UNIVERSIDADE FEDERAL DO ESPÍRITO SANTO
CENTRO DE CIÊNCIAS EXATAS, NATURAIS E DA SAÚDE
DEPARTAMENTO DE COMPUTAÇÃO

MELISSA SILVA DE SOUZA

USO DA MINERAÇÃO DE DADOS NA IDENTIFICAÇÃO DE ALUNOS COM PERFIL DE EVASÃO NOS CURSOS CIÊNCIA DA COMPUTAÇÃO E SISTEMAS DE INFORMAÇÃO DA UFES

ALEGRE – ES
2016

MELISSA SILVA DE SOUZA

USO DA MINERAÇÃO DE DADOS NA IDENTIFICAÇÃO DE ALUNOS COM PERFIL DE EVASÃO NOS CURSOS CIÊNCIA DA COMPUTAÇÃO E SISTEMAS DE INFORMAÇÃO DA UFES

Trabalho de conclusão de curso apresentado ao Departamento de Computação do Centro de Ciências Exatas, Naturais e Saúde da Universidade Federal do Espírito Santo, como requisito parcial para obtenção do grau de Bacharel em Ciência da Computação.

Orientador: Prof. M.Sc. Jacson Rodrigues Correia da Silva

ALEGRE - ES

2016

USO DA MINERAÇÃO DE DADOS NA IDENTIFICAÇÃO DE ALUNOS COM PERFIL DE EVASÃO NOS CURSOS CIÊNCIA DA COMPUTAÇÃO E SISTEMAS DE INFORMAÇÃO DA UFES

Trabalho de conclusão de curso apresentado ao Departamento de Computação do Centro de Ciências Exatas, Naturais e da Saúde da Universidade Federal do Espírito Santo, como requisito parcial para obtenção do grau de Bacharel em Ciência da Computação.

Aprovado em 25 de novembro de 2016.

Banca Examinadora

Jacson R. C. Silva

Prof. M.Sc. Jacson Rodrigues Correia da Silva
Universidade Federal do Espírito Santo
Orientador

Rodrigo FL

Prof. M.Sc. Rodrigo de Freitas Silva
Universidade Federal do Espírito Santo

Prof. Dr. Bruno Vilela Oliveira
Universidade Federal do Espírito Santo

ALEGRE - ES

2016

“Só eu conheço os planos que tenho para vocês: prosperidade e não desgraça e um futuro cheio de esperança. ”

Jeremias 29:11

AGRADECIMENTOS

Primeiramente gostaria de agradecer a Deus pelo cuidado com minha vida, agradecer meus pais Ernandes e Marilane que nunca deixaram eu desistir, mesmo eu escolhendo estudar fora de casa num momento tão delicado, minha irmã, confidente e amiga Milena que passou aperto para que eu pudesse estudar, minha irmã postiça Michelle que me ouve e tem sempre os melhores conselhos.

Um obrigado a Lorrane que me tirou de Rio das Ostras e me acolheu em Alegre, ao Maycown que foi a primeira pessoa a socializar comigo, ao Egle que com todo seu desespero, lançando raios desesperadores muita das vezes chegava de manhã e me acordava com pão e café para a gente estudar, Marília me ensinou teste de mesa e me ajudava em matérias quando eu estava cansada, ao Raphael que acampava na minha casa em semana de provas finais e a gente estudava e assistia vídeos idiotas.

Um agradecimento especial a todas as pessoas que me alimentaram, sendo elas Caíque, que não tinha hora ruim para ir para cozinha, Heloiza que ia para casa e trazia café para mim, Cássio e Emerson que apesar implicantes sempre me chamavam para comer pizzas e comida de mãe, Thaís que juntamente com a Marília me recebiam com café e comidas na madrugada de estudos e a Luciene que nunca me deixa sem pipoca e bolos.

Também a aqueles que me aguentaram e estudaram comigo para as maratonas de programação, Talles, Renan, Denis, Wagner, Caíque, passamos muito aperto e aprendemos muito juntos. Ao Lázaro que apesar de ser com interesse, consertou tomadas e deu jeito no meu notebook me salvando.

E as pessoas que me proporcionaram a oportunidade de crescer como pessoa, Larice e Antonio me chamando para o projeto Info +, Ana Paula que mesmo sem me conhecer bem me acolheu e sem saber se tornou um exemplo a seguir, sempre paciente e animada com seu “Bom dia flor do dia! ”. Logo em seguida me incentivou a entrar na empresa júnior que igualmente me ensinou muito.

Ao forró pé de serra que aprendi nesse Estado que passei a amar.

E não menos importante meu orientador Jacson, que tenho orgulho de ter como a pessoa para me guiar nesse momento que eu tinha tanto medo, e com humildade e paciência me ensinou, me acalmou e me guiou nesse trabalho, e com imenso carinho que faço propaganda para todos do professor incrível que ele é.

SUMÁRIO

1 INTRODUÇÃO 12
2 OBJETIVOS 14
2.1 Objetivo geral 14
2.2 Objetivos específicos 14
3 REVISÃO DE LITERATURA 15
4 METODOLOGIA 21
5 ANÁLISE DOS RESULTADOS 25
5.1 Perfil dos alunos 28
6 CONCLUSÃO 31
7 REFERÊNCIAS BIBLIOGRÁFICAS 32
APÊNDICE A - Formulário de calouros 34

LISTA DE TABELAS

Tabela 1: Índice de preenchimento de vagas por ano nos cursos de Ciência da Computação e Sistemas de Informação. Fonte: Sarcinelli (2014) 16
Tabela 2: Taxa de evasão por período nos cursos de Ciência da Computação e Sistemas de Informação até 2012/1. Fonte: Sarcinelli (2014) 16
Tabela 3: Taxa de evasão por período nos cursos de Ciência da Computação e Sistemas de Informação depois de 2012/1. Fonte: Sarcinelli (2014) 17
Tabela 4: Acurácia dos classificadores com todas as perguntas 25
Tabela 5: Perguntas com peso maior que 0,01 25
Tabela 6: Perguntas com peso maior que 0,0 e menor que 0,01 26
Tabela 7: Acurácia dos classificadores com perguntas com peso maior que 0,01 (Tabela 5) 26
Tabela 8: Perguntas que são decisivas para os classificadores 28

LISTA DE FIGURAS

Figura 1: Mineração de dados é a interseção de diferentes áreas de conhecimento. Fonte: Silva (2004) 18
Figura 2: Definição das perguntas no arquivo .arff 22
Figura 3: Definição das respostas no arquivo .arff..... 23

LISTA DE GRÁFICOS

Gráfico 1: Número de vagas ofertadas por ano e número de alunos ingressantes por ano no curso de Ciência da Computação. Fonte: Sarcinelli (2014) 15
Gráfico 2: Número de vagas ofertadas por ano e número de alunos ingressantes por ano no curso de Sistemas de Informação. Fonte: Sarcinelli (2014) 16
Gráfico 3: Rede Bayesiana 27
Gráfico 4: SMO 27
Gráfico 5: J48 27
Gráfico 6: Porcentagem de alunos por curso 28
Gráfico 7: Porcentagem de alunos por sexo 29
Gráfico 8: Porcentagem de alunos por idade 29
Gráfico 9: Porcentagem de alunos que gostam de Física, Biologia, Artes e Sociologia 30
Gráfico 10: Porcentagem de alunos que possuem Facebook e Twitter 30

LISTA DE SIGLAS

WEKA	Waikato Environment for Knowledge Analysis
UFES	Universidade Federal do Espírito Santo
IES	Instituições de Ensino Superior
CCENS	Centro de Ciências Exatas Naturais e da Saúde
REUNI	Reestruturação e expansão das Universidades Federais
SMO	Otimização Mínima Sequencial
SVM	Support Vector Machines
UFRJ	Universidade Federal do Rio de Janeiro
TI	Tecnologia da Informação

RESUMO

A evasão de estudantes é um problema enfrentado pelas instituições de ensino e a permanência deles é uma preocupação atual, sendo necessárias medidas para garantir a finalização escolar e acadêmica. O objetivo principal deste trabalho foi identificar razões para a evasão no ensino superior da Universidade Federal do Espírito Santo nos cursos Ciência da Computação e Sistemas de Informação no campus de Alegre, comparando a acurácia de classificadores amplamente utilizados neste contexto, propiciando assim subsídios para os gestores aplicarem medidas que possam minimizar tal problema. Portanto, ao utilizar os classificadores J48, SMO e Rede Bayesiana que são os mais usados para a análise da evasão, percebeu-se que é possível identificar alunos com perfil de evasão no primeiro período, alunos do curso de Sistemas de Informação, com mais de 21 anos e que gostam de Física, Artes e Sociologia, que são matérias não relacionadas diretamente aos cursos de computação, são os com mais risco de deixar o curso.

Palavras Chave: Evasão, Mineração de dados, WEKA, UFES.

1 INTRODUÇÃO

Um grande problema enfrentado pelas instituições de ensino é a saída dos estudantes sem a conclusão dos estudos. Passar no vestibular não assegura a inclusão efetiva desses estudantes e sua permanência, sendo necessárias medidas mais eficazes que garantam a finalização escolar e acadêmica (AZEVEDO; SANTOS, 2015).

Por isso, a evasão é um dos maiores problemas de qualquer nível de ensino. Não finalizar os estudos representa uma perda de recursos e de tempo de todos participantes do processo de ensino, ou seja, uma grande perda para nação (LOBO, 2012).

Segundo Motejunas et al. (2007), no setor público, são recursos públicos investidos sem o devido retorno. No setor privado, é uma importante perda de receitas. Em ambos os casos, a evasão é uma fonte de ociosidade de professores, funcionários, equipamentos e espaço físico.

Nesse contexto, entende-se ser importante o desenvolvimento de técnicas que permitam auxiliar na identificação de alunos que apresentam tendências a abandonar os estudos, possibilitando ao professor e a instituição tomarem as devidas providências para reduzir a evasão desses alunos (MARTINS; LOPES; RAABE, 2012).

A perda com a evasão dos alunos pode ser mensurável em termos quantitativos, mas em termos qualitativos, torna-se uma tarefa com maior grau de complexidade. Isso porque a saída do aluno, sem a consequente conclusão do seu curso, traz perdas implícitas (AZEVEDO; SANTOS, 2015) e verifica-se, em todo o mundo, que a taxa de evasão no primeiro ano de curso é duas a três vezes maior do que a dos anos seguintes (MOTEJUNAS et al., 2007).

Portanto, o estudo das causas da evasão e a tomada de medidas preventivas estão fortemente ligados ao contexto de cada instituição de ensino (MANHÃES et al., 2011), e ainda é preciso que os resultados considerem comparativamente a visão daqueles que evadem contra aqueles que decidem permanecer na instituição, ao mesmo tempo em que se busque controlar qualquer viés existente na seleção dos alunos analisados (SILVA, 2013).

Neste trabalho foi realizada a coleta de dados de alunos ingressantes nos cursos de Ciência da Computação e Sistemas de Informação, alunos no primeiro período do curso, onde eles responderem questões pessoais, que juntamente com dados informados pelos coordenadores dos cursos, foram analisados e pôde-se perceber o perfil dos alunos de TI, onde a evasão está diretamente relacionada com o curso, matérias que mais gostavam no ensino médio, sua relação com as redes sociais e a internet. Trazendo assim informações para os coordenadores, professores, podendo assim utilizar esses indicadores para melhorias na recepção desses alunos e permanência deles.

Este trabalho está estruturado da seguinte forma: No Capítulo 2 são apresentados os objetivos do trabalho. No Capítulo 3 discutem-se os trabalhos relacionados referentes ao problema da evasão escolar utilizando mineração de dados. No Capítulo 4 descreve-se a metodologia. No Capítulo 5 descreve-se a análise dos resultados. No Capítulo 6 é apresentada a conclusão do trabalho.

2 OBJETIVOS

2.1 Objetivo geral

O objetivo principal deste trabalho foi identificar razões para a evasão no ensino superior da Universidade Federal do Espírito Santo nos cursos Ciência da Computação e Sistemas de Informação no campus de Alegre, comparando a acurácia de classificadores amplamente utilizados neste contexto, propiciando assim subsídios para os gestores aplicarem medidas que possam minimizar tal problema.

2.2 Objetivos específicos

a. Identificar e apresentar as variáveis, conseguidas através de um questionário aplicado aos calouros, que representam os principais fatores que influenciam a conclusão e a evasão.
b. Utilizar métodos de mineração de dados já conhecidos para buscar quais as características e as variáveis que estão mais relacionadas à evasão dos alunos.
c. Particionar os dados em conjuntos de treinamento e teste para analisar o comportamento do modelo.
d. Identificar o perfil do aluno que evade e o perfil do concluinte.

3 REVISÃO DE LITERATURA

A evasão de alunos no ensino superior não é um problema apenas para a Universidade Federal do Espírito Santo (UFES), mas está presente em todas as Instituições de Ensino Superior (IES), tanto públicas quanto privadas. Segundo Reinert e Gonçalves (2010), a evasão escolar compreende, em linhas gerais, o abandono da escola durante o período letivo, ou seja, o aluno se matricula, inicia suas atividades escolares, porém, em seguida deixa de frequentar a escola, por uma ou mais razões.

O Centro de Ciências Exatas, Naturais e da Saúde (CCENS) da UFES enfrenta ainda outro problema: a pouca procura pelos cursos de Ciência da Computação e de Sistemas de Informação.

Sarcinelli (2014) alerta sobre a situação dos cursos do REUNI, em que podemos verificar uma grande oferta de vagas, porém um preenchimento insatisfatório (Gráfico 1 e Gráfico 2). Podemos analisar que nos 2 cursos o número de ingressantes só foi diminuindo, e apesar das vagas de SI aumentarem para 70, menos alunos se matricularam. No percentual de vagas preenchidas (Tabela 1), é notado que o ano de 2014 foi o pior para os dois cursos, com 47,5% e 41,4% apenas de aproveitamento das vagas para CC e SI.

Gráfico 1: Número de vagas ofertadas por ano e número de alunos ingressantes por ano no curso de Ciência da Computação. Fonte: Sarcinelli (2014)

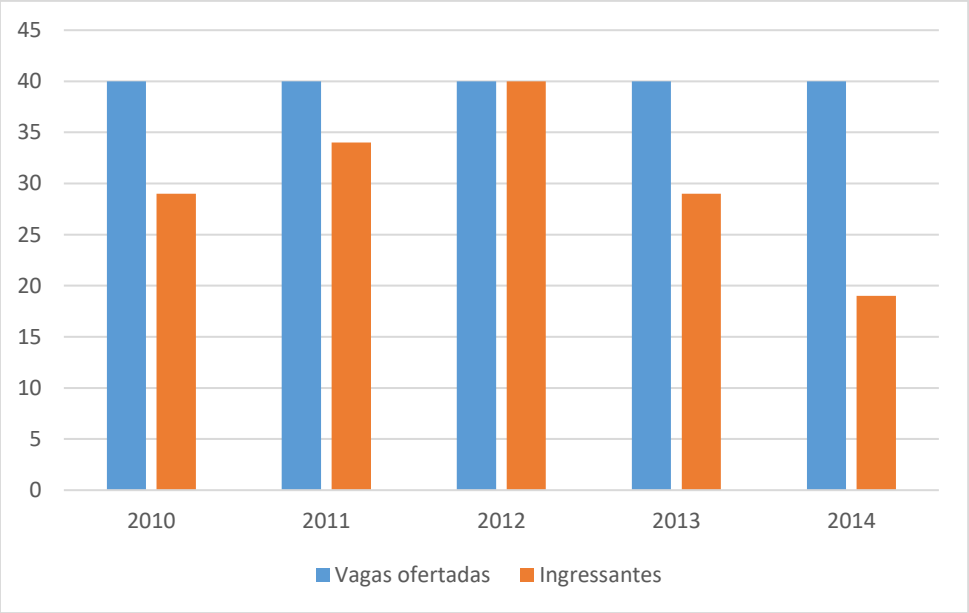


Gráfico 2: Número de vagas ofertadas por ano e número de alunos ingressantes por ano no curso de Sistemas de Informação. Fonte: Sarcinelli (2014)

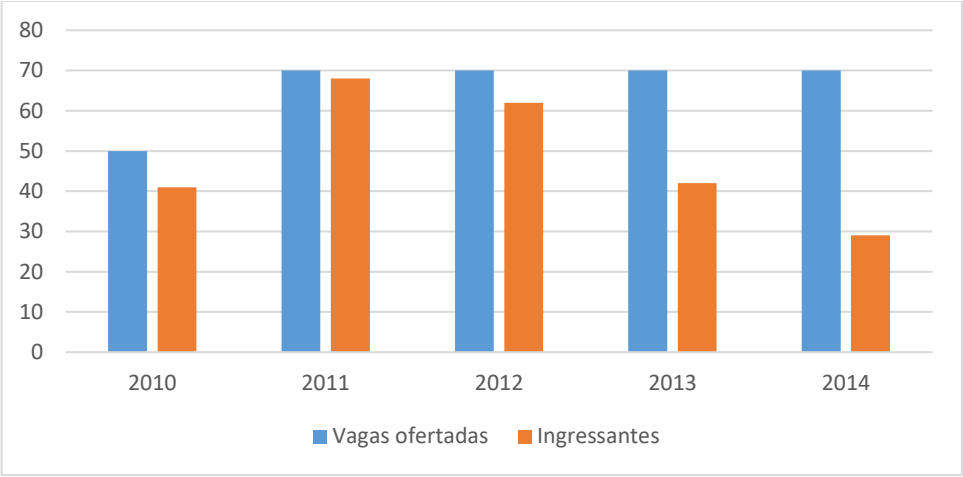


Tabela 1: Índice de preenchimento de vagas por ano nos cursos de Ciência da Computação e Sistemas de Informação. Fonte: Sarcinelli (2014)

Curso	Percentual de Vagas Preenchidas (%)				
	2010	2011	2012	2013	2014
CC	72,5	85,0	100,0	74,5	47,5
SI	82,0	97,1	88,6	60,0	41,4

Se percebe que a situação é ainda mais problemática ao analisar o número de alunos evadidos, poucos alunos se matriculam e muitos deles abandonam o curso (Tabela 2 e Tabela 3). A taxa de evasão é calculada dividindo-se o número de alunos evadidos pelo número de alunos matriculados. Ela pode ser vista como um indicador que mede a capacidade do curso de manter o aluno vinculado. Cursos com elevadas taxas de evasão podem ser considerados cursos com baixa capacidade de manter o aluno vinculado. Entende-se por alunos matriculados aqueles alunos que estejam cursando pelo menos uma disciplina (SARCINELLI, 2014).

Tabela 2: Taxa de evasão por período nos cursos de Ciência da Computação e Sistemas de Informação até 2012/1. Fonte: Sarcinelli (2014)

Curso	Taxa de Evasão (%)				
	2010/1	2010/2	2011/1	2011/2	2012/1
CC	9,1	3,3	1,8	6,1	23,1
SI	5,0	2,7	6,3	5,1	23,1

Tabela 3: Taxa de evasão por período nos cursos de Ciência da Computação e Sistemas de Informação depois de 2012/1. Fonte: Sarcinelli (2014)

Curso	Taxa de Evasão (%)				
	2012/2	2013/1	2013/2	2014/1	2014/2
CC	4,3	8,6	7,1	26,7	10,2
SI	24,6	5,8	6,1	6,7	3,6

Lobo (2012) apresenta algumas medidas bem-sucedidas que foram divulgadas internacionalmente:

1. Estabelecer um grupo de trabalho encarregado de reduzir a evasão: levantar níveis de satisfação dos alunos e estabelecer programas acadêmicos de integração e recuperação dos alunos novos.
2. Avaliar as estatísticas da evasão: levantar épocas críticas para a evasão e criar ações a partir dos achados.
3. Determinar as causas da evasão: comparar prioridades dos alunos com avaliação dos serviços educacionais, administrativos e comunitários.
4. Estimular a visão da IES centrada no aluno: envolver coordenadores, professores e funcionários de maneira obsessiva com o sucesso e bem-estar do aluno.
5. Criar condições que atendam aos objetivos que atraíram os alunos: não decepcionar os calouros é essencial.
6. Tornar o ambiente e o trânsito na IES agradáveis aos alunos: campus limpo e arrumado, com boas condições de trabalho e climáticas, também é cultura. Processos de Avaliação Institucional demostram o grau de insatisfação dos alunos com as condições de muitas IES.
7. Criar programa de aconselhamento e orientação dos alunos: tem que ser proativo e permanente.

Para realizar pesquisas, experimentos e análise do perfil dos alunos que evadem do curso, podem ser utilizadas técnicas de mineração de dados, como pré-processamento, transformação dos dados, classificação e validação.

A aplicação dessas técnicas sobre a base de dados necessita de técnicas e algoritmos de diferentes áreas do conhecimento, principalmente inteligência artificial (especialmente aprendizagem de máquina), banco de dados (recursos para manipular

grandes bases de dados) e estatística (comumente na avaliação e validação de resultados) (SILVA, 2004), conforme a Figura 1.

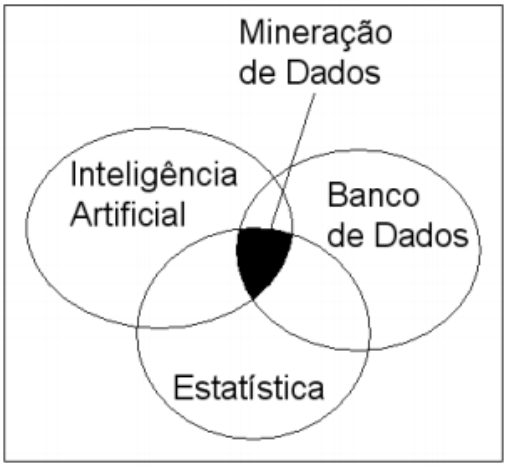


Figura 1: Mineração de dados é a interseção de diferentes áreas de conhecimento. Fonte: Silva (2004)

A ferramenta que se destaca para tal utilização dos dados é: WEKA (FRANK, 2016).

A ferramenta *WEKA* foi desenvolvida na Universidade de *Waikato* na Nova Zelândia. Os algoritmos podem ser aplicados diretamente a um conjunto de dados ou chamado a partir de seu próprio código Java. Contém ferramentas para pré-processamento de dados, classificação, regressão, *clustering*, regras de associação, e visualização. É também adequada para o desenvolvimento de novos sistemas de aprendizagem máquina e possui vídeo aulas gratuitas disponibilizadas pela universidade explicando como utilizar a plataforma (FRANK, 2016).

Amorim, Barone e Mansur (2008) apresentam as principais fases para implementação de um sistema de previsão testando a acurácia de três classificadores amplamente utilizados: J48, SMO e Rede Bayesiana. Com classificações corretas de mais de 80% e mostra estatísticas referentes a evasão em cada curso. Concentrando nos principais aspectos que possam levar um aluno a trancar o curso ou abandoná-lo.

J48 é uma árvore de decisão simples para a classificação. Com esta técnica, uma árvore binária é construída para modelar o processo de classificação. É aplicado a cada variável no banco de dados e ignora os valores em falta, ou seja, o valor para esse item pode ser previsto com base em o que se sabe sobre os valores de atributos para os outros registros. A ideia básica é dividir os dados em faixa com base nos

valores de atributo para esse item que se encontra na amostra de formação (PATIL; SHEREKAR, 2013).

O algoritmo de SMO é utilizado na implementação otimizada do classificador SVM do WEKA. Permite lidar com uma quantidade de arquivos maiores nos testes. Esse algoritmo transforma a saída do classificador em probabilidades através da aplicação de uma função sigmoide padrão. Os atributos nominais são convertidos em binários e todos os valores numéricos são normalizados. Ao final do processo, o algoritmo produz um modelo de aprendizagem, utilizado na classificação dos atributos (BARUFALDI et al., 2011).

Uma rede bayesiana é um método de representar uma probabilidade através do relacionamento entre proposições ou variáveis sempre que esta relação envolve incerteza ou imprecisão. A rede bayesiana foi implementada para sistemas especialistas probabilísticos. É utilizada para realizar tipos de inferência utilizando algoritmos e apresenta um ótimo desempenho além da facilidade de implementação (BELTRÃO1; SCHMITT; DIAS, 2002).

Dekker, Pechenizkiy e Vleeshouwers (2009) descrevem os resultados do estudo de mineração de dados educacional que visa prever a desistência de alunos de Engenharia Elétrica. Os resultados apresentam que classificadores bastante simples e intuitivos (árvores de decisão) identificaram já no primeiro ano letivo os alunos com risco de evasão, com acertos entre 75% e 80%. Também mostram algumas maneiras de melhorar ainda mais a previsão sem ter que coletar dados adicionais sobre os alunos.

Manhães et al. (2011) com objetivo de identificar precocemente alunos em risco de evasão, apresenta um estudo utilizando dados acadêmicos de alunos de graduação de uma universidade brasileira (UFRJ), avaliando a técnica de mineração de dados por meio de três experimentos nos quais foram aplicados dez algoritmos de classificação sobre uma base de dados dos alunos. Sendo percebido que com as primeiras notas semestrais dos calouros é possível identificar com precisão 80% a situação final do aluno no curso, e que o atributo mais importante é o coeficiente de rendimento do primeiro semestre letivo, o segundo é a nota na disciplina de Cálculo Diferencial e Integral I.

Rigo, Cazella e Cambruzzi (2012) apresentam uma análise de melhorias possíveis na aplicação de mineração de dados educacionais para que seus resultados

possam reter os efeitos que estão relacionados com questões financeiras e com a diminuição do número de alunos formados no ensino superior, que geram impactos para toda a cadeia produtiva do país e destaca que a utilização de conjuntos combinados de técnicas como aprendizado de máquina e mineração de dados.

4 METODOLOGIA

Optou-se em utilizar o WEKA, pois a equipe de desenvolvimento tem lançado periodicamente correções e releases do software, além de manter uma lista de discussões sobre a ferramenta. Grande parte de seus componentes de software são resultantes de teses e dissertações de grupos de pesquisa da universidade de *Waikato* na Nova Zelândia. Também possui uma interface gráfica amigável e seus algoritmos fornecem relatórios com dados analíticos e estatísticos do domínio minerado (SILVA, 2004).

A principal característica do WEKA está voltada na classificação de atributos por intermédio de seus algoritmos de aprendizagem. No entanto, também inclui seletores de atributos, e implementações de regras com associação e agrupamento de dados (clusters) (BARUFALDI et al., 2011).

Os classificadores J48, SMO e Rede Bayesiana foram escolhidos por causa da grande utilização deles no contexto da evasão escolar e nos trabalhos estudados na revisão de literatura.

Foi elaborado um questionário baseado nos trabalhos de Ribeiro (2005) e Alves e Pereira (2012), sendo aplicado para 71 alunos ingressantes nos cursos de Ciência da Computação e Sistemas de Informação que responderam nas primeiras semanas de aulas nos períodos de 2014/2, 2015/2 e 2016/2. Este formulário foi elaborado no Google Drive (Apêndice A). Para completar os dados os coordenadores dos cursos forneceram o coeficiente regular dos alunos e quais deles de fato evadiram.

Com isso, os dados foram tratados e segundo Silva (2004), seguindo o primeiro passo de pré-processamento dos dados, foi usada a técnica de limpeza, que consiste em suprir valores ausentes, reduzir discrepâncias de valores ruidosos e corrigir inconsistências dos dados. Assim, os poucos valores ausentes foram supridos manualmente baseados em outras respostas parecidas, caso em que o aluno respondia que o pai era falecido, mas não respondia se os pais eram casados, informava a matrícula, mas não o ano que entrou na UFES, assim como alguns tipos de inconsistências foram corrigidas, caso onde o aluno respondia que o curso é matutino sendo que era noturno, sendo assim, é de fácil identificação o erro ao preencher o formulário.

O segundo passo é a transformação dos dados para o formato de entrada do WEKA, em que foi feita a integração de dados (combinação de diferentes bases de

dados) e a transformação destes (modificações de formato e enriquecimento semântico) (SILVA, 2004).

O formulário do Google Drive nos oferece diversas formas de exportar os dados que foram obtidos, optou-se por exportar no *excel*, sendo assim as respostas ficaram de fácil conferência e modificação, após isso foi feita a integração do que já se tinha com os dados informados pelos coordenadores, a média das notas dos alunos conhecida como coeficiente regular e informação sobre a evasão desse aluno, com isso já tinha a base de dados completa.

Então, para maior facilidade em quantificar os dados, respostas do formulário e analisa-los foi feita a transformação dos mesmos, maioria das respostas passaram a obedecer um padrão em que a primeira resposta seria a 1, a segunda 2, e assim por diante, sendo assim teríamos apenas respostas simples para o WEKA reconhecer. Foi preciso tirar todo tipo de ambiguidade e respostas com acentos, para o sistema reconhecer corretamente a base de dados.

Também se decidiu em separar as perguntas onde pode marcar mais de uma resposta em várias entradas no WEKA, por exemplo cada matéria vira uma entrada na plataforma e quando o aluno marca que gosta daquela disciplina a entrada fica “sim”, e aquelas que ele não marcou fica como “não”. Nas figuras 2 e 3 podemos ver como fica o arquivo .arff, arquivo padrão do WEKA. No começo do arquivo definimos o @RELATION que são as perguntas que vão estar presentes nas respostas, e após isso definimos as respostas em @DATA.

```
1   @RELATION DADOS
2
3   @attribute matricula numeric
4   @attribute curso {SI,CC}
5   @attribute Q37acessoInternet {sim,no}
6   @attribute Q38facebook {sim,no}
7   @attribute Q41twitter {sim,no}
8   @attribute Q52fisica {sim,no}
9   @attribute Q54biologia {sim,no}
10  @attribute Q56artes {sim,no}
11  @attribute Q59sociologia {sim,no}
12  @attribute evade {sim,no}
13
```

Figura 2: Definição das perguntas no arquivo .arff

```
15  @DATA
16  2014204453,SI,sim,sim,no,sim,no,no,no,sim
17  2014204452,SI,sim,sim,sim,sim,no,no,no,sim
18  2014204216,SI,sim,sim,sim,sim,no,no,no,sim
19  2013203863,SI,sim,sim,no,no,no,sim,no,sim
20  2014204211,SI,sim,sim,no,sim,no,no,no,sim
21  2012204302,CC,sim,sim,no,no,no,no,no,no
22  2012204285,CC,sim,sim,no,sim,no,no,no,no
23  2010204097,CC,sim,sim,sim,no,no,no,no,no
24  2011204158,CC,sim,sim,no,sim,no,no,no,no
25  2015208132,CC,sim,sim,sim,no,no,no,no,no
26  2012204309,CC,sim,sim,no,sim,no,no,no,no
27  2012204327,SI,sim,sim,sim,sim,sim,sim,no,no
28  2010204114,SI,sim,sim,no,sim,no,no,no,no
29  2010204321,SI,sim,sim,no,sim,no,no,no,no
30  2010204313,SI,sim,sim,sim,no,sim,no,no,no
31  2012204320,SI,sim,sim,sim,no,sim,no,no,no
32  2011204429,SI,sim,sim,no,no,no,no,no,no
33  2013203829,SI,sim,sim,sim,no,no,no,no,no
34  2013203838,SI,sim,sim,no,no,no,no,no,no
```

Figura 3: Definição das respostas no arquivo .arff

O terceiro passo é a classificação dos dados utilizando os classificadores J48, SMO e Rede Bayesiana (AMORIM; BARONE; MANSUR, 2008) e a seleção de atributos que têm um peso maior, o WEKA disponibiliza vários algoritmos para esta categoria de mineração, dentre eles *InfoGainAttributeEval*, que informa a variância dos atributos (SILVA, 2004).

E por fim, a validação dos resultados, porque é muito importante que os resultados e modelos possam ser avaliados e comparados. Alguns elementos relevantes neste domínio: teste e validação, que fornecem parâmetros de validade e confiabilidade nos modelos gerados (*cross validation, suplied test set*, *use training set*, *percentage split*) (SILVA, 2004).

Na opção *Cross-validation,* os dados serão divididos em um número n de conjuntos definidos pelo usuário, conhecidos como partições, ou *folds*. Cada um destes conjuntos será utilizado uma vez como teste do modelo. Se o número de conjuntos definido foi, por exemplo, 5, significa dizer que haverá 5 testes sobre o modelo. É a forma mais recomendada para avaliar modelos (AMARAL, 2016).

Na opção *Suplied test set,* o usuário deve escolher um arquivo externo e ainda não carregado para testar o modelo. Dessa forma, o modelo será treinado nos dados carregados, e terá sua performance avaliada nos dados externos selecionados.

Ambos conjuntos de dados deverão ter a mesma estrutura, muito embora nos dados de teste, os valores de classe não necessitam estar preenchidos (AMARAL, 2016).

Ao escolher *Use training set,* o modelo será avaliado sobre o mesmo conjunto de dados em que foi treinado. Em outras palavras, todos os registros são utilizados para construção do modelo, e em seguida os mesmos registros são usados para avaliar a performance do modelo construído (AMARAL, 2016).

A última opção, *Percentage Split* é também conhecida como *hold out* e é uma das mais simples e comuns. Neste modelo um percentual definido pelo usuário dos dados é usado para treinar o modelo, e o restante para testá-lo. Normalmente, utiliza-se aproximadamente 30% para teste. O WEKA por padrão traz o percentual de treino preenchido com 66%, o que significa que 34% serão usados para teste se este percentual for mantido (AMARAL, 2016).

Com isso, para catalogar as características e as variáveis e aplicá-las sobre questionários não utilizados na etapa anterior, é comparado os resultados com os dados de alunos que de fato evadiram, para analisar o comportamento do sistema.

5 ANÁLISE DOS RESULTADOS

Ao iniciar os testes, foram usados todos os atributos, ao todo 59, sendo esses eles todas perguntas do formulário aplicado aos calouros e as respostas, ao todo 71 instancias, em que os dados são divididos em 2 classes, alunos que evadem e que não evadem, com isso obtemos as acurácias dos classificadores J48, SMO e Rede Bayesiana (AMORIM; BARONE; MANSUR, 2008) (Tabela 4).

A acurácia é um indicador de quanto o modelo acertou a classificação da instancia, por padrão no WEKA o último atributo é a classe, em que fica definido se o aluno evadiu ou não, com isso, após a classificação das instâncias, a plataforma verifica se a classificação foi correta ou não e nos fornece a porcentagem de acerto.

Tabela 4: Acurácia dos classificadores com todas as perguntas

Método	Rede Bayesiana (%)	SMO (%)	J48 (%)
Training set	88,7324	95,7746	84,5070
Cross-validation	74,6400	67,6056	69,0141
Percentagem split 10%	76,5625	78,1250	78,1250
Suplied teste set	83,3300	70,8333	70,8333

Contudo, desde o começo a intenção era fazer um formulário menor, com as perguntas mais importantes, que definissem melhor o aluno desistente.

Após seguir o primeiro e segundo passos descritos por Silva (2004), duas perguntas foram eliminadas por todas as respostas serem iguais, não contribuindo assim para analisar os alunos, sendo elas: Possui algum tipo de deficiência física limitante? e Seus pais apoiam sua decisão de estudar?.

O terceiro passo, segundo Silva (2004), envolve a seleção de atributos. O WEKA disponibiliza o algoritmo para esse propósito, o *InfoGainAttributeEval* que nos fornece a variância dos atributos (Tabela 5 e 6), em que foi possível definir a variância das perguntas.

Tabela 5: Perguntas com peso maior que 0,01

Curso?	0,0473194
Gostava de Artes na escola?	0,0436354
Possui uma conta no twitter?	0,0379318
Possui acesso à internet onde mora?	0,0335913
Gostava de Sociologia na escola?	0,0312410

Gostava de Física na escola?	0,0226925
Possui uma conta no facebook?	0,0137150
Gostava de Biologia na escola?	0,0102985

Tabela 6: Perguntas com peso maior que 0,0 e menor que 0,01

Gostava de História na escola?	0,0090480
Seu irmão já fez algum curso superior?	0,0083725
Ingressou pelo sistema de reserva de vagas?	0,0080884
Segue amigos no twitter?	0,0041510
Gostava de Português na escola?	0,0035462
Gostou da cidade onde está estudando?	0,0033200
Gostava de Matemática na escola?	0,0022139
Gostava de Química na escola?	0,0020217
Segue páginas de humor no twitter?	0,0008814
Gostava de Geografia na escola?	0,0008806
Gostava de Espanhol na escola?	0,0007169
Tem irmãos?	0,0004041
Possui outras especializações?	0,0002934
Gostava de Inglês na escola?	0,0002874
Sexo?	0,0002005
Segue políticos no twitter?	0,0001254
Segue famosos no twitter?	0,0000865
Gostava de Educação Física na escola?	0,0000196

Demais perguntas foram eliminadas por terem variância 0,0.

Decidiu-se por diminuir a quantidade de perguntas para que o formulário ficasse menor e mais rápido de ser respondido, com isso, foi utilizado a variância previamente calculada no terceiro passo e decidiu-se utilizar apenas as perguntas com peso maior que 0,01 (Tabela 5).

Tabela 7: Acurácia dos classificadores com perguntas com peso maior que 0,01 (Tabela 5)

Método	Rede Bayesiana (%)	SMO (%)	J48 (%)
Training set	83,0986	81,6901	84,507
Cross-validation	78,8732	76,0563	76,0563
Percentagem split 50%	85,7143	85,7143	82,8571
Suplied teste set	83,3333	83,3333	79,1667

Pode-se notar que usando as perguntas com maior peso (Tabela 5), os classificadores se mostram mais eficientes na maioria dos métodos:

Gráfico 3: Rede Bayesiana

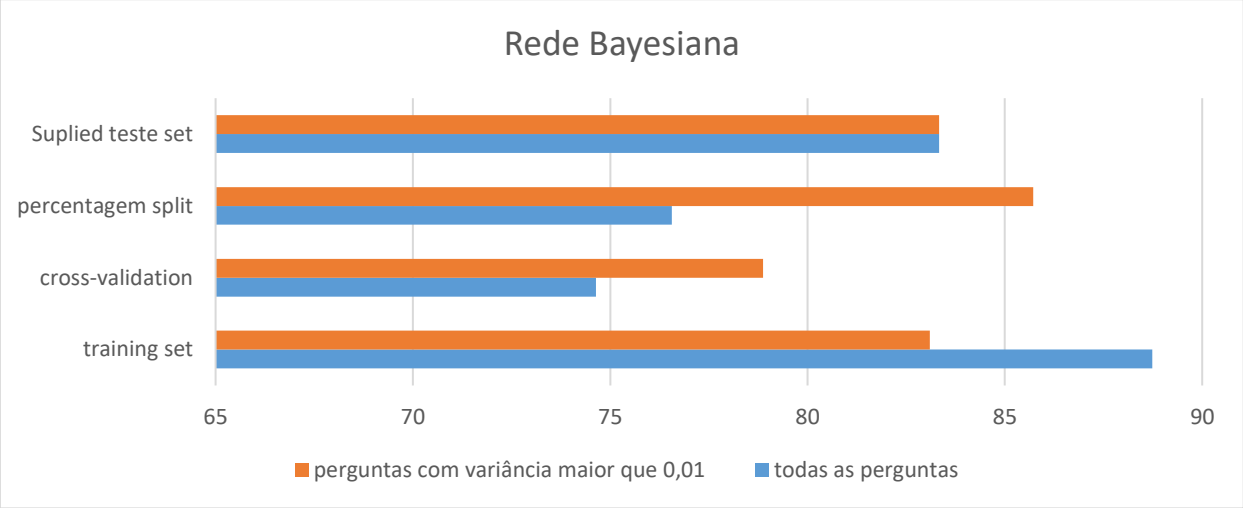


Gráfico 4: SMO

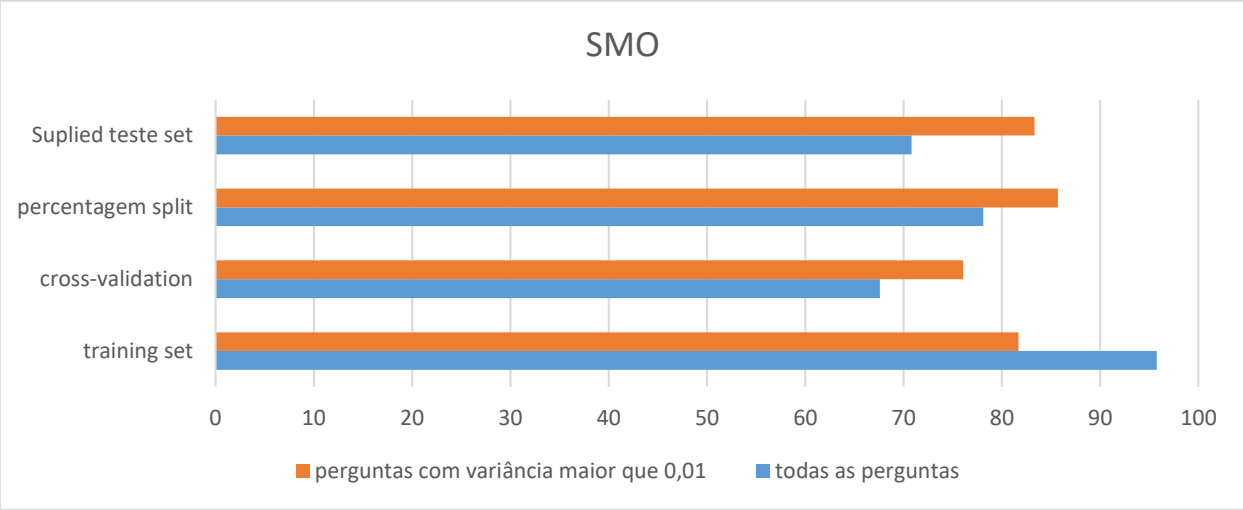
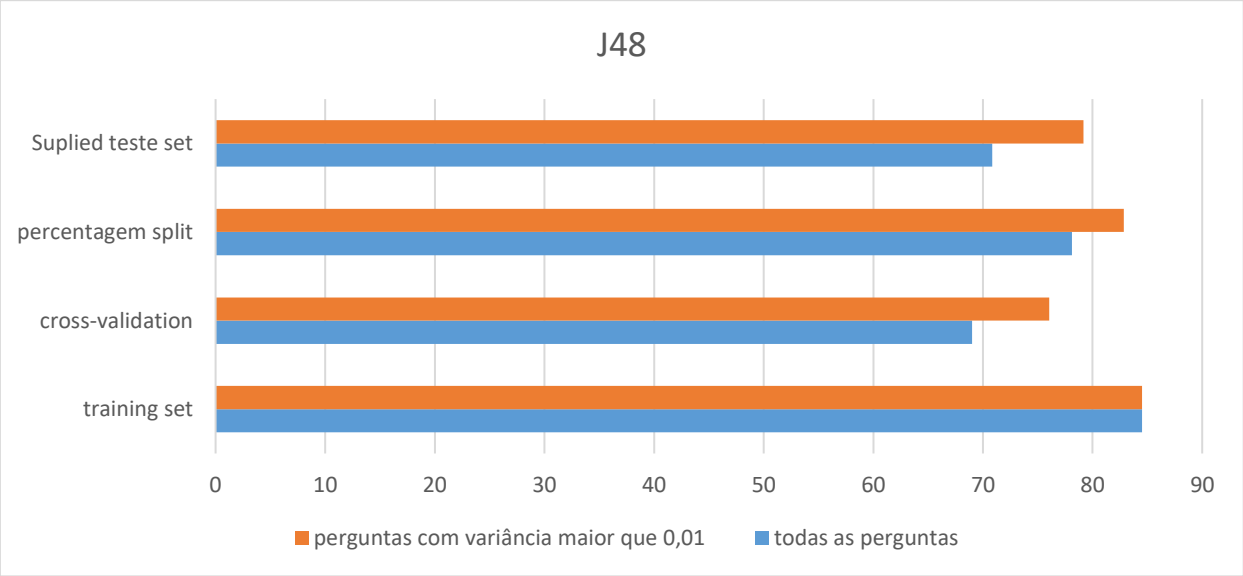


Gráfico 5: J48



Com isso, podemos entender melhor o perfil do aluno que evade e daquele que conclui o curso, utilizando apenas as 5 perguntas com maior variância (Tabela 8), e apesar de algumas respostas da pergunta “Quais matérias mais gostava no Ensino Médio?” apresentarem maior variância, decidiu por manter a pergunta geral, podendo ser marcadas todas as respostas, português, matemática e assim por diante, ou podendo não marcar nenhuma opção.

Tabela 8: Perguntas que são decisivas para os classificadores

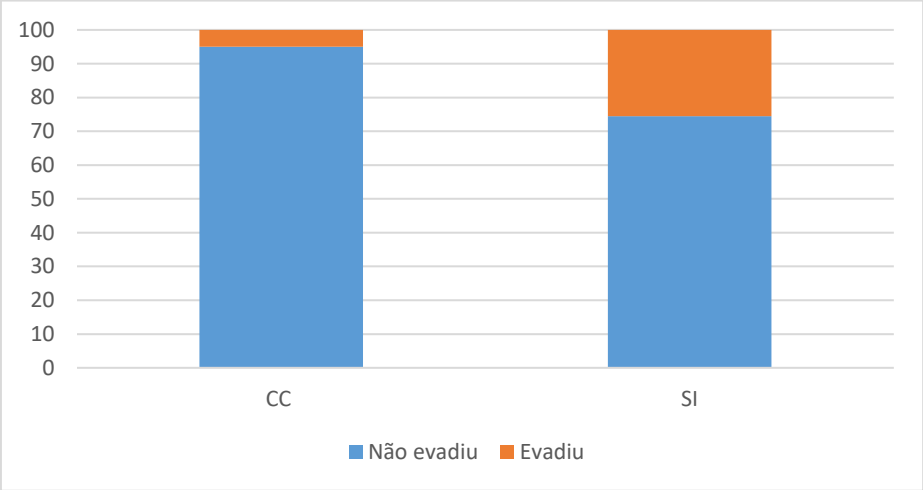
Curso?
Quais matérias mais gostava no Ensino Médio?
Você possui acesso à internet onde mora?
Você possui uma conta no facebook?
Você possui uma conta no twitter?

Após o quarto passo descrito por Silva (2004) como sendo a validação dos dados, é possível traçar o perfil dos alunos.

5.1 Perfil dos alunos

De forma a se ter uma melhor visão foi realizada uma abordagem quantitativa sobre os dados obtidos, cruzando os dados tabelados. 71 alunos ingressantes nos cursos de Ciência da Computação e Sistemas de Informação, sendo 20 de CC e 51 de SI.

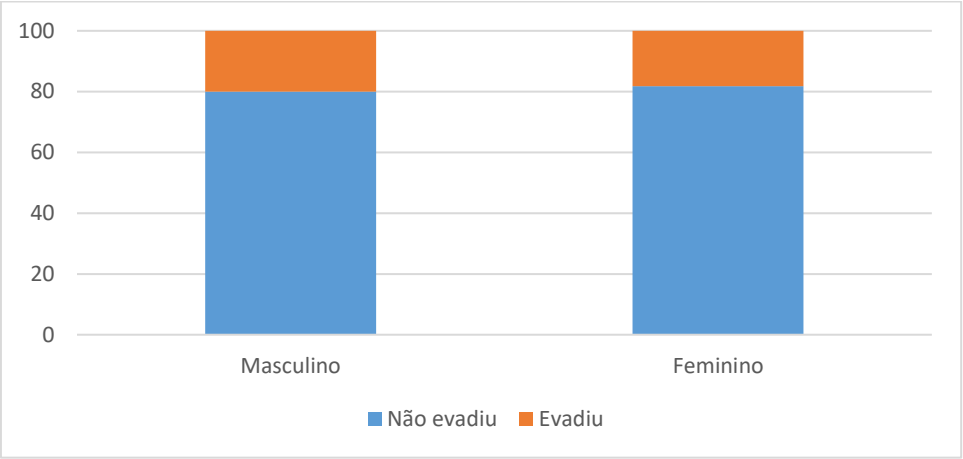
Gráfico 6: Porcentagem de alunos por curso



Pode-se perceber que a maior taxa de evasão está presente no curso de SI (Gráfico 6).

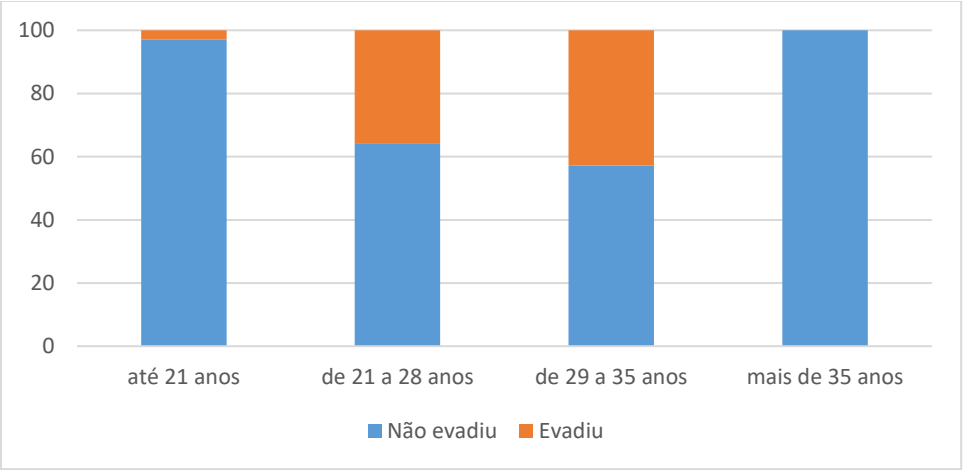
Ao analisar por sexo, a diferença se torna quase imperceptível (Gráfico 7).

Gráfico 7: Porcentagem de alunos por sexo



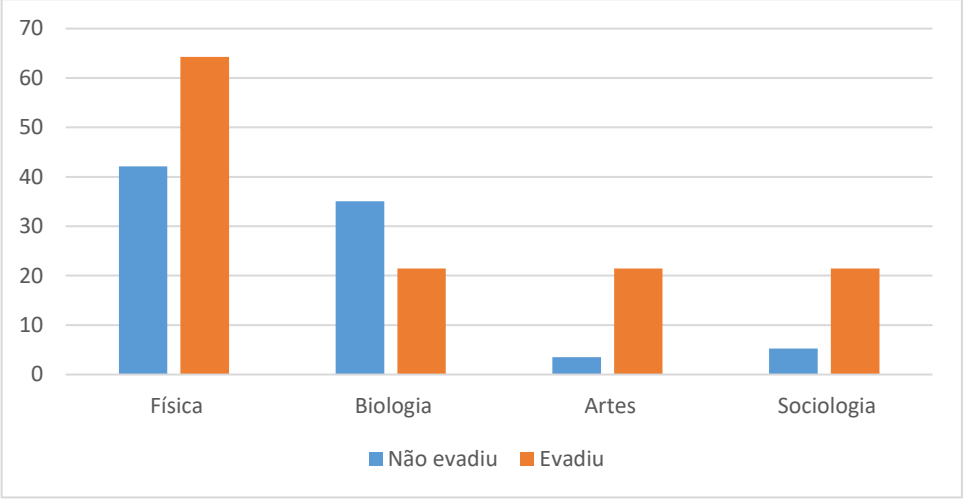
Quanto a idade, entre 21 e 35 anos está presente a maior taxa de evasão (Gráfico 8).

Gráfico 8: Porcentagem de alunos por idade



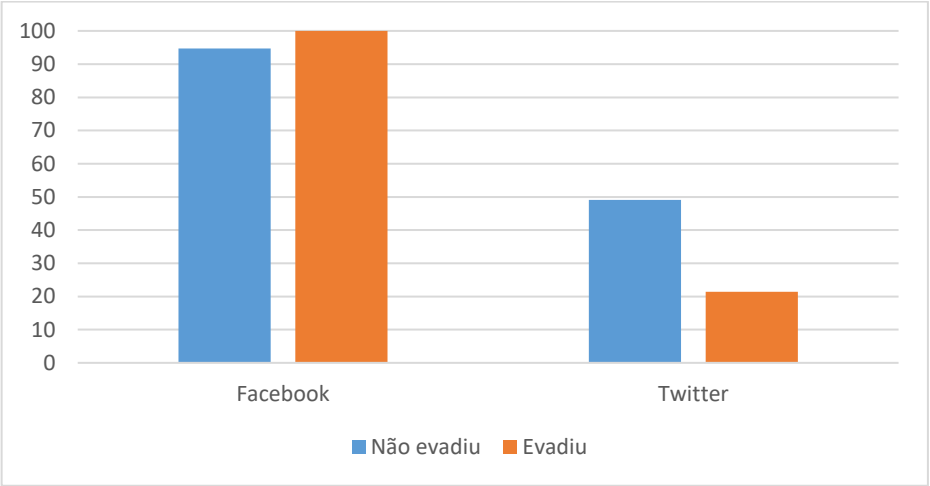
Ao analisar as matérias que os alunos gostavam no Ensino Médio pode-se perceber que as com maior variância, Física, Biologia, Artes e Sociologia, os alunos que evadem possuem uma taxa maior dos que gostam de Física, Artes e Sociologia, enquanto os que não evadem apresentam uma taxa maior para Biologia (Gráfico 9).

Gráfico 9: Porcentagem de alunos que gostam de Física, Biologia, Artes e Sociologia



As redes sociais se mostraram decisivas para os classificadores, sendo assim, pode-se observar que os alunos que evadem possuem mais afinidade como facebook, enquanto os alunos que não evadem possuem a maior taxa no twitter (Gráfico 10).

Gráfico 10: Porcentagem de alunos que possuem Facebook e Twitter



6 CONCLUSÃO

Portanto, ao utilizar os classificadores J48, SMO e Rede Bayesiana que são os mais usados para a análise da evasão, percebeu-se que é possível identificar alunos com perfil de evasão no primeiro período, alunos do curso de Sistemas de Informação, com mais de 21 anos e que gostam de Física, Artes e Sociologia, que são matérias não relacionadas diretamente aos cursos de computação, são os com mais risco de deixar o curso.

Por fim, sugere-se uma forma mais otimizada de tratar os dados, visto que neste trabalho o volume de dados era reduzido e pôde-se processar e transformar os dados utilizando o *excel*, assim, deixar disponível aos alunos do primeiro período o formulário, assim que eles responderam o modelo analisa e responde em que perfil esse aluno se encaixa.

Também entende-se ser importante trabalhar com os alunos que não se identificam com matérias fundamentais para o curso, como matemática, e fazer uma revisão para aqueles que sentem dificuldades ou mesmo para alunos que tiverem uma educação mais fraca.

7 REFERÊNCIAS BIBLIOGRÁFICAS

ALVES, A. P. V.; PEREIRA, S. B. S. **A evasão escolar na modalidade de ensino a distância: o polo presencial de Itapemirim–es**. Anais do SIED: EnPED-ISSN 2316-8722, v. 1, n. 1, 2012.

AMARAL, F. **Aprenda Mineração de Dados: Teoria e prática**. Alta Books Editora, 2016.

AMORIM, M. J.; BARONE, D.; MANSUR, A. U. **Técnicas de aprendizado de máquina aplicadas na previsão de evasão acadêmica.** In: Anais do Simpósio Brasileiro de Informática na Educação. [S.l.: s.n.], 2008. v. 1, n. 1, p. 666–674.

AZEVEDO, L. A.; SANTOS, Y. S. **Mineração de dados aplicada ao estudo da evasão e desempenho dos alunos do bacharelado em ciência da computação da universidade de Brasília**. 2015.

BARDAGI, M. P.; HUTZ, C. S. **Não havia outra saída: percepções de alunos evadidos sobre o abandono do curso superior**. PsicoUSF, Universidade São Francisco, v. 14, n. 1, p. 95–105, 2009.

BARUFALDI, B.; SANTANA, E. F.; BATISTA, L. V.; et al. **Sistema para controle de qualidade de mamógrafos através da análise computacional de imagens de phantom utilizando a ferramenta weka.** WVC 2011, p. 16, 2011.

BELTRÃO, C. J.; SCHMITT, E. J.; DIAS, J. S. **Utilização de rede bayesiana para auxílio de tomada de decisão na triagem de atendimento odontológico.** 2002.

DEKKER, G.; PECHENIZKIY, M.; VLEESHOUWERS, J. **Predicting students drop out: a case study. In: Educational Data Mining 2009**. [S.l.: s.n.], 2009.

FRANK, E.; HALL, M. A.; WITTEN, I. H. **The WEKA Workbench. Online Appendix for "Data Mining: Practical Machine Learning Tools and Techniques",** Morgan Kaufmann, Fourth Edition, 2016.

HAN, J.; KAMBER, M. **Data mining: concepts and techniques. United States of America: Morgan Kauff mann Publishers**, 2001.

LOBO, M. B. C. M. **Panorama da evasão no ensino superior brasileiro: aspectos gerais das causas e soluções**. Associação Brasileira de Mantenedoras de Ensino Superior. Cadernos, n. 25, 2012.

MANHÃES, L. M. B., DA CRUZ, S. M. S., COSTA, R. J. M.; et al. **Previsão de estudantes com risco de evasão utilizando técnicas de mineração de dados**. Anais do XXII SBIE-XVII WIE, Aracaju, 2011.

MARTINS, L. C.; LOPES, D. A.; RAABE, A. **Um assistente de predição de evasão aplicado a uma disciplina introdutória do curso de ciência da computação**. In: Anais do Simpósio Brasileiro de Informática na Educação. [S.l.: s.n.], 2012. v. 23, n. 1.

MOTEJUNAS, P. R.; SILVA FILHO, R. L. L.; HIPÓLITO, O.; et al. **A evasão no ensino superior brasileiro.** *Cadernos de Pesquisa*, SciELO Brasil, v. 37, n. 132, p. 641–659, 2007.

PATIL, T. R.; SHEREKAR, S. **Performance analysis of naive bayes and j48 classification algorithm for data classification.** ***International Journal of Computer Science and Applications***, v. 6, n. 2, p. 256–261, 2013.

REINERT, J. N.; GONÇALVES, W. J. **Evasão escolar: percepção curricular como elemento motivador no ensino para os cursos de administração–estudo de caso**. INPEAU, 2010.

RIBEIRO, M. A. **O projeto profissional familiar como determinante da evasão universitária: um estudo preliminar.** *Revista Brasileira de Orientação Profissional*, Associação Brasileira de Orientadores Profissionais, v. 6, n. 2, p. 55–70, 2005.

RIGO, S. J.; CAZELLA, S. C.; CAMBRUZZI, W. **Minerando dados educacionais com foco na evasão escolar: oportunidades, desafios e necessidades**. In: *Anais do Workshop de Desafios da Computação Aplicada à Educação*. [S.l.: s.n.], 2012. p.168–177.

SARCINELLI, M. R. **Ufes – indicadores de desempenho**. 2014. Disponível em: <http://www.proplan.ufes.br/indicadores>. Acesso em: 5 novembro 2015.

SILVA, G. P. D. **Análise de evasão no ensino superior: Uma proposta de diagnóstico de seus determinantes**. SciELO Brasil, 2013.

SILVA, M. P. d. S. **Mineração de dados: Conceitos, aplicações e experimentos com weka**. *Sociedade Brasileira de Computação*, v. 1, 2004.

APÊNDICE A - FORMULÁRIO DE CALOUROS

Respostas são mantidas em sigilo, e não é realizada a identificação do aluno (nome <-> matrícula).

*Obrigatório

( ) -> aceita apenas 1 resposta

| | -> pode marcar quantas respostas quiser

Questões pessoais

1 - Número de matrícula *

_____

2 - Curso *

( ) SI

( ) CC

3 – Sexo

( ) Feminino

( ) Masculino

4 - Idade

( ) ate 21 anos

( ) de 22 a 28 anos

( ) de 29 a 35 anos

( ) mais de 35 anos

5 - Possui algum tipo de deficiência física limitante?

( ) Auditiva

( ) Física

( ) Visual

( ) Múltipla

( ) Nenhuma

6 - Estado civil

( ) Solteiro(a)

( ) Casado(a)/ mora com um(a) companheiro(a)

( ) Divorciado(a)

( ) Viúvo(a)

7 - Quantidade de filhos

( ) Nenhum

( ) Um

( ) Dois

( ) Três

( ) Quatro ou mais

8 - Onde e com quem mora atualmente

( ) Em residência, com meus pais /parentes

( ) Em residência, com esposa(o)/ filho(s)

( ) Em casa ou apartamento, sozinho(a)

( ) Em habitação coletiva: hotel, hospedaria, quartel, pensionato, república

( ) Outra:

9 - Indique quantas pessoas habitam na residência

( ) Moro sozinho(a)

( ) Duas pessoas

( ) Três

( ) Quatro

( ) Cinco ou mais

10 - Assinale a situação abaixo que melhor descreve seu caso

( ) Meus gastos são financiados pela família

( ) Trabalho e recebo ajuda da família

( ) Trabalho, me sustento e/ou contribuo com o sustento da família

( ) Recebo bolsa na Universidade

( ) Recebo bolsa na Universidade e ajuda externa

( ) Outra:

11 - A sua formação escolar (ensino fundamental e Médio) realizou-se em:

( ) Toda na escola pública

( ) Toda na escola particular

( ) Parte na escola pública e parte na particular

( ) Escola pública ou particular através de ensino supletivo ou educação de jovens e adultos

( ) Escola pública diferenciada (rural, indígena ou quilombola)

12 - Meio de transporte/locomoção mais utilizado para ir a universidade:

( ) Carro

( ) Moto

( ) Ônibus/ van

( ) Bicicleta

( ) A pé/ carona

13 - Onde você reside durante a semana?

( ) Na cidade onde estudo

( ) Em cidade diferente da que estudo

14 - Por que você optou por estudar na UFES?

( ) Por ser uma Universidade Pública e gratuita

( ) Devido a localização geográfica

( ) Por oferecer o curso pretendido

( ) Por indicação de terceiros ou mídia

( ) Pela qualidade do curso/instituição

15 - Em que ano ingressou na UFES?

_____

16 - Qual (ais) matéria (s) mais gostava no ensino médio?

| | Português

| | Matemática

| | História

| | Geografia

| | Física

|| Química

| | Biologia

| | Educação Física

| | Artes

| | Espanhol

| | Inglês

| | Sociologia

17 - Ingressou pelo sistema de reserva de vagas?

( ) Sim

( ) Não

18 - Possui outras especializações?

( ) Sim

( ) Não

19 - Qual seu principal motivo na escolha do curso?

( ) Gostava da área

( ) Mercado de trabalho atraente

( ) Visando concurso

( ) Era sua segunda opção no vestibular

( ) Por incentivo da família

( ) Baixa concorrência

( ) Visando mais conhecimento

( ) Trabalhava na área

( ) Outra:

20 - Em que turno você estuda na UFES?

( ) Matutino

( ) Vespertino

( ) Noturno

( ) Integral

Questões familiares

21 - Qual a sua renda mensal familiar total?

( ) Até 2 salários mínimos (até R$ 1.448,00 inclusive)

( ) De 2 a 5 salários mínimos (acima de R$ 1.448,00 até R$ 3.620,00 inclusive)

( ) De 5 a 10 salários mínimos (de R$ 3.620,00 até R$ 7.240,00 inclusive)

( ) De 10 a 30 salários mínimos (de R$ 7.240,00 até R$ 21.720,00 inclusive)

( ) Mais de 30 salários mínimos (mais de R$ 21.720,00)

22 - Situação dos pais

( ) Casados

( ) Divorciados/ não estão juntos

23 - Situação do pai

( ) Contribui para renda familiar

( ) Falecido

( ) Não conheço

( ) Não convivo

( ) Outra:

24 - Situação do pai no mercado de trabalho

( ) Empregado

( ) Desempregado

( ) Estudante

( ) Aposentado

( ) Outra:

25 - Qual foi o grau de instrução mais alto que seu pai obteve? / Qual o último ano de escola que o seu pai cursou?

( ) Não estudou / Primário incompleto

( ) Primário completo / Ginasial incompleto

( ) Ginasial completo / Colegial incompleto
( ) Colegial completo / Universitário incompleto
( ) Universitário completo
( ) Mestrado
( ) Doutorado
( ) Não sei

26 - Situação da mãe
( ) Contribui para renda familiar
( ) Falecida
( ) Não conheço
( ) Não convivo
( ) Outra:

27 - Situação da mãe no mercado de trabalho
( ) Empregada
( ) Desempregada
( ) Estudante
( ) Aposentada
( ) Outra:

28 - Qual foi o grau de instrução mais alto que sua mãe obteve? / Qual o último ano de escola que a sua mãe cursou?
( ) Não estudou / Primário incompleto
( ) Primário completo / Ginasial incompleto
( ) Ginasial completo / Colegial incompleto
( ) Colegial completo / Universitário incompleto
( ) Universitário completo
( ) Mestrado
( ) Doutorado
( ) Não sei

29 - Tem irmãos?

( ) Sim

( ) Não

30 - Algum deles faz ou já fez algum curso superior?

( ) Sim

( ) Não

Questões gerais

31 - Seus pais apoiam sua decisão de estudar?

( ) Sim

( ) Não

32 - Você gostou da cidade onde está estudando?

( ) Sim

( ) Não

33 - Nota de avaliação para a cidade:

_____

34 - Você possui computador (PC)? Quantos?

_____

35 - Você possui laptop? Quantos?

_____

36 - Você possui smartphones? Quantos?

_____

37 - Você possui acesso à Internet onde mora (republica ou em casa)?

( ) Sim

( ) Não

Questões virtuais

38 - Você possui uma conta no Facebook?

( ) Sim

( ) Não

39 - Quanto tempo você se conecta por dia?

( ) até 2 horas

( ) de 2 horas até 5 horas

( ) mais de 5 horas

40 - Qual é a quantidade de contatos que você possui aproximadamente?

( ) até 500 contatos

( ) de 500 contatos até 1000

( ) mais de 1000 contatos

41 - Você possui uma conta no twitter?

( ) Sim

( ) Não

42 - Quais tipos de pessoas você segue (políticos, músicos, etc.)?

| | Políticos

| | Amigos

| | Famosos

| | Páginas de humor

| | Outra: