UNIVERSIDADE FEDERAL DO ESPÍRITO SANTO

CENTRO DE CIÊNCIAS AGRÁRIAS

DEPARTAMENTO DE COMPUTAÇÃO

SAULO BRAGA MULIN

APLICAÇÃO DE TÉCNICAS DE MINERAÇÃO DE DADOS PARA ESTUDO DA EVASÃO DE ALUNOS NO CCA-UFES

ALEGRE – ES

2015

SAULO BRAGA MULIN

APLICAÇÃO DE TÉCNICAS DE MINERAÇÃO DE DADOS PARA ESTUDO DA EVASÃO DE ALUNOS NO CCA-UFES

Trabalho de conclusão de curso apresentado ao Departamento de Computação do Centro de Ciências Agrárias da Universidade Federal do Espírito Santo, como requisito parcial para obtenção do grau de Bacharel em Sistemas de Informação.
Orientador: Prof. D.Sc. Bruno Vilela Oliveira.

ALEGRE – ES
2015

SAULO BRAGA MULIN

APLICAÇÃO DE TÉCNICAS DE MINERAÇÃO DE DADOS PARA ESTUDO DA EVASÃO DE ALUNOS NO CCA-UFES

Trabalho de conclusão de curso apresentado ao Departamento de Computação do Centro de Ciências Agrárias da Universidade Federal do Espírito Santo, como requisito parcial para obtenção do grau de Bacharel em Sistemas de Informação.

Aprovado em 26 de junho de 2015.

BANCA EXAMINADORA

Prof. D.Sc. Bruno Vilela Oliveira
Universidade Federal do Espírito Santo
Orientador

Profª. M.Sc. Simone Dornelas Costa
Universidade Federal do Espírito Santo

Profª. M.Sc. Valéria Alves da Silva
Universidade Federal do Espírito Santo

AGRADECIMENTOS

A Deus, por ter me dado forças para chegar tão longe.

Aos meus pais por terem acreditado em mim desde o começo e por todo suporte prestado durante minha jornada.

Aos amigos que ganhei ao longo do curso e que estiveram do meu lado.

RESUMO

A evasão de estudantes é um problema recorrente em diversas universidades federais do Brasil. Tanto a sociedade quanto as universidades são prejudicadas, por não obterem retorno de seus investimentos que são comprometidos com a evasão. Nesse trabalho são aplicadas técnicas de mineração de dados na tentativa de se identificar e descrever fatores que possam estar associados à evasão de estudantes dos cursos de graduação de Ciência da Computação e de Sistemas de Informação do Centro de Ciências Agrárias da Universidade Federal do Espírito Santo. A identificação dos fatores associados é uma atividade importante no processo de estabelecimento de estratégias que visa reduzir a evasão de estudantes.

Palavras-chave: Evasão Escolar, Mineração de Dados, Regras de associação.

ABSTRACT

The evasion of students is a recurring problem in several federal universities in Brazil. This problem affects both society and universities that do not get adequate return on their investments due to high rates of evasion. The present work applies data mining techniques in an attempt to identify and describe the factors that can be associated with the evasion of students of Computer Science and Information Systems courses at the Center for Agricultural Sciences - Federal University of Espírito Santo. The identification of associated factors is an important activity in the process of establishing strategies that aim to reduce the evasion of students.

Keywords: School Failure, Data Mining, Association Rules.

LISTA DE FIGURAS

Figura 1 - Três Critérios Diferentes de Formação de Agrupamentos 18
Figura 2 - Cesta de Compras 21
Figura 3 - Regras de Associação 22
Figura 4 - Representação Numérica de cada artigo do supermercado 23
Figura 5 - Um BD de Transações de Clientes 24
Figura 6 - Exemplo de um conjunto de *Itemsets* 29
Figura 7 - Construção de um *Itemset* Candidato à Partir de dois Frequentes de Tamanho *K-1* 30
Figura 8 - Construção de um *Itemset* Podado à Partir do Conjunto de *Itemsets* Candidato 31
Figura 9 - BD a ser Utilizado para Varredura.....32
Figura 10 - *Itemsets* Podados com seus Respectivos Incrementos 33
Figura 11 - Evasão de alunos por ano 35
Figura 12 - Estado de Origem dos Estudantes (CC e SI) 36
Figura 13 - Forma de Evasão dos Estudantes (CC e SI) 37
Figura 14 - Distribuição de Evasão - Cotistas x Não Cotistas (CC e SI) 38
Figura 15 - Distribuição da Evasão por Sexo (CC e SI) 39
Figura 16 - Regras de CC para o Experimento de Confiança.....42
Figura 17 - Regras de CC para o Experimento de Lift 42
Figura 18 - Regras de SI para o Experimento de Confiança 43
Figura 19 - Regras de SI para o Experimento de Lift 44

LISTA DE QUADROS

Quadro 1 - Exemplo de Técnicas Utilizadas nas Tarefas 17

LISTA DE SIGLAS

BD - Banco de Dados

CC - Ciência da Computação

CCA - Centro de Ciências Agrárias

DW - *Data Warehouse*

ES - Espírito Santo

GPL - GNU Public License

INEP - Instituto Nacional de Estudos e Pesquisas

MD - Mineração de Dados

RA - Regras de Associação

REUNI - Reestruturação e Expansão das Universidades Federais

SI - Sistemas de Informação

TID - Identificador Único de Transação

UFES - Universidade Federal do Espírito Santo

WEKA - *Waikato Environment for Knowledge Analysis*

Sumário

1. INTRODUÇÃO 12
1.1 O PROBLEMA E SUA IMPORTÂNCIA 12
1.2 OBJETIVOS 13
1.2.1 Objetivo Geral 13
1.2.2 Objetivos Específicos 13
2. REVISÃO DA LITERATURA 14
2.1 DEFINIÇÕES PARA MINERAÇÃO DE DADOS 14
2.2 APLICAÇÕES 15
2.3 MINERAÇÃO DE DADOS EM AMBIENTE ACADÊMICO 16
2.4 TAREFAS DE MINERAÇÃO DE DADOS 17
2.4.1 Análise de Agrupamentos 17
2.4.2 Classificação 18
2.4.3 Aproximação de Funções ou Regressão (Predições Numéricas) 19
2.4.4 Análise de Associações 20
2.5 REGRAS DE ASSOCIAÇÃO 21
2.5.1 Base de Transações 23
2.5.2 *Itemsets* Frequentes e Regras de Associação 24
2.5.3 Métricas de Interesse 25
2.5.3.1 Confiança 26
2.5.3.2 Suporte 27
2.5.3.3 *Lift* (Medida de Correlação) 27
2.6 ALGORITMO APRIORI 28
2.6.1 Fase de Geração do Conjunto de *Itemsets* Candidatos 29
2.6.2 Fase de poda conjunto de *itemsets* candidatos 30
2.6.3 Fase do Cálculo do Suporte 31
3. FERRAMENTA WEKA 34

4. DESCRIÇÃO DA BASE DE DADOS 34
5. METODOLOGIA 39
5.1 CONFIGURAÇÃO DOS EXPERIMENTOS.....39
6. RESULTADOS – REGRAS DE ASSOCIAÇÃO EXTRAÍDAS 42
7. CONCLUSÃO 45
7.1 DIFICULDADES 46
8. REFERÊNCIAS.....47
9. APÊNDICES 51

1. INTRODUÇÃO

A evasão escolar é o desligamento do aluno, por sua própria responsabilidade, da instituição de ensino a qual está matriculado (SOUZA, 2008 apud MILANI; CAMARGO, 2010, 2010).

O Governo Federal, por meio do Programa de Apoio a Planos de Reestruturação e Expansão das Universidades Federais – REUNI (Governo Federal, 2007 apud MANHÃES et al., 2011) relata que “os índices de evasão de estudantes nos cursos de graduação atingem, em alguns casos, níveis alarmantes”. No entanto, a meta do governo é a elevação gradual da taxa de conclusão média dos cursos de graduação presenciais para noventa por cento (MANHÃES et al., 2011).

O sistema educacional brasileiro possui um grande número de estudantes que iniciam um curso universitário, mas não conseguem obter êxito em cumprir as exigências curriculares e se graduar (MANHÃES et al., 2012). A evasão dos alunos que não completam o curso de graduação se configura como um dos grandes problemas que ocorrem em instituições públicas e particulares (INEP, 2009 apud MANHÃES et al., 2012).

Diante do alto índice de evasão de alunos na graduação, este trabalho se baseia na extração e análise de Regras de Associação (RA) a partir de uma base de dados acadêmica contendo dados de alunos que evadiram nos cursos de Ciência da Computação (CC) e Sistemas de Informação (SI) do Centro de Ciências Agrárias - Universidade Federal do Espírito Santo (CCA-UFES). Os dados fornecidos, foram utilizados como entrada para as estratégias de mineração de dados (MD) na tentativa de identificar fatores que possam estar relacionados à evasão de alunos, encontrando assim, indicadores de probabilidade de evasão.

1.1 O PROBLEMA E SUA IMPORTÂNCIA

A questão da evasão discente é um dos problemas mais relevantes que atinge as instituições de ensino. Elevados índices de evasão fazem com que se levantem questões sobre as causas desse fenômeno e as estratégias que

podem ser adotadas para minimizar o desperdício de recursos públicos e o tempo de permanência dos discentes na universidade (Stoffel e Ziza, 2014). No Centro de Ciências Agrárias da Universidade Federal do Espírito Santo (CCA-UFES), de acordo com informações obtidas com as coordenações dos cursos de Ciência da Computação (CC) e de Sistemas de Informação (SI), a evasão discente atingiu no período que vai do segundo semestre letivo de 2009 até o primeiro semestre letivo do ano de 2013, o total percentual de 29,38% dos alunos do curso de SI e de 24,41% dos alunos do curso de CC. Esses percentuais podem ser considerados significativos para os dois cursos do CCA-UFES. Considerando os prejuízos que incidem no retorno sobre o investimento de recursos que ficam comprometidos devido a altos índices de evasão, torna-se importante identificar os principais fatores que eventualmente promovem esse cenário.

1.2 OBJETIVOS

O presente trabalho tem por objetivo aplicar estratégias de MD à uma base de dados com informações de alunos dos cursos de SI e CC do CCA-UFES buscando identificar e compreender o contexto e os principais fatores que contribuem para evasão de estudantes desses cursos.

1.2.1 Objetivo Geral

Investigar os fatores que podem estar associados à evasão de alunos no CCA-UFES.

1.2.2 Objetivos Específicos

Os objetivos específicos deste trabalho são:

a) Gerar uma base de dados com informações dos alunos dos cursos de CC e SI do CCA-UFES, incluindo informações sobre evasão.

b) Extrair padrões relacionados com associações (regras de associação) entre as informações da base de dados por meio do algoritmo Apriori que

é disponibilizado com a ferramenta WEKA *(Waikato Environment for Knowledge Analysis*)[1].

c) Analisar e interpretar as regras de associação extraídas na tentativa de se descrever os possíveis fatores que contribuem para a evasão dos alunos.

2. REVISÃO DA LITERATURA

2.1 DEFINIÇÕES PARA MINERAÇÃO DE DADOS

A Mineração de Dados pode ser definida como um conjunto de técnicas automáticas de exploração de grandes massas de dados de forma a descobrir novos padrões e relações que, devido ao volume de dados, não seriam facilmente descobertas de forma manual pelo ser humano (AMORIM, 2006).

Segundo Damasceno (1996), a noção de encontrar padrões úteis em grandes volumes de dados pode ser conhecida por diversos termos como mineração de dados, extração de conhecimento, descoberta de informações, arqueologia de dados e processamento de padrões de dados.

> Hand et al. (2001) apresentam uma definição para a MD a partir de uma perspectiva estatística: “Mineração de Dados é a análise de grandes conjuntos de dados a fim de encontrar relacionamentos inesperados e de resumir os dados de uma forma que eles sejam tanto úteis quanto compreensíveis ao dono dos dados”. Já Cabena et al. (1998) definem o tema de MD de acordo com uma visão computacional: "Mineração de Dados é um campo interdisciplinar que combina técnicas como aprendizagem de máquina, reconhecimento de padrões, estatística, Banco de Dados (BD) e visualização, para conseguir extrair informações de grandes bases de dados".
>
> Adotando uma perspectiva de aprendizagem de máquina, Fayyad et al. (1996) argumentam que "Mineração de Dados é um

[1] Weka (Waikato Environment for Knowledge Analysis) é um ambiente de software escrito em Java, desenvolvido na Universidade de Waikato, Nova Zelândia. Trata-se de um software licenciado sobre de acordo com a General Public License o que permite, portanto, estudar e alterar seu código fonte. O Weka incorpora algoritmos de diferentes abordagens/paradigmas de inteligência artificial voltados ao estudo de aprendizagem de máquina. A ferramenta pode ser obtida em < http://www.cs.waikato.ac.nz/ml/weka>.

> passo no processo de Descoberta de Conhecimento que consiste na realização da análise dos dados e na aplicação de algoritmos de descoberta que, sob certas limitações computacionais, produzem um conjunto de padrões de certos dados.

2.2 APLICAÇÕES

A MD tem sido aplicada a diferentes tipos de massas de dados oriundas de áreas de conhecimentos como Medicina, Economia, Astronomia, Geologia, etc. São muito comuns os exemplos de aplicações na área de negócios. Amorim (2006) destaca que a MD pode ser aplicada aos dados de uma empresa, buscando resultados que permitam melhorar o direcionamento de suas vendas e auxiliando na definição de estratégias de marketing.

Como exemplo, em uma determinada rede de lojas americana, escolheu-se 5 (cinco) categorias de produtos para analisar e melhorar suas vendas: vestuário esportivo, aparelhos de ginástica, decoração, mobília e CD. A ideia da empresa era fazer um maior direcionamento à clientela quando fosse enviar catálogos e ofertas. Uma empresa de Data Warehouse (DW) unificou os 3 (três) BDs dessa empresa para análise. O resultado foi clientes com 42 anos, ganhando US$ 35.000,00 por ano, sendo em maioria (58%) do sexo feminino e 50% casados contra 35% solteiros, possuindo 40% deles casa própria (AMORIM, 2006).

Outro exemplo de mercado é de uma grande revendedora de automóveis que observando sua queda de vendas com clientes insatisfeitos por não encontrarem o modelo desejado em estoque, resolveu desenvolver um sistema de previsão de vendas. A empresa possuía um BD de vendas de carros nos últimos 5 (cinco) anos e desejava um sistema de previsão capaz de avaliar as vendas 15 (quinze) dias a frente pois este era o tempo necessário para encomenda e transporte de novos itens (CARVALHO, 2005 apud AMORIM, 2006). O resultado fez com que o sistema passa-se a prever as vendas dos próximos 15 (quinze) dias com erro médio de 10%, mantendo-se dentro da margem desejável. Auxiliado por essa previsão a revendedora passou a ter mais tempo para a encomenda e transporte dos veículos.

2.3 MINERAÇÃO DE DADOS EM AMBIENTE ACADÊMICO

A previsão do desempenho acadêmico é um tema já estudado por diversos pesquisadores. Estudos mais antigos utilizam métodos estatísticos para compreender os aspectos envolvidos (MOORE et al., 1995 apud MANHÃES et al., 2011).

A utilização de técnicas de MD sobre dados educacionais é relativamente recente (BAKER et al., 2009, apud MANHÃES et al., 2011; DEKKER et al., 2009 apud MANHÃES et al., 2011). No entanto, a maioria dos trabalhos correlatos restringem-se a identificar resultados em pequenos contextos relativos apenas a uma disciplina de curso não presencial (MANHÃES et al., 2011).

Um exemplo da aplicação de MD em âmbito acadêmico é o estudo de Minaei-Bidgoli (2004), no qual foi utilizada RA para extrair padrões de informações de BDs gerados a partir de sistemas educacionais online. Com base na disciplina de Física, os autores demonstraram que com a aplicação de um conjunto de regras foi possível identificar quais os atributos que caracterizam o padrão de desempenho dos grupos de estudantes.

Hall (2009) realizou um trabalho na Universidade de Eindhoven, no qual foram identificados alunos com risco de evasão. Foram utilizados diversos algoritmos da ferramenta de MD WEKA, com intuito de descobrir qual era o algoritmo mais adequado ao problema. O experimento analisou diversos dados dos alunos e obteve entre 75% a 80% de acurácia, o classificador baseado em árvore de decisão foi considerado mais adequado (DEKKER et al., 2009).

Percebe-se na literatura que o emprego da MD em bases de dados educacionais para a previsão da situação acadêmica é um campo que necessita da transformação dos dados, identificação dos atributos mais relevantes e escolha do algoritmo mais adequado (CASTRO et al. 2007, apud MANHÃES et al., 2011; BAKER, 2009 apud MANHÃES et al., 2011; DEKKER et al., 2009 apud MANHÃES et al., 2011).

2.4 TAREFAS DE MINERAÇÃO DE DADOS

As estratégias de MD têm por objetivo a extração de padrões a partir de grandes repositórios de dados. Os padrões são interpretados resultando em algum tipo de conhecimento que pode ser utilizado para processos de tomada de decisões. Diferentes tipos de conhecimentos podem ser extraídos e, para cada tipo, é necessário que se aplique uma função, ou tarefa, de MD. Cada tarefa de MD possui um conjunto de técnicas associadas, que representam os algoritmos que podem ser empregados para a sua execução (GONÇALVES, 2000).

No Quadro 1 são apresentados exemplos de técnicas que podem ser utilizadas para a realização de determinada tarefa de MD.

Quadro 1 – Exemplo de Técnicas Utilizadas nas Tarefas

Tarefa	Técnicas
Regras de Associação	Apriori, FP-Growth, DCI, ECLAT, Closet.
Classificação	Naïve Bayes, Árvores de Decisão (ID3, C4.5, CART, CHAID), Predições Numéricas, Redes Neurais, k-Nearest Neighbor, Support Vector Machines.
Análise de Agrupamentos	k-Médias, k-Medoides, Métodos Hierárquicos.

Fonte: Gonçalves (2000)

Mediante o conceito obtido através da tarefa e da técnica, pode-se dizer que a MD é comumente classificada pela sua capacidade em realizar determinadas tarefas (LAROSE, 2005). A próxima seção apresenta uma breve descrição das principais tarefas de MD.

2.4.1 Análise de Agrupamentos

Essa tarefa tem como objetivo gerar e analisar agrupamentos de dados a partir de um conjunto de registros. A ideia normalmente é utilizar algum critério de similaridade como base para se gerar agrupamentos contendo registros altamente similares (homogêneos) em relação aos demais registros do mesmo grupo e bastante distintos dos registros de outros grupos (heterogêneos) (Camilo e Silva, 2009).

Como exemplo de aplicação dessa tarefa, pode-se utilizar dados do Censo para formar grupos de domicílios, utilizando os atributos escolaridade, profissão, faixa etária, sexo, número de filhos. Observa-se que não existem classes previamente estabelecidas e poderemos ter em um mesmo grupo domicílios de estados geograficamente opostos, porém, semelhantes nestes atributos. Na Figura 1 são exemplificados três possíveis agrupamentos formados a partir de um conjunto de dados (CORTÊS et al., 2002).

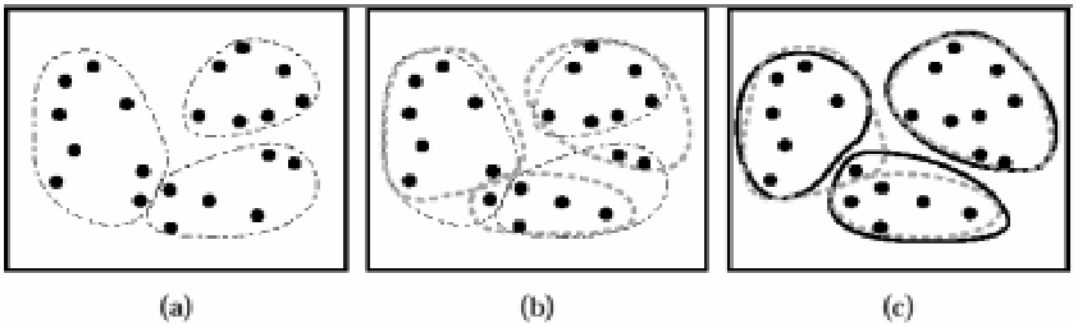

Figura 1 - Três Critérios Diferentes de Formação de Agrupamentos

Fonte: Cortês et al. (2002)

Cortês et al. (2002) enfatizam que cabe ao analisador avaliar se as classes significam algo útil. Por exemplo, agrupar sintomas pode gerar classes que não representem nenhuma doença explicitamente, uma vez que doenças diferentes podem possuir os mesmos sintomas.

2.4.2 Classificação

Segundo Amorim (2006) a classificação é uma das tarefas de mineração de dados mais utilizadas. Pode ser aplicada no auxílio da compreensão do ambiente em que se vive, consistindo do exame de certas características nos dados e da atribuição de uma classe previamente definida. O ser humano está sempre classificando o que percebe a sua volta, criando classes de relações humanas diferentes (colegas de trabalho, amigos, familiares, etc.) e dando a cada classe uma forma diferente de tratamento (AMORIM, 2006).

Pode-se, por exemplo, criar um modelo que classifica os clientes de um banco como especiais ou de risco, um laboratório pode usar sua base histórica de voluntários e verificar em quais indivíduos uma nova droga pode ser melhor ministrada. Em ambos os cenários um modelo é criado para classificar a qual categoria um certo registro pertence: especial ou de risco, voluntários A, B ou C (Camilo e Silva, 2009).

Amorim (2006) ressalva que como no mundo físico nada é exatamente igual, por mais semelhante que pareça, para se criar classes é preciso permitir que detalhes sejam desprezados e somente as características principais sejam observadas. A tarefa de classificar geralmente exige a comparação de um objeto ou dado com outros dados ou objetos que supostamente pertençam a classes previamente definidas. Para comparar dados ou objetos utiliza-se uma métrica ou forma de medida de diferenças entre eles.

2.4.3 Aproximação de Funções ou Regressão (Predições Numéricas)

Os métodos de predição visam descobrir um possível valor futuro de uma variável. As predições numéricas visam prever valores para variáveis contínuas, sendo criada uma função com o objetivo de tentar descrever o comportamento dos dados. Para a predição de variáveis discretas, as técnicas de classificação já apresentadas podem ser aplicadas. Os métodos mais conhecidos para predição numérica são as regressões, desenvolvidas por Sir Frances Galton (1822-1911) (Han e Kamber, 2006).

Alguns autores tratam as predições numéricas e as regressões como sinônimos, porém, alguns métodos de classificação também fazem predições. As técnicas de regressão modelam o relacionamento de variáveis independentes (chamadas preditoras) com uma variável dependente (chamada resposta). As variáveis preditoras são os atributos dos registros, e a resposta é o que se quer predizer através de uma função de acordo com atributos dos registros (Camilo e Silva, 2009).

Como exemplo de aplicação, a tarefa de regressão poderia ser utilizada para prever o volume de chuvas para um determinado mês a partir da humidade do ar, da velocidade do vento e da época do ano.

2.4.4 Análise de Associações

A análise de associações, que pode ser denominada por expressões como grupos de afinidade ou análise de transações de compras, quando voltada para a área de marketing, é a tarefa de MD baseada na busca por relações recorrentes em uma determinada base de dados. O objetivo é encontrar associações interessantes entre conjuntos de itens em bases de dados transacionais e relacionais por meio da mineração de padrões frequentes, que consistem de padrões que ocorrem frequentemente em uma base de dados. Por exemplo, um conjunto de itens como leite e pão, que aparecem frequentemente juntos em uma base de dados de transações, é um conjunto de itens frequente (Han e Kamber, 2006).

A identificação de padrões frequentes tem um papel essencial na mineração de associações e de outros relacionamentos interessantes entre dados. A descoberta de associações pode permitir a geração de conhecimentos importantes para os processos decisórios de empresas e também servir de apoio para outras tarefas de mineração de dados como a classificação e a análise de agrupamentos (Han e Kamber, 2006).

Uma aplicação típica da tarefa de análise de associações é o estudo de preferências e/ou afinidades visando a tomada de decisões sobre organização de mercadorias em prateleiras e sobre a formação de pacotes para consumidores. Esse tipo de aplicação é conhecido como análise de cesta de compras e é realizada em grandes bases de dados transacionais. A “cesta” do supermercado refere-se àquilo que o consumidor compra quando vai ao supermercado. Mediante a análise de transações de compra seria possível a empresa planejar melhor os cartazes de promoções dos produtos, fazer combinações de produtos em uma prateleira, dispondo-os de forma mais acessível aos clientes, buscando aumentar o número de vendas (SILVA, 2004). Associações também podem ser usadas para a busca de entendimento sobre tendências ou ocorrência de eventos como, por exemplo, a compra de carne em conjunto com cerveja pode sugerir que o cliente esteja planejando um churrasco.

2.5 REGRAS DE ASSOCIAÇÃO

Para discutir o que são regras de associação, será utilizado o exemplo motivador da aplicação de mineração de padrões frequentes no problema de análise de cesta de compras (Figura 2).

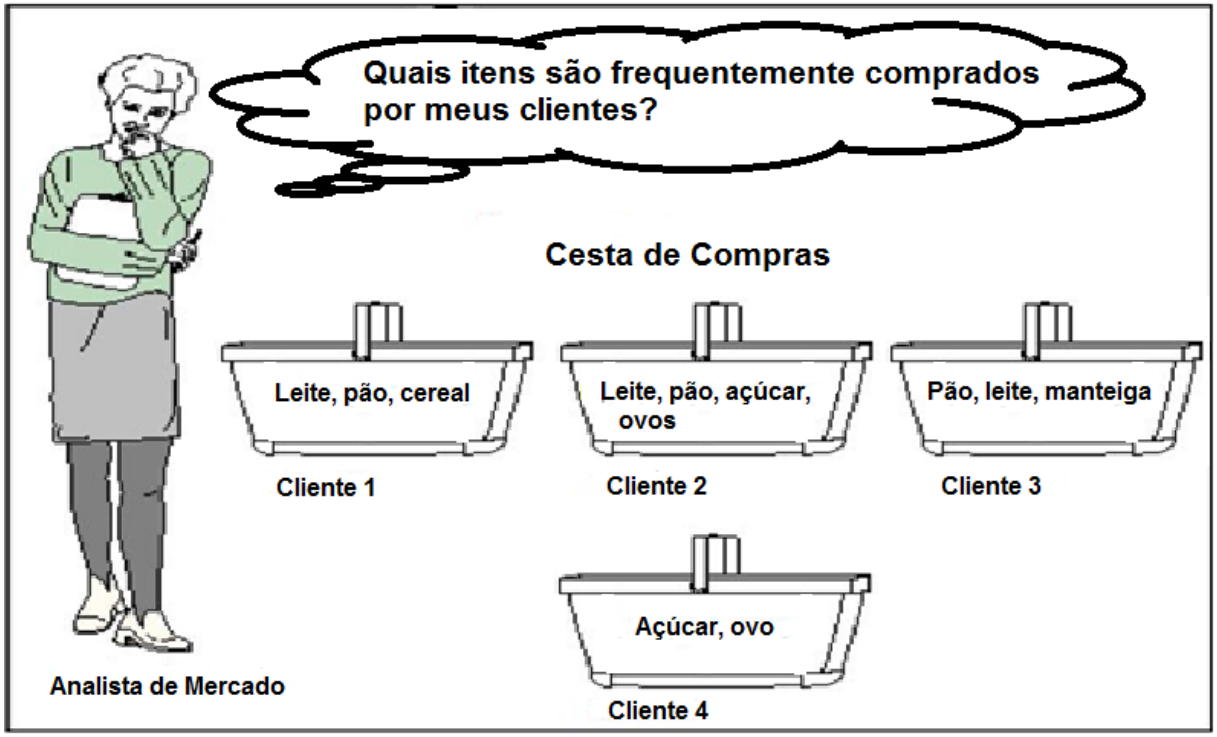

Figura 2 - Cesta de Compras

Fonte: Araújo (2009)

Nesse contexto, gerentes de lojas frequentemente precisam aprender mais sobre os hábitos de compras de seus clientes. Normalmente querem saber quais grupos ou conjuntos de itens serão provavelmente comprados por clientes em uma visita à loja. A análise de cesta de compras pode ser conduzida usando-se os dados de transações dos clientes na loja e os resultados podem ser utilizados para planejar estratégias de propaganda, promoções, leiautes de prateleiras ou para o projeto de um novo catálogo de mercadorias.

Considerando as mercadorias como itens da base de transações e o conjunto universo de itens como o conjunto de itens disponíveis na loja, então cada item tem uma variável booleana representando sua presença ou ocorrência em uma determinada transação. Cada cesta de compras pode ser representada como um vetor booleano de valores associados a essas variáveis (os itens). Os vetores booleanos podem ser analisados para se descobrir padrões de compras que refletem itens que estão frequentemente associados ou comprados juntos. Esses padrões podem ser representados na forma de **regras de associação**. Por exemplo, a informação de que clientes que compram computadores também tendem a comprar ao mesmo tempo software antivírus é representada na regra de associação abaixo (Han e Kamber, 2006):

$$computador \Rightarrow software_antivírus$$

É importante notar que não existe a garantia de que software antivírus sempre será comprado em conjunto com computadores (Camilo e Silva, 2009).

Na Figura 3 são apresentados exemplos de regras que poderiam ser extraídas a partir de diferentes bases de dados.

Regra 1: SE *idade* = *jovem* AND *estudante* = *não* ENTÃO *compra computadores* = *não*
Regra 2: SE *idade* = *jovem* AND *estudante* = *sim* ENTÃO *compra computadores* = *sim*
Regra 3: SE *idade* = *média* ENTÃO *compra computadores* = *sim*
Regra 4: SE *idade* = *adulto* AND *avaliação de crédito* = *excelente* ENTÃO *compra computadores* = *sim*
Regra 5: SE *idade* = *adulto* AND *avaliuação de crédito* = *ruim* ENTÃO *compra computadores* = *não*

Figura 3 - Regras de Associação

Fonte: Camilo e Silva (2009)

De modo geral, define-se então que uma RA é uma expressão da forma **A** → **B**, onde **A** e **B** são conjuntos de itens e **A** ∩ **B** = Ø (ESCOVAR, 2004). A recíproca dessa relação não necessariamente é verdadeira, isto é, não se pode dizer que a existência da associação descrita pela regra **A** → **B** sempre implicará na existência de uma associação descrita pela regra **B** → **A** (ARAÚJO, 2009).

O conjunto de itens que aparecem à esquerda da seta (representado por **A**) é conhecido por **antecedente** da regra e o conjunto de itens que aparece à direita da seta (representado por **B**) é o **consequente** da regra (ESCOVAR, 2004). Assim, uma regra de associação tem a forma:

$$\textit{Antecedente} \rightarrow \textit{Consequente}$$

2.5.1 Base de Transações

Uma base de dados transacional pode ser compreendida a partir do exemplo apresentado na Figura 4, na qual cada artigo (item) é associado a um número de identificação.

Artigo (item)	número que o representa
Pão	1
Leite	2
Açúcar	3
Papel Higiênico	4
Manteiga	5
Fralda	6
Cerveja	7
Refrigerante	8
Iogurte	9
Suco	10

Figura 4 - Representação Numérica de cada artigo do supermercado

Fonte: Amo (2012)

Na Figura 5 ilustra-se os registros de compra de cada cliente em uma loja que possui os artigos listados na Figura 4. Cada compra (transação) realizada pelo cliente é registrada no BD. O conjunto desses registros representam todas as transações efetuadas pelos clientes. Cada transação possui um identificador único de transação (TID) e seus itens associados. Cada transação é composta por um conjunto de itens, os artigos comprados naquela transação, comumente referidos por *itemset* na literatura de MD (Han e Kamber, 2006).

TID	Itens comprados
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}

Figura 5 - Um BD de Transações de Clientes

Fonte: Amo (2012)

Tomando, por exemplo, a transação identificada pelo TID 101 tem-se o *itemset* {1,3,5}. Seja $I = \{1, 2, 3, 4, 5, 6, 7, 8, 9, 10\}$ o conjunto de todos os possíveis itens (artigos do supermercado). A partir desse conjunto pode-se obter 2^{10} (1024) *itemsets* possíveis (ESCOVAR, 2004), ou seja, 1024 possíveis combinações de produtos. Um *itemset* com *k* elementos é chamado de *k-itemset* (AMO, 2012).

2.5.2 *Itemsets* Frequentes e Regras de Associação

Seja $I = \{I_1, I_2, \dots, I_m\}$ um conjunto de itens. Seja *D* uma base de dados contendo transações na qual cada transação *T* é um conjunto de itens tal que $T \subseteq I$. Seja *A* um conjunto de itens. Uma transação *T* é dita conter *A* se e somente se $A \subseteq T$. Uma regra de associação é uma implicação da forma $A \Rightarrow B$, na qual $A \subset I$, $B \subset I$ e $A \cap B = \emptyset$. A regra $A \Rightarrow B$ está presente no conjunto de transações *D* com **suporte** *s*, sendo que *s* é a percentagem de transações em *D* que contêm $A \cup B$ (isto é, a união dos conjuntos *A* e *B*, ou tanto *A* quanto *B*). Essa é tomada como a probabilidade $P(A \cup B)$ (essa notação não deve ser confundida com $P(A\ ou\ B)$, que indica a probabilidade de que uma transação contenha ou *A* ou *B*). A regra $A \Rightarrow B$ tem **confiança** *c* no conjunto de transações *D*, sendo *c* a percentagem de transações em *D* contendo *A* que também contêm *B*. Essa é tomada como a probabilidade condicional $P(A|B)$. Dessa forma, tem-se que:

$$Sup(A \Rightarrow B) = P(A \cup B) \quad (1.1)$$

$$Conf(A \Rightarrow B) = P(A|B) (1.2)$$

Em que **Sup** é o suporte da regra $A \Rightarrow B$ e **Conf** é a confiança da mesma.

Regras que satisfazem limiares mínimos tanto de suporte (SupMin) quanto de confiança (ConfMin) são chamadas de regras **fortes**.

A **frequência de ocorrência** de um *itemset* é o número de transações que contêm o *itemset*. Essa contagem também é conhecida como **frequência**, **contagem de suporte** ou **contagem** do *itemset*. O suporte definido na equação 1.1 é por vezes chamado de **suporte relativo**, enquanto que a frequência de ocorrência é chamada de **suporte absoluto**. Se o suporte relativo de um *itemset* I satisfaz um **limiar mínimo de suporte** predefinido, então I é um ***itemset* frequente**. O conjunto de *itemsets* **frequentes** é comumente denotado por $\boldsymbol{L_k}$.

Tais conceitos e métricas definem um conjunto de critérios que podem ser usados para se extrair regras de associação com maior potencial de utilidade. A seguir, são descritas três métricas de interesse tipicamente usadas para se avaliar o quão interessantes são as regras de associação extraídas.

2.5.3 Métricas de Interesse

As métricas de interesse (ou medidas de interessabilidade) servem para avaliar a qualidade das regras de associação. Essas medidas são usadas na tarefa de mineração de regras de associação com o objetivo de avaliar o quanto uma regra é interessante (o que reflete na utilidade da regra). Segundo Tan et al. (2004), a tarefa central das regras de associação é a descoberta de pares de itens que ocorrem conjuntamente com determinada frequência em um BD transacional, enquanto que o uso das métricas de interesse tem a ver com a identificação dos grupos de variáveis (itens) que estão altamente correlacionadas entre si ou à uma variável alvo específica (SCHONHORST, 2010).

Entre as métricas de interesse que podem ser utilizadas para a geração e avaliação de regras podem ser citadas as métricas **suporte**, **confiança** e ***lift*** (SILVA, 2004).

O suporte e a confiança são as medidas mais utilizadas em regras de associação, sendo que o suporte representa a frequência dos padrões e a confiança a força da implicação (Zhang e Zhang, 2002). O modelo Suporte/Confiança consiste em encontrar todas as regras que possuem confiança maior ou igual a uma confiança mínima (**ConfMin**) e um valor de suporte maior ou igual a um mínimo, ambos estabelecidos (**SupMin**) pelo analista de dados (GONÇALVES, 2005).

2.5.3.1 Confiança

Para toda associação **A** → **B**, determina-se um grau de confiança, denotado por **Conf(A** → **B)**. A confiança de uma regra significa a porcentagem das transações que suportam **B** dentre todas as transações que suportam **A**. A confiança de uma regra é dada por:

$$\textbf{Conf(A} \rightarrow \textbf{B)} = \frac{\text{Número de transações que suportam } (A \cup B)}{\text{Número de transações que suportam } (A)}$$

De acordo com a Figura 4 pode-se considerar **A** = item 2 (leite) e **B** = item 9 (iogurte), e de acordo com os 6 *itemsets* da Figura 5 tem-se que o número de transações que suportam **(A U B)** é 3, representado pelas TIDs 103, 104 e 106, nas quais **A** e **B** estão presentes na mesma cesta. O número de transações que suportam **(A)** é 4, representado pelas TIDs 102, 103, 104 e 106, nas quais **A** encontra-se presente. Dessa forma tem-se que:

$$\text{Conf(\{leite\} } \rightarrow \text{ \{iogurte\})} = \frac{3}{4} = 0{,}75 = 75\%$$

O que significa que 75% dos clientes que compraram leite também compram iogurte. Sendo assim, a confiança de uma regra indica o grau de certeza com o qual podemos esperar de encontrar em uma transação o item **B** dado que essa transação possui o item **A**.

Nem sempre altos valores de confiança representam boas regras de associação. Algumas regras com alta confiança podem relacionar itens que não necessariamente possuem uma relação causal, isto é, não é possível sustentar que a regra descreva uma associação real entre o termo consequente e o termo antecedente da regra.

2.5.3.2 Suporte

Para toda associação **A** → **B** calcula-se o valor de suporte, denotado por **Sup(A → B)**. O suporte de uma regra representa a porcentagem das transações que suportam **A** e **B** sobre o número total de transações **(N)** e é dado por:

$$\textbf{Sup(A} \rightarrow \textbf{B)} = \frac{\text{Número de transações satisfazendo A e B (A} \cup \text{B)}}{\text{Número total de transações (N)}}$$

De acordo com a Figura 4 pode-se considerar **A** = item 2 (leite) e **B** = item 9 (iogurte), e de acordo com os 6 *itemsets* da Figura 5 tem-se que o número de transações que suportam **(A U B)** é 3, representado pelas TIDs 103, 104 e 106, nas quais **A** e **B** estão presentes na mesma cesta. O número total de transações **(N)** é 6, representado pelo total de TIDs e assim:

$$\text{Sup}(\{\text{leite}\} \rightarrow \{\text{iogurte}\}) = \frac{3}{6} = 0{,}5 = 50\%$$

Isto é, 50% dos clientes levaram leite e iogurte em uma mesma compra, sendo assim, de 6 compras realizadas em três o cliente está levando leite e iogurte.

2.5.3.3 *Lift* (Medida de Correlação)

Outro índice estatístico comumente utilizado para definir o grau de interesse de uma regra de associação é denominado *Lift*, também conhecida como *Interest*. Segundo Pinho (2006) essa medida de interesse é utilizada para avaliar as dependências entre o conjunto de itens do antecedente e o do consequente de uma regra de associação. O *Lift* de uma regra de associação **A** → **B** indica o quanto mais frequente torna-se **B**, quando **A** ocorre e é calculado como:

$$\textbf{\textit{Lift}(A} \rightarrow \textbf{B)} = \frac{\text{Conf (A} \rightarrow \text{B)}}{\text{Sup (B)}}$$

De acordo com o resultado obtido no exemplo dos itens 2.3.2.1 e 2.3.2.2 tem-se que **Conf** = 0,75 e **Sup** = 0,5 e, sendo assim:

$$\textbf{\textit{Lift}({leite}} \rightarrow \textbf{{iogurte})} = \frac{0{,}75}{0{,}5} \text{ ou } \frac{75\%}{50\%} = 1{,}5$$

O valor resultante desse cálculo indica que a compra de {iogurte} é 1,5 vez maior quando se compra {leite} e a probabilidade conjunta de leite e iogurte ocorrerem juntos é maior que probabilidade deles ocorrerem de forma independentes na base de dados.

Quando o ***Lift*(A → B) > 1**, **A** e **B** são positivamente correlacionados, ou seja, a presença de **A** aumenta as chances da ocorrência de **B**. Quando ***Lift*(A → B) < 1**, **A** e **B** são negativamente correlacionados, o que significa que a presença de **A** diminui as chances da ocorrência de **B**. ***Lift*(A → B) = 1**, então **A** e **B** são independentes e não há correlação entre eles. Esta medida varia entre 0 e ∞ e quanto maior o valor do *Lift*, mais interessante torna-se a regra. Dessa forma, as regras que possuem *Lift* maior que 1 são mais interessantes que as demais (SCHONHORST, 2010).

Na próxima seção é apresentada a descrição de um algoritmo muito utilizado para extração de RA em bases de dados, o algoritmo Apriori.

2.6 ALGORITMO APRIORI

O algoritmo Apriori é um importante algoritmo proposto por Agrawal e Srikant (1994), para a mineração de itemsets frequentes para geração de regras de associação Booleanas. Dado um banco de transações D e um nível mínimo de suporte B, o algoritmo encontra todos os itemsets frequentes com relação a D e B (SILVA, 2004).

Na Figura 6 pode-se observar o exemplo de um conjunto de itemsets, no qual cada linha da coluna Itemset representa uma compra (transação), sendo cada item referenciado por um número de identificação da lista de itens apresentada na Figura 4.

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

Figura 6 - Exemplo de um conjunto de *Itemsets*

Fonte: Amo (2012)

O algoritmo pode ser dividido em três fases principais: a fase de geração de candidatos, a fase de poda dos candidatos e a fase do cálculo do suporte. As duas primeiras fases são realizadas na memória principal e não necessitam que o BD D seja varrido. A memória secundária só é exigida quando o conjunto de itemsets é grande demais. Apenas na terceira fase, cálculo de suporte dos candidatos, é que o BD D é utilizado (AMO, 2012).

2.6.1 Fase de Geração do Conjunto de *Itemsets* Candidatos

Nesta fase é gerado um conjunto de *k-itemsets* **candidatos** a partir de um conjunto L_{k-1}, ou seja, diante de dois *itemsets* frequentes {1, 3, 4} e {1, 3, 5} de tamanho 3, é gerado um *itemset* candidato de tamanho 4 {1, 3, 4, 5}. O propósito é a geração de *itemsets* candidatos que possuam chance de serem frequentes. Como o algoritmo Apriori é feito de forma iterativa, isto é, os *itemsets* candidatos de tamanho *k* são calculados a partir de *itemsets* frequentes de tamanho *k-1* que já foram calculados num passo anterior, então, o *itemset* candidato provavelmente será frequente, pois se originou de dois *itemsets* frequente. Diante disso, todos os *itemsets* de tamanho *k-1* contidos nos candidatos de tamanho *k* deverão ser frequentes, portanto, deverão pertencer ao conjunto L_{k-1} (AMO, 2012). Na Figura 7 é ilustrado o processo de geração de *itemsets* do Apriori.

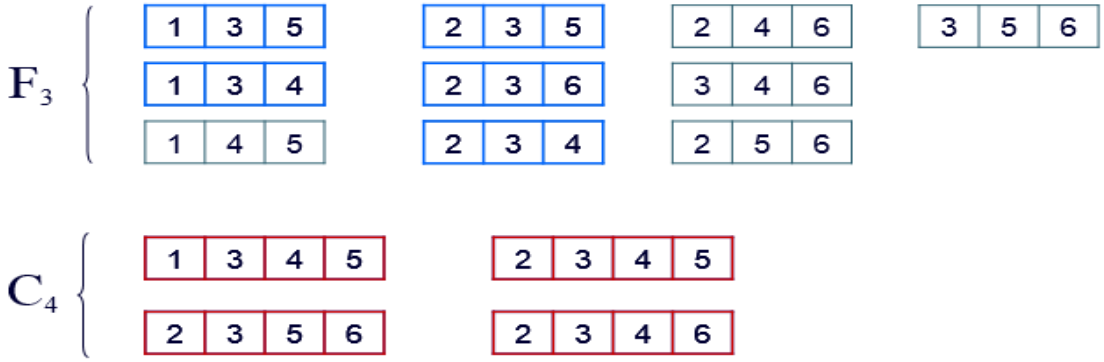

Figura 7 - Construção de um *Itemset* Candidato à Partir de dois Frequentes de Tamanho *K-1*
Fonte: Amo (2012)

Considerando o banco de transações ilustrado na Figura 7, onde **F_3** representa o conjunto de 3-*itemsets* **Frequentes** e **C_4** o conjunto de 4-*itemsets* **Candidatos** frequentes, tem-se:

F_3 = {{1,3,5}, {1,3,4}, {1,4,5}, {2,3,5}, {2,3,6}, {2,3,4}, {2,4,6}, {3,4,6}, {2,5,6}, {3,5,6}}

Onde o conjunto dos *itemsets* **candidatos** (**C_4**) frequentes de tamanho 4 da iteração seguinte será:

C_4= {{1,3,4,5}, {2,3,5,6}, {2,3,4,5}, {2,3,4,6}}

2.6.2 Fase de poda conjunto de *itemsets* candidatos

De acordo com a fase anterior, sabe-se que se um *itemset* **C_4** possuir um subconjunto de itens de tamanho **k-1** que não estiver em **F_3** ele poderá ser descartado, pois se este subconjunto não estiver em nenhum *itemset* de **F_3**, então ele não será frequente, ou seja, se um itemset não é frequente, nenhum conjunto maior que o contenha será frequente, pois de acordo com a propriedade Apriori: **Todos os subconjuntos não vazios de um *itemset* frequente devem também ser frequentes**, propriedade a qual o algoritmo é baseado.

Por definição, se um *itemset* I não satisfaz a um limiar mínimo de suporte, MinSup, então I não é frequente, isto é, P(I) < MinSup. Se um item A é acrescentado ao *itemset* I, então o *itemset* resultante (isto é, I U A) não pode ocorrer mais frequentemente do que I. consequentemente I U A também não será frequente, isto é, P(I U A) < MinSup.

Assim, de acordo com a propriedade Apriori nesta fase é gerado o conjunto de 4-*itemsets* **Candidatos** frequentes podado (**$C_{4\ podado}$**). Considerando a situação apresentada anteriormente, na Figura 8 é ilustrado o processo de geração do conjunto de *itemsets* podado (ARAÚJO, 2009).

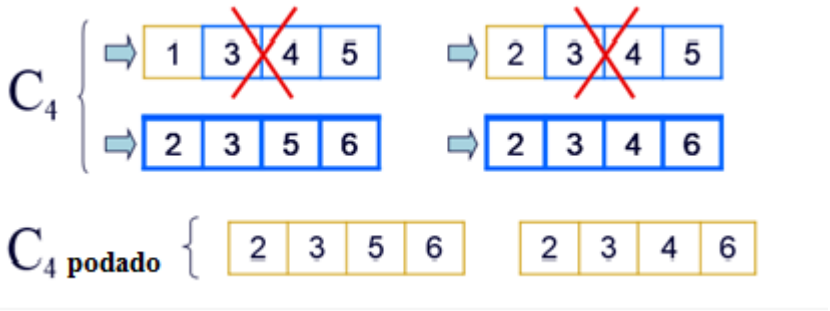

Figura 8 - Construção de um *Itemset* Podado à Partir do Conjunto de *Itemsets* Candidato

Fonte: Amo (2012)

C_4= {{1,3,4,5}, {2,3,5,6}, {2,3,4,5}, {2,3,4,6}}

Os *itemsets* {1,3,4,5} e {2,3,4,5} foram descartados pois não terão chances de serem frequentes. Ambos possuem a sequência {3,4,5} de tamanho 3 que não é frequente, pois não aparece em nenhum *itemset* de **F_3**. Como o *itemset* candidato {2,3,5,6} possui algum *itemset* frequente correspondente em **F_3** para as combinações {2,3,5}, {2,3,6}, {2,5,6} e {3,5,6} ele passa a fazer parte do conjunto candidato podado (ARAÚJO, 2009). O mesmo é válido para o *itemset* candidato {2,3,4,6}, chegando-se assim ao **$C_{4\ podado}$**.

$C_{4\ podado}$ = {{2,3,5,6}, {2,3,4,6}}

2.6.3 Fase do Cálculo do Suporte

Nesta fase é calculado o suporte de cada um dos *itemsets* do conjunto **$C_{4\ podado}$**. Este cálculo é realizado fazendo-se a varredura de todas as transações contidas no BD, onde cada transação é comparada com os dois *itemsets* do **$C_{4\ podado}$**. Caso a transação contenha todos os itens do **$C_{4\ podado}$** esse *itemset* é incrementado em uma unidade. Na Figura 9 é ilustrado um exemplo de BD utilizado para varredura.

Banco de Dados
1 3 5 7 8
1 2 3 4 5 6 7
2 3 4 6 8
2 3 4 5 7 8
1 2 3 4 6 9

Figura 9 - BD a ser Utilizado para Varredura

Fonte: Amo (2012)

No exemplo tem-se 5 transações a serem varridas. No início da varredura o primeiro *itemset* {1,3,5,7,8} é comparado aos dois *itemsets* do **$C_{4\ podado}$** ,note que os *itemset*s {2,3,5,6} e {2,3,4,6} não encontram-se contidos no primeiro *itemset* do BD. Por não estarem contidos, nenhum dos dois *itemsets* são incrementados em uma unidade. Feito isso, o segundo *itemset* é varrida e os mesmos procedimentos realizados até chegar-se ao final do BD. Chegando-se ao término da varredura os *itemsets* podados estariam dispostos conforme ilustrado na Figura 10.

$C_{4\ podado}$ {2,3,5,6} I

Banco de Dados
1 3 5 7 8
1 2 3 4 5 6 7
2 3 4 6 8
2 3 4 5 7 8
1 2 3 4 6 9

$C_{4\ podado}$ {2,3,4,6} I I I

Banco de Dados	Banco de Dados	Banco de Dados
1 3 5 7 8	1 3 5 7 8	1 3 5 7 8
1 2 3 4 5 6 7	1 2 3 4 5 6 7	1 2 3 4 5 6 7
2 3 4 6 8	2 3 4 6 8	2 3 4 6 8
2 3 4 5 7 8	2 3 4 5 7 8	2 3 4 5 7 8
1 2 3 4 6 9	1 2 3 4 6 9	1 2 3 4 6 9

Figura 10 - *Itemsets* Podados com seus Respectivos Incrementos

Fonte: Amo (2012)

Conforme o cálculo de suporte explicado e exemplificado no item 2.5.3.2, o resultado de suporte para $\mathbf{C_{4\ podado}}$ {2,3,5,6} foi:

$$\mathbf{Sup(C_{4\ podado}}\{2,3,5,6\}\mathbf{)} = \frac{1}{5} = 0{,}2 = 20\%$$

E para $\mathbf{C_{4\ podado}}$ {2,3,4,6},

$$\mathbf{Sup(C_{4\ podado}}\{2,3,4,6\}\mathbf{)} = \frac{3}{5} = 0{,}6 = 60\%$$

Definindo-se um suporte mínimo de 50%, apenas $\mathbf{C_{4\ podado}}$ {2,3,4,6} seria aproveitado.

A geração das informações pelo algoritmo Apriori pela ferramenta WEKA é descrita no Apêndice A.

3. FERRAMENTA WEKA

A ferramenta WEKA foi desenvolvida na universidade de Waikato na Nova Zelândia no ano de 1999, possuindo como principais pesquisadores Eibe Frank, Mark Hall, Peter Reutemann e Len Trigg. Também contribuíram significativamente Remco Bouckaert, Richard Kirkby, Ashraf Kibriya, Xin Xu e Malcolm Ware. WEKA é uma ferramenta de MD implementada em Java que oferece suporte para vários sistemas operacionais (Windows, Linux e Mac). WEKA é um software livre composto por um conjunto de implementações algorítmicas e de diversas técnicas de MD. Por ser uma ferramenta que possui o código aberto e tem fins acadêmicos, está disponível gratuitamente na página do próprio WEKA[2] para ser utilizado por qualquer usuário (LIMA, 2005).

A ferramenta WEKA é de licença GPL (GNU Public License), o que significa que este programa é de distribuição e divulgação gratuita. A ferramenta foi desenvolvida em Java e pode ser executada em plataformas para as quais existem versões da máquina virtual Java (MORATE, 2004).

4. DESCRIÇÃO DA BASE DE DADOS

A base de dados é constituída de dados sobre estudantes que evadiram nos cursos de computação do CCA. Esta base possui 119 instâncias, onde cada instância representa um aluno, sendo 77 instâncias para SI e 42 para CC. Esses dados foram obtidos no Portal do Aluno, lugar onde ficam armazenadas todas as informações dos alunos do CCA e só pode ser acessada pelo coordenador do curso, pois são informações extremamente sigilosas contendo todos os dados pessoais de cada aluno. Informações pessoais como CPF, identidade, telefone, nome, nome da mãe, email e matrícula foram descartadas por não serem relevantes para a análise de evasão.

A base de dados foi criada no início do primeiro semestre de 2014 e, por isso, só possui informações referentes aos alunos que ingressaram antes desse semestre, isto é, alunos de 2013 e de anos anteriores.

[2] WEKA: http://www.cs.waikato.ac.nz/ml/weka/

Foram obtidas 8 planilhas do Portal do Aluno, onde cada curso gerou 4 planilhas. Cada planilha representa um ano para CC ou SI, ou seja, as 4 planilhas de CC por exemplo, referenciam os anos que vão de 2010 até 2013. Cada planilha contém as informações dos alunos evadidos naquele ano. As planilhas foram obtidas desta maneira porque o Portal do Aluno não permite extração agrupada por ano de evasão, sendo assim não foi possível extrair somente uma planilha para cada curso.

Conforme pode ser visto no gráfico da Figura 11 a quantidade de alunos evadidos no ano de 2012 foi a maior para ambos os cursos e para SI este ano representou aproximadamente 67,5% do total.

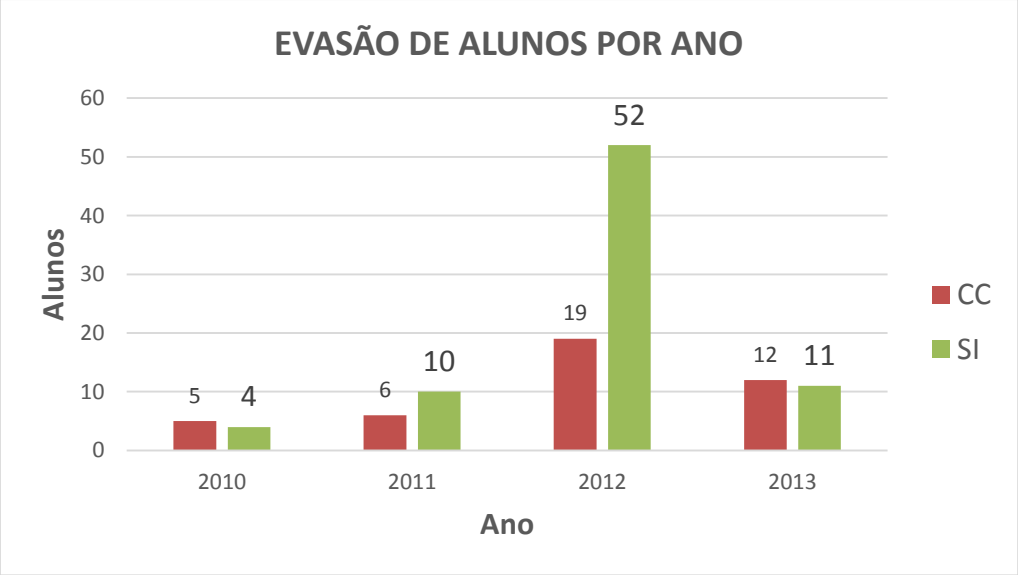

Figura 11 - Evasão de alunos por ano

Por se tratarem de planilhas diferenciadas apenas pelo ano de evasão, as mesmas foram agrupadas em um único registro para cada curso, facilitando assim a análise dos dados, sendo o registro a junção dos dados das 4 planilhas para cada curso. Devido à localização do campus CCA-UFES, o estado de origem da maior parte dos estudantes analisados nos cursos de CC e de SI eram do Espírito Santo, representando 67% dos alunos do curso de CC e 76% dos alunos do curso de SI. Grande parte dos estudantes de outros estados são procedentes da região sudeste, conforme pode ser visto na Figura 12.

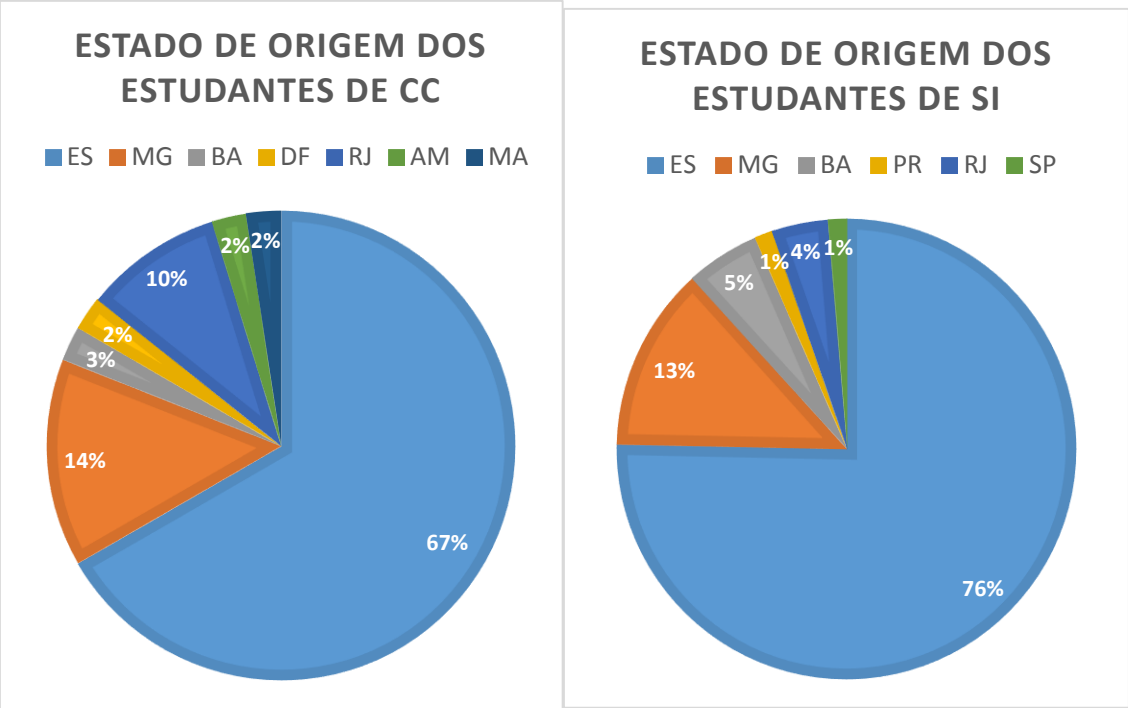

Figura 12 - Estado de Origem dos Estudantes (CC e SI)

Na base de dados, para cada registro (cada estudante) existem 7 atributos, sendo eles:

Cotista:

- Variável nominal que pode assumir os valores **S** ou **N**, sendo **S** usado para indicar que o estudante é cotista e **N** para indicar aqueles que não são cotistas.

Sexo:

- Variável nominal que pode assumir os valores **F** ou **M**, sendo que **F** representa **Feminino** e **M** representa **Masculino**.

Cidade:

- Variável nominal que representa o nome da cidade natal do estudante.

Estado:

- Variável nominal que representa o estado da cidade natal do estudante.

FormaEvasao:

- Variável nominal que representa a forma de evasão do aluno, podendo assumir um dos cinco valores descritos abaixo:
  - **Desistência:** situação em que o aluno desiste do curso;
  - **Desligamento por Abandono:** situação em que o aluno abandona o curso por dois períodos letivos, consecutivos ou não;

- **Desligamento por três reprovações em uma disciplina:** situação em que o aluno é reprovado por três vezes em uma mesma disciplina, sendo que nessa situação um estudante pode, ou não, ser desligado do curso. Estudantes que já tiverem concluído mais de 80% do curso não podem ser desligados por esse motivo;
- **Transferido:** situação em que o estudante deixa o curso para ingressar em um novo curso, podendo ser na mesma universidade ou não;
- **Reopção de Curso:** situação em que o estudante muda seu curso dentro da própria universidade.

AnoEvasao:

- Variável numérica inteira usada para representar o ano em que o aluno evadiu, variando de 2010 a 2013.

SemestreEvasao:

- Variável nominal usada para representar o semestre de evasão**,** podendo ser no primeiro ou segundo semestre.

Os percentuais das evasões por cada forma no período analisado, por curso, são apresentados nos gráficos da Figura 13.

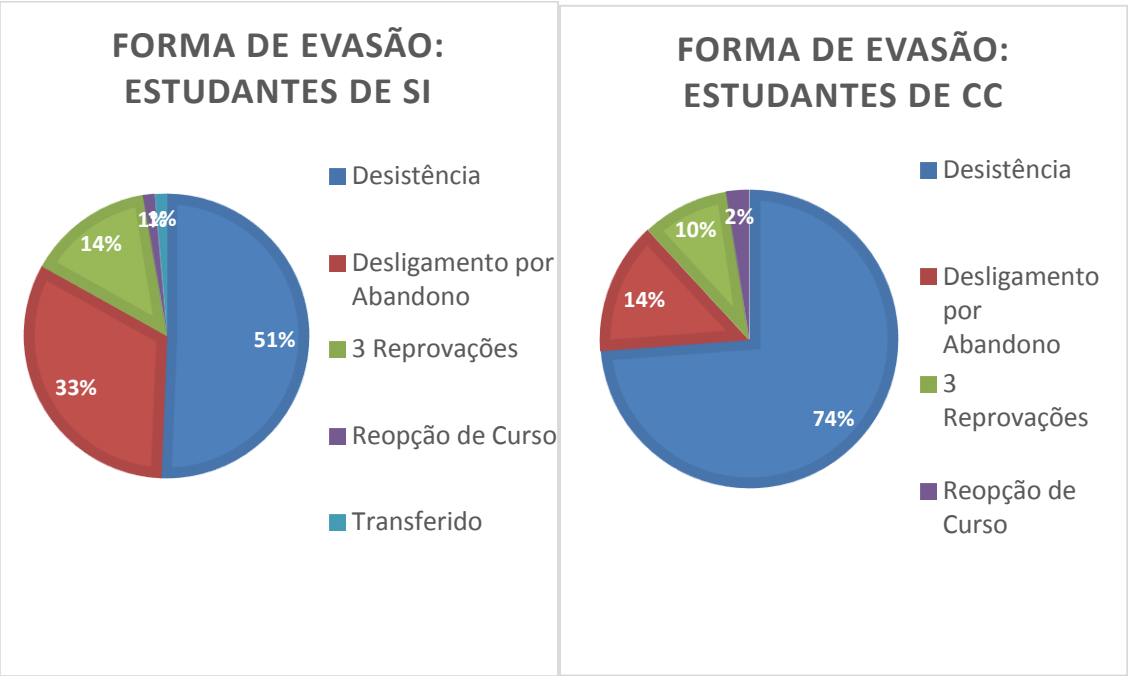

Figura 13 - Forma de Evasão dos Estudantes (CC e SI)

Pode-se observar que a desistência representa um percentual notório em relação as demais modalidades de evasão. O desligamento por abandono que aparece em sequência está fortemente atrelado à modalidade de evasão por desistência, pois o desligamento por abandono é uma desistência não formalizada, pois o aluno deixa a universidade sem dar baixa na sua matrícula, sendo assim, ao invés de entrar na modalidade desistência acaba entrando em situação de desligamento por não oficializar seu desvinculo com a universidade.

Em relação ao percentual de evasões entre cotistas e não cotistas, como pode ser observado no gráfico da Figura 14, tanto para o curso de SI quanto para o curso de CC, a maior parte das evasões registradas são de estudantes não cotistas.

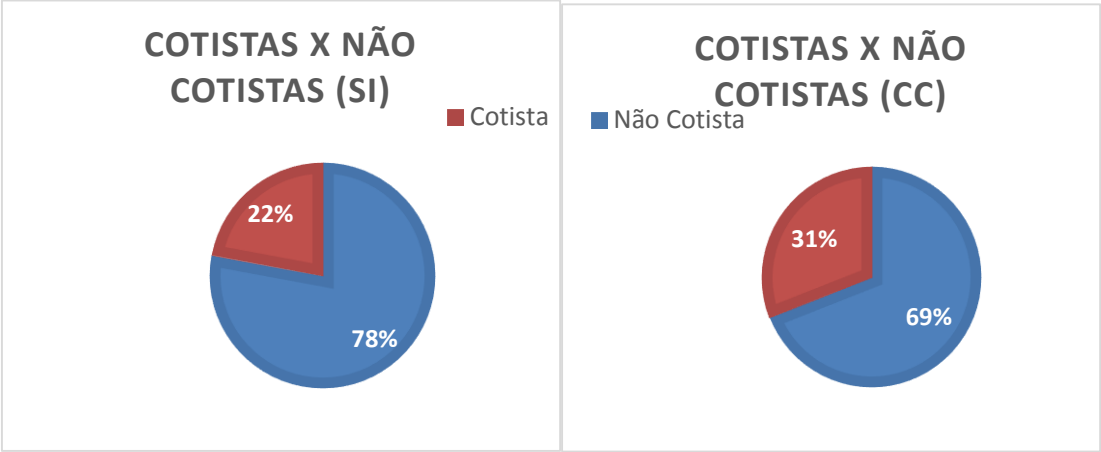

Figura 14 - Distribuição de Evasão - Cotistas x Não Cotistas (CC e SI)

Em relação às evasões por sexo, a maior parte dos estudantes que evadiram são do sexo masculino, como pode ser visualizado na Figura 15.

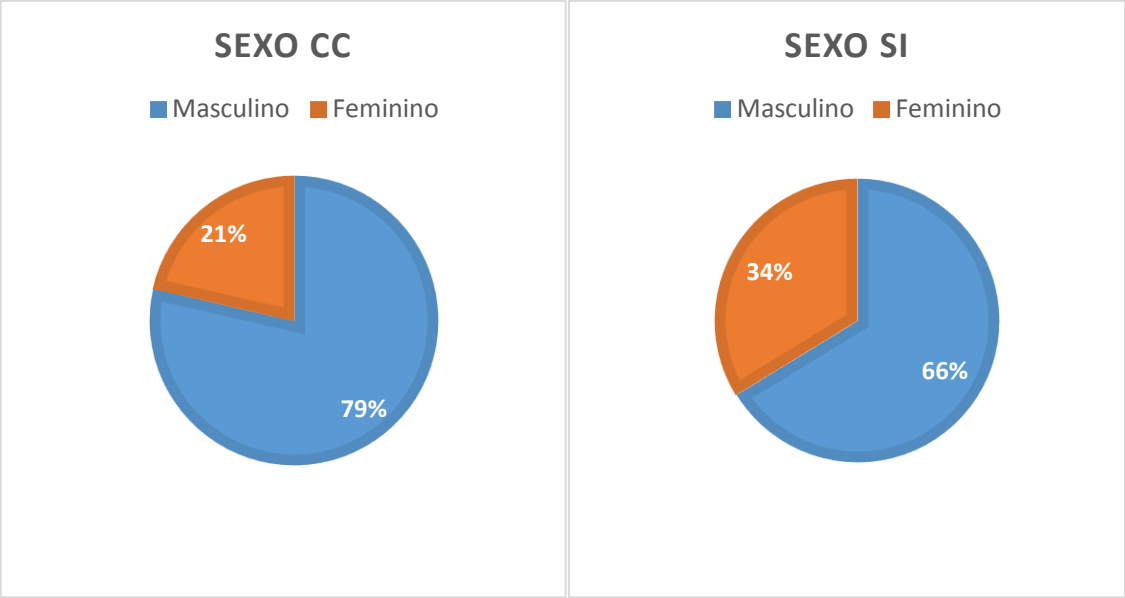

Figura 15 - Distribuição da Evasão por Sexo (CC e SI)

5. METODOLOGIA

O algoritmo usado para extração das regras de associação é o Apriori (Agrawal e Srikant, 1994), em uma implementação disponível na ferramenta WEKA. Essa implementação permite o uso de diferentes métricas de interesse para ranquear as regras geradas e de diferentes valores limiares para se limitar o número de regras de associação que serão geradas.

5.1 CONFIGURAÇÃO DOS EXPERIMENTOS

A geração foi realizada em dois experimentos com configurações diferentes para as métricas de interesse. Para cada experimento foram realizadas execuções preliminares buscando encontrar valores combinados de *Lift*, suporte e confiança que gerassem regras contendo o atributo “FormaEvasao” com Consequente das regras, pois assim se torna possível analisar determinadas tendências de evasão. Os valores para as execuções preliminares foram ajustados para gerarem 100 regras, valor o qual foi definido porque analisar muitas regras consumiria tempo e muitas delas se tratam de regras de pouco interesse, logo, não é produtivo gerar um número grande de regras, já que muitas regras terão valores baixos para suas métricas de interessabilidade ou serão regras redundantes. Com esse valor fixado foi iniciada a busca de valores para extração de regras. Para confiança em CC foi reproduzido o seguinte cenário:

Na primeira execução o suporte foi fixado em 20% e a confiança foi ajustada para 80%, porém com esse valor geravam-se poucas regras. No segundo momento o valor da métrica de confiança foi reduzido para 60% e o suporte foi mantido, o que, por sua vez, fez com que o algoritmo retornasse uma quantidade muito grande de regras (mais de cem regras). Na terceira execução, o valor foi ajustado para 70% e o suporte mantido. Como esse valor resultou uma quantidade de regras próxima ao desejado (99 regras), restou-se ajustar o suporte. Aumentando-se o suporta para 30% apenas 25 regras foram geradas, porém aparentemente interessantes para análise mediante o que foi proposto. Decidindo-se por fim usar os valores de 70% para a confiança e 30% para o suporte, isto é, gerar somente regras com 70% ou mais de confiança e com 30% ou mais de suporte.

No segundo experimento para CC, utilizou-se a métrica *Lift* para a extração das regras, configurando-se o valor mínimo para a extração das melhores regras como 1,2 de *lift* e suporte mínimo de 25% para os *itemsets,* lembrando que o mesmo foi calibrado de forma aleatória buscando-se os possíveis melhores valores que gerassem regras interessantes para análise. Esses valores foram fixados desta maneira pois ao aumentar o suporte para 30% apenas 4 regras eram geradas e ajustando-se o *lift* para 1,3 algumas regras com consequente contendo “FormaEvasao” desapareciam.

Para SI o cenário de confiança foi parecido com o de CC, onde na primeira execução o suporte foi fixado em 20% e a confiança foi ajustada para 80%, porém com esse valor não foi gerada nenhuma regra contendo “FormaEvasao” como consequente da regra. No segundo momento o valor da métrica de confiança foi reduzido para 70% e o suporte foi mantido, porém mais uma vez nenhuma regra interessante foi gerada. Na terceira execução, o valor foi ajustado para 60% e o suporte mantido, porém se manteve o cenário sem o consequente da regra desejado. Na quarta execução, o valor foi mantido em 60% e o suporte aumentado para 30% o qual gerou uma regra interessante. Em uma quinta passagem o suporte foi ajustado em 40% e a confiança mantida, porém nenhuma regra desejada foi encontrada. Na sexta passagem percebendo-se que o aumento do suporte não iria gerar regras desejadas o mesmo foi novamente fixado em 30% e a confiança ajustada para 55%, gerando-se 100 regras das quais 3 possuíam o critério estabelecido, sendo este critério, regras contendo “FormaEvasao” como consequente. Desta forma, optou-se por fim usar os valores de 55% para a confiança e 30% para o suporte, pois outros testes não geraram um número maior do que 3 regras interessantes.

No segundo experimento para SI, foi utilizado a métrica *Lift* para a extração das regras, onde foi configurando o valor mínimo para a extração das regras como 1,2 de *lift* e suporte mínimo de 30% para os *itemsets,* lembrando que o mesmo foi calibrado de forma aleatória buscando-se os possíveis melhores valores que gerassem regras interessantes para análise. Esses valores foram fixados desta maneira pois ao aumentar o suporte para 40% nenhuma regra era gerada e ajustando-se o *lift* para 1,3 algumas regras com consequente contendo “FormaEvasao” desapareciam.

Para cada experimento, das regras resultantes, selecionou-se apenas as regras com *itemsets* contendo informação sobre a evasão: somente as regras com consequente contendo Forma de Evasão foram avaliadas, já que para o presente trabalho tenta-se estabelecer somente as associações entre a evasão dos estudantes e outros atributos da base de dados em cada curso.

6. RESULTADOS – REGRAS DE ASSOCIAÇÃO EXTRAÍDAS

Na Figura 16 são ilustradas as regras de CC para o experimento configurado para filtrar as regras por meio da métrica confiança. Todas as regras da Figura 16 possuem como consequente a Forma de Evasão por Desistência.

```
 1. SemestreEvasao=SegundoSemestre 16 ==> FormaEvasao=Desistencia 15    conf:(0.94)
 8. Cotista=N Sexo=M 22 ==> FormaEvasao=Desistencia 17                 conf:(0.77)
11. Sexo=M 33 ==> FormaEvasao=Desistencia 25                           conf:(0.76)
17. Sexo=M Estado=ES 19 ==> FormaEvasao=Desistencia 14                 conf:(0.74)
19. Cotista=N 29 ==> FormaEvasao=Desistencia 21                        conf:(0.72)
22. Estado=ES 28 ==> FormaEvasao=Desistencia 20                        conf:(0.71)
24. Cotista=N Estado=ES 21 ==> FormaEvasao=Desistencia 15              conf:(0.71)
```

Figura 16 - Regras de CC para o Experimento de Confiança

Tomando como exemplo a regra 8 "`Cotista=N Sexo=M 22 ==> FormaEvasao=Desistencia 17`", a interpretação literal dessa regra seria que não cotistas do sexo masculino tendem a evadir por desistência com confiança de 77%.

Pode diagnosticar-se a existência de uma associação entre alunos não cotistas do sexo masculino do estado do Espírito Santo que evadiram no segundo semestre terem evadido na forma de desistência.

No segundo experimento, o qual extraiu regras utilizando a métrica de interesse de *Lift* com valor mínimo 1.2 e suporte mínimo de 25%, foram obtidas regras como as apresentadas na Figura 17.

```
 6. SemestreEvasao=SegundoSemestre 16 ==> Cotista=N FormaEvasao=Desistencia 11    conf:(0.69) < lift:(1.38)>
 8. Cotista=N SemestreEvasao=SegundoSemestre 11 ==> FormaEvasao=Desistencia 11    conf:(1)    < lift:(1.35)>
16. SemestreEvasao=SegundoSemestre 16 ==> FormaEvasao=Desistencia 15              conf:(0.94) < lift:(1.27)>
22. AnoEvasao=2013 12 ==> FormaEvasao=Desistencia 11                              conf:(0.92) < lift:(1.24)>
24. Sexo=M SemestreEvasao=SegundoSemestre 12 ==> FormaEvasao=Desistencia 11       conf:(0.92) < lift:(1.24)>
```

Figura 17 - Regras de CC para o Experimento de Lift

Nesse experimento foram geradas 5 regras obedecendo o critério estabelecido. Neste experimento tomando como exemplo a regra 8 "`Cotista=N SemestreEvasao=SegundoSemestre 11 ==> FormaEvasao=Desistencia 11`", a regra pode ser interpretada como: alunos não cotistas que evadiram no segundo semestre tendem a evadir por desistência com confiança de 100%, isto é, todos os alunos que evadiram no segundo semestre e que não eram cotistas evadiram por desistência.

A regra 6 "`SemestreEvasao=SegundoSemestre 16 ==> Cotista=N FormaEvasao=Desistencia 11`", inclui em seu consequente tanto "FormaEvasao = Desistência" quanto "Cotista=N", o que leva à interpretação de que alunos evadidos no segundo semestre tendem a ser não cotistas evadidos por desistência.

Utilizando as regras restantes e, conforme interpretado anteriormente para o experimento de confiança, neste experimento se tem algo parecido, pois no experimento de confiança foram geradas regras contendo o estado dos estudantes e com o uso da métrica Lift não aparece estado, sendo assim, neste experimento foi diagnosticado a existência de uma forte associação entre alunos não cotistas do sexo masculino que evadiram no segundo semestre terem evadido na forma de desistência.

Após realizadas as interpretações de regras geradas a partir da base de dados dos alunos de CC, os dados dos alunos de SI foram utilizados para a extração de regras. Na Figura 18 são ilustradas as regras de SI para o experimento baseado na métrica de confiança. Todas as regras deste experimento possuíam como consequente alguma forma de evasão.

```
84. Cotista=N Sexo=M Estado=ES 38 ==> FormaEvasao=Desistencia 23    conf:(0.61)
96. Cotista=N Sexo=M 44 ==> FormaEvasao=Desistencia 25              conf:(0.57)
89. Sexo=M Estado=ES 42 ==> FormaEvasao=Desistencia 25              conf:(0.6)
```

Figura 18 - Regras de SI para o Experimento de Confiança

Uma interpretação para a regra 84 "`Cotista=N Sexo=M Estado=ES 38 ==> FormaEvasao=Desistencia 23`", é que alunos não cotistas do sexo masculino e do estado do Espírito Santo tendem a evadir por desistência com confiança de 61%. Nessa regra ocorre um *itemset* no lado do antecedente da regra que possui itens que já ocorreram isoladamente em outras regras, como é o caso das regras 89 e 96. Apesar de possuir confiança mais elevada que as regras 89 e 96, a regra 84 é mais específica do que essas duas e, portanto, costuma-se considera-la menos interessante.

No segundo experimento, no qual extraiu-se regras utilizando Lift com valor mínimo de 1.2 e suporte mínimo de 30%, foram obtidas regras como as da Figura 19.

```
 2. AnoEvasao=2012 52 ==> FormaEvasao=DesligamentoPorAbandono 25    conf:(0.48) < lift:(1.48)>
 5. Sexo=M Estado=ES 42 ==> Cotista=N FormaEvasao=Desistencia 23    conf:(0.55) < lift:(1.36)>
 7. Sexo=M 51 ==> Cotista=N Estado=ES FormaEvasao=Desistencia 23    conf:(0.45) < lift:(1.34)>
 9. Cotista=N Sexo=M 44 ==> Estado=ES FormaEvasao=Desistencia 23    conf:(0.52) < lift:(1.3)>
13. Cotista=N Estado=ES 50 ==> Sexo=M FormaEvasao=Desistencia 23    conf:(0.46) < lift:(1.27)>
19. Estado=ES 58 ==> Cotista=N Sexo=M FormaEvasao=Desistencia 23    conf:(0.4)  < lift:(1.22)>
23. Sexo=M 51 ==> Estado=ES FormaEvasao=Desistencia 25              conf:(0.49) < lift:(1.22)>
21. Sexo=M 51 ==> Cotista=N FormaEvasao=Desistencia 25              conf:(0.49) < lift:(1.22)>
```

Figura 19 - Regras de SI para o Experimento de Lift

As regras do segundo experimento servem como complemento para as regras do primeiro experimento, pois o mesmo gerou regras com interpretações parecidas (exceto a regra 2), ou seja, regras contendo os mesmos atributos ("Cotista=N", Sexo=M" e "Estado=ES"), porém de menor confiança. Neste experimento tem-se a regra 2 citada anteriormente "`AnoEvasao=2012 52 ==> FormaEvasao=DesligamentoPorAbandono 25`", em que nela pode-se interpretar que alunos com ano de evasão igual a 2012 tendem a evadir na forma de desligamento por abandono com confiança de 48%. O que neste caso, como o que foi interpretado para a regra 22 de CC, não diz algo interessante, pois se trata de algo já decorrido e que aconteceu em 2012, sendo assim, esta regra não se torna relevante na tentativa de prever possíveis fatores relacionados com novas evasões.

A regra 7 "Sexo=M 51 ==> Cotista=N Estado=ES FormaEvasao=Desistencia 23", traz como consequente os atributos "FormaEvasao=Desistencia", "Cotista=N" e "Estado=ES", e sua interpretação é que alunos do sexo masculino tendem a ser do estado do espírito santo, não cotistas e a evadir por desistência, o que torna essa regra menos genérica, portanto não muito interessante.

A regra 19 "Estado=ES 58 ==> Cotista=N Sexo=M FormaEvasao=Desistencia 23", expressa que estudantes do estado do Espírito Santo tendem a ser do sexo masculino, não cotistas e a evadir por desistência.

Por fim, de acordo com as regras extraídas a partir das bases de dados utilizadas, os fatores que parecem estar mais relacionados com a evasão por desistência, no contexto dos cursos de SI e de CC do CCA-UFES, são o fato do aluno ser não cotista, do sexo masculino e do estado do Espírito Santo. Essa associação, no entanto, pode ocorrer devido ao fato desse perfil de estudante ser o mais presente nas bases de dados utilizadas, tornando maior o valor das métricas de interesse utilizadas.

7. CONCLUSÃO

A partir da análise dos resultados, foi possível diagnosticar que de maneira geral os alunos que evadiram são estudantes do sexo masculino, naturais do Espírito Santo e não cotistas. A principal forma de evasão é por desistência.

Os resultados, entretanto, acabam recebendo influência da prevalência de certos valores desses atributos na base. Um exemplo é o sexo masculino em CC, representando 79% do total de alunos do curso. Outros exemplos seriam a quantidade de alunos do Espírito Santo representando 76% do total da base de SI e não cotistas representando 78% do total dessa mesma base. Sendo assim, esses atributos aparecem em média 3 vezes a cada 4 registros. Por esse motivo tais características aparecem em grande parte das regras geradas, tornando assim, regras não interessantes pois se devem à prevalência e não a cenários inesperados.

Outro aspecto a ser considerado é o tempo de existência dos cursos de CC e SI do CCA. Tratam-se de cursos relativamente recentes, o que implica em uma base de dados contendo poucos dados registrados, consistindo de um histórico de apenas 4 anos.

7.1 DIFICULDADES

As tarefas de mineração são muito úteis para descobrir conhecimento implicitamente existente em grandes volumes de dados. Entretanto, os resultados desse trabalho não revelaram associações tão interessantes por serem inesperadas, inéditas ou de difícil percepção sem o uso de técnicas de mineração de dados. A base de dados elaborada possui poucas informações socioculturais e abrange um número relativamente pequeno de estudantes (registros). Para essa aplicação, aparentemente a pouca quantidade de informações e de registros não contribuiu positivamente com a quantidade e as características e com os padrões que puderam ser extraídos da base. Certos itens não ofereceram contribuição para formação de regras por conta da baixa frequência na base. Como exemplo desse tipo de problema, do total de 119 alunos, somente um aluno foi transferido. Observações raras na base de dados podem indicar exceções interessantes, mas esse não foi o caso para certos valores dos atributos nessa pesquisa. Somente um caso isolado não permitiria a geração de uma conclusão realmente confiável sobre as características que contribuem para um estudante se transferir para outras instituições.

É importante que uma base de dados com mais atributos e com valores mais diversos para esses atributos seja elaborada para esse tipo de estudo, além de se ter disponível uma quantidade maior de registros.

Uma nova pesquisa poderia utilizar mais detalhes sobre os alunos, como características socioeconômicas, disciplinas que o aluno reprovou, quantidade de reprovações por disciplina e uma base contendo cursos com maior tempo de existência, de modo a permitir a exploração de um domínio mais abrangente de informações.

8. REFERÊNCIAS

AGRAWAL, R.; SRIKANT, R. Fast Algorithms for Mining Association Rules. The Annals of pharmacotherapy, 1994. Disponível em: <http://goo.gl/4avlmX>.

AMO, S. DE. Aula 2 - Mineração de Regras de Associação - O algoritmo APRIORI. Curso de Data Mining, p. 1–7, 2012. Disponível em: <http://goo.gl/BS50vh>.

AMORIM, T. Conceitos, técnicas, ferramentas e aplicações de Mineração de Dados para gerar conhecimento a partir de bases de dados. 2006. Disponível em: <http://migre.me/fOlzy>.

ARAÚJO, A. L. V. DE. Aplicação de Regras de Associação para Auxílio na Gestão de Vendas de uma Empresa Varejista Utilizando a ferramenta WEKA. 2009. Disponível em: <http://goo.gl/B9HPj0>.

BAKER, R; YACEF, K. The State of Educational Data Mining in 2009: A Review and Future Visions. . JEDM - Journal of Educational Data Mining, p. 3-17, v. 1, 2009.

CABENA, P.; HADJINIAN, P.; STADLER, R.; JAAPVERHEES; ZANASI, A. Discovering Data Mining: From Concept to Implementation. Prentice Halll, p. 195, 1998.

CAMILO, C. O.; SILVA, J. C. DA. Mineração de Dados: Conceitos, Tarefas, Métodos e Ferramentas. 2009. Disponível em: <http://goo.gl/xrxttU>.

CARVALHO, Allyson. Modelo de dados: Transacional x Dimensional. 2011. Disponível em: <http://goo.gl/celIcO>.

CARVALHO, Luis Alfredo Vidal D. Datamining - A Mineração de Dados no Marketing, Medicina, Economia, Engenharia e Administração. 1. ed. Ciência Moderna, 2005, p. 254.

CORTÊS, S. DA C.; PORCARO, R. M.; LIFSCHITZ, S. Mineração de Dados – Funcionalidades, Técnicas e Abordagens. 2002. Disponível em: <http://goo.gl/C3Lui>.

CASTRO, F., VELLIDO, A., NEBOT, A., MUGICA, F. Applying Data Mining Techniques to e-Learning Problems. Evolution of Teaching and Learning Paradigms in Intelligent Environment. Studies in Computational Intelligence. 62, Springer-Verlag, 2007, p.183-221.

DAMASCENO, M. Introdução a mineração de dados utilizando o Weka. n. 1, 1996.

DEKKER, G. W.; PECHENIZKIY, M.; VLEESHOUWERS, J. M. Predicting Students Drop Out : A Case Study. Educational Data Mining, p. 41–50, 2009.

ESCOVAR, E. L. G. Algoritmo SSDM para a Mineração de Dados Semanticamente Similares. São Carlos: UFSCar, p. 86, 2004. Disponível em: <http://goo.gl/8IyUXB>. .

FAYYAD, U.; PIATETSKY-SHAPIRO, G.; SMYTH, P. From Data Mining to Knowledge Discovery in Databases. , p. 37–54, 1996. Disponível em: <http://migre.me/fOiQc>.

GONÇALVES, E. C. Data Mining com a Ferramenta Weka. Fórum de Software Livre de Duque de Caxias, p. 1–60, 2000. Disponível em: <http://goo.gl/EhONwm>.

GONÇALVES, E. C. Regras de Associação e suas Medidas de Interesse Objetivas e Subjetivas. UFF - Universidade Federal Fluminense, 2005. Disponível em: <http://goo.gl/Vhp4j0>.

GOVERNO FEDERAL. Diretrizes Gerais do Programa de Apoio a Planos de Reestruturação e Expansão das Universidades Federais – REUNI. Plano de Desenvolvimento da Educação, 2007. Disponível em: <http://portal.mec.gov.br/sesu/arquivos/pdf/diretrizesreuni.pdf >.

HALL, M. The WEKA Data Mining Software: An Update. SIGKDD Explorations, v. 11, n. 1, 2009. Disponível em: <http://goo.gl/upi8yy>.

HAN, J.; KAMBER, M. Data Mining - Concepts and Techniques. Second Edi ed.Morgan Kaufmann Publishers, 2006.

HAND, D.; MANNILA, H.; SMYTH, P. Principles of Data Mining. 2001.

INEP. Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira. Investimentos Públicos em Educação, 2009. Disponível em: <http://portal.inep.gov.br/estatisticas-gastoseducacao>.

LAROSE, D. T. Discovering Knowledge in Data: An Introduction to Data Mining. JohnWiley & Sons, Inc., 2005.

LIMA, Tânia dos Santos. ESTUDO COMPARATIVO DOS ALGORITMOS DE CLASSIFICAÇÃO DA FERRAMENTA WEKA. Centro Universitário Luterano de Palmas (ULBRA). p. 1-76, 2005.

MANHÃES, L. M. B.; CRUZ, S. M. S. DA; COSTA, R. J. M.; ZAVALETA, J.; ZIMBRÃO, G. Previsão de Estudantes com Risco de Evasão Utilizando Técnicas de Mineração de Dados. Anais do XXII SBIE - XVII WIE, p. 150–159, 2011. Disponível em: <http://migre.me/fHAFX>.

MANHÃES, L. M. B.; CRUZ, S. M. S. DA; COSTA, R. J. M.; ZAVALETA, J.; ZIMBRÃO, G. Identificação dos Fatores que Influenciam a Evasão em Cursos de Graduação Através de Sistemas Baseados em Mineração de Dados: Uma Abordagem Quantitativa. VIII Simpósio Brasileiro de Sistemas de Informação (SBSI 2012) Trilhas Técnicas, p. 468–479, 2012. Disponível em: <http://goo.gl/RHDgwE>.

MILANI, F.; CAMARGO, S. DA S. Aplicação de Técnicas de Mineração de Dados na Previsão de Propensão à Evasão Escolar. p. 1–2, 2010. Disponível em: <http://goo.gl/4Uxg5N>.

MINAEI-BIDGOLI, B. DATA MINING FOR A WEB-BASED EDUCATIONAL SYSTEM. Michigan State University, 2004. Disponível em: <http://goo.gl/xzEzIK>.

MOORE, R. Retention Rates Research Project: Final Report. Sheffield Hallam University, Division of Access and Guidance, 1995.

MORATE, D. G. MANUAL DE WEKA. 2004. Disponível em: <http://goo.gl/f6nv2o>.

PINHO, F. S.; BRITO, J. A. DE M.; SEMAAN, G. S.; OCHI, L. S. Uma heurística baseada em GRASP para a extração de associações em bases de dados. 2006. Disponível em: <http://goo.gl/j1A4a6>.

SCHONHORST, G. B. Mineração de Regras de Associação Aplicada à Modelagem dos Dados Transacionais de um Supermercado. 2010. Disponível em: <http://goo.gl/0ZQIPj>.

SILVA, M. P. DOS S. Mineração de Dados - Conceitos, Aplicações e Experimentos com Weka. 2004. Disponível em: <http://goo.gl/DL81JN>.

STOFFEL, Wagner Pina; ZIZA, Cleiton Ricardo. EVASÃO ESCOLAR EM CURSOS SUPERIORES: Estudo comparativo entre os pedidos de trancamento e o aproveitamento escolar. IX Simpósio Pedagógico e Pesquisas em Comunicação (Simped), p. 1-11, 2014. Disponível em: <http://www.aedb.br/wp-content/uploads/2015/05/46321536.pdf>.

TAN, P.N., KUMAR V., SRIVASTAVA, J. Selecting the Right Objective Measure for Association Analysis. Information Systems, p. 293-313, 2004.

Weka 3: Data Mining Software in Java. Disponível em: <Weka 3: Data Mining Software in Java>. Acesso em: 1/12/2014.

ZHANG, C.; ZHANG S. Association Rules Mining: Models and Algorithms. Lecture Notes in Artificial Intelligence. v. 2307, p. 244, 2002.

9. APÊNDICES

APÊNDICE A - EXTRAÇÃO DE REGRAS COM O WEKA

Para extração das RA utilizando-se a técnica Apriori do WEKA é necessário realizar os procedimentos inicializados pela Figura 1, onde de antemão deve-se estar com o programa inicializado.

Na Figura 1 é exibida a tela ***WEKA GUI Chooser*** ou tela principal da ferramenta para a versão 3.6.10.

Figura A.1 – Tela Principal Weka

Fonte: Frank, 2013.

Para a extração de regras no WEKA deve-se selecionar o ícone ***Explorer*****,** sendo o mesmo um ambiente para explorar dados com WEKA. Feito isso, será aberto o ***Weka Explorer*** apresentada na Figura 2.

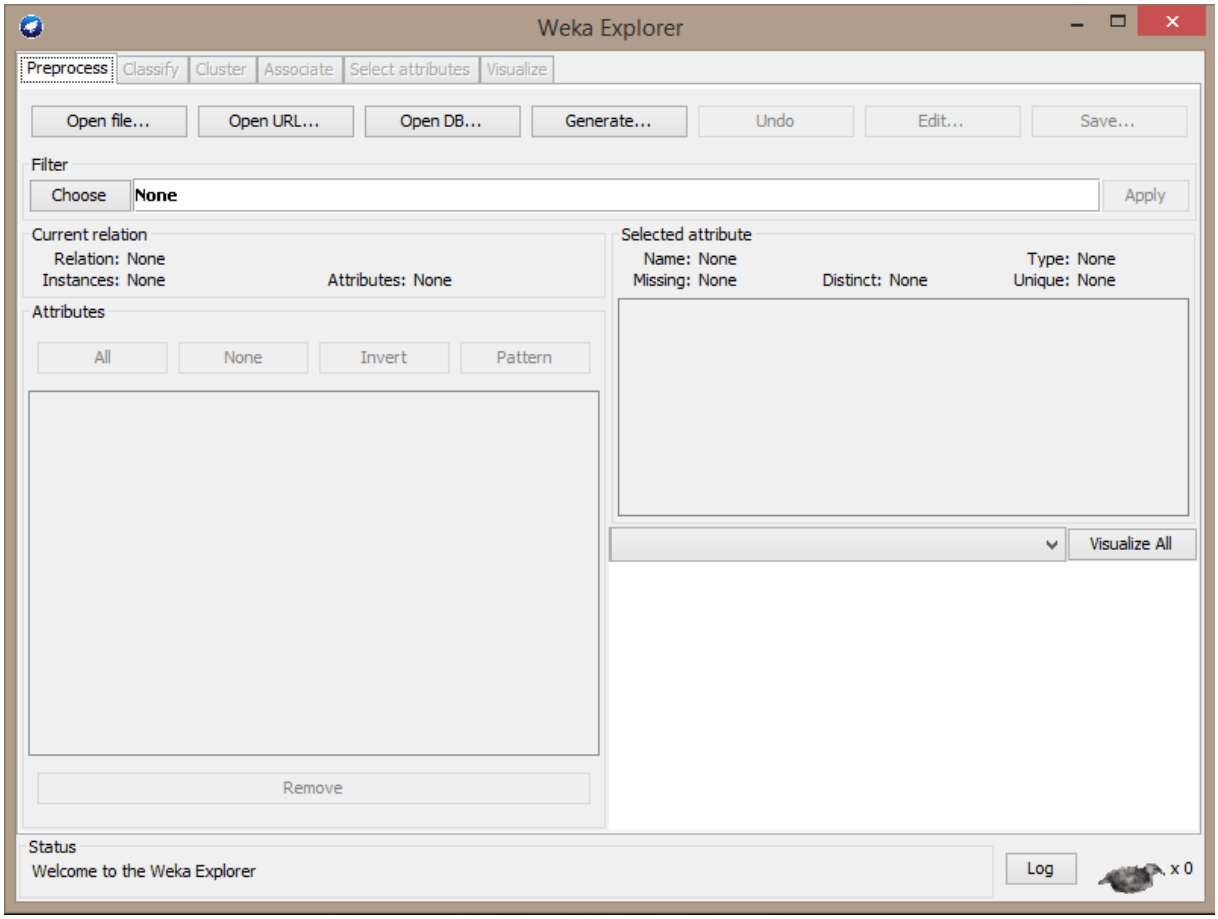

Figura A. 2 – Weka Explorer

Fonte: Frank, 2013.

No o ***Weka Explorer*** deve ser selecionado o ícone *"**Open file...**"*, ícone o qual é utilizado para importação. Antes desta opção ser selecionada é necessário transformar os dados do banco de dados em um arquivo que esteja com a extensão *.arff (*Attribute Relation File Format*) (LIMA, 2005). Para criação do arquivo no formato ARFF é necessário seguir os seguintes passos:

1. Abrir um editor de textos;

2. Adicionar um nome para o conjunto de dados após ***@relation***. Forma genérica: ***@relation*** <nome_do_conjunto>. Por exemplo, ***@relation*** DadosAlunos.

3. Em uma nova linha inserir os nomes dos atributos do conjunto de dados depois de ***@attribute***. Exemplo da forma genérica: ***@attribute*** <nome_do_atributo>. Deve-se respeitar um atributo por linha. No trabalho todos valores são nominais, sendo assim, tem-se atributos como ‘sim’ ou ‘não’, ‘F’ ou ‘M’. Estes atributos devem ser apresentados dentro de chaves como: ***@attribute*** sexo {F,M}. Esses valores vão ser reconhecidos automaticamente como os valores de cada atributo (LIMA, 2005).

4. Acrescentar o ***@data*** para que o algoritmo minerador entenda que vem logo após são os dados processados para a predição.

5. Feito isso o arquivo deve ser salvo como: <nome_do_arquivo>.arff, feito isso, o arquivo já estará pronto para ser analisado.

Na Figura 3 é apresentado um exemplo de arquivo ARFF seguindo a formatação detalhada nos passos acima.

```
1  @relation DadosAlunosSI
2
3  @attribute Cotista {S,N}
4  @attribute Sexo {F,M}
5  @attribute Cidade {ALEGRE,AFONSO_CLAUDIO,AIMORES,ANCHIETA,ARACRUZ,BARREIRAS,BOM_JESUS_DO_ITABAPOANA,CACHOEIRO_DE_ITAPEMIRIM,(
6  @attribute Estado {ES,MG,RJ,SP,BA,PR}
7  @attribute FormaEvasao {Desistencia,DesligamentoPorAbandono,Desligamento3Reprovacoesem1Disciplina,Transferido,ReopcaoDeCurso
8  @attribute AnoEvasao {2010,2011,2012,2013}
9  @attribute SemestreEvasao {PrimeiroSemestre,SegundoSemestre}
10
11 @data
12 N,F,CASTELO,ES,Desistencia,2012,SegundoSemestre
13 N,F,GANDU,BA,Desistencia,2011,SegundoSemestre
14 S,F,VILA_VELHA,ES,Desistencia,2013,SegundoSemestre
15 N,M,VITORIA,ES,Desistencia,2013,PrimeiroSemestre
16 N,M,VENDA_NOVA_DO_IMIGRANTE,ES,Desistencia,2013,PrimeiroSemestre
17 N,F,CASTELO,ES,DesligamentoPorAbandono,2012,SegundoSemestre
18 N,M,VITORIA,ES,Desligamento3Reprovacoesem1Disciplina,2012,PrimeiroSemestre
19 S,F,PANCAS,ES,Desistencia,2011,SegundoSemestre
20 S,F,ROLANDIA,PR,DesligamentoPorAbandono,2012,PrimeiroSemestre
21 N,M,GUACUI,ES,Desligamento3Reprovacoesem1Disciplina,2012,SegundoSemestre
22 S,F,MANHUACU,MG,Desistencia,2013,SegundoSemestre
23 N,M,VITORIA,ES,DesligamentoPorAbandono,2012,SegundoSemestre
24 N,F,GUACUI,ES,DesligamentoPorAbandono,2012,SegundoSemestre
25 N,M,CARANGOLA,MG,Desistencia,2010,PrimeiroSemestre
26 N,F,MANHUMIRIM,MG,Desistencia,2010,SegundoSemestre
27 S,F,ALEGRE,ES,DesligamentoPorAbandono,2012,PrimeiroSemestre
28 S,F,JOAO_MONLEVADE,MG,Transferido,2012,PrimeiroSemestre
29 N,M,ALEGRE,ES,Desistencia,2012,PrimeiroSemestre
30 N,F,VITORIA,ES,Desligamento3Reprovacoesem1Disciplina,2012,PrimeiroSemestre
31 N,M,DOMINGOS_MARTINS,ES,Desistencia,2012,SegundoSemestre
32 N,M,MANHUACU,MG,Desligamento3Reprovacoesem1Disciplina,2012,SegundoSemestre
33 N,F,CACHOEIRO_DE_ITAPEMIRIM,ES,Desligamento3Reprovacoesem1Disciplina,2012,SegundoSemestre
34 N,F,UBERLANDIA,MG,DesligamentoPorAbandono,2012,SegundoSemestre
```

Figura A.3 – Arquivo com Formatação Arff

Fonte: Frank, 2013.

Após a criação do arquivo no formato ARFF, pode-se voltar ao passo da Figura 2 e selecionar o ícone *"**Open file...**"*. Após carregamento do arquivo os campos que antes encontravam-se desabilitados, passam a ficar habilitados, conforme pode ser observado na Figura 4.

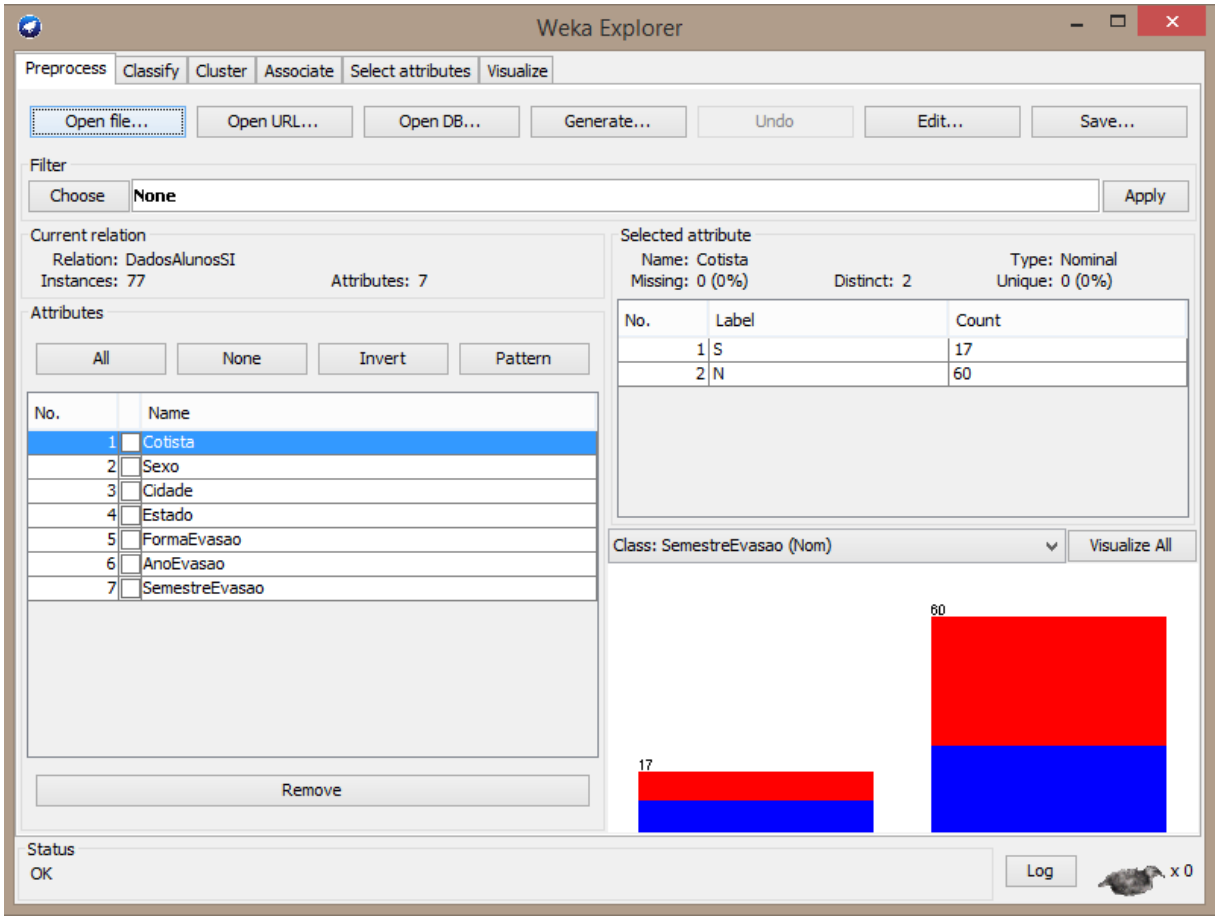

Figura A.4 – Weka Explorer com Arquivo Arff Carregado

Fonte: Frank, 2013.

No passo seguinte, a aba ***Associate*** deve ser selecionada, aba qual será utilizada para extrair as Regras de Associação. Feita a seleção desta aba o ***Weka Explorer*** será exibido conforme a Figura 5.

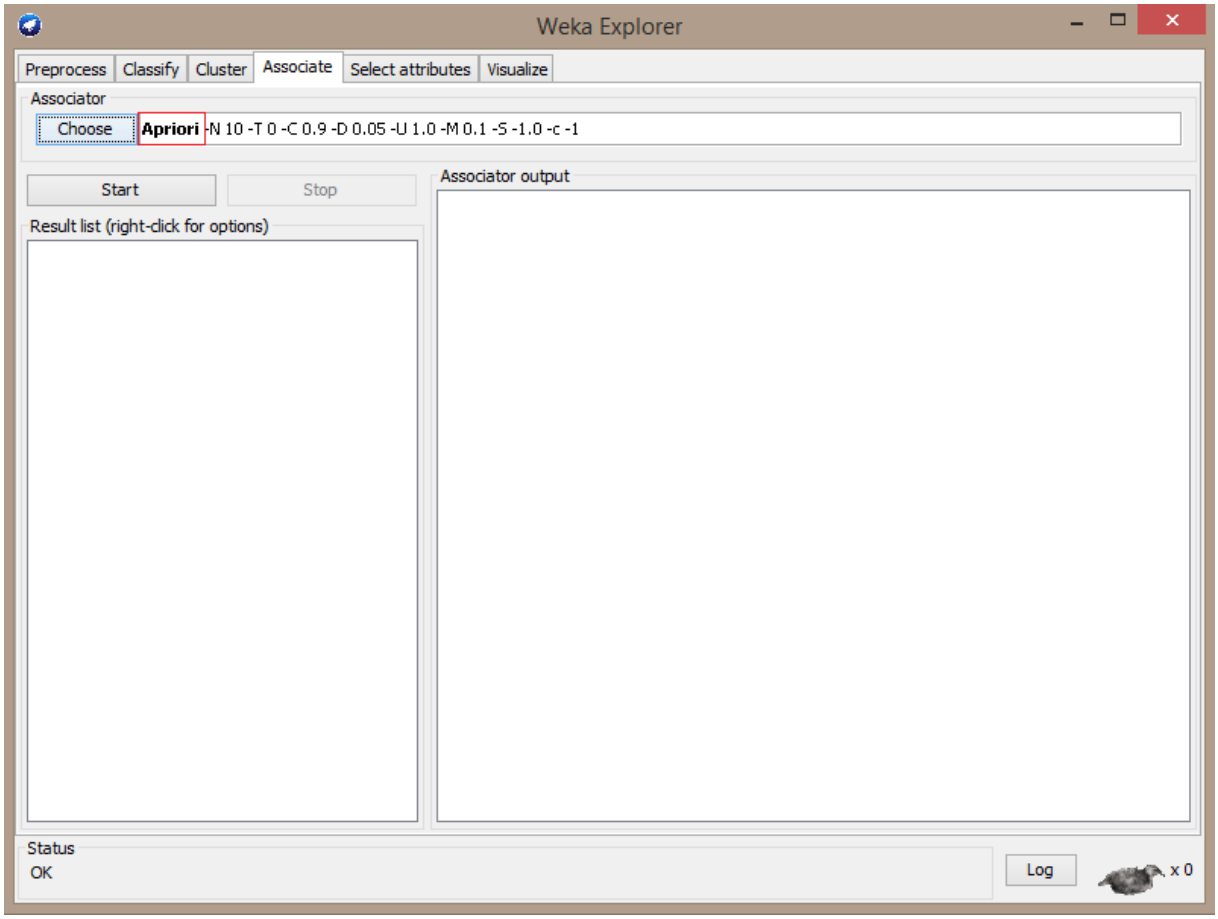

Figura A.5 – Weka Explorer com aba *Associate* Selecionada

Fonte: Frank, 2013.

Inicialmente o ***Weka Explorer*** será encontrada desta maneira. Para calibrar as regras para geração dos resultados o campo ***Apriori*** deve ser selecionado. Nele pode-se realizar ajustes como:

Suporte Mínimo (***lowerBoundMinSupport***), onde 0.1 corresponde a 10%, ou seja, só serão geradas regras que obtiverem o mínimo de 10% de suporte;

Tipo de métrica (***metricType***): Confiança (*Confidence*) e Interesse (*Lift*). Pode-se selecionar *Leverage* e *Conviction* também, porém não serão utilizadas;

Mínimo da métrica (***minMetric***): valor mínimo exigido para a métrica selecionada.

Número de regras (***numRules***): quantidade de regras que serão geradas na análise.

Estas opções podem ser visualizadas na Figura 6.

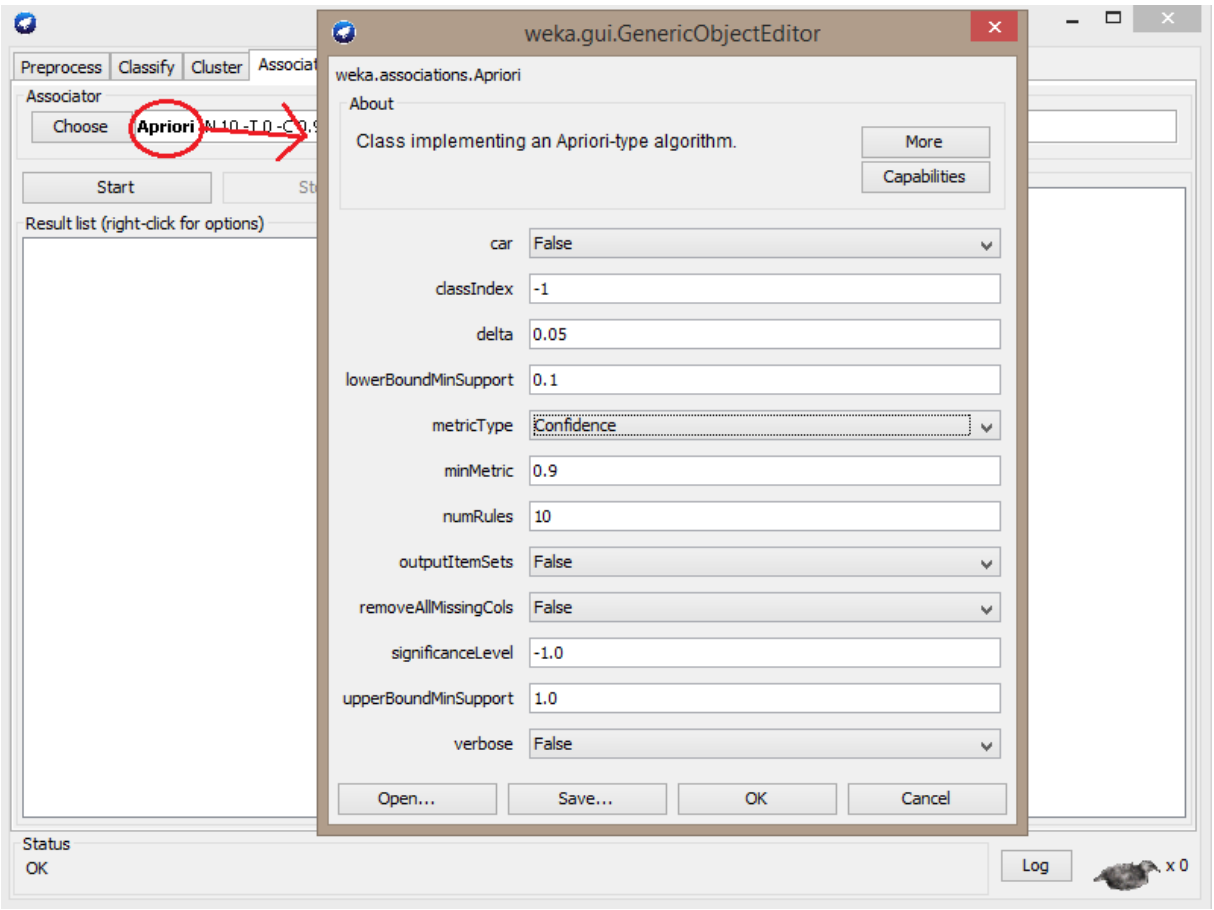

Figura A.6 – Weka Explorer Selecionado para Ajustes

Fonte: Frank, 2013.

No ***weka.gui.GenericObjectEditor*** existem outros ajustes extras visíveis na Figura 6, porém no trabalho editou-se somente os campos, ***lowerBoundMinSupport***, ***metricType***, ***minMetric*** e ***numRules***.

Realizados os ajustes no ***weka.gui.GenericObjectEditor***, deve-se clicar em **OK**. Feito isso a tela atual voltará a ser o ***Weka Explorer***. No ***Weka Explorer*** deve-se clicar em ***Start***, com a seleção deste botão, as regras serão geradas baseadas nos ajustes efetuados no ***weka.gui.GenericObjectEditor***. Na Figura 7 pode ser visualizado um exemplo de saída do ***Weka Explorer***.

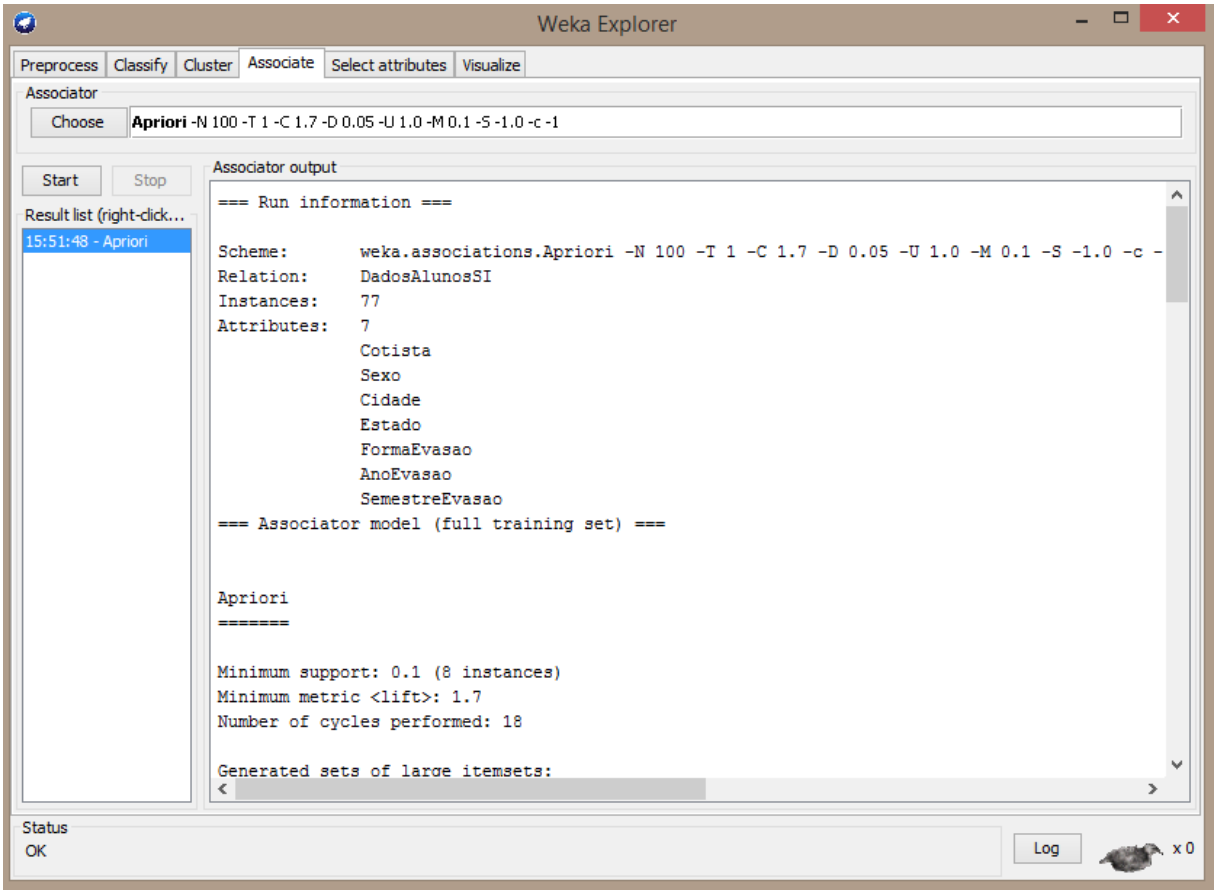

Figura A.7 – Exemplo de Resultado do Weka Explorer

Fonte: Frank, 2013.

É possível e recomendado copiar o resultado para um editor de texto, pois assim, se torna possível realizar uma análise posterior das regras. Ao fechar-se a aplicação as informações são perdidas, por isso recomenda-se salvar o resultado em um editor de texto ou outro tipo de documento genérico.