

# Técnicas de Mineração de Dados Aplicadas na Análise de Dados de Fermentação de Vinhos.

Marcelo Sachet, Scheila de Avila e Silva

Área do Conhecimento de Ciências Exatas e Engenharias  
Campus Universitário da Região dos Vinhedos (CARVI)  
Universidade de Caxias do Sul (UCS)

Alameda João Dal Sasso, 800 – 95705-266 – Bento Gonçalves – RS – Brasil

{marcelosachets@gmail.com, sasilva6@ucs.br}

**Resumo.** *A aplicação de técnicas de mineração de dados é um fator determinante que pode auxiliar a tomada de decisão. O presente trabalho tem por objetivo aplicar técnicas de mineração de dados nas análises de processos de fermentação de vinhos. A exploração destes dados revela informações importantes sobre as características de diferentes leveduras neste processo. Os diferentes aromas, sabores, cores produzidas, alteram a percepção das relações organolépticas dos vinhos. Algoritmos de mineração de dados, como árvore de decisão e Apriori, relevam resultados que antes estavam no conhecimento tácito dos especialistas de domínio. Sendo assim, é possível reforçar a relação de diferentes leveduras que influência no gosto, cheiro e acidez presentes na composição vinífica. É demonstrado ainda, a associação que os diversos compostos químicos provocam no processo de fermentação.*

## 1. Introdução

A sociedade vive um novo paradigma, permeado pelo rápido e fácil acesso a todo tipo de informação. Todas as ações e conexões produzem dados, os quais podem ser armazenados em um banco de dados. Este recurso computacional auxiliar no processo de acesso à informação, podendo ser definido como uma coleção de registros, que possuem um significado implícito, ou seja, os dados possuem uma origem, uma relação com eventos do mundo real e alguém interessado em seu conteúdo (ELMASRI e NAVATHE, 2011).

Embora a tecnologia permita o acesso às informações, isso por si só não consolida o conhecimento. Com a crescente complexidade dos problemas que necessitam ser tratados e, a grande quantidade de dados gerados, se faz necessária a utilização de ferramentas que auxiliam a tomada de decisão e a obtenção de conhecimento. Neste contexto, são aplicadas as técnicas de mineração de dados, as quais podem ser definidas como um processo que analisa e obtém informações passadas ou decorrentes de grandes bases de dados, gerando possíveis predisposições de informação (GAMA et al, 2015; AMARAL, 2016).

No contexto da pesquisa científica aplicada a problemas biológicos, a quantidade de dados das pesquisas gera a informação necessária para o resultado desejado. No entanto, esses elementos podem carregar informações ocultas, gerando possíveis tendências. As técnicas de mineração de dados, aplicadas com os algoritmos corretos, possibilitam a esse conjunto de dados revelar informações ocultas, descobrir tendências, erros, combinações e expor características, sendo os resultados validados por um especialista de domínio. Por exemplo, dados de uma análise de maturação da uva, são

formados por informações que acabam não sendo relevantes para o resultado final, mas se analisadas em conjunto com outros dados podem revelar outras informações. De acordo com Braga (2005), diante do conhecimento adquirido, deve-se ser perspicaz para extrair os objetivos de uma análise, podendo estes serem alcançados pela mineração de dados. Todo o conhecimento obtido poderá contribuir com a sustentabilidade e competitividade da cadeia produtiva de vinhos no Brasil e aprimorar o conhecimento científico a área vinífica. Assim, a questão norteadora deste trabalho foi: “Como a mineração de dados pode contribuir para o conhecimento científico de experimentos realizados nas análises de dados de fermentação de vinhos? Associado a esta questão, foi elencado o seguinte objetivo geral, aplicar as técnicas de mineração de dados a fim de extrair conhecimento e criar novas hipóteses em dados provenientes de duas pesquisas científicas relacionadas à fermentação vinícola.

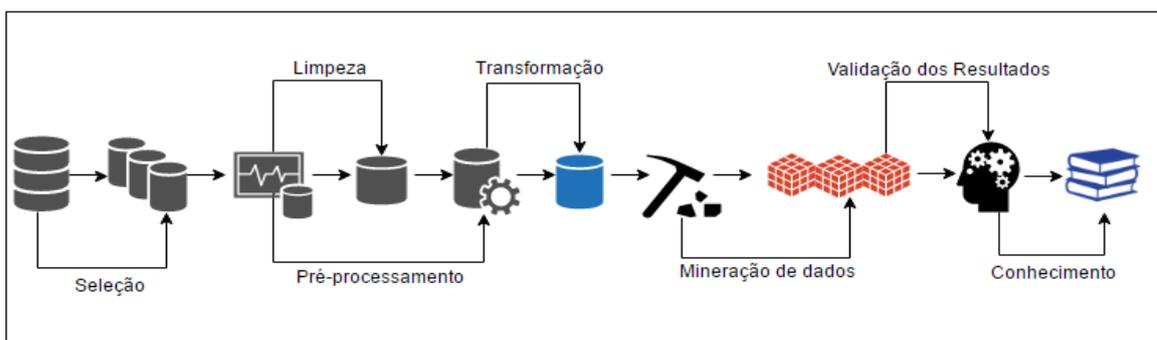
O presente trabalho está dimensionado através da seção 2 que apresenta uma breve relação da descoberta de conhecimento em banco dados KDD (*Knowledge Discovery in Database*) e a mineração de dados. A seção apresenta ainda as etapas do processo de KDD, as técnicas de mineração de dados e os trabalhos relacionados. A seção 3 descreve os dados das pesquisas realizadas, a ferramenta utilizada para o processo de mineração de dados, os algoritmos utilizados e as técnicas de validações e treinamentos dos algoritmos. A seção 4 apresenta os resultados do estudo, e por fim, na seção 5, é apresentada a conclusão do trabalho.

## 2. Fundamentação Teórica

### 2.1. Mineração de Dados e KDD

Conforme Fayyad et al. (1996, p 50) “ KDD é o processo não trivial de identificação de padrões, a partir de dados, que sejam válidos, novos, potencialmente úteis e compreensíveis”. Neste contexto, o resultado deve ser fácil de ser verificado e validado para a tomada de decisão. O autor destaca ainda que o KDD possui estrita relação com a pessoa que analisa estes dados, visto que o processamento do volume dos dados é uma tarefa computacional, mas a realização das análises requer habilidades humanas. O processo de KDD é compreendido por seis fases: seleção dos dados, limpeza de dados, enriquecimento, transformação, mineração de dados e a validação dos resultados, conforme a Figura 1 (ELMASRI e NAVATHE, 2011; FAYYAD, 1996).

Figura 1 – Processos do KDD



Fonte: (Fayyad 1996, modificado)

O primeiro processo do KDD é a seleção dos dados. Segundo Elmasri e Navathe (2011), esse procedimento seleciona os dados que serão utilizadas para a elaboração do processo. Conforme Silva et al (2016), os dados são a fonte para a aplicação dos métodos de mineração de dados. Eles podem estar apresentados de duas formas, dados estruturados e dados não estruturados. Os dados estruturados estão dispostos em estruturas tabuladas, sendo que as linhas armazenam a ocorrência de determinado evento e as colunas representam as características do exemplar, a qual denomina-se de instância. Esses dados podem ser resultantes de processos de medição e observação de determinado ambiente. Outra forma de representação dos dados é da forma não estruturada, sendo este formato representado por textos, imagens, vídeos e sons (SILVA et al, 2016; AMARAL, 2016).

O método de extração dos dados pode ser realizado diretamente no banco de dados. Após a seleção dos dados-alvos, ocorre o pré-processamento. A etapa compreende a limpeza e transformação dos elementos. Essa etapa tem o objetivo selecionar e filtrar conteúdos ruidosos, inconsistentes e também ausentes que podem afetar a qualidade do processo de MD, podendo até anular todo o processo (ELMASRI e NAVATHE, 2011; GOLDSCHMIDT e PASSOS, 2015; SILVA et al, 2016).

Quando os atributos não apresentam valores, este conjunto de dados pode conter relações nulas, tornando assim os atributos inválidos. Problemas de ausência de valores podem ser solucionado com o preenchimento manual das informações. Para aplicar esse método manual deve-se verificar se é possível executar a aquisição dos dados novamente e se o conjunto de preenchimentos não é relativamente grande. Outra forma de solucionar este problema é com o preenchimento automático, para isso deve-se verificar os valores mais frequentes, o valor médio ou o valor mediano (SILVA et al, 2016).

Outro problema tratado no processo de transformação são os valores ruidosos, isto é, elementos que estão fora do conjunto de informações esperado, como um valor consideravelmente diferente da maioria. Conforme Silva et al. (2016), esses ruídos são facilmente percebidos quando possuímos conhecimento sobre os valores resultantes dos atributos. Os autores destacam ainda que esses valores podem ser preenchidos com a inspeção e correção manual ou uma identificação e limpeza automática.

A etapa após a transformação dos dados é a mineração de dados, na qual são aplicados os algoritmos que irão identificar regras e padrões. O processo seguinte à MD é a validação dos resultados e apresentação por meio de gráficos tabelas ou listagens que geram o conhecimento. Como mencionam Elmasri e Navathe (2011), a mineração de dados é considerada uma parte do processo de descoberta do conhecimento, ou seja, está inserido no processo do KDD. No entanto, muitos autores consideram a mineração de dados como sendo o sinônimo do KDD. Essa discordância é comum no meio acadêmico, pois não há um consenso entre os autores (HAND et al, 2001; WANG, 2008; ELMASRI e NAVATHE, 2011).

Na mineração de dados, são definidas as técnicas e algoritmos que irão identificar as regras e padrões. Tais definições serão cruciais para a obtenção dos resultados válidos, gerando assim novos conhecimentos e respostas para o objetivo do estudo. As técnicas preditivas e descritivas são as tarefas possíveis de algoritmos de extração de conhecimento.

A utilização da técnica de classificação é uma forma de descrever determinado exemplar através de modelos de previsão. Neste caso, a base de dados deve estar rotulada em duas ou mais classes, já que o seu principal objetivo é determinar a qual classe um

novo exemplar irá pertencer. Esse novo exemplar pode ser classificado em uma classe já existente, ou poderá gerar uma nova classe com suas novas características. As técnicas de classificação possibilitam a criação dos modelos de previsão. Dentre os algoritmos aplicados na análise de dados biológicos, destaca-se a árvore de decisão (QUILICI-GONZALEZ e DE ASSIS ZAMPIROLI, 2014; AMARAL, 2016).

As árvores de decisão são modelos classificadores que consistem em uma estrutura de árvore com nós folhas e arcos. Cada nó interno da árvore representa um determinado teste em uma característica de uma instância, e os arcos representam os resultados. A árvore é percorrida de cima para baixo, iniciando por um único nó raiz que vai sendo dividido até levar à classe, conhecido como nó-folha. O nó-folha contém a classificação da instância. O modelo de árvore de decisão pode ser interpretado como um modelo SE ENTÃO. Ao percorrer a árvore, regras SE ENTÃO e SE ENTÃO SENÃO são criadas para determinar o caminho. Para o processo de criação da árvore de decisão, o conceito mais utilizado é o de dividir e conquistar. A árvore criada busca o nó raiz o qual é considerado o nó inicial e recursivamente trabalha em cada nó filho buscando criar uma única classe a partir de cada subconjunto. O fator mais crítico para a criação de uma árvore de decisão é a escolha da característica do nó inicial da árvore que será utilizado. Esse processo deve buscar a divisão mais “pura”, ou seja, que o número máximo possível de instâncias em cada subconjunto pertença a uma única classe (ELMASRI e NAVATHE, 2011; CARVALHO, 2015; AMARAL, 2016; SILVA et al, 2016).

Um dos algoritmos utilizados para técnicas de associação é o algoritmo *Apriori*, o qual destaca-se por utilizar o princípio de que, a presença de um conjunto de itens implica em um subconjunto destes itens. Este algoritmo executa várias análises, a fim de encontrar os dados transacionais mais frequentes na base de dados analisada. Com o grande número de itens em uma base de dados, o número de itens frequentes também será grande. Os itens mais frequentes são considerados os itens candidatos e os menos frequentes são eliminados (AMARAL, 2016; DE CASTRO e FERRARI, 2016; SILVA et al, 2016).

## **2.2. Processo de Fermentação de Vinhos e Espumantes.**

O vinho, conhecido hoje no mundo, é resultante do processo de fermentação realizada pelas leveduras. As leveduras são fungos unicelulares responsáveis pela fermentação da uva. A fermentação é um processo natural do vinho que transforma o açúcar em álcool e dióxido de carbono. As leveduras influenciam de forma significativa no gosto, no sabor e no aroma dos vinhos. Diversos tipos de leveduras são utilizados no processo de vinificação, e possuem inúmeras características distintas. (RIBÉREAU-GAYON et al, 2006).

O vinho é composto por água, álcoois, ácidos orgânicos, açúcares, polifenóis, minerais, proteínas e peptídeos polissacarídeos, vitaminas e compostos aromáticos, o qual chega a ter mais de 500 compostos que, juntos, são responsáveis pela formação do seu aroma. Essas características são naturais dos vinhos, e podem sofrer influência pela utilização das leveduras no processo de fermentação. Muitas dessas características estão relacionadas a sua qualidade. Algumas influenciam na cor, sabor e aroma, outras na maturação do vinho. (GUERRA, 2002).

O processo de fermentação do vinho é responsável por extrair complexos químicos naturais da uva, e pela transformação de vários ácidos. Os ácidos formados

durante a fermentação são provenientes de diversas transformações químicas entre vitaminas, dióxidos e enzimas. Algumas leveduras utilizadas no processo são responsáveis por produzir acidez, gostos, cheiros e colorações nos vinhos, em níveis diferentes. No processo de fermentação, as leveduras utilizadas são responsáveis pela formação de acetaldeído ou etanal, que afeta a estabilidade e a coloração dos vinhos. A falta de maturação da uva, influencia nas características frutadas do vinho, acarretando sua diminuição, provocando uma característica herbácea no vinho (GUERRA, 2002; RIBÉREAU-GAYON et al, 2006).

A acidez influencia a estabilidade e coloração do vinho, podendo assim alterar as suas características. Para seu controle, verifica-se a extensão com que os ácidos estão presentes em forma de sais nos vinhos. O controle é feito através de medições da acidez total e acidez volátil, o que influencia nas diferenças organolépticas dos vinhos, que pode ser explicada pelas características percebidas pelos sentidos humanos como, odor, brilho, sabor, cor, textura, luz e som. A acidez volátil no vinho é considerada um importante parâmetro físico-químico, sendo monitorado ao longo do processo de vinificação. Uma acidez volátil excessiva é responsável por uma avaliação negativa do vinho, em dose elevada origina o aroma de vinagre. A acidez do vinho está ligada à presença de acetato de etila (GUERRA, 2002; RIBÉREAU-GAYON et al, 2006).

A acidez total em mostos de vinho se dá pela presença de ácidos tartárico, cítrico, acético entre outros de menor consideração. A contribuição de cada tipo de ácido determina a força total da acidez do vinho (RIBÉREAU-GAYON et al, 2006).

O aroma é influenciado por diversos compostos denominados compostos voláteis. Esses compostos são responsáveis pelos diferentes aromas que o vinho proporciona. As substâncias odoríferas dos vinhos pertencem a diferentes grupos aromático: álcoois, aldeídos, cetonas, fenóis, terpenos e ésteres (GUERRA, 2002).

Os álcoois são originados na fermentação, no qual se destaca o ácido decanol, que produz o aroma de damasco. Nos aldeídos, que podem ser provenientes tanto da própria uva causando odor de limão, quanto da fermentação, causando odor de canela, e ainda do amadurecimento, provocando o odor de baunilha. As cetonas que são responsáveis pelo odor de manteiga. Os fenóis produzem cheiro de couro. Os terpenos provocam odor de cravo, os ésteres, produzido na vinificação como o butirato de etila com aroma de abacaxi, e fruto da ação de leveduras. Já o acetato de isoamila provoca o odor de banana, e é produzido pela ação de enzimas (GUERRA, 2002; RIBÉREAU-GAYON et al, 2006).

### **2.3. Trabalhos Relacionados**

Nesta seção, são apresentados os trabalhos relacionados à mineração de dados em vinhos. Os eventuais trabalhos são apresentados na tabela 1. Nesta tabela, é possível visualizar a aplicação dos trabalhos, os dados utilizados e as técnicas de mineração de dados empregadas em cada estudo.

Nos resultados do estudo de Da Costa (2016), foi evidenciado subgrupos de antocianinas que melhoram a qualidade dos vinhos. Adicionalmente, foi possível classificar os vinhos pela sua qualidade em determinada região, o qual garante a autenticidade do produto, garantindo assim sua alta qualidade (DA COSTA, 2016).

Para o estudo de De Souza Roza (2016) os algoritmos foram aplicados em um conjunto de dados que contém os resultados físico-químico e sensoriais dos boletins de

análise dos vinhos. Este é constituído por vinhos verdes branco, tinto ou rosado, espumantes, vinagres, vinhos licorosos, aguardentes e vinhos regionais, totalizando 23.223 amostras.

De Souza Roza (2016) coloca que o estudo obteve como resultado a identificação de 5 provadores que atribuíam nota diferentes da média, 3 acima da média e 2 abaixo da média. Foi possível verificar ainda que algumas características físico-químicas influenciam significativamente para a determinação da nota e outras não. O autor destaca que as evidências obtidas podem servir para a empresa na qual os avaliadores pertencem, como possibilidade de reestruturação do painel de avaliadores, eliminando aqueles que não cumprem com os critérios cuja viés ultrapasse um dado patamar. O autor destaca ainda que as técnicas de MD utilizadas auxiliaram na comprovação e na aquisição de novos conhecimentos para a área (DE SOUZA ROZA, 2016).

**Tabela 1 - Trabalhos relacionados a Técnica de mineração de dados**

<b>Autor</b>	<b>Aplicação</b>	<b>Objetivo</b>	<b>Dados</b>	<b>Técnicas Aplicadas</b>
Da Costa (2016)	Mineração de dados para classificação de vinhos <i>Vitis vinifera</i> da América do Sul	Evidenciar a qualidade dos vinhos classificando-os em diferentes patamares por região e encontrar compostos que contribuem para a classificação dos vinhos. A aplicação das técnicas abrangeu os vinhos pertencentes a Argentina, Brasil, Chile e Uruguai. As variedades de vinhos são Carménère, Cabernet Sauvignon, Malbec, Merlot, Tannat e Syrah.	Banco de dados do Laboratório de Alimentos Funcionais, Departamento de Alimento e Experimento Nutricional, Faculdade de Ciências Farmacêuticas, Universidade de São Paulo, o qual contém os componentes químicos que descrevem a funcionalidade dos vinhos.	Máquina de Vetores de Suporte, Redes Neurais, Associadores, Normalizadores, Redes bayesianas, Redes Neurais artificiais e Naive Bayes.
De Souza Roza (2016)	Mineração de dados para encontrar características diferentes entre os vinhos.	Encontrar características em vinhos que evidenciem a similaridades entre a região demarcada como Vinhos Verdes localizado no noroeste de Portugal.	Banco de dados com análises físico química e organolépticas dos vinhos.	Técnicas de agrupamento como K-Means, Regressão Linear e árvore de decisão M5P.
De Souza Roza (2016)	Mineração de dados nas avaliações dos vinhos realizadas pela empresa <i>Sogrape Vinhos S.A.</i>	Encontrar a caracterização do porquê da avaliação de determinado vinho por seus provadores.	Conjunto de dados que contém os resultados físico-químico e sensoriais dos boletins de análise dos vinhos. Constituído por vinhos verdes branco, tinto ou rosado, espumantes, vinagres, vinhos licorosos, aguardentes e vinhos regionais.	Técnicas de agrupamento como K-Means, Regressão Linear e árvore de decisão M5P.

Fonte: Criação própria.

Os estudos realizados obtiveram sucesso quanto à aplicação das técnicas de mineração de dados, extraindo informações importantes e contribuindo para o conhecimento, reforçando assim a importância da utilização das técnicas de mineração em conjuntos de dados provenientes de trabalhos relacionados à vinificação.

### **3. Metodologia**

#### **3.1. Dados**

Dois conjuntos de dados foram utilizados no presente estudo. O primeiro é proveniente da pesquisa realizada por Carissimi (2016), na qual o autor realizou experimentos com uvas Merlot/Tannat da safra de 2014 termovinificadas com 8 tipos de leveduras. Informações das características químicas das leveduras e dados da avaliação de degustadores através de características organolépticas, são as informações contidas no conjunto de dados avaliado.

O segundo conjunto de dados é relacionado à pesquisa realizada por Rossi Marcon (2015). A pesquisa teve por objetivo avaliar a influência de diferentes leveduras na composição aromática de espumantes e vinhos do tipo Moscato. Estes dados são de componentes voláteis realizados em 5 diferentes tratamentos. As leveduras utilizadas são; EC1118, Zymaflore alfa, TPI4, TPI4+ EC1118 fermentadas juntas, Zymaflore + EC1118 fermentadas juntas. As informações das leveduras e seus componentes químicos estão presentes no conjunto de dados. Em um segundo momento, os resultados dos tratamentos foram analisados por degustadores na qual diversas características organolépticas como gosto, turbidez entre outras foram avaliadas e pontuadas no conjunto de dados do estudo.

#### **3.2. Ferramentas Computacionais e Algoritmos**

A mineração de dados ocorreu por meio da ferramenta Weka, a qual apresenta uma gama de funcionalidades e aplicabilidades tanto para atributos numéricos quanto atributos nominais, a interface é intuitiva e apresenta suas tarefas organizadas em seqüência.

Os algoritmos utilizados foram: árvores de decisão, A priori. A utilização do algoritmo de árvore de decisão para o objetivo deste trabalho se justifica por possibilitar a fácil leitura e entendimento do conhecimento obtido, além de possibilitar a utilização de atributos numéricos, nominais e realizar previsão do resultado. A sua estrutura é de fácil entendimento, tornando simples para análise do especialista de domínio. Outro algoritmo utilizado foi o A priori, que possibilita a categorização dos dados nominais, realizando a associação dos dados. Este algoritmo complementa a inferência de hipóteses pelo especialista de domínio juntamente com o algoritmo de árvore de decisão. A utilização destes algoritmos complementa a extração do conhecimento.

#### **3.3. Metodologia de Treinamento e Teste**

O método utilizado para este processo é a validação cruzada conhecida como *K-fold Cross Validation*. Neste método, é selecionado aleatoriamente um conjunto de dados, dividindo esse conjunto de  $N$  elementos em  $K$  subconjuntos ou *fold*, sendo  $K-1$  vezes de treinamento e uma vez para teste. Em cada nova interação, um *fold* é usado como conjunto de teste e os  $K-1$  restantes são utilizados para o conjunto de treinamento. Esse processo é repetido com todos os  $K$  subconjuntos. A média dos treinamentos e dos testes é utilizada como

indicador de qualidade do modelo (QUILICI-GONZALEZ e DE ASSIS ZAMPIROLI, 2014; GOLDSCHMIDT e PASSOS, 2015; DE CASTRO e FERRARI, 2016).

Para a validação do modelo, é necessário utilizar medidas de avaliação. Para realizar de forma adequada tal validação destaca-se a matriz de confusão. A matriz pode indicar problemas mais relevantes ou menos relevantes nos resultados dos classificadores. A matriz de confusão é composta por verdadeiros positivos (VP) o qual representa a classificação correta da classe positiva; falso positivos (FP) que representa a classificação incorreta de uma classe positiva; verdadeiro negativo (VN) a qual representa a classificação correta na classe negativa e por fim falso negativo (FN) que representa a classificação incorreta na classe negativa.

A matriz de confusão permite calcular diferentes medidas de performance, como a Exatidão (ACC), Precisão (Pr) e Revocação (Re). A exatidão (fórmula 1) permite verificar a taxa de acertos ou erros para o conjunto de dados, chegando assim à verificação do nível de confiança do modelo adotado. Outra medida de avaliação considerada para os classificadores é a precisão (*precision*) e de revocação (*recall*). A precisão (fórmula 2) mede a probabilidade dos dados recuperados serem relevantes, já a revocação (fórmula 3) é a probabilidade de um dado relevante ser recuperado (QUILICI-GONZALEZ e DE ASSIS ZAMPIROLI, 2014; DE CASTRO e FERRARI, 2016; SILVA et al, 2016).

$$ACC = \frac{VP + VN}{VP + FP + NV + FN} \quad (1)$$

$$Pr = \frac{VP}{FP + VP} \quad (2)$$

$$Re = \frac{VP}{FN + VP} \quad (3)$$

A classificação correta de um conjunto de dados é dada pela taxa de quantidade de verdadeiros positivos, também conhecida como sensibilidade, e a taxa de falsos positivos, conhecida como especificidade. Com esses indicadores é possível avaliar o desempenho do modelo gerado pelos algoritmos de aprendizado utilizados (QUILICI-GONZALEZ e DE ASSIS ZAMPIROLI, 2014; GOLDSCHMIDT e PASSOS, 2015; DE CASTRO e FERRARI, 2016; SILVA et al, 2016).

Adicionalmente, a validação dos resultados se deu por meio dos especialistas de domínio, sendo considerada como uma medida de avaliação qualitativa. As pessoas envolvidas no processo de validade e verificação dos resultados obtidos estavam diretamente envolvidas nos trabalhos de pesquisa.

#### **4. Resultados e Discussão**

Os resultados obtidos com este trabalho devem ser interpretados dentro do contexto específico dos trabalhos relacionados à fermentação de vinhos. Deste modo, considerando que a utilização de algumas leveduras é específica da região, a contextualização dos resultados também deve ser.

Para a aplicação do algoritmo de árvore de decisão, é necessário a definição de um atributo preditor, o qual é orientador dos demais atributos. Nos dados do estudo de Rossi Marcon (2015), foi utilizado o atributo leveduras, com objetivo de identificar quais outros atributos influenciam as 5 leveduras utilizados em seu estudo. A figura 1 mostra o resultado da árvore de decisão, a qual teve um índice de classificações corretas em 100% e classificações incorretas em 0%.

**Figura 1** – Árvore de decisão do atributo leveduras

<b>Buriato_de_etila_(mg/L) &lt;= 0.1</b>	
Etanal_(mg/L) <= 33.59:	MB TPI-4 (2.0)
Etanal_(mg/L) > 33.59:	MB ZYMA FLORE (2.0)
<b>Buriato_de_etila_(mg/L) &gt; 0.1</b>	
Hexanoato_de_etila_(mg/L) <= 0.33:	MB ZYMA+EC1118 (2.0)
Hexanoato_de_etila_(mg/L) > 0.33	
Acetato_de_etila_(mg/L) <= 34.69:	MB EC 1118 (2.0)
Acetato_de_etila_(mg/L) > 34.69:	MB TPI-4 +EC1118 (2.0)

Fonte: Dados da pesquisa.

Os resultados mostram (Figura 1) que o atributo “*Buriato de etila*” é o divisor dos resultados das Leveduras. Sendo seu valor menor ou igual 0.1 resulta duas leveduras MB TPI-4 e MB ZYMA FLORE, tendo elas como resultantes o Etanal. É possível identificar que as leveduras que são compostas, como é o caso da MB ZYMA + EX1118 e a MB TPI-4+EC1118 são resultantes do valor maior que 0.1 do “*Buriato de etila*”. Esse resultado diferencia de forma significativa as leveduras que são compostas. Os mesmos foram validados pelos especialistas.

Outra aplicação utilizada foi do algoritmo *Apriori*. Este algoritmo foi aplicado nos dados das avaliações dos degustadores dos vinhos, e gerou dez regras de associação (Tabela 1).

**Tabela 1** – Regras de associação geradas pelo algoritmo *Apriori*.

Identificação	Regra	Conf.
<b>Regra 1</b>	SE vegetal/herbáceo = zero ENTÃO volátil/acético = zero	1
<b>Regra 2</b>	SE vegetal/herbáceo = zero E gosto indesejável = zero ENTÃO volátil/acético = zero	1
<b>Regra 3</b>	Se cravo da índia = zero ENTÃO volátil/acético = zero	0.96
<b>Regra 4</b>	Se nitidez/franqueza = três ENTÃO volátil/acético = zero	0.96
<b>Regra 5</b>	Se odor indesejável = zero E gosto indesejável = zero ENTÃO volátil/acético = zero	0.94
<b>Regra 6</b>	Se gosto indesejável = zero ENTÃO volátil/acético = zero	0.94

Fonte: Dados da Pesquisa (continua)

**Tabela 1** – Regras de associação geradas pelo algoritmo *Apriori*.

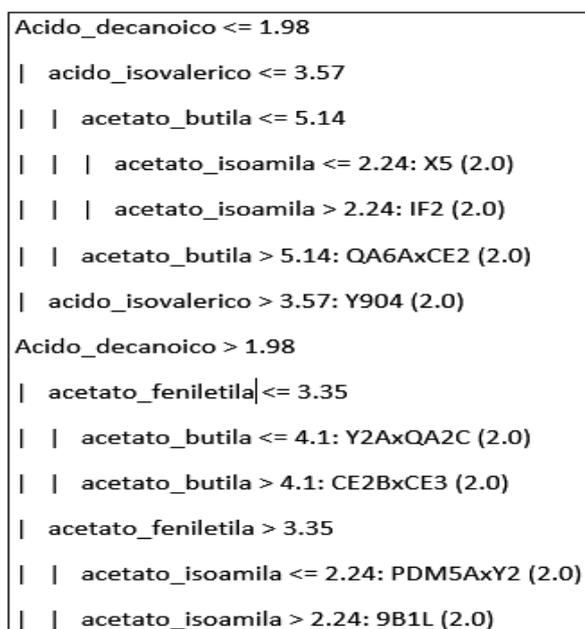
Identificação	Regra	Conf.
<b>Regra 7</b>	Se vegetal/herbáceo = zero ENTÃO gosto indesejável = zero	0.93
<b>Regra 8</b>	Se vegetal/herbáceo = zero E volátil/acético = zero ENTÃO gosto indesejável = zero	0.93
<b>Regra 9</b>	Se vegetal/herbáceo = zero ENTÃO volátil/acético = zero E gosto indesejável = zero	0.93
<b>Regra 10</b>	Se odor indesejável = zero ENTÃO volátil/acético = zero	0.93

Fonte: Dados da pesquisa (conclusão)

A partir da análise destas regras, foi possível identificar que se o atributo “*vegetal/herbáceo*” tiver valor atribuído pelos degustadores igual a zero, o atributo “*volátil/acético*” também será zero. Outra regra é a que se o atributo “*vegetal/herbáceo*” for zero e o gosto indesejável também for zero, o atributo “*volátil/acético*” será zero. Foi possível identificar ainda que o atributo “*gosto indesejável*” e “*odor indesejável*” contribuem significativamente para a avaliação dos degustadores. Esses resultados contribuíram para a afirmação do conhecimento empírico dos especialistas.

No conjunto de dados relacionado ao estudo de Carissimi (2016), é possível observar que através do atributo leveduras o “*ácido decanoico*” e o “*ácido isovalérico*” são os aspectos de maior influência no conjunto de dados. Se o valor do “*ácido decanoico*” for menor ou igual 1.98 classifica-se leveduras X5, IF2, Qa6axCE2, Y904. Se o valor do “*ácido isovalérico*” for maior que 1.98, os resultados, de suas leveduras são Y2AxQA2C, CE2BxCE3, PDM5AxY2, 9B1L. O resultado da Arvore de decisão é mostrado na figura 2, o qual teve um índice de classificações corretas em 93,75% e classificações incorretas em 6,25%.

**Figura 2** – Árvore de decisão do atributo levedura



Fonte: Dados da pesquisa.

No conjunto de dados relacionado às notas dos vinhos, foi possível identificar por meio do atributo “*qualidade geral*” que o aspecto de maior influência são as “*frutas vermelhas*”. Tal atributo é responsável por classificar os vinhos fermentados por diferentes leveduras. Foi possível identificar ainda que o “*odor indesejável*” é o atributo de maior influência no resultado final dos valores das notas aplicadas pelos degustadores, sendo ele influenciado diretamente pelo atributo “*volátil/acético*”.

Outro algoritmo aplicado foi *Apriori*. Este algoritmo foi aplicado nos dados da avaliação dos degustadores, gerando as seguintes regras mostradas na tabela 2.

**Tabela 2** – Regras de associação geradas pelo algoritmo *Apriori*

Identificação	Regra	Conf.
<b>Regra 1</b>	SE odor indesejável = zero E intensidade do sabor = três ENTÃO volátil/acético = zero	1
<b>Regra 2</b>	SE odor indesejável = zero E persistência = dois e meio ENTÃO volátil/acético = zero	1
<b>Regra 3</b>	SE odor indesejável = zero E intensidade do sabor = três E gosto indesejável = zero ENTÃO volátil/acético = zero	1
<b>Regra 4</b>	SE volátil/acético = zero E intensidade do sabor = três E gosto indesejável = zero ENTÃO odor indesejável = zero	1
<b>Regra 5</b>	SE qualidade aroma = três E odor indesejável = zero ENTÃO volátil/acético = zero	1
<b>Regra 6</b>	SE qualidade aroma = três E volátil/acético = zero ENTÃO odor indesejável = zero	1
<b>Regra 7</b>	SE frutas vermelhas = três E odor indesejável = zero ENTÃO volátil/acético = zero	1
<b>Regra 8</b>	SE frutas vermelhas = três E volátil/acético = zero ENTÃO odor indesejável = zero	1
<b>Regra 9</b>	SE pirazinas = zero E volátil/acético = zero ENTÃO odor indesejável = zero	1
<b>Regra 10</b>	SE odor indesejável = zero ENTÃO volátil/acético = zero	0.98

Fonte: Dados da pesquisa.

Através do Algoritmo *Apriori*, foi possível reforçar os resultados obtidos anteriormente. Assim, é possível identificar que o “*odor indesejável*” é influenciado pelo atributo “*volátil/acético*”, além disso, foi identificadas relações entre o “*odor indesejável*” e a “*intensidade do sabor*”. Esses resultados, assim como os resultados da pesquisa anterior, serviram para a afirmação do conhecimento empírico dos especialistas.

## 5. Conclusões

O pressuposto inicial, de que se poderiam extrair informações relevantes e auxiliar o estudo das análises de dados das fermentações de vinhos, pode ser observada com as regras apresentadas e discutidas neste trabalho. Assim, foi possível sintetizar algumas informações sobre os dados resultantes da fermentação dos vinhos. Foram evidenciadas

junto aos especialistas informações relevantes para seus estudos. Novos conhecimentos e inferências sobre os dados foram criadas e comprovadas. Informações que antes poderiam estar no conhecimento prévio dos especialistas, agora podem ser comprovados com dados.

Por fim, a aplicação destas técnicas em dados de fermentação de vinhos de outros anos, torna-se um complemento para este estudo. Podendo reforçar os resultados obtidos, desta forma, reforça a idéia de que a aplicação de técnicas de mineração de dados auxilia no descobrimento de novas hipóteses, gerando assim novos conhecimentos e descobertas.

## Referências

- AMARAL, Fernando. Aprenda Mineração de dados: teoria e prática. Rio de Janeiro: Alta Book, 2016.
- BRAGA, Luis Paulo Vieira. Introdução à Mineração de Dados. 2ª edição ampliada e revisada. Rio de Janeiro: E-Papers, 2005.
- CARISSIMI, Gabriel. Efeito de Cepas de Leveduras na Composição Química e Características Organolépticas de Vinhos Temovinificados. 2016. 53 f. Dissertação (Mestrado em Biotecnologia e Gestão Vitivinícola) – Universidade de Caxias do Sul (UCS), Centro de ciências Biológicas e da Saúde. Caxias do Sul, 2016
- CARVALHO, Hialo Muniz. Aprendizado de máquina voltado para mineração de dados: árvores de decisão. 2015. 98 f. Trabalho de Conclusão de Curso – Universidade de Brasília – UnB Faculdade UnB Gama – FGA, Brasília, 2014.
- COSTA, Evandro et al. Mineração de Dados Educacionais: Conceitos, Técnicas, Ferramentas e Aplicações. Jornada de Atualização em Informática na Educação, v. 1, n. 1, p. 1-29, 2013.
- DA COSTA, Nattane Luíza. Mineração de dados para classificação e caracterização de alguns vinhos Vitis Vinífera da América do Sul. 2016. 99 f. Dissertação (Mestrado em Ciências da Computação) – Universidade Federal de Goiás, Instituto de informática (INF), Programa de Pós-Graduação em Ciências da Computação, Goiânia, 2016.
- DE CASTRO, Leandro Nunes; FERRARI, Daniel Gomes. Introdução à mineração de dados: conceitos básicos, algoritmos e aplicações. São Paulo: Saraiva, 2016.
- DE SOUZA ROZA, Rui Nuno Ribeiro. Aplicação de técnicas de data mining ao setor vinícola. 2016. 89 f. Dissertação (Mestrado em Modelação, Análise de Dados e Sistemas de Apoio à Decisão) – Universidade do Porto, Faculdade de Economia, Porto, 2016.
- ELMASRI, Ramez; NAVATHE, Shamkant B. Sistema de banco de dados. 6 ed. São Paulo: Pearson Addison Wesley, 2011.
- FAYYAD, Usama M.; HAUSSLER, David; STOLORZ, Paul E. KDD for Science Data Analysis: Issues and Examples. In: KDD-96 Proceedings, p. 50-56, 1996
- GALVÃO, Noemi Dreyer; MARIN, Heimar de Fátima. Técnica de mineração de dados: uma revisão da literatura. Acta Paulista de Enfermagem, v. 22, n. 5, p. 686-690, 2009.
- GAMA, João et al. Extração do conhecimento de Dados: Data Mining. 2º Edição. Lisboa: Editora Silabo, 2015.

- GOLDSCHMIDT, Ronaldo; PASSOS, Emmanuel. Data Mining: Conceitos, Técnicas, Algoritmos, Orientações e Aplicações. 2º Edição. Rio de Janeiro: Elsevier, 2015.
- GUERRA, Celito Crivellaro. Maturação da uva e condução da vinificação para a elaboração de vinhos finos. In: Embrapa Uva e Vinho-Artigo em anais de congresso (ALICE). In: Simpósio Mineiro de Viticultura e Enologia, 1., 2002, Andradas, MG. Viticultura e Enologia: atualizando conceitos. Caldas: EPAMIG, 2002.
- HAND, David J.; MANNILA, Heikki; SMYTH, Padhraic. Principles of data mining. London: MIT press, 2001.
- QUILICI-GONZALEZ, José Artur; DE ASSIS ZAMPIROLI, Francisco. Sistemas Inteligentes e Mineração de Dados. Santo André: Triunfal Gráfica e Editora, 2014.
- RIBÉREAU-GAYON, Pascal et al. (Ed.). Handbook of enology, the microbiology of wine and vinifications. John Wiley & Sons, 2006.
- ROSSI MARCON, Ângela. Caracterização e Evolução aromática dos Espumantes Moscatéis de Farroupilha durante o Armazenamento e a Influência das Leveduras nos Vinhos Moscato. 2015. 34 f. Dissertação (Mestrado em Biotecnologia) – Universidade de Caxias do Sul (UCS), Centro de Ciências Biológicas e da Saúde, Programa de Pós-Graduação em Biotecnologia, Caxias do Sul, 2015.
- SILVA, Leandro Augusto; PERES, Sarajane Marques; BOSCARIOLI, Clodis. Introdução à Mineração de dados com aplicação em R. 1ª Edição. Rio de Janeiro: Elsevier, 2016.
- WANG, John (Ed.). Data Warehousing and Mining: Concepts, Methodologies, Tools, and Applications: Concepts, Methodologies, Tools, and Applications. IGI Global, 2008.