Apoio à análise de alunos em situação de acompanhamento acadêmico utilizando mineração de dados

Gisele Germano da Silva

JUIZ DE FORA

DEZEMBRO, 2016

Apoio à análise de alunos em situação de acompanhamento acadêmico utilizando mineração de dados

Gisele Germano da Silva

Universidade Federal de Juiz de Fora
Instituto de Ciências Exatas
Departamento de Ciência da Computação
Bacharelado em Ciência da Computação

Orientador: Eduardo Barrére

JUIZ DE FORA
DEZEMBRO, 2016

Apoio à análise de alunos em situação de acompanhamento acadêmico utilizando mineração de dados

Gisele Germano da Silva

MONOGRAFIA SUBMETIDA AO CORPO DOCENTE DO INSTITUTO DE CIÊNCIAS EXATAS DA UNIVERSIDADE FEDERAL DE JUIZ DE FORA, COMO PARTE INTEGRANTE DOS REQUISITOS NECESSÁRIOS PARA A OBTENÇÃO DO GRAU DE BACHAREL EM CIÊNCIA DA COMPUTAÇÃO.

Aprovada por:

Eduardo Barrére
D. Sc. em Engenharia de Sistemas e Computação, COPPE/UFRJ

Luciana Conceição Dias Campos
D. Sc. em Engenharia Elétrica, PUC - Rio

Daves Márcio Silva Martins
M. Sc. em Sistemas Computacionais, UFRJ

JUIZ DE FORA
13 DE DEZEMBRO, 2016

Resumo

A Mineração de Dados possibilita obter resultados relevantes dado um determinado conjunto de informações, combinando-as e organizando-as de acordo com os algoritmos utilizados para tal. O Instituto de Ciências Exatas da UFJF apresenta cursos em que o número de alunos com baixo desempenho acadêmico é bastante significativo. Este Trabalho de Conclusão de Curso possibilitou a identificação de características relevantes referentes ao desempenho acadêmico, utilizando como base registros de alunos dos cursos de Ciência da Computação noturno e Sistemas de Informação. As características identificadas a partir das técnicas de mineração escolhidas forneceram detalhes que podem servir, por exemplo, para auxiliar tomadas de decisão por coordenadores de curso. Tais detalhes foram obtidos através da execução dos algoritmos J48 e Apriori. A partir do histórico dos alunos obtido pelo Sistema Integrado de Gestão Acadêmica e, via iNtegra, é possível identificar reprovações, índice de rendimento acadêmico, disciplinas cursadas, disciplinas que ainda precisam ser cursadas, notas, ano e semestre em que determinada disciplina foi cursada, créditos concluídos, porém não há dados que possam ser obtidos de forma direta a partir da combinação dos itens citados acima. Atualmente, por exemplo, um coordenador de curso não consegue obter através dos sistemas citados, quais são todas as disciplinas em que seus alunos mais se reprovaram ou aprovaram em um determinado período de tempo.

Palavras-chave: Mineração de Dados, Apoio à tomada de decisão, Apriori, J48

Abstract

Data Mining allows to obtain relevant results given a set of information, combining and organizing them according to the algorithms used for it. The Institute of Exact Sciences of UFJF presents courses in which the number of students with low academic performance is quite significant. This Work of Conclusion of Course made possible the identification of relevant characteristics related to academic performance, using as base the records of students of the courses of Computing Science and Information Systems. The characteristics identified from the mining techniques chosen, provided details that can be used to support for decision-making by course coordinators, for example. These details were obtained through the execution of the J48 and Apriori algorithms. From the students' history obtained by the Integrated Academic Management System, and through iNtegra, it is possible to identify failures, academic achievement index, courses taken, subjects still to be studied, grades, year and semester in which a given course was taken, fineshed credits, but there is no data that can be obtained directly from the combination of the items mentioned above. Currently, for example, a course coordinator can not obtain through the systems mentioned, which are all the disciplines in which his students most fail or have passed in a certain period of time.

Keywords: Data Mining, Support for decision-making, Apriori, J48

Agradecimentos

Primeiramente a Deus, por me dar forças e saúde.

Ao meu pai, Jorge Gomes, pelo imenso incentivo e apoio durante todas as etapas da minha vida. Sua motivação é e sempre será o meu guia.

A minha mãe, Georgina Germano, por todo carinho, amor e orações.

Ao meu namorado, Wellerson Novaes, pela torcida e enorme paciência, principalmente nesta fase final do curso.

Aos amigos feitos em Juiz de Fora por amenizarem a saudade da família e de casa.

Aos amigos de Volta Redonda que mesmo longe se mantiveram presentes em minha vida.

Aos professores que fizeram parte desta longa caminhada, em especial ao meu orientador, professor Eduardo Barrére, por todo suporte e atenção.

"Em qualquer direção que percorras a alma,

nunca tropeçarás em seus limites."

Sócrates

Sumário

Lista de Figuras

Lista de Abreviações

DCC	Departamento de Ciência da Computação
UFJF	Universidade Federal de Juiz de Fora
ICE	Instituto de Ciências Exatas
RAG	Regulamento Acadêmico da Graduação
SIGA	Sistema Integrado de Gestão Acadêmica
CGCO	Centro de Gestão do Conhecimento Organizacional
CEI	Coeficiente de evolução inicial da discente e do discente no curso
CET	Coeficiente de evolução trissemestral da discente ou do discente no curso
SI	Sistemas de Informação

1 Introdução

1.1 Apresentação do Tema

Na UFJF, o Sistema Integrado de Gestão Acadêmica (SIGA) e o iNtegra disponibilizam para os alunos e para os coordenadores de curso (com níveis de acesso diferenciados) dados referentes à formação acadêmica em andamento ou já concluída.

O SIGA comtempla toda a comunidade da UFJF, já o iNtegra é direcionado apenas ao Instituto de Ciências Exatas, que neste caso foi o ambiente de estudo em questão.

Utilizando estes sistemas é possível obter dados correspondentes à graduação de um determinado aluno a partir de seu histórico completo, porém acredita-se que seja possível extrair informações além daquelas que já estão disponibilizadas.

Os dados de entrada para realização das técnicas de mineração foram registros de alunos dos cursos de Ciência da Computação noturno e Sistemas de Informação. Os registros utilizados foram organizados de forma a identificar resultados que possibilitem que novas interpretações a respeito do desempenho acadêmico possam estar ao alcance de coordenadores de curso e chefes de departamento, por exemplo.

Duas importantes variáveis para realização deste trabalho foram o CEI (Coeficiente de evolução inicial da discente e do discente no curso) e o CET (Coeficiente de evolução trissemestral da discente ou do discente no curso), ambos descritos no Art 1º - incisos VIII e IX do Regulamento Acadêmico da Graduação (RAG) aprovado em janeiro de 2016.

1.2 Problema

O principal questionamento proposto neste Trabalho de Conclusão de Curso está em saber como a utilização de registros acadêmicos de alunos dos cursos de Ciência da Computação noturno e Sistemas de Informação, em um determinado período de tempo,

pode contribuir para que novas interpretações sejam estimuladas a partir de um conjunto de relações envolvendo estes registros.

1.3 Justificativa

Os registros referentes aos alunos de graduação da UFJF já se encontram disponíveis via SIGA e/ou iNtegra e podem também ser visualizados como relatórios. O que motivou o estudo aqui proposto foi a forma como esses registros estão dispostos, pois há muita informação não detalhada ou agrupada de forma pouco eficiente que pode ser trabalhada para gerar informação útil aos interessados.

A Mineração de Dados tem apresentado benefícios em diversos cenários com grande volume de dados armazenados. Espera-se que os resultados aqui identificados possam influenciar as rotinas de trabalho dos cursos analisados, e possibilitar o apoio aos discentes de forma mais eficaz e condizente com a realidade de cada curso. A tomada de decisão nestes casos pode influenciar em ajustes de matrículas, quebras de pré-requisitos, acompanhamento acadêmico, entre outras demandas.

1.4 Hipótese

A mineração dos dados a ser estudados aparesentará resultados a partir dos atributos selecionados. Estes resultados permitirão uma análise mais abrangente no que diz respeito ao acompanhamento acadêmico por parte das coordenações de curso. A partir daí as novas informações obtidas através do tratamento dos dados permitirão um estreito relacionamento com o discente.

1.5 Objetivos

1.5.1 Objetivo geral

Mineração nos dados de alunos dos cursos de Ciência da Computação noturno e Sistemas de informação da UFJF para possibilitar apoio à tomada de decisão.

1.5.2 Objetivos específicos

- Identificação dos grupos de alunos a ser estudados
- Definição de técnicas e ferramentas de mineração aplicadas
- Identificação de informações que possam ser relevantes para as coordenações de curso
- Realização de técnicas de mineração de dados com as informações coletadas
- Visualização dos resultados
- Exposição dos resultados de forma que auxiliem tomadas de decisão

1.6 Organização do texto

O trabalho está organizado de forma a apresentar no capítulo 2 a fundamentação teórica necessária para entender o desenvolvimento realizado. Em seguida, no capítulo 3, é descrita a metodologia utilizada para se chegar às interpretações finais obtidas. No capítulo 4 é apresentado com detalhes o processo realizado para que pudesse ser gerado conhecimento relevante a partir de um conjunto de dados iniciais. No capítulo 5, o estudo de caso apresenta todas as técnicas aplicadas e a visualização das interpretações realizadas mediante às minerações. Por fim, no capítulo 6, está a conclusão obtida diante de todo o cenário estudado.

2 Revisão Bibliográfica

Os bancos de dados, em geral, são responsáveis por gerenciar um grande volume de informações de forma segura e eficiente. De acordo com [2], o valor dos dados armazenados está tipicamente ligado à capacidade de se extrair conhecimento de mais alto nível a partir deles, ou seja, informação útil que sirva para apoio à tomada de decisão, e/ou para exploração e melhor entendimento do fenômeno gerador de dados.

Diante disso, é necessário garantir a integridade das informações armazenadas para que posteriormente, quando for necessário gerar algum tipo de conhecimento a partir delas, os dados estejam condizentes com a realidade.

Segundo [13], a obtenção das informações embutidas nas bases de dados não é adquirida de forma direta devido à falta de ferramentas apropriadas para a sua extração e está além da capacidade do ser humano analisar tamanha quantidade de dados e extrair relações significativas entre eles.

A descoberta de novas informações e a identificação de padrões a partir de um conjunto de dados analisado podem ser geradas de diversas formas. Neste trabalho, o foco está na Mineração de Dados. De acordo com [18] a Mineração de Dados trata-se de uma etapa em KDD (*Knowledge Discovery in Databases*[1]) responsável pela seleção dos métodos a ser utilizados para localizar padrões nos dados, seguida da efetiva busca por padrões de interesse numa forma particular de representação, juntamente com a busca pelo melhor ajuste dos parâmetros do algoritmo para a tarefa em questão, ou seja, de maneira mais simples [19] afirma que o objetivo da mineração é, portanto, prover um método automático para descobrir padrões em dados, sem a tendenciosidade e a limitação de uma análise baseada meramente na intuição humana.

Este procedimento é realizado em etapas, são elas: definição do problema, preparação dos dados, exploração dos dados, criação dos modelos e análises. De acordo com [14] o processo consiste primeiramente no entendimento do domínio da aplicação, considerando aspectos como os objetivos dessa aplicação e as fontes de dados (base de dados

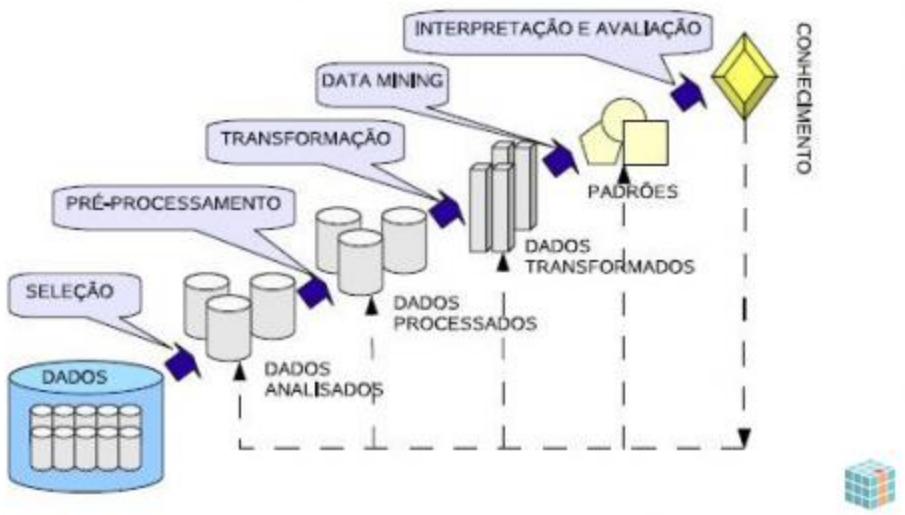
[1]Descoberta de Conhecimento em Banco de Dados

da qual se pretende extrair conhecimento). Em seguida, é realizada uma seleção de dados a partir dessas fontes, de acordo com os objetivos do processo. Os conjuntos de dados resultantes dessa seleção são, então, pré-processados, ou seja, recebem um tratamento para poderem ser submetidos aos métodos e ferramentas na etapa de extração de padrões.

Segundo [20], o propósito do pré-processamento é transformar os dados de entrada brutos em um formato apropriado para análises subsequentes. Os passos envolvidos no pré-processamento de dados incluem a fusão de dados de múltiplas fontes, a limpeza dos dados para remoção de ruídos, observações duplicadas, a seleção de registros e características que sejam relevantes à tarefa de mineração de dados.

Por causa das muitas formas através das quais os dados podem ser coletados e armazenados, o pré-processamento de dados talvez seja o passo mais trabalhoso e demorado no processo geral de descoberta de conhecimento. Segundo [5], algumas partes da fase de pré-processamento implicam na necessidade de se conhecer o domínio analisado para corrigir algumas dificuldades encontradas tais como, inconsistências, poluição e atributos duplicados e redundantes. Outras partes porém, já não levam em consideração o conhecimento do domínio, pois os métodos presentes nas técnicas de mineração se encarregam de resolver o problema automaticamente, isto ocorre, portanto, com os valores em branco e com os dados com classes desbalanceadas.

A Figura 2.1, relaciona todas as etapas envolvidas no processo de mineração:



Fonte: http://imasters.com.br/artigo/10229/tecnologia/mineracao-de-dados-e-web-semantica/?trace=1519021197&source=single

Figura 2.1: Fases do processo de mineração

Tendo os dados em condições de serem utilizados nas técnicas de mineração, é necessário definir quais algoritmos serão executados para explorá-los, pois [10] apud [6] afirmam que não há uma técnica que resolva todos os problemas de mineração de dados. Diferentes métodos servem para diferentes propósitos; cada método oferece suas vantagens e suas desvantagens. A familiaridade com as técnicas é necessária para facilitar a escolha de uma delas de acordo com os problemas apresentados.

Dentre as técnicas disponíveis, duas se destacam perante o problema proposto neste trabalho. Através da Classificação, de acordo com [10], constrói-se um modelo de algum tipo que possa ser aplicado a dados não classificados a fim de categorizá-los em classes, o objetivo é descobrir um relacionamento entre um atributo meta (cujo valor é previsto) e um conjunto de atributos de previsão, utilizando para isto árvores de decisão, que conforme visto em [10], hierarquizam os dados baseados em estágios de decisão (nós) e na separação de classes e subconjuntos. Resumidamente, segundo [17], a Classificação é a técnica recomendada para examinar uma certa característica presente nos dados e atribuí-la a uma classe previamente definida. Já a técnica de Associação, segundo [10], é utilizada para determinar os itens que tendem a ser adquiridos juntos em uma mesma transação. Em [16], a regra de associação é caracterizada como uma expressão da forma

A $\rightarrow B$, onde A e B são conjuntos de itens que ocorrem juntos em uma transação.

A toda regra de associação A $\rightarrow B$ associa-se um grau de confiança :

$$conf(A \rightarrow B) = \frac{\text{número de transações que suportam } (A \cup B)}{\text{número de transações que suportam } A}$$

Figura 2.2: Grau de confiança (Apriori)

O algoritmo leva em consideração também o valor de suporte, que corresponde à fração de transações que contêm A e B. Suporte e confiança são utilizados, portanto, como filtros para diminuir o número de regras geradas, gerando apenas regras de melhor qualidade.

De acordo com [18], o próprio algoritmo utilizado nesta técnica elege os atributos determinantes (lado esquerdo da regra) e os atributos resultantes (lado direito) na tarefa revelando associações entre valores dos atributos, tendo sua ênfase no compromisso entre precisão e cobertura.

No desenvolvimento deste trabalho, dois algoritmos foram escolhidos para execução das técnicas de Classificação e Associação. São eles: o algoritmo J48 e o algoritmo Apriori, repectivamente.

O J48, segundo [12] apud [3], surgiu da necessidade de recodificar o algoritmo C4.5 que, originalmente, é escrito na linguagem C, para a linguagem Java. Caracteriza-se por criar modelos de decisão em formato de árvore. De acordo com [12] a árvore de decisão gerada é baseada em um conjunto de dados de treinamento, sendo este modelo usado para classificar as instâncias no conjunto de teste. Primeiramente, induz a árvore de decisão para em seguida gerar classificações.

Segundo [1], o modelo da árvore é construído mediante o conjunto de treino ou alguma das outras opções fornecidas pelo algoritmo (*use training set, supplied test set, cross-validation, percentage split*). O J48 gera árvores de decisão, em que cada nó da árvore avalia a existência ou significância de cada atributo individual. As árvores de decisão são construídas do topo para a base, através da escolha do atributo mais apropriado para cada situação. Uma vez escolhido o atributo, os dados de treino são divididos em sub-grupos, correspondendo aos diferentes valores dos atributos e o processo

é repetido para cada sub-grupo até que uma grande parte dos atributos em cada sub-grupo pertençam a uma única classe. A indução por árvore de decisão é um algoritmo que habitualmente aprende um conjunto de regras com elevada acuidade.

Em ferramentas como o Weka [2], é possível visualizar graficamente a árvore gerada. Proposto por [10] apud [15], o J48 é considerado o algoritmo que apresenta os melhores resultados na montagem de árvore de decisão.

Já o Apriori, de acordo com [8] apud [11], recebe este nome devido ao fato de utilizar conhecimento prévio (*prior knowledge*) das propriedades de composição e frequência de itens.

Segundo [9], o Apriori é considerado o algoritmo mais utilizado quando se deseja realizar regras de associação. Inicialmente são analisados todos os registros do conjunto de dados em estudo, e através de sub-rotinas do próprio algoritmo, os itens que não são frequentes passam a ser desconsiderados e as regras de associação são montadas a partir dos itens candidatos. Seu objetivo principal é identificar as relações existentes entre os registros na medida em que são separados pela execução do algoritmo.

De acordo com [7] o algoritmo se baseia em duas funções: a Apriori-gen, que gera os itens candidatos levando em consideração o valor do suporte (percentual indicado que fornece a ocorrência mínima de determinada combinação de itens na base de dados), e a função chamada Genrules, que gera as regras de associação considerando o parâmetro de confiança informado.

Segundo [4], estas funções consistem em geração e poda. Na geração, é feita uma varredura sobre o arquivo, a fim de gerar todos os conjuntos de combinações de valores de colunas que aparecem no arquivo. Já na poda, são considerados apenas aqueles conjuntos que aparecem no arquivo com uma frequência não menor que um valor mínimo pré-fixado, são os chamados grandes conjuntos.

A execução do algoritmo Apriori pode ser detalhada em 7 passos:

1. Entrada: coleção de dados associados, suporte mínimo, confiança mínima.
2. Considerar K = 1 para criação de K-itemsets.
3. Analisar os dados associados e criar uma tabela de K-itemsets com suporte acima do

[2]Waikato Environment for Knowledge Analysis. http://www.cs.waikato.ac.nz/ml/weka/

suporte mínimo.

4. Criar com os itemsets filtrados um conjunto de candidatos a (K + 1) itemsets.
5. Usar propriedades do Apriori para eliminar itemsets infrequentes.
6. Repetir desde o passo 3 até que o conjunto gerado seja vazio.
7. Listar regras de associação (com permutações) e aplicar limite de confiança.

3 Metodologia

O desenvolvimento deste trabalho teve como apoio dados referentes aos alunos de Ciência da Computação noturno e Sistemas de Informação, ambos cursos da UFJF, durante os períodos de 2013.1, 2013.3, 2014.1, 2014.3, 2015.1 e 2015.3.

A abordagem da pesquisa caracterizou-se como qualitativa, pois foram adotados métodos de exploração e compreensão para possibilitar a interpretação dos resultados obtidos.

O objetivo foi demonstrar como a exploração de relações entre os dados já conhecidos pode desencadear novas informações que sejam relevantes para tomada de decisão.

Uma entrevista com os coordenadores de curso do ICE, através de um questionário, permitiu levantar quais situações referentes a combinações entre os dados acadêmicos podem ser úteis na rotina das coordenações.

O estudo realizado para as situações abordadas pelos coordenadores que responderam ao questionário concentrou-se na utilização na técnicas de classificação e associação. A ferramenta utilizada para execução de tais técnicas foi o Weka, e a escolha desta ferramenta se deu principalmente por ser um software livre e pelo desempenho apresentado por seus algoritmos.

Para as técnicas decritas foram utilizados os algoritmos J48 e Apriori, respectivamente. Os gráficos apresentados foram feitos através do Excel mediante interpretação dos resultados obtidos durante aplicação das técnicas de mineração.

Em anexo a este trabalho consta um manual com as instruções necessárias para que todas as minerações realizadas no Estudo de Caso possam ser refeitas.

4 O processo

4.1 Identificação do grupo de alunos a ser estudado

Como a proposta inicial era promover novas interpretações a partir de um conjunto de dados já existentes para auxílio de acompanhamento acadêmico, os cursos de graduação do ICE se tornaram a maior motivação para a realização deste trabalho, visto que os índices de reprovação são consideralvemente altos nestes cursos, e não há atualmente medidas que auxiliem a coordenação de cada curso a atuar de forma eficaz com os alunos em situação de baixo rendimento acadêmico.

Dentre os cursos ofertados neste instituto, para os cursos de Ciência da Computação noturno e Sistemas de Informação já se tinha um conhecimento mais aprofundado acerca das dificuldades, problemas e necesidades existentes. Diante disto, estes foram, portanto, os cursos escolhidos para o desenvolvimento deste Trabalho de Conclusão de Curso, visto que era um diferencial que permitia dar início as atividades envolvidas, possibilitando ao final a obtenção de conhecimento de novas informações a partir dos registros utilizados como base.

4.2 Abordagens e ferramentas utilizadas

A base de dados utilizada para estudo foi adquirida em formato .csv e extraída do iNtegra. Inicialmente, foi importada para um gerenciador de banco de dados. Para isto, foi escolhido o phpMyAdmin, por ser um software livre e apresentar uma série de recursos relevantes para o andamento deste trabalho. Após a importação foi necessário um tratamento dos dados armazenados, pois a ocorrência de campos com registros vazios e inconsistentes poderiam comprometer os resultados encontrados durante a execução das técnicas de mineração.

Como o volume de dados é extenso, esta etapa demandou bastante estudo e tempo para que os dados estivessem em condições de gerar conhecimentos relevantes além dos

já conhecidos pela comunidade acadêmica.

Para cada aluno registrado na base foram calculados os índices de CEI e CET a fim de caracterizar quais deles se encontravam ou não em situação de acompanhamento acadêmico, o que é a principal motivação deste projeto de conclusão de curso. O cálculo dos índices foi realizado conforme descrição presente no RAG.

O Weka é um software que a partir dos padrões encontrados é capaz de gerar hipóteses para soluções e análise dos dados em questão, e foi a ferramenta escolhida para execução das técnicas de mineração aqui apresentadas. Dentre seus algoritmos disponíveis, foram utilizados o J48 e o Apriori. Embora ambos tenham sido utilizados para responder todas as situações detalhadas neste trabalho, em determinadas situações um se mostrou mais eficiente do que o outro possibilitando assim, interpretações mais exatas que estão detalhadas na seção do Estudo de Caso.

Para a execução dos algoritmos foi preciso separar a base de dados por curso, pois como cada curso possui sua especificidade, analisá-los de forma conjunta não atenderia a proposta inicial deste projeto que é auxiliar à tomada de decisão por parte dos coordenadores de curso. Em determinados casos, trabalhou-se apenas com as instâncias referentes aos alunos caracterizados em situação de acompanhamento acadêmico.

4.3 Coordenações dos cursos de graduação

A partir das técnicas e ferramentas de mineração é possível identificar diversas hipóteses para determinados resultados, porém sabe-se que nem todo resultado obtido é relevante em determinadas situações. Sendo assim, tornou-se essencial a participação dos coordenadores de curso antes das etapas de mineração. O objetivo desta participação foi identificar de forma mais eficiente quais informações a partir dos dados analisados poderiam de fato contribuir para o apoio à tomada de decisão em relação aos alunos do ICE.

Foi elaborado um questionário destinado a todos os coordenadores dos cursos deste instituto. No questionário foi solicitado que eles respondessem o grau de relevância de possíveis situações que poderiam ser extraídas do processo de mineração, deixando-os a vontade também para sugerir situações não incluídas aos questionamentos e que julgassem

ser de grande importância para decisões em suas coordenações.

A Figura 4.1 apresenta o questionário enviado por email aos coordenadores dos cursos do ICE e a quantidade de respostas obtidas em cada item.

Questionamento	Irrelevante	Pouco Relevante	Relevante	Muito Relevante
Para um determinado período de ingresso, quantos alunos apresentam CEI ou CET insuficientes			3	3
Para os alunos que apresentam CEI ou CET insuficientes, quais disciplinas eles foram aprovados		1	2	3
Para os alunos que apresentam CEI ou CET insuficientes, quais disciplinas eles foram reprovados			1	5
Qual a relação entre alunos ativos e CEI ou CET insuficiente		1	4	1
Qual a relação entre alunos ativos e CEI ou CET suficiente		2	4	
Para os alunos com IRA entre X e Y, quantos apresentam CEI ou CET insuficiente			3	3
Quantos alunos com CEI ou CET insuficiente estão aprovados/reprovados na disciplina X			5	1
Se os alunos foram reprovados nas disciplinas X e/ou Y e/ou Z do 1º período, qual a chance dele ter um CEI insuficiente			3	3
Se os alunos foram reprovados nas disciplinas X e/ou Y e/ou Z do 1º ou 2º período, qual a chance dele ter um CET insuficiente			4	2
Quantas vezes um aluno com CEI ou CET insuficiente faz a disciplina X até ser aprovado		1	2	3
Qual a carga horária média de um aluno com CEI ou CET insuficiente			3	3
Qual a possibilidade de um aluno com CEI insuficiente trancar o curso		2	3	1
Por quantos períodos, em média, um aluno com CET insuficiente permanece no curso			1	5
Qual a possibilidade de um aluno que nunca teve CEI ou CET insuficiente passar a ter	1	2	1	2
Qual a possibilidade de um aluno que teve CEI e/ou CET insuficiente passar a ter CET suficiente e em qual período do curso			2	4

Figura 4.1: Respostas dos coordenadores dos cursos de graduação do ICE

Obteve-se retorno dos coordenadores dos cursos de: Ciência da Computação (diurno e noturno), Estatística, Bacharelado em Física (diurno), Bacharelado em Química (diurno) e Sistemas de Informação, totalizando 6 respostas para cada questão do questionário. Dentre as sugestões apresentadas estavam:

- O relatório "Para os alunos que apresentam CEI ou CET insuficientes, quais disciplinas eles foram reprovados" deveria apresentar a quantidade de vezes que o aluno foi reprovado em cada disciplina
- Além de relatórios, apresentar gráficos que mostrem a evolução do aluno no curso também auxilia na análise
- Qual o perfil de reprovações em relação às disciplinas (um aluno com CET insuficiente reprova só em obrigatórias e passa em eletivas?)
- Se o aluno for reprovado em 50% ou mais dos créditos no primeiro semestre, qual a chance dele possuir CEI insuficiente?
- Quantos alunos com CEI ou CET insuficientes fazem estágio não obrigatório.

O critério utilizado para responder aos questionamentos propostos foi o maior número de votos nos quesitos relevante e muito relevante, pois representam uma necessidade comum à maioria dos entrevistados. Feita a mineração dos itens mais votados, foram escolhidas também algumas situações que embora não votadas pela maioria, podem apoiar tomada de decisão de forma ampla.

Sendo assim, foram desenvolvidas no Estudo de Caso, as seguintes situações:

- Qual a relação entre alunos ativos com CEI ou CET insuficiente/suficiente?
- Para os alunos que apresentam CEI ou CET insuficientes, quais disciplinas eles foram reprovados?
- Por quantos períodos, em média, um aluno com CET insuficiente permanece no curso?
- Quantos alunos com CEI ou CET insuficiente estão aprovados/reprovados na disciplina X?
- Para um determinado período de ingresso, quantos alunos apresentam CEI ou CET insuficientes?
- Para os alunos que apresentam CEI ou CET insuficientes, em quais disciplinas eles foram aprovados?
- Se os alunos foram reprovados nas disciplinas X e/ou Y e/ou Z do 1º ou 2º período, qual a chance dele ter um CET insuficiente?

4.4 Dificuldades encontradas

Durante todo o processo, diversas dificuldades foram encontradas, sendo as principais descritas abaixo:

- Na fase de preparação dos dados, algumas disciplinas que constam na base de dados não apresentavam o valor correspondente às suas cargas horárias. Foi preciso consultar estes valores ausentes com o CGCO, e nas grades curriculares de seus respectivos cursos.

- CEI e CET não são armazenados no histórico dos alunos, pois são valores calculados a cada semestre. Sendo assim, foi necessário calcular estes indíces de forma manual para cada aluno presente na base de dados e consequentemente identificar se eles estavam ou não em situação de acompanhamento acadêmico de acordo com os resultados obtidos para estes índices.

- Alguns alunos reprovados por frequência possuíam valores inconsistentes para suas notas na disciplina em questão. Foram encontrados notas com valor em branco, com valor igual a 0.0 e com valor igual a RI.

- Inicialmente, os dados a serem estudados estavam em arquivos no formato .csv. Para gerar o arquivo a ser lido no Weka (formato .arff), foram removidas da base de dados todas as ocorrências dos seguintes caracteres: , = ”’ *+ - % . Esta modificação se fez necessária, visto que para gerar um arquivo .arff a partir do CSVLoader do Weka, tais caracteres impossibilitam a conversão.

- Para utilização do algoritmo Apiori, por exemplo, durante a fase de pré-processamento foi aplicado a cada atributo os filtros StringToNominal e NumericToNominal. Isto tornou-se necessário pois o algoritmo não aceita valores do tipo numeric e String.

5 Estudo de caso

Buscando obter respostas relevantes a partir dos questionamentos enviados aos coordenadores de curso (Figura 4.1), foi efetuada uma mineração nos dados de alunos dos cursos de Ciência da Computação noturno e Sistemas de Informação. Os dados foram separados por curso de forma a estudar e interpretar cada realidade isoladamente. Em alguns momentos foram considerados para as minerações apenas os registros referentes aos alunos caracterizados em situação de acompanhamento acadêmico.

A base de dados continha os seguintes campos: IDALUNO (código de identificação para distinguir os alunos presentes na base e evitar que os mesmos fossem identificados caso tivessem sido utilizados números de matrícula, por exemplo), SITUACAO (como o aluno está caracterizado no curso: Ativo, Trancado, Em Análise, Cancelado e Sem matrícula), MOTIVOSAIDA (campo normalmente preenchido quando o aluno está caracterizado como Cancelado), IDCURSO (código de identificação dos cursos, neste caso foram utilizados os códigos 35A e 76A para Ciência da Computação noturno e Sistemas de Informação, respectivamente), CURSO (nome dos cursos presentes na base dados), DISCIPLINA (código de identificação de cada disciplina), NOME (nome das disciplinas), CH (carga horária de cada disciplina), TURMA (turma referente à disciplina analisada), ANO, SEMESTRE, NOTA (nota obtida pelo aluno ao final da disciplina), FREQUENCIA (frequência do aluno em uma determinada disciplina durante o ano e semestre considerados), STATUS (resultado obtido pelo aluno ao final da disciplina: Aprovado, Reprovado por nota, Reprovado por frequência ou Trancado), CEI, CET e ACOMPANHAMENTO. Os três últimos campos não constavam na base de dados original, porém foi necessário criá-los e acrescentá-los à base, pois os coeficientes de CEI e CET é que permitem identificar se um aluno está ou não em situação de acompanhamento acadêmico. Estes coeficientes foram calculados para cada IDALUNO de acordo com a descrição presente no RAG. O campo ACOMPANHAMENTO recebe, portanto, os valores de SIM ou NÃO de acordo com os resultados encontrados para os coeficientes CEI e CET. Os registros correspondiam aos períodos de: 2013.1, 2013.3, 2014.1, 2014.3, 2015.1 e 2015.3. Foram solicitados

registros referentes a estes períodos para que os valores CET pudessem ser calculados por pelo menos duas vezes.

Os algoritmos foram utilizados com parâmetros padronizados, pois a alteração de alguns deles não porporcionou mudanças significativas nos resultados das minerações. No caso do J48, que foi o algoritmo com resultados mais satisfatórios, a mudança de parâmetro alterava pouquissímas classificações incorretas.

A seguir são apresentadas as abordagens utilizadas para os questionamentos selecionados.

5.1 Relação entre alunos ativos com CEI ou CET insuficiente/suficiente

Foi questionado aos coordenadores de curso se a relação entre alunos ativos com CEI ou CET insuficiente/suficiente poderia apoiá-los de alguma forma. Dos 6 entrevistados, 4 se manifestaram considerando o questionamento como relevante.

Dados: a mineração se deu a partir dos campos IDCURSO, SITUAÇÃO (Ativos, Trancados, Cancelados, Em análise, Sem matrícula) e ACOMPANHAMENTO (SIM, NÃO) contidos na base de dados. Este último é decorrente dos valores calculados para CEI e CET de cada aluno.

J48: utilizando apenas os registros dos alunos em situação de acompanhamento acadêmico, o Weka não considerou os atributos selecionados em condições de gerar uma possível árvore de decisão, desativando assim a possibilidade de escolha do algoritmo J48.

Considerando a base completa (alunos com CEI e CET suficientes/infucientes), o algoritmo gerou uma árvore de decisão, mas sem relacionar os atributos envolvidos. Impossibilitando, portanto, visualizar a relação desejada.

```
Attributes:   3
              SITUACAO
              IDCURSO
              ACOMPANHAMENTO
Test mode:evaluate on training data

=== Classifier model (full training set) ===

J48 pruned tree
------------------
: Ativo (1235.0/420.0)

Number of Leaves  :     1

Size of the tree :      1
```

Figura 5.1: Resultado da execução do J48 para a relação 5.1

Apriori: este algoritmo, por se tratar de uma técnica de associação, permitiu visualizar os atributos de forma relacionada possibilitando a identificação do quantitativo referente a cada SITUAÇÃO(Ativos, Trancados, Cancelados, Em análise, Sem matrícula).

Para a configuração dos parâmetros do algoritmo foram utilizados os valores padronizados que podem ser vistos no Anexo I, tabela I.1.

O algoritmo apresentou resultados relevantes em relação aos dados selecionados, conforme Figura 5.2, porém, como pode-se observar, o quantitativo apresentado nos dois gráficos não condiz com a quantidade de alunos dos cursos no período de tempo considerado. Os números obtidos representam a quantidade de registros existentes para cada IDALUNO e não a quantidade exata de alunos presentes na base. Isto ocorreu, pois na base de dados os registros referentes a cada IDALUNO normalmente aparecem mais de uma vez, visto que correspondem às disciplinas cursadas por ele, o mesmo ocorre para os valores dos campos SITUAÇÃO e ACOMPANHAMENTO. Para cada registro do IDALUNO a situação do aluno e o valor de SIM ou NÃO referente à situação de acompanhamento acadêmico são repetidos.

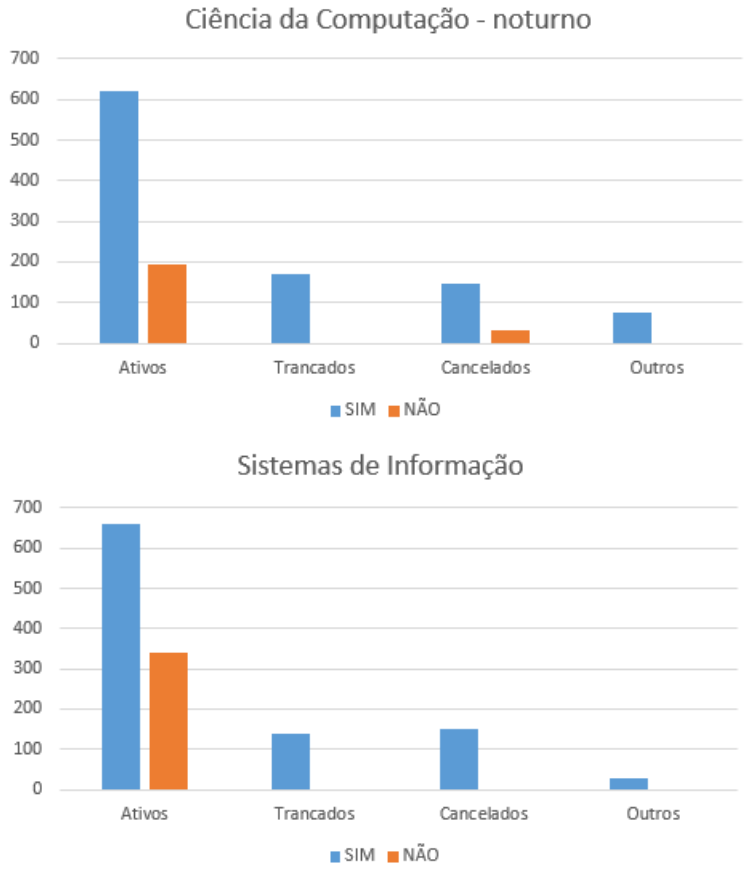


Figura 5.2: Situação x Acompanhamento acadêmico

Sendo assim, foi necessária a realização de outro procedimento para que a questão pudesse ser respondida de forma mais precisa. O procedimento adotado foi apenas uma consulta SQL, visto que este era um resultado que poderia ser encontrado sem utilização de técnicas de mineração. Para o curso de Ciência da Computação noturno, as consultas realizadas foram:

```
SELECT COUNT(DISTINCT IDALUNO) FROM tblbase where IDCURSO='35A'
 AND ACOMPANHAMENTO='SIM'

SELECT COUNT(DISTINCT IDALUNO) FROM tblbase where IDCURSO='35A' AND
ACOMPANHAMENTO='SIM' AND SITUAÇÃO= 'ATIVO'
```

As consultas acima são referentes aos alunos com CEI e CET insuficientes, ou seja, aqueles que estão em acompanhamento acadêmico. Para analisar os caracterizados como suficientes basta alterar nas consultas ACOMPANHAMENTO='SIM' para ACOM-

PANHAMENTO='NÃO'.

A primeira consulta retornou um valor de 63 alunos em acompanhamento acadêmico sendo que, deste total, 31 encontram-se ativos no curso. Já o total encontrado para os alunos fora da situação de acompanhamento foi de 9 alunos, sendo que 8 deles estão ativos. Os resultados obtidos permitem identificar que para o curso analisado, existem registros de um total de 72 alunos e que a maior parte deles está, portanto, em situação de acompanhamento acadêmico.

Para o curso de Sistemas de Informação foram realizadas as mesmas consultas, alterando apenas os valores do campo IDCURSO para 76A que é o código utilizado para representar o curso de SI. Neste caso, as consultas restornaram um valor de 73 alunos com CEI e CET insuficientes, onde 39 estão ativos. Já dos 15 alunos encontrados com CEI e CET suficientes, todos estão caracterizados como ativos. Os resultados obtidos permitem identificar que para o curso de Sistemas de Informação a base de dados continha, então, registros de um total de 88 alunos.

No período de tempo presente na base de dados, o curso de Sistemas de Informação apresentou um quantitativo de alunos maior do que o curso de Ciência da Computação noturno, mas diante das consultas observadas é possível identificar que em ambos os cursos, a quantidade de alunos caracterizados em situação de acompanhamento acadêmico é bastante significativa.

A Figura 5.3 ilustra os resultados obtidos:

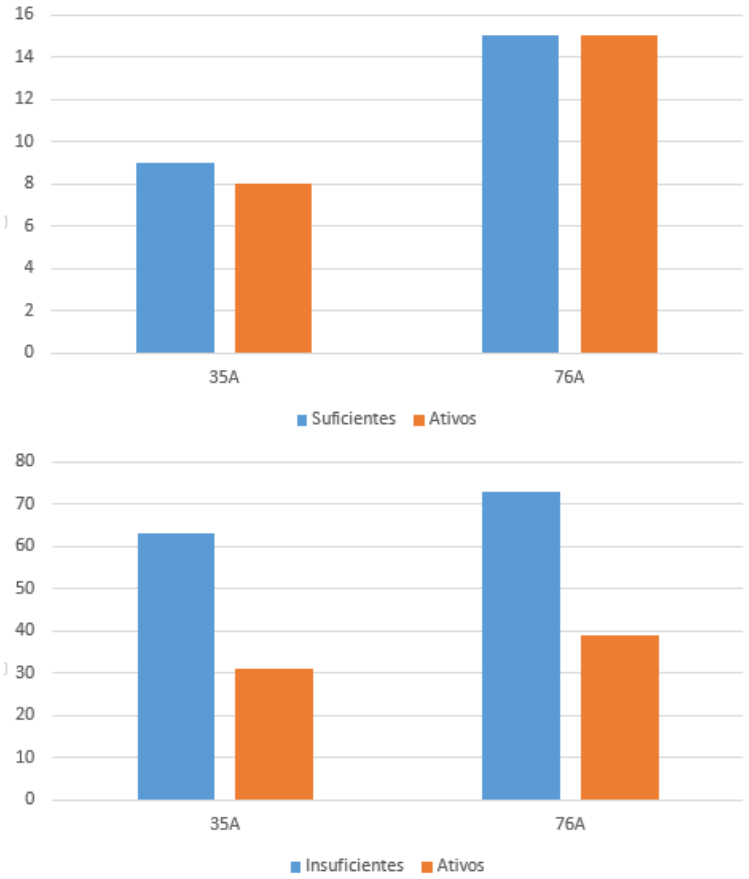


Figura 5.3: Relação entre alunos ativos com CEI e CET suficientes/insuficientes via consulta SQL

5.2 Maiores reprovações entre os alunos com CEI ou CET insuficientes

Dentre os 6 entrevistados, tal questionamento recebeu 5 votos no quesito muito relevante. O conhecimento desta relação pode apoiar principalmente os ajustes de matrícula.

Dados: para a mineração foram selecionados os campos NOME (nome da disciplina), STATUS(Aprovado, Reprovado, Reprovado por nota, Reprovado por frequência e Trancado) e ACOMPANHAMENTO (SIM e NÃO).

Apriori: o algoritmo não obteve resultados satisfatórios, pois não realizou as associações entre os atributos selecionados da forma esperada. As associações concentraram-se mais no atributo STATUS, e apenas uma disciplina foi retornada. Neste caso, a utilização do algoritmo se tornou ineficiente, pois não foi possível obter resultados referentes a todas

as diciplinas presentes na base de dados. Este tipo de relação por necessitar categorizar todas as disciplinas já permite imaginar que a melhor técnica a ser utilizada seria a de classificação e não a técnica que associação, pois esta última visa apresentar resultados apenas de registros mais frequentes e não de todos os existentes.

```
Best rules found:

 1. STATUS=Aprovado 394 ==> ACOMPANHAMENTO=SIM 394    conf:(1)
 2. STATUS=Rep Nota 303 ==> ACOMPANHAMENTO=SIM 303    conf:(1)
 3. STATUS=Rep Freq 175 ==> ACOMPANHAMENTO=SIM 175    conf:(1)
 4. STATUS=Trancado 139 ==> ACOMPANHAMENTO=SIM 139    conf:(1)
 5. NOME=CÁLCULO I 111 ==> ACOMPANHAMENTO=SIM 111    conf:(1)
```

Figura 5.4: Resultado da execução do Apriori para a relação 5.2

J48: neste questionamento, o algoritmo J48 permitiu através da árvore de decisão gerada após a execução, uma melhor visualização dos dados a serem interpretados. Para a configuração dos parâmetros do algoritmo foram utilizados os valores padronizados que podem ser consultados no Anexo I, tabela I.2.

A árvore de decisão permitiu identificar o número de aprovações e reprovações em todas as disciplinas presentes na base de dados. A Figura 5.5 e a Figura 5.6 ilustram os resultados obtidos através das porcentagens calculadas referentes às reprovações para cada resultado retornado durante a mineração.

Para os dois cursos analisados foram consideradas maiores reprovações as disciplinas que mediante interpretação dos resultados obtidos pela técnica de mineração apresentaram mais que 50% de reprovação. Os gráficos permitem identificar algumas disciplinas bastante conhecidas por suas reprovações nesses cursos, confirmando assim o conhecimento prévio que já se tinha para situações como essas.

Alguns dados presentes nos gráficos podem estar distorcidos devido à pequena quantidade de alunos nos períodos mais avançados do curso. Por exemplo, a disciplina de Italiano Instrumental I apresentou, de acordo com as minerações realizadas, alto índice de reprovação, isto demonstra que se apenas 1 aluno cursou a disciplina e esse único aluno foi reprovado, a taxa de reprovação foi, portanto, de 100%.

Observações/Justificativas: observando a Figura 5.5, para a disciplina de Cálculo

Numérico, por exemplo, foram obtidas as seguintes informações durante a execução do algoritmo:

NOME = CALCULO NUMERICO: Rep Nota (9.0/5.0)

Isto significa que nesta disciplina, para o período de tempo considerado, a classificação retornou um total de 9 alunos reprovados por nota, mas o algoritmo classificou incorretamente 5 deles, ou seja, apenas 4 foram realmente reprovados por nota. Consultando a base de dados é possível identificar de fato que os 5 classificados incorretamente possuíam na verdade outros status: 1 deles foi reprovado por infrequência, 3 trancaram a disciplina e 1 foi aprovado. A interpretação precisa deste resultado exigiu uma análise em conjunto tanto das informações obtidas através da execução do algoritmo quanto da base de dados propriamente dita.

Figura 5.5: Disciplinas com maiores índices de reprovação em Ciência da Computação noturno

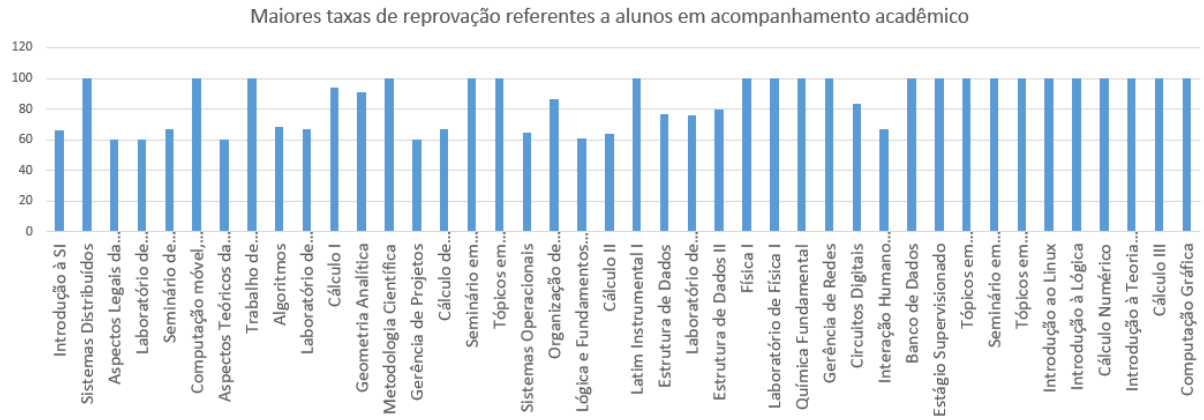


Figura 5.6: Disciplinas com maiores índices de reprovação em Sistemas de Informação

5.3 Relação entre permanência no curso e alunos com CET insuficiente

Dentre os 6 entrevistados, 5 coordenadores classificaram este questionamento como muito relevante. Tal resultado provavelmente é decorrente do número significativo de evasões que os cursos do ICE apresentam.

Dados: para se chegar ao resultado deste questionamento foram selecionados os atributos SITUAÇÃO (Ativo, Cancelado, Sem Matrícula e Trancado), ANO e SEMESTRE.

Apriori: o algoritmo conseguiu associar apenas dados referentes ao terceiro semestre do ano de 2013, mas sem identificar a situação da maior parte dos alunos durante este período (conforme Figura 5.7). Isto se deu, provavelmente, pela forma como o algoritmo funciona. Sua execução visa identificar itens mais frequentes e não apresentar resultados de forma a contemplar todos os registros presentes na base de dados

```
Best rules found:

 1. ANO=2013 203 ==> SEMESTRE=3 187    conf:(0.92)
```

Figura 5.7: Resultado da execução do Apriori para a relação 5.3

J48: este foi o algoritmo escolhido para estudo, pois a árvore de decisão gerada

possiblitou identificar relações consideráveis entre os atributos selecionados e responder ao questionamento proposto. Para a configuração dos parâmetros do algoritmo foram utilizados os valores padronizados que podem ser observados no Anexo I, tabela I.2.

A Figura 5.8 apresenta os resultados provenientes da execução do algoritmo J48. Para o semestre 1, o algoritmo classificou apenas o ano de 2015, já para o semestre 3 foram classificados registros referentes aos demais anos presentes na base. É possível identificar, por exemplo, em que determinado período uma situação acadêmica apresentou mais registros. Os valores entre parênteses representam a quantidade total/quantidade de classificações incorretas.

A mineração representada na Figura 5.8 mostra que dos 101 alunos identificados como trancados e com a maior parte dos registros em 2014-3, 50 estão realmente trancados, e os demais 51 podem não estar caracterizados como trancados ou não necessariamente com registros somente até este período.

```
J48 pruned tree
------------------

SEMESTRE = 1: 2015 (367.0/151.0)
SEMESTRE = 3
|   SITUACAO = Ativo: 2014 (384.0/206.0)
|   SITUACAO = Cancelado: 2013 (103.0/45.0)
|   SITUACAO = Sem Matricula: 2014 (50.0/21.0)
|   SITUACAO = Trancado: 2014 (101.0/51.0)
|   SITUACAO = Em Analise: 2014 (7.0)
```

Figura 5.8: Resultado da execução do J48 para a relação 5.3

Observações/Justificativas: de acordo com a base de dados, os períodos (ANO + SEMESTRE) registrados são: 2013.1, 2013.3, 2014.1, 2014.3, 2015.1 e 2015.3 totalizando, então, 6 períodos. Feita a mineração dos dados referentes a estes períodos, a interpretação dos resultados obtidos e sabendo que apenas os alunos com SITUAÇÃO= ATIVO permanecem matriculados no curso, bastou identificar o número correspondente dos alunos caracterizados como TRANCADO ou CANCELADO, visto que a base contém

informações referentes ao último período cursado destes alunos.

A Figura 5.9 apresenta as interpretações obtidas a partir da Figura 5.8. O quantitativo apresentado diz a respeito a quantidade de registros existentes para cada IDALUNO. Neste caso este quantitativo não interferiu a interpretação dos dados, pois o objetivo não era obter a quantidade exata de alunos em cada situação acadêmica presente na base dados.

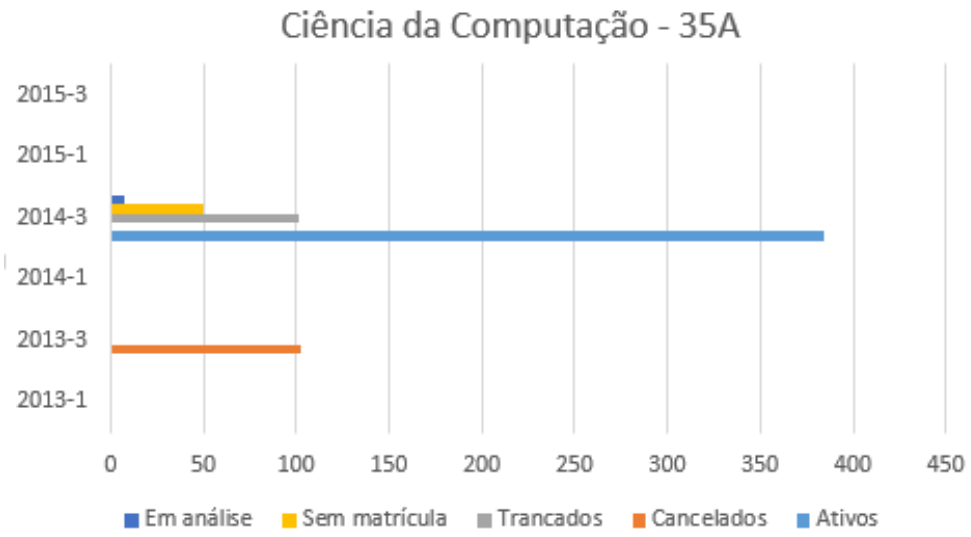


Figura 5.9: Permanência no curso de Ciência da Computação noturno

O algoritmo classificou a maior parte dos alunos trancados ao período 2014-3, conforme observado na Figura 5.8. O quantitativo obtido para este período não significa que dos resultados obtidos para esta situação acadêmica todos os trancados estão apenas com registros em 2014-3. Como os valores entre parênteses representam quantidade total/-quantidade classificada incorretamente, as classificações incorretas das demais situações podem conter registros de alunos trancados. Considerando que para a base de dados utilizada, o início das atividades acadêmicas foi em 2013-1, estes alunos permaneceram cerca de 4 períodos ativos no curso.

Em relação aos alunos caracterizados como cancelados, o algoritmo classificou os registros referentes a maior parte deles ao período 2013-3, o que corresponde a 2 períodos.

É importante ressaltar o significado das classificações incorretas realizadas pelo algoritmo para este caso, visto que há registros para alunos trancados em semestres posteriores a 2014-3, por exemplo.

O mesmo algoritmo não apresentou resultados relevantes para o curso de Sistemas de Informação. Os resultados obtidos foram apenas os presentes na Figura 5.10. Tal comportamento pode ter como motivação quantitativos pouco expressivos em cada situação acadêmica presente neste curso.

```
J48 pruned tree
------------------

SEMESTRE = 1: 2015 (360.0/162.0)
SEMESTRE = 3: 2014 (619.0/307.0)
```

Figura 5.10: Permanência no curso de Sistemas de Informação

5.4 Alunos com CEI ou CET insuficiente aprovados/-reprovados em X

No questionário enviado aos coordenadores de curso, esta situação apresentou 5 votos no quesito relevante. Com isso, foi possível identificar que analisar as disciplinas que os alunos em acompanhamento acadêmico são aprovados é tão importante quanto analisar suas reprovações

Dados: para responder este questionamento foram selecionados os atributos NOME (nome da disciplina), STATUS (Aprovado, Reprovado, Reprovado por nota, Reprovado por frequência, Trancado) e ACOMPANHAMENTO (SIM, NÃO). Deveriam existir na base de dados apenas dois tipos de regristros referentes à reprovações (Reprovado por nota e Reprovado por frequência), porém devido a existência de dados inconsistentes, foram encontrados também STATUS como Reprovado.

Apriori: o algoritmo não obteve resultados satisfatórios, pois não realizou as associações entre os atributos selecionados da forma esperada. As associações concentraram-se mais no atributo STATUS, e apenas uma disciplina foi retornada. Isto provavelmente ocorreu devido a forma como o algoritmo funciona, tendo como objetivo considerar apenas os itens mais frequentes nas transações analisadas. Neste caso, a utilização do algoritmo se

tornou ineficiente, pois não foi possível identificar o aproveitamento em cada disciplina presente na base de dados.

```
Best rules found:

 1. STATUS=Aprovado 394 ==> ACOMPANHAMENTO=SIM 394    conf:(1)
 2. STATUS=Rep Nota 303 ==> ACOMPANHAMENTO=SIM 303    conf:(1)
 3. STATUS=Rep Freq 175 ==> ACOMPANHAMENTO=SIM 175    conf:(1)
 4. STATUS=Trancado 139 ==> ACOMPANHAMENTO=SIM 139    conf:(1)
 5. NOME=CÁLCULO I 111 ==> ACOMPANHAMENTO=SIM 111    conf:(1)
```

Figura 5.11: Resultado da execução do Apriori para a relação 5.4

J48: neste questionamento, o algoritmo J48 permitiu através da árvore de decisão gerada após a execução, uma melhor visualização dos dados a ser interpretados. Para a configuração dos parâmetros do algoritmo foram utilizados os valores padronizados, conforme Anexo I tabela I.2.

Observações/ Justificativas: neste caso, os dados analisados foram somente dos alunos em situação de acompanhamento acadêmico e a interpretação foi feita através dos valores encontrados para aprovações e reprovações em cada uma das disciplinas presentes na base. Utilizando como exemplo uma disciplina comum aos dois cursos obteve-se o resultado apresentado na Figura 5.12.

Os gráficos fazem referência a todos os alunos em acompanhamento acadêmico que cursaram a disciplina de Cálculo II nos cursos analisados. No curso de Ciência da Computação noturno, dos 33 alunos (em situação de acompanhamento acadêmico) que cursaram a disciplina, 9 foram aprovados. Já no curso de Sistemas de Informação, dos 11 alunos (em situação de acompanhamento acadêmico) que cursaram, apenas 4 obtiveram aprovação.

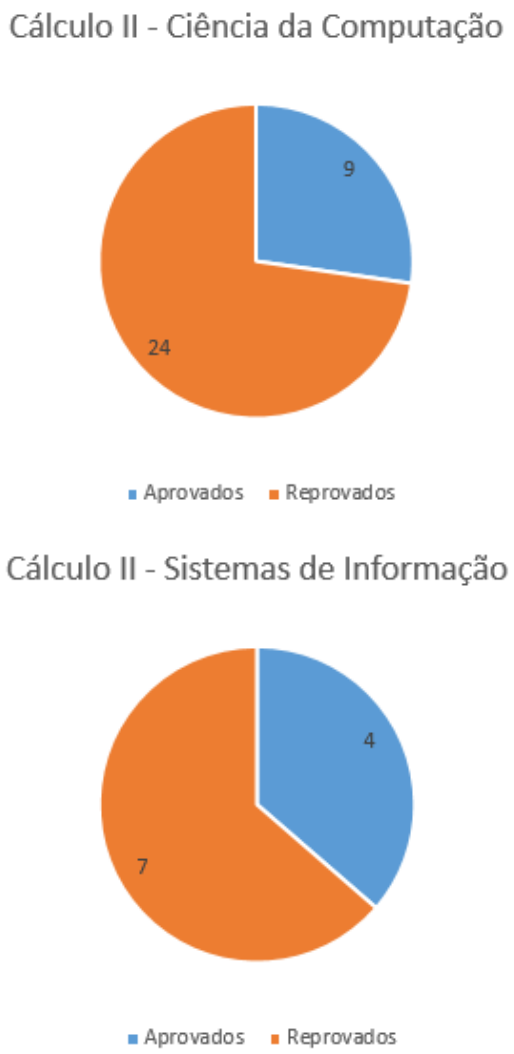


Figura 5.12: Aprovação x Reprovação em Cálculo II

O estudo também foi realizado para a disciplina de Estrutura de dados oferecida pelo departamento de Ciência da Computação. Em Ciência da Computação noturno, dos 55 alunos (em situação de acompanhamento acadêmico) que cursaram esta disciplina, 30 foram aprovados e 25 foram reprovados. Já em Sistemas de Informação, o resultado encontrado foi de 24 aprovados e 33 reprovados para um total de 57 alunos (em situação de acompanhamento acadêmico).

A Figura 5.13 ilustra estes valores.

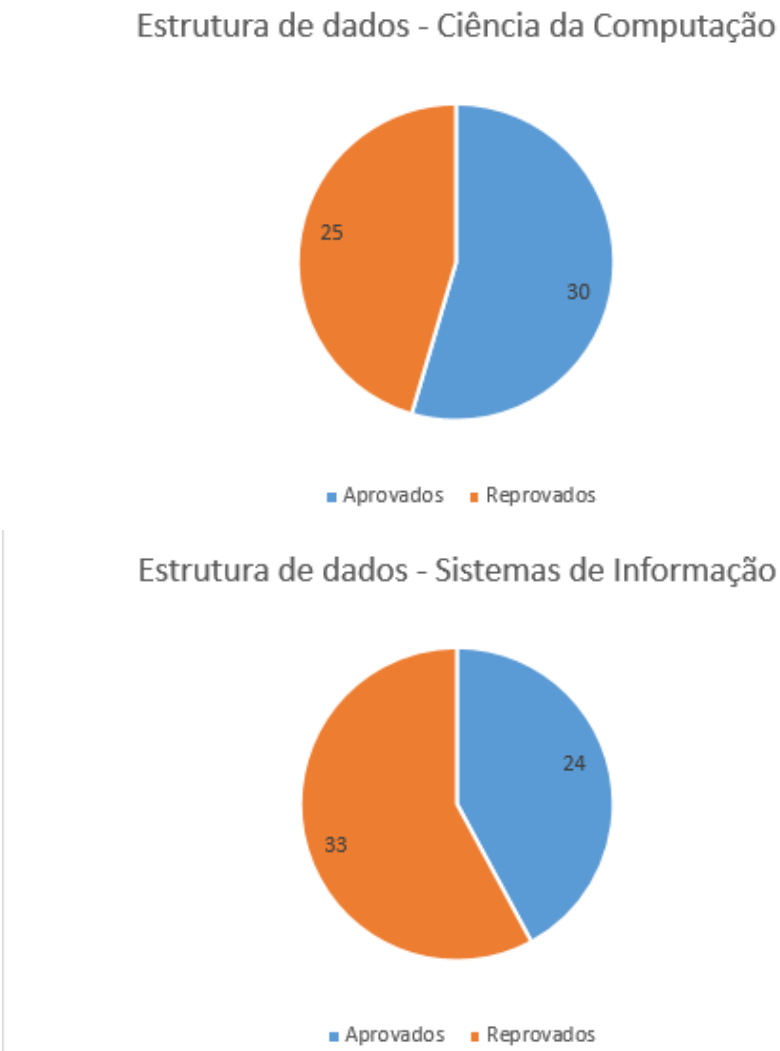


Figura 5.13: Aprovação x Reprovação em Estrutura de dados

5.5 Quantidade de alunos que apresentam CEI ou CET insuficientes

Dentre os entrevistados, houve um empate na avaliação de relevância desta situação: 3 consideram muito relevante e 3 consideram relevante.

Abordagem: esta foi uma questão que pôde ser analisada sem a necessidade da utilização de técnicas de mineração. Consultas no banco de dados permitiram obter valores exatos para tal questionamento. Diante disso, as consultas realizadas no banco de dados foram:

```
SELECT COUNT(DISTINCT IDALUNO) FROM tblbase where IDCURSO='35A'
SELECT COUNT(DISTINCT IDALUNO) FROM tblbase where IDCURSO='35A' AND
ACOMPANHAMENTO='SIM'

SELECT COUNT(DISTINCT IDALUNO) FROM tblbase where IDCURSO='76A'
SELECT COUNT(DISTINCT IDALUNO) FROM tblbase where IDCURSO='76A' AND
ACOMPANHAMENTO='SIM'
```

As consultas acima permitiram identificar a total de alunos presentes em cada curso e quantos deles se enquandram em situação de acompanhamento acadêmico.

Levando em consideração que os alunos que apresentam o atributo ACOMPANHAMENTO= SIM são caracterizados como alunos com CEI/CET insuficientes, foram obtidos os seguintes resutados: para o curso de Ciência da Computação, do total de 72 alunos presentes na base de dados, 63 se encontram em situação de acompanhamento acadêmico. Já para o curso de Sistemas de Informação, dos 88 alunos presentes na base dados, 73 estão em acompanhamento acadêmico. A Figura 5.14 ilustra os resultados obtidos.

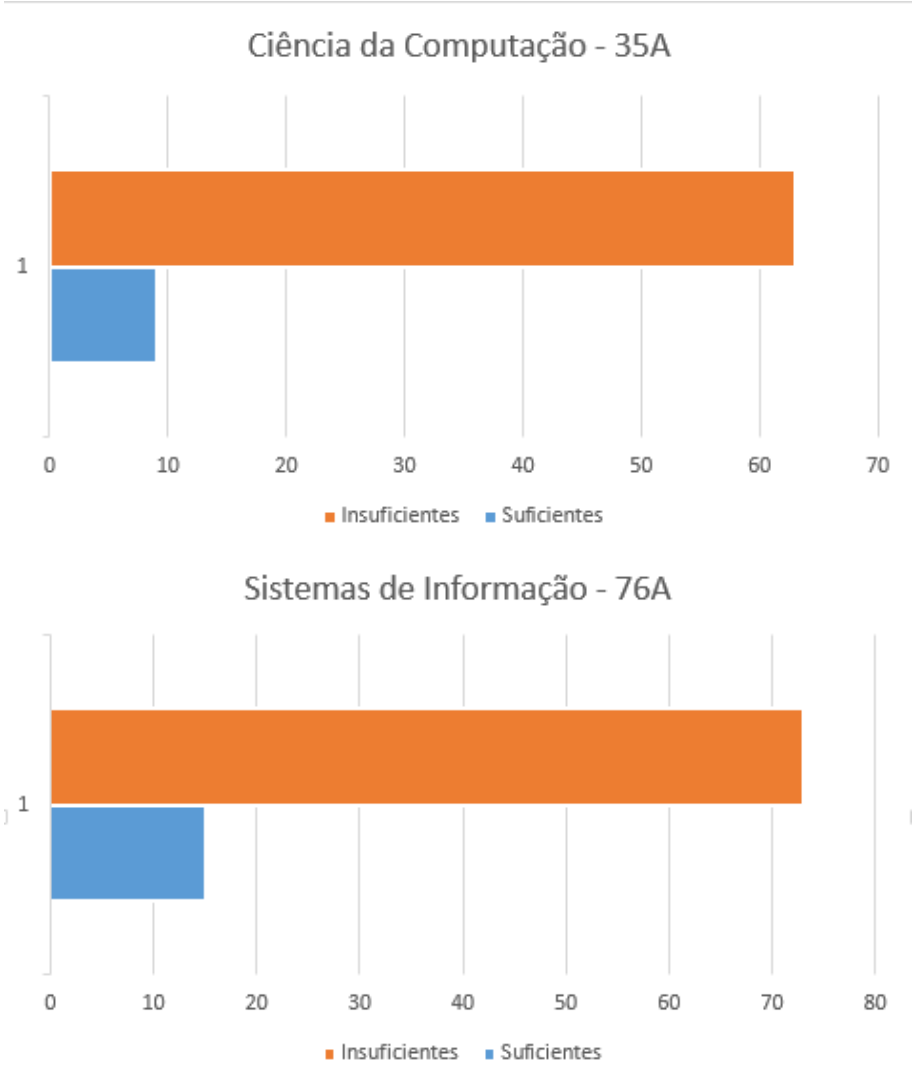


Figura 5.14: Alunos com CEI ou CET insuficientes

5.6 Maiores aprovações entre os alunos com CEI e CET insuficientes

Esta situação apresentou 3 votos no quesito muito relevante, 2 no quesito relevante e 1 no quesito pouco relevante.

Apriori: o algoritmo não obteve resultados satisfatórios, pois não realizou as associações entre os atributos selecionados da forma esperada. A Figura 5.15 mostra que as associações concentraram-se mais no atributo STATUS, e apenas uma disciplina foi retornada. A forma como o algoritmo funciona tem por objetivo identificar os itens mais frequentes nas transações analisadas, por este motivo neste contexto não foi possível obser-

var resultados referentes a todas as disciplinas, visto que na base de dados umas são mais frequentes do que as outras. Neste caso, a utilização do algoritmo se tornou ineficiente, pois não foi possível identificar as disciplinas em que os alunos mais se aprovaram.

```
Best rules found:

 1. STATUS=Aprovado 394 ==> ACOMPANHAMENTO=SIM 394    conf:(1)
 2. STATUS=Rep Nota 303 ==> ACOMPANHAMENTO=SIM 303    conf:(1)
 3. STATUS=Rep Freq 175 ==> ACOMPANHAMENTO=SIM 175    conf:(1)
 4. STATUS=Trancado 139 ==> ACOMPANHAMENTO=SIM 139    conf:(1)
 5. NOME=CÁLCULO I 111 ==> ACOMPANHAMENTO=SIM 111    conf:(1)
```

Figura 5.15: Resultado da execução do Apriori para a relação 5.6

J48: neste questionamento, o algoritmo J48 permitiu através da árvore de decisão gerada após a execução, uma melhor visualização dos dados a ser interpretados. Para a configuração dos parâmetros do algoritmo foram utilizados os valores padronizados que podem ser observados no Anexo I, tabela I.2.

Para encontrar as disciplinas com maior índice de aprovação foi necessário verificar todas as matérias cursadas pelos alunos de cada um dos cursos presentes na base de dados. Os resultados ilustrados na Figura 5.16 foram obtidos através da porcentagem de aprovação identificada nos resultados da técnica aplicada, correspondente a cada uma das disciplinas presentes na base. Alguns valores podem estar distorcidos devido à pequena quantidade de alunos cursando algumas das disciplinas apresentadas.

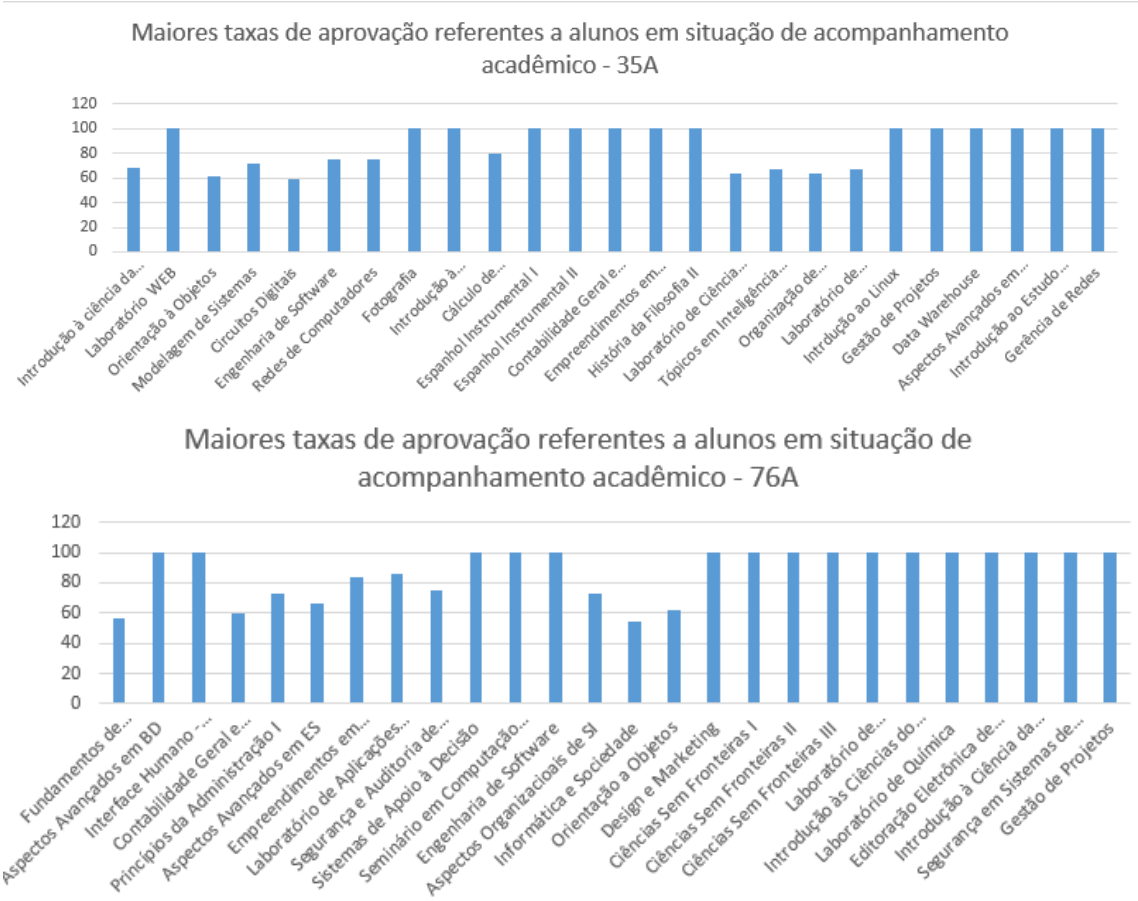


Figura 5.16: Disciplinas com maior número de aprovações

5.7 CET insuficiente após reprovação em disciplinas do 1º ou 2º período

Dentre os entrevistados, 4 consideram esta situação muito relevante.

Apriori: nesta situação, o algoritmo associou os atributos selecionados de forma que o questionamento envolvido não pôde ser respondido, pois seu objetivo é identificar os itens mais frequentes nas transações analisadas e não apresentar resultados de forma que todos os itens sejam detalhados. Dessa forma, conforme pode ser visto na Figura 5.17, apenas uma disciplina dentre todas as presentes na base foi identificada durante a mineração.

```
Best rules found:

 1. STATUS=Aprovado 394 ==> ACOMPANHAMENTO=SIM 394    conf:(1)
 2. STATUS=Rep Nota 303 ==> ACOMPANHAMENTO=SIM 303    conf:(1)
 3. STATUS=Rep Freq 175 ==> ACOMPANHAMENTO=SIM 175    conf:(1)
 4. STATUS=Trancado 139 ==> ACOMPANHAMENTO=SIM 139    conf:(1)
 5. NOME=CÁLCULO I 111 ==> ACOMPANHAMENTO=SIM 111    conf:(1)
```

Figura 5.17: Resultado da execução do Apriori para a relação 5.7

J48: executando o algoritmo J48 para a base de dados original (Com ou sem CET insuficientes), é possível verificar para cada disciplina a quantidade de alunos aprovados e reprovados. Para isto, foram escolhidos os atributos NOME, ACOMPANHAMENTO e STATUS com os parâmetros do algoritmo padronizados que podem ser observados no Anexo I, tabela I.2.

Utilizando registros de alunos apenas do curso de Ciência da Computação noturno e escolhendo como exemplo uma matéria do 1º período, o algoritmo retornou o seguinte resultado:

```
NOME = CÁLCULO I
| ACOMPANHAMENTO = NÃO: Aprovado (13.0/4.0)
| ACOMPANHAMENTO = SIM: Rep Nota (111.0/58.0)
```

Isto significa que, de um total de 124 alunos, 13 estão classificados como fora da situação de acompanhamento acadêmico e 111 estão em acompanhamento. Os 111 representam 89,51% dos alunos que cursaram esta disciplina, ou seja, neste caso, a chance desses alunos adquirirem CET insuficiente foi muito acima da média. Com isso, é possível identificar que dos alunos que reprovaram nesta disciplina no início do curso a maior parte deles passou a integrar o grupo em situação de acompanhamento acadêmico. Analisando os registros referentes aos alunos de Sistemas de Informação, o algoritmo retornou para a mesma disciplina, o seguinte resultado:

```
NOME = CÁLCULO I
| ACOMPANHAMENTO = NÃO: Aprovado (20.0/10.0)
```

```
| ACOMPANHAMENTO = SIM: Rep Nota (104.0/51.0)
```

A Figura 5.18 ilustra o resultado deste estudo obtido para outras duas disciplinas do 1º período:

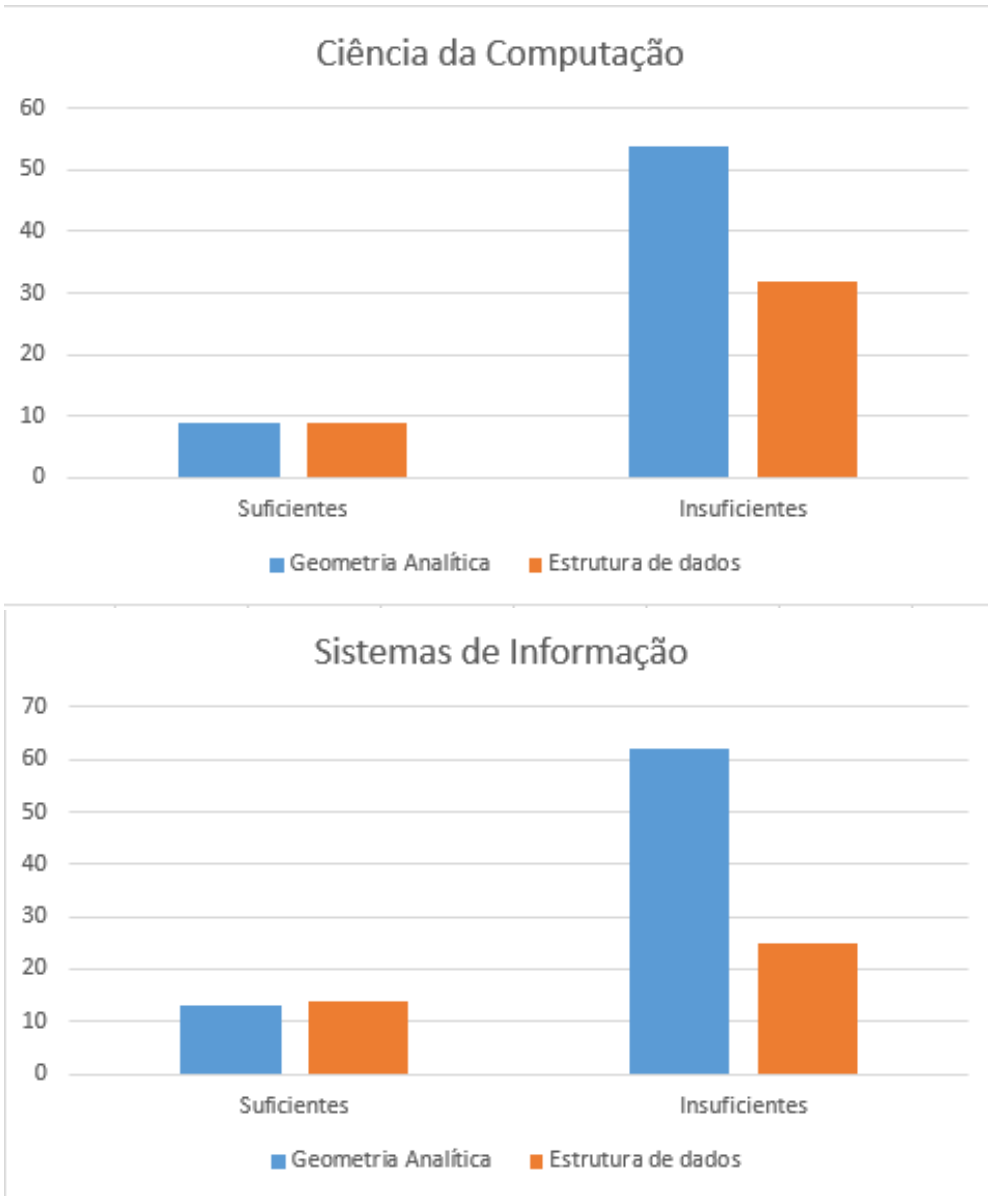


Figura 5.18: Disciplinas do primeiro período x Acompanhamento acadêmico

6 Conclusão

O Instituto de Ciências Exatas apresenta um histórico preocupante em relação ao desempenho acadêmico em algumas disciplinas ofertadas por seus departamentos. Tal conhecimento é comum a todos, porém pouco se tem feito para reverter esta situação, pois o detalhamento dos problemas e a conexão entre eles ainda não ocorrem de forma a apoiar tomadas de decisão.

Este trabalho teve como principal motivação viabilizar técnicas que permitam analisar os registros referentes aos dados acadêmicos dos alunos do ICE, para que alguma solução pudesse ser proposta e consequentemente alterar o cenário atual em que alunos em situação de acompanhamento acadêmico estão sujeitos.

O uso de mineração de dados se mostrou viável para interpretar as informações referentes aos cursos escolhidos, pois seria muito custoso e pouco eficiente realizar o estudo destas informações de forma manual.

A ideia é que, independente do período de tempo considerado nos registros, consiga-se obter resultados que apoiem tomadas de decisão, apenas apresentando para isto os mesmos atributos aqui considerados. O que possibilita, portanto, que o estudo aqui realizado possa ser aplicado também para os demais cursos do ICE.

A opinião dos coordenadores de curso permitiu identificar o tipo de informação que realmente poderia auxiliá-los. Sendo assim, pretende-se com este trabalho permitir que qualquer coordenador munido de uma base de dados, consiga extrair conhecimento necessário para apoiá-lo em situações adversas.

Alunos se enquadram em situação de acompanhamento acadêmico por diversos motivos, cabe então aos coordenadores tomarem decisões que sejam de responsabilidade da coordenação, como orientações durante fase de matrícula, para assim diminuírem evasões e perda do interesse dos alunos que ainda se matém ativos.

Os resultados obtidos neste trabalho possibilitaram identificar de forma mais precisa problemas bastante conhecidos no ambiente analisado. A forma como os algoritmos adotados se comportaram permitiram um conhecimento e visualização mais detalhados e

sem tendenciosiade.

Como proposta de trabalho futuro, está a obtenção de dados seguintes ao período analisado para identificar se mediante ações decorrentes da realização destas técnicas, os alunos conseguiram obter melhores índices CET e com isso saírem da situação de acompanhamento acadêmico. Outro trabalho interessante seria o desenvolvimento de uma ferramenta que concentre toda a interação do coordenador de curso com os dados calculados, tornando a manipulação dessas informações mais intuitiva.

Referências Bibliográficas

[1] Martins A. C.; Marques J. M.; Costa P. D. Estudo comparativo de três algortimos de machine learning na classificação de dados eletrocardiográficos. 2009.

[2] Goldschmidt E.; Passos E.; Bezerra E. *Data Mining - Conceitos, técnicas, algoritmos, orientações e aplicações*. ELSEVIER, 2015.

[3] WITTEN I. H.; FRANK E. *Data mining: practical machine learning tools and techniques*. Morgan Kaufmann Publishers, 2005.

[4] Carvalho J. V.; Sampaio M. C.; Mongiovi G. Utilização de técnicas de "data mining" para o reconhecimento de caracteres manuscritos. 1999.

[5] Castanheira L. G. Aplicação de técnicas de mineração de dados em problemas de classificação de padrões. Master's thesis, Universidade Federal de Minas Gerais, Belo Horizonte, 2008.

[6] Harrison T. H. *Intranet data warehouse*. Editora Berkeley Brasil, 1998.

[7] Gonçalves A. L. Utilização de técnicas de mineração de dados em bases de ct: uma análise de grupos de pesquisa no brasil. Master's thesis, Universidade Federal de Santa Catarina, Florianópolis, 2000.

[8] Souza S. L. Evasão no ensino superior: Um estudo utilizando a mineração de dados como ferramenta de gestão do conhecimento em um banco de dados referente à graduação de engenharia. Master's thesis, Universidade Federal do Rio de Janeiro, Rio de Janeiro, 2008.

[9] Vasconcelos L. M. R.; Carvalho C. L. Aplicação de regras de associação para mineração de dados na web.

[10] Dias M. M. Parâmetros na escolha de técnicas e ferramentas de mineração de dados. 2002.

[11] HAN J.; KAMBER M. *Data Mining concepts and techniques.* Morgan Kaufmann Publishers, 2001.

[12] Librelotto S. R.; Mozzaquatro P. M. Análise dos algoritmos de mineração j48 e apriori aplicados na detecção de indicadores da qualidade de vida e sáude.

[13] Mitra S.; Pal S. K.; P. M. Data mining in soft computing framework: A survey. 2002.

[14] Rezende S. O. Mineração de dados. *XXV Congresso da Sociedade Brasileira de Computação*, 2005.

[15] QUINLAN J. R. *C4.5: Programs for machine learning.* Morgan Kaufmann Publishers, San Francisco, 1993.

[16] Amo S. Técnicas de mineração de dados. 2004.

[17] Côrtes S. C.; Porcaro R. M.; Lifschitz S. Mineração de dados - funcionalidades, técnicas e abordagens. 2002.

[18] Silva M. P. S. Mineração de dados - conceitos, aplicações e experimentos com weka. 2004.

[19] Braga L. P. V. *Introdução à Mineração de Dados.* E-Papers Serviços Editoriais, 2005.

[20] TAN Pang Ning; STEINBACH Michael; KUMAR Vipin. *Introdução ao DATAMING Mineração de Dados 6ed:.* Editora Ciência Moderna Ltda, 2009.

I Parâmetros dos algoritmos

Tabela I.1: Parâmetros utilizados para o algoritmo Apriori

Parâmetro	Valor
classIndex	-1
delta	0.05
lowerBoundMinSupport	0.1
metricType	Confidence
minMetric	0.9
numRules	10
outputItemSets	false
removeAllMissingCols	false
significanceLevel	-1.0
upperBoundMinSupport	1.0
verbose	false

Tabela I.2: Parâmetros utilizados para o algoritmo J48

Parâmetro	Valor
binarySplits	false
confidenceFactor	0.25
debug	false
minNumObj	2
numFolds	3
reducedErrorPruning	false
saveInstanceData	false
seed	1
subtreeRaising	true
unpruned	false
useLaplace	false

A Manual de Instruções

Manual de instruções para aplicação de técnicas de mineração em dados acadêmicos

Gisele Germano da Silva

JUIZ DE FORA

DEZEMBRO, 2016

Base de dados

1. Obtenha a base de dados do curso desejado em formato .csv com os atributos: "IDALUNO", "SITUACAO", "MOTIVOSAIDA", "IDCURSO", "CURSO", "DISCIPLINA", "NOME", "CH", "TURMA", "ANO", "SEMESTRE", "NOTA", "FREQUENCIA" e "STATUS", conforme demonstrado na Figura 1.

IDALUNO	SITUACAO	MOTIVOSAIDA	IDCURSO	CURSO	DISCIPLINA	NOME	TURMA	ANO	SEMESTRE	NOTA	FREQUENCIA	STATUS
126878	Ativo	A Especificar	35A	CIÊNCIA DA COMPUTAÇÃO	DCC119	ALGORITMOS	H	2013	1	93	100	Aprovado

Figura 1

2. Importe o arquivo .csv em um banco de dados
3. Crie mais três atributos no banco de dados: CEI, CET e ACOMPANHAMENTO. Os dois primeiros devem receber valores numéricos, e o último deve ser limitado a SIM e NÃO como resposta.
4. Neste trabalho, os valores de CEI e CET foram calculados de forma manual conforme fórmula prevista no RAG. Dependendo do número de registros, a realização deste calculo demandará bastante tempo. Calcule, então, CEI e CET para cada IDALUNO e atribua o resultado aos respectivos atributos. Se o aluno ainda não tiver CET, e o CEI for menor do que a carga horária média do curso em análise, atribua SIM para o atributo ACOMPANHAMENTO. Caso contrário, se o CET >=1,5*CHM atribua NÃO, se CET<1,5*CHM atribua SIM
5. Verifique as possíveis inconsistências presentes na base.
    - Atribua algum código para valores em branco encontrados (exemplo: EMBRANCO, SEMANO, SEMSEMESTRE, etc) .
    - Verifique se há mais de uma informação representando uma mesma característica e unifique-a (por exemplo, RI e Reprovado por infrequência).
6. Utilize o Weka para efetuar a mineração dos dados. Download disponível em: http://www.cs.waikato.ac.nz/ml/weka/downloading.html

Pré-processamento

7. É necessário converter o arquivo .csv em .arff que é o formato reconhecido pelo Weka. Para isto, na tela inicial do Weka vá em *Tools – ArffViewer.*
8. Uma tela em branco surgirá. Nesta tela vá em *File – Open*. E abra o arquivo desejado em formato .csv
9. Após abrir o arquivo você verá uma tela semelhante à figura 2:

File Edit View

35A.csv

Relation: 35A

No.	IDALUNO Numeric	SITUACAO Nominal	MOTIVOSAIDA Nominal	IDCURSO Nominal	CURSO Nominal	DISCIPLINA Nominal	NOME Nominal	CH Numeric	TURMA Nominal	ANO Numeric	SEMESTRE Numeric	NOTA Nominal	FREQUENC Numeric
1	126878.0	Ativo	A Especificar	35A	CIÊNC...	DCC119	ALGO...	60.0	H	2013.0	1.0	93.0	10(
2	126878.0	Ativo	A Especificar	35A	CIÊNC...	DCC120	LABO...	30.0	E	2013.0	1.0	91.0	10(
3	126878.0	Ativo	A Especificar	35A	CIÊNC...	MAT154	CÁLC...	60.0	I	2013.0	1.0	98.0	10(
4	126878.0	Ativo	A Especificar	35A	CIÊNC...	MAT155	GEOM...	60.0	I	2013.0	1.0	92.0	10(
5	126878.0	Ativo	A Especificar	35A	CIÊNC...	DCC013	ESTRU...	60.0	A	2013.0	3.0	97.0	10(
6	126878.0	Ativo	A Especificar	35A	CIÊNC...	DCC107	LABO...	30.0	AA	2013.0	3.0	97.0	10(
7	126878.0	Ativo	A Especificar	35A	CIÊNC...	DCC116	INTRO...	30.0	A	2013.0	3.0	94.0	10(
8	126878.0	Ativo	A Especificar	35A	CIÊNC...	DCC121	LABO...	30.0	AA	2013.0	3.0	93.0	10(
9	126878.0	Ativo	A Especificar	35A	CIÊNC...	FIS073	FISICA I	60.0	D	2013.0	3.0	85.0	10(
10	126878.0	Ativo	A Especificar	35A	CIÊNC...	FIS077	LABO...	30.0	N	2013.0	3.0	88.0	10(
11	126878.0	Ativo	A Especificar	35A	CIÊNC...	MAT067	INTRO...	60.0	A	2013.0	3.0	98.0	10(
12	126878.0	Ativo	A Especificar	35A	CIÊNC...	MAT156	CÁLC...	60.0	E	2013.0	3.0	90.0	10(
13	126878.0	Ativo	A Especificar	35A	CIÊNC...	DCC012	ESTRU...	60.0	A	2014.0	1.0	77.0	10(
14	126878.0	Ativo	A Especificar	35A	CIÊNC...	DCC025	ORIEN...	60.0	A	2014.0	1.0	75.0	10(
15	126878.0	Ativo	A Especificar	35A	CIÊNC...	DCC065	COMP...	60.0	A	2014.0	1.0	89.0	10(
16	126878.0	Ativo	A Especificar	35A	CIÊNC...	EST029	CÁLC...	60.0	E	2014.0	1.0	92.0	10(
17	126878.0	Ativo	A Especificar	35A	CIÊNC...	MAT157	CÁLC...	60.0	E	2014.0	1.0	93.0	10(
18	126878.0	Ativo	A Especificar	35A	CIÊNC...	DCC008	CALC...	60.0	E	2014.0	3.0	90.0	10(
19	126878.0	Ativo	A Especificar	35A	CIÊNC...	DCC059	TEORI...	60.0	A	2014.0	3.0	71.0	10(
20	126878.0	Ativo	A Especificar	35A	CIÊNC...	DCC117	MODE...	60.0	A	2014.0	3.0	87.0	10(
21	126878.0	Ativo	A Especificar	35A	CIÊNC...	FIS075	FISIC...	60.0	C	2014.0	3.0	68.0	10(
22	126878.0	Ativo	A Especificar	35A	CIÊNC...	MAT029	EQUA...	60.0	D	2014.0	3.0	78.0	10(
23	126878.0	Ativo	A Especificar	35A	CIÊNC...	DCC033	FLUXO...	60.0	A	2015.0	1.0	100.0	10(
24	126878.0	Ativo	A Especificar	35A	CIÊNC...	DCC060	BANC...	60.0	A	2015.0	1.0	80.0	10(
25	126878.0	Ativo	A Especificar	35A	CIÊNC...	DCC061	ENGE...	60.0	A	2015.0	1.0	87.0	10(
26	126878.0	Ativo	A Especificar	35A	CIÊNC...	DCC063	LINGU...	60.0	A	2015.0	1.0	69.0	10(

Figura 2

Vá em *File* – *Save as*, e escolha o formato Arff data files (*.arff)

Carregando os dados no Weka

- Na tela inicial do Weka, vá em *Explorer*, conforme Figura 3

Figura 3

- Você verá a seguinte tela (Figura 4)

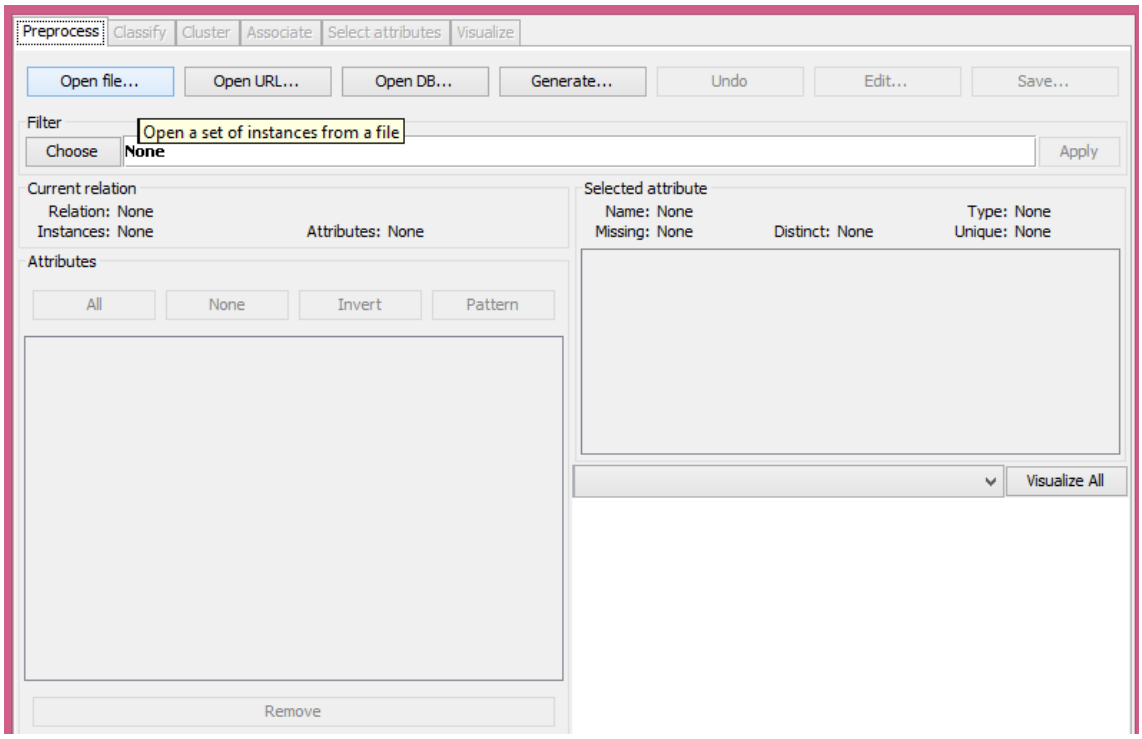


Figura 4

- Vá em *Open file...* e abra o arquivo desejado no formato .arff
- Com o arquivo carregado, selecione os atributos necessários para a mineração

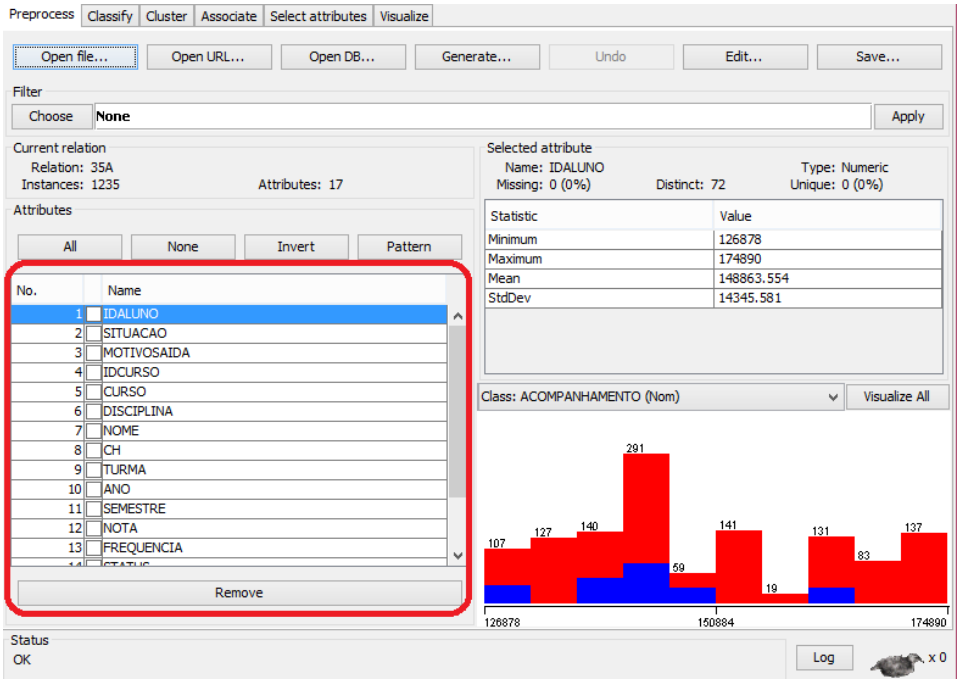


Figura 5

Processamento

10. Para identificar dentre os alunos que apresentam CEI ou CET insuficientes, quais disciplinas eles foram reprovados, realize os seguintes passos:
    - Neste caso, mantenha apenas os atributos NOME, STATUS e ACOMPANHAMENTO. E em seguida, clique na aba *Classify*.
    - Defina o Classifier, como J48 (*Choose – weka – classifiers – trees – J48*)
    - Marque a opção *Use training set*
    - Selecione o atributo *STATUS* e clique em *Start*

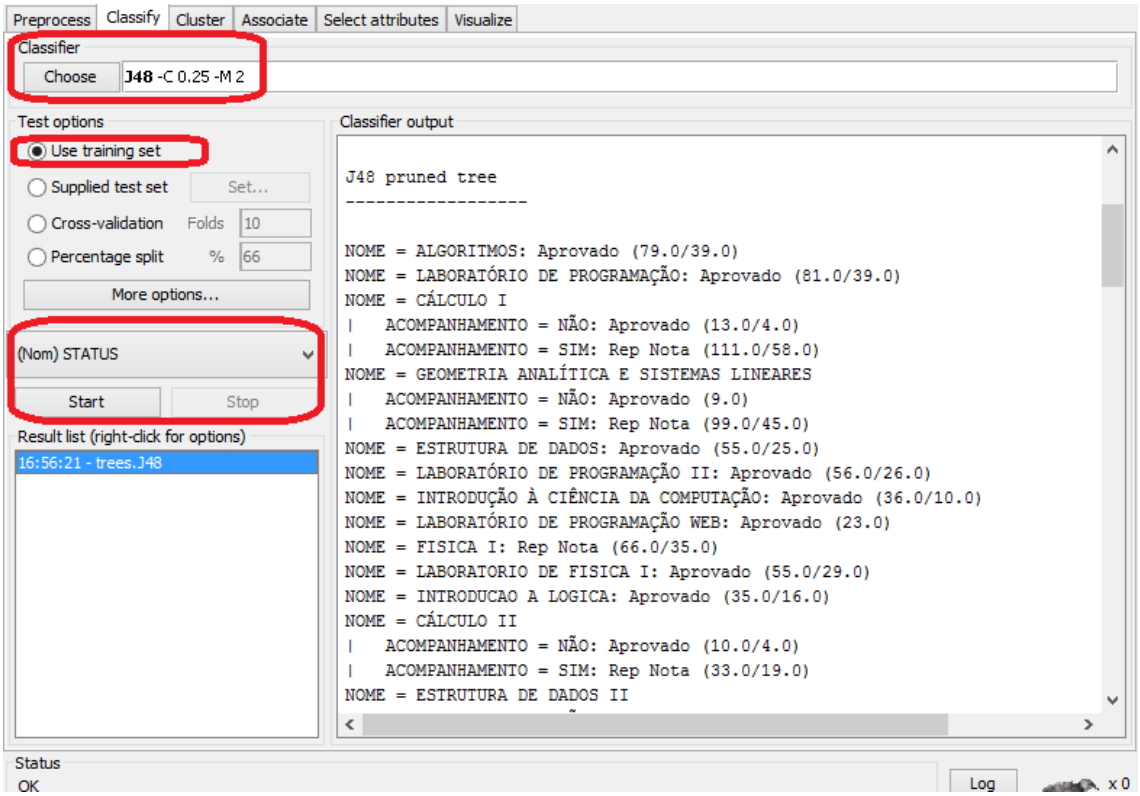


Figura 6

- Em *Classifier output*, ao centro, serão apresentados os resultados da mineração realizada.

11. Para saber por quantos períodos, em média, um aluno com CET insuficiente permanece no curso, execute os seguintes passos:

    - Mantenha apenas os atributos SITUAÇÃO, ANO e SEMESTRE
    - Em *Filter*, escolha o filtro NumericToNominal (*Choose - weka – filters – unsupervised – atribute – NumericToNominal*)
    - Selecione os atributos que forem do tipo numeric e clique em *Apply*

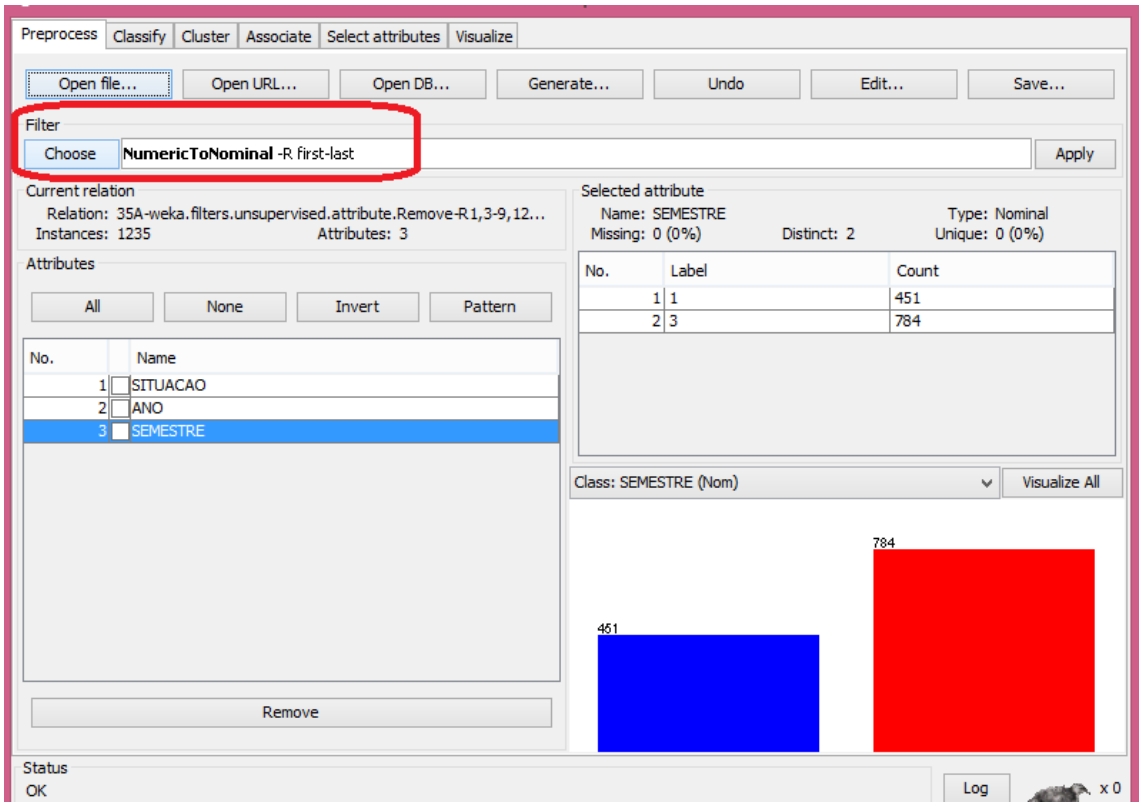


Figura 7

- Vá em *Classify* e defina o *Classifier*, como J48 (*Choose – weka – classifiers – trees – J48*)
- Marque a opção *Use training set*
- Selecione o atributo *ANO* e clique em *Start*
- Em *Classifier output*, ao centro, serão apresentados os resultados da mineração realizada.

12. Para saber quantos alunos com CEI ou CET insuficiente estão aprovados/reprovados na disciplina X:
    - Realize todos os passos presentes no item 11, e interprete os resultados obtidos para a disciplina X desejada

13. Para saber dentre os alunos que apresentam CEI ou CET insuficientes, em quais disciplinas eles aprovaram, faça:
    - Repita os passos do item 11, e verifique dentre as disciplinas apresentadas, as que possuem maiores índices de aprovação

14. Para saber qual a chance de um aluno reprovado em disciplinas do 1º ou 2º período ter um CET insuficiente, faça:
    - Carregue a base de dados no Weka
    - Matenha apenas os atributos NOME, ACOMPANHAMENTO e STATUS
    - Vá em *Classify* e defina o *Classifier*, como J48 (*Choose – weka – classifiers – trees – J48*)
    - Marque a opção *Use training set*

- Selecione o atributo *STATUS* e clique em *Start*
- Em *Classifier output*, ao centro, serão apresentados os resultados da mineração realizada.
- Interprete os resultados referentes às disciplinas de 1º e 2º período.

Os questionamentos abaixo puderam ser realizados sem a aplicação de técnicas de mineração, visto que se tratavam de informações que podem ser obtidas com precisão através de consultas ao banco de dados.

15. Para saber qual a relação entre alunos ativos com CEI ou CET insuficiente/suficiente**:**

- Neste caso é interessante estudar o grupos isoladamente para melhor interpretação dos dados
- Para obter resultados em relação àqueles que possuem CEI ou CET suficientes execute as seguintes consultas

```
SELECT COUNT(DISTINCT IDALUNO) FROM tblbase where IDCURSO='35' and
ACOMPANHAMENTO=’NÃO'
```

Obs: IDCURSO deve corresponder ao código associado ao curso analisado e tblbase deve ser substituído pelo nome da sua tabela. A consulta acima se faz necessária para identificar o total de alunos com CEI ou CET suficientes. Em seguida faça:

```
SELECT COUNT(DISTINCT IDALUNO) FROM tblbase where IDCURSO='35A' AND SITUAÇÃO=
'ATIVO' AND ACOMPANHAMENTO='NÃO'
```

Após essa segunda consulta, será retornado o número de alunos ativos dentre o total apresentado anteriormente.

- Para obter resultados em relação àqueles que possuem CEI ou CET insuficientes execute as seguintes consultas:

```
SELECT COUNT(DISTINCT IDALUNO) FROM tblbase WHERE IDCURSO='35' AND
ACOMPANHAMENTO='SIM'
```

Obs: IDCURSO deve corresponder ao código associado ao curso analisado e tblbase deve ser substituído pelo nome da sua tabela. A consulta acima se faz necessária para identificar o total de alunos com CEI ou CET insuficientes. Em seguida faça:

```
SELECT COUNT(DISTINCT IDALUNO) FROM tblbase WHERE IDCURSO='35A' AND
SITUAÇÃO= 'ATIVO' AND ACOMPANHAMENTO='SIM'
```

Após essa segunda consulta, será retornado o número de alunos ativos dentre o total apresentado anteriormente.

16. Para um determinado período de ingresso, quantos alunos apresentam CEI ou CET insuficientes?
    - Neste caso, utilize a base de dados original do curso a ser analisado (contendo registros tanto de alunos em acompanhamento acadêmico quanto dos que estão fora deste quadro)
    - Execute as seguintes consultas ao banco de dados:

      SELECT COUNT(DISTINCT IDALUNO) FROM tblbase WHERE IDCURSO='35A'

      Para saber quantos alunos estão na base de dados analisada e,

      SELECT COUNT(DISTINCT IDALUNO) FROM tblbase WHERE IDCURSO='35A' AND ACOMPANHAMENTO='SIM'

      Para saber do total encontrado na primeira consulta, quantos estão com CEI ou CET insuficiente.