CENTRO UNIVERSITÁRIO LUTERANO DE PALMAS
Recredenciado pela Portaria Ministerial nº 1.162, de 13/10/16, D.O.U nº 198, de 14/10/2016
ASSOCIAÇÃO EDUCACIONAL LUTERANA DO BRASIL
Joel dos Santos Silva

UTILIZAÇÃO DE DATA MINING EM UMA BASE DE DADOS DA ÁREA FARMACÊUTICA PARA ANÁLISE DE DADOS

Palmas – TO
2018/2

Joel dos Santos Silva

UTILIZAÇÃO DE DATA MINING EM UMA BASE DE DADOS DA ÁREA FARMACÊUTICA PARA ANÁLISE DE DADOS

Projeto de Pesquisa elaborado e apresentado como requisito parcial para aprovação na disciplina de Trabalho de Conclusão de Curso (TCC) II do curso de bacharel em Sistemas de Informação pelo Centro Universitário Luterano de Palmas (CEULP/ULBRA).

Orientador: Prof. M.e Fernando Luiz de Oliveira.

Palmas – TO

2018/2

Joel dos Santos Silva

UTILIZAÇÃO DE DATA MINING EM UMA BASE DE DADOS DA ÁREA FARMACÊUTICA PARA ANÁLISE DE DADOS

Projeto de Pesquisa elaborado e apresentado como requisito parcial para aprovação na disciplina de Trabalho de Conclusão de Curso (TCC) II do curso de bacharel em Sistemas de Informação pelo Centro Universitário Luterano de Palmas (CEULP/ULBRA).

Orientador: Prof. M.e Fernando Luiz de Oliveira.

Aprovado em: _____/_____/_____

BANCA EXAMINADORA

_____

Prof. M.e Fernando Luiz de Oliveira.

Orientador

Centro Universitário Luterano de Palmas – CEULP

_____

Prof. M.e Fabiano Fagundes

Centro Universitário Luterano de Palmas – CEULP

_____

Prof. M.e Heloise Acco Tives Leão

Centro Universitário Luterano de Palmas – CEULP

Palmas – TO

2018/2

AGRADECIMENTOS

Agradeço primeiramente a DEUS que me deu força, coragem, saúde e pessoas para ajudar neste presente trabalho de conclusão de curso. Agradeço minha querida esposa pela grande motivação que me dispensou, ao meu cunhado Gabriel pela tradução do resumo e o Heloniel por avaliar as regras de associação.

Sou muito agradecido aos meus queridos pais que através de muito trabalho preparam o terreno para essa realização. Agradeço ao meu orientador Fernando Luiz e a banca examinadora composta por Fabiano Fagundes e Heloise Acco pelas considerações que fizeram sobre o trabalho, que por efeito, enriqueceu o conteúdo do mesmo e deu direção em meio a tanta confusão evidente na primeira qualificação do TCC I.

“Examinai tudo. Retende o bem.”

Primeira epístola do Apóstolo Paulo aos Tessalonicenses, cap. 5, vers. 21.

RESUMO

SILVA, Joel dos Santos. **Utilização de data mining em uma base de dados da área farmacêutica para análise de dados**. 2018. 82 f. Trabalho de Conclusão de Curso (Graduação) – Curso de Sistemas de Informação, Centro Universitário Luterano de Palmas, Palmas/TO, 2018[1].

O rápido avanço tecnológico junto com o crescimento da capacidade de armazenamento tornou possível o gerenciamento de dados em grande quantidade por parte das organizações e contribuiu para o crescimento do volume de dados. Por outro lado, analisar e entender esses dados ainda é um problema. O ser humano possui capacidade limitada de extrair conhecimento válido nessas imensas bases de dados (CRUZ, 2007). Em face dessa limitação, surge o processo *Knowledge Discovery in Databases* (KDD), em português descoberta de conhecimento em bases de dados (CRUZ, 2007). Esse processo automatizado possui algumas etapas que abrange desde o motivo de usá-lo (problema a ser resolvido) até a análise dos conhecimentos encontrados (possíveis soluções). Dentre essas etapas encontra-se o *data mining* (DM) mineração de dados, que por sua vez possui técnicas de análise de dados. Este trabalho teve por objetivo fazer o levantamento bibliográfico do processo KDD com foco maior na etapa de DM e ao final aplicou as técnicas de regressão linear simples e associação em uma base da área farmacêutica, especificamente nos dados de vendas.

Palavras chave: Mineração de dados. Banco de dados. Processo KDD. Apriori. Linear model.

ABSTRACT

Rapid technological advancement coupled with the growth of storage capacity has made it possible to manage data in large numbers by organizations and contributed to the growth of data volume. On the other hand, analyzing and understanding these data is still a problem. Human being has limited capacity to extract valid knowledge in these immense databases (CRUZ, 2007). In the face of this limitation, the Knowledge Discovery in Databases (KDD) process comes up (CRUZ, 2007). This automated process has some steps that range from the reason for using it (the problem to be solved) to analyzing the knowledge found (the possible solutions). Among these steps is data mining (DM), which in turn has data analysis techniques. This study goal was to make a bibliographic survey of the KDD process with a major focus on the DM stage and, in the end, it applied the simple linear regression and association techniques in a pharmaceutical area, specifically in the sales data.

Keywords: data mining, databases, knowledge, algorithms, apriori, pharmacy, linear model.

LISTA DE FIGURAS

LISTA DE TABELAS

LISTA DE GRÁFICOS

LISTA DE ABREVIATURAS E SIGLAS

DM – *Data mining*

IA – Inteligência artificial

KDD - *Knowledge Discovery in Databases*

SGBD – Sistema de gerenciamento de banco de dados

CRM - *Customer relationship management*

(RH) - Recursos humanos

(GPS) - Sistema de posicionamento global

(LM) - Linear model

SUMÁRIO

1 INTRODUÇÃO

O rápido avanço tecnológico junto com o crescimento da capacidade de armazenamento, tornou possível o gerenciamento de dados em vasta quantidade por parte das organizações e contribuiu para o aumento de acumulo de dados (TAN, STEINBACH e KUMAR, 2009).

Em paralelo ao enorme crescimento de dados armazenados, também aumentou a complexidade das relações entre os dados, ou seja, o modo como os dados são organizados para representar um problema real. Esses eventos limitam a capacidade humana de extrair conhecimento válido nessas imensas bases de dados (CRUZ, 2007).

Existe a possibilidade de fazer buscas e filtros nos dados através dos Sistemas de Gerenciamento de Banco de Dados (SGBD) e posteriormente visualizar as informações de forma amigável por meio de sistemas tradicionais em forma de gráficos, planilhas, relatórios, entre outros. Porém, esses processos tradicionais de exploração de dados são fundamentados basicamente na manipulação direta dos dados pelo homem. Os SGBDs oferecem funções para armazenar e fazer buscas em grandes bases de dados, mas resta ainda o problema de como analisar e entender esses registros (SINGH, 2001).

Em face dessas limitações, surge o processo *Knowledge Discovery in Databases* (KDD) em português: descoberta de conhecimento em bases de dados (CRUZ, 2007). Esse processo possui algumas etapas e abrange desde o motivo de usá-lo (problema a ser resolvido) até a análise dos conhecimentos encontrados (possíveis soluções).

Dentre essas etapas encontra-se o *data mining* (DM) ou mineração de dados em português, técnica que reúne métodos tradicionais de descoberta de informação e algoritmos robustos de inteligência artificial (IA) para processar enormes quantidades de dados com o objetivo de permitir que diferentes tipos de dados sejam analisados de uma nova maneira (TAN; STEINBACH; KUMAR, 2009).

Diante desse cenário, esse trabalho tem por objetivo extrair conhecimento de uma base de dados da área farmacêutica. Esta base de dados contém centenas de registros e, devido a esse volume de dados, a tarefa

de descobrir conhecimento de forma manual tornou-se demasiadamente cansativa para uma pessoa e inviável quanto ao custo e tempo.

Diante dos fatos expostos, usar uma ferramenta para extrair conhecimento de forma automatizada é uma saída conveniente uma vez que permitirá extrair conhecimento de maneira eficiente, rápida e automatizada.

Ao final, o gestor da farmácia poderá usar o conhecimento extraído para customizar a exposição dos produtos nas prateleiras, criar promoções direcionadas a clientes com características específicas e ter uma noção mais abrangente das rotinas de vendas da empresa. Para isso, objetivou-se em específico:

- organizar as informações de estudo de caso;
- extrair regras de associação dos registros de vendas usando o algoritmo Apriori;
- aplicar a técnica de regressão linear nos registros de vendas;
- verificar os resultados para validar se as informações extraídas são adequadas ao objeto de estudo.

Esse trabalho está divido da seguinte forma: a seção 2 possui a revisão da literatura do processo KDD e suas subseções tratam das etapas desse processo, dentre elas o DM; a seção 3 aborda a metodologia da referida pesquisa; a seção 4 apresenta os resultados da discussão; e, por fim, a seção 5 dispõe as considerações finais do desenvolvimento e desafios do trabalho.

2 REFERENCIAL TEÓRICO

Para que os objetivos deste trabalho fossem alcançados, fez-se necessário a abordagem do conceito *Knowledge Discovery in Databases* (KDD). Nesta seção, serão abordados os conceitos e as etapas do processo KDD com o foco maior na fase de DM.

2.1 PROCESSO KNOWLEDGE DISCOVERY IN DATABASES (KDD)

A procura por informações estratégicas acontece nos mais diferentes ramos. Geneticistas buscam padrões em genomas de seres vivos, empresas procuram descobrir a preferência dos clientes para realizar marketing dirigido, economistas geram previsões de mercado usando dados históricos como base, entre outros dos muitos exemplos do campo acadêmico e empresarial (ITAKURA, 2004).

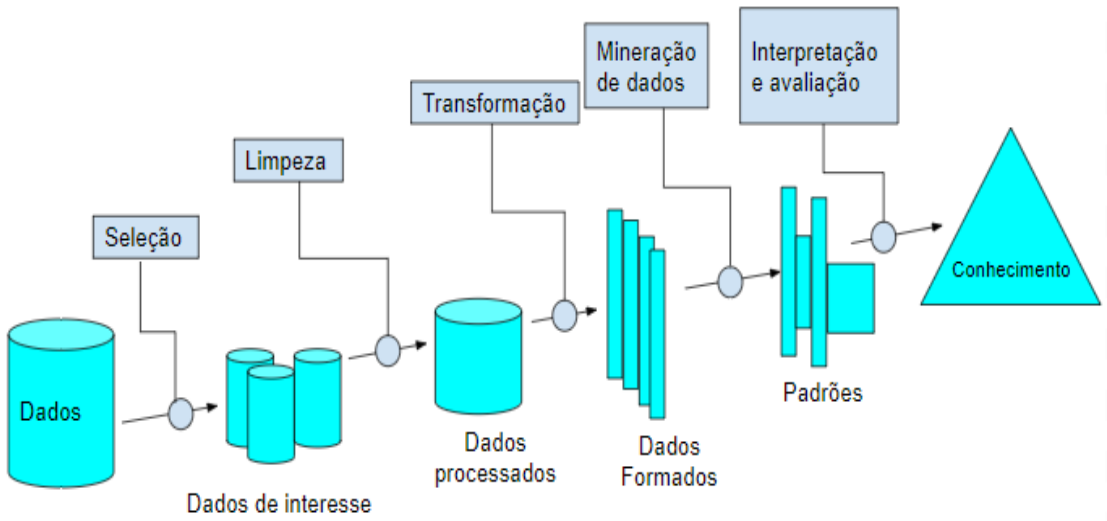
O processo (KDD) surge como uma possibilidade para a descoberta de informações em bases de dados. Para Fayyad (1996) o KDD é um processo incomum de descoberta de padrões válidos, novos, potencialmente utilizáveis e compreensíveis em coleções de dados. Itakura (2004) examina os termos dessa definição com mais detalhes conforme pode ser observado nos tópicos abaixo:

- dados: coleção de fatos, por exemplo, registros no banco de dados;
- processo: habitualmente o KDD é processo e possui várias etapas, que abrange o preparo de dados, procura por padrões, análise do conhecimento e refinamento;
- validade: os padrões encontrados devem possuir um determinado grau de certeza para que seja validado e usado posteriormente em outra etapa;
- novo: os padrões encontrados são novos ou atuais ao menos para o *software*. Para medir a novidade são levados em consideração:
    - modificações nos dados: comparam-se os valores atuais com o inicial ou já previstos;
    - conhecimento: qual a relação do novo conhecimento com o velho.

- potencialmente útil: os padrões novos encontrados têm de ser úteis em futuras atividades;
- compreensível: o processo KDD tem por finalidade descobrir conhecimento que seja entendível pelo ser humano. Se o usuário não conseguir compreender ou validar os dados não terá confiança para usá-lo na tomada de decisão. É difícil medir o quanto compreensível é uma regra ou padrão, para saber o grau é levado em consideração a medida da sua simplicidade. Existem vários padrões de simplicidade e variam de medidas completamente sintáticas (exemplo: tamanho do padrão de *bits*) e medidas semânticas (exemplo: simples de ser entendido pelo ser humano em aplicações).

O processo KDD é geralmente dividido em cinco etapas: seleção, limpeza, transformação, DM e avaliação. Estas etapas, bem como suas interações, podem ser visualizadas através da figura 1 apresentada abaixo.

Figura 1 - Etapas do processo KDD



Fonte: Adaptado de Itakura (2004)

Na figura 1 é apresentado o relacionamento das cinco etapas do processo KDD. Cada etapa desse processo será abordada nas seções seguintes para explicar sua importância e destacar alguns dos possíveis tratamentos para os problemas relacionados aos dados como: dados

incompletos, redundantes, inconsistentes, dados com falta de valor e dados com ruídos, entre outros.

A primeira etapa do processo KDD é a seleção, a mesma será abordada na próxima subseção.

2.1.1 SELEÇÃO

Na etapa de seleção ocorre a redução de dados. Os dados devem ser identificados e classificados como relevantes ou irrelevantes para serem usadas no processo KDD, em especial na fase de DM (MENDES, 2011). A relevância dos dados deve ser levada em consideração, os mesmos devem ser úteis para se chegar a um objetivo definido. Assim, caso não haja necessidade de determinado dado é feito o descarte do mesmo.

Segundo Cornelius Junior (2015) nesta etapa os objetivos de DM já devem estar definidos, essa atividade é feita junto com o especialista de domínio. Desse modo, a seleção dos dados acontece de forma alinhada para o alcance dos propósitos. A redução de dados acontece com a aplicação de tarefas como: eliminação manual dos atributos e redução da dimensionalidade, ambas tratadas nas subseções seguintes (2.1.1.1) e (2.1.1.2).

2.1.1.1 ELIMINAÇÃO MANUAL DOS ATRIBUTOS

Nessa etapa a participação do especialista do domínio é essencial, visto que o mesmo sabe com mais propriedade quais dados são necessários para chegar a determinadas informações e quais atributos são desnecessários e não agregam para o objetivo final.

Faceli, Lorena, Gama e Carvalho (2011) exemplificam essa tarefa usando uma tabela de nome “paciente” de uma base de dados de um determinado hospital. O objetivo de DM em questão é chegar a um diagnóstico com base nos dados de entrada do paciente; a tabela 1 apresenta os dados em seu formato original armazenado no banco de dados.

Tabela 1 - Tabela de dados completa

Id	Nome	Idade	Sexo	Peso	Manchas	Temp	Qt-Internacão	Est	Diagnóstico
1	João	28	M	79	Concentradas	38,0	2	SP	Doente
2	Maria	18	F	67	Inexistentes	39,5	4	MG	Doente
3	Luiz	9	M	92	Espalhadas	38,0	2	RS	Saudável
4	José	18	M	43	Inexistentes	38,5	8	MG	Doente
5	Claúdia	21	M	52	Uniformes	37,6	1	PE	Saudável
6	Ana	22	F	72	Inexistentes	38,0	3	RJ	Doente
7	Marta	19	F	87	Espalhadas	39,0	6	AM	Doente
8	Paulo	34	M	67	Uniformes	38,4	2	GO	Saudável

Fonte: Adaptado de Faceli, Lorena, Gama e Carvalho (2011)

Os especialistas do domínio decidiram que para se chegar ao diagnóstico do paciente não são necessários os atributos id, estado e nome. A partir dessa informação é feita a remoção manual desses atributos, a tabela 2 apresenta o resultado da tabela 1 com a ausência dos atributos apontados como desnecessários.

Tabela 2 – Tabela de dados após a redução de atributos

Idade	Sexo	Peso	Manchas	Temp	Qt-Internacão	Diagnóstico
28	M	79	Concentradas	38,0	2	Doente
18	F	67	Inexistentes	39,5	4	Doente
9	M	92	Espalhadas	38,0	2	Saudável
18	M	43	Inexistentes	38,5	8	Doente
21	M	52	Uniformes	37,6	1	Saudável
22	F	72	Inexistentes	38,0	3	Doente
19	F	87	Espalhadas	39,0	6	Doente
34	M	67	Uniformes	38,4	2	Saudável

Fonte: Adaptado de Faceli, Lorena, Gama e Carvalho (2011)

Em comparação com a tabela 1, a tabela 2 pode ser melhor compreendida pelo fato de possuir menos atributos. Essa tarefa contribui para a criação de um modelo de dados mais simples, de fácil visualização e com um menor tempo de processamento na aplicação dos algoritmos de DM (TAN; STEINBACH; KUMAR, 2009).

Existem outros casos em que um atributo passa ser dispensável como, por exemplo, um atributo que tem o mesmo valor em todos os objetos, não

contribuindo para uma diferenciação dos objetos (FACELI, LORENA, GAMA e CARVALHO, 2011). A seção seguinte trata da redução da dimensionalidade dos dados por meio de outras tarefas.

2.1.1.2 REDUÇÃO DA DIMENSIONALIDADE

O tamanho da dimensionalidade de uma coleção de dados é proporcional à quantidade de atributos que representa esse modelo. Uma quantidade grande de atributos pode desencadear uma série de problemas ao aplicar algoritmos de DM: a análise de dados se torna consideravelmente mais difícil, os dados se tornam dispersos, aumenta o tempo de processamento dos dados, algoritmos de classificação não conseguem atribuir de forma leal uma classe de todos os objetos potenciais e os algoritmos de agrupamento geram grupos de qualidade ruim (TAN; STEINBACH; KUMAR, 2009).

Cada atributo é tratado conforme uma coordenada em espaço d-dimensional, onde d é a quantidade de atributos, o acréscimo de novos atributos ocasiona o crescimento de forma exponencial do volume que representa esse dado. Para tornar mais compreensível, deve ser considerado que em uma coleção de dados todos os objetos têm apenas um atributo e esse atributo pode receber 1 entre 10 valores. Esse modelo de dados pode assumir então 10^1, ou seja, 10 objetos distintos, um para cada possível valor (FACELI, LORENA, GAMA e CARVALHO, 2011).

Para reduzir a dimensionalidade, Tan, Steinbach e Kumar (2009) destacam três abordagens que têm por papel selecionar características relevantes nos dados para a etapa de DM, sendo elas: interna, filtro e envoltório.

- abordagens internas: essa abordagem faz uso do próprio algoritmo de DM para fazer a seleção de características de forma natural, o algoritmo decide quais atributos serão necessários e quais podem ser ignorados;
- abordagens de filtro**:** possuem por característica o fato de que os atributos são devidamente selecionados antes da aplicação de técnicas de DM, por alguma abordagem isolada da tarefa de DM;
- abordagem envoltório**:** nesse caso é feito uso de algoritmos de DM, mas diferente da abordagem interna o foco do algoritmo é estabelecer um

conjunto de dados alvo e definir os atributos que tem potencial colaborativo para se chegar a esses dados, funciona de forma semelhante a uma caixa preta identificando os subconjuntos de atributos válidos.

Após a etapa de seleção dos dados inicia-se a etapa de limpeza, fase do processo KDD que tem sua importância na criação de um modelo de dados válido, a qual será abordada na seção seguinte.

2.1.2 LIMPEZA

Após a etapa de seleção, os dados escolhidos podem conter problemas que comprometem sua qualidade. O conjunto de dados pode ter objetos incompletos (com atributos faltando valor), dados com ruídos (possui conteúdo distinto do esperado), inconsistentes (com valores que contradizem aos demais atributos do mesmo objeto) e redundantes (atributos com mesmo valor) (MENDES, 2011).

Os objetos que possuem atributos sem valor ou redundantes são de fácil percepção, mas objetos com atributos com valor inconsistente ou com dados ruidosos exigem mais atenção para que sejam mapeados pois exige que se tenha conhecimento do domínio e que seja levado em consideração a combinação de todos os atributos de cada objeto.

Para Faceli, Lorena, Gama e Carvalho (2011) os problemas relacionados ao conteúdo dos dados são: dados incompletos, redundantes, inconsistentes e com ruídos, a causa e possíveis saídas para tratá-los; serão abordados nas subseções seguintes.

2.1.2.1 DADOS INCOMPLETOS

Segundo Prass (2004) este problema é caracterizado pela falta de valor em atributos de determinados objetos. Para Faceli, Lorena, Gama e Carvalho (2011) essa ausência pode ser causada por alguns motivos como:

- o atributo em determinado momento não era tratado como relevante, um exemplo é o atributo *e-mail* que na década de 1990 era pouco conhecido e usado, mas com o passar dos anos ganhou importância e foi incorporado no objeto, os objetos criados anteriormente ficaram com o valor do atributo em branco;
- falta de conhecimento do atributo na hora de inseri-lo, com o tipo sanguíneo do paciente;
- falta de atenção no momento do preenchimento;
- falta de obrigatoriedade de preenchimento do campo;
- a não de necessidade de preenchimento do campo, um caso seria do campo número de partos para um paciente homem;
- problema no equipamento no momento da transmissão, coleta ou gravação dos dados.

Segundo Prass (2004) existem várias formas de tratar esses atributos vazios, sendo as alternativas mais aplicadas:

- eliminar os objetos: normalmente é aplicada quando os atributos vazios do objeto são determinantes, ou seja, definem a classe do objeto. Essa alternativa não é sugerida quando o número de atributos vazios do objeto for pequeno ou quando a quantidade de objetos for pequena;
- preencher manualmente o valor dos atributos: essa opção não é praticável quando existem muitos objetos com atributos vazios, o tempo para realizar essa tarefa pode ser muito longa e cansativa;
- utilizar a média ou moda: é usado a média dos valores dos atributos para imputar no atributo vazio. Essa técnica pode levar a valores inconsistentes, por exemplo o objeto pessoa com o atributo idade igual a 2 anos, e a média do peso total das pessoas para colocar no atributo peso vazio igual a 72;
- Definir o valor de forma indutiva: nesse caso é usado um atributo alvo, os demais atributos do objeto são tratados como variáveis de entrada para o atributo alvo. Como base para inferir o valor do atributo alvo é usado os objetos com entradas semelhantes.

2.1.2.2 DADOS INCONSISTENTES

Dados inconsistentes são aqueles que têm valor que não corresponde a todos os atributos do objeto (FACELI, LORENA, GAMA e CARVALHO, 2011). Como por exemplo o CEP inserido não corresponde ao da cidade do endereço da pessoa ou o atributo peso é igual a 72 e a idade igual 2 anos. Alguns problemas de inconsistência ocorrem devido a integração de dados, fontes de dados diferentes podem usar diferentes tipos de medidas como metros, centímetros (MENDES, 2011).

Quando a quantidade de dados inconsistentes é pequena a remoção pode ser feita de forma manual. O problema está na identificação dos mesmos que depende da busca linha por linha a procura de inconsistências.

2.1.2.3 DADOS REDUNDANTES

Para Tan, Steinbach e Kumar (2009) um conjunto de dados pode possuir tantos objetos como atributos redundantes. Um objeto é redundante quando seus atributos têm valores idênticos a outro objeto. No caso de atributos, é redundante quando é possível inferir o seu valor com base em um ou mais atributos, um caso ainda mais grave ocorre quando determinado atributo tem o mesmo valor para todos os objetos.

Dados redundantes podem ser gerados quando acontecem problemas na coleta de dados, na entrada, na integração, no modelo de armazenamento ou na transmissão de dados (FACELI, LORENA, GAMA e CARVALHO, 2011). A tabela 3 apresenta um exemplo de redundância.

Tabela 3 – Tabela de dados com redundância

Idade	Sexo	Peso	Manchas	Temp	Qt-Internacão	Diagnóstico
28	M	79	Concentradas	38,0	2	Doente
18	**F**	**65**	**Inexistentes**	**39,5**	**4**	**Doente**
9	M	92	Espalhadas	38,0	2	Saudável
18	**F**	**65**	**Inexistentes**	**39,5**	**4**	**Doente**
21	M	52	Uniformes	37,6	1	Saudável
22	F	72	Inexistentes	38,0	3	Doente
19	F	87	Espalhadas	39,0	6	Doente
34	M	67	Uniformes	38,4	2	Saudável

Fonte: Adaptado de Faceli, Lorena, Gama e Carvalho (2011)

Observa-se que na tabela 3, o segundo e o quarto objeto possuem os mesmos valores, por esse motivo é considerado redundante. Um exemplo claro de atributo redundante seria um objeto “pessoa” possuir os atributos “ano” de nascimento e o atributo “idade”, pois pelo ano de nascimento pode ser inferido a idade com base no ano atual.

2.1.2.4 DADOS COM RUÍDOS

Para Faceli, Lorena, Gama e Carvalho (2011) um dado é tratado como ruidoso quando possui um valor de atributo que foi aparentemente gerado de forma aleatória, dados inconsistentes podem ser consequência da existência de ruídos. É importante salientar que não é possível entender se o valor de determinado atributo é resultado da existência de ruídos, o que se tem é apenas um indicio que o dado foi gerado com ruído.

Os indicadores de um possível ruído são chamados de *outliers*, que consistem em valores que apresentam discrepâncias, diferentes dos valores aceitáveis ou incomuns em comparação com os demais objetos (PRASS ,2004). A tabela 4 apresenta um exemplo de um dado com ruído.

Tabela 4 – Tabela de dados com ruído

Idade	Sexo	Peso	Manchas	Temp	Qt-Internacão	Diagnóstico
28	M	79	Concentradas	38,0	2	Doente
18	**F**	**615**	**Inexistentes**	**39,5**	**4**	**Doente**
9	M	92	Espalhadas	38,0	2	Saudável
18	M	43	Inexistentes	38,5	8	Doente
21	M	52	Uniformes	37,6	1	Saudável
22	F	72	Inexistentes	38,0	3	Doente
19	F	87	Espalhadas	39,0	6	Doente
34	M	67	Uniformes	38,4	2	Saudável

Fonte: Adaptado de Faceli, Lorena, Gama e Carvalho (2011)

Ao observar a tabela 4, nota-se que o atributo peso do segundo objeto tem um valor anormal em comparação com os demais objetos do conjunto de dados, pois o atributo peso com valor de 615 quilos é desproporcional para o objeto pessoa.

A exclusão de um objeto que possui um atributo com valor anormal deve ser feita de forma cuidadosa, ou seja, exige uma averiguação para constatar a veracidade do fato. Às vezes, o valor do atributo é discrepante ou um valor atípico verdadeiro que mostra um comportamento anormal, como uma tendência ou ainda uma possível transação fraudulenta; sendo que mapear essas discrepâncias muitas vezes é um objetivo de DM (PRASS, 2004).

Após a fase de limpeza, vem a etapa de transformação de dados que será abordada na subseção seguinte.

2.1.3 TRANSFORMAÇÃO

A etapa de transformação tem por objetivo preparar os dados para a aplicação de algoritmos de DM, os mesmos devem ser interpretáveis pela ferramenta de DM e os atributos carecem de estar conforme requerido pelo algoritmo que será aplicado na técnica de DM (MENDES, 2011).

Por exemplo, os algoritmos de classificação e associação precisam que os atributos dos dados sejam de tipo simbólico para terem um melhor desempenho, já o de agrupamento requer que os atributos estejam de forma

binária, desta maneira a transformação dos dados é feita em função da técnica de DM que será aplicada (TAN; STEINBACH; KUMAR, 2009).

Nesta etapa também é feita a padronização dos dados, pois dependendo da situação os dados vem de fontes diferentes, ou seja, de outros bancos que por sua vez possuem finalidades distintas (FACELI, LORENA, GAMA e CARVALHO, 2011). Ao juntar esses dados na etapa de seleção pode ocorrer que determinado atributo de um banco seja de um tipo diferente do mesmo atributo de outro banco. Um exemplo seria o atributo sexo que no banco **A** pode assumir um de dois valores M ou F e no banco **B** o mesmo atributo pode assumir os valores 0 ou 1. Outro tipo de divergência pode ocorrer da seguinte forma: no banco **A** o atributo “distância” é medido em centímetros e no banco **B** em metros.

A importância dessa etapa está relacionada a qualidade dos resultados das tarefas de DM, se dados não passar pelo devido preparo de adequação e padronização o modelo de dado pode não corresponder a realidade. Os algoritmos terão resultados de baixa qualidade por possuir características restritivas a determinados tipos de dados.

Algumas técnicas de conversão e transformação de atributos serão abordadas nas próximas seções, sendo que a primeira técnica (subseção 2.1.3.1) é utilizada para converter atributos de tipo simbólico em numérico, a segunda (subseção 2.1.3.2) converte atributos numérico em simbólico e a última (subseção 2.1.3.3) faz a transformação de atributos numéricos.

2.1.3.1 CONVERSÃO SIMBÓLICO – NUMÉRICO

Segundo Faceli, Lorena, Gama e Carvalho (2011) algoritmos de agrupamento, técnicas de redes neurais artificiais e *support vector machines* utilizam apenas dados numéricos. Dessa forma, quando a coleção de dados que será submetida a uma dessas técnicas possui atributos simbólicos é necessário que sejam convertidos para numéricos.

No caso de atributos do tipo nominal que podem assumir um de dois valores, o valor pode ser representado por 0 ou 1. Seria o caso do atributo “ativo” da tabela cliente, os possíveis valores simbólicos a serem assumidos

seria “sim” ou “não”, com a conversão poderia ser representado por atributos numéricos 0 ou 1, a tabela 5 abaixo apresenta a situação mencionada (FACELI, LORENA, GAMA e CARVALHO, 2011).

Tabela 5 - Conversão simbólico – numérico

id	nome	ativo
1	João	sim
2	Pedro	não
3	José	não
4	Ricardo	sim

Dados simbólicos

id	nome	ativo
1	João	1
2	Pedro	0
3	José	0
4	Ricardo	1

Dados numéricos

Quando o atributo simbólico pode assumir mais de dois valores e o mesmo é do tipo ordinal, o valor do atributo pode ser representado por um número de uma de terminada sequência, como pode ser observado na tabela 6 (FACELI, LORENA, GAMA e CARVALHO, 2011).

Tabela 6 – Conversão de valor ordinal para inteiro

Valor ordinal	Valor inteiro
Janeiro	1
Fevereiro	2
Março	3
Abril	4
Maio	5

Outra forma de equivaler um atributo simbólico nominal a um numérico é representar cada possível valor do atributo por uma sequência *C* de bits, onde *C* é igual ao número de possíveis valores desse atributo. Cada sequência de bits pode possuir apenas um bit 1 os demais valores devem ser 0, então a diferença das sequências de bits se dá pela posição do bit de valor 1 (FACELI, LORENA, GAMA e CARVALHO, 2011).

As diferentes sequências de bits formadas correspondem a um valor do atributo nominal, como mostra a tabela 7, onde a transformação simbólica

nominal para numérico é feito em cima do atributo cor (FACELI, LORENA, GAMA e CARVALHO, 2011).

Tabela 7 – Codificação 1 – de - C

Valor nominal	Código *C* - de - 1
Azul	100000
Amarelo	010000
Verde	001000
Marrom	000100
Branco	000010
Preto	000001

Fonte: Adaptado de Faceli, Lorena, Gama e Carvalho (2011)

2.1.3.2 CONVERSÃO NUMÉRICO – SIMBÓLICO

Para Faceli, Lorena, Gama e Carvalho (2011) a conversão de atributos numéricos em atributos simbólicos se dá pela necessidade de que uma parcela de algoritmos de classificação e associação trabalham melhor se os dados forem do tipo qualitativo. Quando esses algoritmos lidam com dados quantitativos, os mesmos têm sua performance diminuída.

Segundo Faceli, Lorena, Gama e Carvalho (2011) para um melhor aproveitamento dessas técnicas de DM é necessário fazer a conversão (discretização) dos dados para o tipo simbólico. Se o atributo a ser convertido possuir característica binária, a discretização é feita de forma comum, onde simplesmente é associado um nome para cada valor possível, no caso do atributo "sexo" representado por 0 ou 1 passaria a ser representado por M ou F.

Quando o atributo for constituído por sequências binárias e não possuir uma relação de ordem, cada sequência será representada por uma categoria ou nome (FACELI, LORENA, GAMA e CARVALHO, 2011). Um atributo de nome "departamento" representado por sequências de bits poderia ser representado por um nome como: recursos humanos (RH), contabilidade, etc.

Se a quantidade de possíveis valores a ser assumido por um determinado atributo for grande, pode ser proveitoso para alguns algoritmos de DM reduzir esse número juntando alguns valores e formar categorias para

representa-los. Um caso real seria a categorização dos departamentos de uma grande empresa, onde os departamentos, recursos humanos e administrativo passariam a fazer parte da categoria "humanas".

2.1.3.3 TRANSFORMAÇÃO DE ATRIBUTOS NUMÉRICOS

Segundo Elmasri e Navathe (2005), a transformação de atributos refere-se ao ato de transformar o valor de um determinado atributo para todos os objetos. Tal ato é necessário quando a diferença do menor valor do atributo para o maior valor é demasiadamente grande ou quando existe caso em que os atributos estão em diferentes escalas.

Considere-se um atributo de valor de tipo inteiro, em que para chegar ao modelo de dados pretendido não seja necessário o sinal, mas sim o seu valor. Nesse caso existe a necessidade de transformar o valor do atributo para o seu valor pleno sem sinal (FACELI, LORENA, GAMA e CARVALHO, 2011).

Por outro lado, às vezes surge a necessidade de traduzir um valor de um determinado atributo para uma forma mais entendível. No caso, transformar o atributo data de nascimento para idade, de escala graus *Celsius* para *Fahrenheit*, local dado pelo sistema de posicionamento global (GPS) para código de caixa postal (FACELI, LORENA, GAMA e CARVALHO, 2011).

Após a etapa de transformação inicia-se a fase de DM que será abordada na seção seguinte.

2.2 DATA MINING

De acordo com Carvalho (2001) DM ou mineração de dados em português, é o termo referente a um conjunto de técnicas extraídas da estatística e inteligência artificial com foco único em descobrir novos conhecimentos que esteja oculto em grandes bases de dados.

“*Data mining* é o processo de extração de informações desconhecidas, porém significativas, de bancos de dados extensos para serem utilizados na tomada de decisões do negócio” (SINGH, 2001, p.29). Segundo

Tan, Steinbach e Kumar (2009) DM é uma maneira automatizada de descobrir informações aproveitáveis em bases de dados volumosas. Na próxima subseção será abordado de forma rápida alguns motivos que ajudaram a popularizar o uso de DM.

2.2.1 MOTIVOS DO CRESCENTE AUMENTO NO USO DE TÉCNICAS DE DM

Carvalho (2001) destaca alguns motivos que justifica o aumento e a necessidade do uso de técnicas de DM nos últimos anos:

- atualmente o número de dados disponível é imenso: o DM é um processo que para seu uso deve dispor de uma grande quantidade de dados, porque seus algoritmos necessitam calibrar e tirar dos dados informações confiáveis. Grandes empresas do ramo de telefonia, bancária, cartões de crédito, serviços *online*, etc. Criam enormes quantidade de dados sobre suas tarefas e clientes. Tais dados podem ser submetidos a análise por DM;
- recursos computacionais: o DM carece de máquinas dotadas de muitos recursos computacionais para trabalhar seus algoritmos em dados volumosos. Com o aumento da potência computacional, possibilitado pelo avanço da microeletrônica, e a baixa nos preços dos computadores tornou possível a prática de DM;
- competição empresarial: a competição cria a necessidade de medidas mais modernas de tomada de decisão. Algumas empresas da área de finanças, telecomunicações e seguros adotaram o uso de DM por já deterem uma enorme quantidade de dados. Para empresas de serviços de vendas, obter dados para extrair conhecimento é fundamental, pois com base no conhecimento obtido será possível ofertar melhor seus produtos ao público certo. Por outro lado, algumas empresas vendem as informações como produto.

2.2.2 APLICADABILIDADE DE DM

A aplicabilidade de técnicas de DM é muito vasta já que, muitas áreas podem tirar um bom proveito desse processo, dentre essas aplicações, algumas são Cruz (2007):

- *customer relationship management* (CRM), em português (gestão de relacionamento com o cliente) é mais vantajoso para uma organização manter os clientes atuais do que obter novos clientes, nesse sentido a DM ajuda a traçar o perfil dos clientes possibilitando prever algumas necessidades. Dessa maneira é possível criar serviços e produtos personalizados que ajuda na fidelização dos consumidores;
- suporte a decisão: o futuro de uma empresa é afetado pelas decisões que os gestores tomam, normalmente essas escolhas são tomadas com base em tendências de mercado, preços de produtos, entre outros. O DM proporciona predizer dados de vendas, lucros, prever a necessidade e interesse dos clientes, entre outros;
- finanças: DM é usado na área de financeira para análise de crédito, identificação de fraudes, previsões de mercado, os bancos usam DM para definir o perfil de consumo de seus clientes de cartão de crédito com o objetivo de inferir um desvio de padrão de consumo que pode ser causado por um suposto intruso;
- análise cientifica: quando existe uma grande quantidade de dados é possível usar DM em áreas como medicina e biologia, na medicina pode ser aplicado para inferir diagnóstico de paciente, identificar a terapia apropriada para o paciente, descobrir formas modernas de tratamento de doenças, etc. Um exemplo de uso de DM na biologia pode ser dado na exploração do genoma humano.

2.2.3 A METODOLOGIA DO DM

Podem ser usados diferentes tipos de metodologias para minerar dados. A escolha da mesma depende do grau de conhecimento que se tem sobre o

problema a ser analisado. No caso, se não há nenhuma informação do comportamento do evento a saída é deixar que as técnicas prontas de DM busquem novas relações implícitas nos dados que humanamente seriam difíceis de serem identificadas. Esse método é chamado de “descoberta não supervisionada de relações” (CARVALHO, 2001).

Na ocasião em que se tem algum entendimento sobre a área de trabalho da empresa ou ciência da relação nova que está procurando, pode-se formular uma hipótese e averiguar sua comprovação ou contestação através da metodologia testagem de hipótese. Por último, na ocasião que se tem um conhecimento avançado sobre o problema e área de atuação da empresa usa o método denominado de “modelagem dos dados” (CARVALHO, 2001).

2.2.4 TAREFAS DE DM

É importante saber a diferença entre tarefa de DM e técnica. A tarefa de DM apenas diz "o que" se pretende procurar nos dados, que tipo de padrões, categorias ou regularidades. A técnica de DM aponta o método que responde a "como" chegar aos padrões de interesse, algumas das principais técnicas de DM advêm da estatística e da inteligência artificial (AMO, 2003).

De forma resumida as tarefas de DM podem ser executadas por suas devidas técnicas, sendo essas tarefas: associação, classificação, regressão, sumarização e agrupamento.

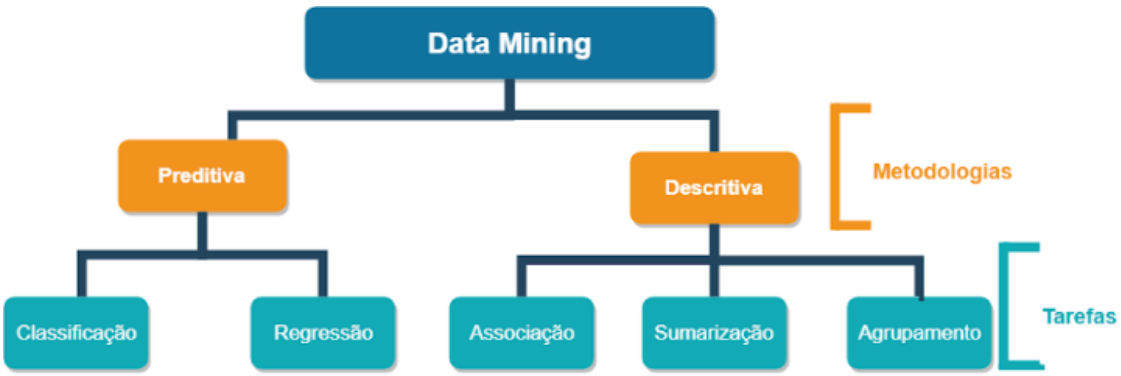
- associação: determina os fatos ou objetos que tendem a acontecer juntos na mesma transação ou mesmo evento. Um exemplo seria associar quais produtos são vendidos juntos numa mesma compra (PEREIRA, 2005);
- classificação: o objetivo dessa tarefa é classificar os objetos conforme as suas características comuns com base em classes pré-definidas (HARRISON, 1998). Um exemplo seria classificar clientes como honestos ou desonestos com base nos dados de histórico de pagamento de contas;
- regressão: é utilizada para definir um valor para uma determinada variável, por exemplo predizer a chance de um paciente superar determinada enfermidade baseado nos resultados de diagnósticos ou predizer quantos

filhos uma determinada família poderá ter (FAYYAD; SHAPIRO; SMYTH, 1996);

- sumarização: utiliza técnicas para descobrir descrições de forma detalhada de um subconjunto de dados; as técnicas de sumarização modernas advêm de regras de resumo, visualização e descobrimento de relacionamento entre variáveis (FAYYAD; SHAPIRO; SMYTH, 1996);
- agrupamento: também conhecida como *clustering* em inglês, tem por função dividir grupos heterogêneos em subgrupos homogêneos. Essa tarefa se difere de classificação porque não possui classes predefinidas. Um exemplo seria agrupar pacientes com doenças semelhantes (PEREIRA, 2005).

O DM é muito abrangente possuindo diversas técnicas e programas para resolver o problema de forma adequada, o problema define o tipo de metodologia, tarefa e técnica que será usada. A figura 2 abaixo apresenta a divisão do tipo de metodologia de DM e suas tarefas.

Figura 2 - Metodologia de DM e subcategorias



Fonte: Adaptado de Correia (2017)

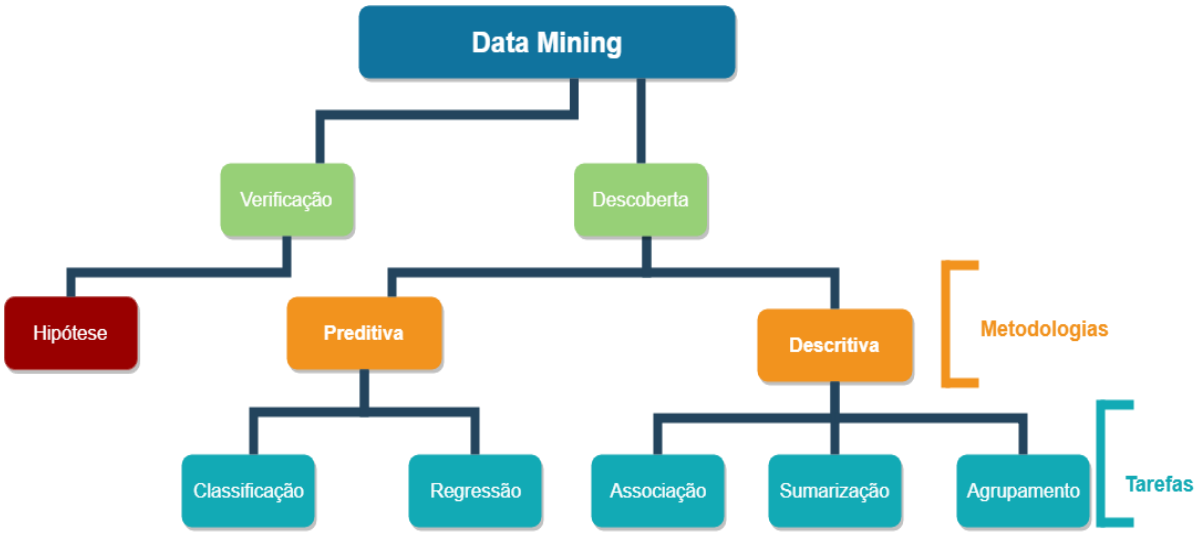
Como pode ser observado na figura 2 o DM possui duas metodologias, preditiva e descritiva, e suas respectivas tarefas a serem executadas pelas técnicas adequadas. Mendes (2011) descreve de forma breve essas metodologias:

- preditiva: pode predizer o valor futuro de uma determinada variável, os dados futuros são conhecidos com base no conhecimento antigo;

- descritiva: descobre valores e padrões interpretáveis que explicam um conjunto de dados, ou seja, os padrões descobertos explicam os dados atuais.

Pra Cruz (2007) DM é dividido da seguinte forma conforme apresenta a figura 3.

Figura 3 – Metodologia de DM e subcategorias segundo Cruz



Com base na demonstrado na figura 3 nota-se que a organização de DM segundo Cruz (2007) acrescenta as subcategorias “descoberta” para organizar e enquadrar as descobertas de cunho: preditiva e descritiva e cria a subcategoria verificação para indicar a metodologia de testagem de hipótese já mencionada anteriormente neste trabalho na seção 2.2.5.

2.2.5 TÉCNICAS DE DM

Segundo Cruz (2007) existe uma grande variedade de técnicas de DM disponíveis onde cada uma responde a uma determinada tarefa, nesta seção serão abordadas apenas algumas, sendo elas: associação, classificação, regressão linear e agrupamento.

2.2.5.1 ASSOCIAÇÃO

A tarefa de associação compreende descobrir relações em atributos distintos que ocorrem no mesmo evento; é muito utilizada quando se minera dados em bases comerciais onde se deseja descobrir relacionamentos de produtos distintos que são vendidos na mesma compra (ITAKURA, 2004).

Essa tarefa também é denominada como análise de afinidade, o objetivo é encontrar fatos que acontecem de forma simultânea e com a probabilidade de recorrência (CARVALHO, 2001).

Para Mendes (2011) essa técnica é bastante útil para descobrir relacionamentos em bases de dados grandes, porém o processo de descoberta pode ser computacionalmente custoso e alguns padrões podem surgir do acaso.

Para assegurar um certo grau de confiança sobre os padrões descobertos, Tan, Steinbach e Kumar (2009) destacam dois fatores conhecidos como suporte e confiança, ambos explicados abaixo.

- suporte: define a frequência que uma regra é apropriada a uma coleção de dados. Sua importância se dar pelo fato de mostrar as regras sem interesse e identificar regras que oferecem pouca similaridade;
- confiança: define a frequência que um objeto Y surge em eventos que possuem X. Mede a precisão da inferência feita por uma regra de associação; como pode ser entendido a regra de associação compreende a seguinte expressão: se X então Y. O nível de confiança atribuído mostra a probabilidade de Y encontrar-se presente nos eventos que possuem X.

Segundo Itakura (2004) X e Y são objetos, os fatores suporte e confiança são medidos da seguinte forma:

$$\text{Suporte} = \frac{nº\ de\ transações\ com\ X\ e\ Y}{nº\ total\ de\ transações}$$

$$\text{Confiança} = \frac{nº\ de\ transações\ com\ X\ e\ Y}{nº\ de\ transações\ com\ X}$$

A partir da observação das fórmulas de suporte e confiança, a afirmação de Mendes (2011) de que se a base de dados for muito grande o tempo computacional para descobrir as afinidades será longo ganha fundamentos matemáticos. O algoritmo terá que identificar as transações com os itens frequentes e depois efetuar as divisões para chegar aos valores de “suporte’ e “confiança” das relações descobertas, o tempo de processamento vai depender do tamanho do modelo de dados e da potência da máquina.

Para diminuir o problema de tempo computacional Tan, Steinbach e Kumar (2009) dividiram o processamento dessa técnica em duas partes:

- gerar conjunto de itens frequentes: tem por objetivo estabelecer todos os conjuntos de objetos que atendam o limite do suporte;
- gerar regras: extrai totalmente as regras com alto nível de confiança do conjunto de afinidade alcançados, ou seja, separar as regras válidas.

Amo (2003) exemplifica a aplicação da técnica de associação em um banco de dados de um supermercado, onde o objetivo é descobrir os itens que são frequentemente vendidos juntos, cada transação (compra) é chamada de *Itemset*, a tabela abaixo apresenta o conjunto de itens que são ofertados no supermercado.

Tabela 8 – Lista de produtos

Produto	Número que o representa
Pão	1
Leite	2
Açucar	3
Papel higiênico	4
Manteiga	5
Fralda	6
Cerveja	7
Refrigerante	8
Iogurte	9
Suco	10

Fonte: Adaptado de Amo (2003)

A tabela 8 apresentou a lista de produtos de um supermercado associando cada um com um número, a associação é apenas para efeito de resumo para a sequência da ilustração de Amo (2003).

A seguir, a tabela 9 apresenta uma representação de compras e seus respectivos itens.

Tabela 9 – Representação de compras

TID	Compra
101	{1,3,5}
102	{2,1,3,7,5}
103	{4,9,2,1}
104	{5,2,1,3,9}
105	{1,8,6,4,3,5}
106	{9,2,8}

Fonte: Adaptado de Amo (2003)

A tabela 9 possui duas colunas TID e Compra, a coluna TID é apenas um identificador da compra, os itens da compra são representados pelos seus números associados na tabela 8, no caso a compra {1,3,5} representa {Pão, Açúcar, Manteiga}.

Como já dito anteriormente, cada transação (compra) é chamada de *Itemset*. Se para um *Itemset* ser considerado como constante ele tenha que aparecer no mínimo em 50% de todas as compras, nesse caso o *Itemset* válido que satisfaria essa regra na tabela 9 seria {1,3}, essa é uma definição prática do fator “suporte” (AMO, 2003).

A tabela 10 apresenta os *Itemsets* e o valor de suporte de cada um.

Tabela 10 – Representação do suporte

Itemset	Suporte
{1,3}	0,6666
{2,3}	0,3333
{1,2,7}	0,16666
{2,9}	0,5

Fonte: Adaptado de Amo (2003)

Conforme apresentado na tabela 9, se o suporte definido é no mínimo 50% os *Itemsets* considerados frequentes são os {1,3} e {2,9}. Porém se o suporte for de no mínimo 60% apenas o *Itemset* {1,3} será válido, ou seja, considerado frequente (AMO, 2003).

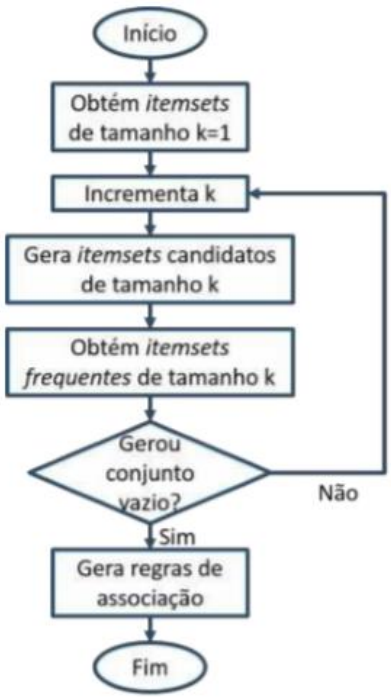
2.2.5.1.1 ALGORITMO APRIORI

Para Cornelios Junior (2015) o algoritmo Apriori é muito popular na aplicação da técnica de associação. O Apriori implementa a seguinte regra, se um grupo de dados é constante então todos os seus subconjuntos devem ser frequentes (AGRAWAL; IMIELINSKI; SWAMI, 1993).

Imagine que em um supermercado o conjunto {arroz, feijão, carne} é frequente nas compras, então com base no Apriori os seus subconjuntos também são frequentes, sendo eles: {arroz}, {feijão}, {carne}, {arroz, feijão}, {arroz, carne} e {feijão, carne} menos o conjunto vazio {} que também é um subconjunto.

Para uma maior compreensão da lógica que o algoritmo Apriori implementa, analise o fluxograma apresentado pela figura 4.

Figura 4 - Fluxograma Apriori



Fonte: (SCHMITZ, 2015)

A figura 4 apresenta o conjunto de passos lógicos que o algoritmo Apriori executa; de início no passo 1 é identificado os potenciais conjuntos com a quantidade de itens (k) mínima, onde os subconjuntos possuem apenas um elemento. A cada interação é incrementado o valor de (k) para obter os demais subconjuntos de *Itemsets* candidatos.

Os *Itemsets* candidatos são todos as combinações de subconjuntos de tamanho definido por (k). *Itemsets* frequentes são todos os subconjuntos gerados que atendem o valor do suporte estabelecido previamente.

A condição de parada do algoritmo é quando o mesmo chegar no *Itemset* vazio, neste momento os conjuntos frequentes frutos das interações anteriores serão usados para a gerar as regras de associação.

Para criar as regras de associação cada *Itemset* é dividido de forma a individualizar todos os seus subconjuntos possíveis, depois é feito um filtro com base na confiança mínima estabelecida.

2.2.5.2 CLASSIFICAÇÃO

Para Tan, Steinbach e Kumar (2009) a técnica de classificação constitui-se do propósito de dividir um conjunto de objetos em grupos pré-definidos. Um exemplo, para classificar *e-mails* como inofensivo ou *spam*, é necessário primeiro criar uma classe de *spam* com características de uma mensagem de *spam* real, uma eventual mensagem de *e-mail* nova será classificada com base nas similaridades com a classe de *spam* criada anteriormente, se a mesma possuir características semelhantes será classificada como *spam*, os atributos usados na comparação serão o assunto e o corpo do *e-mail*.

Para Mendes (2011) essa técnica permite a criação de modelos de classificação, isto é, aptos para identificar a função que descreva uma determinada classe que um objeto pertence com base no conjunto de dados de entrada.

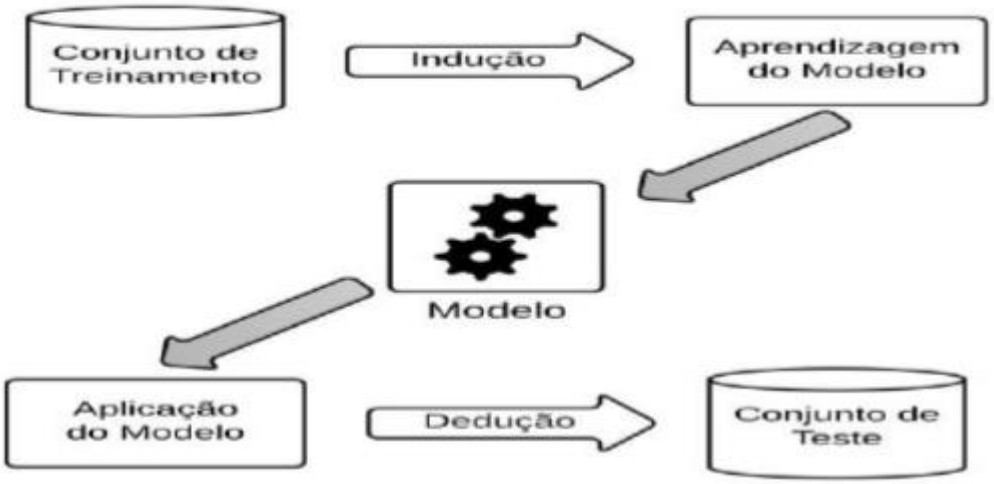
O conjunto de dados de entrada pode ser retratado por objetos ou registros, onde cada registro é constituído por dois atributos (X e Y) onde:

- X: é o conjunto de atributos do tipo discreto (número finito) ou contínuo (número infinito de valores possíveis);
- Y: é um atributo especial do tipo discreto denominado de classe rótulo.

Diante dos fatos apresentados Mendes (2011) conclui que a técnica de classificação identifica padrões nos registros e os classificam com base na similaridade das classes predefinidas.

Para Cornelios Junior (2015) o modelo de classificação pode ser representado conforme apresenta a figura 5.

Figura 5 - Classificador



Fonte: Costa, Baker, Amorim, Magalhães e Marinho (2013).

A figura 5 apresenta o “conjunto de treinamento” como entrada para o algoritmo de aprendizagem, esse conjunto é constituído por dados de amostra onde a classe já é conhecida, em cima desses dados na tarefa de “aprendizagem do modelo” é feito a indução de um modelo classificador que em seguida será submetido a teste juntamente com outro “conjunto de teste”, que é composto de amostras de dados sem classe predefinida e precisam ser classificadas e comparadas ao resultado do modelo criado.

Para Cornelios Junior (2015) a técnica de classificação é separada em duas fases:

- treinamento: essa etapa também é conhecida como aprendizagem, nela é utilizado um conjunto de amostragem de dados já associados as classes

predefinidas (rótulos) para se construir um modelo classificador. Como o conjunto de dados é conhecido essa etapa se enquadra no tipo de aprendizagem supervisionado;

- classificação: nessa etapa é feito o uso do modelo classificador criado na fase anterior para avaliar o grau de exatidão. Para isso, é usado um conjunto de dados teste sem nenhuma classe predefinida, dessa forma o modelo classificador pode ser testado ao comparar os dois resultados: o algoritmo classificador recebendo como entrada o conjunto de dados com classes predefinidas e recebendo como entrada o conjunto de dados teste sem classe predefinida.

2.2.5.3 REGRESSÃO LINEAR

Segundo Witten e Frank (2005) a técnica de regressão é parecida com a técnica de classificação, a diferença é que na regressão linear os atributos alvos são do tipo numérico e continuo e na classificação são do tipo discreto.

Para Elmasri e Navarthe (2005) regressão linear é um tipo de regra de classificação especial, pois se determinada regra de classificação é vista como uma função que atua sobre variáveis e as organiza em uma classe alvo (padrão) a regra é vista como regressão.

Para Elmasri e Navarthe (2005) a técnica de regressão linear pode ser vista quando, ao invés de organizar um conjunto de registros em uma classe singular, o valor do registro que é predefinido com base naquele conjunto. Para exemplificar considere a seguinte relação:

{id_paciente, exame1, exame2, exame3, exame4 ..., exame n}

Essa relação contém os resultados de n exames de um determinado paciente de um hospital. A variável alvo que se pretende prever é P que é a chance de o paciente sobreviver, nesse caso a regra de regressão linear assumiria a seguinte forma (ELMASRI e NAVARTHE, 2005):

(exame1 na faixa$_1$) e (exame2 na faixa$_2$) e ... (exame n na faixa$_n$) => $\mathrm{P} = \mathrm{x}$

ou

$$\mathrm{x} < \mathrm{P} \backslash le\ Y$$

A função deve ser adaptada de acordo com o objetivo, caso o interesse seja prever uma faixa de valores de P, P será considerada uma função:

$$P\ =\ f\ (exame\ 1,\ exame\ 2,\ \ldots\ exame\ n)$$

Essa função é geralmente representada da seguinte forma:

$$Y = f\ (x_{1,}\ x_{2,\ \ldots,}\ x_n)$$

Onde Y é a variável alvo que se deseja descobrir o valor, f é a função linear no domínio dos argumentos x_i, onde a tarefa de derivar f de um grupo de tuplas $< x_1, x_{2, \ldots,} x_n>$ é denominado regressão linear.

As regras de regressão linear podem ser aplicadas em um conjunto de dados para alcançar objetivos do tipo preditivo como: prever a quantidade aproximada de vendas de uma determinada mercadoria, predizer a quantidade de aproximada de lucro em um determinado período, entre outros.

A seguir, será abordada a técnica de agrupamento que também é muito parecida com a de classificação.

2.2.5.4 AGRUPAMENTO

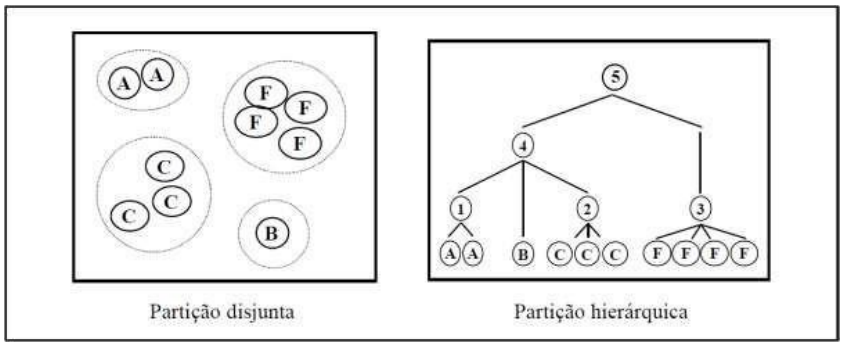
A técnica de agrupamento faz parte do processo de aprendizado não supervisionado, é aplicado para encontrar grupos homogêneos em uma base de dados (Kotsiantis & Pintelas, 2004). Essa técnica tem por função dividir a base de dados em subconjuntos menores de acordo com a similaridade dos objetos, esses subconjuntos são conhecidos como de *clusters* ou grupos em português (SILVA, 2016).

Diferente da técnica de classificação, o algoritmo que implementa a técnica de agrupamento não recebe nenhum grupo pré-definido para fazer os agrupamentos (SCHMITZ, 2015). O agrupamento acontece em razão das informações descritivas contidas no objeto e seus relacionamentos, o objetivo é que cada grupo possua objetos semelhantes entre si e distintos dos objetos dos demais grupos, quanto mais idênticos forem os objetos entre si mais distintos serão os grupos (MENDES, 2011).

Agrupamento possui duas categorias, sendo elas, "plana" e "hierárquica" (SCHMITZ, 2015). A seguir a figura 6 apresentará a distinção entre as duas

categorias, sendo à esquerda a representação do agrupamento plano (partição disjunta) e a direita a representação do agrupamento hierárquico.

Figura 6 - Agrupamento plano e hierárquico



Fonte: Schmitz (2015) *apud* Wives (2002, p. 93)

Segundo Mendes (2011) o agrupamento hierárquico é criado e organizado em forma de árvore onde os grupos são denominados nós e possuem como filhos os seus subgrupos e o nó raiz da árvore possui todos os grupos da árvore como filhos.

A figura 6 apresenta o nó 5 como a raiz da árvore, o nó 4 possui como filhos os nós (subgrupos) 1 e 2, os nós A's, B, C's e F's são os nós folhas, ou seja, não possuem filhos. O nó 5 representa todo o conjunto de dados o nó 4 representa a junção dos nós 1 e 2 e seus filhos.

Segundo Diniz e Louzada Neto (2000) o agrupamento plano busca de forma objetiva dividir os *n* elementos em *k* grupos sem o uso de ligações hierárquicas. O objetivo é limitado em apenas dividir o conjunto total de dados em subconjuntos em função da similaridade dos dados descritivos dos objetos.

Uma aplicação bastante útil dessa técnica no processo KDD pode ser feita na ocasião de identificar de forma automática os dados com ruídos e os dados redundantes da base de dados, esses problemas foram abordados nas seções (2.1.2.3) e (2.1.2.4). Averiguar uma base de dados extensa de forma manual com o objetivo de encontrar objetos redundantes (objetos duplicados) e com ruídos (que possui algum valor anormal em relação aos demais objetos) demanda muito tempo, é exaustivo e têm o risco da falha humana.

Ao término da aplicação das técnicas de DM, os resultados devem ser interpretados e avaliados.

2.3 INTERPRETAÇÃO E AVALIAÇÃO

Segundo Prass (2004) essa etapa é feita junto com o especialista do domínio, todo o conteúdo resultado da aplicação das técnicas de DM deve ser interpretado e avaliado para saber se o objetivo que impulsionou aplicar DM foi alcançado.

Caso os resultados sejam satisfatórios, é feita a aplicação dos mesmos, do contrário o processo de DM pode retroceder a qualquer uma das fases. Quando o resultado é considerado ruim, duas ações são muito usadas, a primeira é alterar o conjunto da dados de entrada, a segunda é trocar algoritmo de DM aplicado.

O usuário especialista do domínio necessita de alguma ferramenta para visualizar e interagir com os dados resultante do KDD, esse *software* deve ser capaz de fazer filtros e pesquisas nos resultados (SCHMITZ, 2015). Utilizar ferramentas de filtros pode não ser o ideal, pois os resultados podem ser muito extensos e de difícil entendimento pelo ser humano.

Nesse sentido, o uso de gráficos é mais benéfico pois permite saber as respostas de várias perguntas sobre o modelo de dados de forma visual e amigável. Perguntas como: qual a quantidade de padrões extraídos sobre o assunto x? Qual a quantidade e grupos extraídos dos dados? Cada resultado pode ser visualizado por diferentes tipos de gráficos (SCHMITZ, 2015).

3 MATERIAIS E MÉTODOS

Esta seção expõe a metodologia que foi empregada neste trabalho, quais ferramentas e tecnologias foram usadas para alcançar os objetivos específicos estabelecidos.

3.1 MATERIAIS

Para executar as tarefas de DM foi utilizado o *software* RStudio Desktop na versão 1.1.456. O RStudio está disponível gratuitamente sobe a licença AGPL v3 e pode ser instalado em plataformas como Linux, Windows e MacOS. Esta ferramenta permite a instalação de um conjunto vasto de pacotes e bibliotecas necessárias para a execução das tarefas de DM de forma gráfica e amigável.

A escolha do R Studio se deu pela sua popularidade no que tange a vasta quantidade de conteúdo documental na internet e a boa velocidade de processamento de dados promovida pela linguagem R, que no caso será utilizada na versão 3.5.1.

Para exploração da base de dados do estudo de caso foi usado o programa SGBD *SQL Server Management Studio* 2014 na versão gratuita. Foi utilizado o *framework* Laravel em conjunto com o banco de dados SQLite para preparar os dados para os algoritmos Apriori e Linear Model.

O pacote *office* da *Microsoft* foi utilizado para os cálculos trimestrais das vendas manipulação e visualização dos dados.

3.2 MÉTODOS

As tarefas de DM implementadas foram: regressão linear e associação. Para tanto, foi necessário fazer o levantamento bibliográfico do assunto para permitir um entendimento claro dos conceitos envolvidos.

De início, foi feito a importação da base de dados no SGBD para analisar a qualidade dos dados no que tange identificar as tabelas com dados ausentes

e dados discrepantes, aferir a dimensionalidade dos dados, observar e selecionar os atributos relevantes para a criação do modelo de dados.

Logo após, foi dado início ao preparo dos dados para que fosse possível tirar maior proveito dos algoritmos que implementam as tarefas de associação e regressão linear simples. Caso a quantidade de dados com problemas relacionados a qualidade fosse grande e a identificação dos problemas dificultosa, seria executado a tarefa de agrupamento para dar agilidade nesse processo, pois a mesma gera grupos de objetos com similaridades o que evidenciaria os grupos de dados com discrepâncias, porém, não foi necessário.

Em seguida foram utilizados o algoritmo Apriori para a tarefa de associação e a função *linear model* (LM) para a tarefa de regressão linear simples. Ao final, os resultados foram analisados juntamente com o especialista do domínio e validados. Para uma maior compreensão dos resultados os mesmos foram apresentados em forma de gráficos gerados pelo RStudio e todas as etapas do processo de DM documentados.

4 RESULTADOS E DISCUSSÃO

Os resultados apresentados nesta seção são derivados da utilização de um banco de dados de uma farmácia como objeto de estudo para aplicação de DM, em específico as tarefas de associação e regressão linear simples.

As seções posteriores expõem de forma detalhada as etapas do trabalho, a partir da exportação do banco de dados original até a análise dos resultados.

4.1 EXPORTAÇÃO E IMPORTAÇÃO DO BANCO DE DADOS

O primeiro passo realizado foi a exportação do banco de dados da base original, que é gerido pelo SGBD SQL *Serve Management Studio* 2014, o uso da instrução abaixo gerou um arquivo de *backup* de extensão .bak que pode ser importado em outro SGBD.

Figura 7 - Instrução SQL para fazer backup

```
BACKUP DATABASE INOVAFarma TO DISK='C:\BACKUP\BANCO_BACKUP-FULL.BAK'
GO
```

Para importação do SGBD foram executados os seguintes passos:

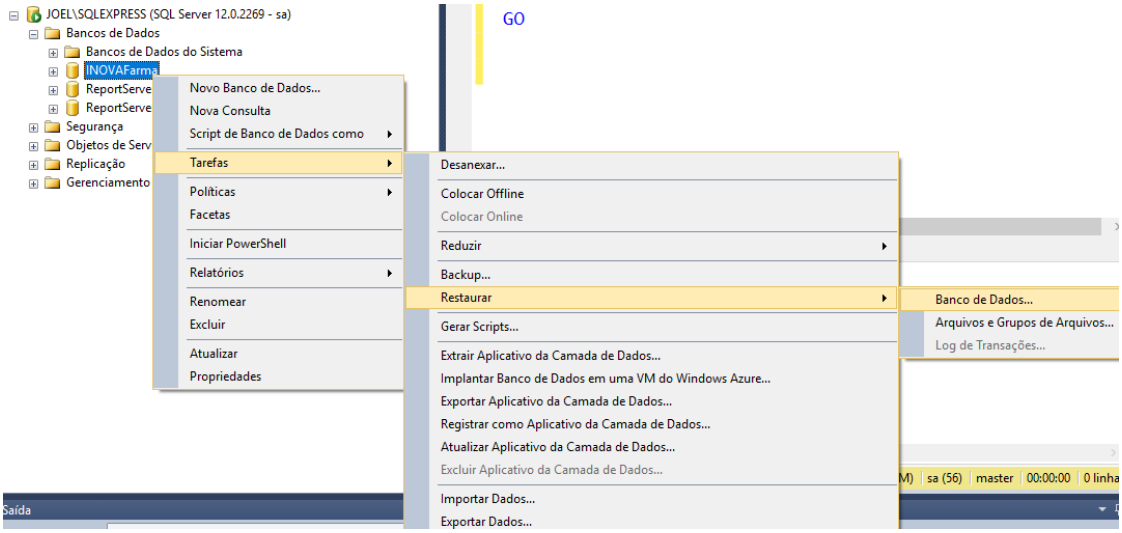
- Passo 1: criar um banco de dados, a figura 8 apresenta a instrução para criar a base de dados de nome INOVAFarma.

Figura 8 - Instrução SQL para criar um banco de dados

```
CREATE DATABASE INOVAFarma
GO
```

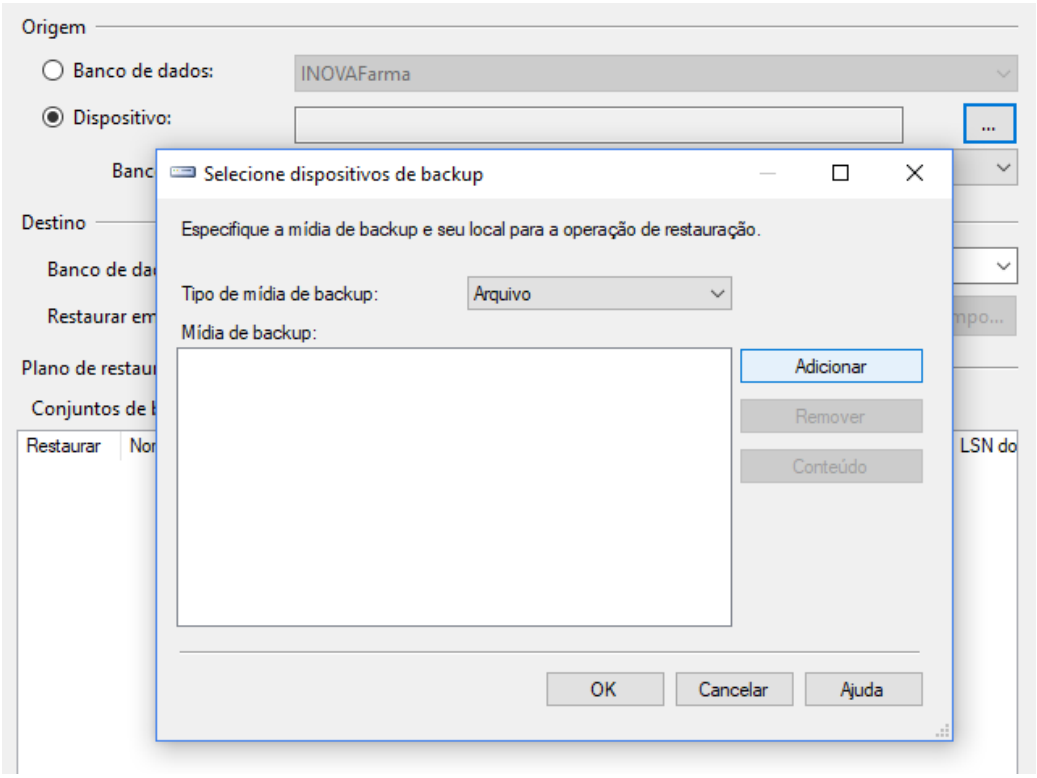
- Passo 2: restaurar a base de dados, esse processo se faz necessário para proteger a base de dados original. Todas as manipulações dos dados que foram efetuadas no preparo dos dados para aplicar as tarefas de DM ocorreram em outro SGBD com a cópia exata do banco de dados original. A figura 9 apresentará a primeira etapa do processo de restauração.

Figura 9 - Etapa 1 do processo de restauração



A figura 9 apresenta a primeira etapa do passo 2 para restauração da base de dados que, no caso, a ação é clicar com o botão direito do mouse em cima da base de dados criada, escolher a opção ‘tarefas’, ‘restaurar’ e clicar na opção ‘Banco de dados’. A figura 10 apresenta a próxima tela a ser exibida pelo SGBD.

Figura 10 - Etapa 2 do processo de restauração



A figura 10 apresenta as ações: escolher a opção ‘Dispositivo’ e clicar no botão ‘**...**’ **.** Ao clicar em adicionar será apresentada uma janela para escolher o arquivo de restauração a ser anexado para iniciar a restauração do banco de dados.

A seguir na seção 4.2 será exibido a etapa de pré-processamento dos dados, onde será apresentado as fases: seleção, limpeza e transformação dos dados.

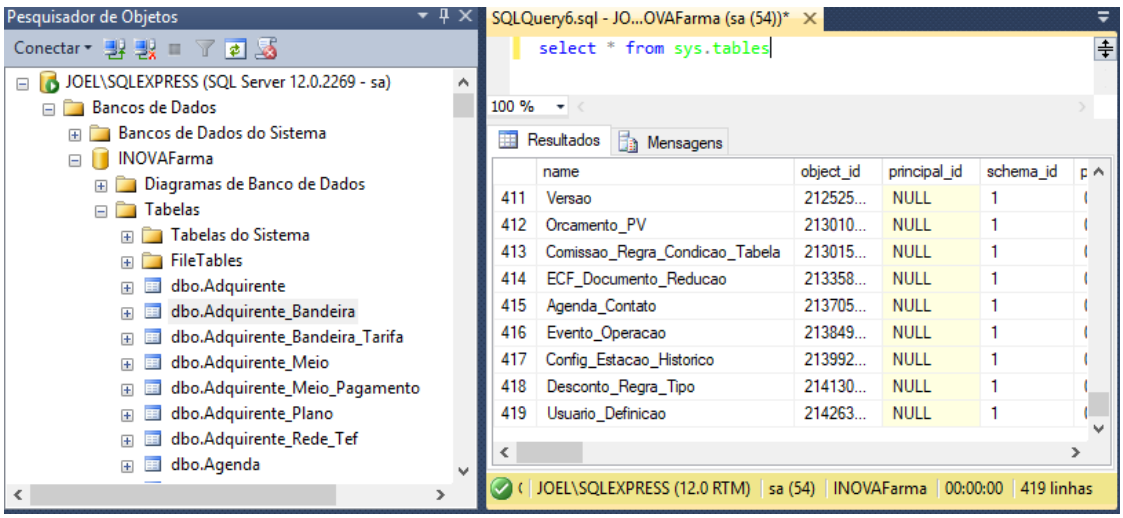
4.2 ETAPA DE PRÉ-PROCESSAMENTO

Nesta etapa, os dados foram preparados para posteriormente serem submetidos aos algoritmos de DM, nesse caso, o algoritmo Apriori para a tarefa de associação e a função LM para a tarefa de regressão linear simples.

4.2.1 SELEÇÃO DE DADOS

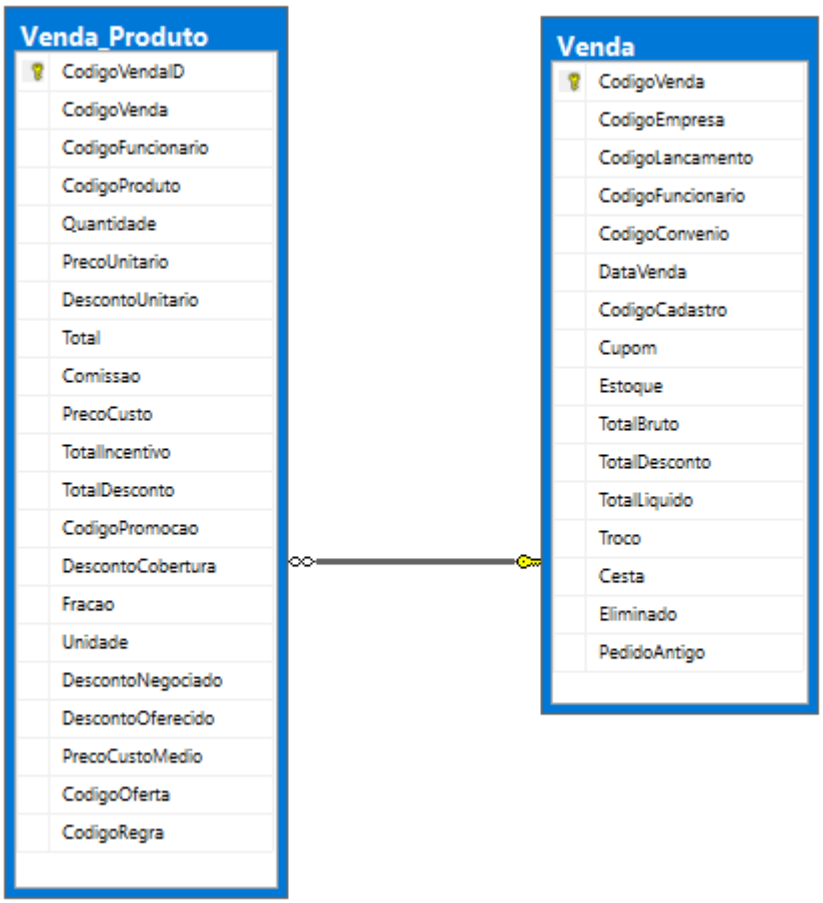
A base de dados completa do estudo de caso possui 419 tabelas, conforme apresenta a figura 11.

Figura 11 - Quantidade de tabelas



As tabelas necessárias para alcançar os objetivos do presente trabalho são apenas duas: a tabela de Venda e a tabela Venda_Produto. Na tabela Venda foi executado a tarefa de regressão linear simples e na tabela Venda_Produto foi aplicado a tarefa de associação. A figura 12 apresenta as mesmas com seus respectivos atributos.

Figura 12 - Tabelas de interesse



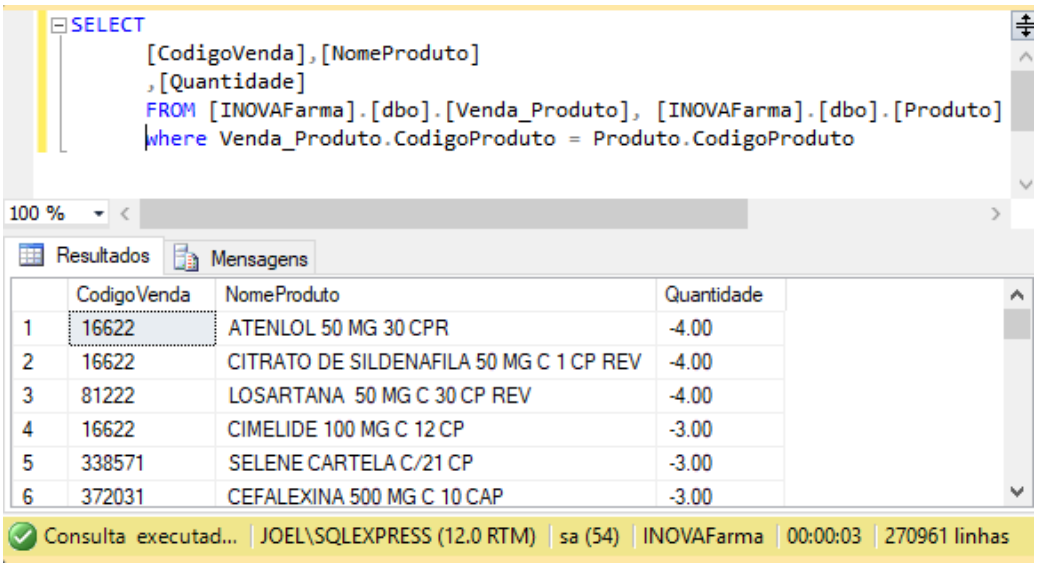
A figura 12 apresenta as tabelas Venda_Prodoto e Venda com seus respectivos atributos, a seguir na seção 4.2.1.1 será apresentado quais atributos foram utilizados para criar o modelo de dados.

4.2.1.1 ELIMINAÇÃO MANUAL DOS ATRIBUTOS

A quantidade de atributos nas tabelas Venda e Venda_Produto é desnecessária, muitos atributos não são necessários para alcançar os objetivos

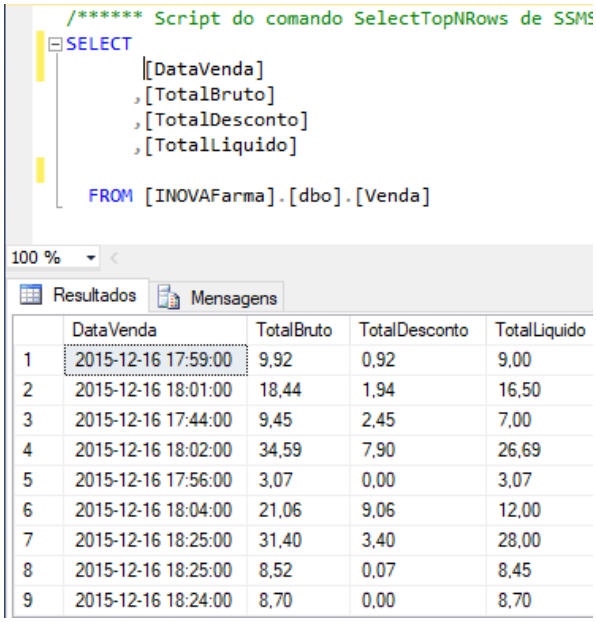
propostos. Para resolver esse problema foi feita a redução da dimensionalidade dos dados através da eliminação manual dos atributos irrelevantes. A figura 13 apresenta a consulta no banco de dados para retornar apenas os dados dos atributos importantes da tabela Venda_Produto.

Figura 13 - Redução de atributos da tabela Venda_Produto



A tabela Venda_Produto era composta por 21 atributos, sendo que os atributos necessários são apenas 3: CodigoVenda, NomeProduto e Quantidade. Essa tabela possui duzentos e setenta mil novecentos e sessenta e um registros. A seguir na figura 14 será demonstrada a eliminação manual dos atributos na tabela Venda.

Figura 14 - Redução de atributos da tabela Venda



	DataVenda	TotalBruto	TotalDesconto	TotalLiquido
1	2015-12-16 17:59:00	9,92	0,92	9,00
2	2015-12-16 18:01:00	18,44	1,94	16,50
3	2015-12-16 17:44:00	9,45	2,45	7,00
4	2015-12-16 18:02:00	34,59	7,90	26,69
5	2015-12-16 17:56:00	3,07	0,00	3,07
6	2015-12-16 18:04:00	21,06	9,06	12,00
7	2015-12-16 18:25:00	31,40	3,40	28,00
8	2015-12-16 18:25:00	8,52	0,07	8,45
9	2015-12-16 18:24:00	8,70	0,00	8,70

O objetivo a ser alcançado com os dados da tabela Vendas é predizer a quantidade em valor líquido que será vendido no segundo trimestre de 2019; a tabela tinha dezesseis atributos, com a eliminação dos desnecessários ficaram apenas quatro, sendo eles: DataVenda, TotalBruto,TotalDesconto e TotalLiquido. Essa tabela possui cento e oitenta e sete mil quinhentos noventa e seis registros.

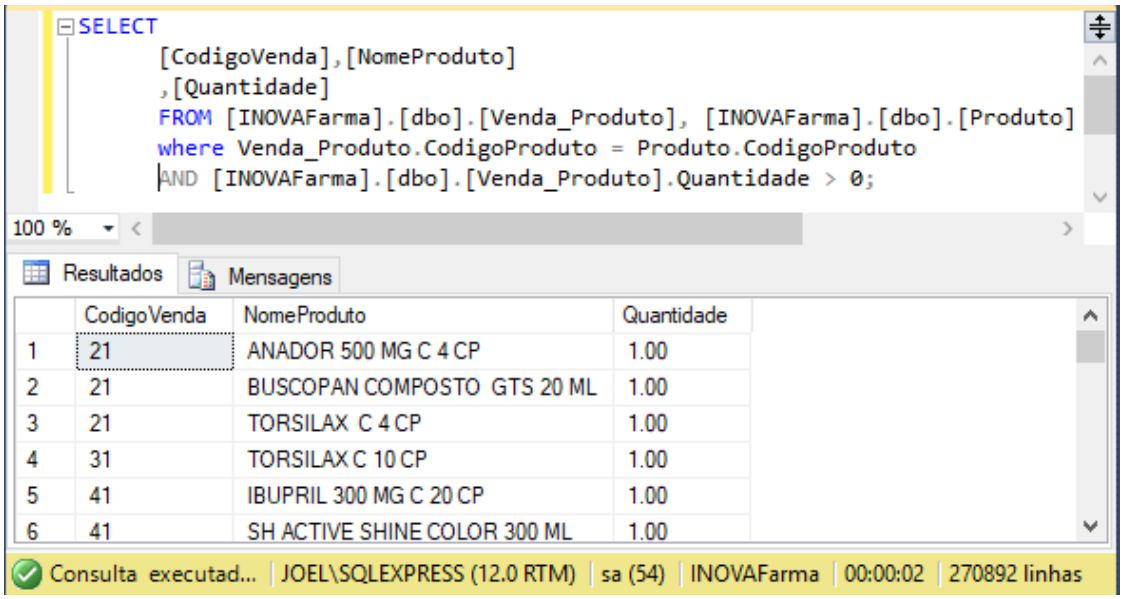
O atributo DataVenda foi mantido apenas para calcular os lucros trimestrais, os atributos TotalBruto e Total desconto foram retidos apenas para constatar se há alguma irregularidade na subtração que resulta no TotalLiquido, tais averiguações foram feitas na etapa de limpeza que será apresentada a seguir.

4.2.2 LIMPEZA

Conforme pode se observar na figura 13 alguns produtos estão apresentando venda negativa, ou seja, a quantidade de produto vendida é menor que um, dessa forma fica evidente que esse dado é um *outlier*, um ruído.

A eliminação desses registros com o valor de quantidade negativa no banco de dados foi feita através da instrução SQL conforme ilustra a figura 15.

Figura 15 - Eliminação de ruídos da tabela Venda_Produto



A figura 15 apresenta a consulta de eliminação dos ruídos com 270892 linhas resultantes, a tabela original possuía 270961 linhas, ou seja, 69 registros foram eliminados, o que correspondem a cerca de 0,02% do total. Os dados de todas as consultas realizadas foram exportados para arquivos de extensão .csv.

A tabela Vendas apresentou alguns ruídos e uma inconsistência, como: o desconto maior que o preço bruto do produto e valores negativos, a tabela 11 apresenta a tabela Vendas com os valores descritos.

Tabela 11 - Tabela Venda com inconsistência e ruídos

1	TotalBruto	TotalDesconto	TotalLiquido
2	19,75	781,97	- 762,22
3	- 345,54	- 216,75	- 128,79
4	- 167,64	- 48,63	- 119,01
5	- 106,74	- 36,74	- 70,00
6	- 87,66	- 24,66	- 63,00
7	- 110,99	- 57,99	- 53,00
8	- 87,94	- 36,94	- 51,00
9	- 111,49	- 61,50	- 49,99
10	- 111,49	- 61,50	- 49,99
11	- 53,73	- 8,74	- 44,99
12	- 47,79	- 7,14	- 40,65
13	- 43,22	- 3,22	- 40,00
14	- 64,86	- 25,86	- 39,00

Apenas a primeira linha apresentou um dado inconsistente onde o desconto é maior que o valor do produto, por esta razão a linha foi excluída. Os demais dados com valores negativos foram aproveitados pois o valor do Total-Liquido é consistente com o cálculo do TotalBruto menos o TotalDesconto. A ação utilizada foi apenas passar os valores negativos para positivos, a tabela 12 apresenta o resultado.

Tabela 12 - Tabela Venda sem inconsistência e ruídos

	A	B	C
1	TotalBruto	TotalDesconto	TotalLiqui
2	345,54	216,75	128,79
3	167,64	48,63	119,01
4	106,74	36,74	70,00
5	87,66	24,66	63,00
6	110,99	57,99	53,00
7	87,94	36,94	51,00
8	111,49	61,5	49,99
9	111,49	61,5	49,99
10	53,73	8,74	44,99
11	47,79	7,14	40,65
12	43,22	3,22	40,00
13	64,86	25,86	39,00

Após a adequação dos dados, foi voltado a etapa de seleção e eliminados os atributos TotalBruto e TotalDesconto, pois só serviram de validação para o atributo TotalLiquido que será aproveitado na tarefa de regressão, o atributo DataVenda também foi eliminado depois do cálculo dos lucros trimestrais.

A seguir na etapa de transformação de dados será demonstrado as atividades dessa fase para fazer os alinhamentos finais dos dados para a execução das tarefas de DM.

4.2.3 TRANSFORMAÇÃO DOS DADOS

Os dados exportados possuíam uma estrutura inadequada para aplicar os algoritmos de associação e regressão linear, pois os mesmos exigem uma estrutura e tipagem dos dados diferente, dessa forma, se fez necessário a manipulação dos mesmos para o devido ajuste. A seguir a tabela 13 apresenta a estrutura dos dados da tabela venda-produto antes da transformação.

Tabela 13 - Tabela Venda_Produto antes da transformação

1	CodigoVenda	NomeProduto
2	11	TORSILAX C 4 CP
3	11	ANADOR 500 MG C 4 CP
4	21	ANADOR 500 MG C 4 CP
5	21	BUSCOPAN COMPOSTO GTS 20 ML
6	21	TORSILAX C 4 CP
7	31	TORSILAX C 10 CP
8	41	IBUPRIL 300 MG C 20 CP
9	41	SH ACTIVE SHINE COLOR 300 ML
10	51	ANADOR 500 MG C 4 CP
11	61	TORSILAX C 4 CP
12	61	ANGITENS 25 MG C 30 CP
13	61	ABSORVENTE SYM COB SUAVE C ABAS TRANSP LILAS
14	71	FLORENT 100 MG C 12 CAP

Conforme ilustra a tabela 13 os dados da tabela venda-produto estão organizados da seguinte forma: cada item de venda é organizado em uma linha diferente repetindo apenas o código da venda para referenciar a transação. O

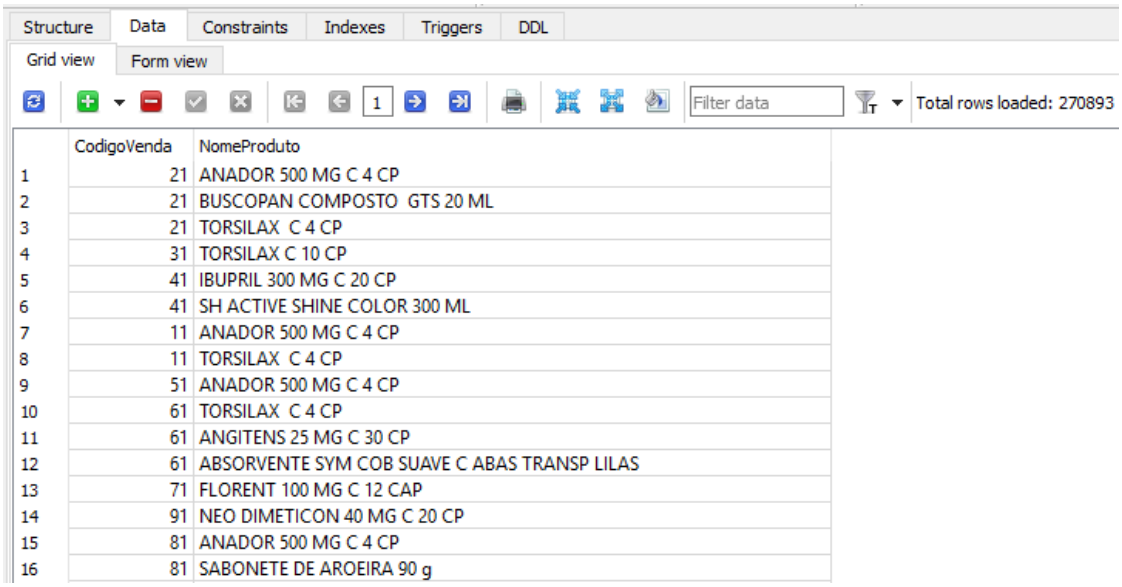
algoritmo de associação entende cada linha como uma transação, dessa forma, surge a necessidade de estruturar cada item de venda em uma única linha e colunas diferentes. A tabela 14 apresenta a mesma tabela organizada da maneira como requer o algoritmo de associação.

Tabela 14 - Tabela Venda_Produto organizada

	CodigoVenda	Produto1	Produto2	Produto3
2	11	TORSILAX C 4 CP	ANADOR 500 MG C 4 CP	
3	21	ANADOR 500 MG C 4 CP	BUSCOPAN COMPOSTO GTS 20 ML	TORSILAX C 4 CP
4	31	TORSILAX C 10 CP		
5	41	IBUPRIL 300 MG C 20 CP	SH ACTIVE SHINE COLOR 300 ML	
6	51	ANADOR 500 MG C 4 CP		
7	61	TORSILAX C 4 CP		
8	61		ANGITENS 25 MG C 30 CP	ABSORVENTE SYM COB SUAVE C ABAS TRANSP LILAS
9	71	FLORENT 100 MG C 12 CAP		

Para organizar os dados conforme ilustra a tabela 14 foi adotado uso do *framework* Laravel, o banco de dados SQLite, e a linguagem de programação PHP. A base de dados anteriormente exportada da base original para csv foi importada no SGBD SQLite. A seguir a figura 16 exibe a base de dados importada.

Figura 16 - Tabela Venda_Produto SQLite



	CodigoVenda	NomeProduto
1	21	ANADOR 500 MG C 4 CP
2	21	BUSCOPAN COMPOSTO GTS 20 ML
3	21	TORSILAX C 4 CP
4	31	TORSILAX C 10 CP
5	41	IBUPRIL 300 MG C 20 CP
6	41	SH ACTIVE SHINE COLOR 300 ML
7	11	ANADOR 500 MG C 4 CP
8	11	TORSILAX C 4 CP
9	51	ANADOR 500 MG C 4 CP
10	61	TORSILAX C 4 CP
11	61	ANGITENS 25 MG C 30 CP
12	61	ABSORVENTE SYM COB SUAVE C ABAS TRANSP LILAS
13	71	FLORENT 100 MG C 12 CAP
14	91	NEO DIMETICON 40 MG C 20 CP
15	81	ANADOR 500 MG C 4 CP
16	81	SABONETE DE AROEIRA 90 g

Depois de importar a base de dados foi feita a manipulação da mesma. Para tanto, foi criado um projeto PHP com o Laravel e conectado a base de

dados SQLite, o nome do banco cujo os dados foram importados é TCC. A figura 17 apresenta o arquivo de conexão da base de dados no projeto Laravel.

Figura 17 – Arquivo de conexão do banco de dados

```
 9   DB_CONNECTION=sqlite
10   DB_HOST=127.0.0.1
11   DB_PORT=3306
12   DB_DATABASE="C:\Users\Joel dos Santos\Documents\TCC\database\TCC.sqlite"
13   DB_USERNAME=homestead
14   DB_PASSWORD=secret
```

A figura 17 exibe os parâmetros necessários para configurar a base de dados em um projeto Laravel, na linha 9 é informado o tipo de banco de dados, na linha 10 a máquina, que no caso é a local, linha 11 a porta da máquina que será utilizada pelo banco, linha 12 o caminho do arquivo do banco na máquina, linha 13 o nome do usuário do banco e na linha 14 é informado a senha do banco de dados.

Depois de importados, os dados foram organizados por meio de um algoritmo php criado para este propósito. A figura 18 apresenta o código que estrutura todos os itens de cada venda em uma única linha, separando cada elemento por ponto e vírgula.

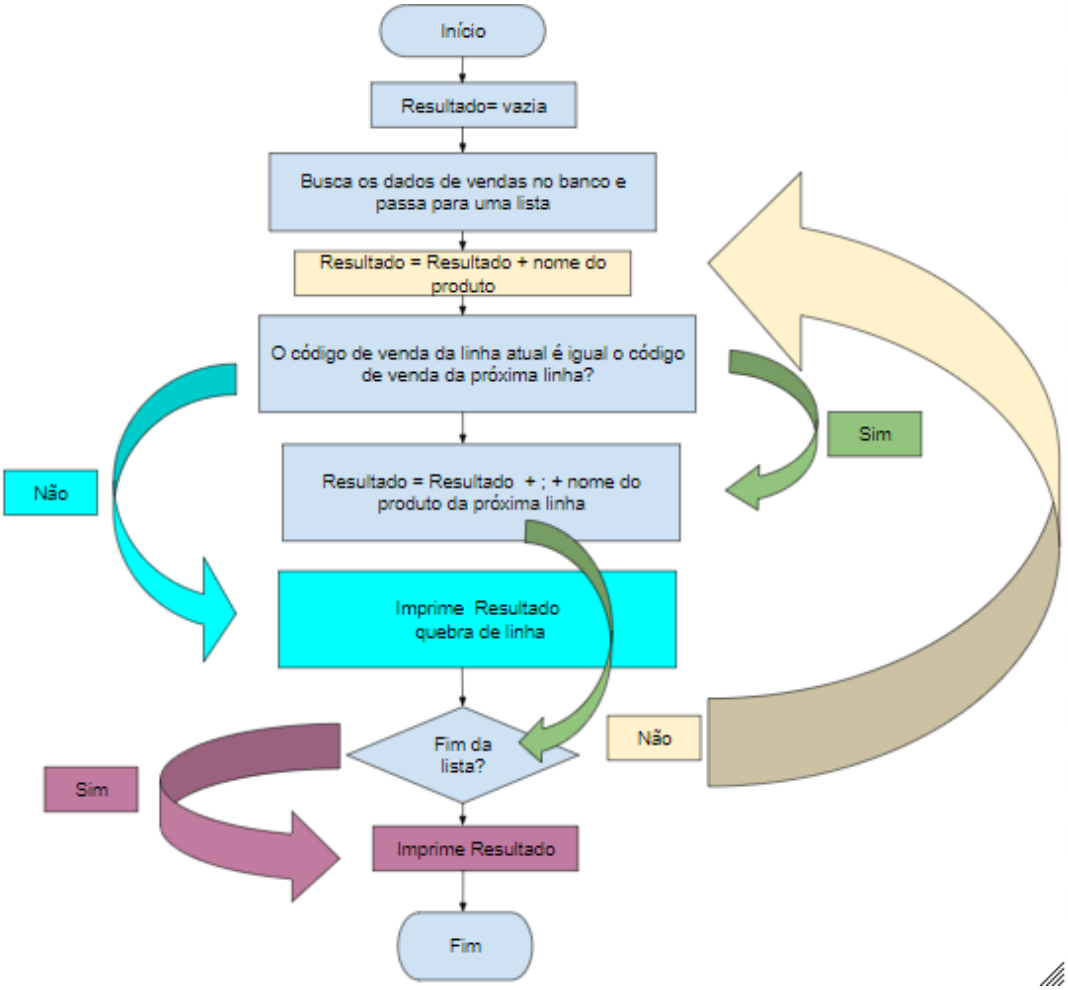
Figura 18 – Algoritmo PHP organizador dos dados de venda-produto

```php
<?php
Use App\vendas;
use Illuminate\Support\Facades\DB;

Route::get('/', function () {
    $vendas = DB::table('vendas')->get();
    $resultado = "";
    $t = count($vendas)-1;
    for ($i = 0; $i < count($vendas); $i ++)
    {
        if($i < $t and $vendas[$i]->CodigoVenda != $vendas[$i+1]->CodigoVenda and $resultado == "")
        {
            $resultado = $vendas[$i]->NomeProduto;
        }
        if($i < $t and $vendas[$i]->CodigoVenda == $vendas[$i+1]->CodigoVenda)
        {
            if($resultado == "")
            {
                $resultado = $vendas[$i]->NomeProduto;
            }
            if($resultado != "")
            {
                $resultado = $resultado.";".$vendas[$i+1]->NomeProduto;
            }
        }
        else
        {
            echo $resultado."<br>";
            $resultado = "";
        }
    }
});
```

Para um melhor entendimento do algoritmo apresentado, a figura 19 exibe o pseudocódigo do mesmo.

Figura 19 – Pseudocódigo do algoritmo organizador dos dados de venda-produto



Conforme apresentado na figura 19 os nomes dos produtos contidos em cada linha foram separados por ponto e vírgula, cada linha representa uma venda e cada produto da mesma linha representa um item de venda. A saída impressa pode ser copiada para qualquer editor de texto e salva em diferentes extensões como txt e csv. Esses foram os ajustes feitos na tabela venda-produto e depois foi dado início no ajustamento dos dados de vendas.

Quanto a tarefa de regressão linear simples, o objetivo a ser alcançado é predizer o valor total líquido das vendas do segundo trimestre de 2019. Para tanto, foi realizado os cálculos das vendas trimestrais desde do ano da abertura da farmácia, que no caso foi em 2015.

Os trimestres escolhidos ocorreram a partir de janeiro de 2016 até junho de 2018, os mesmos foram aproveitados na tarefa de regressão pois os dados

estavam completos. A seguir a tabela 15 apresentará o cálculo das vendas de cada trimestre.

Tabela 15 – Regras de associação

TRIMESTRE	VALOR
SUB TOTAL 2 - 1 TRIM. 2016	254.471,71
SUB TOTAL 2 - 2 TRIM. 2016	249.338,95
SUB TOTAL 2 - 3 TRIM. 2016	208.588,85
SUB TOTAL 2 - 4 TRIM. 2016	249.856,40
SUB TOTAL 2 - 1 TRIM. 2017	279.050,00
SUB TOTAL 2 - 2 TRIM. 2017	304.824,51
SUB TOTAL 2 - 3 TRIM. 2017	307.305,62
SUB TOTAL 2 - 4 TRIM. 2017	328.085,79
SUB TOTAL 2 - 1 TRIM. 2018	330.501,90
SUB TOTAL 2 - 2 TRIM. 2018	385.556,67

Para executar o algoritmo linear model o modelo de dados ficou da seguinte forma conforme apresenta a figura 20.

Figura 20 – Tabela Venda transformada

	trimestre	VALOR
1	1	254.471,71
2	2	249.338,95
3	3	208.588,85
4	4	249.856,40
5	5	279.050,00
6	6	304.824,51
7	7	307.305,62
8	8	328.085,79
9	9	330.501,90
10	10	385.556,67

Tal alteração no formato da representação dos dados trimestrais de vendas se deu pelo motivo que a tarefa de regressão linear necessita que os dados sejam do tipo numérico conforme abordado na seção 2.1.3.1.

O arquivo de texto contém as mesmas informações da figura 29, onde cada linha representa um trimestre com seu respectivo valor separado por um espaço. Como o trimestre 10 representa o segundo trimestre de 2018 o 14 trimestre representará o segundo trimestre de 2019.

Após a transformação dos dados, foi dado início a etapa de DM. A seção 4.3 demonstrará o uso do *software* RStudio para executar as tarefas de associação e regressão linear simples, tal como a exibição dos resultados.

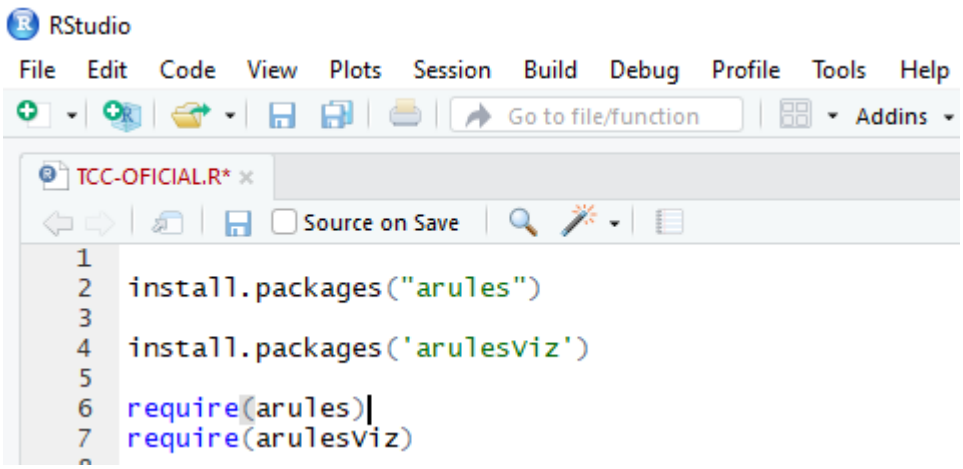
4.3 DATA MINING

Nesta seção será demonstrada a etapa de DM, em específico as tarefas de associação executada com o algoritmo Apriori e a tarefa de regressão linear simples pelo algoritmo Linear Model.

4.3.1 TAREFA DE ASSOCIAÇÃO

O RStudio usa pacotes de terceiros para executar o máximo de tarefas de DM, no caso, é necessário a instalação do pacote Arules, o que pode ser feito de forma simples pois o mesmo possui um gerenciador de pacotes integrado. A instalação pode ser feita por linha de comando ou por interface gráfica, a figura 21 apresenta o método por linha de comando.

Figura 21 – Instalação do pacote Arules e ArulesViz

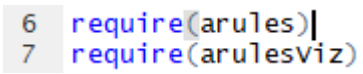


A instalação de qualquer pacote necessário segue a mesma estrutura demonstrada pela figura 21, o nome do pacote desejado é passado como

parâmetro, para executar a instalação é necessário clicar em cima do comando e depois apertar as teclas Ctrl e Enter.

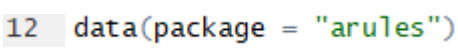
O pacote ArulesAviz estende o pacote Arules com várias técnicas de visualização para regras de associação e conjuntos de itens. Depois de instalado é necessário fazer a importação dos pacotes, pois esta é a forma de informar ao RStudio que se pretende usar um determinado recurso que depende de um pacote. A seguir a figura 22 apresenta o comando de importação.

Figura 22 – Importação de pacote no RStudio



Conforme demonstrado na figura 22 foram realizadas as importações dos pacotes Arules e AruleViz, resta agora informar que a base de dados que será utilizada é do tipo *groceries* - “mantimentos, cesta de supermercado” em português. Esse tipo de base de dados não possui número de colunas fixas em cada linha, porque pode haver compras com número de itens diferentes. A figura 23 mostra o comando que comunica ao RStudio esse tipo de base de dados.

Figura 23 – Informa o tipo de dados



A figura 23 apresentou o comando necessário para informar que a base de dados que será importada é do tipo “*groceries*”. O próximo passo dado foi importar a base de dados e executar os algoritmos necessários para minerar e gerar os relatórios.

4.3.1.1 IMPORTAÇÃO DA BASE DE DADOS E EXECUÇÃO DA TAREFA DE ASSOCIAÇÃO

Despois de instalar e importar os pacotes necessários foi feita a importação do arquivo com os dados organizados na etapa de transformação, a figura 24 apresenta a instrução que importa a base de dados.

Figura 24 – Importação da base de dados

```
17  Compras <- read.transactions(file = "C:\\Users\\Joel dos Santos\\Music\\TCC.csv", sep = ";")
```

A base de dados foi passada para uma variável de nome Compras. Para tanto, é necessário informar o caminho do arquivo, a extensão, que no caso é csv e o separador que distingue os itens de compra que é o ponto e vírgula.

Depois que os dados são importados, o RStudio exibe a quantidade de transações e o tamanho da base de dados em megabytes, mas é possível obter um resumo mais rico em informações, a seguir a figura 25 apresentará o comando que revela mais detalhes da base de dados.

Figura 25 – Comando para ver os detalhes de forma resumida da base de dados

```
24  #exibe os detalhes da base importada
25  summary(Compras)
26
23:1  (Top Level)
```

```
Console  Terminal
~/
> #exibe os detalhes da base importada
> summary(Compras)
transactions as itemMatrix in sparse format with
 200835 rows (elements/itemsets/transactions) and
 8426 columns (items) and a density of 0.0001600268

most frequent items:
               TORSILAX C 10 CP                DORALGINA C 4 DRG
                           4519                             4188
        CIMELIDE 100 MG C 12 CP       GASTROL 1 ENV EFERV TODOS
                           3410                             3386
CICLO 21 0,03 MG 0,15 MG C 21 C                         (Other)
                           3193                           252107

element (itemset/transaction) length distribution:
sizes
     1      2      3      4      5      6      7      8      9     10     11
153936  32048   9850   3178   1110    396    165     71     35     22     10
    12     13     14     16     17     20     26
     3      4      2      1      1      2      1

   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  1.000   1.000   1.000   1.348   1.000  26.000
```

Conforme pode ser observado na figura 25, a base de dados possui 200835 linhas (compras) e 8426 itens (produtos), os cinco produtos mais frequentes nas compras são apresentados na tabela 16 e a distribuição das vendas apresentado na tabela 17.

Tabela 16 – Itens mais frequentes nas compras

Produto	Qt. De compras presentes
TORSILAX C 10 CP	4519
DORALGINA C 4 DRG	4188
CIMELIDE 100 MG C 12 CP	3410
GASTROL 1 ENV EFERV TODOS	3386
CICLO 21 0,03 MG 0,15 MG C 21 C	3193

A tabela 16 exibe os produtos em ordem decrescente em função da quantidade de compras em o mesmo faz parte, nota-se que o Torsilax lidera estando presente em 4519 compras. A seguir a tabela 17 apresentará a distribuição das vendas.

Tabela 17 – Distribuição de frequência das vendas

Quantidade de produto na venda	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	20	26
Quantidade de vendas	153936	32048	9850	3178	1110	396	165	71	35	22	10	3	4	2	2	1	1	2	1

Conforme pode ser observado na tabela 17 existem 153936 compras com apenas 1 item, o que representa 76,64 % das transações. Isso significa que para encontrar as relações entre os itens que ocorrem na mesma compra será necessário colocar o suporte e a confiança muito baixo. A seguir a figura 26 apresentará o comando que gera as regras de associação.

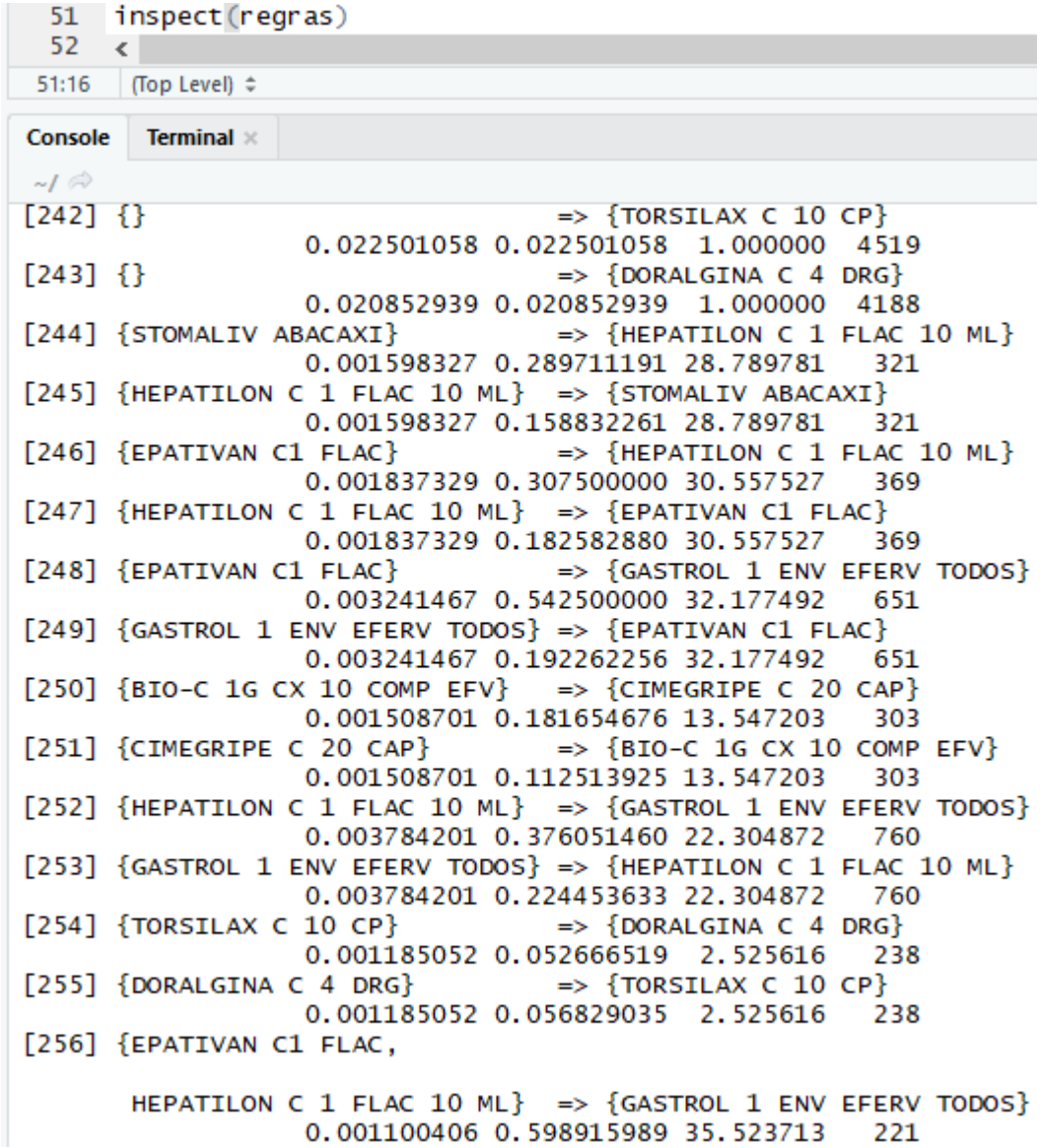
Figura 26 – Comando para ver os detalhes de forma resumida da base de dados

```
47  #gera regras de associação de acordo os parâmetros
48  regras <- apriori(Compras, parameter = list(supp = 0.0010, conf = 0.0010))
```

O conjunto regras geradas foram atribuídas a variável “regras”, onde o suporte e a confiança são de 0.0010; com esse ajuste do suporte e confiança

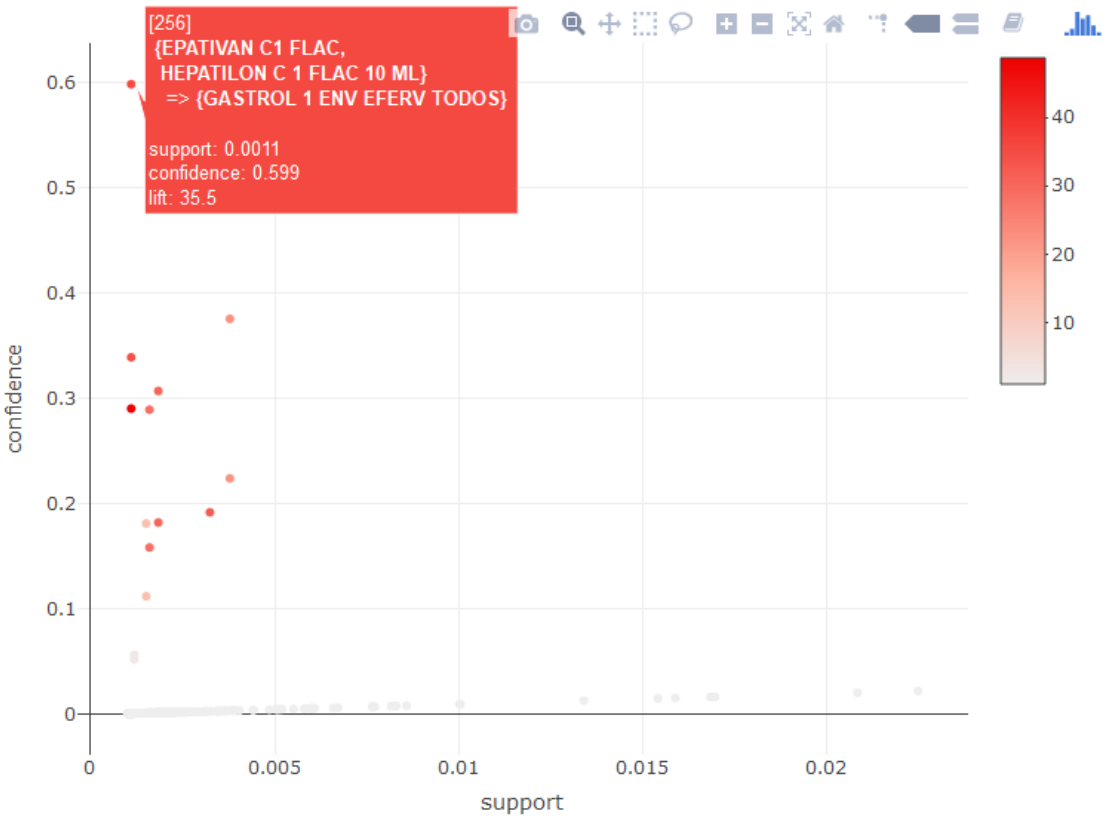
o resultado foi 258 regras, a seguir a figura 27 apresentará o conteúdo das regras.

Figura 27 – Comando para ver os detalhes de forma resumida da base de dados



Segundo apresenta a figura 27, da primeira regra até a 243 são formadas por apenas um item, o que não interessa para esse trabalho onde se buscou descobrir as relações entre os itens de compras que são vendidos juntos, dessa forma, a quantidade de regras analisadas foram 15. A seguir serão apresentadas no gráfico 1 as regras criadas ordenas pela confiança em ordem decrescente.

Gráfico 1 – Regras de associação



O gráfico 1 apresenta as 15 regras válidas ordenadas pelo suporte em ordem decrescente, a regra com maior suporte é a 256, onde a confiança é 0.599 e o suporte é 0.0011. Após a etapa de mineração de dados foi dado início a fase de avaliação dos resultados abordada na seção 4.3.2.

4.3.2 IMPORTAÇÃO DA BASE DE DADOS E EXECUÇÃO DA TAREFA DE REGRESSÃO LINEAR SIMPLES

O processo de importação segue o mesmo padrão já documentado anteriormente na seção 4.3.1.1 na tarefa de associação. A figura 28 exibe a importação dos dados.

Figura 28 – Importação dos dados de venda

```
#Importação dos dados
vendas = read.table("C:\\Users\\Joel dos Santos\\Music\\vendas.txt",header = T)
attach(vendas)
```

A base de dados foi passada para a variável “vendas”, o parâmetro header indica que a primeira linha do arquivo é o cabeçalho de títulos. A função attach faz com que o R entenda que “vendas” é um objeto e o coloca no ambiente de variáveis globais. Com o método View é possível ver a base de dados importada, a seguir a tabela 18 apresentará a base de dados importada.

Tabela 18 – Dados de vendas importada

	trimestre	venda
1	1	254.4717
2	2	249.3390
3	3	208.5889
4	4	249.8564
5	5	279.0500
6	6	304.8245
7	7	307.3056
8	8	328.0858
9	9	330.5019
10	10	385.5567

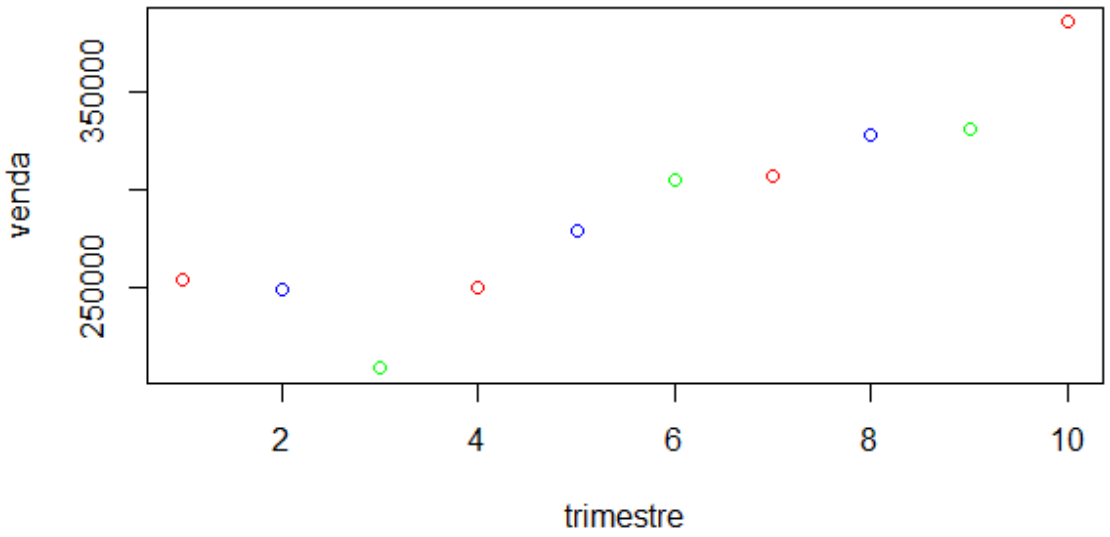
Foi criado um diagrama de dispersão para averiguar a existência da relação dos atributos “trimestre” e “venda”. A figura 29 apresenta a criação do diagrama.

Figura 29 – Diagrama de dispersão

```
#Diagrama de dispersão
plot(trimestre, venda, col=c('red', 'blue', 'green'))
```

O diagrama de dispersão foi criado, onde, a cor vermelha representa o ‘trimestre’ e a azul a ‘’venda”. A saída da instrução da figura 29 é o gráfico 2.

Gráfico 2 – Dispersão



Ao observar o gráfico 2 é possível notar a presença de uma correlação positiva entre a variável ‘trimestre’ X e ‘venda’ Y. De forma geral o passar dos trimestres acarreta aumento nas vendas, dessa forma, é possível saber que X influencia positivamente em Y e que o resultado da predição será positiva para o aumento das vendas.

Para calcular a força da relação entre as variáveis foi utilizada a função ‘cor’, a figura 30 apresenta sua aplicação bem como a saída da mesma.

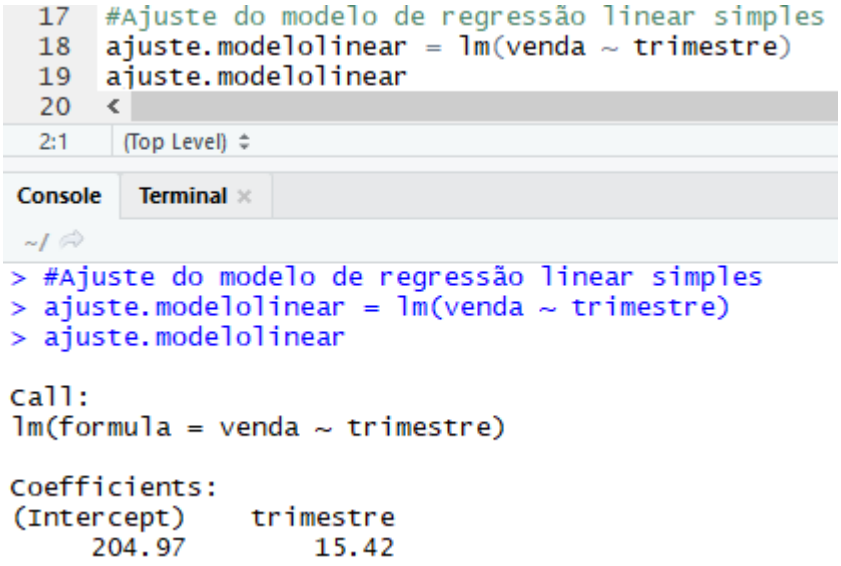
Figura 30 – Cálculo da correlação entre x e y

```
10  #Coeficiente de correlação
11  cor(trimestre,venda)
12
13
8:36  (Top Level)
Console  Terminal
~/
> #Coeficiente de correlação
> cor(trimestre,venda)
[1] 0.9035071
```

Conforme apresenta a Figura 30 a correlação entre X e Y é de 0.90, a relação fica sempre entre -1 e 1 onde 1 ou -1 é a correlação perfeita e 0,9 é

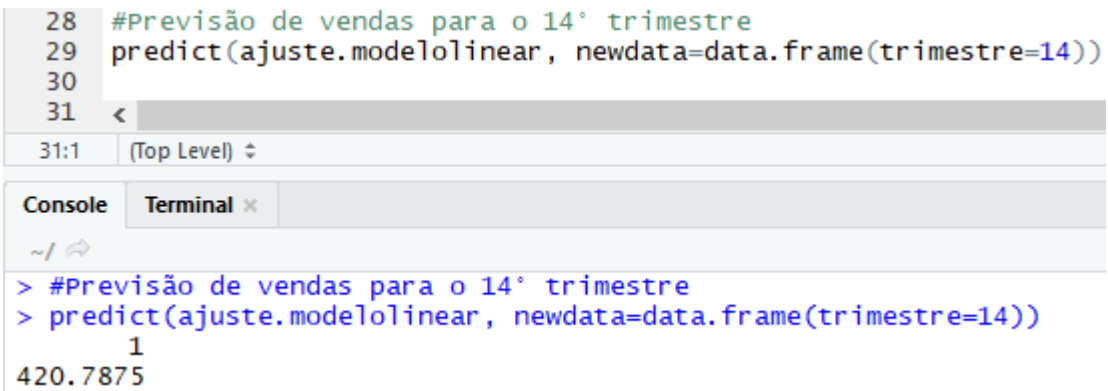
uma relação forte, logo após foi feito o ajuste da regressão linear simples com a função LM. A figura 31 apresenta o uso da função é o resultado como saída.

Figura 31 – Uso da função LM – linear model



A função LM retorna os coeficientes da equação $Y = \beta 0 + \beta 1X + \varepsilon$, onde Y é a variável alvo 'venda' e X 'trimestre' é a variável explicativa. A equação fica montada dessa forma: $venda = 204.97 + 15.42 * trimestre$. A seguir a figura 32 apresenta a previsão de vendas para o segundo trimestre de 2019 que corresponde ao trimestre 14.

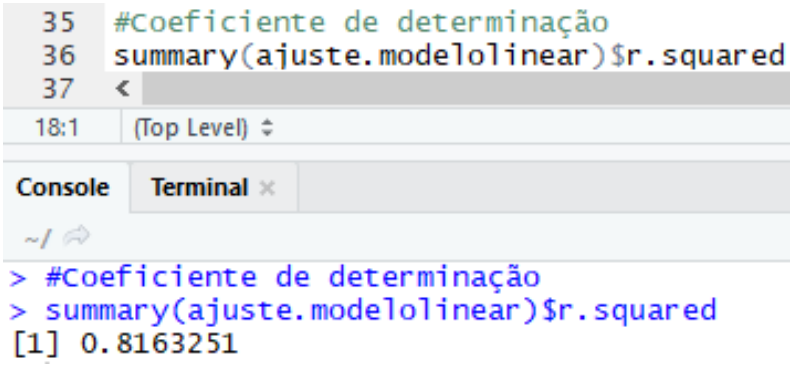
Figura 32 – Previsão de vendas



A saída mostra que no segundo trimestre de 2019 a farmácia poderá vender quatrocentos e vinte mil e setecentos e oitenta e sete reais e setenta e

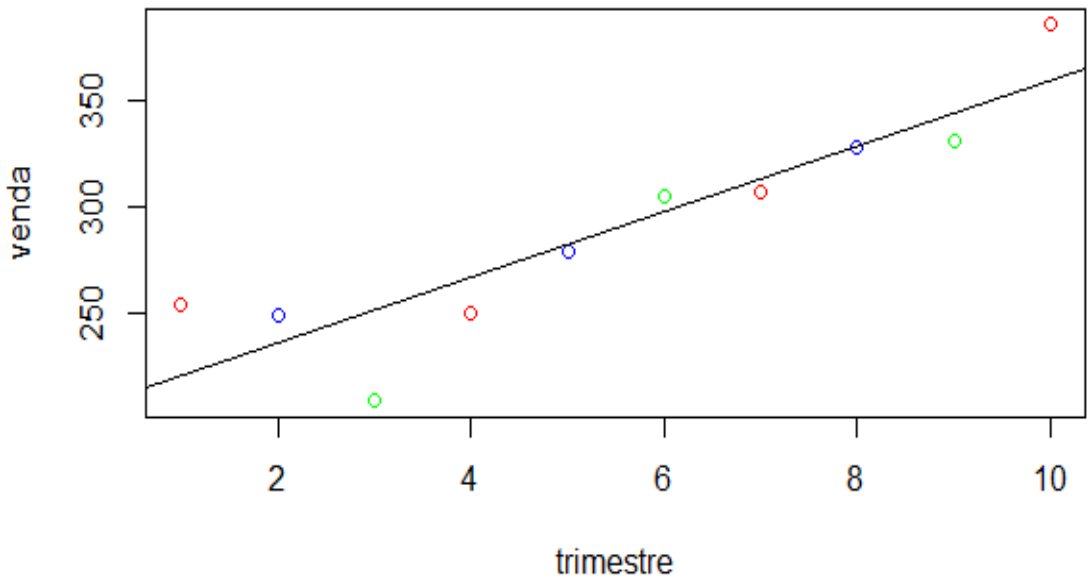
cinco centavos. A qualidade do modelo de dados criada pode ser avaliada conforme apresenta a figura 33.

Figura 33 – Avaliação do modelo de dados



O resultado apresentado na figura na figura 33 mostra que o modelo de dados pode explicar 81% dos dados analisados. O gráfico 3 apresenta o diagrama de dispersão com a reta devidamente ajustada.

Gráfico 3 – Dispersão com reta ajustada



A seguir será abordada a etapa de avaliação dos resultados.

4.4 AVALIAÇÃO DOS RESULTADOS

Nesta seção é abordada a fase de avaliação dos resultados, onde a participação do especialista do domínio foi requerida e empregada, no caso um farmacêutico em atividade na área de vendas analisou as regras de associação e fez algumas colocações, as quais serão expostas nas subseções a seguir.

4.4.1 ANÁLISE DAS REGRAS DE ASSOCIAÇÃO

A base de dados original possui 76.64% de transações com apenas um item, o que gerou regras fracas. Ao conversar com o especialista do domínio o mesmo sugeriu que fossem retiradas as transações com apenas um elemento e minerados novamente os dados. A seguir a tabela 19 apresenta o conjunto de regras geradas antes da eliminação das transações com apenas um item.

Tabela 19 – Regras de associação antes da eliminação das vendas de apenas um item

Se X	Então Y	Suporte % (x e y) vs (U)	Confiança (x e y) vs (x)	Comentário
STOMALIV ABAC AXI	HEPATILON C 1 FL AC 10 ML	0.001598327	0.289	Medicamento indicado Pelos farmacêuticos (venda conhecida)
EPATIVAN C1 FL AC	HEPATILON C 1 FL AC 10 ML	0.001837329	0.307	Medicamento para ressaca (venda conhecida)
EPATIVAN C1 FL AC	GASTROL 1 ENV E FERV TODOS	0.003241467	0.542	Venda conhecida, ressaca
BIO-C 1G CX 10 COMP EFV	CIMEGRIPE C 20 CAP	0.001508701	0.181	Para gripe (venda conhecida)
HEPATILON C 1 FLAC 10 ML	GASTROL 1 ENV E FERV TODOS	0.003784201	0.376	
TORSILAX C 10 CP	DORALGINA C 4 DRG	0.001185052	0.052	Para dor de cabeça e dor no corpo (venda conhecida)
EPATIVAN C1 FL AC, HEPATILON C 1 FLAC 10 ML	GASTROL 1 ENV E FERV TODOS	0.001100406	0.598	Ressaca (venda conhecida)

DORALGINA C 4 DRG	TORSILAX C 10 CP	0.001185052	0.056829035	Para dor de cabeça e dor no corpo (venda conhecida)

A seguir a tabela 20 apresenta o conjunto de regras geradas com as maiores taxas de confiança depois da eliminação das transações com apenas um item.

Tabela 20 – Regras de associação depois da eliminação das vendas de apenas um item

Se X	Então Y	Suporte % (x e y) vs (U)	Confiança (x e y) vs (x)	Comentário
COND.CAVALO FORTE HASKELL 300 ML	SH.CAVALO FORTE HASKELL 300 ML	0.001086470	0.8360656	Fortalecer o couro cabeludo (venda conhecida)
BROMIDRATO DE FENOTEROL 5 MGML GTS 20 ML	BROMETO DE IPRATROPIO 0,25 MG ML SOL INAL 20 M	0.002172940	0.5964912	Asma, reduzir as crises, produtos associados (venda conhecida)
SECNIDAZOL 450 MG SUSP 30 ML	ALBENDAZOL 40 MGML SUSP 10 ML	0.003110287	0.7724868	Para verme, venda associada (venda conhecida)
MEBENIX 400 MG C 1 CP MAST	SECNIMAX 1000 MG C 2 CP	0.001299504	0.7721519	Para verme, venda associada (venda conhecida)
ALBENDAZOL 400MG C3 CP	SECNIMAX 1000 MG C 2 CP	0.002449884	0.6804734	Para verme, venda associada (venda conhecida)
BENZOL 400MG C1	SECNIMAX 1000 MG C 2 CP	0.001214290	0.6333333	Para verme, venda associada (venda conhecida)
HEPATOVIT BOLDO 10ML FLACONETES	EPATIVAN C1 FLAC	0.001107774	0.6046512	Ressaca (venda conhecida)
HEPATOVIT ABACAXI 10ML C/60 FLACONETES, STOMALIV A BACAXI	HEPATILON C 1 FLAC 10 ML	0.001086470	0.6891892	Ressaca (venda conhecida)
HEPATOVIT BOLDO 10ML FLACONETES	GASTROL 1 ENV E FERV TODOS	0.001384717	0.7558140	Ressaca e queimação no estômago (venda conhecida)

EPATIVAN C1 FL AC	GASTROL 1 ENV E FERV TODOS	0.013868473	0.6112676	Ressaca e queimação no estômago (venda conhecida)

A análise da regra de regressão linear simples foi feita conforme documentado na seção 4.4.1, através das funções plot, cor e summary. Os resultados poderão ser analisados no decorrer do ano de 2019 e então será possível levantar hipóteses sobre o resultado esperado e o alcançado.

5 CONSIDERAÇÕES FINAIS

O *software* RStudio de fato possui muito material rico disponível na internet e exige poucos passos para alcançar largos resultados. Neste trabalho foi utilizado linha de comando para demonstrar os passos de importação da base de dados, instalação e importação de pacotes, porém o RStudio permite executar as mesmas atividades de forma gráfica, essa abordagem não foi utilizada pelo motivo do aumento fluxo de telas.

A etapa de pré-processamento foi a etapa mais delicada e trabalhosa de todo o processo KDD a mesma demanda mais tempo que as demais etapas, essa fase é crucial para se obter resultados que condizem com a realidade do domínio.

Os dados de venda-produto possuem 76,64% das compras com apenas um produto, consequentemente o suporte e a confiança das regras encontradas nesse modelo de dados foram pequenas. Como o processo KDD permite voltar a qualquer etapa, a pedido do especialista de domínio foi regressado até a fase de seleção de dados e executado mais uma redução da dimensionalidade, foram excluídas as vendas que possuíam apenas um item e submetido a mineração de dados por meio do Apriori.

Como esperado, as novas regras geradas possuem o suporte e a confiança muito mais elevado, porém, a veracidade das regras é passível de novos estudos para averiguar se as mesmas condizem com a realidade.

As regras de associação descobertas que possuem os maiores graus de confiança são todas conhecidas pelo especialista do domínio. Para outros

estudos futuros pode ser trazido para a discursão a quantidade de medicamentos para ressaca e contraceptivos que são vendidos e fazer um comparativo por região e período.

6 REFERÊNCIAS

AGRAWAL, Rakesh; IMIELINSKI, Tomasz; SWAMI, Arun. **Mining association rules between sets of items in large databases.** 1993. Disponível em: <https://goo.gl/kjs9nU>. Acesso em: 13 jun. 2018.

AMO, S. **Curso de data mining**: programa de mestrado em ciência da computação. Uberlândia: Universidade Federal de Uberlândia, 2003. Disponível em: < https://goo.gl/uH3Cw3 >. Acesso em: 29 mai. 2018.

CARVALHO, Luís Alfredo Vidal de. **Datamining:** A mineração de Dados no Marketing, Medicina, Economia, Engenharia e Administração. São Paulo: Érica, 2001. 234 p.

CORNELIUS JUNIOR, Romeu. **Uso da Mineração de Dados na Identificação de Alunos com Perfil de Evasão do Ensino Superior.** 2015. Disponível em: <https://goo.gl/kneMtQ>. Acesso em: 31 maio 2018.

CORREIA, Carlos Daniel Dias. **Data Mining e Data Quality em Dados da Saúde.** 2017. Disponível em: <https://goo.gl/6oEmzg>. Acesso em: 29 maio 2018.

COSTA, Evandro; BAKER, Ryan S.j.d; AMORIM, Lucas; MAGALHÃES, Jonathas; MARINHO, Tarsis. **Mineração de Dados Educacionais: Conceitos, Técnicas, Ferramentas e Aplicações.** 2012. Disponível em: <https://goo.gl/guLuAr>. Acesso em: 14 jun. 2018.

CRUZ, Armando Jorge Ribeiro da. **Data Mining via Redes Neuronais Artificiais e Máquinas de Vectores de Suporte.**2007. Disponível em: <https://goo.gl/rMBuxT>. Acesso em: 30 maio 2018.

DINIZ, Carlos Alberto; LOUZADA NETO, Francisco. **Data Mining**: uma introdução. São Paulo: ABE, 2000.

ELMASRI, Ramez; NAVATHE, Shamkant B. **Sistemas de banco de dados.** 4. ed. São Paulo: Pearson Education, 2005. 724 p. Vários tradutores.

FAYYAD, Usama; SHAPIRO, Gregory Piatetsky; SMYTH, Padhraic. **From Data Mining to Knowledge Discovery in Databases.** 1996. Disponível em: <https://goo.gl/Ferp9p>. Acesso em: 13 jun. 2018.

FERNANDES, Anita Maria da Rocha. **Inteligência Artificial:** Noções Gerais. Florianópolis: Visual Books, 2003. 160 p. Revisão Ortografia: Carina de Melo.

FIGUEIRA, Rafael Medeiros Andrade. "**Miner**: um Software de Inferência de Dependências Funcionais". Rio de Janeiro, 1998. Trabalho de conclusão de curso – Instituto de Matemática, Universidade Federal do Rio de Janeiro.

GROTH, Robert. **Data Mining a Hands-On for business Professional**. New Jersey: Editora Prentice Hall PTR, 1998.

HARRISON, Thomas H. **Intranet data warehouse**. São Paulo: Berkeley Brasil, 1998.

ITAKURA, Fernando Takashi. **Inteligência Artificial.** Guarapuava: Escola Regional de Informática, 2004. 230 p.

KOTSIANTIS, S. B., & PINTELAS, P. E. **Recent Advances in Clustering** : A Brief Survey. Methods, 2004. Disponível em: < https://goo.gl/fYq7rU > Acesso em: 07 jun. 2018.

MENDES, Luciana. **Data Mining: –**Estudo de Técnicas e Aplicações na Área Bancária. 2011. Disponível em: <https://goo.gl/GQCfZQ>. Acesso em: 11 abr. 2018.

PEREIRA, João José Rodrigues. **Modelos de Data Mining para multi-previsão:** aplicação à medicina intensiva**.** 2005. Disponível em: <https://goo.gl/sjRGaV>. Acesso em: 13 jun. 2018.

PRASS, Fernando Sarturi. **Estudo Comparativo entre Algoritmos de Análise de Agrupamentos em Data Mining.** 2004. Disponível em: <https://goo.gl/qajs3q>. Acesso em: 29 maio 2018.

RAUPP, Fabiano Maury; BEUREN, Ilse Maria. **Metodologia da Pesquisa Aplicável às Ciências Sociais.** 2006. Disponível em: <https://goo.gl/1TULVN>. Acesso em: 22 abr. 2018.

SCHMITZ, F. E. B. **Aplicação da técnica de Text Mining para comentários relacionados ao contexto do turismo**. 2015. Trabalho de Conclusão de Curso (Bacharelado em Sistemas de Informação). Centro Universitário Luterano de Palmas, Palmas, Tocantins, 2015. Disponível em: <https://goo.gl/BGCxUp>. Acesso em: 07 jun. 2018.

SILVA, Rui Flávio Gonçalves da. **Data mining na caracterização geo-espacial e previsão da incidência de pneumonia em Portugal.** 2016. Disponível em: <https://goo.gl/XRD2hB>. Acesso em: 07 jun. 2018.

SINGH, Harry, PhD. **Data Warehouse:** Conceitos, tecnologias, implementação e gerenciamento. São Paulo: Makron Books, 2001. 382 p. Tradução de: Monica Rosemberg.

TAN, Pang Ning; STEINBACH, Michael; KUMAR, Vipin. **Data Mining:** Mineração de dados. Rio de Janeiro: Ciência Moderna, 2009. 900 p. Tradução de: Acauan P. Fernandes.

WITTEN, Ian H.; FRANK, Eibe. **Data Mining**: Practical machine learning tools and techniques**.** 2005. Disponível em: <https://goo.gl/3eoDFG>. Acesso em: 13 jun. 2018.