uma linguagem entendível aos usuários (BERNABEU, 2004).

2.2.2 Tarefas do KDD

Para conceituar o tratamento sobre um conjunto de dados, autores utilizam a nomenclatura de Tarefas do KDD que são aplicadas durante a etapa de Mineração de Dados, a

seguir será abordado as características sobre cada uma das tarefas principais do KDD (BERNABEU, 2004).

As tarefas do KDD são subdivididas em duas categorias: Preditivas e Descritivas. Enquanto as tarefas da primeira categoria visam a abstração do conhecimento através de um conjunto de dados históricos para realizar a predição de novas amostras, as tarefas da segunda categoria buscam identificar padrões de comportamento que sintetizam as relações entre os dados. Dentre as tarefas Preditivas estão as de Descrição, Classificação e Regressão, já as de Associação, *Clustering* e Sumarização pertencem à categoria de tarefas Descritivas (CAMPOS NETO, 2016).

A primeira tarefa é chamada de Descrição, é utilizada para descrever os padrões revelados pelos dados, oferecendo possíveis interpretações sobre os resultados obtidos considerando possíveis influências de determinadas variáveis no produto final. (CAMILO; SILVA, 2009).

A tarefa de Classificação tem como objetivo identificar a qual classe pertence um determinado registro. Para este caso, o modelo construído analisa um conjunto de registros previamente classificados para que o modelo seja capaz de avaliar a classificação indicada e realizar por si o mesmo processo para os demais registros. Ou seja, para cada registro é indicado a qual classificação pertence. O modelo criado analisa esses registros classificados e desta forma é capaz de predizer a qual categoria um novo registro se encaixa (CAMILO; SILVA, 2009; CAMPOS NETO, 2016).

Por sua vez, a tarefa de Regressão (Estimativa), ao contrário da tarefa de Classificação, é utilizada quando é identificado um registro de valor numérico real e não um categórico. Desta forma, é possível estimar o valor de uma variável a partir dos demais valores (LAROSE, 2005).

A Associação visa identificar o relacionamento entre os atributos, descrevendo associações entre as variáveis. Segundo Larose (2005, p. 17, tradução nossa) “As regras de associação têm a forma: SE atributo X ENTÃO atributo Y, juntamente com uma medida do apoio e confiança associados à regra”.

A Clusterização (Segmentação) é a tarefa de agrupamento que visa aproximar os registros similares. Os algoritmos de clusterização têm como objetivo segmentar o conjunto de dados em subgrupos em que a similaridade dos registros dentro do *cluster* seja maximizada em relação à similaridade com registros fora dele (LAROSE, 2005).

A Sumarização consiste em facilitar o entendimento dos dados identificando inúmeras características nos dados sendo estudados, ou seja, suas similaridades nos registros da base de dados. Uma das principais abordagens para descrição de informações é a visualização dos

dados, principalmente quando os dados não estão organizados em uma forma padrão. Os resultados obtidos nessa técnica são utilizados em conjunto com outras funcionalidades (CORTÊS, PORCARO; LIFSCHITZ, 2002).

2.3 Ferramentas de Mineração de Dados

Segundo Rangra e Bansal (2014), ao longo dos anos um vasto número de ferramentas de mineração de dados foi desenvolvido por uma comunidade de pesquisadores e entusiastas de análise de dados. Desta união foram originadas diversas ferramentas de código aberto como RapidMiner, Tanagra, WEKA e Orange, sendo estas duas últimas as ferramentas de foco deste projeto. Conforme Wahbeh et al. (2011, p. 18, tradução nossa) "Essas ferramentas e *softwares* fornecem um conjunto de métodos e algoritmos que ajudam na melhor utilização dos dados e informações disponíveis para os usuários". Abaixo serão descritas algumas características das ferramentas de mineração de dados mencionadas anteriormente.

2.3.1 RapidMiner

A ferramenta RapidMiner é um *software* de código aberto desenvolvido em linguagem Java que oferece um ambiente integrado para procedimentos de mineração de dados e aprendizado de máquinas, incluindo: carga e transformação de dados (ETL), pré-processamento dos dados, modelagem, avaliação e implantação. Para os processos de mineração de dados, estes podem ser descritos em XML e criados em interface gráfica do usuário (GUI). Além disso, o RapidMiner pode ser usado para processos de mineração de texto, mineração de multimídia, análise preditiva e análise de negócios (RAMAMOHAN et al., 2012; RANGRA, BANSAL, 2014).

2.3.2 Tanagra

O Tanagra é um *software* gratuito de mineração de dados desenvolvido para fins acadêmicos e de pesquisa. Ele possui vários métodos de mineração de dados como análise exploratória de dados, aprendizado estatístico, aprendizado de máquina e banco de dados. Ele é um projeto de *software* livre (projeto de código aberto), pois todo usuário pode acessar o seu código fonte e adicionar os seus próprios algoritmos, desde que concorde e esteja em conformidade com a licença de distribuição do *software* (PATEL; DESAI, 2015; RAMAMOHAN et al., 2012; WAHBEH et al., 2011).

A ferramenta possui como foco três objetivos, o primeiro objetivo considerado como principal é fornecer uma plataforma para pesquisadores e estudantes usarem de maneira fácil

de utilizar, em conformidade com as normas atuais do desenvolvimento de *software* e permitindo analisar dados reais ou sintéticos. O segundo objetivo consiste é propor aos pesquisadores uma arquitetura que lhes permita adicionar seus próprios métodos de mineração de dados, para comparar seus desempenhos. O terceiro objetivo é que os desenvolvedores iniciantes desfrutem do acesso gratuito ao código-fonte, para compreender a sua arquitetura de desenvolvimento, os problemas a serem evitados, as principais etapas do projeto e quais ferramentas e bibliotecas de código utilizar (PATEL; DESAI, 2015; RAMAMOHAN et al., 2012; WAHBEH et al., 2011).

2.3.3 WEKA

O WEKA (*Waikato Environment for Knowledge Analysis*) é um *software* de código livre desenvolvido na Nova Zelândia, pela Universidade de Waikato, seu nome faz referência a um pássaro que não voa, de natureza curiosa, encontrado apenas nas ilhas da Nova Zelândia. Com ele, pesquisadores, cientistas industriais, estudantes e professores de Universidades, são capazes de utilizar o *Machine Learning* (Aprendizado de Máquina) para obter conhecimento de grandes bases de dados, eliminando a necessidade de uma análise totalmente manual (WITTEN; FRANK, 2005; WAIKATO, 2019).

O WEKA é uma coleção de algoritmos de aprendizado de máquina e ferramentas de preparação, classificação, regressão, *clustering*, mineração de regras de associação e visualização de dados. Ele foi projetado para que seja possível testar rapidamente os métodos existentes em novos conjuntos de dados de maneira flexível (WITTEN; FRANK, 2005; WAIKATO, 2019).

Conforme Witten e Frank (2005), o WEKA fornece amplo suporte para todo o processo de mineração de dados experimental, incluindo a preparação dos dados de entrada, a avaliação estatística dos esquemas de aprendizado e a visualização dos dados de entrada e do resultado do aprendizado. Seu conjunto de ferramentas diversificado e abrangente é acessado por meio de uma interface comum, para que seus usuários possam comparar diferentes métodos e identificar os mais adequados para o problema analisado.

O sistema foi desenvolvido em Java e distribuído de forma livre, conforme os termos da GNU, *General Public License*. Possui compatibilidade com plataformas *Linux*, *Windows e MacOS*, o que ajuda em sua disseminação (WITTEN; FRANK, 2005).

2.3.4 Orange

O Orange, assim como o WEKA, é um *software* de código aberto para aprendizado de máquina e mineração de dados. Mesmo destinando-se a usuários experientes e pesquisadores, a ferramenta possui uma comunidade ativa que auxilia novatos e especialistas em suas análises, além de fornecer cursos (pagos) e vídeos rápidos e gratuitos no *Youtube* com legendas disponíveis em português (PATEL, DESAI, 2015; ORANGE, 2019).

O desenvolvimento do Orange começou em 1997 por Janez Demšar e Zupan, até então membros do Laboratório de Inteligência Artificial na Universidade de Ljubljana, na Eslovênia. Como informado pelos próprios desenvolvedores no artigo *Orange: Data Mining Fruitful and Fun* de 2012, o *software* foi desenvolvido inicialmente como uma biblioteca C ++ de algoritmos de aprendizado de máquina.

Segundo Demšar e Zupan (2012), o Orange era utilizado principalmente para exploração de dados em conjunto com diferentes algoritmos de pré-processamento e aprendizado para testes e combinação de validações. Os componentes eram incorporados a programas que poderiam ser usados via linha de comando. Com isso se mostrando limitante, o Orange começou a fornecer *scripts* em *Python*. Com *scripts Python*, o usuário é capaz de prototipar novos algoritmos enquanto reutiliza o máximo de código possível, além de possuir interface de fácil utilização.

O Orange possui suporte às tarefas que vão desde o pré-processamento de dados até modelagem e avaliação. O *software* pode ser utilizado em sistemas operacionais *Linux, Windows* e *MacOS* (DEMŠAR et. al., 2004).

2.4 Base de Dados

No início do processo do KDD é necessário determinar aquilo que se deseja extrair de conhecimento e a base de dados a ser utilizada para que seja possível obter o resultado esperado ao final do projeto. Tendo em vista a escolha do problema a ser abordado, é fundamental a correta definição do conjunto de dados-alvo, a qual o analista terá como foco determinados atributos ou instâncias de dados, pois esta poderá ser determinante para o sucesso do projeto (AMARAL, 2001; BERNABEU, 2004).

Tendo em vista esta etapa, foi definida uma base genérica e a escolhida foi a base de dados de ocorrências aeronáuticas em solo brasileiro, disponibilizada pelo Portal Brasileiro de Dados Abertos, este é uma ferramenta disponibilizada pelo Governo para que todos possam encontrar e utilizar dados e informações públicas, onde preza pela simplicidade e organização

para que possa encontrar facilmente os dados e informações que precisa (PORTAL BRASILEIRO DE DADOS ABERTOS, 2019).

Na subseção a seguir é abordado o contexto em que a base de dados é utilizada, o que ajuda a compreender os dados registrados contidos nela, além de interpretação da informação e auxílio nas próximas etapas de descoberta de conhecimento.

2.4.1 Ocorrências Aeronáuticas

A história dos acidentes aeronáuticos e da aviação tiveram inícios simultâneos, em virtude da falta de conhecimentos para o desenvolvimento das aeronaves com condições mínimas de segurança. Dessa forma, as tentativas de colocar uma máquina mais pesada que o ar se seguiram e com isso os primeiros acidentes aeronáuticos (CENIPA, 2000 apud SOUZA, 2012).

O primeiro acidente aeronáutico brasileiro foi de um balão pilotado pelo Tenente Juventino, do Exército Brasileiro, em 20 de maio de 1908. A comissão de investigação desta ocorrência concluiu que houve um rompimento do cabo que limitava a subida do balão. Na tentativa de subir o balão, o Tenente Juventino acionou a válvula de gás, como ação corretiva, cujo mal funcionamento acabou ocorrendo uma queda violenta contra o solo, que por consequência ocasionou a sua morte (CENIPA, 2000 apud SOUZA, 2012).

As definições das atividades de investigação de acidentes aeronáuticos no Brasil tiveram início na década de 20 nas Forças Armadas da Marinha e do Exército. A Marinha realizava as suas investigações através do Inquérito Policial Militar (IPM), já o Exército utilizava o Inquérito de Acidente Aeronáutico (IAA) a fim de apurar as suas ocorrências. Em ambos os casos, os motivos das apurações eram compreender as causas responsáveis pelo acidente (BRASIL, 2000 apud CAMARGO 2010).

Após a criação do Ministério da Aeronáutica, em 1941, ocorreu uma reformulação dos procedimentos que já vinham sendo executados pela Marinha e do Exército, os procedimentos foram unificados e ficaram sob a jurisdição da então Inspetoria Geral da Aeronáutica. Dessa forma foi criado o Inquérito Técnico Sumário como substituto do IPM e do IAA na apuração de acidentes aeronáuticos com o objetivo de encontrar as causas que provocaram o acidente aeronáutico (BRASIL, 2000 apud CAMARGO 2010).

Em 1951, através do novo Regulamento da Inspetoria Geral da Aeronáutica é criada a sigla SIPAER (Serviço de Investigação e Prevenção de Acidentes Aeronáuticos) que manteve o objetivo de encontrar as causas responsáveis pelos acidentes aeronáuticos. A partir de 1965, teve a reestruturação do SIPAER, deixando de ser Serviço e passando a se tornar um Sistema,

dessa forma, a palavra "inquérito" foi substituída definitivamente por "investigação". As investigações conduzidas pelo SIPAER passaram a ter uma única finalidade, a prevenção de acidentes aeronáuticos (CAMARGO, 2010; SOUZA, 2012).

A Inspetoria Geral de Aeronáutica, responsável pelo SIPAER, utiliza conceitos de ocorrências aeronáuticas por meio das definições que foram estabelecidas no Anexo nº 13 da Convenção de Chicago datado em 11 de abril de 1951. A responsável pela elaboração e edição do Anexo nº 13 da Convenção de Chicago é a ICAO (*International Civil Aviation Organization*). A ICAO é uma agência especializada da ONU (Organização das Nações Unidas), criada por Estados em 1944 para gerenciar a administração e o governo da Convenção sobre a Aviação Civil Internacional (ICAO, 2001; ICAO, 2019).

A agência trabalha com os 193 Estados Membros e grupos industriais da Convenção para alcançar um consenso sobre os Padrões e Práticas Recomendadas da aviação civil internacional e políticas de apoio a um setor de aviação civil seguro, eficiente, protegido, economicamente sustentável e ambientalmente responsável (ICAO, 2019).

A seguir serão conceituados, conforme Anexo nº 13 da Convenção de Chicago, termos utilizados quando se trata o assunto de ocorrências aeronáuticas (ICAO, 2001):

Inicialmente a aeronave é definida como qualquer máquina que possa obter apoio na atmosfera a partir de reações do ar que não sejam as reações do ar contra a superfície da terra. Quando se trata de ocorrências aeronáuticas, ela é dividida entre acidentes e incidentes O acidente compreende-se como uma ocorrência associada à operação de uma aeronave que ocorre entre o momento em que alguém embarca na aeronave com a intenção de voar até o momento em que todas essas pessoas desembarcarem; o incidente é uma ocorrência aeronáutica associada à operação de uma aeronave que afeta ou pode afetar a segurança da operação.

Entretanto as ocorrências aeronáuticas não acontecem sem um fator contribuinte. Ele é definido como uma ação, omissão, evento, condição ou a combinação destes que se fossem eliminados, evitados ou ausentes poderia reduzir a probabilidade de uma ocorrência aeronáutica ou até mesmo reduzido a severidade das consequências da ocorrência aeronáutica (BRASIL, 2017).

As investigações das ocorrências aeronáuticas são realizadas pelo CENIPA (Centro de Investigação e Prevenção de Acidentes Aeronáuticos). O CENIPA é um órgão do Comando da Aeronáutica responsável pelas atividades de investigação de ocorrências aeronáuticas da aviação civil e da Força Aérea Brasileira. As investigações são realizadas de acordo com Anexo 13 à Convenção Internacional de Aviação Civil da ICAO. Por fim, através das investigações realizadas pelo CENIPA (Centro de Investigação e Prevenção de Acidentes Aeronáuticos) tem-

se a descoberta dos fatores contribuintes das ocorrências aeronáuticas que se tornam dados estatísticos de bases de dados disponíveis na página *web* do CENIPA[1] (CENIPA, 2019).

Torna-se necessário extrair conhecimento dessas bases de dados a fim de verificar se os fatores contribuintes das ocorrências aeronáuticas uma vez já conhecidos continuam a provocar novas ocorrências aeronáuticas, aumentando dessa forma a consciência situacional do panorama da aviação brasileira.

[1] Disponível em: https://dados.gov.br/dataset/ocorrencias-aeronauticas-da-aviacao-civil-brasileira

3 RESULTADOS

Esta Seção visa demonstrar os resultados obtidos ao longo do projeto. Para isso ele foi subdividido etapas para ter uma compreensão cronológica dos resultados alcançados.

3.1 Trabalhos Correlatos

Foi realizada uma pesquisa com trabalhos correlatos da área científicas como trabalho de conclusão de curso, dissertações, teses e artigos, foram encontrados quatro principais trabalhos que abordam o tema análogo de "análise de ferramentas de mineração de dados" citados abaixo:

- Wahbeh et al. (2011) aborda a descrição básica de quatro ferramentas de mineração (WEKA, Orange, Tanagra e KNIME) e a comparação do desempenho das ferramentas realizando a tarefa de classificação.
- Ramamohan et al. (2012) realiza um estudo comparativo de seis ferramentas de mineração de dados (WEKA, RapidMiner, Tanangra, DBMiner, Witness Miner e Orange) abordando assuntos como seus principais objetivos e aplicações.
- Rangra e Bansal (2014) descreve um estudo comparativo contendo as especificações técnicas, os recursos e a especialização de seis ferramentas de mineração de dados (WEKA, KEEL, R, KNIME, RapidMiner e Orange) junto com suas aplicações.
- Patel e Desai (2015) apresenta um estudo comparativo de seis ferramentas de mineração de dados (WEKA, R, Orange, RapidMiner, Tanangra e KNIME) tratando da descrição geral de cada ferramenta e seus prós e contras.

Notou-se que os quatros trabalhos correlatos realizam descrições, comparações, especificações das ferramentas de mineração de dados, porém nenhum dos autores descrevem as ferramentas sendo utilizadas no processo de descoberta de conhecimento em banco de dados.

Ainda ao analisar as justificativas das escolhas das ferramentas de mineração de dados de 24 trabalhos científicos incluindo Trabalhos de Conclusão de Curso, Dissertações e Teses que possuía o tema associado a mineração de dados foram contatados que as justificativas foram pautadas em cima de escolhas pessoais sendo apresentadas no quadro 1.

Quadro 1. Justificativas das Escolhas das Ferramentas de Mineração

Instituição de Ensino	Ferramenta de mineração	Justificativas	Ano	Autor
ITA	WEKA	*Open Source* e menos complexa	2015	(DIAS, 2015)
ITA	WEKA	sem justificativa	2010	(WINTER, 2010)
ITA	WEKA	Escolha de um algoritmo fornecido pela ferramenta	2004	(BERNABEU, 2004)
ITA	WEKA	Interface Amigável e utilizada em muitos projetos	2010	(PARREIRA, 2010)
ITA	WEKA	Conhecimento de um método fornecido pela ferramenta	2008	(ALBUQUERQUE, 2008)
ITA	WEKA	Integração com o Java	2015	(JACINTO, 2015)
PUC-GO	WEKA	Conhecimento dos métodos que a ferramenta fornece	2017	(ROCHA, 2017)
UFMG	R	*Open Source* e facilidade de uso	2015	(BRANQUINHO, 2015)
UFMG	-	Utilizou o MATLAB	2008	(CASTANHEIRA, 2008)
UFMG	WEKA	Conhecimento dos métodos que a ferramenta fornece	2008	(MAIA, 2008)
UFPE	WEKA	*Open Source*, conhecimentos dos algoritmos que a ferramenta fornece e utilizada em muitos projetos	2018	(SANTANA JUNIOR, 2018)
UFPE	KNIME, WEKA e R	*Open Source*	2017	(SILVA FILHO, 2017)
UFRJ	-	Utilizou o MATLAB	2017	(ALMEIDA, 2017)
UNB	WEKA	Conhecimento dos algoritmos que a ferramenta fornece, boa usabilidade e experiência agradável	2016	(SILVA, 2016)
UNB	WEKA	Conhecimento dos algoritmos que a ferramenta fornece	2017	(ASSIS, 2017)
UNESP	WEKA	Experimentos com as ferramentas RapidMiner, Pentaho e WEKA	2018	(TAMAE, 2018)
USP	WEKA	Conhecimento dos algoritmos que a ferramenta fornece	2018	(OLIVEIRO, 2018)

USP	ORANGE e WEKA	*Open Source*, muito utilizadas no meio acadêmico e bem avaliadas	2016	(CAMPOS NETO, 2016)
USP	WEKA	Não justificado	2019	(MACEDO, 2019)
USP	DAMICORE	Não justificado	2018	(SOUSA, 2018)
USP	DAMICORE	Não justificado	2018	(RECHE, 2018)
USP	WEKA	Conhecimento dos algoritmos	2019	(CALÇADA, 2019)
UNIEVANGELICA	WEKA	Parametrização de funcionalidades	2018	(VIEIRA, 2018)
UNIEVANGELICA	WEKA	Facilidade de uso e conhecimentos dos algoritmos que a ferramenta fornece	2017	(SANTOS, 2017)

Fonte: SANTOS, PEREIRA (2020)

No quadro 2 são apresentados de forma sintética os parâmetros científicos utilizados nos 24 trabalhos científicos para justificar a escolha da ferramenta de mineração de dados que foi aplicada. Portanto, no presente trabalho foram utilizadas as ferramentas WEKA e Orange. A ferramenta WEKA foi escolhida por ter se destacado na quantidade de vezes em que foi utilizada nos trabalhos correlatos analisados, e a ferramenta Orange por ter sido bastante comentada nos mesmos trabalhos.

Quadro 2. Parâmetros Científicos Utilizados

Ferramenta	**Parâmetros Científicos utilizados**						
	Open Source	**Facilidade de uso**	**Flexibilidade**	**Muito Utilizada**	**Projeto semelhante**	**Sem justificativa**	**Estudo Comparativo**
WEKA	3	13	2	3	3	3	1
ORANGE	1	1		1	1		
Linguagem R	2	1					
KNIME	1						
MATLAB						2	
DAMICORE						2	

Fonte: SANTOS, PEREIRA (2020)

A partir de pesquisas relacionadas ao tema de "processo de descoberta de conhecimento em banco de dados" foi possível adquirir conhecimentos teóricos a respeito do processo do KDD e suas etapas, visão geral das ferramentas de mineração de dados processo de descoberta de conhecimento em banco de dados, etapas do KDD e das ferramentas de mineração de dados Orange, RapidMiner, Tanagra e WEKA.

Nas pesquisas relacionadas ao tema “ocorrências aeronáuticas” foram adquiridos conhecimentos sobre o histórico das ocorrências aeronáuticas, os principais conceitos que envolvem um acidente aeronáutico como acidente, incidente e fator contribuinte, e algumas das regulamentações e agências internacionais e nacionais que regulamentam as investigações de ocorrências aeronáuticas sendo respectivamente o Anexo nº 13 da Convenção de Chicago e NSCA 3-13, ICAO e o CENIPA.

3.2 Ambiente de Execução do Processo

Durante a execução deste trabalho, algumas ferramentas adicionais foram utilizadas, para limpeza e organização dos dados analisados bem como a geração de resultados visuais na etapa final. Este item descreve o ambiente computacional utilizado para a execução do processo de Descoberta de Conhecimento.

a) Configuração de *Hardware*

Desktop, com processador Intel(R) Core (RM) i3-3240, 3.40GHz. Memória RAM de 12,0 GB.

b) Configuração de *Software*

O quadro 3 contém e descreve as ferramentas adicionais e sua utilização

Quadro 3. Configuração de *Software*

Ferramenta	Objetivo
Windows 10 Pro 64 *bits*	● Sistema Operacional onde as demais ferramentas foram instaladas.
Planilhas Google	● Seleção dos dados; ● Derivação dos atributos; ● Conversão dos atributos; ● Transformação dos dados.
PgAdmin 4	● Armazenamento dos dados; ● Avaliação dos dados; ● Geração de comandos para utilização nas consultas e seleção dos dados nas ferramentas de mineração.
Bloco de Notas	● Visualização de dados da clusterização da ferramenta WEKA.

Fonte: SANTOS, PEREIRA (2020)

3.3 Aplicação do Processo de Descoberta de Conhecimento

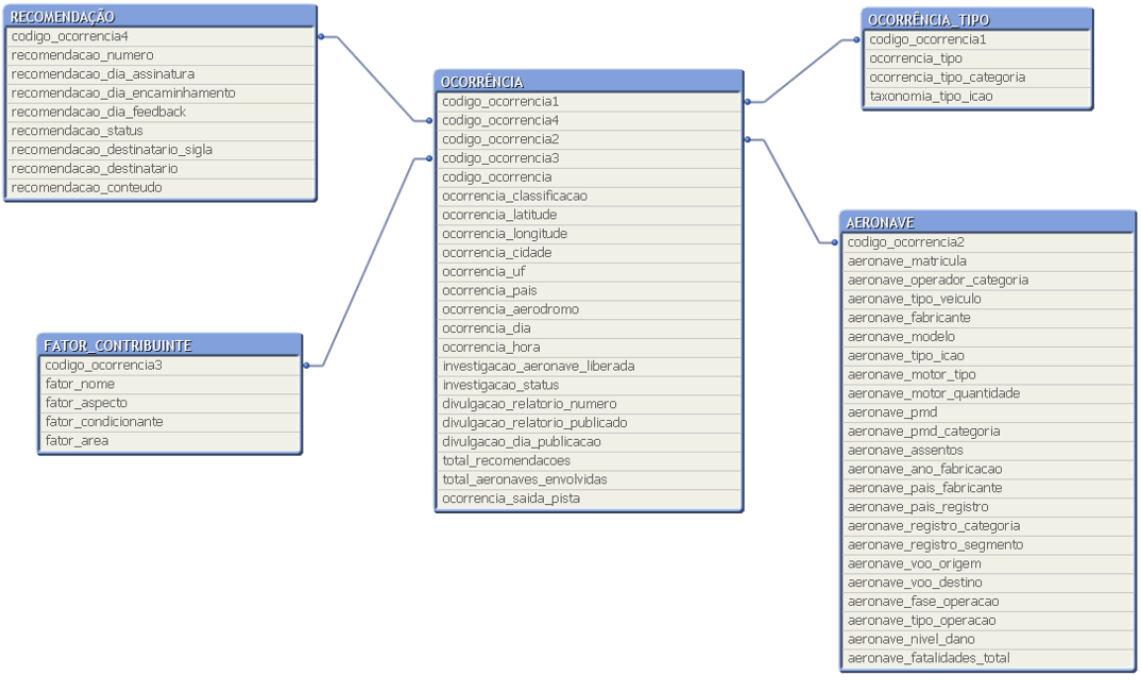
Nesta Seção são descritos os esforços realizados para a realização do processo de descoberta de conhecimento na base de dados.

3.3.1 Seleção

Para realizar a etapa de descoberta de conhecimento foi escolhida a base de dados de ocorrências aeronáuticas que é gerenciada pelo CENIPA. Nela constam registros de ocorrências aeronáuticas notificadas ao CENIPA entre janeiro de 2010 a abril de 2019 que ocorreram em solo brasileiro. Dentre as informações disponíveis estão os dados sobre as aeronaves envolvidas, fatalidades, local, data e horário dos eventos e informações taxonômicas típicas das investigações de acidentes.

A base de dados é composta por informações iniciais, proveniente de Ficha de Notificação de Ocorrência Aeronáutica denominado Formulário CENIPA-05, e consolidadas, provenientes dos relatórios de investigações publicados. Dados provenientes de ocorrências relacionadas a risco de fauna, emissões de raio laser e risco baloeiro não constam nesta base de dados, pois possuem formulários próprios para coleta de dados com foco no gerenciamento de risco.

Figura 3. Modelo de Dados - CENIPA



Fonte: CENIPA (2020)

Conforme Figura 3, a base de dados é composta por cinco tabelas ocorencia.csv (informações gerais sobre as ocorrências), ocorrencia_tipo.csv (informações sobre os tipos de ocorrências), aeronave.csv (informações sobre as aeronaves envolvidas nas ocorrências), fator_contribuinte.csv (informações sobre os fatores contribuintes das ocorrências que tiveram as investigações finalizadas) e recomendacoes.csv (informações sobre as recomendações de segurança geradas nas ocorrências).

3.3.2 Pré-processamento

Nesta Seção é descrita a etapa de pré-processamento dos dados, contendo as subetapas deste processo a fim realizar a análise e manipulação da base de dados, garantir a qualidade dos dados, reduzir o espaço de busca (sem que comprometa a integridade dos dados) e definir o algoritmo a ser utilizado na etapa de Mineração.

3.3.2.1 Limpeza e Seleção dos Dados

Conforme informado na Seção 3.3.1, inicialmente a base de dados utilizada estava disposta em cinco tabelas, disponibilizadas em arquivos .csv, desta forma, a intenção inicial foi de unir essas tabelas de modo que fosse possível utilizar o mesmo formato em ambas as ferramentas.

A primeira tentativa foi incluir os arquivos em um banco de dados e para que fosse possível relacionar as tabelas e assim extrair as informações. Porém, durante a importação dos dados foi notado a duplicidade dos códigos de chave primária das tabelas (com exceção da tabela ocorrencia), conforme mostrado na figura 4 com exemplo da tabela fator_contribuinte onde há casos de 15 fatores diferentes para 1 registro de ocorrência.

Figura 4. Tabela fator_contribuinte

	A	B	C	D	E
1	codigo_ocorrencia3	fator_nome	fator_aspecto	fator_condicionante	fator_area
2	39115	APLICAÇÃO DE COMANDOS	DESEMPENHO DO SER HUMANO	OPERAÇÃO DA AERONAVE	FATOR OPERACIONAL
3	39115	JULGAMENTO DE PILOTAGEM	DESEMPENHO DO SER HUMANO	OPERAÇÃO DA AERONAVE	FATOR OPERACIONAL
4	39115	PERCEPÇÃO	ASPECTO PSICOLÓGICO	INDIVIDUAL	FATOR HUMANO
5	39115	PLANEJAMENTO DE VOO	DESEMPENHO DO SER HUMANO	OPERAÇÃO DA AERONAVE	FATOR OPERACIONAL
6	39115	POUCA EXPERIÊNCIA DO PILOTO	DESEMPENHO DO SER HUMANO	OPERAÇÃO DA AERONAVE	FATOR OPERACIONAL
7	39115	PROCESSO DECISÓRIO	ASPECTO PSICOLÓGICO	INDIVIDUAL	FATOR HUMANO
8	39156	MANUTENÇÃO DE AERONAVE	DESEMPENHO DO SER HUMANO	MANUTENÇÃO DA AERONAVE	FATOR OPERACIONAL
9	39156	PLANEJAMENTO GERENCIAL	DESEMPENHO DO SER HUMANO	OPERAÇÃO DA AERONAVE	FATOR OPERACIONAL
10	39156	SUPERVISÃO GERENCIAL	DESEMPENHO DO SER HUMANO	OPERAÇÃO DA AERONAVE	FATOR OPERACIONAL
11	39235	PLANEJAMENTO DE VOO	DESEMPENHO DO SER HUMANO	OPERAÇÃO DA AERONAVE	FATOR OPERACIONAL
12	39235	POUCA EXPERIÊNCIA DO PILOTO	DESEMPENHO DO SER HUMANO	OPERAÇÃO DA AERONAVE	FATOR OPERACIONAL
13	39275	APLICAÇÃO DE COMANDOS	DESEMPENHO DO SER HUMANO	OPERAÇÃO DA AERONAVE	FATOR OPERACIONAL
14	39275	CULTURA DO GRUPO DE TRABALHO	ASPECTO PSICOLÓGICO	PSICOSSOCIAL	FATOR HUMANO
15	39275	CULTURA ORGANIZACIONAL	ASPECTO PSICOLÓGICO	ORGANIZACIONAL	FATOR HUMANO
16	39275	ESTRESSE	ASPECTO PSICOLÓGICO	INDIVIDUAL	FATOR HUMANO
17	39275	FADIGA	ASPECTO MÉDICO	***	FATOR HUMANO
18	39275	FORMAÇÃO			
19	39275	INSTRUÇÃO	DESEMPENHO DO SER HUMANO	OPERAÇÃO DA AERONAVE	FATOR OPERACIONAL
20	39275	MANUTENÇÃO DE AERONAVE	DESEMPENHO DO SER HUMANO	MANUTENÇÃO DA AERONAVE	FATOR OPERACIONAL
21	39275	MEDICAMENTO	ASPECTO MÉDICO	***	FATOR HUMANO
22	39275	ORGANIZAÇÃO DO TRABALHO	ASPECTO PSICOLÓGICO	ORGANIZACIONAL	FATOR HUMANO

Fonte: SANTOS, PEREIRA (2020)

Sendo assim, foram encontradas duas alternativas distintas para trabalhar com os dados disponíveis: mesclar os arquivos em um ou criar tabelas associativas para que fosse possível realizar os relacionamentos.

As duas alternativas foram analisadas em paralelo para otimizar tempo e esforço. Para a primeira, foi desenvolvido um *script* simples em *python* em que o algoritmo recebesse os arquivos a serem mesclados, gerando uma nova tabela única de nome accidents.csv (Figura 5). Entretanto, observou-se que esta alternativa também não apresentava uma solução sobre os códigos duplicados, para que fosse possível utilizar o arquivo gerado, seria necessário ajustes

manuais ao longo dos mais de 267.000 registros, criando repetições para os códigos de ocorrência fizessem referência aos registros de fatores contribuintes, tipo de ocorrência e aeronaves. Criar mais registros duplicados do que os já existentes trariam problemas para as etapas seguintes.

Figura 5. Algoritmo para mesclar os arquivos .csv

mesclar.py

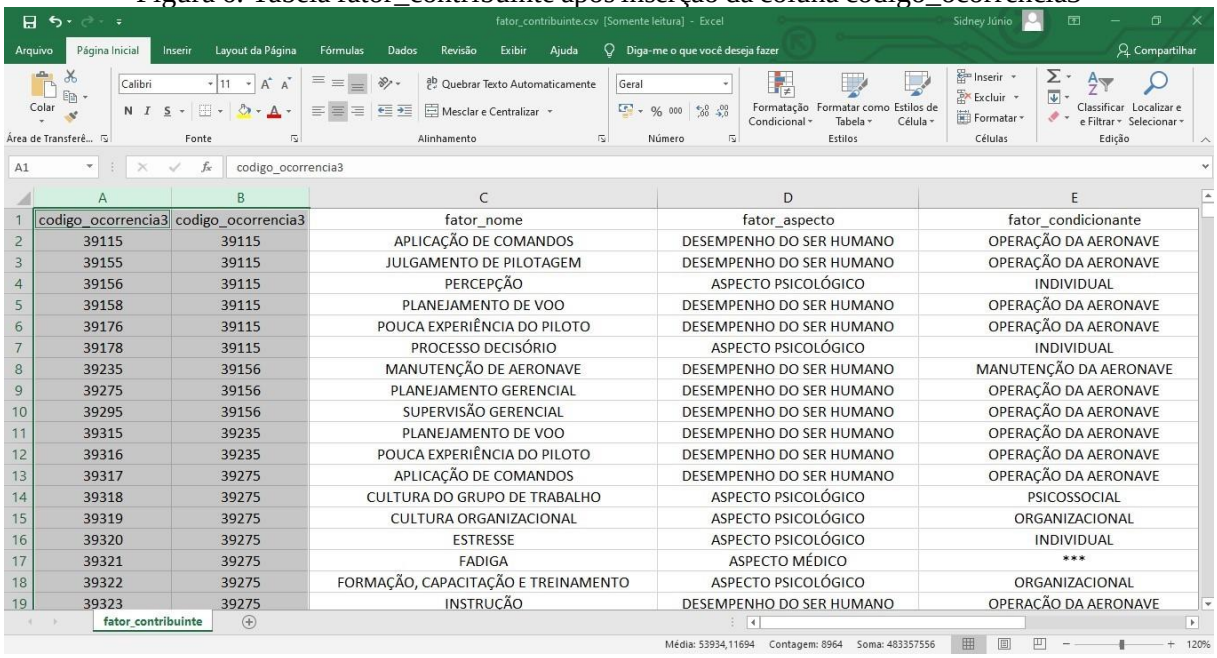
```
import pandas as pd

o = pd.read_csv('ocorrencia.csv',sep=';', encoding='utf-8')
a = pd.read_csv('aeronave.csv',sep=';', encoding='utf-8')
ot = pd.read_csv('ocorrencia_tipo.csv',sep=';', encoding='utf-8')
f = pd.read_csv('fator_contribuinte.csv',sep=';', encoding='utf-8')

base = pd.concat([o,a,ot,f], axis=1, sort=False)
base.to_csv("acidentes.csv", index=False)
```

Fonte: SANTOS, PEREIRA (2020)

A segunda alternativa, a que mostrou resultado a curto prazo, foi a criação das tabelas associativas nos relacionamentos de n:n existentes entre as tabelas ocorrencia_tipo, aeronave, fator_contribuinte com a tabela ocorrencia. Como o foco da mineração é descobrir quais os fatores contribuintes que continuam a ocorrer nas ocorrências aeronáuticas nos últimos 10 anos, a limpeza da base de dados foi focada em cima da tabela fator_contribuinte.

Inicialmente foi necessário encontrar as ocorrências aeronáuticas que possui fator contribuinte cadastrado, para isso foi utilizado o Microsoft Excel para auxiliar neste processo. Na tabela fator_contribuinte foi criada uma coluna onde foram inseridos os dados da coluna codigo_ocorrencia3 da tabela ocorrencia. Na tabela fator_contribuinte ficou com duas colunas chamadas codigo_ocorrencia3, conforme a Figura 6 abaixo.

Figura 6. Tabela fator_contribuinte após inserção da coluna codigo_ocorrencia3

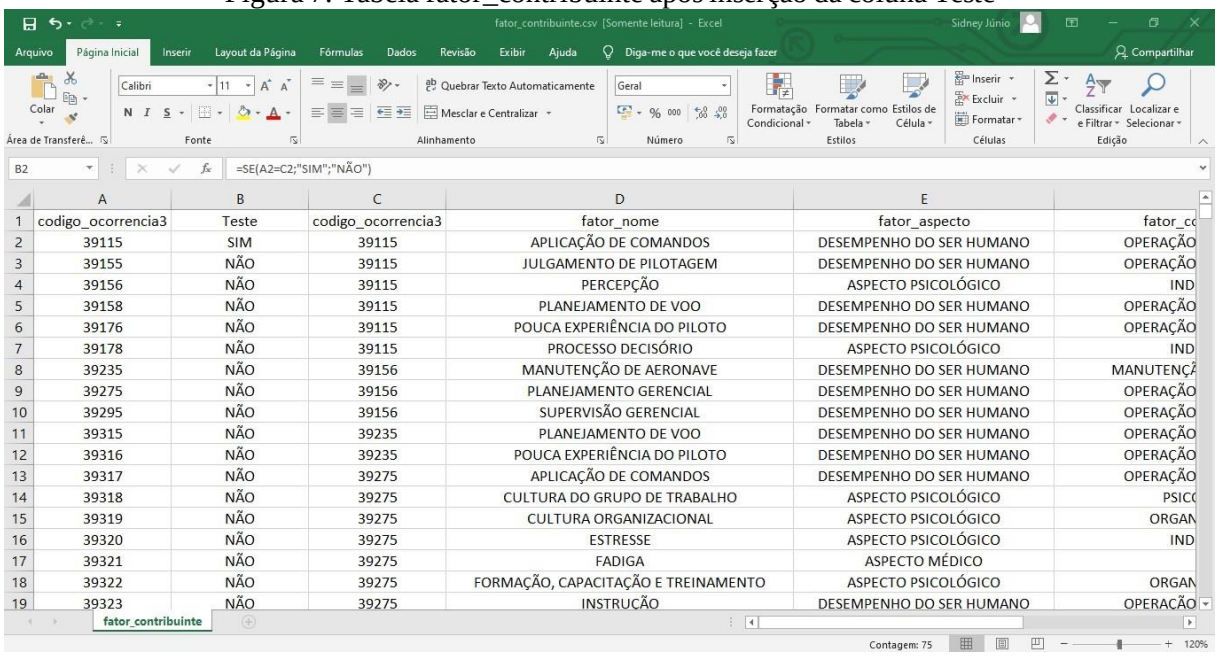


	A	B	C	D	E
1	codigo_ocorrencia3	codigo_ocorrencia3	fator_nome	fator_aspecto	fator_condicionante
2	39115	39115	APLICAÇÃO DE COMANDOS	DESEMPENHO DO SER HUMANO	OPERAÇÃO DA AERONAVE
3	39155	39115	JULGAMENTO DE PILOTAGEM	DESEMPENHO DO SER HUMANO	OPERAÇÃO DA AERONAVE
4	39156	39115	PERCEPÇÃO	ASPECTO PSICOLÓGICO	INDIVIDUAL
5	39158	39115	PLANEJAMENTO DE VOO	DESEMPENHO DO SER HUMANO	OPERAÇÃO DA AERONAVE
6	39176	39115	POUCA EXPERIÊNCIA DO PILOTO	DESEMPENHO DO SER HUMANO	OPERAÇÃO DA AERONAVE
7	39178	39115	PROCESSO DECISÓRIO	ASPECTO PSICOLÓGICO	INDIVIDUAL
8	39235	39156	MANUTENÇÃO DE AERONAVE	DESEMPENHO DO SER HUMANO	MANUTENÇÃO DA AERONAVE
9	39275	39156	PLANEJAMENTO GERENCIAL	DESEMPENHO DO SER HUMANO	OPERAÇÃO DA AERONAVE
10	39295	39156	SUPERVISÃO GERENCIAL	DESEMPENHO DO SER HUMANO	OPERAÇÃO DA AERONAVE
11	39315	39235	PLANEJAMENTO DE VOO	DESEMPENHO DO SER HUMANO	OPERAÇÃO DA AERONAVE
12	39316	39235	POUCA EXPERIÊNCIA DO PILOTO	DESEMPENHO DO SER HUMANO	OPERAÇÃO DA AERONAVE
13	39317	39275	APLICAÇÃO DE COMANDOS	DESEMPENHO DO SER HUMANO	OPERAÇÃO DA AERONAVE
14	39318	39275	CULTURA DO GRUPO DE TRABALHO	ASPECTO PSICOLÓGICO	PSICOSSOCIAL
15	39319	39275	CULTURA ORGANIZACIONAL	ASPECTO PSICOLÓGICO	ORGANIZACIONAL
16	39320	39275	ESTRESSE	ASPECTO PSICOLÓGICO	INDIVIDUAL
17	39321	39275	FADIGA	ASPECTO MÉDICO	***
18	39322	39275	FORMAÇÃO, CAPACITAÇÃO E TREINAMENTO	ASPECTO PSICOLÓGICO	ORGANIZACIONAL
19	39323	39275	INSTRUCÃO	DESEMPENHO DO SER HUMANO	OPERACÃO DA AERONAVE

Fonte: SANTOS, PEREIRA (2020)

Com as duas colunas na mesma tabela foi necessário realizar a comparação entre ambas para saber quais as ocorrências aeronáuticas que possuem fator contribuinte cadastrado. Para isso foi criada uma nova coluna na tabela fator_contribuinte chamada Teste apresentando um resultado de uma fórmula do Excel conforme Figura 7 abaixo:

Figura 7. Tabela fator_contribuinte após inserção da coluna Teste



	A	B	C	D	E	
1	codigo_ocorrencia3	Teste	codigo_ocorrencia3	fator_nome	fator_aspecto	fator_cc
2	39115	SIM	39115	APLICAÇÃO DE COMANDOS	DESEMPENHO DO SER HUMANO	OPERAÇÃO
3	39155	NÃO	39115	JULGAMENTO DE PILOTAGEM	DESEMPENHO DO SER HUMANO	OPERAÇÃO
4	39156	NÃO	39115	PERCEPÇÃO	ASPECTO PSICOLÓGICO	IND
5	39158	NÃO	39115	PLANEJAMENTO DE VOO	DESEMPENHO DO SER HUMANO	OPERAÇÃO
6	39176	NÃO	39115	POUCA EXPERIÊNCIA DO PILOTO	DESEMPENHO DO SER HUMANO	OPERAÇÃO
7	39178	NÃO	39115	PROCESSO DECISÓRIO	ASPECTO PSICOLÓGICO	IND
8	39235	NÃO	39156	MANUTENÇÃO DE AERONAVE	DESEMPENHO DO SER HUMANO	MANUTENÇÃ
9	39275	NÃO	39156	PLANEJAMENTO GERENCIAL	DESEMPENHO DO SER HUMANO	OPERAÇÃO
10	39295	NÃO	39156	SUPERVISÃO GERENCIAL	DESEMPENHO DO SER HUMANO	OPERAÇÃO
11	39315	NÃO	39235	PLANEJAMENTO DE VOO	DESEMPENHO DO SER HUMANO	OPERAÇÃO
12	39316	NÃO	39235	POUCA EXPERIÊNCIA DO PILOTO	DESEMPENHO DO SER HUMANO	OPERAÇÃO
13	39317	NÃO	39275	APLICAÇÃO DE COMANDOS	DESEMPENHO DO SER HUMANO	OPERAÇÃO
14	39318	NÃO	39275	CULTURA DO GRUPO DE TRABALHO	ASPECTO PSICOLÓGICO	PSICC
15	39319	NÃO	39275	CULTURA ORGANIZACIONAL	ASPECTO PSICOLÓGICO	ORGAN
16	39320	NÃO	39275	ESTRESSE	ASPECTO PSICOLÓGICO	IND
17	39321	NÃO	39275	FADIGA	ASPECTO MÉDICO	
18	39322	NÃO	39275	FORMAÇÃO, CAPACITAÇÃO E TREINAMENTO	ASPECTO PSICOLÓGICO	ORGAN
19	39323	NÃO	39275	INSTRUÇÃO	DESEMPENHO DO SER HUMANO	OPERAÇÃO

Fonte: SANTOS, PEREIRA (2020)

A fórmula Excel a seguir demonstra quais as ocorrências que apresentam as ocorrências que possuem fator contribuinte.

$$= SE(A2 = C2; "SIM"; "NÃO")$$

Onde os dados da fórmula são:

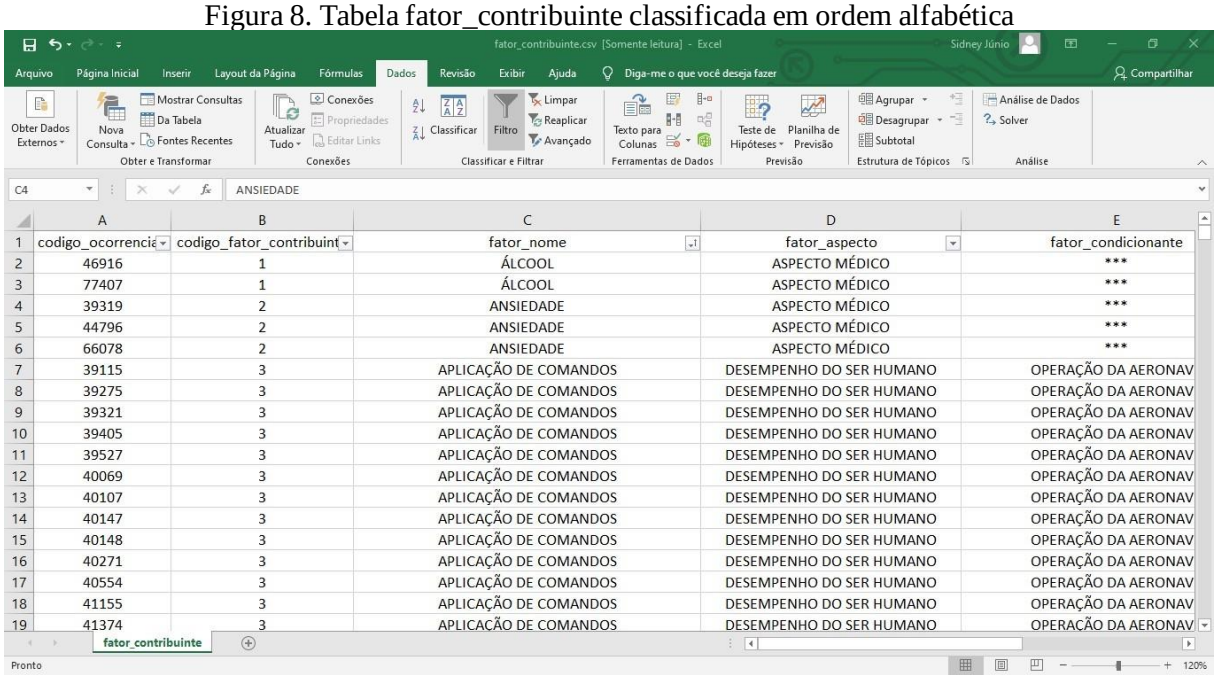
- A2: coluna que contém todos os códigos de ocorrências da tabela ocorrência;
- C2: coluna que contém todos os códigos de ocorrência que possuem fator contribuinte cadastrado;
- Resultado SIM: codigo_ocorrencia que possui fator contribuinte cadastrado;
- Resultado NÃO: codigo_ocorrencia que não possui fator contribuinte cadastrado.

Com o resultado da fórmula foram excluídas as linhas uma por uma que demonstravam o resultado NÃO para as ocorrências que não possuíam fator contribuinte cadastrado. Após excluir uma linha era repassada novamente a fórmula do Excel na coluna Teste para atualizar os dados de referência da fórmula para não efetuar uma exclusão indevida.

Concluindo as ações mencionadas acima foi obtido todas as ocorrências que possuíam fator contribuinte cadastrado. Com esses dados foi executado um processo análogo nas tabelas ocorrencia, tipo_ocorrencia e aeronave. Desta forma todas as tabelas continham somente os dados de ocorrências de interesse para o processo de Mineração de Dados.

Antes da criação das tabelas associativas foi necessário criar as chaves primárias das tabelas fator_contribuinte, tipo_ocorrencia e aeronave. Na tabela fator_contribuinte foi criada a coluna codigo_fator_contribuinte para criação das chaves primárias e assim foi inserido filtros em todas as colunas para realizar a classificação em ordem alfabética conforme Figura 8 a seguir:

Figura 8. Tabela fator_contribuinte classificada em ordem alfabética

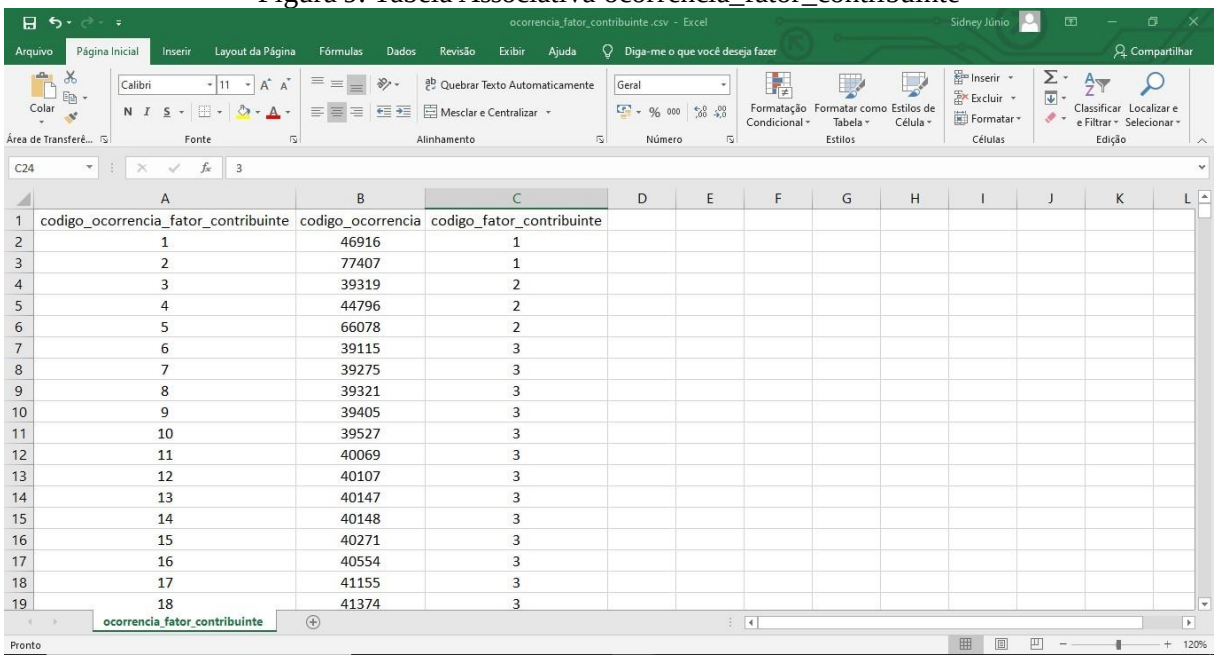


	A	B	C	D	E
1	codigo_ocorrencia	codigo_fator_contribuint	fator_nome	fator_aspecto	fator_condicionante
2	46916	1	ÁLCOOL	ASPECTO MÉDICO	***
3	77407	1	ÁLCOOL	ASPECTO MÉDICO	***
4	39319	2	ANSIEDADE	ASPECTO MÉDICO	***
5	44796	2	ANSIEDADE	ASPECTO MÉDICO	***
6	66078	2	ANSIEDADE	ASPECTO MÉDICO	***
7	39115	3	APLICAÇÃO DE COMANDOS	DESEMPENHO DO SER HUMANO	OPERAÇÃO DA AERONAV
8	39275	3	APLICAÇÃO DE COMANDOS	DESEMPENHO DO SER HUMANO	OPERAÇÃO DA AERONAV
9	39321	3	APLICAÇÃO DE COMANDOS	DESEMPENHO DO SER HUMANO	OPERAÇÃO DA AERONAV
10	39405	3	APLICAÇÃO DE COMANDOS	DESEMPENHO DO SER HUMANO	OPERAÇÃO DA AERONAV
11	39527	3	APLICAÇÃO DE COMANDOS	DESEMPENHO DO SER HUMANO	OPERAÇÃO DA AERONAV
12	40069	3	APLICAÇÃO DE COMANDOS	DESEMPENHO DO SER HUMANO	OPERAÇÃO DA AERONAV
13	40107	3	APLICAÇÃO DE COMANDOS	DESEMPENHO DO SER HUMANO	OPERAÇÃO DA AERONAV
14	40147	3	APLICAÇÃO DE COMANDOS	DESEMPENHO DO SER HUMANO	OPERAÇÃO DA AERONAV
15	40148	3	APLICAÇÃO DE COMANDOS	DESEMPENHO DO SER HUMANO	OPERAÇÃO DA AERONAV
16	40271	3	APLICAÇÃO DE COMANDOS	DESEMPENHO DO SER HUMANO	OPERAÇÃO DA AERONAV
17	40554	3	APLICAÇÃO DE COMANDOS	DESEMPENHO DO SER HUMANO	OPERAÇÃO DA AERONAV
18	41155	3	APLICAÇÃO DE COMANDOS	DESEMPENHO DO SER HUMANO	OPERAÇÃO DA AERONAV
19	41374	3	APLICAÇÃO DE COMANDOS	DESEMPENHO DO SER HUMANO	OPERAÇÃO DA AERONAV

Fonte: SANTOS, PEREIRA (2020)

Concluindo a ação acima foi criada a tabela associativa ocorrencia_fator_contribuinte utilizando as colunas codigo_ocorrencia e codigo_fator_contribuinte da tabela fator_contribuinte como chaves estrangeiras, a coluna codigo_ocorrencia foi utilizada como chave estrangeira da tabela ocorrencia pois os codigo_ocorrencia estão na tabela ocorrencia. Na tabela associativa ocorrencia_fator_contribuinte foi criada a Excel coluna codigo_ocorrencia_fator_contribuinte para a chave primária da tabela associativa conforme Figura 9:

Figura 9. Tabela Associativa ocorrencia_fator_contribuinte

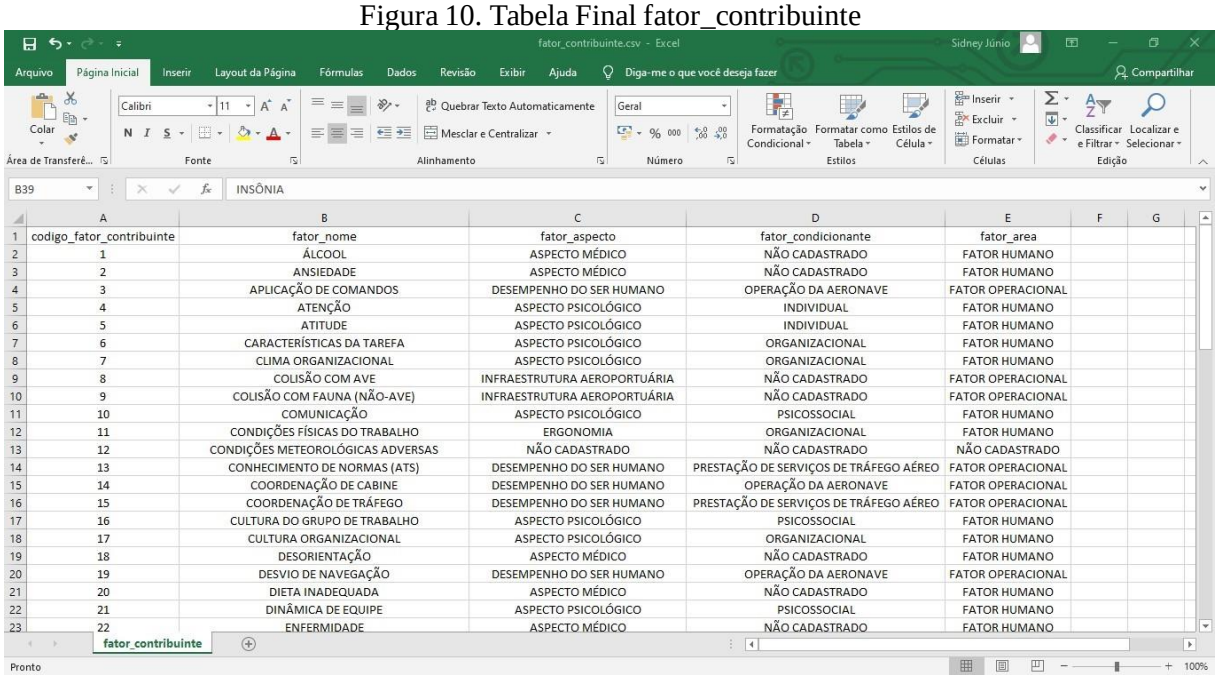


codigo_ocorrencia_fator_contribuinte	codigo_ocorrencia	codigo_fator_contribuinte
1	46916	1
2	77407	1
3	39319	2
4	44796	2
5	66078	2
6	39115	3
7	39275	3
8	39321	3
9	39405	3
10	39527	3
11	40069	3
12	40107	3
13	40147	3
14	40148	3
15	40271	3
16	40554	3
17	41155	3
18	41374	3

Fonte: SANTOS, PEREIRA (2020)

Com a tabela associativa criada, foram excluídos os dados duplicados e as colunas codigo_ocorrencia3 e Teste da tabela fator_contribuinte conforme Figura 10:

Figura 10. Tabela Final fator_contribuinte



codigo_fator_contribuinte	fator_nome	fator_aspecto	fator_condicionante	fator_area
1	ÁLCOOL	ASPECTO MÉDICO	NÃO CADASTRADO	FATOR HUMANO
2	ANSIEDADE	ASPECTO MÉDICO	NÃO CADASTRADO	FATOR HUMANO
3	APLICAÇÃO DE COMANDOS	DESEMPENHO DO SER HUMANO	OPERAÇÃO DA AERONAVE	FATOR OPERACIONAL
4	ATENÇÃO	ASPECTO PSICOLÓGICO	INDIVIDUAL	FATOR HUMANO
5	ATITUDE	ASPECTO PSICOLÓGICO	INDIVIDUAL	FATOR HUMANO
6	CARACTERÍSTICAS DA TAREFA	ASPECTO PSICOLÓGICO	ORGANIZACIONAL	FATOR HUMANO
7	CLIMA ORGANIZACIONAL	ASPECTO PSICOLÓGICO	ORGANIZACIONAL	FATOR HUMANO
8	COLISÃO COM AVE	INFRAESTRUTURA AEROPORTUÁRIA	NÃO CADASTRADO	FATOR OPERACIONAL
9	COLISÃO COM FAUNA (NÃO-AVE)	INFRAESTRUTURA AEROPORTUÁRIA	NÃO CADASTRADO	FATOR OPERACIONAL
10	COMUNICAÇÃO	ASPECTO PSICOLÓGICO	PSICOSSOCIAL	FATOR HUMANO
11	CONDIÇÕES FÍSICAS DO TRABALHO	ERGONOMIA	ORGANIZACIONAL	FATOR HUMANO
12	CONDIÇÕES METEOROLÓGICAS ADVERSAS	NÃO CADASTRADO	NÃO CADASTRADO	NÃO CADASTRADO
13	CONHECIMENTO DE NORMAS (ATS)	DESEMPENHO DO SER HUMANO	PRESTAÇÃO DE SERVIÇOS DE TRÁFEGO AÉREO	FATOR OPERACIONAL
14	COORDENAÇÃO DE CABINE	DESEMPENHO DO SER HUMANO	OPERAÇÃO DA AERONAVE	FATOR OPERACIONAL
15	COORDENAÇÃO DE TRÁFEGO	DESEMPENHO DO SER HUMANO	PRESTAÇÃO DE SERVIÇOS DE TRÁFEGO AÉREO	FATOR OPERACIONAL
16	CULTURA DO GRUPO DE TRABALHO	ASPECTO PSICOLÓGICO	PSICOSSOCIAL	FATOR HUMANO
17	CULTURA ORGANIZACIONAL	ASPECTO PSICOLÓGICO	ORGANIZACIONAL	FATOR HUMANO
18	DESORIENTAÇÃO	ASPECTO MÉDICO	NÃO CADASTRADO	FATOR HUMANO
19	DESVIO DE NAVEGAÇÃO	DESEMPENHO DO SER HUMANO	OPERAÇÃO DA AERONAVE	FATOR OPERACIONAL
20	DIETA INADEQUADA	ASPECTO MÉDICO	NÃO CADASTRADO	FATOR HUMANO
21	DINÂMICA DE EQUIPE	ASPECTO PSICOLÓGICO	PSICOSSOCIAL	FATOR HUMANO
22	ENFERMIDADE	ASPECTO MÉDICO	NÃO CADASTRADO	FATOR HUMANO

Fonte: SANTOS, PEREIRA (2020)

Para a criação das tabelas associativas ocorrencia_aeronave e ocorrencia_ocorrencia_tipo e remoção de dados duplicados das tabelas aeronave e ocorrencia_tipo foram utilizados os mesmos processos já descritos.

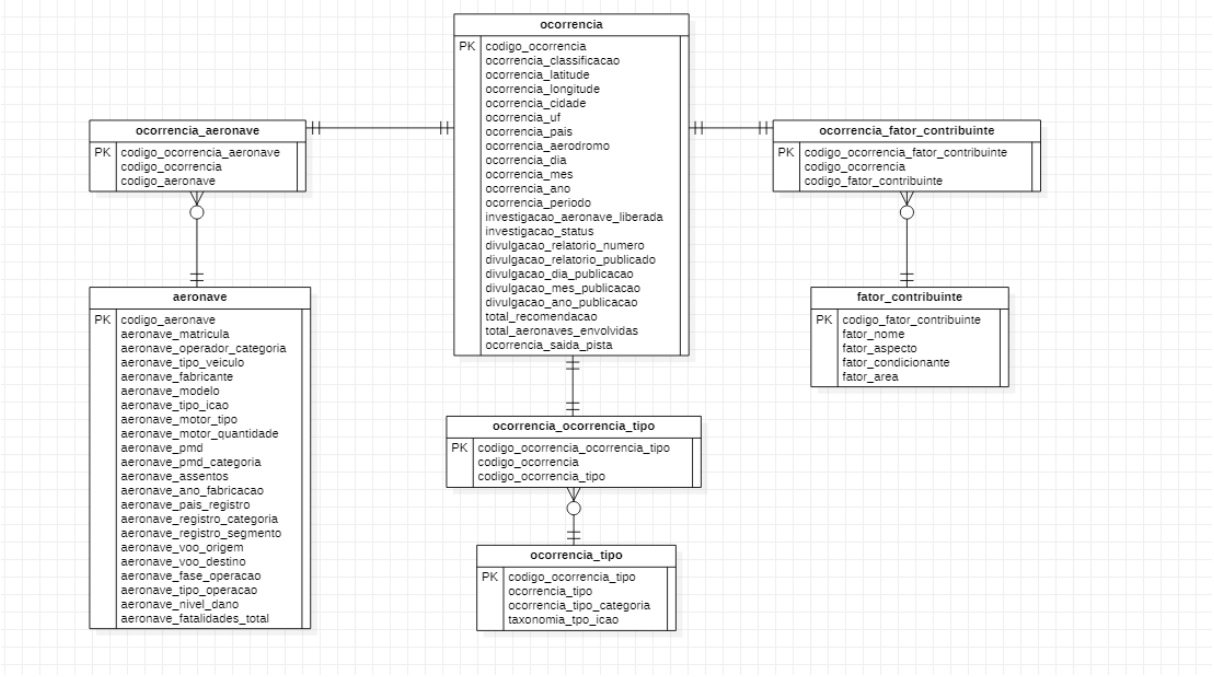
Com as tabelas criadas foi verificado que possuíam registros com informações que influenciaria no resultado na tarefa de clusterização da base de dados, portanto os registros ***,

NULL, ####, ****, *****, 0 e colunas em branco foram substituídas por NÃO CADASTRADO. Não foi optado por realizar a exclusão das linhas das tabelas pois reduziria muito a quantidade de registros da base de dados.

Os *softwares* de mineração estavam apresentando erro de leitura dos registros da coluna ocorrencia_hora da tabela ocorrencia foi substituída por uma coluna ocorrencia_periodo que contém os dados de acordo com os períodos do dia. Sendo no período de 00:00 às 05:59 foi atribuído o registro MADRUGADA, no período de 06:00 às 11:59 o registro MANHÃ, no período de 12:00 às 17:59 o registro TARDE e no período de 18:00 às 23:59 o registro NOITE.

Ao final da Preparação e Seleção dos dados, o modelo de dados sofreu alterações em seu relacionamento entre as tabelas para que fosse possível realizar as consultas de forma satisfatória (Figura 11). As planilhas em formato .cvs originais e após as alterações mencionadas estão disponíveis para acesso no Google Drive[2].

Figura 11. Modelo de Entidade e Relacionamento após modificação das tabelas



Fonte: SANTOS, PEREIRA (2020)

Quanto à seleção dos atributos a serem considerados no processo de mineração, foi avaliada a relevância de cada um conforme a intenção de identificar o relacionamento entre acidentes com fator contribuinte para agrupar padrões baseados nos eventos dos anos de 2010 a 2019. No quadro abaixo também são consideradas as tabelas as quais teve a necessidade de transformação dos dados:

[2] Disponível em: https://drive.google.com/drive/folders/10Y6Abk98bjDfglYhEakHOAnD-bOeAqtf?usp=sharing

Quadro 4. Atributos selecionados

Atributo	Descrição
fator_nome	Nome do fator contribuinte
aeronave_fase_operacao	Fase de operação no momento da ocorrência
ocorrencia_ano	Ano da ocorrência aeronáutica
ocorrencia_tipo	Tipo da ocorrência aeronáutica

Fonte: SANTOS, PEREIRA (2020)

3.3.2.2 Extração dos Padrões

Levando em consideração a base de dados escolhida e os atributos selecionados na etapa anterior, fica definido o objetivo de encontrar os fatores contribuintes (e demais agentes) mais frequentes nas ocorrências aeronáuticas entre os anos de 2010 e 2019.

Para isso foi realizado um estudo entre as tarefas de mineração que teoricamente poderiam desempenhar o melhor comportamento para que fosse alcançado o objetivo definido, considerando as características e particularidades do conjunto a ser analisado. Neste sentido, para o cenário escolhido (analisar um grupo de fatores que conduzem a uma ocorrência aeronáutica), foi definido a utilização da tarefa de Clusterização.

Com a tarefa definida, faz-se necessário avaliar o algoritmo a ser utilizado seguindo o pré-requisito de o mesmo algoritmo coexistir nas duas ferramentas e apresentarem uma saída dos dados de forma satisfatória em ambos os ambientes. Seguindo as condições anteriores, foram identificados dois algoritmos em potencial: *Hierarchical Clusterer* e *Kmeans*.

Entre esses algoritmos, seguindo a documentação das ferramentas, foram identificadas as particularidades do Quadro 5. Logo foi definido que para as necessidades e características dos dados disponíveis para análise a escolha mais assertiva seria a utilização do *Kmeans* por permitir maior autonomia para manipulação dos dados e análise dos resultados.

Quadro 5. Comparativo entre os algoritmos de Clusterização

Ferramenta	*Hierarchical Clusterer*	*K-means*
Orange	1. Entrada de Dados: Matriz de distância; 2. Saída dos Dados: Instâncias em gráfico hierárquico (considerando a distância entre os dados); 3. Sem possibilidade de selecionar a quantidade de *clusters*.	1. Entrada de Dados: Conjunto de Dados; 2. Saída dos Dados: Conjunto de Dados com índice de *cluster* como atributo de classe; 3. Possibilidade de parametrizar até 30 *clusters*.

WEKA	1. Entrada de Dados: Conjunto de Dados; 2. Saída de Dados: Relatório estatístico dos dados; 3. Possibilidade de parametrizar a quantidade desejada de *clusters*.	1. Entrada de Dados: Conjunto de Dados; 2. Saída de Dados: Gráficos de Distribuição e Dispersão; 3. Possibilidade de parametrizar a quantidade desejada de *clusters.*

Fonte: SANTOS, PEREIRA (2020)

3.3.3 Mineração dos Dados e Avaliação do Conhecimento

Assim como o descrito na Seção anterior, na etapa de Mineração dos dados, para que fosse possível atingir os objetivos deste trabalho de forma assertiva as consultas, parâmetros e atributos foram selecionados da mesma forma em ambas as ferramentas.

Apesar das ferramentas serem distintas em sua utilização, durante todo o trabalho procurou-se similar ao máximo os processos, como por exemplo, a entrada dos dados onde em ambas as ferramentas foi utilizada conexão direta com o banco de dados.

Ao longo desta etapa, conforme são relatadas, foram identificadas incompatibilidades dos resultados entre as ferramentas, desta forma, a análise que inicialmente seria feita considerando apenas o cenário geral dos últimos 10 anos foi estendida para uma análise específica que complementasse os resultados adquiridos, ou seja, optou-se por realizar a mineração com dois cenários: ocorrências dos anos de 2010 a 2019 e ocorrências do ano de 2010. Nas próximas subseções estes processos são relatados separadamente.

3.3.3.1 Orange

A ferramenta Orange apresenta uma interface mais amigável, auxiliando na manipulação dos dados e visualização dos resultados (Figura 12). Entretanto, isto não significa que a ferramenta não exige um estudo e conhecimento prévios sobre os processos a qual o usuário queira realizar. Como exemplo da entrada dos dados, para que fosse possível a conexão com a base de dados, foi necessário a configuração e preparação para utilizar o *widget SQL Table* (Anexo B).

Figura 12. Visualização dos Dados - Orange

Fonte: SANTOS, PEREIRA (2020)

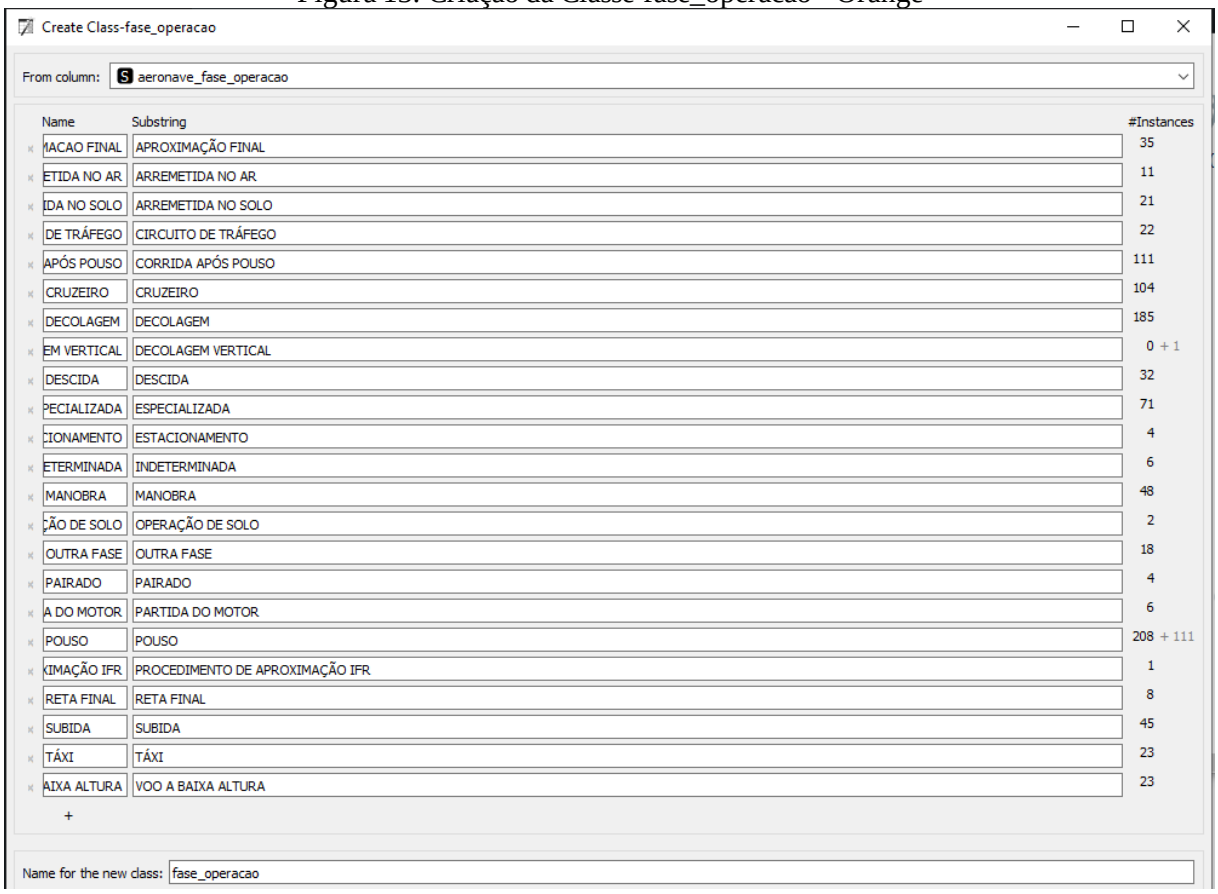
3.3.3.1.1 Clusterização - *K-means*

Com a conexão com a base estabelecida, iniciou-se a preparação dos dados na própria ferramenta, processo realizado de acordo com a base e as informações a qual deseja-se utilizar. Com o *widget SQL Table* é possível inserir um *script* de consulta a qual irá possibilitar a filtragem inicial dos dados. Essa filtragem (ou nova seleção) poderá ser realizada também com o *widget Select Columns*.

Devido a alguns dos atributos considerados (conforme Quadro 4), como fator_nome, ocorrencia_tipo e aeronave_fase_operacao, possuírem uma grande quantidade diversas de possibilidades, para esses atributos o Orange não os considerou como atributos classificatórios, sendo necessário a criação manual de classes através do *widget Create Class* (Figura 13). Desta forma, seguiu-se com as configurações a seguir:

- Para a coluna fator_nome: criou-se 51 classes;
- Para a coluna ocorrencia_tipo: criou-se 51 classes;
- Para a coluna aeronave_fase_operacao: criou-se 23 classes;

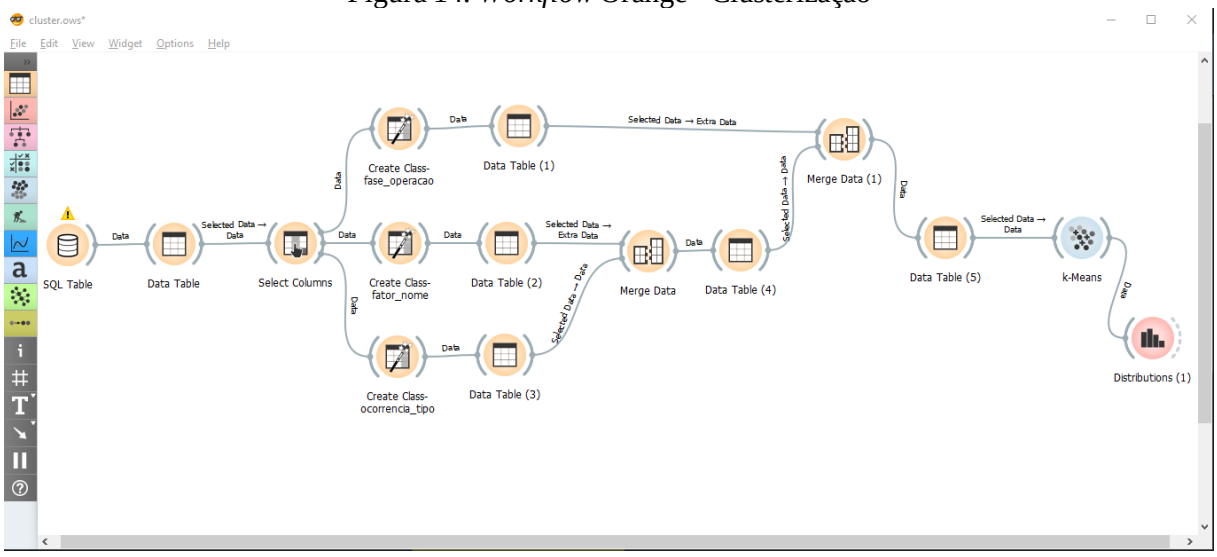
Figura 13. Criação da Classe fase_operacao - Orange



Fonte: SANTOS, PEREIRA (2020)

Com a necessidade da criação das classes e a limitação da ferramenta em possibilitar a configuração de mais de uma classe por *widget*, com a criação das classes e cada uma ter uma saída de dados individual, foi necessário utilizar também o *widget Merge Data* para que as classes criadas fossem incorporadas entre si e o restante dos atributos em uma mesma tabela e esta tabela ser usada finalmente para a aplicação do algoritmo *K-means*. Todo o fluxo descrito acima é demonstrado na Figura 14.

Figura 14. *Workflow* Orange - Clusterização



Fonte: SANTOS, PEREIRA (2020)

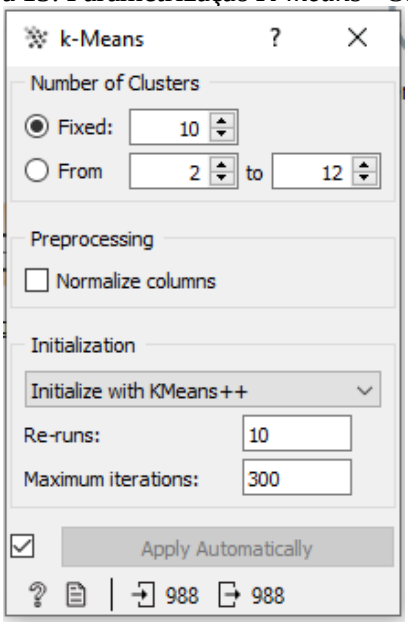
Após a criação das classes e união dos dados em uma única tabela, os dados selecionados são considerados para a execução do algoritmo de clusterização *K-means* que irá agrupar os itens de acordo com uma classificação principal escolhida pelo próprio algoritmo com base na quantidade de instâncias dos atributos escolhidos (a coluna com menor número de variações é considerada a principal para realizar os agrupamentos).

O *widget* aplica o algoritmo conforme configurado nos parâmetros e produz como saída o conjunto de dados tendo o índice de *cluster* como um atributo de classe. Segue abaixo a descrição dos parâmetros e a Figura 15 demonstrando a tela de parametrização do algoritmo:

- Número de *Clusters*:
  - Fixo: informar um dado fixo de *clusters* a serem executados;
  - De/para (Otimizado): quando selecionado mostra as pontuações de cada *cluster* considerando a distância média para elementos no mesmo *cluster* com a distância média para elementos em outros *clusters* utilizando o método *Silhouette*.
- Pré-processamento:
  - Normalizar Colunas: Eliminar redundâncias e/ou anomalias;
- Inicialização:
  - Inicie com *Kmeans* ++: iniciar o algoritmo tendo como regra, selecionar aleatoriamente as instâncias para o primeiro cluster, os seguintes são escolhidos dos atributos restantes;
  - Inicialização Aleatória: os *clusters* são atualizados automaticamente conforme a entrada de dados;

- Reexecução: quantidade de vezes em que o algoritmo será executado;
- Interações Máximas: número máximo de iterações em cada execução do algoritmo.

Figura 15. Parametrização *K-means* - Orange



Fonte: SANTOS, PEREIRA (2020)

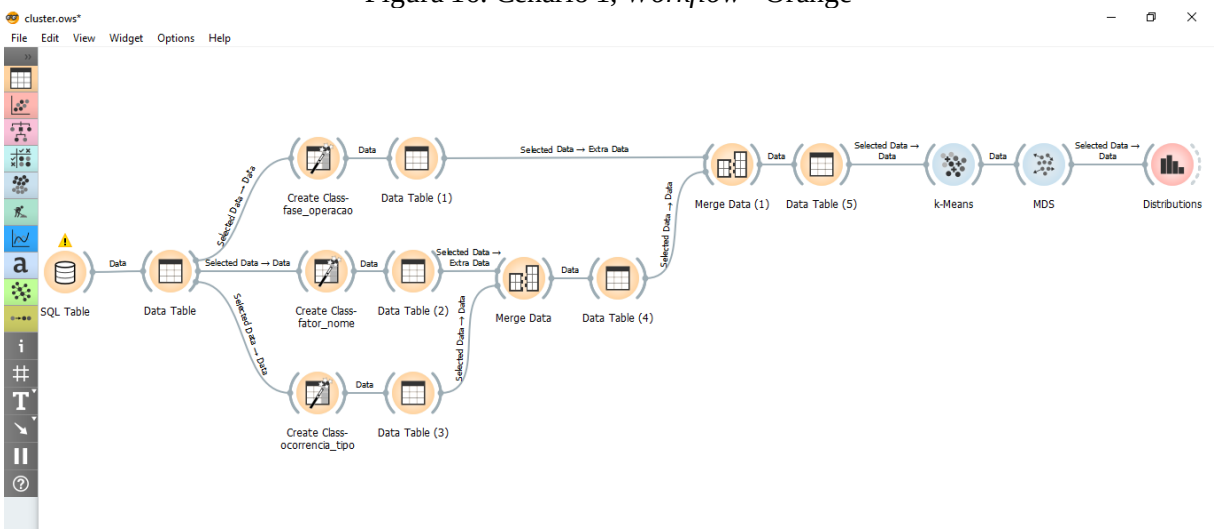
Para avaliação dos resultados gerados, o Orange dispõe de uma série de opções de visualização que são utilizadas conforme as características do conjunto de dados a ser analisado. Para este cenário foram utilizadas os widgets Distribuitions (para visualização dos dados em gráfico de barras), MDS (escalonamento multidimensional que projeta os dados em gráfico de dispersão) e *Scatter Plot* (gera um gráfico de dispersão bidimensional para atributos contínuos).

3.3.3.1.1.1 Cenário 1 - Geração de 10 *clusters* no período de 2010 a 2019

Nesta subseção são demonstradas as visualizações e saída dos dados levando em consideração o primeiro cenário com ocorrências de 2010 a 2019. Conforme será possível observar a seguir (Figura 17), para o cenário abordado utilizando a quantidade de 10 *clusters*, a visualização através do *widget Scatter Plot* não é satisfatória e dificulta a análise individual dos *clusters*. Para as demais opções de visualização utilizadas, não houve o mesmo problema.

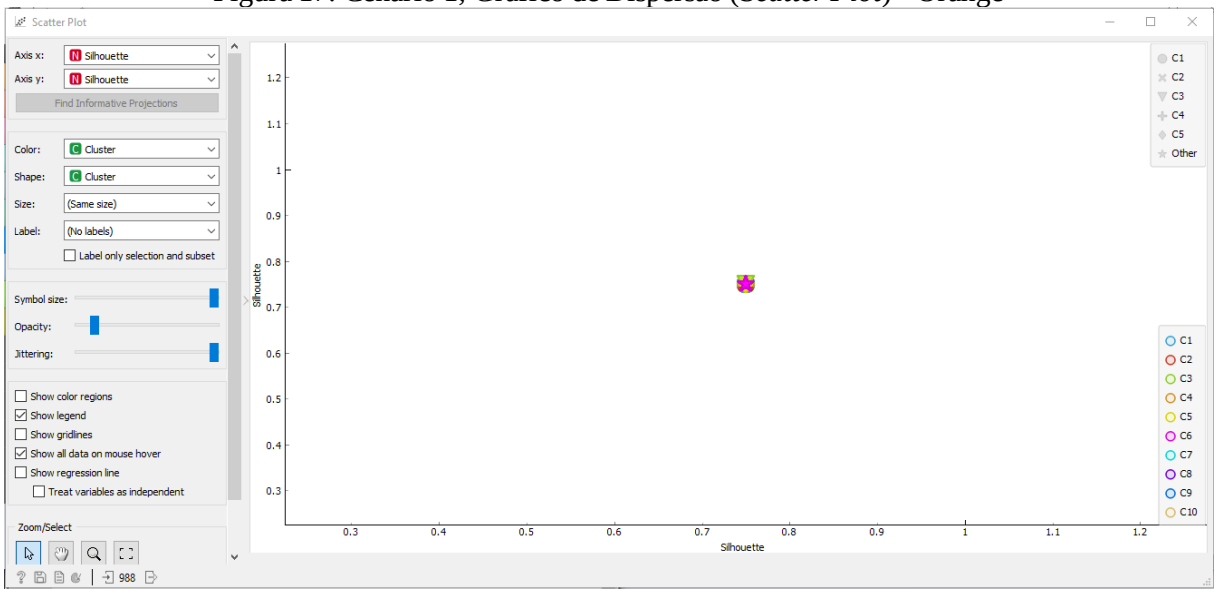
As figuras a seguir demonstram o esquema de área de trabalho utilizado bem como a disposição dos *widgets* e a visualização da saída dos dados com o *widget Scatter Plot* (figuras 16 e 17).

Figura 16. Cenário 1, *Workflow* - Orange



Fonte: SANTOS, PEREIRA (2020)

Figura 17. Cenário 1, Gráfico de Dispersão (*Scatter Plot*) - Orange



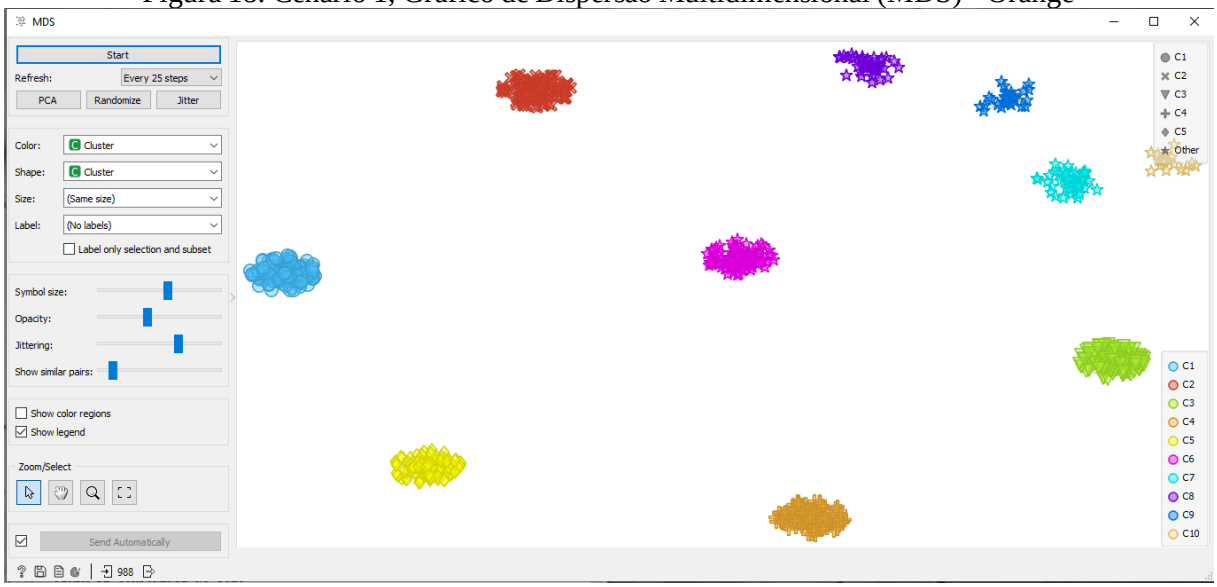
Fonte: SANTOS, PEREIRA (2020)

3.3.3.1.1.1.1 Avaliação dos Resultados

Para o cenário em questão foram utilizados os *widgets* MDS e *Distributions* em sequência. Com o emprego do *widget* MDS para a visualização da saída dos dados (Figura 18) é possível observar que em cada *cluster* há diversas possibilidades de associações (usando a classificação principal do ano da ocorrência). Cada possibilidade associativa é ilustrada como um símbolo e cor.

Com a aplicação de um número fixo de 10 *clusters* no parâmetro do *widget* (da forma como mostra a figura 15) e o comportamento do algoritmo em considerar a classificação com o número menor de instâncias como atributo principal dos *clusters*, nesta configuração cada *cluster* mantém como atributo principal o ano da ocorrência.

Figura 18. Cenário 1, Gráfico de Dispersão Multidimensional (MDS) - Orange

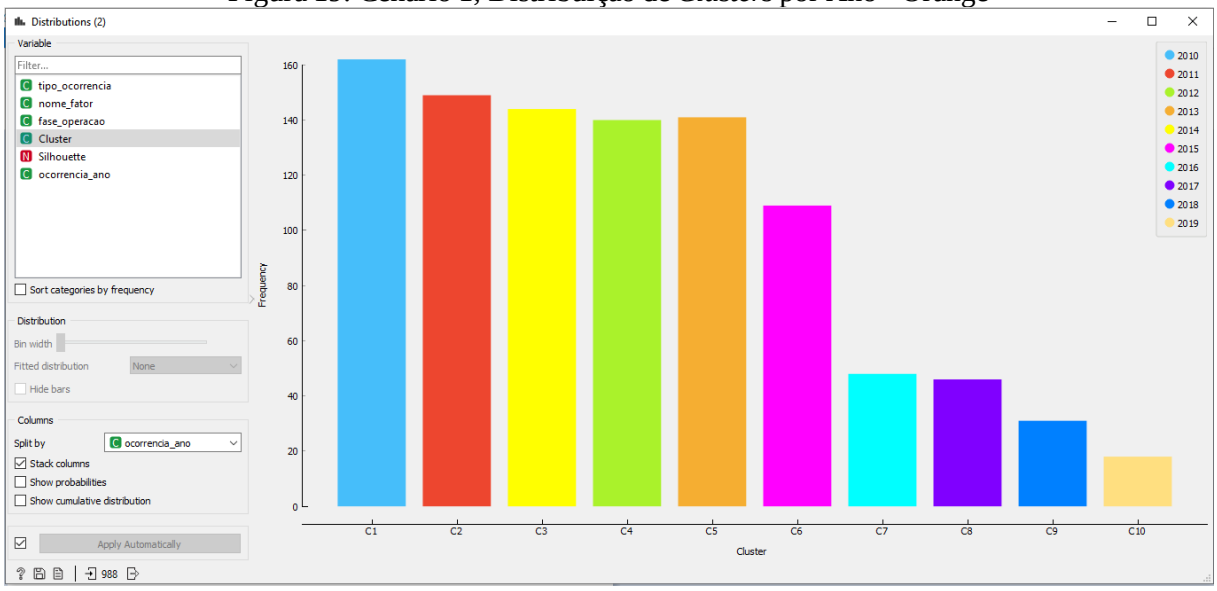


Fonte: SANTOS, PEREIRA (2020)

Para melhor análise, pode ser usada uma segunda alternativa de análise para complementar o resultado demonstrado, que neste caso foi escolhido o *widget Distributions* porque foi possível avaliar melhor os resultados gerados.

Das 988 instâncias dos registros de ocorrências aeronáuticas entre os anos de 2010 a 2019 geradas pelo algoritmo, pode-se observar no gráfico de distribuição da figura 19, os anos com maior número de ocorrências foram 2010 (162 instâncias), 2011 (149 instâncias) e 2014 (144 instâncias), estes números chegam a representar 46% do total do período. A baixa expressiva destes números ocorreu apenas em 2016 onde foram registradas 48 instâncias.

Figura 19. Cenário 1, Distribuição de *Clusters* por Ano - Orange

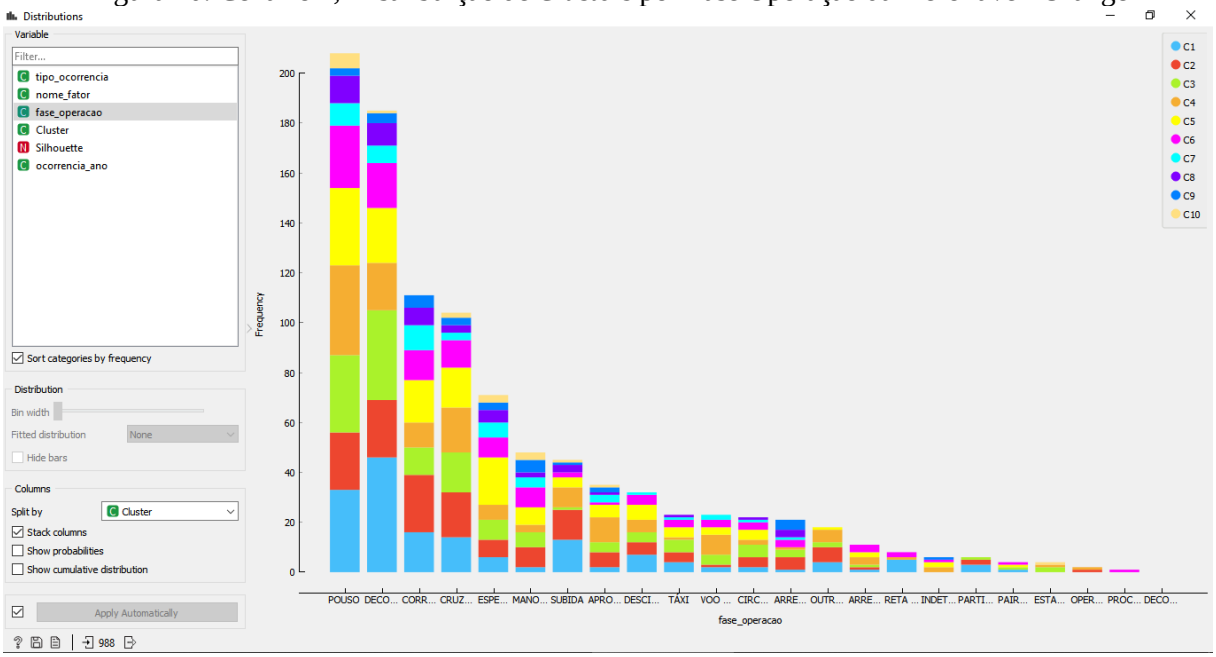


Fonte: SANTOS, PEREIRA (2020)

Em relação a fase de operação da aeronave no momento da ocorrência, as fases que mais se destacaram foram: Pouso, Decolagem e Corrida após pouso, com respectivamente 208, 185

e 111 instâncias (Figura 20). Quanto aos fatores contribuintes, 3 demonstraram maior frequência que os demais, são eles: Julgamento de Pilotagem (158 instâncias), Equipamento de Apoio (ATS) (152 instâncias) e Intoxicação Alimentar (133 instâncias). Para os tipos de ocorrência a maior frequência ocorre para: Perda de Controle no Solo com 158 instâncias, Falha do Motor em Voo com 152 instâncias e Perda de Controle em Voo com 133 instâncias.

Figura 20. Cenário 1, Distribuição de *Clusters* por Fase Operação da Aeronave - Orange



Fonte: SANTOS, PEREIRA (2020)

Devido a quantidade de instâncias para as colunas fator_nome e ocorrencia_tipo a visualização dos dados não é demonstrada de forma completa, sendo necessário selecionar e arrastar a coluna de legendas para que todas as instâncias possam ser visualizadas.

3.3.3.1.1.2 Cenário 2 - Geração de 5 *clusters* no ano de 2010

De acordo com análise realizada no cenário anterior, pôde ser observado que o ano onde mais ocorreram ocorrências aeronáuticas foi em 2010. Sendo assim, o segundo cenário para análise específica, foi escolhido avaliar os fatores chaves das ocorrências de 2010.

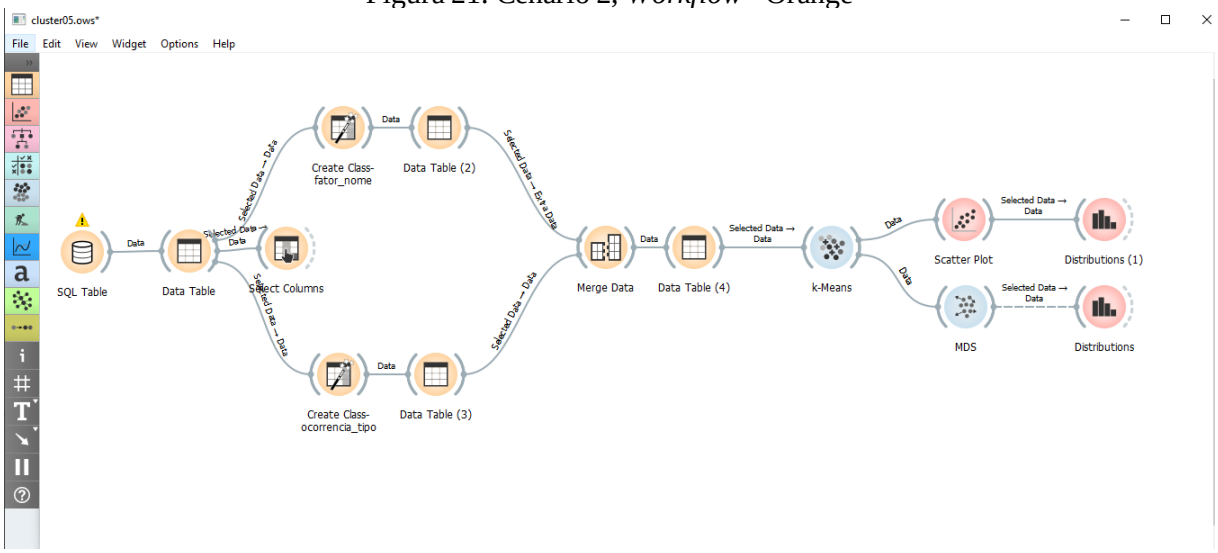
Para este cenário houve um novo fluxo de trabalho por apresentar algumas diferenças em relação ao conjunto de dados analisado anteriormente, como por exemplo, não houve a necessidade de criação de classe para o argumento aeronave_fase_operacao por este apresentar 18 instâncias das 23 criadas para o Cenário 1. Desta forma, o Orange considerou o argumento como *feature* (característica de determinada informação do conjunto de dados).

Entretanto, para os demais argumentos utilizados, ainda foram criadas as classificações da mesma forma que no cenário anterior, necessitando do auxílio do *widget Merge Data* para união de todos os argumentos para aplicar o conjunto de dados ao algoritmo *K-means*.

A configuração dos parâmetros do algoritmo também foi alterada para se adequar ao cenário atual, desta vez, foram utilizados apenas 5 *clusters*.

Para a visualização da saída dos dados foram utilizados os mesmos *widgets* do Cenário 1 com a adição do *widget Scatter Plot*. Seguem abaixo as figuras que demonstram a disposição da área de trabalho após a preparação para leitura e análise do conjunto de dados e a saída de dados após a aplicação do algoritmo (Figura 21).

Figura 21. Cenário 2, *Workflow* - Orange



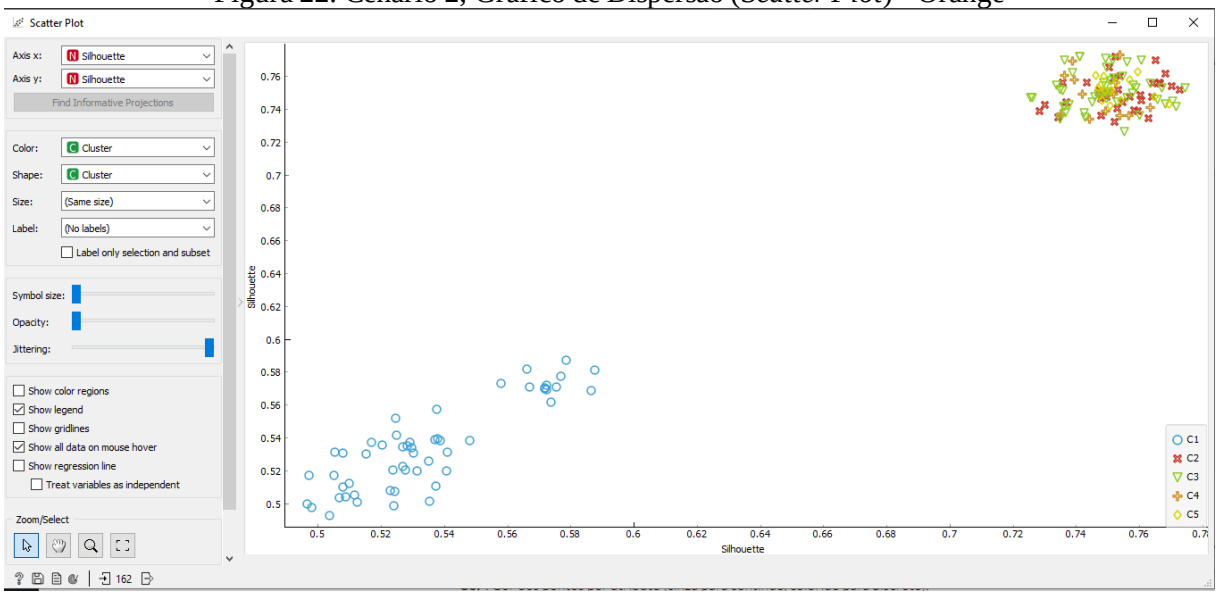
Fonte: SANTOS, PEREIRA (2020)

3.3.3.1.1.2.1 Avaliação dos Resultados

Ao contrário do Cenário 1, neste contexto o *widget Scatter Plot* demonstrou a saída dos dados de forma adequada para a interpretação dos resultados. Neste contexto, por existir menos *clusters* do que a quantidade de instâncias para a coluna fase_operacao o C1 (primeiro *cluster*) recebeu as instâncias com menor frequência, enquanto as 4 fases com maior ocorrência foram dispostas separadamente em cada um dos grupos restantes.

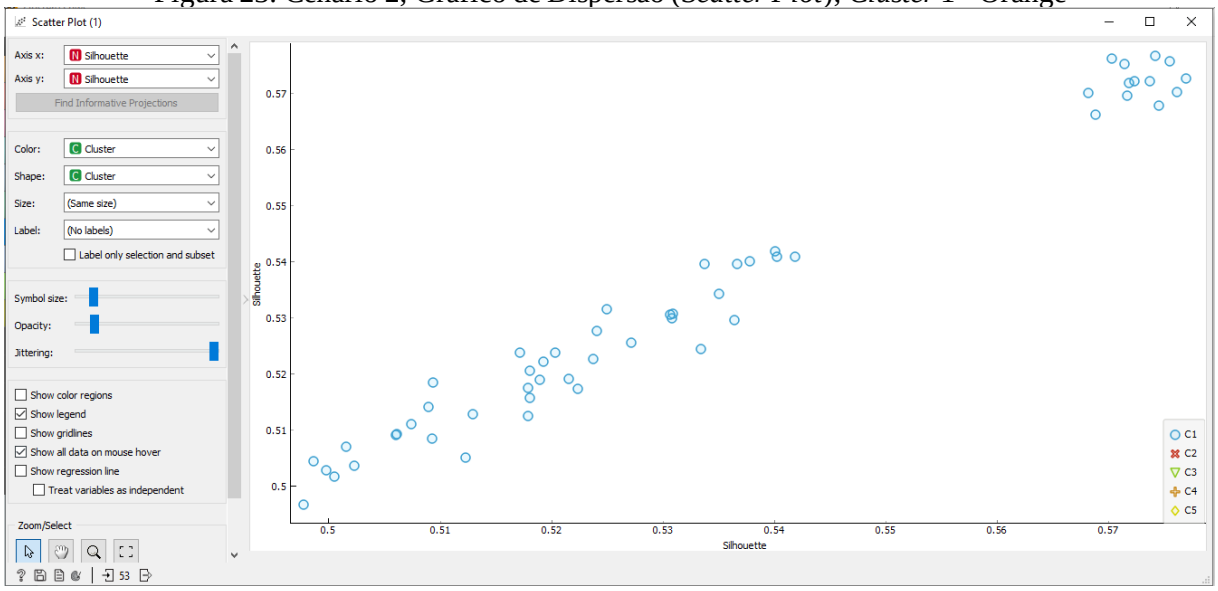
Abaixo, nas figuras 22 e 23, é possível observar a disposição dos dados, com exemplo do *cluster* 1, os círculos mais próximos e/ou entrelaçados compartilham da mesma fase_operacao (atributo principal para a classificação) ou possuem outras informações semelhantes, caso dos *clusters* de 2 a 5.

Figura 22. Cenário 2, Gráfico de Dispersão (*Scatter Plot*) - Orange



Fonte: SANTOS, PEREIRA (2020)

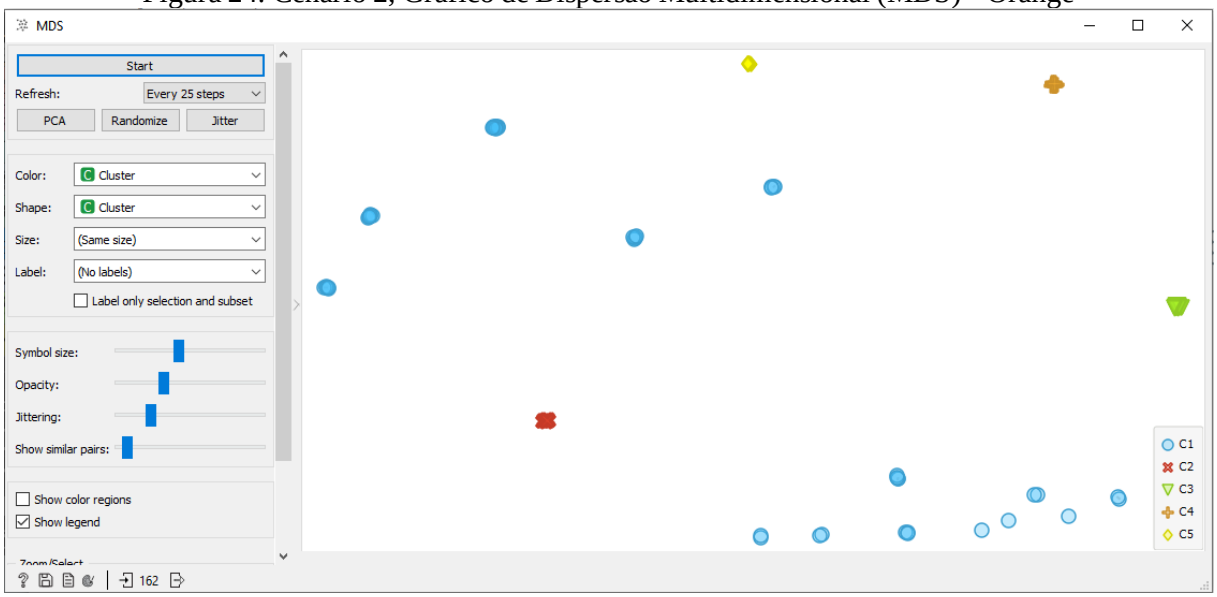
Figura 23. Cenário 2, Gráfico de Dispersão (*Scatter Plot*), *Cluster* 1 - Orange



Fonte: SANTOS, PEREIRA (2020)

Neste caso, a visualização dos *clusters* pelo *widget* MDS se aproximou da visualização do *Scatter Plot* (Figura 24). Cada círculo ilustrado representa uma fase de operação, a tonalidade do círculo é mais presente conforme a quantidade de ocorrências que compartilham da mesma classificação.

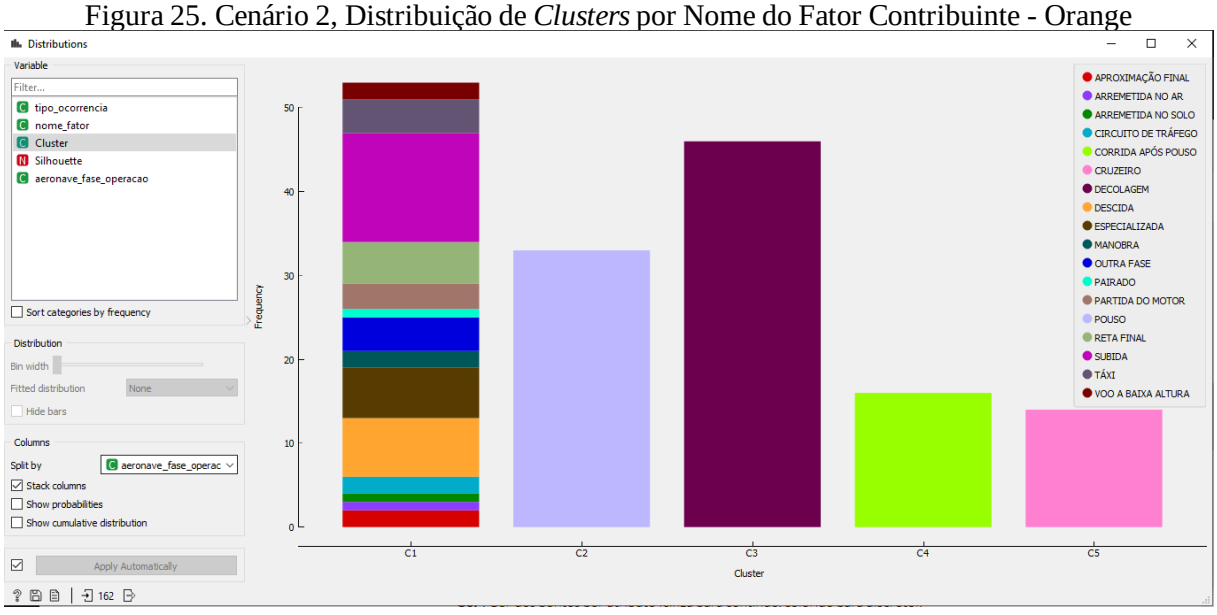
Figura 24. Cenário 2, Gráfico de Dispersão Multidimensional (MDS) - Orange



Fonte: SANTOS, PEREIRA (2020)

Assim como no Cenário 1, neste também foi usado o *widget Distributions* para reforçar os resultados e facilitar o entendimento sobre eles. Conforme dados analisados, foi identificado o conjunto de dados com maior número de instâncias.

Figura 25. Cenário 2, Distribuição de *Clusters* por Nome do Fator Contribuinte - Orange



Fonte: SANTOS, PEREIRA (2020)

Ao que se refere a fase de operação da aeronave no momento da ocorrência, os momentos que mais se destacaram foram: Decolagem, Pouso, e Corrida após Pouso, com respectivamente 46, 33 e 16 instâncias (Figura 25). Quanto aos fatores contribuintes: Equipamento de Apoio (ATS) (29 instâncias), Enfermidade (26 instâncias) e Intoxicação Alimentar (22 instâncias). Para os tipos de ocorrência a maior frequência no ano de 2010: Falha do Motor em Voo com 29 instâncias, Excursão de Pista com 26 instâncias e Perda de Controle em Voo com 22 instâncias.

3.3.3.2 WEKA

A ferramenta WEKA apresenta uma interface limpa, objetiva e pouco intuitiva, porém não foi impedimento para realizar as configurações da ferramenta e parametrização do algoritmo de Clusterização (*SimpleKmeans*). Já a resultado da clusterização é apresentada de forma textual e quantitativa o que necessita de uma análise morosa. O resultado da clusterização foi exportado para o bloco de notas e dividida em partes para uma visualização completa através das figuras 26, 27, 28 e 29.

Figura 26. *Run Information* - WEKA

```
=== Run information ===

Scheme:       weka.clusterers.SimpleKMeans -init 0 -max-candidates 100 -periodic-pruning 10000 -min-density 2.0 -t1 -1.25 -t2 -1.0 -N 10 -A
"weka.core.EuclideanDistance -R first-last" -I 500 -num-slots 1 -S 10
Relation:     QueryResult
Instances:    988
Attributes:   4
              fator_nome
              aeronave_fase_operacao
              ocorrencia_ano
              ocorrencia_tipo
Test mode:    evaluate on training data
```

Fonte: SANTOS, PEREIRA (2020)

Figura 27. *Clustering model* - WEKA

```
=== Clustering model (full training set) ===


kMeans
======

Number of iterations: 4
Within cluster sum of squared errors: 1886.0

Initial starting points (random):

Cluster 0: 'JULGAMENTO DE PILOTAGEM','CORRIDA APÓS POUSO',2013,'PERDA DE CONTROLE NO SOLO'
Cluster 1: 'CULTURA ORGANIZACIONAL',TÁXI,2010,'COM TREM DE POUSO'
Cluster 2: 'CONHECIMENTO DE NORMAS (ATS)',DECOLAGEM,2011,'COM PARA-BRISAS / JANELA / PORTA'
Cluster 3: 'INTOXICAÇÃO ALIMENTAR','ARREMETIDA NO AR',2015,'PERDA DE CONTROLE EM VOO'
Cluster 4: 'ILUSÕES VISUAIS',SUBIDA,2011,'OPERAÇÃO A BAIXA ALTITUDE'
Cluster 5: 'ILUSÕES VISUAIS',MANOBRA,2019,'OPERAÇÃO A BAIXA ALTITUDE'
Cluster 6: 'CARACTERÍSTICAS DA TAREFA','VOO A BAIXA ALTURA',2012,'COLISÃO COM OBSTÁCULO DURANTE A DECOLAGEM E POUSO'
Cluster 7: PERCEPÇÃO,CRUZEIRO,2010,'VOO CONTROLADO CONTRA O TERRENO'
Cluster 8: ENFERMIDADE,'CORRIDA APÓS POUSO',2017,'EXCURSÃO DE PISTA'
Cluster 9: 'EQUIPAMENTO DE APOIO (ATS)',ESPECIALIZADA,2014,'FALHA DO MOTOR EM VOO'

Missing values globally replaced with mean/mode
```

Fonte: SANTOS, PEREIRA (2020)

Figura 28. *Final cluster centroids* - WEKA

```
Final cluster centroids:
                                                                                          Cluster#
Attribute                                                    Full Data                           0
                                                               (988.0)                     (237.0)
==================================================================================================
fator_nome                                     JULGAMENTO DE PILOTAGEM     JULGAMENTO DE PILOTAGEM
aeronave_fase_operacao                                           POUSO          CORRIDA APÓS POUSO
ocorrencia_ano                                                    2010                        2013
ocorrencia_tipo                              PERDA DE CONTROLE NO SOLO   PERDA DE CONTROLE NO SOLO


Time taken to build model (full training data) : 0.06 seconds
```

Fonte: SANTOS, PEREIRA (2020)

Figura 29. *Model and evaluation on training set* - WEKA

```
=== Model and evaluation on training set ===

Clustered Instances

0      237 ( 24%)
1      143 ( 14%)
2       74 (  7%)
3      140 ( 14%)
4       29 (  3%)
5       15 (  2%)
6      113 ( 11%)
7       22 (  2%)
8       69 (  7%)
9      146 ( 15%)
```

Fonte: SANTOS, PEREIRA (2020)

Contudo é necessário um conhecimento prévio dos processos que o usuário necessita realizar, pois para realizar a comunicação com o banco de dados PostgreSQL é necessário a alteração textual dos arquivos de instalação da ferramenta e *download* de arquivos adicionais, conforme demonstrados no anexo A.

3.3.3.2.1 Clusterização - *SimpleKmeans*

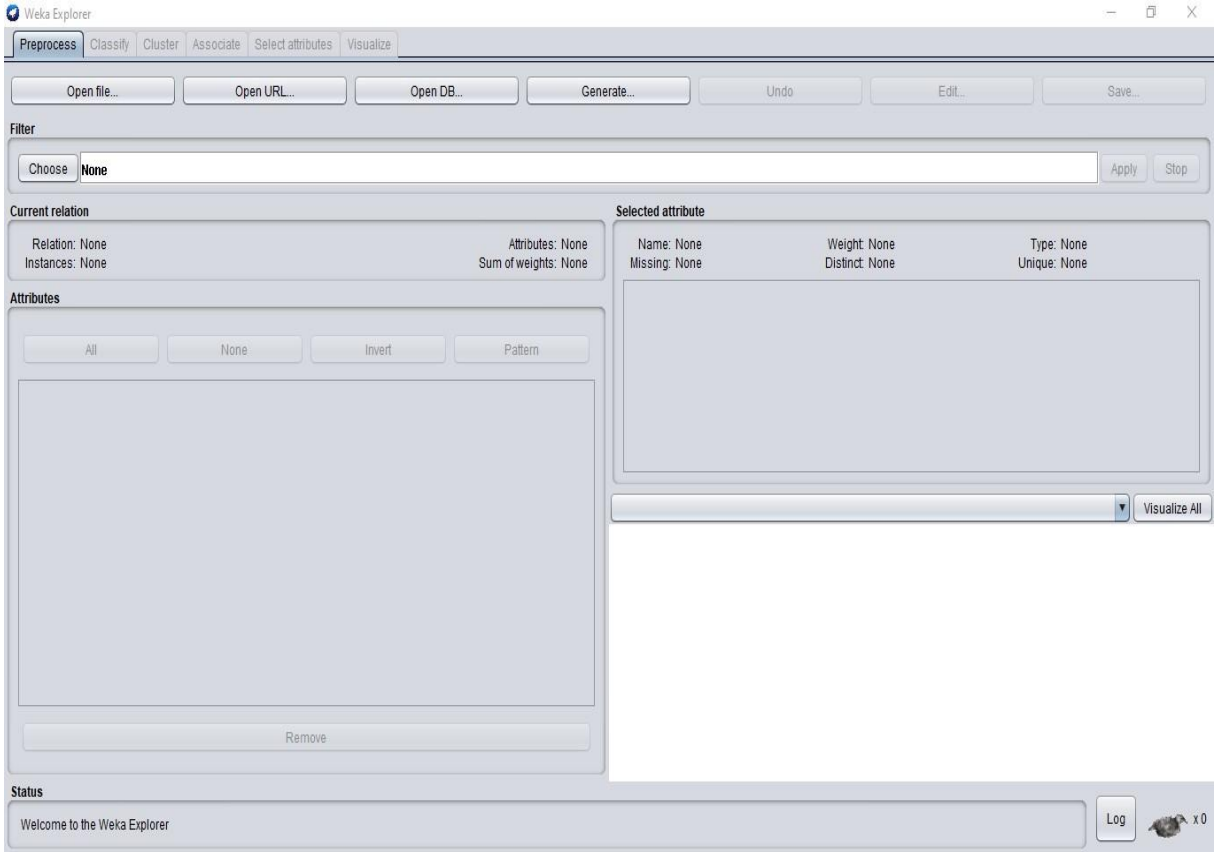
Antes de iniciar a etapa de pré-processamento dos dados é necessário realizar a comunicação com o banco de dados PostgreSQL que contém a base de dados do CENIPA. A figura 30 exibe a tela inicial da ferramenta WEKA.

Figura 30. Tela inicial - WEKA

Fonte: SANTOS, PEREIRA (2020)

No menu *Explorer* agrupa as funcionalidades utilizadas na etapa de pré-processamento, mineração e visualização dos dados que foram utilizadas, figura 31.

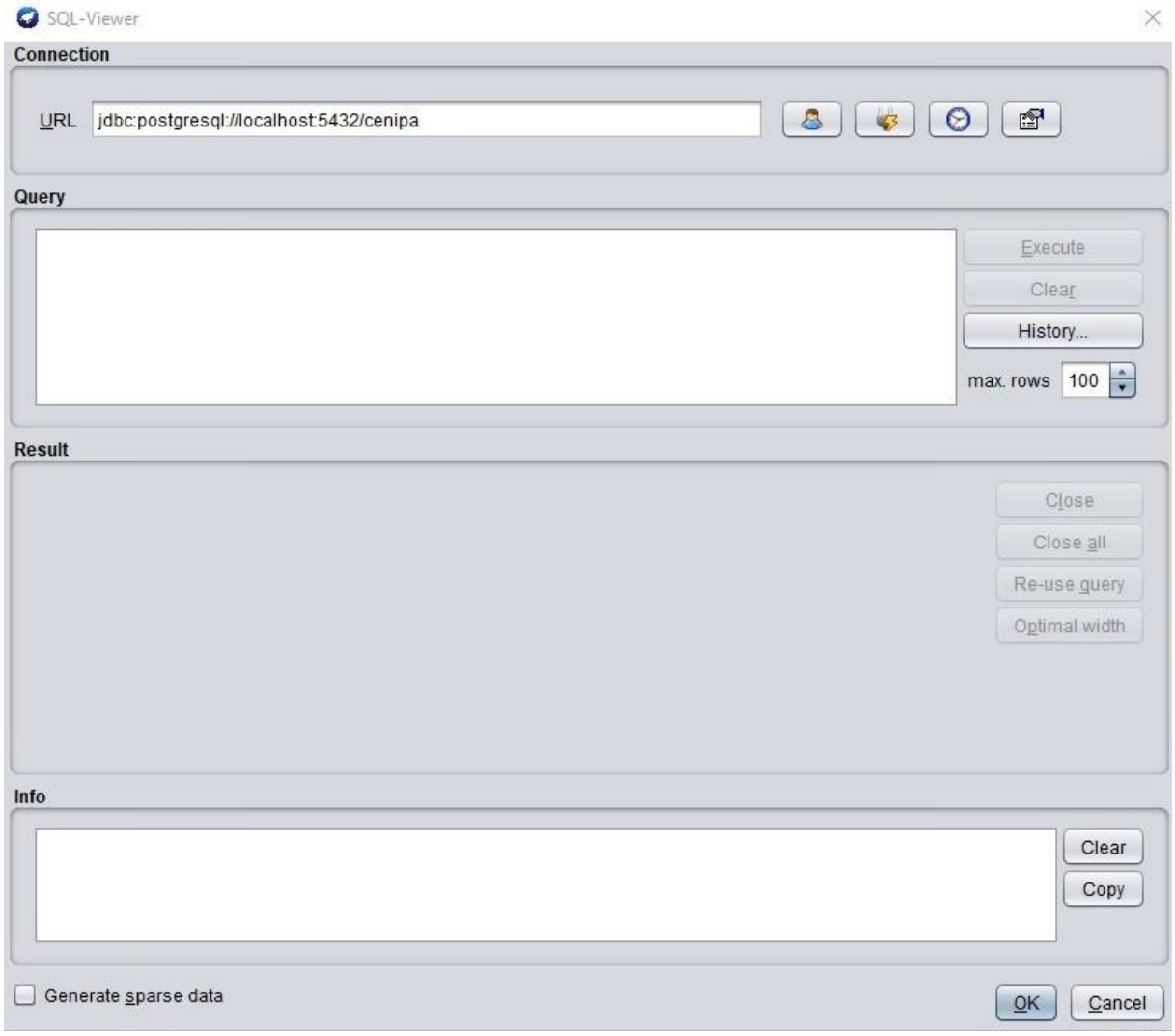
Figura 31. Tela *Explorer* - WEKA



Fonte: SANTOS, PEREIRA (2020)

Ao clicar no botão *Open DB...* é demonstrada a tela *SQL-Viewer* utilizada para a comunicação com o banco de dados PostgreSQL, figura 32.

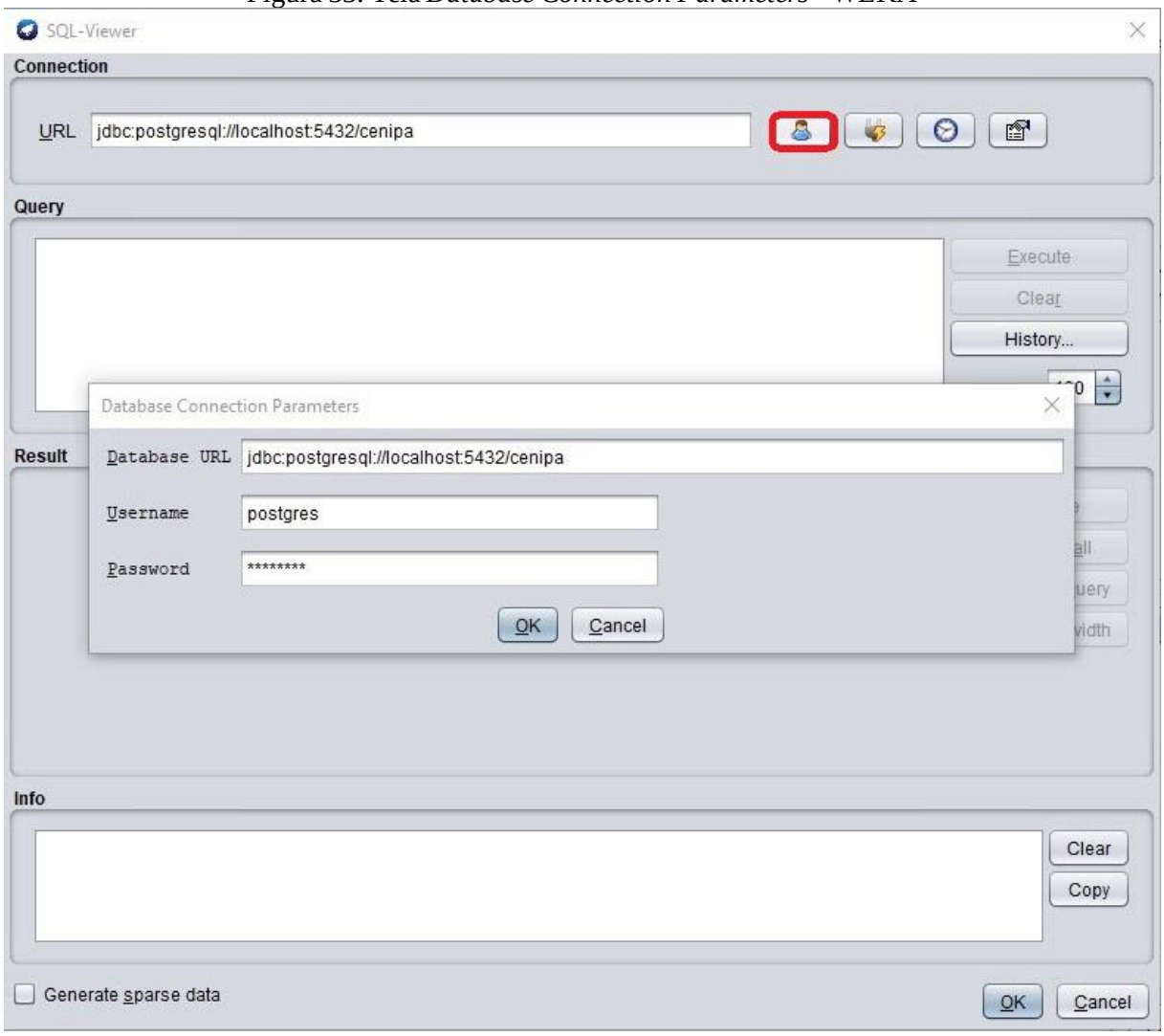
Figura 32. Tela *SQL-Viewer* - WEKA



Fonte: SANTOS, PEREIRA (2020)

Ao acionar o botão *Set user and password* é aberta uma tela *Database Connection Parameters* para inserção das credenciais de acesso ao banco de dados, figura 33.

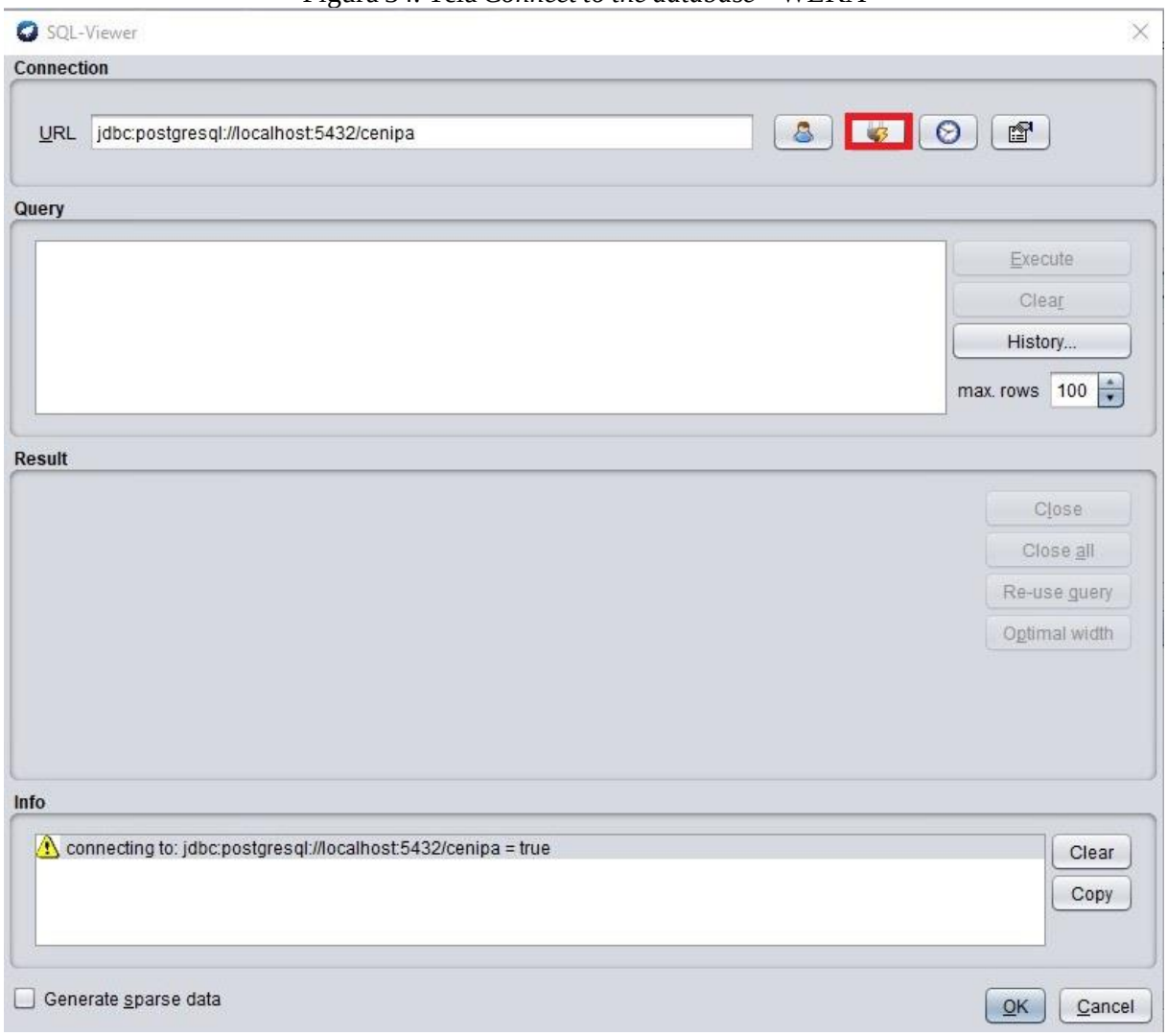
Figura 33. Tela *Database Connection Parameters* - WEKA



Fonte: SANTOS, PEREIRA (2020)

Clicando no botão *Connect to the database* é realizada a comunicação com o banco de dados e é demonstrado no quadro *Info* se a comunicação ocorreu com sucesso, figura 34.

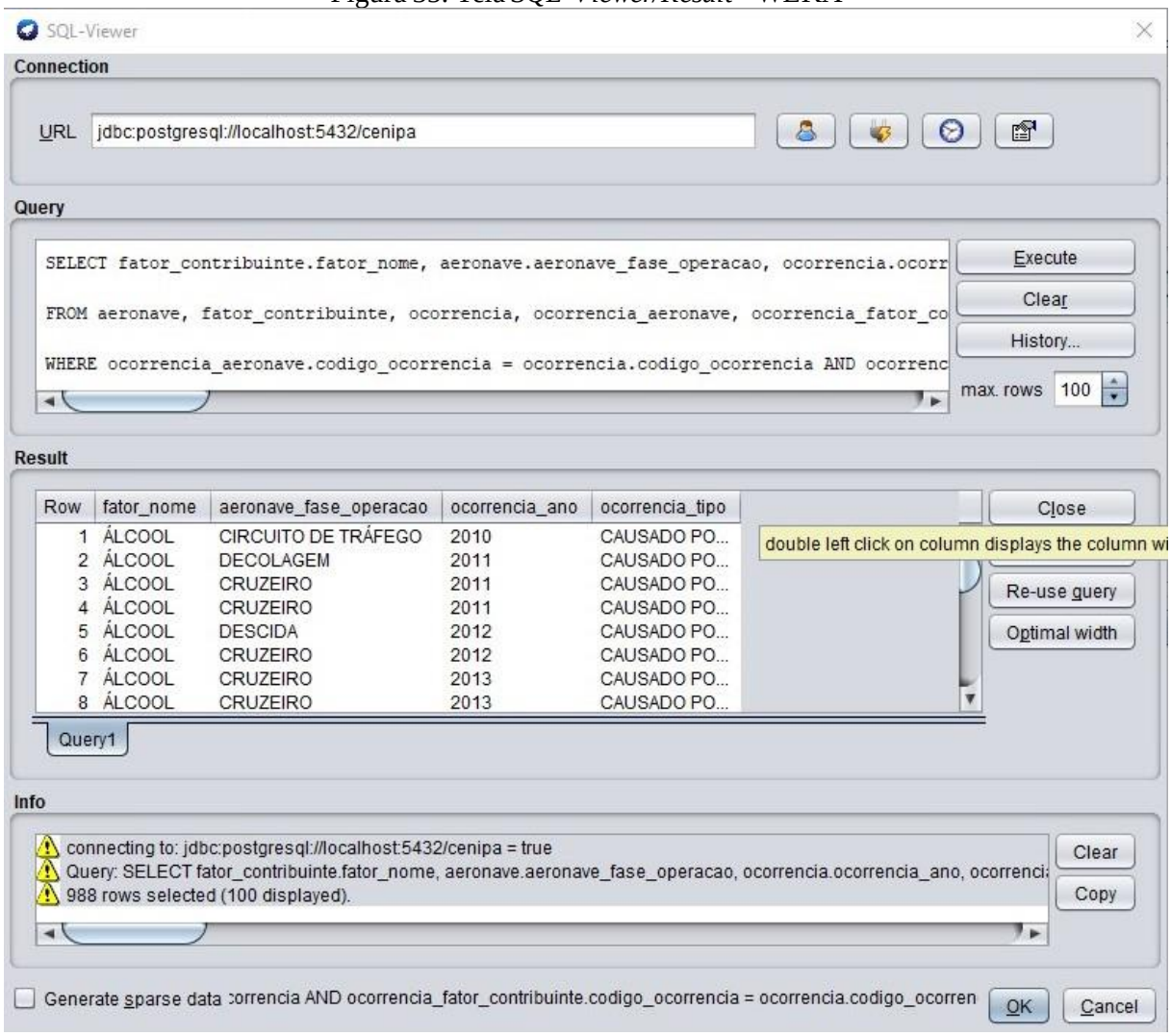
Figura 34. Tela *Connect to the database* - WEKA



Fonte: SANTOS, PEREIRA (2020)

Na tela de *SQL-Viewer*, no quadro *Query* é o campo que insere a *query* a ser utilizada para realizar a busca das informações da base de dados armazenadas no banco de dados PostgreSQL e ao clicar no botão *Execute* são retornadas as informações da base de dados a serem utilizadas na mineração de dados e o resultado é apresentado no quadro *Result*, figura 35.

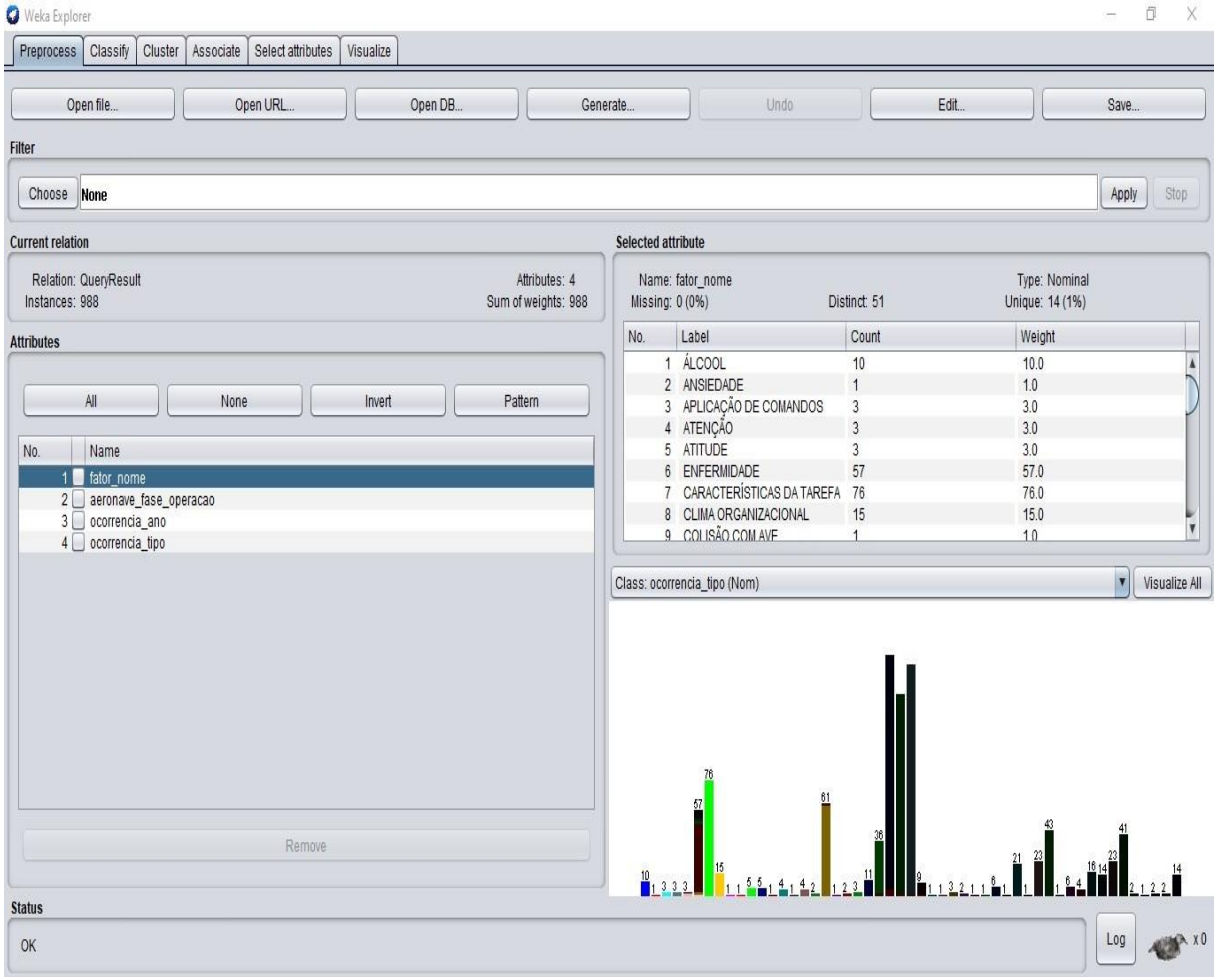
Figura 35. Tela *SQL-Viewer/Result* - WEKA



Fonte: SANTOS, PEREIRA (2020)

Ao clicar no botão *OK* é demonstrada a tela de pré-processamento de dados. No quadro *Filters* possui uma Seção de opções de filtros para tratamento de dados. Ao clicar em algum atributo do quadro *Attibutes* pode-se visualizar as informações do conjunto de dados no quadro *Select attribute*. Por fim, no canto inferior esquerdo é gerado um gráfico de acordo com o atributo classe que foi selecionado, figura 36.

Figura 36. WEKA *Explorer*, aba *Preprocess* - WEKA



Fonte: SANTOS, PEREIRA (2020)

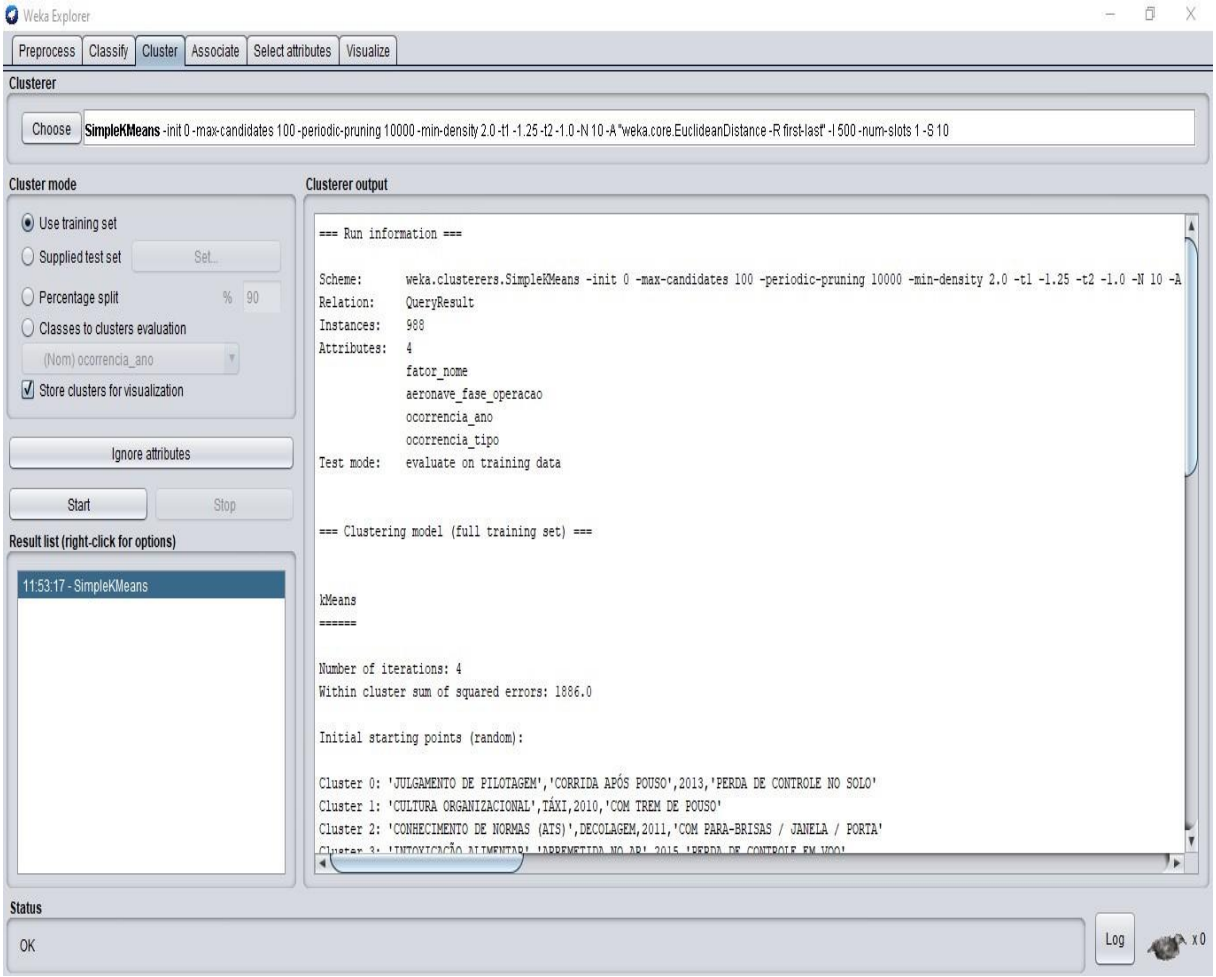
Na parte superior da tela do WEKA *Explorer* são demonstrados grupos de algoritmos de acordo com a Técnica de Mineração a ser utilizada. O grupo de algoritmos que fazem parte da Técnica de Clusterização fica na aba *Cluster*. No quadro *Clusterer*, acionando o botão *Choose* abre-se uma janela que contém os tipos de algoritmos de clusterização, esquema de agrupamento. No quadro *Cluster Mode* possui opções usadas para escolher o que agrupar e como analisar os resultados apresentados, figura 37, conforme detalhado abaixo:

- *Use training set*: o algoritmo de clusterização classifica as instâncias de treinamento em cluster de acordo com a representação do cluster e calcula a porcentagem de instâncias que caem em cada *cluster*;
- *Supplied test set*: o algoritmo de clusterização é avaliado em quão bem ele agrupa a classe de um conjunto de instâncias carregadas de um arquivo;
- *Percentage split*: o algoritmo de clusterização é avaliado com base no quão bem ele agrupa uma certa porcentagem dos dados que são apresentados para teste. A quantidade de dados apresentados depende do valor inserido no campo %;

- *Classes to clusters evaluation*: o algoritmo de clusterização compara o quão bem os *clusters* escolhidos correspondem a uma classe pré-atribuída nos dados. A caixa suspensa abaixo dessa opção seleciona a classe;
- *Store clusters for visualization*: após a execução do algoritmo de clusterização é possível visualizar os *clusters* após a conclusão do treinamento.

No botão *Ignore attributes* abre-se uma janela que permite selecionar quais atributos serão ignorados durante a execução do algoritmo de clusterização, e o botão *Start* inicia a execução do algoritmo de clusterização. No quadro *Clusterer output* mostra o resultado da execução do algoritmo de clusterização, representação dos dados, figura 37.

Figura 37. WEKA *Explorer*, aba *Cluster* - WEKA



Fonte: SANTOS, PEREIRA (2020)

Após selecionar o algoritmo de clusterização a ser utilizado, no quadro *Clusterer*, é exibido o nome do algoritmo e ao clicar no campo que exibe no nome do algoritmo, ao lado do botão *Choose* é exibida a janela weka.gui.GenericObjectEditor que permite editar o modo de execução do algoritmo de clusterização, contendo as seguintes opções, com descrição detalhada fornecida pela ferramenta de cada opção, conforme segue a seguir:

- *seed*: o valor inicial do número aleatório a ser usado;
- *displayStdDevs*: exibe desvios padrão de atributos numéricos e contagens de atributos nominais;
- *numExecutionSlots*: o número de *slots* de execução (*threads*) a serem usados. Definir igual ao número de CPU / núcleos disponíveis;
- *canopyMinimumCanopyDensity*: se usar agrupamento de *canopy* para inicialização e / ou aumento de velocidade, esta é a densidade baseada em T2 mínima abaixo da qual um *canopy* será podado durante a poda periódica;
- *dontReplaceMissingValues*: substitua os valores ausentes globalmente pela média / modo;
- *debug*: se definido como verdadeiro, o *clusterer* pode enviar informações adicionais para o *console*;
- *canopyT2*: a distância T2 a ser usada ao usar agrupamento de *canopy*. Valores <0 indicam que isso deve ser definido usando uma heurística baseada no desvio padrão do atributo;
- *numClusters*: definir o número de *clusters*;
- *doNotCheckCapabilities*: Se definido, os recursos do *clusterer* não são verificados antes que o *clusterer* seja construído;
- *maxIterations*: defina o número máximo de iterações;
- *preserveInstancesOrder*: preserve a ordem das instâncias;
- *canopyPeriodicPruningRate*: se estiver usando agrupamento de *canopy* para inicialização e / ou *speedup*, esta é a frequência com que podar os *canopies* de baixa densidade durante o treinamento;
- *canopyMaxNumCanopiesToHoldInMemory*: se estiver usando o agrupamento de *canopy* para inicialização e / ou aumento de velocidade, este é o número máximo de *canopies* candidatos a reter na memória principal durante o treinamento do *clusterer* de *canopy*. A distância T2 e as características de dados determinam quantas *canopies* candidatos são formados antes da poda periódica e final ser realizada. Pode não haver memória suficiente disponível se T2 estiver definido como muito baixo;
- *initializationMethod*: o método de inicialização a ser usado. Aleatório, *K-means ++*, *Canopy* ou mais distante primeiro;
- *distanceFunction*: a função de distância a ser usada para comparação de instâncias (padrão: weka.core.EuclideanDistance);

- *canopyT1*: A distância T1 a ser usada ao usar agrupamento de *canopy*. Valores <0 são tomados como um multiplicador positivo para a distância T2;
- *fastDistanceCalc*: usa valores de corte para acelerar o cálculo da distância, mas suprime também o cálculo e a saída da soma dos erros quadrados / soma das distâncias dentro do *cluster*;
- *reduzNumberOfDistanceCalcsViaCanopies*: use o agrupamento de *canopy* para reduzir o número de cálculos de distância realizados por *K-means*.

Para avaliação dos resultados gerados, a ferramenta WEKA dispões apenas de duas formas de visualização, quantitativa e visual (*plot*). Porém será utilizada somente a visualização quantitativa dos dados conforme as características do conjunto de dados a ser analisados e por ser compreensível o modo que é exibida a representação do conhecimento disponibilizado pela ferramenta.

3.3.3.2.1.1 Cenário 1 - Geração de 10 *clusters* no período de 2010 a 2019

Nesta subseção são demonstradas as visualizações e saída dos dados levando em consideração o primeiro cenário com ocorrências de 2010 a 2019. Para o cenário abordado foi utilizado a quantidade de 10 *clusters*, e como o gráfico de distribuição da ferramenta Orange apresenta os dados por um atributo classe, os *clusters* gerados no WEKA foi considerado somente um atributo classe por vez.

As figuras 38, 39 e 40 apresentam o resultado do quadro *Clusterer Output* considerando apenas o atributo classe ocorrencia_ano. Foi utilizada a ferramenta bloco de notas para uma melhor visualização dos dados gerados.

Figura 38. Cenário 1, *Run Information*, atributo classe ocorrencia_ano - WEKA

```
=== Run information ===

Scheme:       weka.clusterers.SimpleKMeans -init 0 -max-candidates 100 -periodic-pruning 10000 -min-density 2.0 -t1 -1.25 -t2 -1.0 -N 10 -A
"weka.core.EuclideanDistance -R first-last" -I 500 -num-slots 1 -S 10
Relation:     QueryResult
Instances:    988
Attributes:   4
              ocorrencia_ano
Ignored:
              fator_nome
              aeronave_fase_operacao
              ocorrencia_tipo
Test mode:    evaluate on training data
```

Fonte: SANTOS, PEREIRA (2020)

Figura 39. Cenário 1, *Clustering model*, atributo classe ocorrencia_ano - WEKA

```
=== Clustering model (full training set) ===


kMeans
======

Number of iterations: 2
Within cluster sum of squared errors: 0.0

Initial starting points (random):

Cluster 0: 2013
Cluster 1: 2010
Cluster 2: 2011
Cluster 3: 2015
Cluster 4: 2019
Cluster 5: 2012
Cluster 6: 2017
Cluster 7: 2014
Cluster 8: 2016
Cluster 9: 2018

Missing values globally replaced with mean/mode
```

Fonte: SANTOS, PEREIRA (2020)

Figura 40. Cenário 1, *Model and evaluation on training set*, atributo classe ocorrencia_ano - WEKA

```
=== Model and evaluation on training set ===

Clustered Instances

0      141 ( 14%)
1      162 ( 16%)
2      149 ( 15%)
3      109 ( 11%)
4       18 (  2%)
5      140 ( 14%)
6       46 (  5%)
7      144 ( 15%)
8       48 (  5%)
9       31 (  3%)
```

Fonte: SANTOS, PEREIRA (2020)

As figuras 41, 42 e 43 apresentam o resultado do quadro *Clusterer Output* considerando apenas o atributo classe fator_nome. Foi utilizada a ferramenta bloco de notas para uma melhor visualização dos dados gerados.

Figura 41. Cenário 1, *Run Information*, atributo classe fator_nome - WEKA

```
=== Run information ===

Scheme:       weka.clusterers.SimpleKMeans -init 0 -max-candidates 100 -periodic-pruning 10000 -min-density 2.0 -t1 -1.25 -t2 -1.0 -N 10 -A
"weka.core.EuclideanDistance -R first-last" -I 500 -num-slots 1 -S 10
Relation:     QueryResult
Instances:    988
Attributes:   4
              fator_nome
Ignored:
              aeronave_fase_operacao
              ocorrencia_ano
              ocorrencia_tipo
Test mode:    evaluate on training data
```

Fonte: SANTOS, PEREIRA (2020)

Figura 42. Cenário 1, *Clustering model*, atributo classe fator_nome - WEKA

```
=== Clustering model (full training set) ===


kMeans
======

Number of iterations: 2
Within cluster sum of squared errors: 308.0

Initial starting points (random):

Cluster 0: 'JULGAMENTO DE PILOTAGEM'
Cluster 1: 'CULTURA ORGANIZACIONAL'
Cluster 2: 'CONHECIMENTO DE NORMAS (ATS)'
Cluster 3: 'INTOXICAÇÃO ALIMENTAR'
Cluster 4: 'ILUSÕES VISUAIS'
Cluster 5: 'CARACTERÍSTICAS DA TAREFA'
Cluster 6: PERCEPÇÃO
Cluster 7: ENFERMIDADE
Cluster 8: 'EQUIPAMENTO DE APOIO (ATS)'
Cluster 9: INSTRUÇÃO

Missing values globally replaced with mean/mode
```

Fonte: SANTOS, PEREIRA (2020)

Figura 43. Cenário 1, *Model and evaluation on training set*, atributo classe fator_nome- WEKA

```
=== Model and evaluation on training set ===

Clustered Instances

0      466 ( 47%)
1       61 (  6%)
2        4 (  0%)
3      133 ( 13%)
4       21 (  2%)
5       76 (  8%)
6       14 (  1%)
7       57 (  6%)
8      152 ( 15%)
9        4 (  0%)
```

Fonte: SANTOS, PEREIRA (2020)

As figuras 44, 45 e 46 apresentam o resultado do quadro *Clusterer Output* considerando apenas o atributo classe aeronave_fase_operacao. Foi utilizada a ferramenta bloco de notas para uma melhor visualização dos dados gerados.

Figura 44. Cenário 1, *Run Information*, atributo classe aeronave_fase_operacao - WEKA

```
=== Run information ===

Scheme:       weka.clusterers.SimpleKMeans -init 0 -max-candidates 100 -periodic-pruning 10000 -min-density 2.0 -t1 -1.25 -t2 -1.0 -N 10 -A
"weka.core.EuclideanDistance -R first-last" -I 500 -num-slots 1 -S 10
Relation:     QueryResult
Instances:    988
Attributes:   4
              aeronave_fase_operacao
Ignored:
              fator_nome
              ocorrencia_ano
              ocorrencia_tipo
Test mode:    evaluate on training data
```

Fonte: SANTOS, PEREIRA (2020)

Figura 45. Cenário 1, *Clustering model*, atributo classe aeronave_fase_operacao - WEKA

```
=== Clustering model (full training set) ===


kMeans
======

Number of iterations: 2
Within cluster sum of squared errors: 160.0

Initial starting points (random):

Cluster 0: 'CORRIDA APÓS POUSO'
Cluster 1: TÁXI
Cluster 2: DECOLAGEM
Cluster 3: 'ARREMETIDA NO AR'
Cluster 4: SUBIDA
Cluster 5: MANOBRA
Cluster 6: 'VOO A BAIXA ALTURA'
Cluster 7: CRUZEIRO
Cluster 8: ESPECIALIZADA
Cluster 9: POUSO

Missing values globally replaced with mean/mode
```

Fonte: SANTOS, PEREIRA (2020)

Figura 46. Cenário 1, *Model and evaluation on training set*, atributo classe aeronave_fase_operacao - WEKA

```
=== Model and evaluation on training set ===

Clustered Instances

0      271 ( 27%)
1       23 (  2%)
2      184 ( 19%)
3       11 (  1%)
4       45 (  5%)
5       48 (  5%)
6       23 (  2%)
7      104 ( 11%)
8       71 (  7%)
9      208 ( 21%)
```

Fonte: SANTOS, PEREIRA (2020)

As figuras 47, 48 e 49 apresentam o resultado do quadro *Clusterer Output* considerando apenas o atributo classe ocorrencia_tipo. Foi utilizada a ferramenta bloco de notas para uma melhor visualização dos dados gerados.

Figura 47. Cenário 1, *Run Information*, atributo classe ocorrencia_tipo - WEKA

```
=== Run information ===

Scheme:       weka.clusterers.SimpleKMeans -init 0 -max-candidates 100 -periodic-pruning 10000 -min-density 2.0 -t1 -1.25 -t2 -1.0 -N 10 -A
"weka.core.EuclideanDistance -R first-last" -I 500 -num-slots 1 -S 10
Relation:     QueryResult
Instances:    988
Attributes:   4
              ocorrencia_tipo
Ignored:
              fator_nome
              aeronave_fase_operacao
              ocorrencia_ano
Test mode:    evaluate on training data
```

Fonte: SANTOS, PEREIRA (2020)

Figura 48. Cenário 1, *Clustering model*, atributo classe ocorrencia_tipo - WEKA

```
=== Clustering model (full training set) ===


kMeans
======

Number of iterations: 2
Within cluster sum of squared errors: 308.0

Initial starting points (random):

Cluster 0: 'PERDA DE CONTROLE NO SOLO'
Cluster 1: 'COM TREM DE POUSO'
Cluster 2: 'COM PARA-BRISAS / JANELA / PORTA'
Cluster 3: 'PERDA DE CONTROLE EM VOO'
Cluster 4: 'OPERAÇÃO A BAIXA ALTITUDE'
Cluster 5: 'COLISÃO COM OBSTÁCULO DURANTE A DECOLAGEM E POUSO'
Cluster 6: 'VOO CONTROLADO CONTRA O TERRENO'
Cluster 7: 'EXCURSÃO DE PISTA'
Cluster 8: 'FALHA DO MOTOR EM VOO'
Cluster 9: 'PERDA DE COMPONENTE NO SOLO'

Missing values globally replaced with mean/mode
```

Fonte: SANTOS, PEREIRA (2020)

Figura 49. Cenário 1, *Model and evaluation on training set*, atributo classe ocorrencia_tipo - WEKA

```
=== Model and evaluation on training set ===

Clustered Instances

0      466 ( 47%)
1       61 (  6%)
2        4 (  0%)
3      133 ( 13%)
4       21 (  2%)
5       76 (  8%)
6       14 (  1%)
7       57 (  6%)
8      152 ( 15%)
9        4 (  0%)
```

Fonte: SANTOS, PEREIRA (2020)

Os dados dos atributos classe referentes a *Final cluster centroids* não foram demonstrados nesta subseção devido a ferramenta disponibilizar a informação organizada no sentido horizontal, assim a informação contida em uma imagem ficaria incompleta, conforme pode ser percebida na imagem 28. Os mesmos dados são quase idênticos em quantidade de informações comparado ao *Model and evaluation on training set* que são organizados em sentido vertical e não possuindo informações de porcentagem e quantidade de instâncias que possui em *Final cluster centroids*.

3.3.3.2.1.1.1 Avaliação dos Resultados

É possível observar que na geração de 10 *clusters* com um atributo classe por vez das 988 instâncias dos registros de ocorrências aeronáuticas entre os anos de 2010 a 2019 geradas pelo algoritmo, os anos com maior número de instâncias foram 2010 (162 instâncias), 2011 (149 instâncias), e 2014 (144 instâncias), estes números chegam a representar 46 % do total do

período. A baixa expressiva destes números ocorreu apenas em 2016 onde foram geradas 48 instâncias.

Em relação à fase de operação da aeronave no momento da ocorrência, as fases que mais se destacaram foram corrida após pouso, pouso e decolagem, com respectivamente 271, 208 e 184 instâncias. Quanto aos fatores contribuintes, 3 demonstraram maior frequência que os demais, são eles: Julgamento de pilotagem (466 instâncias), Equipamento de Apoio (ATS) (152 instâncias) e Intoxicação Alimentar (133 instâncias).

Para os tipos de ocorrência a maior frequência ocorre para: Perda de Controle no Solo com 466 instâncias, Falha do Motor em Voo com 152 instâncias e Perda de Controle em Voo com 133 instâncias.

3.3.3.2.1.2 Cenário 2 - Geração de 5 *clusters* no ano de 2010

De acordo com análise realizada no cenário anterior, pôde ser observado que o ano onde mais ocorreram ocorrências aeronáuticas foi em 2010. Sendo assim, o segundo cenário para análise específica, foi escolhido avaliar os fatores chaves das ocorrências de 2010. A parametrização do algoritmo foi alterada para gerar a quantidade de 5 *clusters*.

Em comparação ao cenário 1, não foi necessário executar o algoritmo de clusterização para o atributo classe ocorrencia_ano, pois está sendo analisado somente o ano de 2010.

As figuras 50, 51 e 52 apresentam o resultado do quadro *Clusterer Output* considerando apenas o atributo classe fator_nome. Foi utilizada a ferramenta bloco de notas para uma melhor visualização dos dados gerados.

Figura 50. Cenário 2, *Run Information*, atributo classe fator_nome - WEKA

```
=== Run information ===

Scheme:       weka.clusterers.SimpleKMeans -init 0 -max-candidates 100 -periodic-pruning 10000 -min-density 2.0 -t1 -1.25 -t2 -1.0 -N 5 -A "weka.core.EuclideanDistance
-R first-last" -I 500 -num-slots 1 -S 10
Relation:     QueryResult
Instances:    162
Attributes:   4
              fator_nome
Ignored:
              aeronave_fase_operacao
              ocorrencia_ano
              ocorrencia_tipo
Test mode:    evaluate on training data
```

Fonte: SANTOS, PEREIRA (2020)

Figura 51. Cenário 2, *Clustering model*, atributo classe fator_nome - WEKA

```
=== Clustering model (full training set) ===


kMeans
======

Number of iterations: 2
Within cluster sum of squared errors: 74.0

Initial starting points (random):

Cluster 0: 'INTOXICAÇÃO ALIMENTAR'
Cluster 1: INSTRUÇÃO
Cluster 2: 'CULTURA ORGANIZACIONAL'
Cluster 3: ENFERMIDADE
Cluster 4: 'EQUIPAMENTO DE APOIO (ATS)'

Missing values globally replaced with mean/mode
```

Fonte: SANTOS, PEREIRA (2020)

Figura 52. Cenário 2, *Model and evaluation on training set*, atributo classe fator_nome- WEKA

```
=== Model and evaluation on training set ===

Clustered Instances

0       96 ( 59%)
1        1 (  1%)
2       10 (  6%)
3       26 ( 16%)
4       29 ( 18%)
```

Fonte: SANTOS, PEREIRA (2020)

As figuras 53, 54 e 55 apresentam o resultado do quadro *Clusterer Output* considerando apenas o atributo classe aeronave_fase_operacao. Foi utilizada a ferramenta bloco de notas para uma melhor visualização dos dados gerados.

Figura 53. Cenário 2, *Run Information*, atributo classe aeronave_fase_operacao - WEKA

```
=== Run information ===

Scheme:       weka.clusterers.SimpleKMeans -init 0 -max-candidates 100 -periodic-pruning 10000 -min-density 2.0 -t1 -1.25 -t2 -1.0 -N 5 -A "weka.core.EuclideanDistance
-R first-last" -I 500 -num-slots 1 -S 10
Relation:     QueryResult
Instances:    162
Attributes:   4
              aeronave_fase_operacao
Ignored:
              fator_nome
              ocorrencia_ano
              ocorrencia_tipo
Test mode:    evaluate on training data
```

Fonte: SANTOS, PEREIRA (2020)

Figura 54. Cenário 2, *Clustering model*, atributo classe aeronave_fase_operacao - WEKA

```
=== Clustering model (full training set) ===


kMeans
======

Number of iterations: 2
Within cluster sum of squared errors: 48.0

Initial starting points (random):

Cluster 0: DECOLAGEM
Cluster 1: 'CORRIDA APÓS POUSO'
Cluster 2: 'RETA FINAL'
Cluster 3: POUSO
Cluster 4: CRUZEIRO

Missing values globally replaced with mean/mode
```

Fonte: SANTOS, PEREIRA (2020)

Figura 55. Cenário 2, *Model and evaluation on training set*, atributo classe aeronave_fase_operacao - WEKA

```
=== Model and evaluation on training set ===

Clustered Instances

0       94 ( 58%)
1       16 ( 10%)
2        5 (  3%)
3       33 ( 20%)
4       14 (  9%)
```

Fonte: SANTOS, PEREIRA (2020)

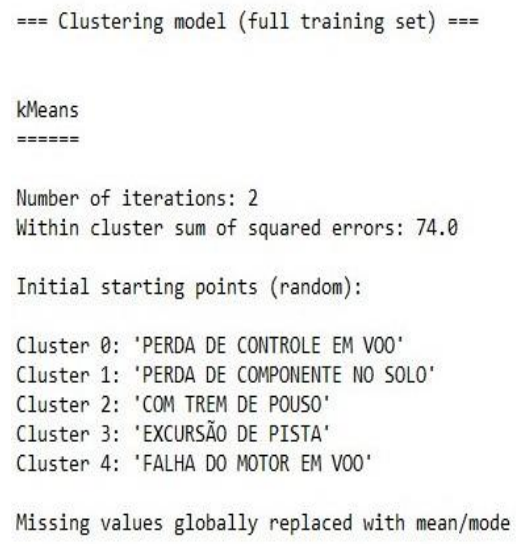
As figuras 56, 57 e 58 apresentam o resultado do quadro *Clusterer Output* considerando apenas o atributo classe ocorrencia_tipo. Foi utilizada a ferramenta bloco de notas para uma melhor visualização dos dados gerados.

Figura 56. Cenário 2, *Run Information*, atributo classe ocorrencia_tipo - WEKA

```
=== Run information ===

Scheme:       weka.clusterers.SimpleKMeans -init 0 -max-candidates 100 -periodic-pruning 10000 -min-density 2.0 -t1 -1.25 -t2 -1.0 -N 5 -A "weka.core.EuclideanDistance
-R first-last" -I 500 -num-slots 1 -S 10
Relation:     QueryResult
Instances:    162
Attributes:   4
              ocorrencia_tipo
Ignored:
              fator_nome
              aeronave_fase_operacao
              ocorrencia_ano
Test mode:    evaluate on training data
```

Fonte: SANTOS, PEREIRA (2020)

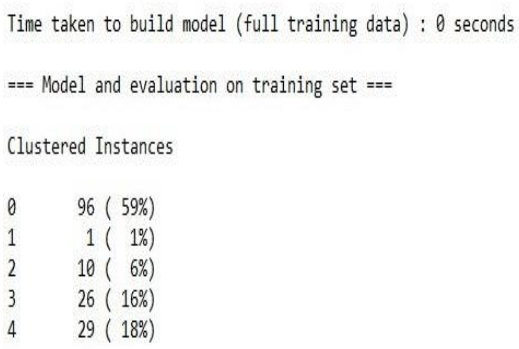
Figura 57. Cenário 2, *Clustering model*, atributo classe ocorrencia_tipo - WEKA



```
=== Clustering model (full training set) ===


kMeans
======

Number of iterations: 2
Within cluster sum of squared errors: 74.0

Initial starting points (random):

Cluster 0: 'PERDA DE CONTROLE EM VOO'
Cluster 1: 'PERDA DE COMPONENTE NO SOLO'
Cluster 2: 'COM TREM DE POUSO'
Cluster 3: 'EXCURSÃO DE PISTA'
Cluster 4: 'FALHA DO MOTOR EM VOO'

Missing values globally replaced with mean/mode
```

Fonte: SANTOS, PEREIRA (2020)

Figura 58. Cenário 2, *Model and evaluation on training set*, atributo classe ocorrencia_tipo - WEKA



```
Time taken to build model (full training data) : 0 seconds

=== Model and evaluation on training set ===

Clustered Instances

0       96 ( 59%)
1        1 (  1%)
2       10 (  6%)
3       26 ( 16%)
4       29 ( 18%)
```

Fonte: SANTOS, PEREIRA (2020)

Os dados dos atributos classe referentes *Final cluster centroids* não foram demonstrados nesta subseção devido a ferramenta disponibilizar a informação organizada no sentido horizontal, assim a informação contida em uma imagem ficaria incompleta, conforme pode ser percebida na imagem 28. Os mesmos dados são quase idênticos em quantidade de informações comparado ao *Model and evaluation on training set* que são organizados em sentido vertical e não possuindo informações de porcentagem e quantidade de instâncias que possui em *Final cluster centroids*.

3.3.3.2.1.2.1 Avaliação dos Resultados

É possível observar que na geração de 5 *clusters* com um atributo classe por vez das 162 instâncias dos registros de ocorrências aeronáuticas geradas pelo algoritmo, no que se refere a fase de operação da aeronave no momento da ocorrência, os momentos que mais se destacaram foram: Decolagem, Pouso, e Corrida após Pouso, com respectivamente 94, 33 e 16 instâncias. Quanto aos fatores contribuintes: Intoxicação Alimentar (96 instâncias),

Equipamento de Apoio (ATS) (29 instâncias) e Enfermidade (26 instâncias). Para os tipos de ocorrência a maior frequência no ano de 2010: Perda de Controle em Voo com 96 instâncias, Falha do Motor em Voo com 29 instâncias e Excursão de Pista com 26 instâncias.

3.5 Paralelo entre as ferramentas

Enfim, após análise e avaliação do processo em ambas as ferramentas, permitiu-se a indicação de alguns critérios para justificar a escolha para cada uma das ferramentas em trabalhos científicos. Abaixo segue análise do contraste entre as características, semelhanças e diferenças entre os processos de cada uma.

As ferramentas WEKA e Orange possuem documentação detalhada que explicam suas características e funcionalidades operacionais. A documentação da ferramenta Orange é visualmente organizada, facilitando a leitura e a navegação em busca de outras informações na documentação, além de possuir exemplos de aplicação, e seus vídeos tutoriais são dinâmicos e diretos. A ferramenta WEKA possui uma documentação para cada versão da ferramenta e apêndices gratuitos, as informações das funcionalidades são detalhadas e com exemplos de aplicação, e seus vídeos tutoriais são detalhados e abrangentes. As documentações de ambas ferramentas estão disponíveis no *site* oficial de cada ferramenta na língua inglesa.

A navegação no *site* da ferramenta Orange[3] é mais ágil e visualmente elaborada para fins comerciais, e assim encontra-se com poucos passos a documentação da ferramenta. A navegação no *site* da ferramenta WEKA[4] precisa de atenção para encontrar o que procura ao acessar as outras páginas do *site* oficial, nota-se ainda que as páginas do *site* possuem uma interface limpa e objetiva, porém nota-se que foram elaboradas para um público científico, com mais detalhes e encontrando uma dificuldade para encontrar a documentação da ferramenta.

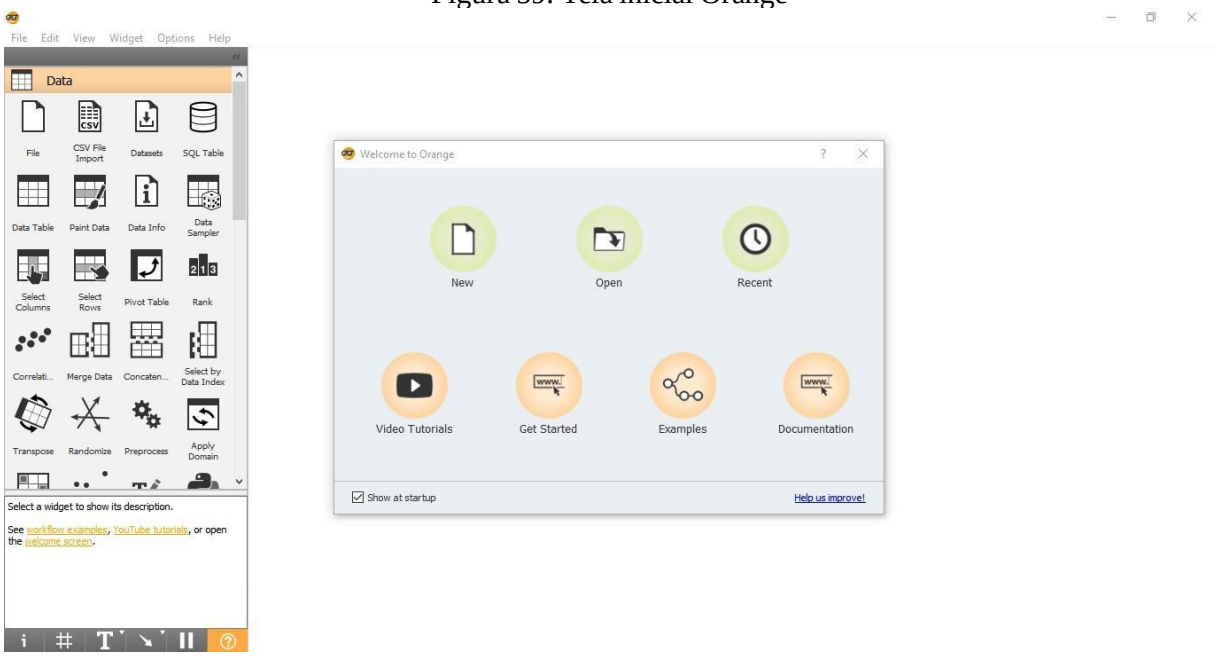
Ainda sobre a documentação da ferramenta WEKA possui uma diversidade de materiais não oficiais disponibilizados em *sites* e vídeos tutoriais de terceiros que explicam suas funcionalidades, utilização e formas de estabelecer comunicação com banco de dados. E quanto a ferramenta Orange mostrou-se dificuldade em encontrar informações adicionais da ferramenta em *sites* e vídeos tutoriais de terceiros, um exemplo foi a dificuldade enfrentada neste trabalho em estabelecer a comunicação da ferramenta com o banco de dados PostgreSQL.

A tela inicial da ferramenta Orange apresenta uma tela de bem-vindo auto explicativa que permite a criação de um novo projeto, abrir um projeto em andamento ou recém aberto, acessar vídeos tutoriais, exemplos pré-definidos ou acessar a documentação. No lado esquerdo

[3] Disponível em: https://orange.biolab.si/
[4] Disponível em: https://www.cs.waikato.ac.nz/ml/weka/

é possível ver diversas categorias que agrupam as funcionalidades da ferramenta, os chamados *widgets*. Assim demonstra-se logo no primeiro acesso à ferramenta um visual mais intuitivo de como utilizar a ferramenta.

Figura 59. Tela inicial Orange



Fonte: SANTOS, PEREIRA (2020)

A tela inicial da ferramenta WEKA é limpa e objetiva para quem possui um conhecimento prévio da ferramenta. A tela inicial no primeiro acesso a ferramenta fornece dificuldade para um iniciante compreender por onde deve iniciar a configuração e uso da ferramenta, conforme Figura 30 já apresentada.

Conforme relatado sobre as características das documentações, funcionalidades e usabilidades das ferramentas Orange e WEKA, notou-se uma diferença na forma que é exibida a representação do conhecimento por cada ferramenta. A ferramenta Orange fornece os resultados visuais através dos diversos *widgets* de visualização em formas de gráficos necessitando pouca parametrização do algoritmo de clusterização para gerar os gráficos através dos *widgets*. Assim percebe-se que resultados apresentados pela ferramenta são rápidos e com características de relatórios empresariais, o que não impede de a ferramenta ser utilizada em estudos científicos, porém houve dificuldade em compreender os resultados apresentados de forma visual pela ferramenta.

Contudo a ferramenta WEKA possui mais opções de parametrização do algoritmo de clusterização e apresenta duas formas de visualização dos resultados gerados pelo algoritmo de clusterização que são de forma quantitativa e textual, e gráfico de dispersão, sendo o relatório quantitativo e textual possuir uma maior facilidade de compreensão. Os resultados apresentados

são de interesse do público científico pois apresentam mais informações e detalhes de performance podendo ainda serem inseridos em outros *softwares*, como planilha Excel.

A consulta das informações da base de dados armazenada no banco de dados PostgreSQL foi utilizada a mesma *query*, entretanto os resultados apresentados no final da execução do algoritmo de clusterização de cada ferramenta divergem em algumas informações. No cenário 1 que foi gerado 10 *clusters* no período de 2010 a 2019, os anos que apresentaram mais ocorrências aeronáuticas foram consecutivamente 2010, 2011 e 2014, ambas ferramentas apresentaram os mesmos resultados. Porém as fases de operação que mais se destacaram pela ferramenta Orange foram consecutivamente Pouso, Decolagem e Corrida após Pouso; já pela ferramenta WEKA foram consecutivamente Corrida após Pouso, Pouso e Decolagem. Pode-se visualizar as divergências dos resultados de ambas ferramentas no quadro 6. Os registros dos resultados foram inseridos de forma decrescente, de maior para menor quantidade de instância dos registros dos atributos classes.

Quadro 6. Divergência de resultados

Cenário	Atributo Classe	Orange	WEKA	Resultados
1	ocorrencia_ano	2010	2010	convergem
		2011	2011	convergem
		2014	2014	convergem
	aeronave_fase_operacao	Pouso**	Corrida após pouso	divergem
		Decolagem	Pouso**	divergem
		Corrida após pouso	Decolagem	divergem
	fator_nome	Julgamento de Pilotagem*	Julgamento de Pilotagem*	convergem
		Equipamento de Apoio (ATS)	Equipamento de Apoio (ATS)	convergem
		Intoxicação Alimentar	Intoxicação Alimentar	convergem
	ocorrencia_tipo	Perda de Controle no Solo	Perda de Controle no Solo	convergem
		Falha do Motor em Voo	Falha do Motor em Voo	convergem
		Perda de Controle em Voo	Perda de Controle em Voo	convergem

2	aeronave_fase_operacao	Decolagem*	Decolagem*	convergem
		Pouso	Pouso	convergem
		Corrida após Pouso	Corrida após Pouso	convergem
	fator_nome	Equipamento de Apoio (ATS)**	Intoxicação Alimentar	divergem
		Enfermidade**	Equipamento de Apoio (ATS)**	divergem
		Intoxicação Alimentar	Enfermidade**	divergem
	ocorrencia_tipo	Falha do Motor em Voo**	Perda de Controle em Voo	divergem
		Excursão de Pista**	Falha do Motor em Voo**	divergem
		Perda de Controle em Voo	Excursão de Pista**	divergem

Legenda:
*: quantidades de instâncias diferentes.
**: quantidades de instâncias iguais.

Fonte: SANTOS, PEREIRA (2020)

Conforme os dados apresentados no quadro 6, levando em consideração os três primeiros resultados dos atributos classes com maior quantidade de instâncias que foi gerada pelo algoritmo, 12 resultados convergem e 9 divergem. Por fim, calcula-se que aproximadamente 57 % dos resultados analisados, no quadro 6, convergem em ambas ferramentas e aproximadamente 43 % dos resultados analisados divergem em ambas ferramentas.

No quadro 7 são apresentadas algumas características das ferramentas que foram notadas durante a execução deste trabalho.

Quadro 7. Paralelo entre as ferramentas

	Orange	WEKA
Usabilidade		
Auxílio Gráfico	Sim	Não
Intuitivo	Sim	Não
Plataforma de execução	Máquina do usuário	Máquina do usuário
Documentação	*On-line* e impresso	*On-line* e impresso
Criação de Classes dos Atributos (mais de 20 instâncias)	Manual	Automática
Versão corrente	3.27.1	3.8.4
Tutoriais em vídeos	Sim	Sim
Entrada de Dados		
Conexão com o Banco/Tabelas	Conexão ODBC	Conexão ODBC
	Documentos .csv	Documentos .csv
	Documentos .xls/.xlsx	Documentos. arrf
	Google Docs	
Necessário configuração/instalação de terceiros	Instalação psycopg2	Configuração da conexão em arquivo de instalação da ferramenta
	Instalação *backend*	
K-means		
Quantidade de campos de parametrização	Poucos	Muitos
Quantidade de *Clusters* Disponível	30 *clusters*	Sem limitação
Saída de Dados		
Alternativas de visualização dos dados	Sim	Não
Gráficos	Sim	Sim
Relatório	Não	Sim

Fonte: SANTOS, PEREIRA (2020)

4. CONSIDERAÇÕES FINAIS

O Processo de Descoberta de Conhecimento em Base de Dados com foco na Mineração de Dados permite resolver problemas de diversas áreas e analisar um problema em diversas perspectivas.

Este trabalho buscou analisar os resultados apresentados pelas ferramentas de Mineração de Dados Orange e WEKA onde foi utilizada a Técnica de Clusterização (algoritmo *K-means*), no cenário de Ocorrências Aeronáuticas Brasileiras no período de 2010 a 2019. Durante o processo do KDD procurou-se similar ao máximo os processos, como as mesmas consultas e conexão com o banco de dados, PostgreSQL.

Porém no final do processo do KDD, verificou-se que os resultados apresentados por ambas ferramentas divergem em aproximadamente 43 % levando em consideração a ordem de apresentação do resultado de cada atributo classe e suas respectivas quantidades de instâncias. Não se pode afirmar que a estrutura lógica e dados dos algoritmos de clusterização utilizadas por ambas ferramentas são idênticas, uma vez que o WEKA e Orange possuem suas particularidades, sendo o primeiro usando recursos em Java e o segundo em *Python*, no WEKA para a edição e configuração das variáveis do algoritmo há mais opções do que as disponíveis no Orange.

Portanto a escolha da ferramenta de mineração de dados pautada em escolhas pessoais influência no resultado obtido em um estudo científico, devendo ser levado em consideração parâmetros científicos oriundos de estudos realizados com cada ferramenta de mineração de dados.

4.1 Trabalhos Futuros

Sugere-se como trabalhos futuros:

- Análise assertiva dos resultados apresentados por cada ferramenta;
- Descoberta de conhecimento utilizando o mesmo cenário com aplicação de algoritmos de clusterização diferentes;
- Descoberta de conhecimento utilizando o mesmo cenário com uso de diferentes técnicas de mineração de dados para enriquecimento do conhecimento gerado;
- Descoberta de conhecimento utilizando o mesmo cenário com utilização de outras ferramentas de mineração de dados;
- Avaliação dos algoritmos escolhidos para identificar semelhanças e diferenças em cada ferramenta.

REFERÊNCIAS BIBLIOGRÁFICAS

ALBUQUERQUE, Vitor Hugo Oliveira. **Procedimentos de mineração de dados aplicados à análise de riscos.** 47 f. TCC (Graduação) - Curso de Engenharia de Computação, Instituto Tecnológico de Aeronáutica, São José dos Campos, 2008.

ALMEIDA, Victor Andrade de. **Uma metodologia para tratamento de dados de curvas de carga baseada em técnicas de inteligência artificial.** 2017.

AMARAL, Fernanda Cristina Naliato do. **Data Mining: Técnicas e aplicações para o marketing direto**. São Paulo: Berkley Brasil, 2001. 110 p.

ASSIS, Lucas Rocha Soares de. **Perfil de evasão no ensino superior brasileiro: uma abordagem de mineração de dados.** 2017.

BRANQUINHO, Lucelia Pinto; BARACHO, Renata Maria Abrantes; ALMEIDA, Mauricio Barcellos. **Modelo Para Suporte À Descoberta De Conhecimento Em Base De Dados (KDD): Aplicação Em Estratégias No Mercado De Medicina Diagnóstica.** 2015.

BERNABEU, Francisco Guirado. **Mineração de dados aplicada ao estudo do perfil de tráfegos aéreos desconhecidos.** 2004. 94 f. Tese (Doutorado) - Curso de Engenharia Eletrônica e Computação - Área de Informática, Instituto Tecnológico de Aeronáutica, São José dos Campos, 2004.

BITTENCOURT, Rogério Gonçalves. **Aspectos Básicos de Banco de Dados.** Florianópolis: Adobe Acrobat Ebookreader, 2004.

BRAMER, Max. **Undergraduate Topics in Computer Science:** Principles of Data Mining. 2. ed. New York: Springer, 2007.

BRASIL. Ministério da Defesa. Comando da Aeronáutica. Estado Maior da Aeronáutica. Centro de Investigação e Prevenção de Acidentes Aeronáuticos. **Protocolos de Investigação de Ocorrências Aeronáuticas da Aviação Civil Conduzidas pelo Estado Brasileiro: NSCA 3-13**. Brasília, 2017.

BUGNION, Pascal; MANIVANNAN, Arun; NICOLAS, Patrick R.. **Scala: Guide for Data Science Professionals.** Birmingham: Packt Publishing Ltd, 2017.

CALÇADA, Dario Brito. **Redes de regras de associação filtradas e multialvo**. 2019. Tese de Doutorado. Universidade de São Paulo.

CAMARGO, Fernando Silva Alves de. **Modelo de gestão da investigação de acidente aeronáutico**. 2010. 117 f. Dissertação (Mestrado) - Curso de Mestrado Profissional em Segurança de Aviação e Aeronavegabilidade Continuada, Programa de Pós-graduação em Engenharia Aeronáutica e Mecânica, Instituto Tecnológico de Aeronáutica, São José dos Campos, 2010.

CAMILO, Cássio Oliveira; SILVA, João Carlos da. **Mineração de dados: Conceitos, tarefas, métodos e ferramentas.** Goiânia: Universidade Federal de Goiás, 2009.

CAMPOS NETO, Cantídio de Moura. **Análise inteligente de dados em um banco de dados de procedimentos em cardiologia intervencionista.** 148 f. Tese (Doutorado) - Curso de Programa de Medicina, Tecnologia e Intervenção em Cardiologia, Instituto Dante Pazzanese de Cardiologia, Universidade de São Paulo, São Paulo, 2016.

CASTANHEIRA, Luciana Gomes. **Aplicação de técnicas de mineração de dados em problemas de classificação de padrões.** Belo Horizonte: UFMG, 2008.

CENIPA. **O que fazemos. 2019**. Disponível em: <http://www2.fab.mil.br/cenipa/index.php/o-cenipa>. Acesso em: 16 out. 2019.

CIOS, Krzysztof J; KURGAN, Lukasz A.; PEDRYCZ, Witold; SWINIARSKI, Roman W. **Data Mining: A Knowledge Discovery Approach**. New York: Springer, 2007.
CÔRTES, Sérgio da Costa; PORCARO, Rosa Maria; LIFSCHITZ, Sérgio. **Mineração de dados – Funcionalidades, Técnicas e Abordagens**, Puc-RioInf. MCC10/02, 2002.

DIAS, Fabio Rodrigo da Costa. **Mineração de dados sobre as solicitações de serviços: estudo de casos sobre a GetNinjas.** 42 f. TCC (Graduação) - Curso de Ciência da Computação, Instituto Tecnológico de Aeronáutica, São José dos Campos, 2015.

DIAS, Maria Madalena. **Parâmetros na escolha de técnicas e ferramentas de mineração de dados.** Acta Scientiarum, Maringá, v. 24, n. 6, p.1715-1725, jan. 2002.

DEMŠAR, Janez; ZUPAN, Blaž, Gregor, Leban; TOMAZ, Curk, **Orange: From Experimental Machine Learning to Interactive Data Mining,** Faculty of Computer and Information Science, University of Ljubljana, Slovenia. Dep. of molecular and Human Genetics, Baylor College of Medicine, Houston, USA, 2004.

DEMŠAR, Janez; ZUPAN, Blaž. **Orange: Data Mining Fruitful And Fun**. University of Ljubljana, Faculty of Computer and Information Science, 2012.

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

FAYYAD, Usama M.; SHAPIRO, Gregory Piatetsky; SMYTH, Padhraic; UTHURUSAMY, Ramasamy. **Advances in Knowledge Discovery and Data Mining**. MIT Press, Cambridge, MA, 1996, p. 40.

HAN, Jiawei; KAMBER, Micheline; PEI, Jian. **Data mining: concepts and techniques**. 3. ed. Waltham: Elsevier, 2012.

ICAO (Montreal). **About ICAO**. 2019. Disponível em: https://www.icao.int/about-icao/Pages/default.aspx>. Acesso em: 12 out. 2019.

ICAO, **"Aircraft Accident and Incident Investigation, Annex 13"**, 9th Edition, 2001.
LAROSE, D. T. **Discovering Knowledge in Data: An Introduction to Data Mining**. John Wiley and Sons, Inc, 2005, p. 11 e 17.

JACINTO, Adriana da Silva. **Análise da relevância semântica na seleção de atributos para a mineração de dados.** 2015. PhD Thesis. Instituto Tecnológico de Aeronáutica, São José dos Campos, Brazil.

JOHNSTONE, Mark. **Data cake.** Disponível em: <http://markjohnstone.co/data-cake/>. Acesso em: 05 nov. 2019.

MACEDO, Charles Mendes de. **Aplicação de algoritmos de agrupamento para descoberta de padrões de defeito em software JavaScript**. 2019. Tese de Doutorado. Universidade de São Paulo.

MAIA, Luiz Cláudio Gomes. **Uso de sintagmas nominais na classificação de documentos eletrônicos.** 103 f. Tese (Doutorado) - Curso de Ciência da Informação, Universidade Federal de Minas Gerais, Belo Horizonte, 2008.

MAGRANI, Eduardo. **Entre dados e robôs: Ética e privacidade na era da hiperconectividade**. Arquipélago Editorial, Porto Alegre, 2019.

MURATA, T., et al. **Mineração de dados aplicada ao estudo do perfil de tráfegos aéreos desconhecidos**. 2004. PhD Thesis. Tese de Mestrado–Instituto Tecnológico de Aeronáutica, São Jose dos Campos, SP, Brazil.

OLIVERIO, Vinícius. **Inteligência artificial aplicada ao auxílio no diagnóstico da dor pélvica crônica**. 2018. Tese de Doutorado. Universidade de São Paulo.

ORANGE. **Home.** Disponível em: <http://orange.biolab.si/>. Acesso em: 30 out. 2019.

PARREIRA, Michelle de Oliveira. **Avaliação experimental da imputação múltipla e composta de valores ausentes no processo de mineração de dados.** 124 f. Dissertação (Mestrado) - Curso de Engenharia Eletrônica e Computação, Instituto Tecnológico de Aeronáutica, São José dos Campos, 2010.

PATEL, Priti S.; DESAI, S. G. **A comparative study on data mining tools.** International Journal of Advanced Trends in Computer Science and Engineering, v. 4, n. 2, 2015.

RANGRA, Kalpana; BANSAL, K. L. **Comparative study of data mining tools.** International journal of advanced research in computer science and software engineering, v. 4, n. 6, 2014.

RAMAMOHAN, Y.; VASANTHARAO, K.; CHAKRAVARTI, C. K.; RATNAM, A. S. K. **A study of data mining tools in knowledge discovery process.** International Journal of Soft Computing and Engineering (IJSCE) ISSN, v. 2, n. 3, p. 2231-2307, 2012.

RAUTENBERG, Sandro; CARMO, Paulo Ricardo Viviurka do. **Big Data e Ciência de Dados: complementariedade conceitual no processo de tomada de decisão.** Brazilian Journal Of Information Science. Marilia, p. 56-67. 2019.

RECHE, Evandro Agostinho. **Metodologia baseada em mineração de dados para redução de múltipla estimação na localização de faltas em alimentadores de distribuição radiais**. 2018. Tese de Doutorado. Universidade de São Paulo.7

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 Sistemas de Informação e Gestão do Conhecimento, Universidade Fumec, Belo Horizonte, 2014.

REZENDE, Solange Oliveira. **Sistemas Inteligentes: Fundamentos e Aplicações**. Barueri: Manole Ltda, 2003, p. 317.

ROCHA, Santiago Meireles. **Mineração de dados aplicada à classificação dos contribuintes de ICMS da SEFAZ-GO.** 76 f. Dissertação (Mestrado) - Curso de Engenharia de Produção e Sistemas, Pontifícia Universidade Católica de Anápolis, Goiânia, 2017.

SANTANA JÚNIOR, Wilton Moreira de. **Mineração em dados do ENEM para a predição do desempenho acadêmico no âmbito da Rede Federal de Educação Tecnológica.** 2018. Master's Thesis. Universidade Federal de Pernambuco.

SANTOS, Tatiane Gomes dos. **Análise De Opiniões Utilizando Técnicas De Mineração De Dados Em Redes Sociais**. Estudo De Caso: Twitter. 2017.

SILVA, Renan Monteiro da. **Modelo de mineração de dados em bases de dados acadêmicas.** 2016.

SILVA FILHO, Rogério Luiz Cardoso. **Modelo de análise e predição do desempenho dos alunos dos Institutos Federais de Educação usando o ENEM como indicador de qualidade escola.** 93 f. Dissertação (Mestrado) - Curso de Ciência da Computação, Universidade Federal de Pernambuco, Recife, 2017.

SOUSA, Jeovane Vicente de. **Computação em nuvem no contexto das smart grids: uma aplicação para auxílio à localização de faltas em sistemas de distribuição**. 2018. Tese de Doutorado. Universidade de São Paulo.

SOUSA, Marília Maria Bastos de Araújo Cavalcanti Feitoza Fava de. **Mineração de Dados Educacionais: Previsão de Notas Parciais Utilizando Classificação.** 2017. 82 f. Dissertação (Mestrado) - Curso de Informática, Universidade Federal do Amazonas, Manaus, 2017.

SOUZA, Raul de. **Guia técnico de ação inicial de investigação de acidentes aeronáuticos com aeronaves de asas fixas de acordo com técnicas recomendadas internacionalmente.** 2012. 206 f. Dissertação (Mestrado) - Curso de Mestrado Profissional em Segurança de Aviação e Aeronavegabilidade Continuada, Programa de Pós-graduação em Engenharia Aeronáutica e Mecânica, Instituto Tecnológico de Aeronáutica, São José dos Campos, 2012.

TAMAE, Rodrigo Yoshio. **Técnicas de Mineração de Dados em Educação Híbrida desenvolvida segundo a abordagem CCS.** 2018.

VIEIRA, Raphael dos Santos Guedes. **Descoberta De Conhecimento Na Relação Entre Acidentes De Trânsito Rodoviário E Fatores Climáticos, No Eixo Goiânia-distrito Federal.** 2018.

WAHBEH, Abdullah H., AL-RADAIDEH, Qasem A., AL-KABI, Mohammed N., AL-SHAWAKFA, Emad M. A comparison study between data mining tools over some classification methods. **International Journal of Advanced Computer Science and Applications**, v. 8, n. 2, p. 18-26, 2011.

WAIKATO, Department Of Computer Science - University Of. **Machine Learning at Waikato University.** Disponível em: <https://www.cs.waikato.ac.nz/ml/index.html>. Acesso em: 20 out. 2019.

WINTER, Rogério; FORSTER, Carlos Henrique Quartucci. **Método para identificar intrusão por anomalia em host com o sistema operacional Windows™ usando o algoritmo de método de boosting.** 2010.

WITTEN, Ian H.; FRANK, Eibe. **Data Mining: practical machine learning tools and techniques**. 2. ed. Boston: Morgan Kaufmann, 2005.

APÊNDICE A - *QUERY* DE CRIAÇÃO DO BANCO DE DADOS E SUAS TABELA

```
CREATE DATABASE CENIPA;

CREATE TABLE public.ocorrencia (
    codigo_ocorrencia integer NOT NULL primary key,
    ocorrencia_classificacao character varying(15) COLLATE pg_catalog."default",
    ocorrencia_latitude character varying(30) COLLATE pg_catalog."default",
    ocorrencia_longitude character varying(30) COLLATE pg_catalog."default",
    ocorrencia_cidade character varying(35) COLLATE pg_catalog."default",
    ocorrencia_uf character varying(15) COLLATE pg_catalog."default",
    ocorrencia_pais character varying(6) COLLATE pg_catalog."default",
    ocorrencia_aerodromo character varying(15) COLLATE pg_catalog."default",
    ocorrencia_dia character varying(15) COLLATE pg_catalog."default",
    ocorrencia_mes character varying(15) COLLATE pg_catalog."default",
    ocorrencia_ano character varying(15) COLLATE pg_catalog."default",
    ocorrencia_periodo character varying(10) COLLATE pg_catalog."default",
    investigacao_aeronave_liberada character varying(15) COLLATE pg_catalog."default",
    investigacao_status character varying(15) COLLATE pg_catalog."default",
    divulgacao_relatorio_numero character varying(40) COLLATE pg_catalog."default",
    divulgacao_relatorio_publicado character varying(3) COLLATE pg_catalog."default",
    divulgacao_dia_publicacao character varying(15) COLLATE pg_catalog."default",
    divulgacao_mes_publicacao character varying(15) COLLATE pg_catalog."default",
    divulgacao_ano_publicacao character varying(15) COLLATE pg_catalog."default",
    total_recomendacoes integer,
    total_aeronaves_envolvidas integer,
    ocorrencia_saida_pista character varying(3) COLLATE pg_catalog."default"
);

CREATE TABLE public.aeronave (
    codigo_aeronave integer NOT NULL primary key,
    aeronave_matricula character varying(10) COLLATE pg_catalog."default",
    aeronave_operador_categoria character varying(30) COLLATE pg_catalog."default",
    aeronave_tipo_veiculo character varying(15) COLLATE pg_catalog."default",
    aeronave_fabricante character varying(50) COLLATE pg_catalog."default",
    aeronave_modelo character varying(30) COLLATE pg_catalog."default",
    aeronave_tipo_icao character varying(15) COLLATE pg_catalog."default",
    aeronave_motor_tipo character varying(15) COLLATE pg_catalog."default",
    aeronave_motor_quantidade character varying(20) COLLATE pg_catalog."default",
    aeronave_pmd integer,
    aeronave_pmd_categoria integer,
    aeronave_assentos character(15) COLLATE pg_catalog."default",
    aeronave_ano_fabricacao character(15) COLLATE pg_catalog."default",
    aeronave_pais_fabricante character varying(30) COLLATE pg_catalog."default",
    aeronave_pais_registro character varying(30) COLLATE pg_catalog."default",
    aeronave_registro_categoria character varying(30) COLLATE pg_catalog."default",
    aeronave_registro_segmento character varying(30) COLLATE pg_catalog."default",
    aeronave_voo_origem character varying(100) COLLATE pg_catalog."default",
    aeronave_voo_destino character varying(100) COLLATE pg_catalog."default",
```

```
    aeronave_fase_operacao character varying(40) COLLATE pg_catalog."default",
    aeronave_tipo_operacao character varying(15) COLLATE pg_catalog."default",
    aeronave_nivel_dano character varying(15) COLLATE pg_catalog."default",
    aeronave_fatalidades_total integer
);

CREATE TABLE public.fator_contribuinte (
    codigo_fator_contribuinte integer NOT NULL primary key,
    fator_nome character varying(50) COLLATE pg_catalog."default",
    fator_aspecto character varying(35) COLLATE pg_catalog."default",
    fator_condicionante character varying(40) COLLATE pg_catalog."default",
    fator_area character varying(35) COLLATE pg_catalog."default"
);

CREATE TABLE public.ocorrencia_tipo (
    codigo_ocorrencia_tipo integer NOT NULL primary key,
    ocorrencia_tipo character varying(80) COLLATE pg_catalog."default",
    ocorrencia_tipo_categoria character varying(100) COLLATE pg_catalog."default",
    taxonomia_tipo_icao character varying(15) COLLATE pg_catalog."default"
);

CREATE TABLE public.ocorrencia_aeronave (
    codigo_ocorrencia_aeronave integer NOT NULL primary key,
    codigo_ocorrencia integer NOT NULL,
    codigo_aeronave integer NOT NULL,

        FOREIGN KEY (codigo_ocorrencia) REFERENCES ocorrencia (codigo_ocorrencia),
        FOREIGN KEY (codigo_aeronave) REFERENCES aeronave (codigo_aeronave)
);

CREATE TABLE public.ocorrencia_fator_contribuinte (
    codigo_ocorrencia_fator_contribuinte integer primary key NOT NULL,
    codigo_ocorrencia integer NOT NULL,
    codigo_fator_contribuinte integer NOT NULL,

        FOREIGN KEY (codigo_ocorrencia) REFERENCES ocorrencia (codigo_ocorrencia),
        FOREIGN KEY (codigo_fator_contribuinte) REFERENCES fator_contribuinte
(codigo_fator_contribuinte)
);

CREATE TABLE public.ocorrencia_ocorrencia_tipo (
    codigo_ocorrencia_ocorrencia_tipo integer NOT NULL primary key,
    codigo_ocorrencia integer NOT NULL,
    codigo_ocorrencia_tipo integer NOT NULL,

        FOREIGN KEY (codigo_ocorrencia) REFERENCES ocorrencia (codigo_ocorrencia),
        FOREIGN KEY (codigo_ocorrencia_tipo) REFERENCES ocorrencia_tipo
(codigo_ocorrencia_tipo))
```

APÊNDICE B - *QUERY* DE IMPORTAÇÃO DOS DADOS DOS ARQUIVOS CSV PARA AS TABELAS DO BANCO DE DADOS

```
COPY ocorrencia FROM 'C:\Users\Bruno\Documents\_Ketlen\TCC\Base\_limpa\ocorrencia.csv'
DELIMITER ';'
CSV HEADER;

COPY aeronave FROM 'C:\Users\Bruno\Documents\_Ketlen\TCC\Base\_limpa\aeronave.csv'
DELIMITER ';'
CSV HEADER;

COPY fator_contribuinte FROM
'C:\Users\Bruno\Documents\_Ketlen\TCC\Base\_limpa\fator_contribuinte.csv'
DELIMITER ';'
CSV HEADER;

COPY ocorrencia_tipo FROM
'C:\Users\Bruno\Documents\_Ketlen\TCC\Base\_limpa\ocorrencia_tipo.csv'
DELIMITER ';'
CSV HEADER;

COPY ocorrencia_aeronave FROM
'C:\Users\Bruno\Documents\_Ketlen\TCC\Base\_limpa\ocorrencia_aeronave.csv'
DELIMITER ';'
CSV HEADER;

COPY ocorrencia_ocorrencia_tipo FROM
'C:\Users\Bruno\Documents\_Ketlen\TCC\Base\_limpa\ocorrencia_ocorrencia_tipo.csv'
DELIMITER ';'
CSV HEADER;

COPY ocorrencia_fator_contribuinte FROM
'C:\Users\Bruno\Documents\_Ketlen\TCC\Base\_limpa\ocorrencia_ocorrencia_tipo.csv'
DELIMITER ';'
CSV HEADER;
```

APÊNDICE C - *QUERY* DE CONSULTA E PROCESSO DE KDD

Cenário 1

```
SELECT fator_contribuinte.fator_nome,
       ocorrencia.ocorrencia_ano,
       ocorrencia_tipo.ocorrencia_tipo,
       aeronave.aeronave_fase_operacao

FROM aeronave,
     fator_contribuinte,
     ocorrencia,
     ocorrencia_aeronave,
     ocorrencia_fator_contribuinte,
     ocorrencia_ocorrencia_tipo,
     ocorrencia_tipo

WHERE ocorrencia_aeronave.codigo_ocorrencia = ocorrencia.codigo_ocorrencia AND
     correncia_fator_contribuinte.codigo_ocorrencia = ocorrencia.codigo_ocorrencia AND
     ocorrencia_ocorrencia_tipo.codigo_ocorrencia = ocorrencia.codigo_ocorrencia AND
     aeronave.codigo_aeronave = ocorrencia_aeronave.codigo_aeronave AND
     fator_contribuinte.codigo_fator_contribuinte =
     ocorrencia_fator_contribuinte.codigo_fator_contribuinte AND
     ocorrencia_tipo.codigo_ocorrencia_tipo =
     ocorrencia_ocorrencia_tipo.codigo_ocorrencia_tipo
```

Cenário 2

```
SELECT fator_contribuinte.fator_nome,
       ocorrencia.ocorrencia_ano,
       ocorrencia_tipo.ocorrencia_tipo,
       aeronave.aeronave_fase_operacao

FROM aeronave,
     fator_contribuinte,
     ocorrencia,
     ocorrencia_aeronave,
     ocorrencia_fator_contribuinte,
     ocorrencia_ocorrencia_tipo,
     ocorrencia_tipo

WHERE ocorrencia_aeronave.codigo_ocorrencia = ocorrencia.codigo_ocorrencia AND
     correncia_fator_contribuinte.codigo_ocorrencia = ocorrencia.codigo_ocorrencia AND
     ocorrencia_ocorrencia_tipo.codigo_ocorrencia = ocorrencia.codigo_ocorrencia AND
     aeronave.codigo_aeronave = ocorrencia_aeronave.codigo_aeronave AND
     fator_contribuinte.codigo_fator_contribuinte =
     ocorrencia_fator_contribuinte.codigo_fator_contribuinte AND
     ocorrencia_tipo.codigo_ocorrencia_tipo =
     ocorrencia_ocorrencia_tipo.codigo_ocorrencia_tipo AND ocorrencia.ocorrencia_ano
     = '2010'
```

ANEXO A - INTEGRAÇÃO ENTRE WEKA E PGADMIN 4

Para realizar a comunicação entre a base de dados e a ferramenta WEKA, foram necessárias configurações adicionais[5], conforme descritas abaixo:

1. Salve o arquivo JDBC Driver na pasta de instalação do WEKA, o arquivo pode ser encontrado no *site* oficial do PostgreSQL;
2. Edite o arquivo RunWEKA.ini (que se encontra na mesma pasta de instalação do WEKA) incluindo o nome do driver JDBC baixado ao *Classpath*, este será utilizado para a inicialização da ferramenta. A configuração deve estar semelhante à da figura 1 abaixo:

Figura 1 - Configuração do arquivo RunWEKA.ini

```
# Version $Revision: 1.3 $

# setups (prefixed with "cmd_")
cmd_default=javaw -Dfile.encoding=#fileEncoding# -Xss#maxstack# -Djava.net.useSystemProxies=#systemProxies# #javaOpts# -classpath "#wekaja
cmd_console=cmd.exe /K start cmd.exe /K "java --add-opens java.base/java.lang=ALL-UNNAMED -Xss#maxstack# -Dfile.encoding=#fileEncoding# -D
cmd_explorer=java -Dfile.encoding=#fileEncoding# -Xss#maxstack# -Djava.net.useSystemProxies=#systemProxies# #javaOpts# -classpath "#wekaja
cmd_knowledgeFlow=java -Dfile.encoding=#fileEncoding# -Xss#maxstack# -Djava.net.useSystemProxies=#systemProxies# #javaOpts# -classpath "#we

# placeholders ("#bla#" in command gets replaced with content of key "bla")
# Note: "#wekajar#" gets replaced by the launcher class, since that jar gets
#       provided as parameter
maxstack=20m
# The MDI GUI
#mainclass=weka.gui.Main
# The GUIChooser
mainclass=weka.gui.GUIChooser
# The file encoding; use "utf-8" instead of "Cp1252" to display UTF-8 characters in the
# GUI, e.g., the Explorer
fileEncoding=Cp1252
# Use global system-wide proxies if set. Set this to false to ignore any system-wide proxy settings
systemProxies=true
# The JAVA_OPTS environment variable (if set). Can be used as an alternative way to set
# the heap size (or any other JVM option)
javaOpts=%JAVA_OPTS%
# The classpath placeholder. Add any environment variables or jars to it that
# you need for your Weka environment.
# Example with an enviroment variable (e.g., THIRD_PARTY_LIBS):
#   cp=%CLASSPATH%;%THIRD_PARTY_LIBS%
# Example with an extra jar (located at D:\libraries\libsvm.jar):
#   cp=%CLASSPATH%;D:\\\\libraries\\\\libsvm.jar
# Or in order to avoid quadrupled backslashes, you can also use slashes "/":
#   cp=%CLASSPATH%;D:/libraries/libsvm.jar
cp=%CLASSPATH%postgresql-42.2.16.jar
```

Fonte: SANTOS, PEREIRA (2020)

3. Ainda na pasta de instalação, localize o arquivo WEKA.jar e abra-o com o auxílio de um descompactador de arquivos (o arquivo não deverá ser extraído). Ao acessar o arquivo serão demonstradas algumas pastas, acesse o caminho WEKA.jar\WEKA\experiment e localize o arquivo DatabaseUtils.props.postgresql e salve-o na pasta de instalação com o nome DatabaseUtils.props (neste arquivo serão definidas as DatabaseUtils.props.postgresql configurações de acesso ao SGBD);
4. Acesse o arquivo DatabaseUtils.props com o auxílio de um editor de textos e informe no campo jdbcDriver o driver a ser utilizado e no campo jdbcURL será informado a

[5] Material produzido com base nas orientações encontradas em DEVMEDIA, disponível em: https://www.devmedia.com.br/mineracao-de-dados-no-mysql-com-a-ferramenta-weka/26360

URL de acesso à base de dados (Figura 2). Ainda pode-se ser configurado as credenciais de usuário e senha utilizado no SGBD no campo jdbcURL;

Figura 2 - Configuração do arquivo DatabaseUtils.props

```
# Database settings for PostgreSQL 7.4
#
# General information on database access can be found here:
# https://waikato.github.io/weka-wiki/databases/
#
# url:      http://www.postgresql.org/
# jdbc:     http://jdbc.postgresql.org/
# author:   Fracpete (fracpete at waikato dot ac dot nz)
# version: $Revision: 15257 $

# JDBC driver (comma-separated list)
jdbcDriver=org.postgresql.Driver

# database URL
jdbcURL=jdbc:postgresql://localhost:5432/Acidentes?user=POSTGRES&password=PSWD
```

Fonte: SANTOS, PEREIRA (2020)

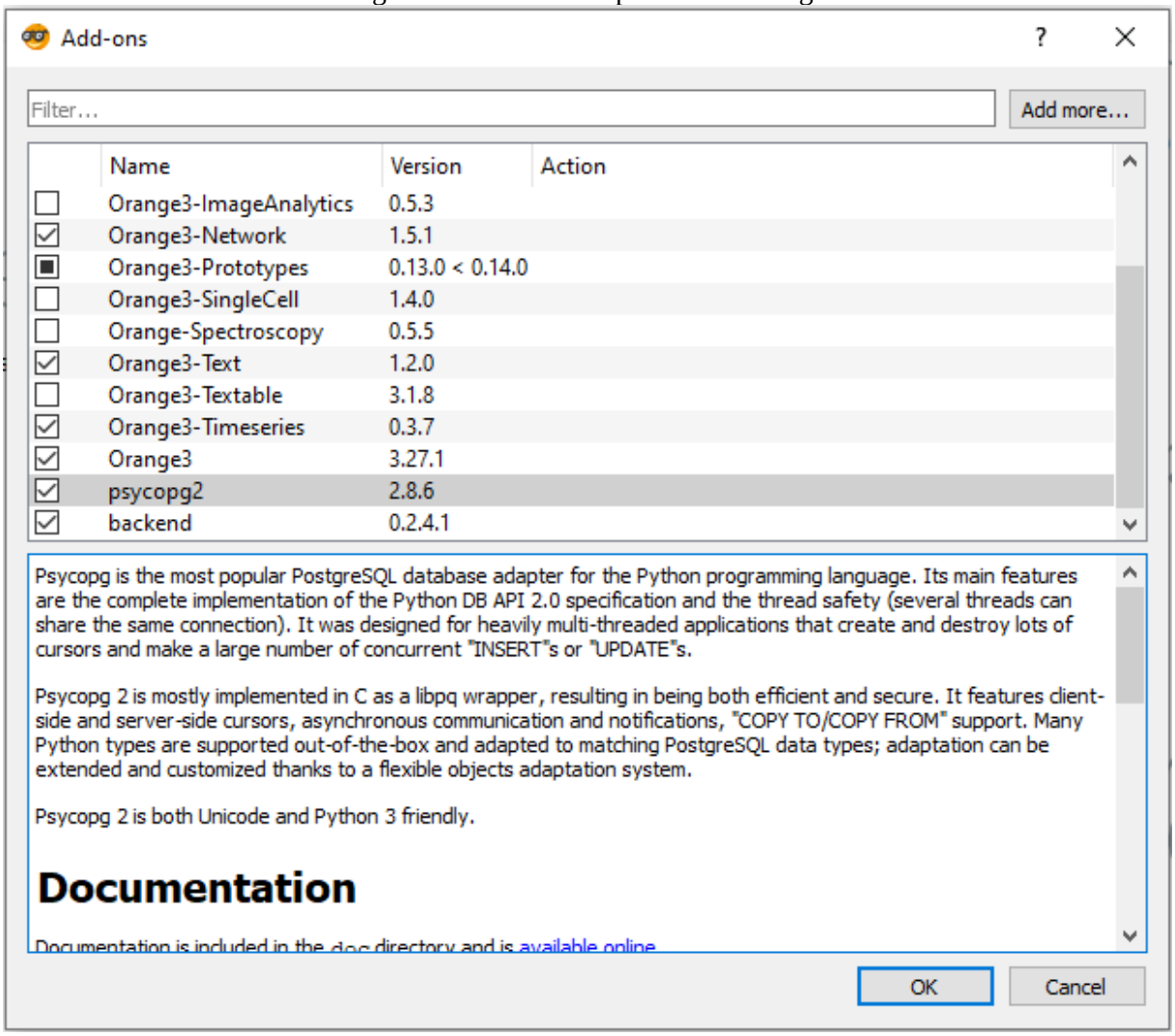
5. Após as configurações acima o WEKA poderá ser inicializado com sucesso.

ANEXO B - CONEXÃO COM A BASE DE DADOS NO ORANGE

A integração entre a base de dados e o Orange é possível a partir da configuração do ambiente instalando a biblioteca *psycopg2* em *Python* através do *site* do *pip* da *psycopg*. Após o download será necessário ir até o diretório do arquivo salvo e executar um prompt de comando da seguinte forma: "*pip install* [nome do arquivo]".

Feito isso, utilizando a opção *Add-ons* (Complementos) do Orange, será adicionado o pacote *psycopg2* ao Orange junto ao pacote *backend*, conforme figura 1 abaixo:

Figura 1 - Incluindo os pacotes ao Orange

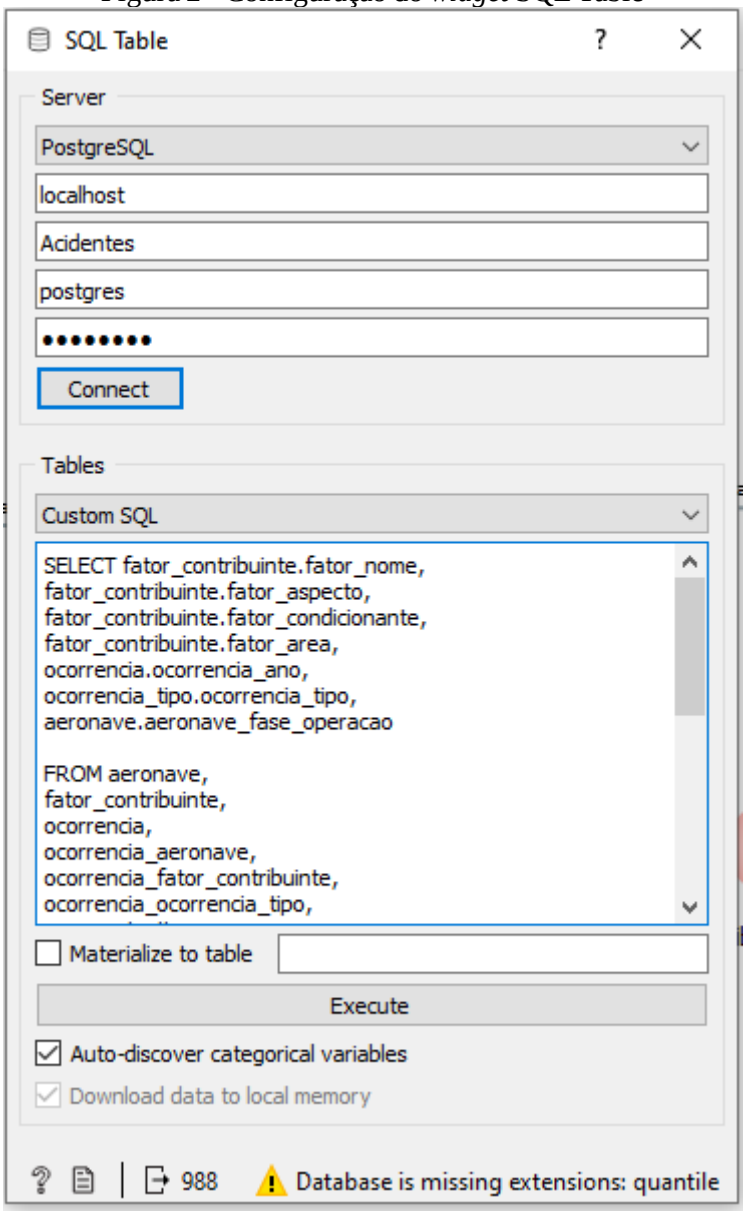


Fonte: SANTOS, PEREIRA (2020)

Configurado o ambiente e reiniciada a ferramenta, será liberada a configuração do *widget* SQL Table que servirá para realizar a conexão via ODBC com a base de dados escolhida. Ainda nesse *widget* é possível escolher entre selecionar apenas um dos atributos da base de dados ou inserir um SQL de Consulta personalizado (Figura 2). No canto inferior será demonstrado a quantidade de registros de saída após seleção da opção de consulta desejada.

Para avaliar as informações o usuário deverá utilizar de outros *widgets* como o *Select Columns* que permite visualizar as informações da forma de tabela com linhas e colunas.

Figura 2 - Configuração do *widget* SQL Table



Fonte: SANTOS, PEREIRA (2020)

A documentação oficial da configuração descrita acima está disponível em: https://orange.biolab.si/blog/2018/02/16/how-to-enable-sql-widget-in-orange/