CENTRO UNIVERSITÁRIO DE ANÁPOLIS – UniEVANGÉLICA
BACHARELADO EM ENGENHARIA DE COMPUTAÇÃO

VICTOR GABRIEL CARDOSO DE MORAIS
YAN AUGUSTO SOUSA SANTOS

APLICABILIDADE DA MINERAÇÃO DE DADOS COMO INSTRUMENTO DE ESTUDO SOCIAL COM DESTAQUE NA RELAÇÃO DO AMBIENTE ESCOLAR E A INGESTÃO DE ÁLCOOL E DROGAS ILÍCITAS

ANÁPOLIS
2020-12

VICTOR GABRIEL CARDOSO DE MORAIS
YAN AUGUSTO SOUSA SANTOS

APLICABILIDADE DA MINERAÇÃO DE DADOS COMO INSTRUMENTO DE ESTUDO SOCIAL COM DESTAQUE NA RELAÇÃO DO AMBIENTE ESCOLAR E A INGESTÃO DE ÁLCOOL E DROGAS ILÍCITAS

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

Orientador(a): Ma. Aline Dayany de Lemos.

ANÁPOLIS
2020-12

VICTOR GABRIEL CARDOSO DE MORAIS
YAN AUGUSTO SOUSA SANTOS

APLICABILIDADE DA MINERAÇÃO DE DADOS COMO INSTRUMENTO DE ESTUDO SOCIAL COM DESTAQUE NA RELAÇÃO DO AMBIENTE ESCOLAR E A INGESTÃO DE ÁLCOOL E DROGAS ILÍCITAS

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

Orientador(a): Ma. Aline Dayany de Lemos.

Aprovado(a) pela banca examinadora em **08** de dezembro de 2020, composta por:

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

_____
Profª. Altino Dantas Basílio Neto

_____
Profª. William Pereira dos Santos Júnior

Resumo

A quantidade de jovens e adolescentes envolvidas com drogas nos últimos anos cresceu substancialmente, assim como o aumento da quantidade e tipos de drogas, chegado cada vez mais fácil aos ambientes escolares. As medidas de restrição e punição não se apresentam mais eficazes e com isso a informação se tornou mais valiosa que a simples intervenção. Entretanto os dados disponíveis atualmente na *internet* vêm se multiplicando constantemente, o que promove diversos benefícios para a sociedade, mas gera alguns impactos quando não tratados de forma coerente e precisa, pois as informações desestruturadas causam dubiedade. E acerca da devida dimensão das drogas nos ambientes escolares, há que se elaborar um pensamento estruturado e preciso, gerando assim conhecimento aplicável que por meio da grande gama de dados disponíveis na *internet* e através da aplicação de técnicas por meio dos *softwares* de mineração como *Weka*, logo obtém-se resultados, indicadores e padrões para a elaboração de uma política efetiva no combate as drogas.

Palavras-Chave: Ambiente Escolar; Utilização de Drogas; Mineração de Dados; Drogas; *Weka;* Sociedade.

Abstract

The number of young people and adolescents involved in drugs in recent years has grown substantially, as has the increase in the number and types of drugs, which has reached school environments more and more easily. Restriction and punishment measures are no longer effective and as a result, information has become more valuable than simple intervention. However, the data currently available on the internet has been constantly multiplying, which promotes several benefits for society, but it generates some impacts when not treated in a coherent and accurate way, as the unstructured information causes doubt. And regarding the proper dimension of drugs in school environments, it is necessary to elaborate a structured and precise thinking, thus generating applicable knowledge that through the wide range of data available on the internet and through the application of techniques through mining software such as Weka, soon, results, indicators and standards are obtained for the elaboration of an effective policy in the fight against drugs.

Keywords: *School environment; Use of Drugs; Data Mining; Drugs; Weka; Society.*

Lista de Figuras

Lista de Quadros

Lista de Abreviaturas e Siglas

ART	Artigo
ARFF	*Attribute-Relation File Format*
CEBRID	Centro Brasileiro de Informações sobre Drogas Psicotrópicas
CSV	*Comma-separated values*
DENARC	Divisão Estadual de Narcóticos
EUA	Estados Unidos da América
FIOCRUZ	Fundação Oswaldo Cruz
FMI	Fundo Monetário Internacional
IBGE	Instituto Brasileiro de Geografia e Estatística
IC	Intervalo de Segurança
ICICT	Instituto de Comunicação e Informação Científica e Tecnológica em Saúde
JDBC	*Java DataBase Connectivity*
KDD	*Knowledge Discovery in Databases*
LI	Limites Inferiores
LS	Limites Superiores
LSD	Dietilamina do ácido lisérgico
PeNSE	Pesquisa Nacional de Saúde do Escolar
PIB	Produto Interno Bruto
PISA	Programa Internacional de Avaliação de Alunos
SENAD	Secretaria Nacional de Políticas Sobre Drogas
SGBD	Sistema de gerenciamento de banco de dados
SNC	Sistema Nervoso Central
UNESCO	Organização das Nações Unidas para a Educação, a Ciência e a Cultura
WEKA	*Wiakato Environment for Knowledge Anal*

Sumário

1. INTRODUÇÃO

De acordo com Zeitoune et al. (2010), a utilização de substâncias lícitas e ilícitas no Brasil tende a mostrar uma realidade cada vez mais comum entre os adolescentes, principalmente em função da facilidade de obtenção e ausência de informação de grande porção da população sobre os malefícios da utilização desses tipos de substâncias, como o álcool e as drogas.

Segundo o Dicionário Infopédia da Língua Portuguesa (2003), droga é "1. nome comum a todas as substâncias ou ingredientes aplicados em farmácias ou nas indústrias; 2. Substância alucinógena e que pode causar dependência química; estupefaciente; narcótico".

Já de acordo com o Centro Brasileiro de Informações sobre Drogas Psicotrópicas (CEBRID) drogas são "quaisquer substâncias capazes de modificar a função dos organismos vivos, resultando em mudanças fisiológicas ou de comportamento" (CEBRID, 2012).

É válido ressaltar que a bebida alcoólica também é considerada uma droga por ter essas mesmas características, de acordo com a Lei nº 13.106, de 17 de março de 2015, que altera a Lei nº 8.069, de 13 de julho de 1990 – Estatuto da Criança e do Adolescente, "passa a ser considerado crime vender, fornecer, servir, ministrar ou entregar bebida alcoólica a criança ou a adolescente" (Brasil, 2015).

Porém, essa não é uma prática adotada fielmente nos comércios de drogas lícitas. Além disso, segundo uma pesquisa realizada por Inácio et al. (2017, p.81), conforme demonstrado na Figura 1, aproximadamente 101 milhões de brasileiros, de um total de 209 milhões já consumiram ao menos uma dose de algum tipo de bebida alcoólica durante a vida. É válido notar que a maior parte está entre os homens, 55 milhões enquanto 46 milhões são mulheres.

Conforme a pesquisa nos 12 meses que antecederam a realização 65 milhões de brasileiros assumiram o uso de bebida alcoólica, e esse número pode ser confirmado pelo Intervalo de confiança IC 95%, que é o intervalo de segurança de 95% que existe entre os Limites inferiores (LI) e os Limites superiores (LS).

Figura 1 - Número de consumidores de 12 a 65 anos e prevalência de consumo de bebidas alcoólicas na vida, nos últimos 12 meses, nos últimos 30 dias e em binge, segundo o sexo.

Sexo	Vida				12 meses			
	Pessoas (1.000)	%	IC95%		Pessoas (1.000)	%	IC95%	
			LI	LS			LI	LS
Total	**101.615**	**66,4**	**64,8**	**68,0**	**65.943**	**43,1**	**41,8**	**44,4**
Homens	55.085	74,3	72,3	76,2	38.296	51,6	49,6	53,6
Mulheres	46.530	59,0	56,8	61,1	27.647	35,0	33,4	36,7

Fonte: (ICICT, Fiocruz. III levantamento Nacional sobre o Uso de Drogas pela População Brasileira, 2017).

Além disso, de acordo com Inácio et al. (2017, p.113) conforme a Figura 2, cerca de 15 milhões de brasileiros já fizeram o uso de drogas ilícitas no decorrer da vida, quase 5 milhões já admitiram o uso durante os últimos 12 meses e 2,5 milhões durante os últimos 30 dias que precederam o estudo.

Figura 2 - Número de consumidores e prevalência de pessoas de 12 a 65 anos que consumiram alguma substância ilícita na vida, nos últimos 12 meses e nos últimos 30 dias, segundo sexo.

Sexo	Na vida				12 meses				30 dias			
	Pessoas (1.000)	%	IC95%		Pessoas (1.000)	%	IC95%		Pessoas (1.000)	%	IC95%	
			LI	LS			LI	LS			LI	LS
Total	**15.197**	**9,9**	**9,2**	**10,6**	**4.906**	**3,2**	**2,8**	**3,6**	**2.566**	**1,7**	**1,3**	**2,0**
Homens	11.087	15,0	13,7	16,1	3.712	5,0	4,2	5,8	2.032	2,7	2,1	3,4
Mulheres	4.110	5,2	4,6	5,8	1.194	1,5	1,2	1,8	534	0,7	0,5	0,9

Fonte: (ICICT, Fiocruz. III levantamento Nacional sobre o Uso de Drogas pela População Brasileira, 2017).

É pertinente evidenciar que existem orientações associadas ao consumo e porte de drogas que são descritas no Art. 243 da Lei nº 11.343, de 23 de agosto de 2006, em que "prescreve medidas para prevenção do uso indevido, atenção e reinserção social de usuários e dependentes de drogas; estabelece normas para repressão à produção não autorizada e ao tráfico ilícito de drogas; define crimes e dá outras providências" (Brasil, 2006).

Na página da Saúde Naval, são classificadas drogas ilícitas aquelas substâncias cuja fabricação e comercialização designam crime, como a maconha, cocaína, crack, dentre outras. Já as drogas lícitas, segundo o Capitão, são aquelas que a fabricação, comercialização e o uso não são apontados como uma conduta criminosa, como o álcool e o tabaco (GONÇALVES, 2017).

Os entorpecentes como maconha, cocaína dentre outras drogas ilícitas e as drogas lícitas como bebidas alcoólicas e cigarros são em diversas ocasiões utilizadas entre os adolescentes e jovens em busca de sua aprovação em círculos sociais, como condição de atestar seu valor ou merecimento em meio aos integrantes. É possível perceber essa busca pela aceitação e consequentemente o consumo de drogas lícitas e ilícitas, como descreve Inácio et al. (2017, p.82) dessa forma essa busca é construída já na adolescência, e nota-se um aumento dos 18 aos 34 anos conforme Figura 3.

Figura 3 - Número de consumidores de 12 a 65 anos e prevalência de consumo de bebidas alcoólicas na vida, nos últimos 12 meses, segundo a faixa etária - Brasil, 2015.

Faixa etária	Vida				12 meses			
	Pessoas (1.000)	%	IC95%		Pessoas (1.000)	%	IC95%	
			LI	LS			LI	LS
Total	**101.615**	**66,4**	**64,8**	**68,0**	**65.943**	**43,1**	**41,8**	**44,4**
12 a 17 anos	6.951	34,3	30,6	38,0	4.510	22,2	19,0	25,5
18 a 24 anos	16.089	72,1	69,0	75,1	11.883	53,2	50,1	56,3
25 a 34 anos	23.587	74,5	72,0	77,1	16.434	51,9	49,5	54,3
35 a 44 anos	21.861	71,9	69,6	74,2	14.049	46,2	44,0	48,4
45 a 54 anos	18.562	70,1	67,9	72,3	11.369	43,0	40,7	45,2
55 a 65 anos	14.565	66,3	63,5	69,1	7.698	35,0	32,5	37,6

Fonte: (ICICT, Fiocruz. III levantamento Nacional sobre o Uso de Drogas pela População Brasileira, 2017).

Os dados apresentados na Figura 3 das pessoas com idades entres 18 aos 34 anos mostram um crescimento na porcentagem dos que tiveram ou mantém uso de álcool, com o ciclo para alguns adolescentes se iniciando com os 12 anos de idade, e com isso, a curva se mantém a partir dos 18 e atinge seu índice mais alto entre 25 e 34 anos, sofrendo uma leve queda após os 44 anos. Contudo o pico predominante, conforme Figura 3 mantém-se na faixa etária de 25 a 34 anos, e as demais faixas etárias após os 34 anos começam a reduzir a quantidade do consumo de bebida alcoólica, mas o consumo até os 65 anos ainda é maior que no início da fase adolescente.

A Figura 4 apresenta o nível de escolaridade nos quais ocorre o prevalecimento do uso das drogas durante toda a vida, assim nos mais afetados temos grupos sem instrução e ensino fundamental incompleto (69,5 %) e grupos com ensino médio completo e ensino superior incompleto (71,9%) . Segundo Silva et al. (2005) fatores como estrutura, renda familiar e qualidade de ensino, são considerados como evidências para o consumo de drogas e álcool, tal

prevalência de grupos menos instruídos reforçam a importância do papel escolar no enfrentamento aos problemas com drogas, tanto lícitas quanto ilícitas.

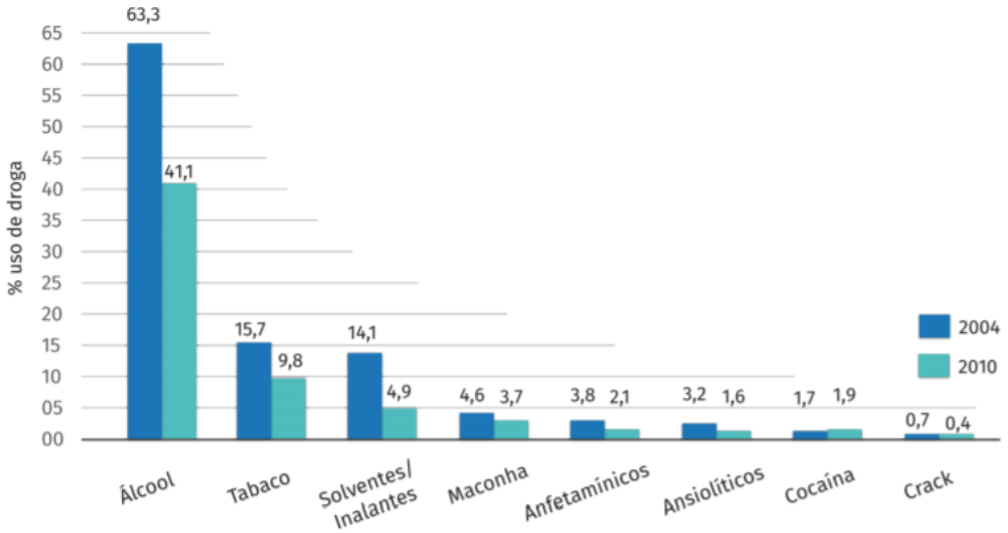
Figura 4 - Número de consumidores de 18 a 65 anos e prevalência de consumo de bebidas alcoólicas na vida, nos últimos 12 meses, segundo o nível de escolaridade.

Nível de escolaridade	Vida				12 meses			
	Pessoas (1.000)	%	IC95%		Pessoas (1.000)	%	IC95%	
			LI	LS			LI	LS
Total	**94.664**	**71,3**	**69,5**	**73,1**	**61.433**	**46,3**	**44,8**	**47,7**
Sem instrução e fundamental incompleto	30.046	69.5	67,0	72,0	16.427	38,0	36,0	40,0
Fundamental completo e médio incompleto	18.801	70.1	67,6	72,5	12.331	46,0	43,7	48,3
Médio completo e superior incompleto	34.043	71.9	69,6	74,2	23.497	49,6	47,5	51,7
Superior completo ou mais	11.774	76.5	73,4	79,5	9.178	59,6	56,4	62,8

Fonte: (ICICT, Fiocruz. III levantamento Nacional sobre o Uso de Drogas pela População Brasileira, 2017).

Conforme a Figura 5 temos um comparativa do avanço na utilização de drogas dos últimos anos.

Figura 5 - Infográfico que apresenta dados sobre o uso de drogas entre escolares brasileiros.



Fonte: SENAD e CEBRID (2003 e 2010) adaptado por NUTE-UFSC (2016).

De acordo com Cardoso e Malbergier (2012, apud CEBRID, 2004; HORTA, HORTA, PINHEIRO, MORALES, & STREY, 2007; LATIMER & ZUR, 2010; SALAZAR, UGARTE, VASQUEZ, & LOAIZA, 2004) “tanto estudos nacionais quanto internacionais têm mostrado que faltas, repetências, evasão escolar, dificuldade de aprendizagem e pouco comprometimento

com essas atividades estão associados ao uso de álcool, tabaco e drogas ilícitas entre os adolescentes".

> "A defasagem escolar dos alunos foi uma das principais consequências do uso de drogas ilícitas entre os estudantes. Aponta-se que a ausência das aulas e faltas entre os adolescentes envolvidos com drogas ilícitas, eram mais frequentes do que com adolescentes sem envolvimento" (CARDOSO E MALBERGIER; 2012, apud CEBRID, 2004).

Um dos grandes motivos para uso de drogas é a ausência de envolvimento dos alunos em eventos e tarefas extra curriculares, e segundo a Pesquisa Drogas nas Escolas realizada em parceria com Organização das Nações Unidas para a Educação, a Ciência e a Cultura (UNESCO), em que são avaliados o conjunto escolar de maneira abrangente, "e visto que, a diminuição do uso de drogas ilícitas não vem de determinações uniforme e de natureza repressiva como o aumento de câmeras, detectores de metais e fumaça nas escolas" (ABRAMOVAY; CASTRO, 2005)

Portanto a tomada de medidas rigorosas e de caráter repreensivas abruptamente são na realidade ineficazes, assim como todos os fatores que envolvam educação, que são basilares para o desenvolvimento da economia, nota-se que, o desempenho escolar brasileiro quando comparado com outras nações, ganha posição muita baixa levando em consideração a disparidade econômica, que:

> Os dados divulgados pelo Programa Internacional de Avaliação de Alunos (PISA) edição de 2018 na qual foram analisados 79 países, mostram que os resultados do Brasil são pouco animadores quando reveladas as posições, entre 58º e 60º lugar em leitura, entre 66º e 68º em ciências e entre 72º e 74º em matemática, as variações de posição ocorrem devidos a erros percentuais adotadas pela pesquisa [...] Ainda em comparação aos resultados apresentados pelo PISA alguns países onde o PIB é abaixo do brasileiro exibiram maiores escores em relação ao Brasil (PINTO, 2019).

Assim a diferença financeira acerca dos países não expressa obrigatoriamente que as maiores economias têm os maiores índices. Segundo o Fundo Monetário Internacional (FMI) o "Brasil em relação ao Produto Interno Bruto (PIB) ocupa a 8º posição em um ranking com mais de 170 nações" (FUNDO MONETÁRIO INTERNACIONAL, 2019).

Portanto uma análise dos dados disponibilizados pelo Instituto Brasileiro de Geografia e Estatística (IBGE) em levantamentos realizados por meios de censos, foi necessário para a identificação de problemas e busca de soluções para integração com o sistema de educação brasileira. De modo que, todo o contexto escolar é avaliado, analisando o consumo de drogas,

quais fatores sugerem o início da utilização de drogas ilícitas e lícitas, qual classe se mostra mais vulnerável e tem maior acesso as drogas, como o fator socioeconômico afeta a utilização e qual a relação com a influência família-escola (IBGE, 2015).

E os dados obtidos através da disponibilização eletrônica das bases de dados do governo, pelo IBGE sobre a Pesquisa Nacional de Saúde do Escolar (PeNSE) que podem ser acessados através do portal de estatísticas do IBGE[1], e que possuem diversos levantamentos, dentre alguns abordados leva-se em consideração os níveis de escolaridade por região, os fatores socioeconômicos por estado e município, fatores como núcleo familiar, diversos outros pontos de idades e qualidade de ensino são abordados (IBGE, 2015).

> "De acordo com o site Sistema de análise estatística, a mineração de dados é o processo de encontrar anomalias, padrões e correlações em grandes conjuntos de dados para prever resultados. Através de uma variedade de técnicas, você pode usar essas informações para aumentar a renda, cortar custos, melhorar o relacionamento com os clientes, reduzir riscos e mais". (SOFTWARES E SOLUÇÕES ANALYTICS 2019).

Em consideração a essa polêmica, essa pesquisa busca sugestionar melhorarias por meio da manifestação de evidências presentes nas bases de conhecimento, indicando padrões e anomalias através da mineração das informações. Por meio da utilização de algoritmos e ferramentas de mineração, possibilitando assim a geração de conhecimento.

Com o tratamento das informações sendo antecedido por uma série de transformações como a padronização, unificação e ajustes nas informações, uma vez que, a análise das informações individuais não agrega conhecimento. De modo a impossibilitar a concepção de resultados de cunho científico.

Ao longo das demais páginas é apresentada a fundamentação teórica, embasando a motivação que precedeu a elaboração de tal estudo, ainda no decorrer do documento são argumentadas as escolhas que levaram a definição de *softwares*, técnicas, algoritmos e os resultados alcançados, assim como todas as considerações para se chegar à conclusão.

[1] https://www.ibge.gov.br/estatisticas/sociais/educacao/9134-pesquisa-nacional-de-saude-do-escolar.html?=&t=downloads

2. FUNDAMENTAÇÃO TEÓRICA

Os estudos e levantamentos de tal pesquisa tem como base a necessidade do processamento de dados e análise dos conteúdos dispostos nas bases obtidas através do IBGE, de forma a politizar os efeitos das drogas e suas consequências nos ambientes escolares.

Por meio do conjunto banco de dados e ferramentas de análise, os padrões de comportamento e consumo são fortes indicadores para a elaboração de políticas de enfrentamento as drogas.

2.1 Drogas: O que são?

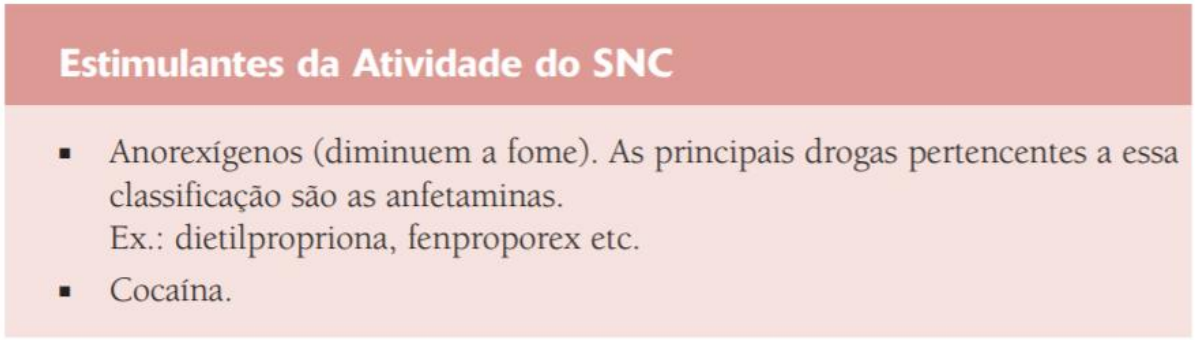
São medicamentos vendidos em drogarias ou farmácias, também são descritos como drogas, de acordo com a Divisão Estadual de Narcóticos (DENARC) “droga é o nome genérico dado a todos os tipos de substâncias, naturais ou não, que ao serem ingeridas provocam alterações físicas e psíquicas” (DENARC, [2015?]).

De acordo com o DENARC as “drogas que atuam diretamente no sistema nervoso central e causam modificações no estado mental são chamadas de drogas psicotrópicas”. As drogas ainda são classificadas em estimuladoras, perturbadoras e depressoras. As drogas estimuladoras agem no Sistema Nervoso Central (SNC), através das substâncias encontradas em seu composto, provocando um aumento dos níveis de adrenalina e consequentemente elevando o estado de vigília (DENARC, [2015?]).

> “As drogas perturbadoras modificam qualitativamente o SNC, causando em um primeiro momento sensação de bem estar, diminuição da fadiga e cansaço corporal, entretanto são responsáveis por modificar o funcionamento do cérebro levando-o a alucinações e delírios” (ABRANTES, 2018).

E finalmente as drogas depressoras que segundo o CEBRID (2012) estimulam a inatividade no SNC, provocando um desligamento dos incentivos e reduzindo seu funcionamento. Logo abaixo de acordo com as Figuras 6, 7 e 8 são apresentadas as drogas mais utilizadas de forma abusiva conforme a classificação descrita na Figura 6.

Figura 6 - Psicotrópicos estimuladores utilizados de forma abusiva.



Fonte: CEBRID (2012).

As drogas estimulantes são sobretudo tidas pela sua possibilidade de manter as pessoas que fazem uso em alerta, ativa e estimular a ausência de fome, são conhecidas ainda por sua capacidade de aumentar o funcionamento do cérebro durante seu período no organismo, causando fortes perdas dessas funções ao seu término no corpo (CEBRID, 2012).

Uma das drogas que constam nessa classe são as anfetaminas de acordo com a Figura 6 e no Brasil alguns medicamentos que possuem em sua fórmula anfetaminas são comercializados, conforme a Figura 7 abaixo, facilitando o acesso aos jovens e adolescentes.

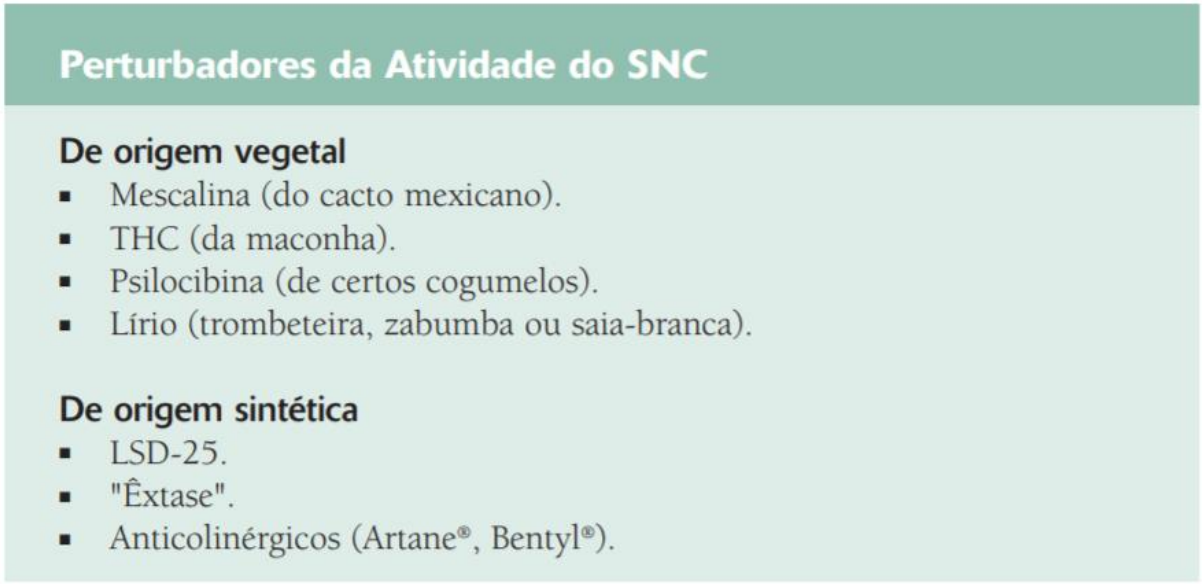
Figura 7 - Nomes comerciais na venda de remédios à base de anfetaminas vendidos nas farmácias.

Anfetamina	Produtos (nomes comerciais) vendidos em farmácias
Dietilpropiona ou Anfepramona	Dualid S®; Hipofagin S®; Inibex S®; Moderine®
Fenproporex	Desobesil-M®
Mazindol	Fagolipo®; Absten-Plus®
Metanfetamina	Pervitin®*
Metilfenidato	Ritalina

Fonte: CEBRID (2012).

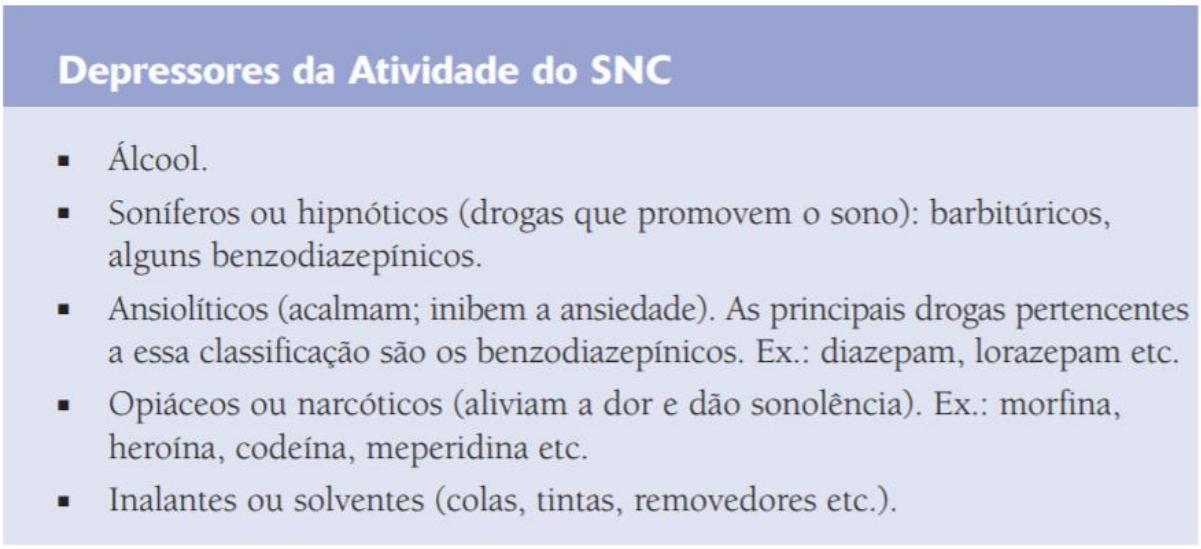
Segundo CEBRID (2012) os psicotrópicos perturbadores pelo contrário dos estimuladores não são de livre comercialização, e grande parte são proibidos, o mais comum a maconha chamada cientificamente de *Cannabis Sativa* foi proibida em praticamente todo o mundo ocidental nos últimos 60 anos devido a sua grande capacidade de causar malefícios ou danos à saúde, no entanto pesquisa recentes vem demostrando seus benefícios em vários tratamentos. Outra droga do mesmo grupo acima o Dietilamina do ácido lisérgico (LSD-25) uma droga sintética criada em laboratório e que gera alucinações, do mesmo modo proibida em todo o território nacional, tem o seu uso mais frequente entre os usuários mais favorecidos, raramente é produzida no Brasil vindo em sua maioria do exterior.

Figura 8 - Psicotrópicos perturbadores utilizados de forma abusiva.



Fonte: CEBRID (2012).

Figura 9 - Psicotrópicos depressores utilizados de forma abusiva.



Fonte: CEBRID (2012).

Diferente das demais drogas apresentadas as depressoras são amplamente utilizadas nas mais inúmeras áreas da sociedade, dentre elas segundo a Figura 9 temos os calmantes, o álcool, morfina o ópio entre outros, segundo CEBRID (2012) a utilização do álcool é fortemente aceita pela sociedade e em alguns casos incentivada, por se tratar de uma droga lícita para venda no livre comércio.

O álcool a morfina e o ópio são de uso liberado, contudo a morfina e medicamentos que tenham ópio em sua preparação são liberados exclusivamente por meio de prescrição médica,

mas são facilmente encontrados em farmácias, conforme a Figura 10 a quantidade de drogas comercializadas com tais compostos e muito grande, tornando mais acessível.

Figura 10 - Medicamentos encontrados em farmácias a base de ópio e outros compostos opiáceos.

Opiáceo ou opióide	Indicação de uso médico	Nomes comerciais dos medicamentos	Preparações farmacêuticas
Naturais		**Naturais**	
Morfina	Analgésico	Dimorf Morfina	Ampola; comprimidos
Pó de ópio	Antidiarréico; Analgésico	Tintura de ópio; Elixir paregórico; Dover	Elixir de tintura alcoólica
Codeína	Antitussígeno	Belacoclid; Belpar; CodeinCodelasa; Binelli; Naquinto; Setux; Tussaveto; Tussodina; Tylex; Pastilhas Veabon; Pastilhas Warton; Benzotiol	Gotas; comprimidos; supositórios
Sintéticos			
Meperidina ou Petidina	Analgésico	Dolantina; Demerol; Meperidina	Ampolas; comprimidos
Propoxifeno	Analgésico	Algafan; Doloxene A; Febutil; Previum Compositum; Femidol	Ampolas; comprimidos
Fentanil	Analgésico	Fentanil; Inoval	Ampolas
Semi-sintético		**Semi-sintético**	
Heroína	Proibido o uso médico	Metadon	
Metadona	Tratamento de dependentes de morfina e heroína		

Fonte: CEBRID (2012 Apud Dicionário de Especialidades Farmacêuticas – DEF 2002/2003).

2.1.1 O Consumo de Drogas

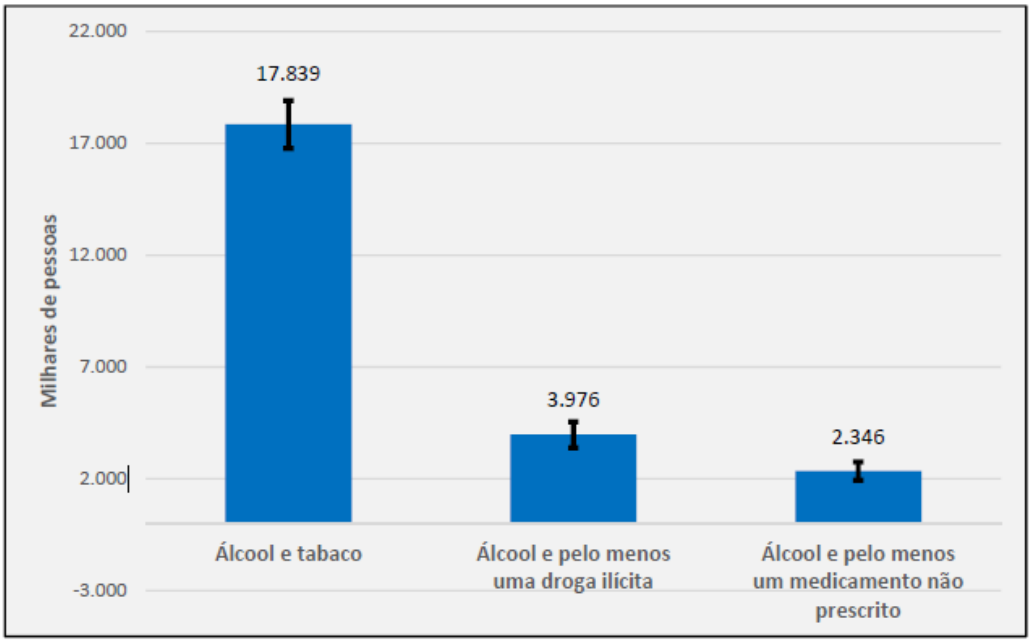
> "Os fatores que influenciam a dependência de drogas possuem três eixos de origem o sujeito com suas características e singularidade biológica; a substância psicoativa (droga), com propriedades farmacológicas específicas; e o contexto sociocultural (meio ambiente) no qual se realiza o encontro entre sujeito e droga" (SILVEIRA; DOERING-SILVEIRA, 2017).

O contexto sociocultural conforme citado acima tem uma influência importante na procura pelo uso de drogas. O exemplo familiar em relação às drogas, estimulam à busca pelo

uso ou distanciamento dessas substâncias. Assim o uso de drogas está associado a diversos infortúnios sociais, de saúde e segurança pública, sendo primordial a melhoria, tanto na terapia do dependente, como na perspectiva de diminuir a busca (COLÉGIO WEB, 2012).

Logo os aspectos sociais e socioeconômicos podem cooperar para a busca e utilização de drogas como forma de salvação ou encorajamento pela classe, "no Brasil em torno de 11,7% dos brasileiros de 12 a 65 anos ingeriram álcool e tabaco, 2,6% usou álcool e pelos menos uma substância ilícita e 1,5% consumiu álcool e algum fármaco não receitado nos últimos 12 meses" (CEBRID, 2012).

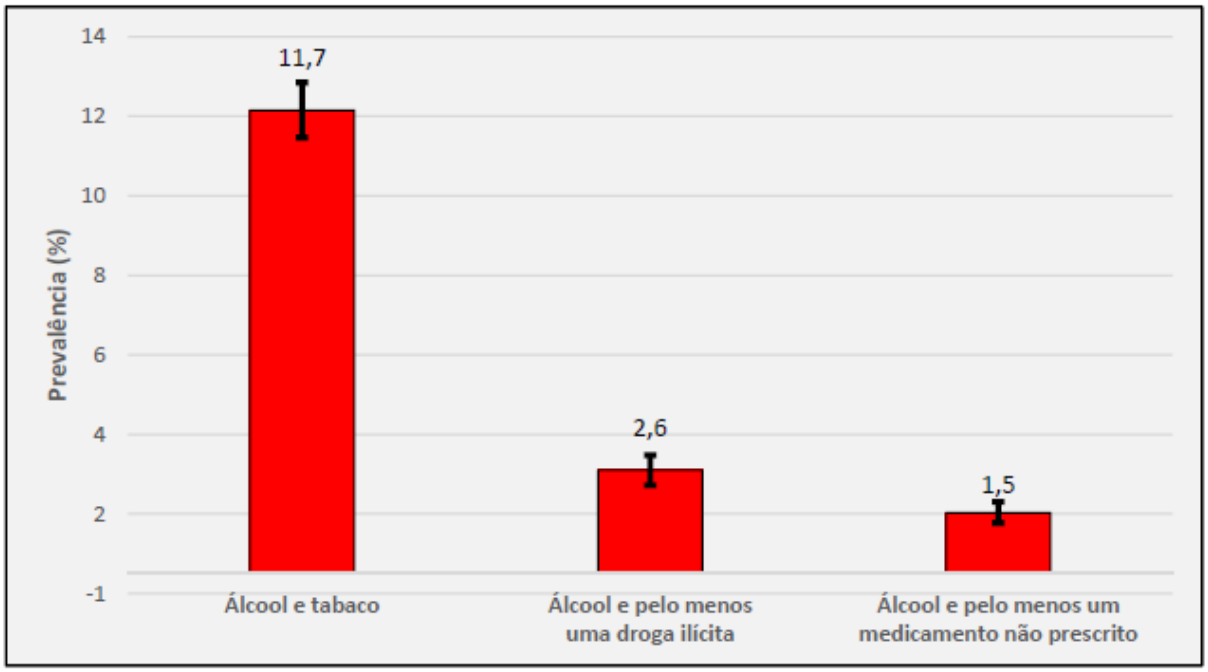
Figura 11 - Número de pessoas entre 12 e 65 que consumiram múltiplas substâncias nos últimos 12 meses.



Fonte: (ICICT, Fiocruz. III levantamento Nacional sobre o Uso de Drogas pela População Brasileira).

Os levantamentos dispostos na Figura 11 representam em milhares a quantidade de pessoas que usaram algum tipo de droga nos últimos 12 meses da realização da pesquisa, tanto as lícitas quanto as ilícitas, levando em consideração ainda os medicamentos sem prescrição.

Figura 12 - Prevalência de consumo de múltiplas drogas nos últimos 12 meses.



Fonte: (ICICT, Fiocruz. III levantamento Nacional sobre o Uso de Drogas pela População Brasileira).

Com relação aos números das Figuras 11 e 12 expressa a quantidade prevalecente, a proporção de caso que mantiveram o uso das drogas nos últimos 12 meses, representando em porcentagem a parcela dos milhares de pessoas.

2.2 Base de Pesquisa (Banco de Dados)

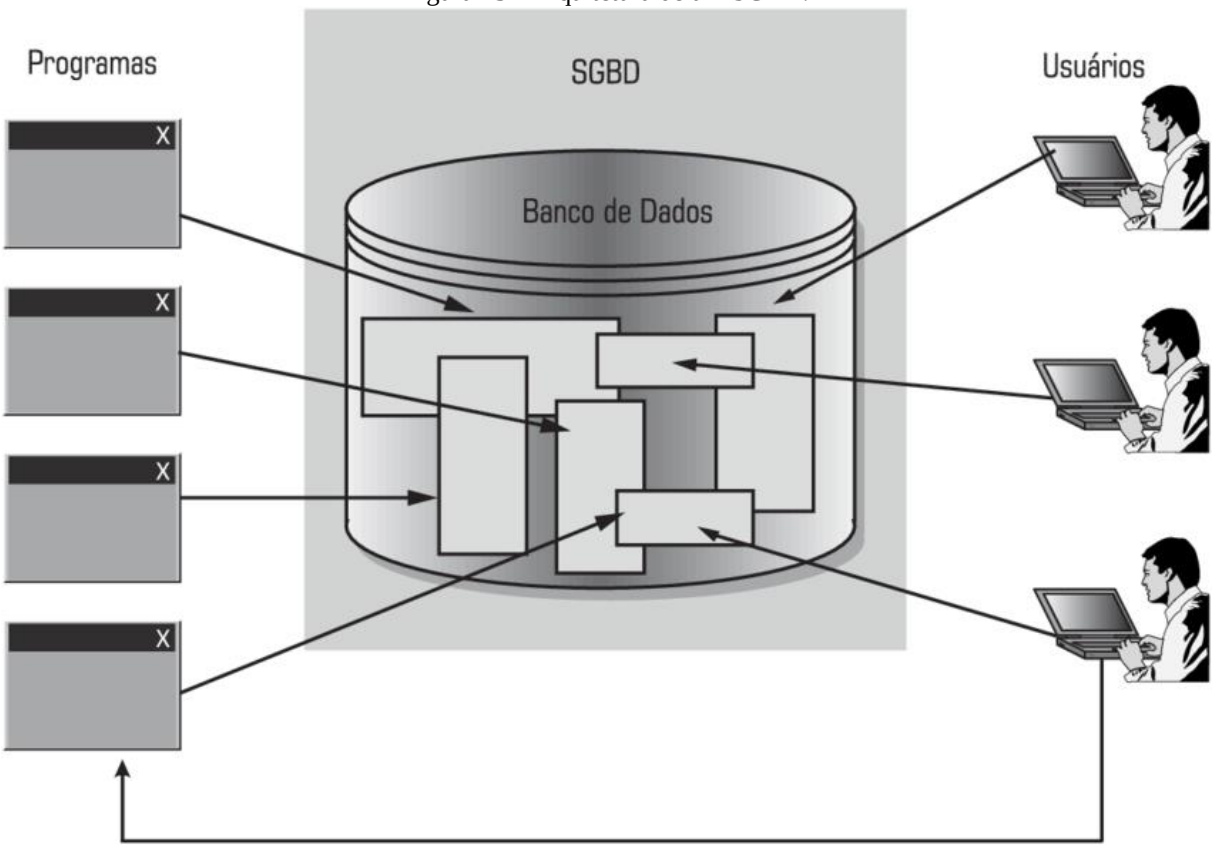
As bases de dados são essenciais quando se procura realizar a análise pois é através delas que são apontados os padrões. Segundo Elmasri e Navathe (2004), um banco de dados é uma coleção de informações relacionadas, que possui algumas propriedades implícitas como ser uma coleção lógica e coerente de dados com algum significado inerente.

Ou ainda conforme Medium (2018) um banco de dados é “um conjunto de arquivos que possuem relações entre si, que apresentam dados sobre um mesmo assunto e que permitem a extração de informações dele”.

Logo os dados contêm o fundamental para as pesquisas, sendo o principal fator um banco de dados de qualidade, bem relacionado e que permitem a sua análise, de modo que os bancos de dados relacionais, relacionam as bases disponíveis entre si através de posições de dados e são embasados no padrão relacional, forma na qual os dados são apresentados em tabelas (ORACLE, 2014).

Sua elaboração é um método para guardar as informações em certa mídia adequada e controlada pelo Sistema Gerenciador de Bando de Dados (SGBD), que são conjuntos de aplicações utilizadas para assegurar o funcionamento de um banco de dados, sendo, portanto, um sistema de aplicações facilitadoras dos processos de descrição, elaboração, controle e distribuição do banco de dados entre diversos utilizadores e programas. (ELMASRI; NAVATHE, 2004)

Figura 13 - Arquitetura de um SGBD.



Fonte: Arquitetura simples de um SGBD (JÚNIOR, 2018).

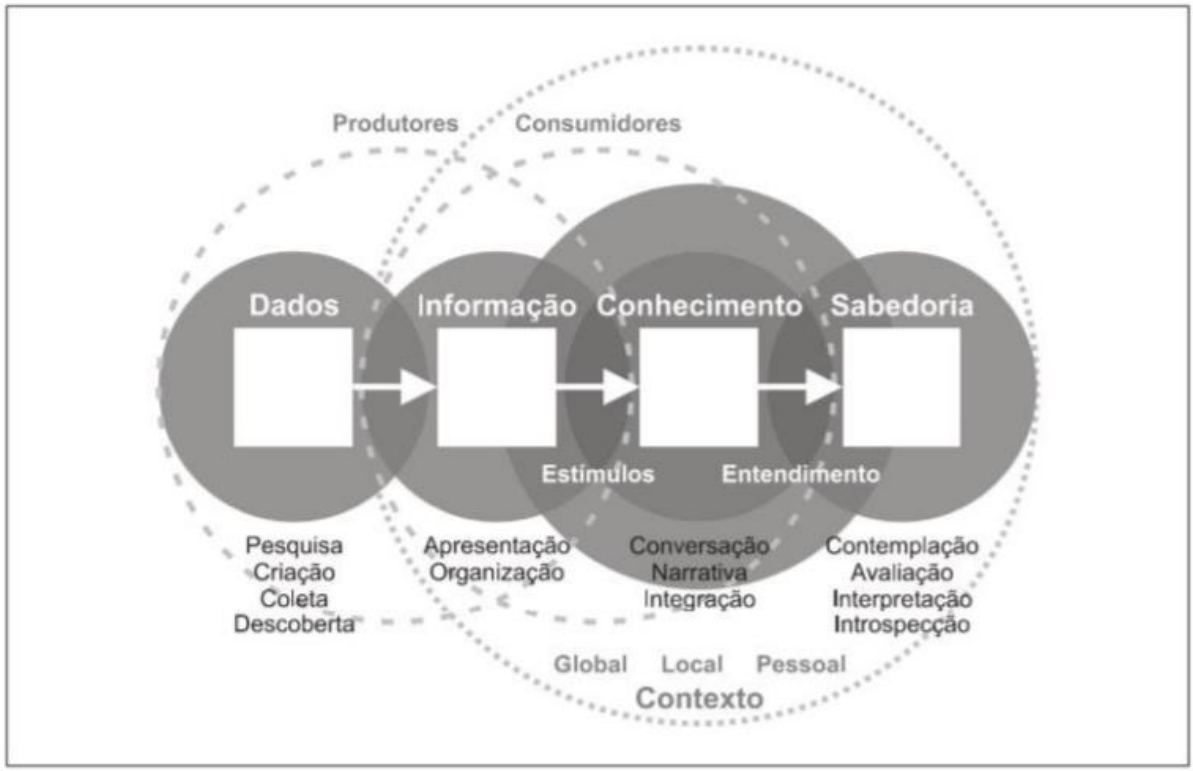
A Figura 13 apresenta um esquema da arquitetura de um SGBD. Conforme visto o banco de dados centraliza as informações que são gerenciadas pelo SGBD afim de permitir que, alterações e o compartilhamento dos dados possam ocorrer de forma rápida e eficiente tanto para as aplicações e programas, quanto para os usuários (JÚNIOR, 2018).

2.3 Dados, informação e conhecimento.

Os dados são informações desordenadas, desconexas e que separadas não tem finalidade instrutiva, quando conexas dão início ao processo de elaboração das informações. Portanto os

dados são apenas uma ponte para à busca por informações estruturadas, que sem um processo de análise não gera resultados. Os resultados são baseados nas informações, que são compostas pelos significados que a coleção de dados tem para o indivíduo que as obtém e manipula. Sendo assim, os conceitos acima apontam que os dados dão início a estrutura primordial para a chegada até o conhecimento, conforme ilustrado na Figura 14 (REUTERS, 2017; HASHIMOTO, 2009).

Figura 14 - Processo de continuidade da Informação.



Fonte: SHEDROFF (1999).

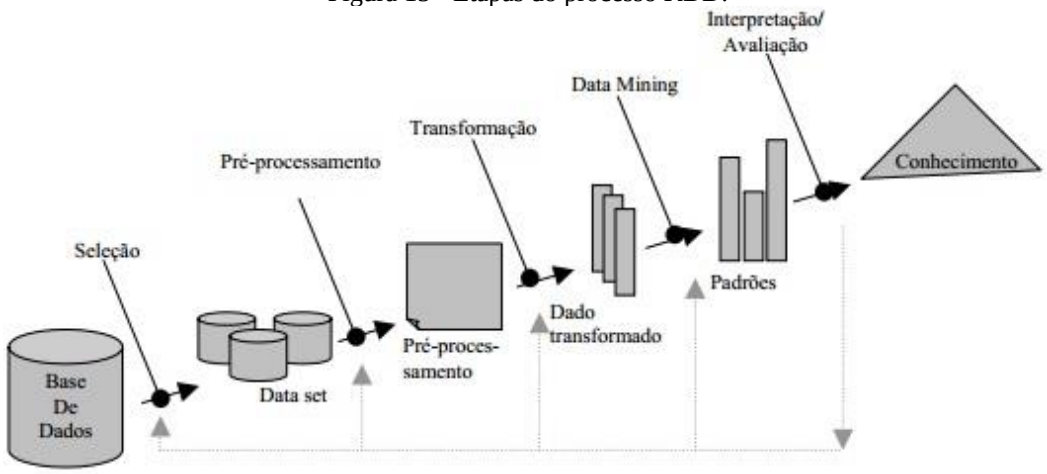
O processo de descoberta do conhecimento ilustrados acima remetem a busca pela sabedoria que segundo Davenport e Prusak (1998, p. 6 e 7) é definido como algo caracterizado por uma mistura de diversas partes constantes e estruturadas, que por fim acaba se tornado difícil de ser compreendido em termos lógicos. Em outras palavras o conhecimento é o processamento de toda a informação obtidas através dos dados, que em algum momento é convertida em experiência vivida pelo indivíduo.

2.4 Mineração de Dados

> "A mineração de dados é o processo de descoberta de informações acionáveis em grandes conjuntos de dados. A mineração de dados usa análise matemática para derivar padrões e tendências que existem nos dados. Normalmente, esses padrões não podem ser descobertos com a exploração de dados tradicional pelo fato de as relações serem muito complexas ou por haver muitos dados" (MICROSOFT, 2019).

De acordo com Cios et al. (2007) e Fayyad (1996) *Knowledge Discovery in Databases* (KDD) ou simplesmente descoberta de conhecimento em bancos de dados que compete a todo o processo de descoberta de conhecimento, e a Mineração de Dados a uma das funções do processo. O processo KDD é definido por Fayyad et al. (1996) como um "processo não trivial de identificação de padrões, a partir de dados, que sejam válidos, novos, potencialmente úteis e compreensíveis", sendo um conjunto de atividades sequenciais compostas por cinco etapas: seleção dos dados, pré-processamento, transformação, mineração de dados e interpretação, conforme ilustrado na Figura 15.

Figura 15 - Etapas do processo KDD.



Fonte: Fayyad et al. (1996).

A primeira etapa do processo KDD é a seleção de dados, nela serão determinados quais os bancos de dados serão considerados para que sejam obtidos resultados com conhecimentos pertinentes, período em que são determinados os objetivos pretendidos, tendo em vista à execução dos algoritmos de mineração (DILLY, 1996; REZENDE, 2003).

Em seguida, o pré-processamento, nesta etapa dados ausentes, imprecisas ou infundado nas bases de dados devem ser reparadas de forma a não prejudicar a qualidade dos protótipos de conhecimento a serem tirados ao final do processo de KDD, restringindo a dimensão da base

de dados e assumindo-os de forma que não sejam exibidos erros ou inconsistências. (DILLY, 1996; GONÇALVES, 2000).

Posteriormente, a transformação ou formatação, nesta etapa ocorre análise dos dados obtidos da etapa anterior e as remodela para um modelo específico para que possam ser interpretados na etapa seguinte (GONÇALVES, 2000).

Logo após a transformação ocorre a etapa considerada o núcleo do processo que é a mineração de dados, momento em que são determinadas as tarefas e técnicas, além de aplicar os algoritmos selecionados sobre o modelo obtido. Todavia, no decorrer desse procedimento pode ser necessário acessar dados adicionais como também modificar dados selecionados (BERRY & LINOFF, 1997).

A etapa final, de interpretação é onde as prescrições indicadas pelo processo precedente serão interpretadas e examinadas. Após a interpretação poderão surgir padrões, relacionamentos e descoberta de novos elementos, que podem ser empregues para pesquisas, otimização e outros (DILLY, 1996).

2.4.1 Tarefas

As tarefas da mineração de dados são os tipos de descoberta que se planeja verificar-se em um conjunto de dados, isto é, são os elementos que se pretende obter. Dentre as várias tarefas de mineração de dados existentes, as principais são a associação, classificação, clusterização e padrões sequenciais (FAYYAD, 1996; JOHN, 1997).

A associação que dentre todas as tarefas pode ser considerada a mais conhecida, engloba a busca por itens que constantemente ocorram de forma síncrona em transações de bancos de dados. Um exemplo prático é a aplicação no setor de Marketing, onde uma rede de supermercados consegue analisar se uma pessoa compra um produto X, e sempre que compra esse produto, acaba levando uma mercadoria Y, podem simplesmente aproximar as seções desses produtos, para que facilitem para o cliente e para que ele seja induzido a levar as duas mercadorias (BERRY; LINOFF, 1997; GOLDSCHMIDT; PASSOS, 2005).

A classificação consiste em investigar uma função que mapeie um agrupamento de registros em um grupo de rótulos definitivos predefinidos designados classes. Visto esta definição nota-se que é possível utilizar classificação para desenvolver uma ideia do tipo de cliente, item ou objeto. Descrevendo vários atributos para identificar uma determinada classe, utilizando por exemplo para classificar diferentes tipos de carros, com atributos diferentes, com

isso dada a chegada de um novo veículo, aplicado àquele tipo (HAN; KAMBER, 2001; GOLDSCHMIDT; PASSOS, 2005).

A clusterização reconhece um grupo finito de categorias ou agrupamento para descrever os dados, ou seja, é uma classificação não-supervisionada. Tem por finalidade particionar a base de dados em um número máximo de clusters (grupos), em que as áreas de um *cluster* sejam semelhantes. As categorias podem ser respectivamente exclusivas, hierárquicas ou ainda possuir atributos em comum. Como técnicas empregadas para associar dados tem-se segmentação demográfica e redes neurais (NEVES, 2003).

Os padrões sequenciais são geralmente utilizados sobre dados de longo prazo, os padrões sequenciais são um método útil para identificar tendências ou ocorrências regulares de eventos semelhantes. Com isso é possível identificar, por exemplo, que o cliente sempre troca de carro em determinada época do ano, através dessa informação o vendedor pode ligar para o cliente mostrando possíveis ofertas de veículos (BROWN, 2012).

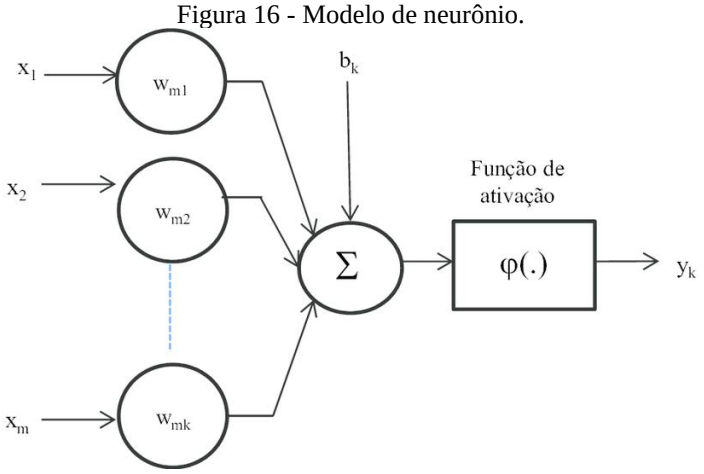
2.4.2 Técnicas

Após abordar as principais tarefas da mineração de dados, será discutido sobre as técnicas e a existência de uma correlação com as tarefas descritas na seção 2.4.1. A técnica é parte de um todo Mineração de Dados que apresenta um designo particular, em que existem várias implementações distintas através de diversos algoritmos. Assim, as tarefas dividem os algoritmos de acordo com o propósito de implementação, sendo que os algoritmos de uma mesma tarefa dispõem de uma mesma finalidade (FURLAN, 2018).

A primeira técnica a ser exemplificada são as redes neurais que está associada a classificação e, de acordo com Haykin (2001) considera-se a equivalência das redes neurais artificiais com os padrões de processo paralelo distribuídos. As redes neurais são compostas por unidades simples de ajustes, que assimilam um tipo de relacionamento complexo experimental, capacitando-o para supostas aplicações, como a de prever características ou situações.

Além disso, segundo Perera et al. (2011) nas redes neurais há uma relação de pesos sinápticos com conexões entre neurônios. Conforme novas informações são incorporadas nas redes os pesos mudam, a partir de algoritmos de aprendizagem, os pesos sinápticos possibilitam a fixação de uma variável de saída ou de reposta com base de dados de entrada. Logo abaixo na Figura 16 é demonstrado um neurônio de uma rede neural.

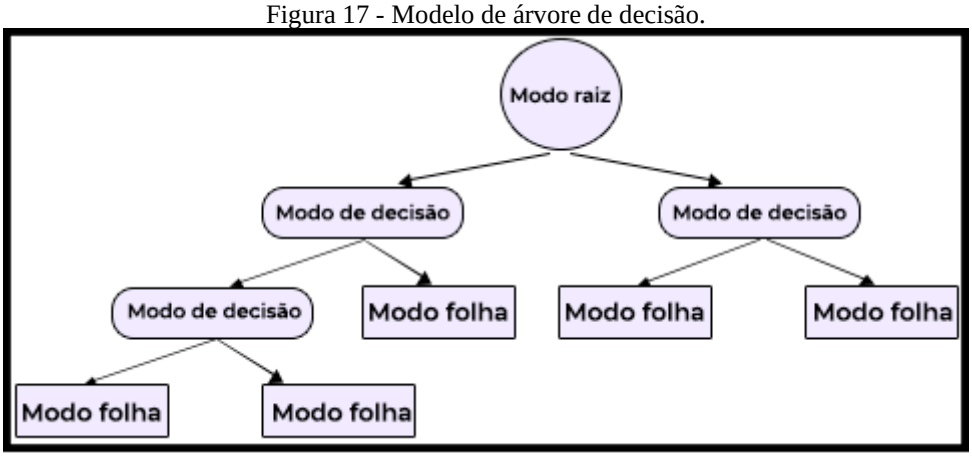
Figura 16 - Modelo de neurônio.



Fonte: HAYKIN (2001, p. 36).

As árvores de decisão de acordo com Han e Lamber (2001) possuem os padrões estatísticos que empregam treinamento supervisionado para classificação. O protótipo cria uma base de decisão em forma de árvore, onde cada nó, pode ser tido como uma propriedade de um teste, e respectivamente cada nó-folha tem um lugar na classe. A árvore inicia-se por um único nodo, que vai sendo fragmentada até chegar na classe. A Figura 17 apresenta um modelo de árvore de decisão com o modo raiz, modo de decisão e o modo folha.

Figura 17 - Modelo de árvore de decisão.



Fonte: Adaptado de SYACHRANI; JEONG; JUNG (2012, p.636).

A última técnica a ser abordada são regras de associação, uma configuração comumente usada das regras de associação, é a anotação condicional onde SE A ENTÃO B, historicamente demonstrada em um padrão de implicação A → B, tendo dois fundamentos. O que é exibido antecedentemente que é representado por A e o que vem posteriormente, que é denotado como consequência, representado pelo B. (AGRAWAL; IMIELINSKI; SWAMI, 1993)

Além disso segundo Levy (1999) as regras de associação são relativas a dois aspectos chamados suporte (sup) e confiança (conf), que são incumbidos pela análise de, por exemplo, o quanto que determinado produto A acarreta a compra de um produto B. O valor de suporte retrata o número de vezes que uma associação sucedeu na relação à totalidade de registros, ou seja, é a expectativa de uma transação condizer com a condição A. Ao mesmo tempo que o valor de confiança evidência a porcentagem de eventos do antecedente onde o produto consequente está relacionado, quer dizer, é a possibilidade de que uma operação corresponda a condição B, se ela satisfaz a condição A

2.4.3 Algoritmos

Algoritmo é uma sequência finita de passos para uma possível solução de um problema, tido também como uma sequência minuciosa de ações para cumprir-se uma tarefa (MEDINA; FERTIG, 2005).

O C4.5 que é um algoritmo de árvore de decisão e que foi publicado em 1987 por John Rosa. Quilan (1993). O algoritmo tem o objetivo de estabelecer um modelo classificador na forma de árvore de decisão, por meio da folha que mostra um ponto final na classificação, em que as classes são atribuídas, e do nó de decisão, que pode envolver uma ramificação. Apresentando também algumas contribuições, em que as principais são lidar com os atributos categóricos e contínuos e tratar de valores desconhecidos.

O algoritmo Apriori desenvolvido em 1993 por Agrawal e Srikant. Este algoritmo é utilizado para tarefas de associação com o intuito de buscar itens semelhantes em um intervalo de tempo ou não. De acordo com Pasta (2011) há dois passos que dividem o algoritmo. A pesquisa de itens frequentes é o primeiro passo, onde o usuário define uma margem mínima para o suporte, com isso o algoritmo procura por todos os conjuntos de dados que se demonstram com maior apoio a esse limite. O segundo é a construção de regras com base nos conjuntos do que estão presentes no primeiro passo.

Segundo Pasta (2011) a ideia principal do Apriori é fazer com qualquer que seja um subgrupo de um grupo de itens e que ele seja constante, acabando assim por eliminar dos itens que tem algum subgrupo que não seja contínuo.

Segundo Costa et al. (2013) o algoritmo *K-means* que foi proposto por Lloyd em 1957 e que se tratando de tarefas de agrupamento é o mais conhecido, seu funcionamento se discorre da seguinte maneira.

De acordo com Costa et al. (2013) inicialmente são apresentados o número K de conjuntos que está se procurando. Posteriormente X pontos são escolhidos aleatoriamente para destacar os centroides dos conjuntos, com isso, um grupo de vetores é particionado de modo que cada componente é atribuído ao conjunto de centroide mais perto. A todo momento que ocorrer iteração do algoritmo, os K centroides ou médias, são formatados conforme os elementos pertinentes ao conjunto subsequente, os elementos são realocados para o conjunto onde o novo centroide localiza-se mais próximo. Algoritmo esse que é executado várias vezes, e ao final é escolhido o melhor resultado.

2.4.4 Ferramentas de mineração

Na mineração de dados podemos encontrar várias ferramentas, essa seção demonstra algumas das ferramentas utilizada para mineração de dados.

O *DataMelt* é um programa aberto que pode ser usado em várias áreas como ciências naturais, engenharia de modelagem e análise de mercados financeiros. Além de ser um *software* para computação numérica é também para análise de grandes volumes de dados e visualização científica, executado em plataforma *Java*, mas também pode ser aplicado com *Python*. Ao final do processamento o *DataMelt* entrega imagens vetoriais de alta qualidade para que posteriormente possam ser utilizadas em sistemas de processamento de texto (DATAMELT, 2005).

Outra ferramenta é a *Oracle Data Mining*, que conforme *Oracle* (2020) fornece algoritmos robustos de mineração de dados para que os analíticos possam fazer previsões e retirar informações mais precisas de seus dados e investimentos *Oracle*. Tendo também a oportunidade de desenvolver e empregar modelos preditivos dentro do banco de dados, para que assim possam realizar previsões.

O *Orange Data Mining* é uma ferramenta aberta com foco principal em *machine leraning* podendo criar um projeto de trabalho do início ao fim, sem a utilização de qualquer tipo de código. Funciona em sistemas operacionais como *Windows, Linux e MacOs* (BATISTA, 2019).

Já a outra ferramenta desenvolvida em 1999 por estudantes da Nova Zelândia, o *Wiakato Environment for Knowledge Analysis, (Weka),* foi desenvolvida em *Java*, com isso ela é considerada um *software* multiplataforma com suporte a *Windows*, *MacOs X* e *Linux*, apresentando como restrição a instalação da Máquina Virtual *Java*. O *Weka* é constituído de dois pacotes: um pacote autônomo para utilização clara dos algoritmos, empregando um

formato de dados específico e o outro são os pacotes responsáveis pelas classes *Java* que rodam os algoritmos, aplicando um formato de informações únicas. Com isso é provável que construam uma aplicação *Java* que disponha desses algoritmos e aplicá-los nos bancos de dados pela conexão *Java DataBase Connectivity* (JDBC) (PASTA, 2011).

2. RESULTADOS

Nesta seção o ponto chave é esclarecer os resultados obtidos durante este estudo. Foi organizada conforme avanço de cada etapa, e seus respectivos resultados, observado que, durante todas as etapas foram avaliadas as implementações e analisadas cada alteração necessária.

Após o levantamento dos dados obtidos no estudo do IBGE e da fundamentação teórica, foi observada a conveniência da elaboração de tal estudo. De modo que, as estatísticas deixem de se tornar apenas números estáticos e promovam o conhecimento para a elaboração de políticas de enfretamento as drogas.

Isto posto, foram assimilados conhecimentos necessários dos bancos de dados e sobre a mineração de dados. Buscando absorver quais as melhores formas de se obter os padrões esperados por meios das técnicas de mineração.

3.1 Abrangência do Estudo

A análise dos conteúdos dispostos nas bases de informações, viabilizou determinar os locais a serem estudados com mais exatidão, com isso o escopo resultou nas regiões do Brasil sendo elas, o Norte, Centro-Oeste, Nordeste, Sudeste e Sul, com destaque as suas capitais, assim foi possível abranger uma gama maior de regiões, permitindo a abstração das informações. Logo por meio da continuidade das análises, notou-se a necessidade da verificação dos entornos dos ambientes escolares, considerando a segurança do local onde se encontra cada escola.

Além das informações acima, as bases de conhecimento concentram os dados fundamentais para as justificativas dos estudos, nota-se dentre os temas o consumo de drogas lícitas e ilícitas e sua periodicidade.

3.2 Critério para Seleção da Ferramenta de Mineração de Dados

Definido o escopo da área de estudo, partiu-se em busca das demais atividades a serem realizadas e com isso a busca pelo *software* encarregado de realizar a mineração dos dados. Vários fatores foram utilizados durante a seleção, como a indicação por pessoal mais experiente, a adaptabilidade a esse *software*, as condições para adquirir, sua usabilidade, sua interoperabilidade, suas funcionalidades, o grau de instruções disponíveis acerca de como utilizar tal ferramenta e a qualidade das informações entregues pelo *software*.

Após algumas sugestões e pesquisas formou-se uma predileção por dois *softwares*, *Orange Data Mining*[2] *e Weka*[3], que possuem robustez nos seus processos de mineração, assim os demais fatores como qualidade das informações geradas, interoperabilidade, facilidade de aprendizado entre outros conforme descrito no decorrer da avalição, foram estudados a fim de eleger o mais apto.

A busca por informações e instruções a cerca de sua utilização notou-se que ambos possuem uma documentação em seus respectivos sites, no entanto são escritas em inglês algo que dificulta levando em consideração o tempo e a necessidade de uma abordagem prática.

Em seguida percebeu-se que a quantidade de materiais disponíveis para estudo e pesquisa sobre ambos possuem uma diferença evidente, o *Weka* possui uma grande quantidade de materiais não oficiais como cursos, videoaulas, dentre outras ferramentas que facilitam o seu aprendizado. Enquanto a ferramenta *Orange Data Mining* é mais restrita com relação a conteúdos como tutoriais, videoaulas.

Assim os materiais de ambos se concentram em conteúdo não oficial na língua portuguesa (oficialmente em inglês), tendo o *Weka* uma quantidade maior de conteúdo em português.

Em relação a compatibilidade ambos são compatíveis nos sistemas operacionais *Windows, Linux e MacOs*.

Em relação à interface, ambas são intuitivas e possuem fácil identificação dos elementos essenciais, no entanto devido à grande quantidade de matérias disponíveis do *Weka*, o aprendizado, a assimilação e o manuseio de tal ferramenta acaba por ser facilitado, entretanto a adaptação a qualquer ferramenta é algo um tanto que pessoal, não sendo assim um critério concludente para que sim ou para que não, levando somente a uma inclinação na escolha, uma vez obtida a primeira impressão.

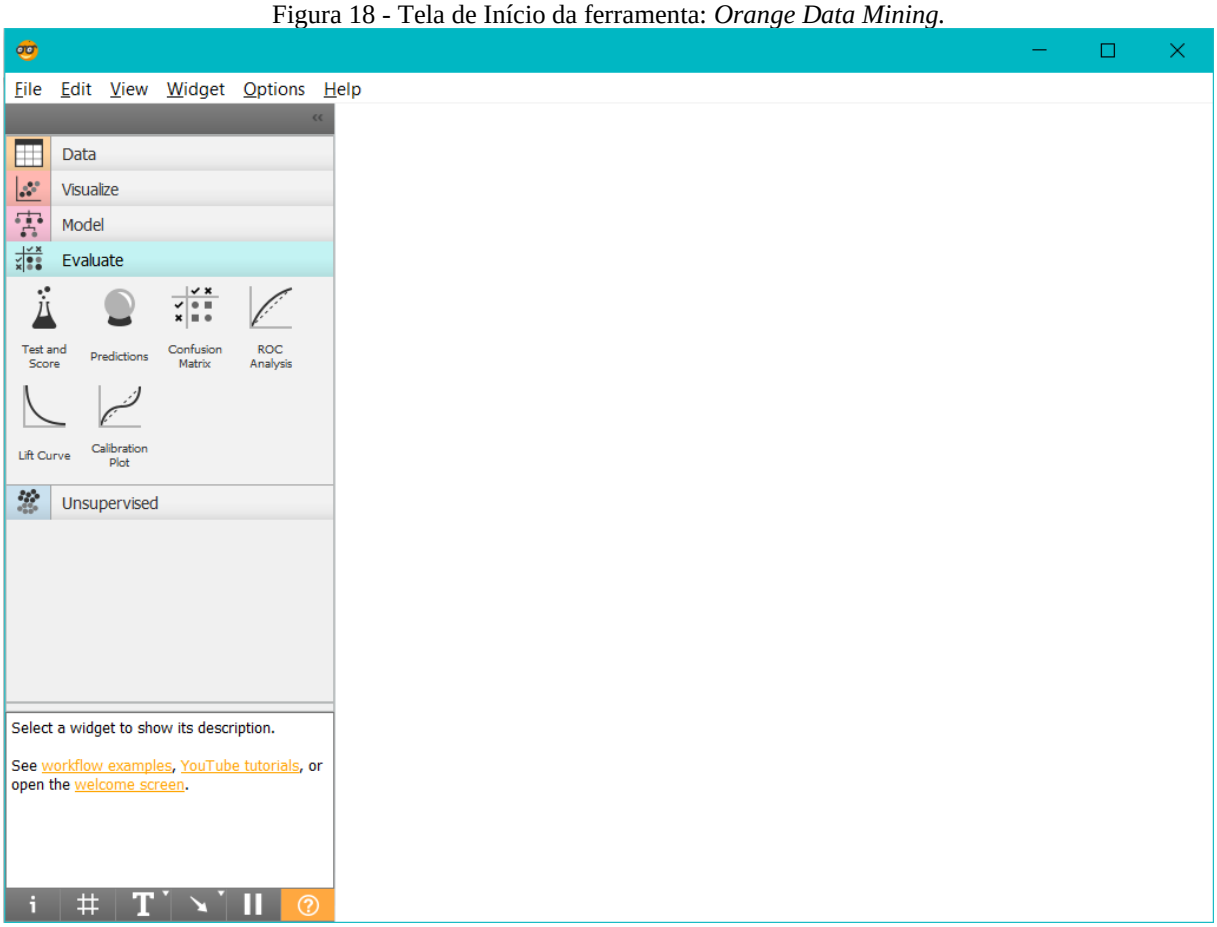
Em seguir será apresentado as interfaces das duas ferramentas *Weka* e *Orange Data Mining.*

A. *Orange Data Mining*

[2] https://orange.biolab.si/download/#windows
[3] https://waikato.github.io/weka-wiki/downloading_weka/#windows

A Figura 18 representa a tela inicial da ferramenta *Oracle Data Mining*, que possui uma interface intuitiva e autoexplicativa. Através da página inicial o usuário pode verificar as principais opções, discorridas de maneira breve em seguida.

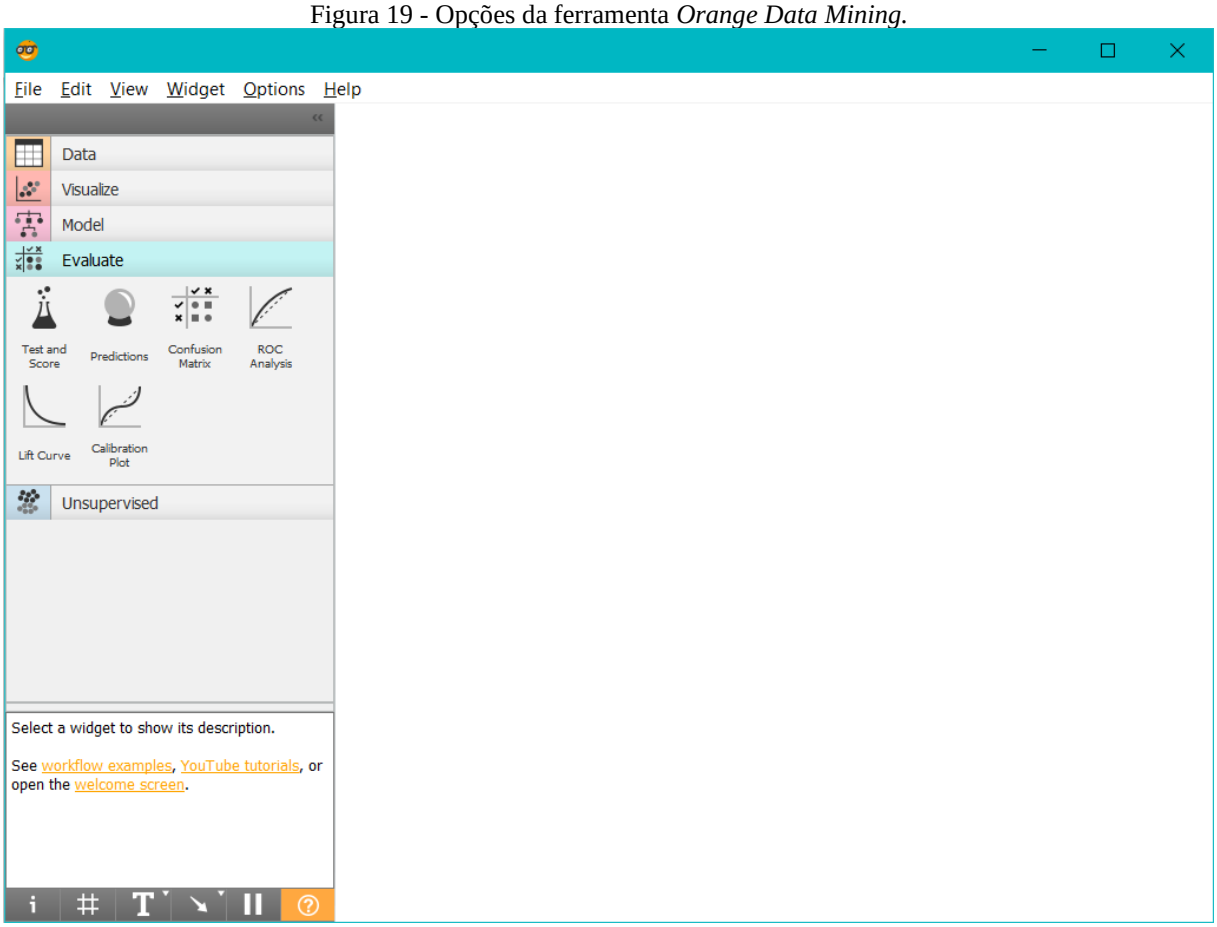
Figura 18 - Tela de Início da ferramenta: *Orange Data Mining.*



Fonte: MORAIS; SANTOS (2020).

Após a apresentação inicial da tela conforme Figura 18 o usuário pode escolher dentre as opções disponíveis, e com isso dar continuidade no processamento das informações, logo se observado fica disposto a esquerda 5 abas na qual são agrupadas várias funcionalidades e técnicas para a mineração das informações.

Figura 19 - Opções da ferramenta *Orange Data Mining.*

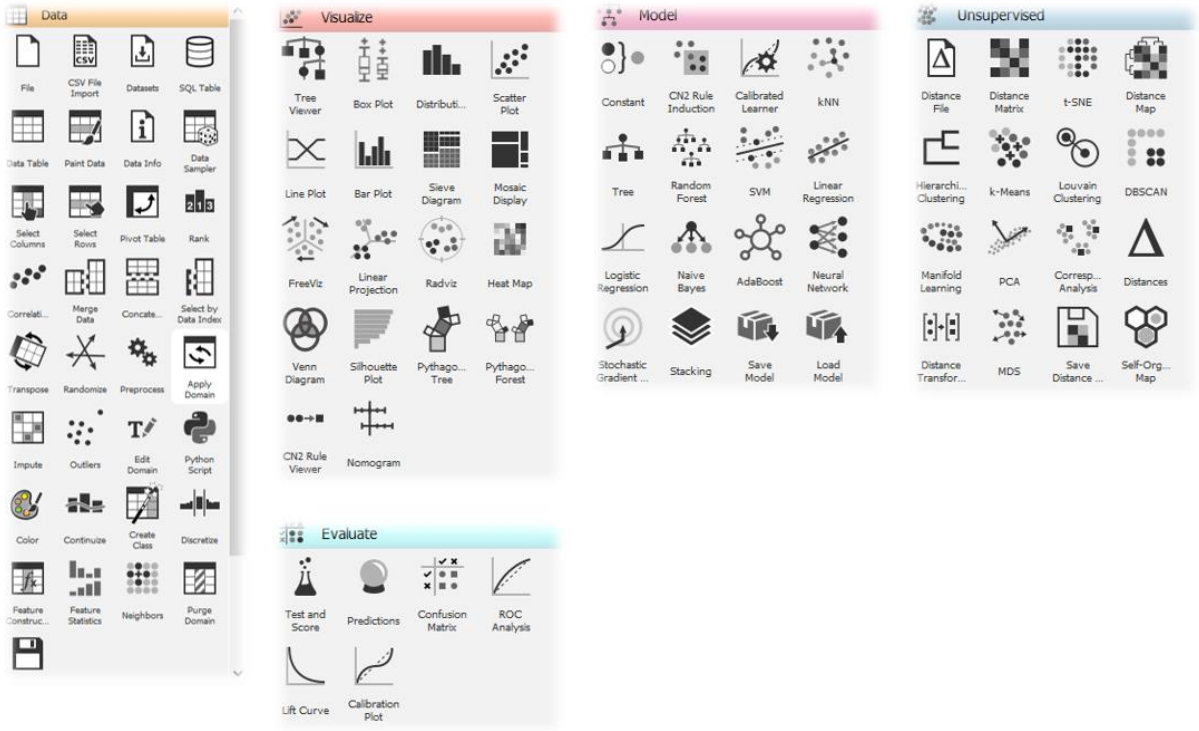


Fonte: MORAIS; SANTOS (2020).

Conforme apresentado nas Figuras 18 e 19 abaixo são descritas de forma sucinta cada umas das 5 abas disponíveis no *Orange Data Mining.*

I. ***Data:*** apresenta as funcionalidades para o processamento de volume e preparo das informações, assim como várias outras técnicas para a mineração.
II. ***Visualize:*** organiza as visualizações, é possível realizar a criação de gráficos de vários tipos, assim como outras visualizações.
III. ***Model:*** apresenta os modelos, demonstra os vários modelos utilizados em *Data Mining* desde a regressão linear até as redes neurais.
IV. ***Evaluate:*** mensura o desempenho dos modelos criados.
V. ***Unsupervised:*** apresenta as transformações de dados e outras funcionalidades.

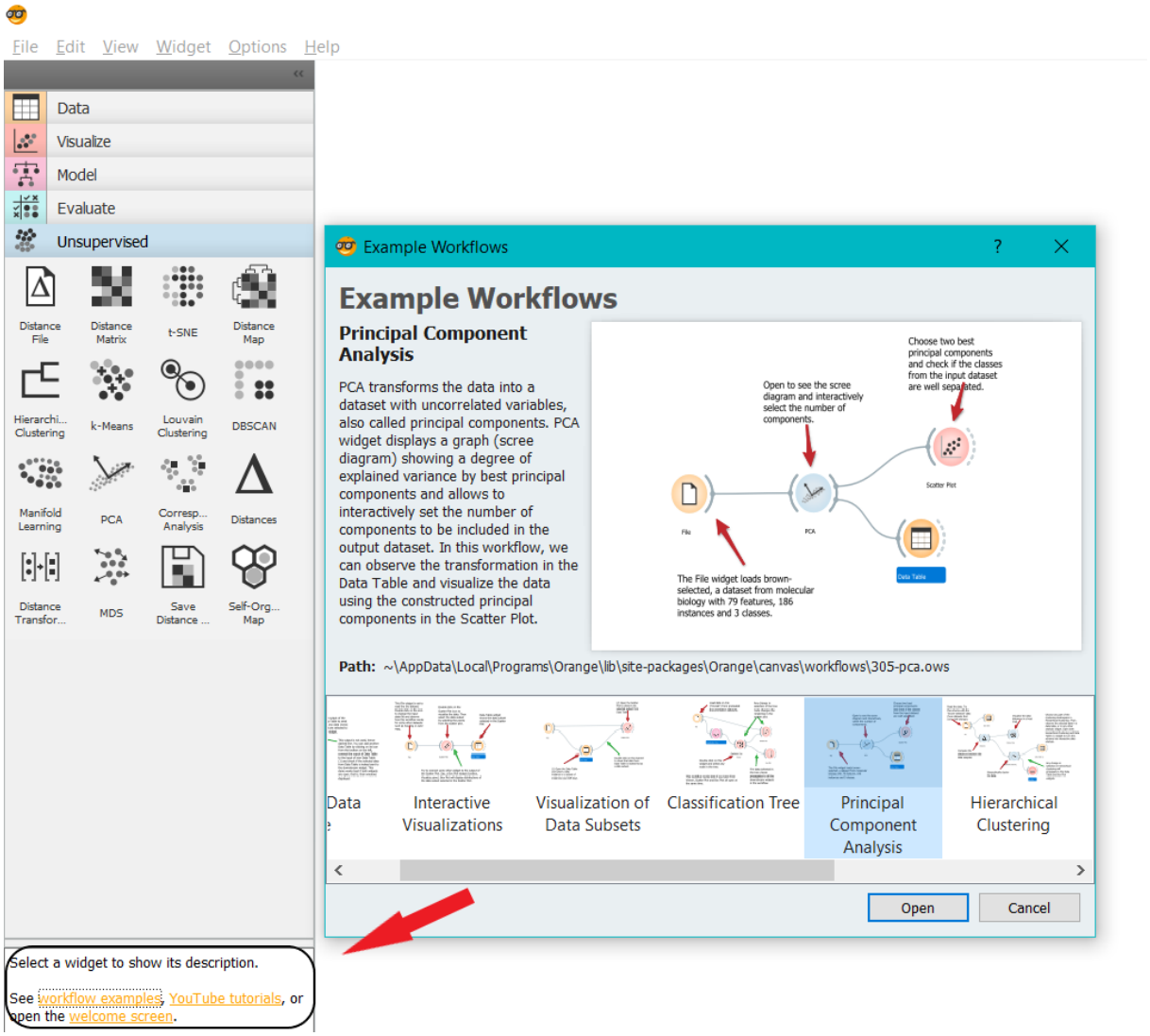
Figura 20 - Secções



Fonte: MORAIS; SANTOS (2020).

Conforme a Figura 20 pode ser visto as funcionalidades e opções que cada aba possui, é possível ainda incluir várias funcionalidades com um pouco de pesquisa e de acordo com a necessidade, assim logo abaixo temos alguns exemplos de *workflows* que podem ser obtidas na tela principal assim como tutoriais e algumas configurações.

Figura 21 - Mais opções e exemplos da ferramenta.

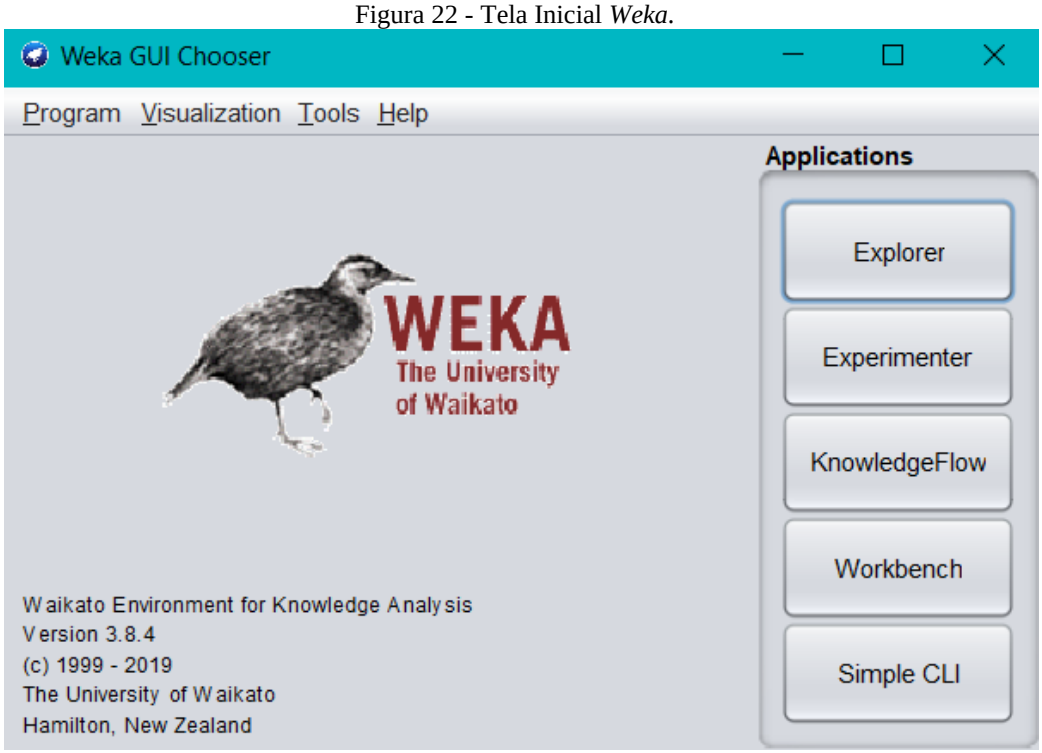


Fonte: MORAIS; SANTOS (2020).

O *workflow* da Figura 21 representa os principais componentes para análise, contudo há diversos outros *workflows* no próprio *software* e conforme destacado no canto inferior esquerdo há como visualizar mais .

B. *Weka*

Figura 22 - Tela Inicial *Weka*.

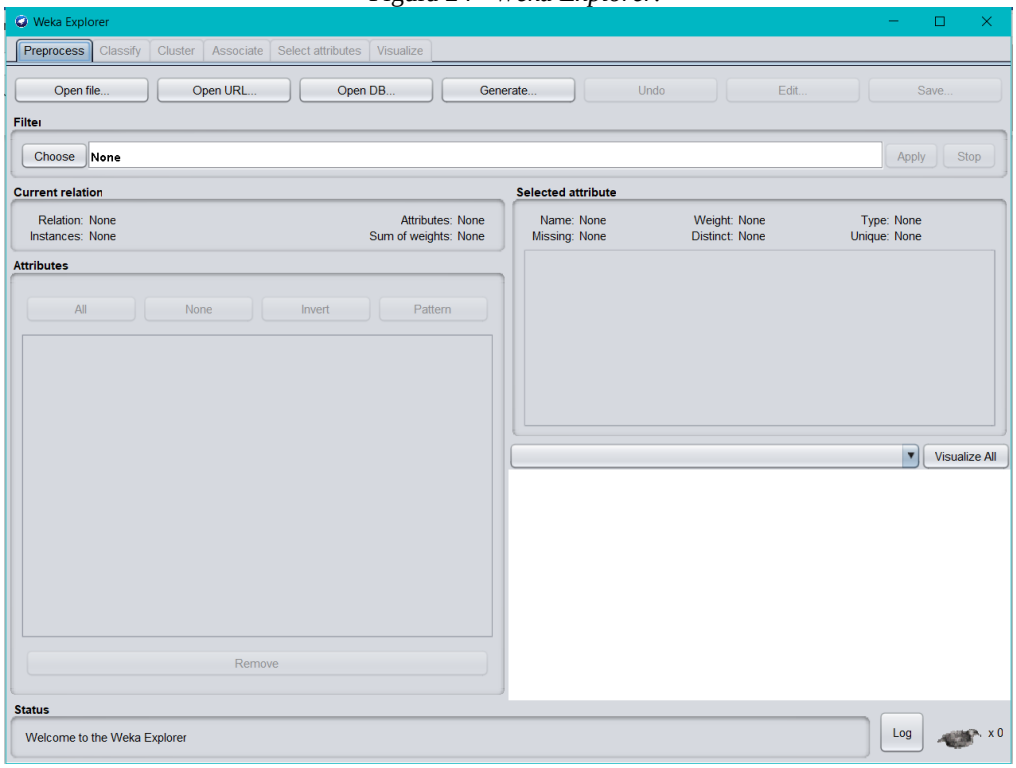


Fonte: MORAIS; SANTOS (2020).

Na tela inicial do *Weka* dispõe de 5 abas com várias funcionalidades. A primeira opção é o *Explorer* que possui várias funcionalidades, desde a visualização até o seu pré-processamento dos dados. Conforme a Figura 24 é possível ver a divisão da interface e é por meios dessas opções que todo o processo de mineração é realizado, em suas funcionalidades estão opções de visualizar, classificar, associar, selecionar atributos, editar, visualizar entre outras.

E válido enfatizar que o *Weka* possui um formato de arquivo próprio, o *Attribute-Relation File Format* (ARFF) que é um tipo de arquivo de dados, dividido em cabeçalho e dado. A ferramenta possui também a funcionalidade de importar arquivos *.csv*, ou ainda a consulta direta em bancos de dados por meio de um endereço ou ainda a realização de consultas direta nas bases de dados para diversos SGBDs utilizando um *driver* JDBC.

Figura 24 - *Weka Explorer*.



Fonte: MORAIS; SANTOS (2020).

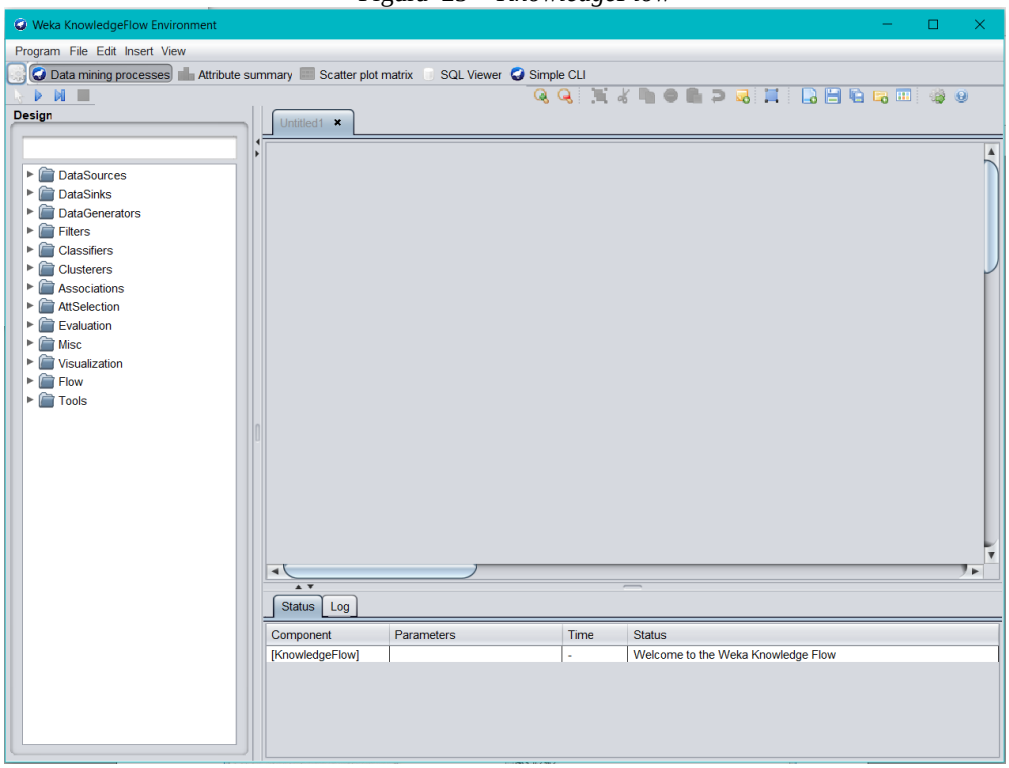
A opção *Experimenter* fornece os testes de desempenho, automatiza os processos e é voltada para o desempenho dos algoritmos. Por meio desta os algoritmos de aprendizado de máquina são executados e assim é possível visualizar os resultados e definir a melhor opção para o problema e ainda parametrizar os algoritmos.

A opção *Setup* realizada as configurações e os dados são definidos para análise nos *Datasets*. A ferramenta apresenta ainda funções de parametrização, quantidade de verificações, inserção de algoritmos e após a devida configuração e seleção dos dados é realizada a análise das informações que são apresentadas na aba *Run*, caso algum erro ocorra durante o processamento isso é apresentado indicando onde.

A opção *Experimenter* realiza as configurações, definições e vários ajustes necessários para que o processamento ocorra, assim a extensão de ajustes disponíveis é enorme e cabe a necessidade e o conhecimento para uma parametrização mais fina. Logo também é possível configurar para a apresentação somente daqueles dados essenciais, eliminando muito o ruído.

Em seguida temos a opção *KnowledgeFlow* conforme Figura 25 que realiza a criação dos fluxos de processos através dos dados. É possível ainda realizar várias modificações no fluxo através dessa ferramenta.

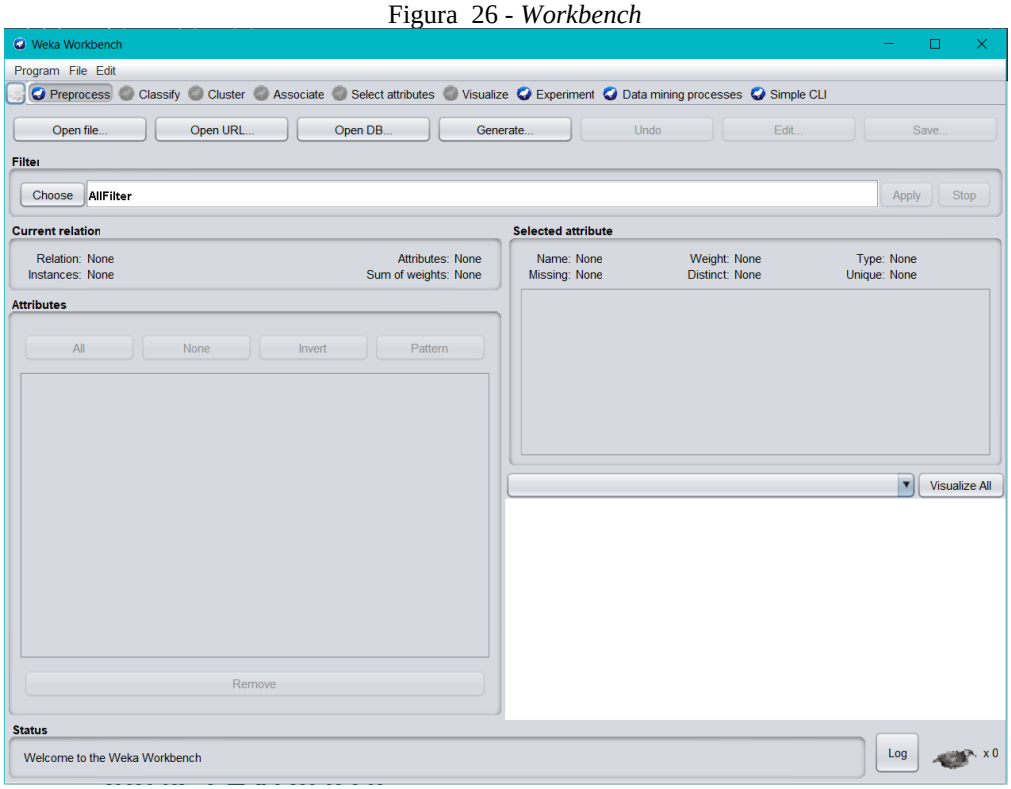
Figura 25 - *KnowledgeFlow*



Fonte: MORAIS; SANTOS (2020).

E finalmente as opções *Workbench e Simple CLI* conforme Figura 26, que reuni a informações da ferramenta em uma única aba, ficando dispostas na parte superior, se torna mais viável para casos em que há necessidade de utilização de múltiplas funções ao mesmo tempo, a segunda opção se resume a uma forma mais manual de utilizar a ferramenta, através do uso de scripts e linhas de comando.

Figura 26 - *Workbench*



Fonte: MORAIS; SANTOS (2020).

Após analisar o *Weka* suas funcionalidades e interface optou-se pela utilização da ferramenta *Weka*. As funcionalidades disponíveis na ferramenta atenderam as necessidades deste estudo, porém a quantidade de materiais de apoio foi o requisito mais impactante.

3.3 Processo para Geração dos Resultados

O processo KDD conforme seção 2.4 é um processo não trivial para a identificação de padrões, com isso optou-se logo ao se pensar no projeto por tal método uma vez que, apresenta etapas simplificadas e que são adequadas ao propósito final.

Com finalidade na descoberta de padrões, informações úteis para a tomada de decisão ou descoberta de anomalias ainda não identificas, as etapas do KDD conseguem gerar vários resultados. Embora haja outros métodos para a geração de conhecimento, devido ao processo de 5 etapas, verificou se que a aplicação do KDD para os resultados buscados é concisa pertinente.

3.4 Escolha da Tarefa, Técnica e Algoritmo

Classificada como uma das etapas primordiais para a definição de todo o processo, as escolhas aqui apresentadas tiveram grande impacto nos resultados, levando em consideração que antes de se pensar no projeto havia-se definido as bases a serem utilizadas, e através das informações pode-se mensurar algumas necessidades e definir objetivos.

As bases de dados adquiridas integram parte de um levantamento realizado pelo IBGE conforme o item 1, no qual grande maioria das informações são relativas e não se conectam ou apontam casualidade entre os fatos apresentados, com isso a necessidade da aplicação de uma tarefa para organizar as informações, as classificar e identificar padrões.

Dentre as diversas tarefas apresentadas conforme a seção 2.4.1 de acordo com as necessidades de se mapear informações, investigando tal função destacasse a classificação, por permitir um trabalho sobre os atributos e classes, de modo que foi a escolhida para os processos desse estudo.

Em sequência, a técnica definida para complementar os processos acima foi árvore de decisão que, em uma descrição sucinta de acordo com a seção 2.4.2 por meio de números matemáticos emprega o treinamento controlado para a classificação e predição dos dados, é valido ressaltar que assim como o *Weka* os conteúdos presentes são bem vastos, facilitando a sua utilização.

E finalmente o algoritmo utilizado para a implementação da técnica acima, logo de início destacou-se o *C.4.5* sendo bem conhecido e constantemente utilizado, o mesmo é um algoritmo de árvores de decisão conforme a seção 2.4.3, e em alguns momentos ele aparece como *J48* principalmente na ferramenta *Weka* que utiliza esse nome em sua implementação.

3.5 Ambiente de Processamento

Para todas as alterações nas bases, foi necessária a utilização do *NotePad* ++ 7.9.1, principalmente para as adequações que sofreram alterações em sua estrutura com intuito de organizá-las para o processo de mineração dos dados. O sistema operacional utilizado foi o *Windows 10 Home Single Language* versão 2004 sobe uma configuração de *hardware Intel Core i7 – 6500U 2.5 GHz*, 16 *GB* de memória *RAM*, abaixo as ferramentas utilizadas:

A. *Microsoft Office* 365 2020 – Modificação e formatação dos dados.
B. *NotePad* ++ 7.9.1– Adaptação e conversão para o formato *.arff* dos dados.

C. *Workbench* 8.0.22– Conversão dos dados e integração com a ferramenta *Weka.*

3.6 Processo KDD

Conforme dissertado durante o embasamento teórico, o processo KDD possui em si algumas características essenciais como, ser iterativo e interativo e ainda um processo não trivial, de modo que as etapas estão interligadas e com isso para se chegar na próxima etapa é necessário que a última tenha sido realizada, no entanto é possível intervir nas atividades e repetir cada etapa do processo várias vezes se necessário.

Essa característica se mostrou essencial e legitimou a execução do processo visto a necessidade de modificações e repetições nas etapas, contudo para melhor entendimento foram discorridas conforme as etapas originais e seguindo assim a cronologia conforme o embasamento teórico (Figura 15).

3.6.1 Necessidade do Estudo

Ao se iniciar a realização da busca do tema da pesquisa, ficou claro o interesse em contribuir com algo que gerasse resultados para a sociedade de modo a identificar algo de valor social. Deste modo ficou decidido contribuir com a diminuição nos índices de consumo de álcool e drogas ilícitas no âmbito escolar.

Assim buscou-se avaliar as situações das áreas de estudo e quais eram as medidas já vigentes para mitigação, assim notou-se:

- Poucos estudos sobre o assunto.
- Medidas ineficazes para o combate.
- Uso inadequado dos dados disponíveis.
- Políticas repreensivas.

Em seguidas alguns objetivos iniciais foram traçados para se alcançar os resultados esperados:

I. Identificar padrões e anomalias por meio da mineração de dados através das informações de escolas públicas e privadas.
II. Mineração das bases públicas do último estudo sobre drogas nas escolas - Pesquisa Nacional de Saúde do Escolar (PeNSE - 2015).
III. Utilização do método de classificação e geração de árvores de decisões para análise dos resultados.

3.6.2 Seleção dos Dados

A. Obtenção das bases

Logo após a definição da necessidade do estudo, foi realizada a busca por dados, assim foram abordados estudos mais recentes sobre o tema e verificado sua compatibilidade com a proposta. Um estudo já realizado há anos ganhou destaque por ser composto por diversas bases com os mais variados assuntos sobre drogas nas escolas, tanto públicas quanto privadas, o PeNSE, que pode ser obtido através do portal de estatísticas do IBGE[1].

B. Conteúdo dos Dados

Seguindo as etapas do processo de KDD a descrição dos dados é essencial para se entender o que está sendo minerado, para demonstrar os dados originais, a Figura 27 e 28 apresenta a dados sobre a utilização das drogas. Os dados foram obtidos em formato *.xls* facilitando a utilização dos mesmos.

Figura 27 - Trecho dos dados obtidos

Brasil	**38,5**	**37,6**	**39,4**	**39,5**	**38,3**	**40,8**	**37,6**	**36,4**	**38,7**
Norte	**36,5**	**34,9**	**38,1**	**38,0**	**35,6**	**40,5**	**35,1**	**32,9**	**37,3**
Rondônia	40,5	37,7	43,3	41,0	36,3	45,7	40,0	37,0	43,0
Acre	34,6	31,8	37,4	36,8	33,1	40,5	32,3	28,7	36,0
Amazonas	36,9	34,1	39,8	37,1	33,5	40,7	36,8	32,9	40,7
Roraima	41,2	38,1	44,2	43,2	39,5	46,9	39,1	35,4	42,9
Pará	35,1	31,9	38,3	37,0	31,7	42,3	33,4	29,0	37,8
Amapá	35,5	33,0	38,0	37,3	34,1	40,5	33,7	30,1	37,4
Tocantins	37,0	33,0	41,0	40,9	35,8	46,0	33,4	28,4	38,3
Nordeste	**34,7**	**33,5**	**35,8**	**38,1**	**36,4**	**39,8**	**31,7**	**30,2**	**33,1**
Maranhão	34,0	30,9	37,1	38,4	34,3	42,5	29,9	25,8	34,0
Piauí	32,5	29,5	35,6	37,7	32,8	42,5	27,6	24,2	31,0
Ceará	33,5	30,9	36,1	35,9	32,8	39,0	31,3	27,5	35,0
Rio Grande do Norte	29,5	26,8	32,1	32,8	28,9	36,8	26,5	23,2	29,8
Paraíba	35,4	32,5	38,3	36,5	32,7	40,3	34,5	30,6	38,4
Pernambuco	35,3	32,6	38,0	39,2	35,4	43,0	31,6	27,6	35,6
Alagoas	33,5	30,2	36,8	36,0	31,1	41,0	31,3	27,7	34,9
Sergipe	33,2	30,9	35,6	38,6	34,9	42,3	29,2	26,0	32,4
Bahia	36,9	33,9	39,9	40,5	35,9	45,2	34,1	30,8	37,4
Sudeste	**39,7**	**37,9**	**41,5**	**39,3**	**36,9**	**41,7**	**40,2**	**37,9**	**42,4**

Fonte: MORAIS; SANTOS (2020).

Figura 28 - Trecho dos dados obtidos

Porto Velho	12,6	10,9	14,2	13,3	11,3	15,3	11,9	9,6	14,2
Rio Branco	10,6	8,5	12,8	12,5	9,4	15,6	8,8	6,5	11,2
Manaus	10,5	9,1	11,9	12,8	10,7	14,9	8,2	6,3	10,0
Boa Vista	13,2	10,2	16,3	14,4	10,9	17,9	12,1	8,9	15,4
Belém	6,2	5,1	7,3	7,6	6,0	9,2	4,9	3,5	6,2
Macapá	8,0	6,5	9,5	9,7	7,5	12,0	6,3	4,4	8,2
Palmas	7,1	5,5	8,6	8,8	6,7	11,0	5,5	3,7	7,3
São Luís	9,5	7,8	11,3	9,0	7,3	10,7	10,0	7,0	12,9
Teresina	5,7	4,4	7,0	7,7	5,7	9,8	3,8	2,4	5,3
Fortaleza	9,8	7,6	12,0	12,1	9,0	15,2	7,7	5,1	10,2
Natal	5,2	3,9	6,6	5,8	3,6	7,9	4,7	3,3	6,2
João Pessoa	8,3	6,8	9,7	9,1	7,1	11,1	7,5	5,9	9,1
Recife	7,8	6,5	9,1	9,1	6,9	11,3	6,7	4,9	8,4
Maceió	8,3	6,0	10,7	8,2	5,7	10,6	8,5	5,8	11,2
Aracaju	6,2	4,6	7,8	7,1	4,6	9,5	5,5	3,7	7,2
Salvador	6,1	4,8	7,4	6,7	4,7	8,8	5,6	3,9	7,4

Fonte: MORAIS; SANTOS (2020).

Figura 29 - Trecho dos dados obtidos

13 a 17 anos									
Brasil	12,0	10,8	13,1	12,5	11,0	14,1	11,4	9,9	12,8
Norte	8,4	5,8	10,9	11,2	7,4	15,0	5,2	3,4	7,1
Nordeste	9,0	7,1	10,9	10,1	7,4	12,7	8,0	5,7	10,2
Sudeste	13,0	10,7	15,3	13,4	10,5	16,4	12,5	9,7	15,3
Sul	16,8	14,5	19,0	15,7	12,6	18,9	17,9	14,8	20,9
Centro-Oeste	13,1	11,0	15,2	13,0	10,2	15,7	13,2	10,9	15,6
13 a 15 anos									
Brasil	9,1	8,1	10,1	8,9	7,4	10,5	9,3	7,9	10,7
Norte	7,2	4,6	9,8	9,1	5,9	12,3	5,1	2,4	7,9
Nordeste	6,5	4,8	8,2	8,0	5,0	10,9	4,9	2,7	7,1
Sudeste	10,0	8,3	11,7	8,9	6,2	11,7	11,1	8,6	13,6
Sul	12,6	10,0	15,1	10,3	6,6	14,0	14,9	11,3	18,5
Centro-Oeste	10,4	8,4	12,4	10,0	7,3	12,6	10,9	8,5	13,3
16 e 17 anos									
Brasil	16,6	14,2	18,9	18,5	15,5	21,6	14,7	11,9	17,5
Norte	10,6	7,6	13,5	15,0	8,6	21,3	5,5	2,6	8,4
Nordeste	13,2	9,1	17,2	13,5	8,9	18,2	12,8	7,7	17,9
Sudeste	17,6	13,1	22,0	20,9	14,8	27,0	14,6	9,4	19,7
Sul	23,8	19,6	28,0	25,2	19,9	30,5	22,5	17,6	27,5
Centro-Oeste	17,6	13,7	21,5	18,2	12,9	23,5	17,1	12,8	21,4

Fonte: MORAIS; SANTOS (2020).

Assim foi realizado uma análise sobre as bases obtidas e foram definidas as tabelas a serem utilizadas, logo abaixo listamos as tabelas utilizadas para o processamento do estudo:

Quadro 1 - Bases utilizadas

Arquivo	Conteúdo	Origem
Amostra_1_Tema_12_Cigarro	Dados gerais sobre vários temas acercas de drogas, escolas públicas e privadas.	IBGE
Amostra_1_Tema_13_Bebidas_Alcoolicas		
Amostra_1_Tema_14_Drogas_Ilicitas		
Amostra_2_Tema_04_Cigarro		
Amostra_2_Tema_05_Bebidas_Alcoolicas		
Amostra_2_Tema_11_Seguranca		
Amostra_2_Tema_06_Drogas_Ilicitas		

Fonte: MORAIS; SANTOS (2020).

Os dados conforme o Quadro 1 foram escolhidos com base nos critérios já discorridos e na abordagem do tema drogas nas escolas. Até essa etapa os dados continuaram como na fonte, sem a intervenção ou alteração da estrutura, como os dados foram obtidos de uma mesma fonte os ajustes realizados foram pontuais.

C. Análise dos Dados

Feitas as escolhas e apresentados os dados foi necessário identificar quais atributos seriam utilizados para as próximas etapas do KDD. Ao analisar cada uma das colunas nas tabelas verificou-se que as colunas intervalo de confiança, sexo, dependência administrativa e uma coluna de totalização não seriam úteis em razão do que se almejava ao final do estudo. As colunas mais promissoras para o estudo foram definidas conforme o Quadro 2 abaixo:

Quadro 2 - Colunas utilizadas

Descrição	Origem
Faixa etária	IBGE
Região	
Tipo de droga	
Dependência administrativa	
Quantidade de usuários	
Estado	

Fonte: MORAIS; SANTOS (2020).

D. Qualidade dos Dados

Ao finalizar as seleções e escolhas dos dados foi avaliado a qualidade dos dados escolhidos. Notou-se que a qualidade dos dados a um primeiro contato se apresentou satisfatória para mineração, e logo após um aprofundamento e limpeza necessária sua qualidade teve uma diminuição.

3.6.3 Processamento dos Dados

Na etapa de processamento de dados os dados são preparados para a etapa posterior, também descrita como uma das etapas mais demoradas do processo KDD.

A. Escolha dos Dados

As escolhas das informações foram realizadas de modo a atingir os objetivos conforme a seção 3.7.1. Os dados foram selecionados de maneira sucinta na tentativa de alçar os melhores resultados, com isso grande parte das informações não essenciais foram removidas, assim a seleção se deu por colunas que continham as informações e os atributos.

B. Remoção dos Dados Desnecessários

Grande parte do processo de processamento realizado sobre as informações foi através do *Excel* e *NotePad++*. Logo após esses processamentos os arquivos eram gerados no formato padrão do *Weka.* A Figura 30 apresenta os dados já com as remoções das colunas que julgamos desnecessários.

Figura 30 - Recorte do arquivo e remoção das colunas não essências

Total	**43,7**	**42,9**	**45,8**	**Total**	**24,4**	**26,9**	**17,5**
Porto Velho	39,7	39,7	39,4	Porto Velho	23,5	25,1	12,6
Rio Branco	32,5	32,8	30,7	Rio Branco	27,0	29,5	11,1
Manaus	36,9	36,4	40,6	Manaus	19,8	21,0	10,7
Boa Vista	39,3	38,8	44,9	Boa Vista	24,6	25,2	18,1
Belém	37,8	39,9	33,0	Belém	17,7	20,5	11,4
Macapá	36,5	37,7	27,6	Macapá	22,3	23,7	12,3
Palmas	38,0	38,5	36,1	Palmas	19,9	22,7	9,0
São Luís	42,3	40,8	46,1	São Luís	16,8	18,5	12,5
Teresina	43,1	40,8	47,7	Teresina	22,4	26,3	14,5
Fortaleza	41,0	43,4	35,9	Fortaleza	21,0	24,7	13,4
Natal	37,0	34,5	41,3	Natal	19,8	24,4	12,1
João Pessoa	38,9	36,5	42,7	João Pessoa	19,6	23,2	14,1
Recife	46,6	45,4	48,6	Recife	20,9	24,9	14,0
Maceió	46,4	44,0	48,9	Maceió	18,5	24,4	12,2
Aracaju	46,0	43,1	50,1	Aracaju	17,4	20,4	13,1
Salvador	46,6	47,6	44,6	Salvador	15,3	17,7	10,5

Fonte: MORAIS; SANTOS (2020).

3.6.4 Transformação dos Dados

Algumas colunas precisaram ser adaptadas como a quantidade de casas decimais em um número. Após a exclusão das colunas desnecessárias dos arquivos originais gerou-se um arquivo em *.csv* e uniu as informações através do *NotePad++* combinando assim as bases sobre drogas, bebidas e cigarros em um arquivo *.arff* conforme Figura 31.

Figura 31 - Trecho dos dados obtidos

```
@RELATION DROGAS_ESTADO

@ATTRIBUTE ESTADO {Rondonia,Acre,Amazonas,Roraima,Para,Amapa,Tocantins,Maranhao,Piaui,Ceara,'Rio Grande Do Norte'
@ATTRIBUTE TIPO_DROGAS {'Drogas ilicitas',Cigarro,Bebida}
@ATTRIBUTE DEPENDENCIA_ADM {Publica,Privada}
@ATTRIBUTE QTD_USUARIOS {1,2}

@DATA
Rondonia,Bebida,Publica,1
Rondonia,Bebida,Publica,1
Rondonia,Bebida,Publica,1
Rondonia,Bebida,Publica,1
Rondonia,Bebida,Publica,1
Rondonia,Bebida,Publica,1
Rondonia,Bebida,Publica,1
Rondonia,Bebida,Publica,1
Rondonia,Bebida,Publica,1
Rondonia,Bebida,Publica,1
Rondonia,Bebida,Publica,1
Rondonia,Bebida,Publica,1
Rondonia,Bebida,Publica,1
Rondonia,Bebida,Publica,1
Rondonia,Bebida,Publica,1
Rondonia,Bebida,Publica,1
Rondonia,Bebida,Publica,1
Rondonia,Bebida,Publica,1
Rondonia,Bebida,Publica,1
Rondonia,Bebida,Publica,1
Rondonia,Bebida,Publica,1
Rondonia,Bebida,Publica,1
Rondonia,Bebida,Publica,1
Rondonia,Bebida,Publica,1
Rondonia,Bebida,Publica,1
Rondonia,Bebida,Publica,1
Rondonia,Bebida,Publica,1
Rondonia,Bebida,Publica,1
Rondonia,Bebida,Publica,1
Rondonia,Bebida,Publica,1
Rondonia,Bebida,Publica,1
```

Fonte: MORAIS; SANTOS (2020).

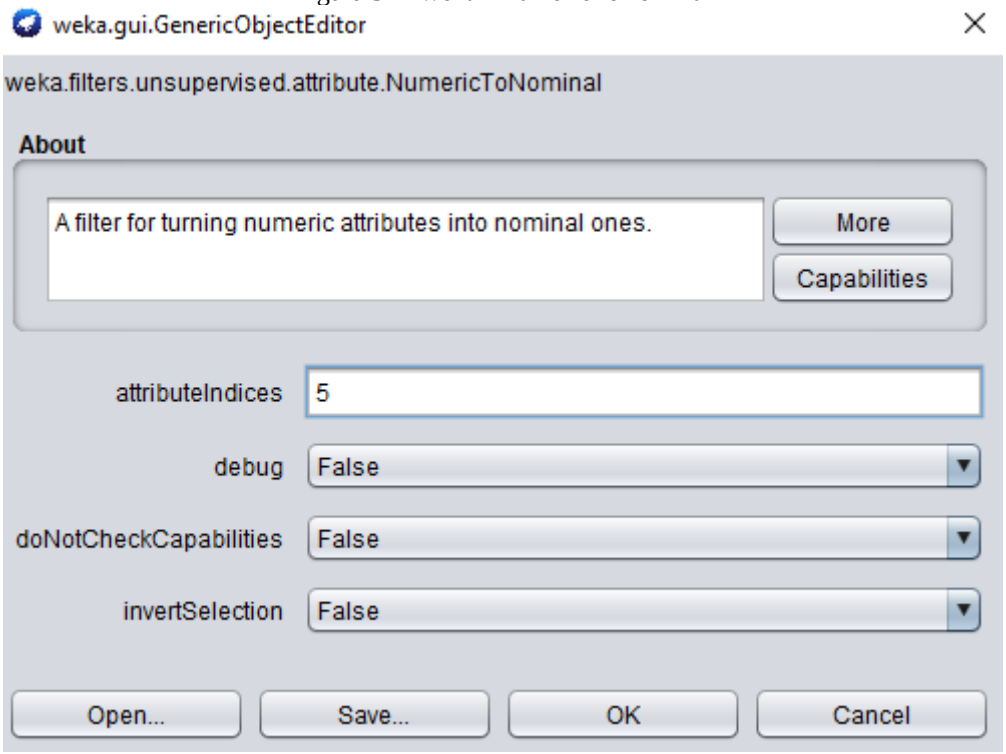
Em razão da definição de como as inserções e manipulações dos dados ocorreriam alguns erros foram previamente corrigidos e não houve a necessidade da utilização de ferramentas do *Weka.*

A. Organização dos dados

Logo após a primeira seleção dos arquivos, já com os dados convertidos para *.arff* foram importados para o *weka*. Nesta etapa do processo de KDD a organização e construção dos dados se mostrou um desafio pela necessidade de solucionar problemas nos arquivos gerados, em razão dos constantes erros na importação no ambiente do *Weka.*

Apesar de várias adaptações terem sido feitas anteriormente alguns ajustes ainda foram necessários. Estes mesmos ajustes foram realizados com a utilização do *Weka* conforme pode ser observado na Figura 32 onde foi utilizado opção de conversão dos dados numéricos para nominais.

Figura 32 - *Weka - NumericToNominal*



Fonte: MORAIS; SANTOS (2020).

B. Junção dos Dados

As bases de dados que estão sendo utilizadas foram dispostas no mesmo arquivo no momento de criação dos arquivos *.arff.*

C. Finalização dos Dados

Como parte final dessa etapa os dados do pré-processamento foram ordenados anteriormente na criação dos arquivos *.arff.* Na inserção destes dados seguiu-se o padrão que pode ser observado na Figura 33, e foram elaborados 3 arquivos individuais.

Figura 33 - Arquivo *.arff*

```
@RELATION DROGAS_ESTADO

@ATTRIBUTE ESTADO {Rondonia,Acre,Amazonas,Roraima,Para,Amapa,Tocantins,Maranhao,Piaui,Ceara,'Rio Grande Do Norte',Paraiba,Pernambuco,
@ATTRIBUTE TIPO_DROGAS {'Drogas ilicitas',Cigarro,Bebida}
@ATTRIBUTE DEPENDENCIA_ADM {Publica,Privada}
@ATTRIBUTE QTD_USUARIOS {1,2}

@DATA
Rondonia,Bebida,Publica,1
Rondonia,Bebida,Publica,1
Rondonia,Bebida,Publica,1
Rondonia,Bebida,Publica,1
Rondonia,Bebida,Publica,1
Rondonia,Bebida,Publica,1
Rondonia,Bebida,Publica,1
Rondonia,Bebida,Publica,1
Rondonia,Bebida,Publica,1
Rondonia,Bebida,Publica,1
Rondonia,Bebida,Publica,1
Rondonia,Bebida,Publica,1
Rondonia,Bebida,Publica,1
Rondonia,Bebida,Publica,1
Rondonia,Bebida,Publica,1
Rondonia,Bebida,Publica,1
Rondonia,Bebida,Publica,1
Rondonia,Bebida,Publica,1
Rondonia,Bebida,Publica,1
Amapa,'Drogas ilicitas',Publica,1
Amapa,'Drogas ilicitas',Publica,1
Amapa,'Drogas ilicitas',Publica,1
Amapa,'Drogas ilicitas',Publica,1
Amapa,'Drogas ilicitas',Publica,1
Amapa,'Drogas ilicitas',Publica,1
Amapa,'Drogas ilicitas',Publica,1
Amapa,'Drogas ilicitas',Privada,1
Amapa,'Drogas ilicitas',Privada,1
Amapa,'Drogas ilicitas',Privada,1
Tocantins,Cigarro,Publica,1
Tocantins,Cigarro,Publica,1
Tocantins,Cigarro,Publica,1
```

Fonte: MORAIS; SANTOS (2020).

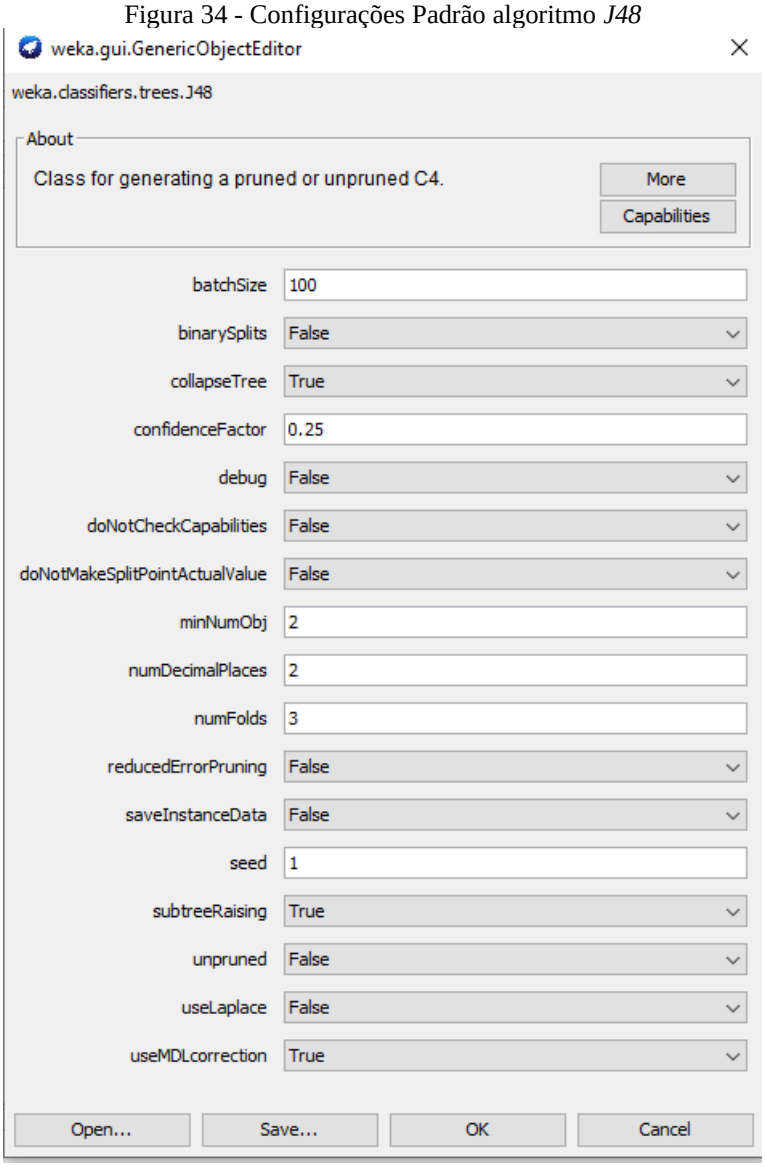
3.6.5 Mineração dos Dados

Após avaliadas as tarefas, os dados e os objetivos da tarefa de classificação, elas foram definidas para realizar a mineração dos dados.

A. Elaboração dos Modelos de Mineração

Conforme descrito na seção 3.5 sobre os modelos de classificação e o algoritmo *J48*, a construção foi realizada através da opção *Experimenter* da ferramenta *Weka*, com isso permitiu-se uma avaliação do modelo, de modo que a visualização dos dados ficou mais clara e objetiva. E assim as configurações foram sendo testadas para que os melhores resultados pudessem se atingidos, conforme a Figura 34 vemos a configuração padrão do *J48*.

Figura 34 - Configurações Padrão algoritmo *J48*



Fonte: MORAIS; SANTOS (2020).

Quadro 3 - Combinações do algoritmo *J48* para avaliação das bases.

trees.J48 '- C 0.25 – M 1 '	**trees.J48 '- C 0.2 – M 1 '**
trees.J48 '- C 0.25 – M 2 '	trees.J48 '- C 0.2 – M 2 '
trees.J48 '- C 0.25 – M 3 '	trees.J48 '- C 0.2 – M 3 '
trees.J48 '- C 0.15 – M 1 '	trees.J48 '-B -C 0.25 – M 1 '
trees.J48 '- C 0.15 – M 2 '	trees.J48 '-B -C 0.25 – M 2 '
trees.J48 '- C 0.15 – M 3 '	trees.J48 '-B -C 0.25 – M 3 '

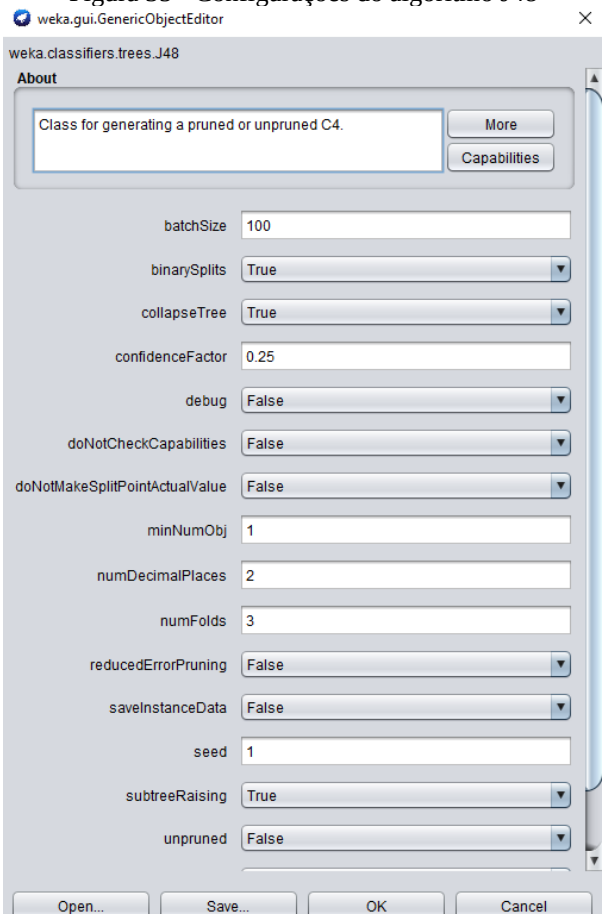
Fonte: MORAIS; SANTOS (2020).

Após os testes com o modelo padrão outros testes foram realizados com novas combinações e logo chegou-se a uma combinação de parâmetros que foram utilizados na avaliação do algoritmo conforme abaixo:

- *trees.J48 '-B -C 0.25 -M 1*

Com isso os outros testes serviram de justificativa para determinar a escolha das configurações e notou-se que, para esse estudo a combinação citada se mostrou adequada. Alguns parâmetros como *binarySplit*, *confidenceFactor e minNumObj* foram utilizados para permitir um melhor controle nas configurações do classificador de modo tentar gerar uma poda mais precisa nas árvores de decisão.

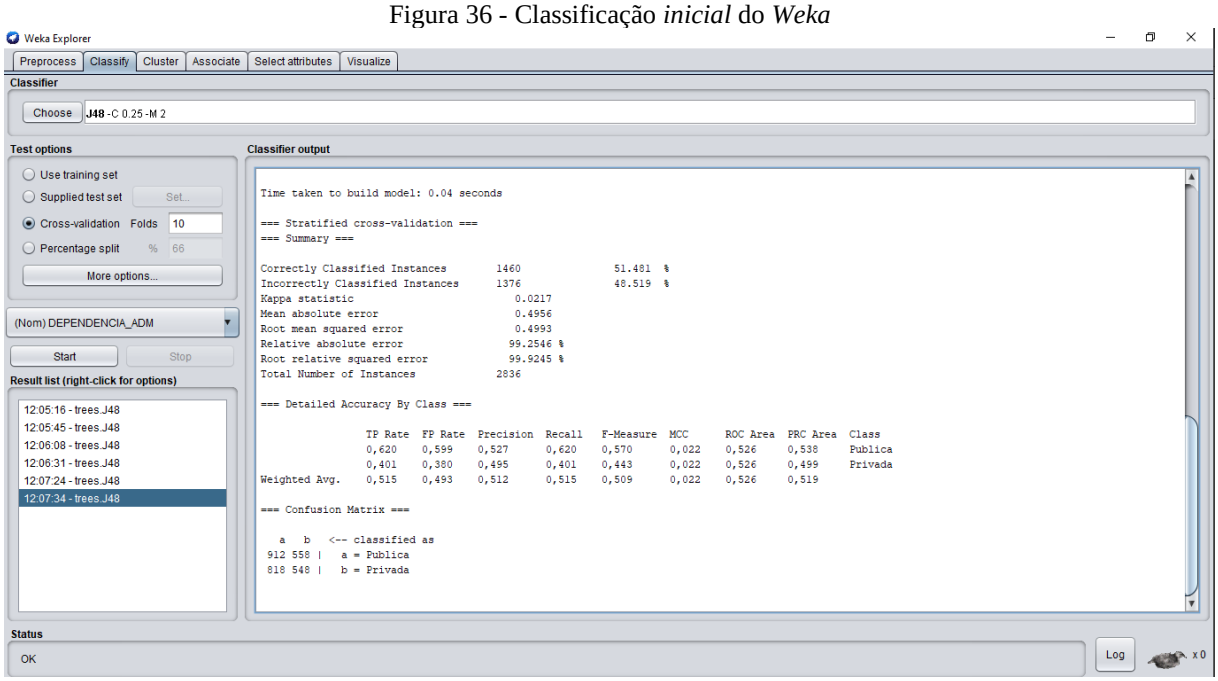
Figura 35 - Configurações do algoritmo *J48*



Fonte: MORAIS; SANTOS (2020).

Logo os dados criados conforme descrito na seção 3.7.4 foram processados no *Weka* através da aba *Classify*, gerando inicialmente alguns modelos.

Figura 36 - Classificação *inicial* do *Weka*



Fonte: MORAIS; SANTOS (2020).

Abaixo são apresentados os resultados dos modelos de classificação dos arquivos criados. Os resultados exibidos são baseados na combinação do algoritmo, técnica e tarefa relatados sendo que a respectiva combinação foi responsável por gerar o nível de acurácia maior dentre todas as testadas para o estudo.

A Figura 37 demonstra a classificação utilizando as colunas de região e dependência administrativa que apresentou uma acurácia de 50.67% .

Figura 37 - Modelo de classificação gerado por região_dependência administrativa

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

Number of Leaves  :     6
Size of the tree :      11

Time taken to build model: 0.02 seconds

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

Correctly Classified Instances        1436               50.6704 %
Incorrectly Classified Instances      1398               49.3296 %
Kappa statistic                          0.0042
Mean absolute error                      0.4956
Root mean squared error                  0.4994
Relative absolute error                 99.2521 %
Root relative squared error             99.9469 %
Total Number of Instances             2834

=== Detailed Accuracy By Class ===

               TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
               0,628    0,624    0,520      0,628    0,569      0,004    0,523     0,541     Publica
               0,376    0,372    0,484      0,376    0,423      0,004    0,523     0,490     Privada
Weighted Avg.  0,507    0,503    0,503      0,507    0,499      0,004    0,523     0,516

=== Confusion Matrix ===

   a   b   <-- classified as
 923 546 |   a = Publica
 852 513 |   b = Privada
```

Fonte: MORAIS; SANTOS (2020).

A Figura 38 demonstra a classificação utilizando as colunas de região e tipo de droga que apresentou uma acurácia de 64.50% .

Figura 38 - Modelo de classificação região_tipo de droga

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

Number of Leaves  :     3
Size of the tree :      5

Time taken to build model: 0.02 seconds

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

Correctly Classified Instances        1828               64.5025 %
Incorrectly Classified Instances      1006               35.4975 %
Kappa statistic                          0.0697
Mean absolute error                      0.3428
Root mean squared error                  0.4142
Relative absolute error                 96.9787 %
Root relative squared error             98.5428 %
Total Number of Instances             2834

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0,000    0,000    ?          0,000    ?          ?        0,510     0,148     Drogas ilicitas
                 0,069    0,012    0,644      0,069    0,125      0,154    0,503     0,271     Cigarro
                 1,000    0,931    0,645      1,000    0,784      0,211    0,527     0,639     Bebida
Weighted Avg.    0,645    0,588    ?          0,645    ?          ?        0,519     0,486

=== Confusion Matrix ===

    a    b    c   <-- classified as
    0   26  346 |    a = Drogas ilicitas
    0   47  634 |    b = Cigarro
    0    0 1781 |    c = Bebida
```

Fonte: MORAIS; SANTOS (2020).

A Figura 39 demonstra a classificação utilizando as colunas de estado e dependência administrativa que apresentou uma acurácia de 54,79%.

Figura 39 - Modelo de classificação estado_dependência

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

Number of Leaves  :     1
Size of the tree :      1

Time taken to build model: 0.06 seconds

=== Evaluation on training set ===

Time taken to test model on training data: 0.03 seconds

=== Summary ===

Correctly Classified Instances        2239               54.7969 %
Incorrectly Classified Instances      1847               45.2031 %
Kappa statistic                          0
Mean absolute error                      0.4954
Root mean squared error                  0.4977
Relative absolute error                 99.9995 %
Root relative squared error            100      %
Total Number of Instances             4086

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 1,000    1,000    0,548      1,000    0,708      ?        0,500     0,548     Publica
                 0,000    0,000    ?          0,000    ?          ?        0,500     0,452     Privada
Weighted Avg.    0,548    0,548    ?          0,548    ?          ?        0,500     0,505

=== Confusion Matrix ===

    a    b   <-- classified as
 2239    0 |    a = Publica
 1847    0 |    b = Privada
```

Fonte: MORAIS; SANTOS (2020).

A Figura 40 demonstra a classificação utilizando as colunas de estado e tipo de droga que apresentou uma acurácia de 69.13%.

Figura 40 - Modelo de classificação estado_tipo_de_droga

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

Number of Leaves  :     1
Size of the tree :      1

Time taken to build model: 0.06 seconds
=== Evaluation on training set ===

Time taken to test model on training data: 0.03 seconds
=== Summary ===

Correctly Classified Instances        2825               69.1385 %
Incorrectly Classified Instances      1261               30.8615 %
Kappa statistic                          0
Mean absolute error                      0.3121
Root mean squared error                  0.395
Relative absolute error                 99.9689 %
Root relative squared error            100      %
Total Number of Instances             4086

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0,000    0,000    ?          0,000    ?          ?        0,500     0,099     Drogas ilicitas
                 0,000    0,000    ?          0,000    ?          ?        0,500     0,210     Cigarro
                 1,000    1,000    0,691      1,000    0,818      ?        0,500     0,691     Bebida
Weighted Avg.    0,691    0,691    ?          0,691    ?          ?        0,500     0,532

=== Confusion Matrix ===

    a    b    c   <-- classified as
    0    0  403 |    a = Drogas ilicitas
    0    0  858 |    b = Cigarro
    0    0 2825 |    c = Bebida
```

Fonte: MORAIS; SANTOS (2020).

A Figura 41 demonstra a classificação utilizando as colunas de segurança e dependência administrativa que apresentou uma acurácia de 69.18%.

Figura 41 - Modelo de classificação segurança_dependência_administrativa

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

Number of Leaves  :     1
Size of the tree :      1

Time taken to build model: 0.06 seconds
=== Evaluation on training set ===

Time taken to test model on training data: 0.03 seconds
=== Summary ===

Correctly Classified Instances        2825               69.1385 %
Incorrectly Classified Instances      1261               30.8615 %
Kappa statistic                          0
Mean absolute error                      0.3121
Root mean squared error                  0.395
Relative absolute error                 99.9689 %
Root relative squared error            100      %
Total Number of Instances             4086

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0,000    0,000    ?          0,000    ?          ?        0,500     0,099     Drogas ilicitas
                 0,000    0,000    ?          0,000    ?          ?        0,500     0,210     Cigarro
                 1,000    1,000    0,691      1,000    0,818      ?        0,500     0,691     Bebida
Weighted Avg.    0,691    0,691    ?          0,691    ?          ?        0,500     0,532

=== Confusion Matrix ===

    a    b    c   <-- classified as
    0    0  403 |    a = Drogas ilicitas
    0    0  858 |    b = Cigarro
    0    0 2825 |    c = Bebida
```

Fonte: MORAIS; SANTOS (2020).

Com base nos modelos acima, algumas árvores foram geradas, com isso obtivemos dentre as árvores geradas o exemplo conforme a Figura 43.

Figura 42 - Árvore de decisão

```
TIPO_DROGAS = Bebida
|   REGIAO = Nordeste: Privada (359.0/158.0)
|   REGIAO != Nordeste
|   |   FAIXA_ETARIA = 16 A 17: Privada (602.0/298.0)
|   |   FAIXA_ETARIA != 16 A 17: Publica (822.0/398.0)
TIPO_DROGAS != Bebida
|   FAIXA_ETARIA = 16 A 17
|   |   REGIAO = Norte: Publica (68.0/30.0)
|   |   REGIAO != Norte: Privada (412.0/196.0)
|   FAIXA_ETARIA != 16 A 17: Publica (573.0/217.0)
```

Fonte: MORAIS; SANTOS (2020).

3.6.6 Avaliação dos Resultados

Após todo o processo de mineração de dados, notou-se que as informações geradas foram inconclusivas, com isso não se obtiveram os resultados esperados, logo algumas medidas foram adotadas durante todo o trabalho para que se pudesse chegar aos resultados propostos, buscando sempre aproximar-se do esperado.

Desde a junção de tabelas, testes de diferentes padrões do algoritmo *C4.5*, até mesmo alterações e adequações pertinentes nas tabelas usadas para a mineração.

Figura 43 - Modelo de arquivos *.arff* gerado

```
@RELATION DROGAS_ILICITAS_IDADE

@ATTRIBUTE REGIAO_IDADE{'Brasil 13 A 17 ANOS','Norte 13 A 17 ANOS','Nordeste 13 A 17 ANOS',
@ATTRIBUTE TOTAL NUMERIC
@ATTRIBUTE TOTAL_MASCULINO NUMERIC
@ATTRIBUTE TOTAL_FEMININO NUMERIC
@ATTRIBUTE TOTAL_PUBLICA NUMERIC
@ATTRIBUTE TOTAL_PRIVADA NUMERIC

@DATA
'Brasil 13 A 17 ANOS',12,13,11,12,12
'Norte 13 A 17 ANOS',8,11,5,9,4
'Nordeste 13 A 17 ANOS',9,10,8,9,9
'Sudeste 13 A 17 ANOS',13,13,13,13,13
'Sul 13 A 17 ANOS',17,16,18,17,15
'Centro-Oeste 13 A 17 ANOS',13,13,13,13,12
'Brasil 13 A 15 ANOS',9,9,9,10,6
'Norte 13 A 15 ANOS',7,9,5,8,2
'Nordeste 13 A 15 ANOS',6,8,5,7,3
'Sudeste 13 A 15 ANOS',10,9,11,11,7
'Sul 13 A 15 ANOS',13,10,15,13,10
'Centro-Oeste 13 A 15 ANOS',10,10,11,11,8
'Brasil 16 A 17 ANOS',17,19,15,16,21
'Norte 16 A 17 ANOS',11,15,5,11,11
'Nordeste 16 A 17 ANOS',13,14,13,13,13
'Sudeste 16 A 17 ANOS',18,21,15,16,27
'Sul 16 A 17 ANOS',24,25,23,24,24
'Centro-Oeste 16 A 17 ANOS',18,18,17,17,20
```

Fonte: MORAIS; SANTOS (2020).

A Figura 43 mostra o primeiro modelo de arquivo *.arff* gerado para importação no *Weka*. Percebeu-se, no entanto, que a quantidade de dados era demasiadamente pequena para obtenção de resultado útil. Após analisado o problema constatou-se a necessidade da junção de tabelas com dados similares, assim foi tomada a decisão de juntas as tabelas: Amostra_1_Tema_12_Cigarro, Amostra_1_Tema_13_Bebidas_Alcoolicas e Amostra_1_Tema_14_Drogas_Ilicitas, formando assim uma única tabela, o mesmo foi feito com as tabelas Amostra_2_Tema_04_Cigarro, Amostra_2_Tema_05_Bebidas_Alcoolicas, Amostra_2_Tema_11_Seguranca, gerando um arquivo *.arff* conforme a Figura 44.

Figura 44 - Modelo de arquivos *.arff* gerado

```
@RELATION DROGAS_CIGARRO_BEBIDA1

@ATTRIBUTE REGIAO_IDADE{'Drogas Brasil 13 A 17 ANOS','Drogas Norte 13 A 17 ANOS','Drogas Nordest
@ATTRIBUTE TOTAL NUMERIC
@ATTRIBUTE TOTAL_MASCULINO NUMERIC
@ATTRIBUTE TOTAL_FEMININO NUMERIC
@ATTRIBUTE TOTAL_PUBLICA NUMERIC
@ATTRIBUTE TOTAL_PRIVADA NUMERIC

@DATA
'Drogas Brasil 13 A 17 ANOS',12,13,11,12,12
'Drogas Norte 13 A 17 ANOS',8,11,5,9,4
'Drogas Nordeste 13 A 17 ANOS',9,10,8,9,9
'Drogas Sudeste 13 A 17 ANOS',13,13,13,13,13
'Drogas Sul 13 A 17 ANOS',17,16,18,17,15
'Drogas Centro-Oeste 13 A 17 ANOS',13,13,13,13,12
'Drogas Brasil 13 A 15 ANOS',9,9,9,10,6
'Drogas Norte 13 A 15 ANOS',7,9,5,8,2
'Drogas Nordeste 13 A 15 ANOS',6,8,5,7,3
'Drogas Sudeste 13 A 15 ANOS',10,9,11,11,7
'Drogas Sul 13 A 15 ANOS',13,10,15,13,10
'Drogas Centro-Oeste 13 A 15 ANOS',10,10,11,11,8
'Drogas Brasil 16 A 17 ANOS',17,19,15,16,21
'Drogas Norte 16 A 17 ANOS',11,15,5,11,11
'Drogas Nordeste 16 A 17 ANOS',13,14,13,13,13
'Drogas Sudeste 16 A 17 ANOS',18,21,15,16,27
'Drogas Sul 16 A 17 ANOS',24,25,23,24,24
'Drogas Centro-Oeste 16 A 17 ANOS',18,18,17,17,20
'Cigarro Brasil 13 A 17 ANOS',23,24,22,23,19
'Cigarro Norte 13 A 17 ANOS',21,23,18,22,9
'Cigarro Nordeste 13 A 17 ANOS',18,19,17,18,17
'Cigarro Sudeste 13 A 17 ANOS',24,26,23,25,22
'Cigarro Sul 13 A 17 ANOS',29,27,30,30,20
'Cigarro Centro-Oeste 13 A 17 ANOS',28,28,27,29,20
'Cigarro Brasil 13 A 15 ANOS',19,19,19,20,14
'Cigarro Norte 13 A 15 ANOS',18,19,16,20,7
'Cigarro Nordeste 13 A 15 ANOS',14,14,14,14,8
```

Fonte: MORAIS; SANTOS (2020).

O arquivo gerado possui 6 atributos e 54 instâncias, que após todo o pré-processamento e as demais etapas da mineração, gerou uma acurácia de 0%. Após este resultado houve a necessidade de ser feita uma nova análise nos dados, com isso constatou-se algumas mudanças necessárias, que foram a retirada dos atributos TOTAL_FEMININO TOTAL_MASCULINO e TOTAL, por conta de algumas inconsistências encontradas nesses dados. A principal inconsistência verificada é que na maioria das vezes a soma dos valores das colunas não resultavam no valor esperado, que seria a soma das colunas TOTAL_PUBLICA e TOTAL_PRIVADA. Outra mudança necessária foi no atributo REGIAO_IDADE que eram descritas da seguinte maneira: ‘Drogas Norte 13 A 17 ANOS’ e teve que ser readequada e desmembrada, o que antes era um único atributo se tornaram três, sendo eles TIPO_DROGA, REGIAO e FAIXA_ETARIA. Uma última alteração nos dados foi necessária que era o formato da disposição das variáveis do tipo numérico. A linha totaliza um conjunto de informações, como exemplo o valor TOTAL_PRIVADA, no campo ‘Drogas Norte 13 A 17 ANOS’, continha o valor 9, essa coluna de valor foi alterada para 1 e a linha copiada 9 vezes. Após todas essas alterações foi gerado o arquivo demonstrado na Figura 45.

Figura 45 - Modelo de arquivos *.arff* gerado

```
@RELATION DROGAS_REGIAO_IDADE

@ATTRIBUTE REGIAO {Norte,Nordeste,Sudeste,Sul,'Centro Oeste'}
@ATTRIBUTE FAIXA_ETARIA {'13 A 17','13 A 15','16 A 17'}
@ATTRIBUTE TIPO_DROGAS {'Drogas ilicitas',Cigarro,Bebida}
@ATTRIBUTE DEPENDENCIA_ADM {Publica,Privada}
@ATTRIBUTE QTD_USUARIOS NUMERIC

@DATA
Norte,'13 A 17','Drogas ilicitas',Publica,1
Norte,'13 A 17','Drogas ilicitas',Publica,1
Norte,'13 A 17','Drogas ilicitas',Publica,1
Norte,'13 A 17','Drogas ilicitas',Publica,1
Norte,'13 A 17','Drogas ilicitas',Publica,1
Norte,'13 A 17','Drogas ilicitas',Publica,1
Norte,'13 A 17','Drogas ilicitas',Publica,1
Norte,'13 A 17','Drogas ilicitas',Publica,1
Norte,'13 A 17','Drogas ilicitas',Publica,1
Norte,'13 A 17','Drogas ilicitas',Privada,1
Norte,'13 A 17','Drogas ilicitas',Privada,1
Norte,'13 A 17','Drogas ilicitas',Privada,1
Norte,'13 A 17','Drogas ilicitas',Privada,1
Nordeste,'13 A 17','Drogas ilicitas',Publica,1
Nordeste,'13 A 17','Drogas ilicitas',Publica,1
Nordeste,'13 A 17','Drogas ilicitas',Publica,1
Nordeste,'13 A 17','Drogas ilicitas',Publica,1
Nordeste,'13 A 17','Drogas ilicitas',Publica,1
Nordeste,'13 A 17','Drogas ilicitas',Publica,1
Nordeste,'13 A 17','Drogas ilicitas',Publica,1
Nordeste,'13 A 17','Drogas ilicitas',Publica,1
Nordeste,'13 A 17','Drogas ilicitas',Privada,1
Nordeste,'13 A 17','Drogas ilicitas',Privada,1
Nordeste,'13 A 17','Drogas ilicitas',Privada,1
Nordeste,'13 A 17','Drogas ilicitas',Privada,1
Nordeste,'13 A 17','Drogas ilicitas',Privada,1
Nordeste,'13 A 17','Drogas ilicitas',Privada,1
Nordeste,'13 A 17','Drogas ilicitas',Privada,1
```

Fonte: MORAIS; SANTOS (2020).

Constituindo um total de 5 atributos e 2836 instâncias, com uma acurácia de 51,48% rodando o algoritmo *C4.5* no padrão -B -C 0.25 -M1, foi realiza então o teste de diferentes padrões do algoritmo *C4.5* que geraram os seguintes resultados, para tentar encontrar outros resultados conforme o Quadro 4.

Quadro 4 - Padrões do algoritmo *J48*.

-C 0.25 -M1 = 50,21%	-C 0.15 -M2 = 50,77%	-C 0.2 -M3 = 50,59%
-C 0.25 -M2 = 49,85%	-C 0.15 -M3 = 50,77%	-B -C 0.25 -M1 = 51,48%
-C 0.25 -M3 = 49,85%	-C 0.2 -M1 = 50,56%	-B -C 0.25 -M2 = 51,45%
-C 0.15 -M1 = 50,77%	-C 0.2 -M1 = 50,59%	-B -C 0.25 -M3 = 51,45%

Fonte: MORAIS; SANTOS (2020).

Após executar todas as etapas do processo KDD de maneira constante e repetitiva sobre os dados a acurácia manteve uma média de 50%. Esperávamos uma acurácia maior, porém estes dados ainda não haviam sido minerados. Assim, verificando o momento e as possibilidades optou-se por finalizar o estudo por falta de tempo, pois seria necessário agregar ao estudo novas bases e que pudessem apresentar mais qualidade nos dados.

3. CONSIDERAÇÕES FINAIS

Com base no estudo realizado, é possível utilizar a mineração de dados para a descoberta de conhecimento em diversas áreas que contenha grandes quantidades de dados. Nas etapas da mineração é possível analisar, moldar e verificar novos caminhos a serem seguidos uma vez que durante o processo várias situações podem ocorrer, modificando a linha de pensamento e sendo necessário um desvio para se chegar aos resultados esperados ou ainda a possibilidade de que os dados minerados não gerem resultados satisfatórios ou conclusivos.

Durante toda essa pesquisa buscou-se identificar ou revelar padrões até o presente momento não identificados para serem utilizados no tratamento do problema drogas nas escolas, empregando a classificação e geração de árvores de decisão e utilização do algoritmo *C4.5 ou J48.*

Contudo durante as etapas de execução da mineração notou-se que os dados coletados não forneceriam dados conclusivos ao ponto de se obter métricas com qualidade suficiente para a declaração de padrões e ou identificação de anomalias, logo a compreensão dos dados seria inconclusiva para os resultados esperados.

Durante a execução das etapas de treinamento os dados apresentaram algumas boas respostas, no entanto não superando os 70% de acurácia. Também durante a etapa de mineração ficou evidente que nem todas as informações contidas seriam úteis para o processo de mineração, o que gerou uma perda significativa de dados.

Contudo conclui-se que a falta de dados atualizados e a baixa qualidade das informações disponíveis sobre o assunto dificulta os processos de descoberta de conhecimento, dificultando assim a aplicação de medidas eficazes no combate ao uso de drogas lícitas e ilícitas por parte dos estudantes.

REFERÊNCIAS BIBLIOGRÁFICAS

ABRAMOVAY, M.; CASTRO, M. G. Drogas nas escolas: versão resumida. **Unesco**, 2005. Disponivel em: <https://unesdoc.unesco.org/ark:/48223/pf0000139387>. Acesso em: 27 Março 2020.

ABRANTES, B. Drogas: o que são, tipos e classificação! **Stoodi**, 09 Novembro 2018. Disponivel em: <https://www.stoodi.com.br/blog/2018/11/09/drogas/>. Acesso em: 28 Abril 2020.

AGRAWAL, R.; IMIELINSKI, T.; SWAMI, A. **Mining Association Rules beween Sets of Items in Large Databases.** Proceedings of the ACM SIGMOD Conference. Washington, USA, 1993.

BATISTA, B. **Machine Learning sem cógido:** Usando Orange Data Mining para criar um modelo preditivo sem usar código, 2019. Disponível em: < https://medium.com/ensina-ai/machine-learning-sem-código-636d1a8f9081#:~:text=Orange%20Data%20Mining%20é%20uma,mining%2C%20sem%20necessidade%20de%20código. >. Acessp em 16 de novembro de 2020

BERRY, M. J. A.; LINOFF, G. **Data Mining Techniques:** for Marketing, Sales, and Customer Support. 3. Ed. Indianapolis: John Wiley & Son, 1997.

BROWN, M. Técnicas de Mineração de Dados . **IBM**, 2012. Disponível em: <https://www.ibm.com/developerworks/br/library/tecnicas-mineracao-de-dados/index.html>. Acesso em: 23 de maio 2020.

BRASIL. LEI Nº 11.343, DE 23 DE AGOSTO DE 2006. **Institui o Sistema Nacional de Políticas Públicas sobre Drogas - Sisnad**, 23 agosto 2006. Disponivel em: <http://www.planalto.gov.br/ccivil_03/_ato2004-2006/2006/lei/l11343.htm#>. Acesso em: 20 novembro 2019.

BRASIL. LEI Nº 13.106, DE 17 DE MARÇO DE 2015. **Estatuto da Criança e do Adolescente**, 17 março 2015. Disponivel em: <http://www.planalto.gov.br/ccivil_03/_Ato2015-2018/2015/Lei/L13106.htm>. Acesso em: 23 novembro 2019.

CARDOSO, L. R. D.; MALBERGIER, A. Problemas escolares e o consumo de álcool e outras drogas entre adolescentes. **Scielo**, 25 Maio 2012. Disponivel em: <http://www.scielo.br/scielo.php?script=sci_arttext&pid=S1413-85572014000100003>. Acesso em: 15 novembro 2019.

CARLINI, E. L. D. A. et al. **VI Levantamento Nacional sobre o Consumo de Drogas Psicotrópicas entre Estudantes do Ensino Fundamental e Médio das Redes Pública e Privada de Ensino nas 27 Capitais Brasileiras**. SENAD - Secretaria Nacional de Políticas sobre Drogas. Brasilia, DF, p. 506. 2010.

CEBRID. Livreto Informativo Sobre Drogas Psicotrópicas. **Cebrid - Centro Brasileiro de Informações sobre Drogas Psicotrópicas**, 12 Dezembro 2012. Disponivel em:

<https://www.cebrid.com.br/wp-content/uploads/2012/12/Livreto-Informativo-sobre-Drogas-Psicotrópicas.pdf>. Acesso em: 28 Abril 2020.

CIOS, K. J. et al. **Data Mining – A Knowledge Dsicovery Approach**. Springer, 2007.

COLÉGIO WEB. Tipos de Drogas, Causas e Tratamento. **Colégio Web**, 20 Junho 2012. Disponivel em: <https://www.colegioweb.com.br/saude/drogas-2.html>. Acesso em: 28 Abril 2020.

CONCEITOS de mineração de dados. **Microsoft**, 2019. Disponível em: <https://docs.microsoft.com/pt-br/analysis-services/data-mining/data-mining-concepts?view=asallproducts-allversions>. Acesso em: 22 de maio 2020.

COSTA, E. et al. **Mineração de dados educacionais:** conceitos, técnicas, ferramentas e aplicações. Jornada de atualização em informática na Educação, 2013.

DAVENPORT, T. H.; PRUSAK, L. **Conhecimento Empresaria**l: como as organizações gerenciam o seu capital intelectual. Rio de Janeiro: Campus, 1998.

DENARC. Drogas. **Denarc - Divisão Estadual de Narcóticos**, [2015?]. Disponivel em: <http://www.denarc.pr.gov.br/modules/conteudo/conteudo.php?conteudo=40>. Acesso em: 28 Abril 2020.

DILLY, R. **Data Mining:** An Introduction. Disponível em: <https://www.adt.unipd.it/corsi/Bianco/www.pcc.qub.ac.uk/tec/courses/datamining/ohp/dm-OHP-final_1.html>. Acesso em: 20 de maio 2020.

ELMASRI, R.; NAVATHE, S. B. **SISTEMAS DE BANCO DE DADOS**. 4º. ed. São Paulo : Pearson , v. II, 2004.

FAYYAD, U. M. et al. From data mining to Knowledge Discovery: na overview. In: **Advances in knowledge discovery and data mining.** California: AAAI/The MIT, 1996.

FUNDO MONETÁRIO INTERNACIONAL. **Relatório Anual do FMI 2019**. Fundo Monetário Internacional. Washington, D.C. 20431 EUA, p. 32. 2019.

GIL, A. C. **Métodos e Técnicas de Pesquisa Social**. 6°. ed. São Paulo: Editora Atlas S.A, 2008.

FURLAN, M. B. **Algoritmos e Técnicas para Mineração de Dados**. Assis: 2018.

GONÇALVES, A. L. UNIVERSIDADE FEDERAL DE SANTA CATARINA, Programa de Pós-graduação em Engenharia de Produção. **Utilização de técnicas de mineração de dados em bases de C&T**: uma análise dos grupos de pesquisa no Brasil, 2000.

GOLDSCHMIDT, R.; PASSOS, E. **Data Mining**: Um guia prático. Rio de Janeiro: Elsevier, 2005.

GONÇALVES, P. T. B. M. B. **Drogas lícitas e ílicitas**. Disponível em: <https://www.marinha.mil.br/saudenaval/teste/content/drogas-l%C3%ADcitas-e-il%C3%ADcitas> Acesso em 29 de maio 2020.

HAN, J.; KAMBER, M. **Data Mining – Concepts ans Techniques.** Morgan Kaufmann Publishers, 2001.

HASHIMOTO, A. N. DADO, INFORMAÇÃO E CONHECIMENTO. **KMOL**, 25 Setembro 2009. Disponivel em: <https://kmol.pt/artigos/2009/09/25/dado-informacao-conhecimento/>. Acesso em: 27 Maio 2020.

HAYKIN, S. **Redes Neurais**: Princípios e Prática. 2. ed. Porto Alegre, Bookman, 2001.

IBGE. Pesquisa Nacional de Saúde do Escolar - PeNSE. **Instituto Brasileiro de Geografia e Estatísticas**, 2015. Disponivel em: <https://www.ibge.gov.br/estatisticas/sociais/educacao/9134-pesquisa-nacional-de-saude-do-escolar.html?=&t=downloads>. Acesso em: 20 fev. 2020.

ICICT;FIOCRUZ. **III Levantamento Nacional Sobre o Uso de Drogas Pela População Brasileira**. FIOCRUZ - Fundação Osvaldo Cruz. [S.l.], p. 528. 2017.

INFOPÉDIA Dicionário da Língua Portuguesa: **Droga** Porto: Porto Editora, 2003. Disponível em <: https://www.infopedia.pt/dicionarios/lingua-portuguesa/droga> Acesso em: 29 de Maio 2020

JÚNIOR, M. C. Arquitetura Simples de um SGBD. **Researchgate**, Setembro 2018. Disponivel em: <https://www.researchgate.net/figure/Figura-11-Arquitetura-Simples-de-um-SGBD_fig1_327582653>. Acesso em: 27 Maio 2020.

LEVY, E. **The Lowndonw on Data Mining**. Teradatareview, Summer, 1999.
MEDINA, M.; FERTIG, C. **Algoritmos e Programação**: Teoria e Prática. São Paulo: Novatec Editora, 2006.

NEVES, R. C. D. **Estudo de Metodologias de Descoberta de Conhecimento em Banco de Dados**. 2003.

ORACLE. O que É um Banco de Dados Relacional? **Oracle**, 31 Julho 2014. Disponivel em: <https://www.oracle.com/br/database/what-is-a-relational-database/>. Acesso em: 27 Maio 2020.

PASTA, Arquelau. **Aplicação de Técnica de Data Mining na base de dados do ambiente de gestão educacional**: um estudo de caso de uma instituição de ensino supervisor de Blumenau – SC. Dissertação de Mestrado, UNIVALI, São José, 2011.

PERERA, L. C. J. et al. **Uma Análise em Data Mining: Árvores de Decisão, Redes Neurais e Support Vector Michines**. Rio de Janeiro, 2011.

PINTO, D. D. O. Pisa – Ranking de educação mundial: entenda os dados do Brasil. **Blog Lyceum**, 05 dezembro 2019. Disponivel em: <https://blog.lyceum.com.br/ranking-de-educacao-mundial-posicao-do-brasil/>. Acesso em: 19 novembro 2019.

ROMAO, L. **Análise do uso de técnicas de pré-processamento de dados em algoritmos para classificação de proteínas General Terms.** 2016.

QUINLAN, J. R. **C4.5**: Programs for Machine Learning. Morgan Kaufmann Publishers, 1993.

REUTERS, L. Dado, informação e conhecimento: você sabe a diferença? **Pandora Soluções**, 13 Fevereiro 2017. Disponivel em: <https://blog.pandora.com.br/dado-informacao-e-conhecimento-voce-sabe-diferenca/>. Acesso em: 27 Maio 2020.

REZENDE, E. Dados, Informação e Conhecimento. O que são? **Eliana Rezende**, 10 Novembro 2015. Disponivel em: <https://eliana-rezende.com.br/dados-informacao-e-conhecimento-o-que-sao/>. Acesso em: 28 Maio 2020.

REZENDE, S. O. **Sistemas Inteligentes**: fundamentos e aplicações. 1. ed. Barueri: Editora Manole Ltda, 2003.

SILVEIRA, D. X. D.; DOERING-SILVEIRA, E. B. PADRÕES DE USO DE DROGAS: Eixo Políticas e Fundamentos. **SENAD**, 24 Abril 2017. Disponivel em: <http://www.aberta.senad.gov.br/medias/original/201704/20170424-094251-001.pdf>. Acesso em: 28 Abril 2020.

SHEDROFF, Nathan. **Information Interaction Design:** A Unified Field Theory. 1999. Disponível em:< https://nathan.com/wp/wp-content/uploads/2014/03/Screen-Shot-2015-07-02-at-2.56.18-PM.png/> Acesso em: 27 maio 2020.

SYACHRANI, S.; JEONG, H. S. D.; CHUNG, C. S. **Decision tree-based deterioration model for buried wastewater pipelines**. Journal of Performance of Constructed Facilities, v. 27, n. 5, p. 636, 2012.