UNIVERSIDADE FEDERAL DE PELOTAS
Centro de Desenvolvimento Tecnológico
Programa de Pós-Graduação em Computação

Dissertação

Geração de Modelos de Predição para Estudantes em Risco de Evasão em Cursos Técnicos a Distância Utilizando Técnicas de Mineração de Dados

Emanuel Marques Queiroga

Pelotas, 2017

Emanuel Marques Queiroga

Geração de Modelos de Predição para Estudantes em Risco de Evasão em Cursos Técnicos a Distância Utilizando Técnicas de Mineração de Dados

Dissertação apresentada ao Programa de Pós-Graduação em Computação da Universidade Federal de Pelotas, como requisito parcial à obtenção do título de Mestre em Ciência da Computação

Orientador: Prof. Dr. Cristian Cechinel
Coorientador: Prof. Dr. Ricardo Matsumura Araujo

Pelotas, 2017

Universidade Federal de Pelotas / Sistema de Bibliotecas
Catalogação na Publicação

Q3g Queiroga, Emanuel Marques

Geração de modelos de predição para estudantes em risco de evasão em cursos técnicos a distância utilizando técnicas de mineração de dados / Emanuel Marques Queiroga ; Cristian Cechinel, orientador ; Ricardo Matsumura Araujo, coorientador. — Pelotas, 2017.

93 f. : il.

Dissertação (Mestrado) — Programa de Pós-Graduação em Computação, Centro de Desenvolvimento Tecnológico, Universidade Federal de Pelotas, 2017.

1. Mineração de dados. 2. Predição. 3. Evasão. 4. Inteligência artificial. I. Cechinel, Cristian, orient. II. Araujo, Ricardo Matsumura, coorient. III. Título.

CDD : 005

Elaborada por Maria Inez Figueiredo Figas Machado CRB: 10/1612

AGRADECIMENTOS

Agradeço aos meus avós Ereni e João, sem eles nada em minha vida teria sido possível. Obrigado, por oferecer-me a oportunidade de estudar e nunca terem desistido de mim.

Agradeço imensamente a minha namorada Tainá, que me acompanhou ao longo dessa jornada e em tantos momentos foi privada de lazer e e das tão sonhadas viagens de férias, para que eu pudesse dar continuidade a esse trabalho.

Devo no mínimo um obrigado a minhas dindas, Nádia e Marcia, por todo o suporte que me deram em minha infância, por todos puxões de orelha que eu mereci e foram dados.

No meio acadêmico, gostaria de agradecer ao apoio incondicionalmente prestado pelos meus orientadores Prof. Dr. Cristian Cechinel e Prof. Dr. Ricardo Matsumura Araujo, que em nenhum momento desistiram deste projeto.

Meus sinceros agradecimentos ao Instituto Federal Sul-rio-grandense Campus Visconde da Graça e a Universidade Federal de Pelotas, pelo suporte dado para execução.

As pessoas mais felizes
não têm o melhor de tudo,
elas apenas fazem o melhor
com tudo o que têm.
— AUTOR DESCONHECIDO

RESUMO

QUEIROGA, Emanuel Marques. **Geração de Modelos de Predição para Estudantes em Risco de Evasão em Cursos Técnicos a Distância Utilizando Técnicas de Mineração de Dados**. 2017. 92 f. Dissertação (Mestrado em Ciência da Computação) – Programa de Pós-Graduação em Computação, Centro de Desenvolvimento Tecnológico, Universidade Federal de Pelotas, Pelotas, 2017.

A evasão é considerada um dos principais problemas relacionados com a Educação a Distância (EAD). Nessa modalidade de ensino, a interação entre estudantes e professores geralmente é mediada por um Ambiente Virtual de Aprendizagem (AVA), onde ficam registradas em seus logs de interações as ações realizadas pelos estudantes e professores durante o processo de ensino-aprendizagem. O grande volume de dados gerados por essas interações permite a utilização de técnicas de mineração para analisar os dados dos estudantes. Este trabalho aplica técnicas de mineração de dados e aprendizagem de máquina em logs das interações dos estudantes de cursos técnicos a distância dentro dos AVAs com o objetivo de identificar estudantes em situação de risco de evasão, utilizando como variáveis principais de entrada para os modelos de predição apenas a contagem dessas interações e atributos variados das mesmas. Foram utilizados dados de logs no AVA (contagens de interações e situação final dos estudantes) de quatro cursos técnicos EAD. As interações dos estudantes foram contabilizadas separadamente da seguinte forma: quantidade de interações diárias, soma das interações semanais, média semanal das interações, desvio padrão e situação final. Foram avaliados dois cenários diferentes, sendo eles: 1) Geração de modelos de predição com treinamento e teste utilizando dados do próprio curso e a partir de validação cruzada e 2) Treinamento dos modelos com dados de três cursos e teste dos modelos com dados do curso restante. No primeiro cenário, foram obtidos modelos de predição da evasão com ACG de até 84% antes da décima semana de curso, alcançado 95% até a semana 52. No segundo, a maioria dos modelos de predição apresentam resultados de até 80% nas primeiras dez semanas de curso alcançando 98% antes da metade do curso. Um dos modelos alcançou uma ACG de até 95% desde as primeiras semanas. Na comparação direta com o modelo estatístico, ambas as técnicas apresentaram resultados próximos nas primeiras semanas. Entretanto, a partir da décima semana, os modelos gerados por meio de mineração de dados apresentaram um crescimento significativo nas ACG, enquanto que o modelo estatístico se manteve estável. Assim a contribuição deste trabalho é a geração de modelos de predição que possam auxiliar de forma mais precisa no combate a evasão.

Palavras-chave: Mineração de dados, Predição, Evasão, Inteligência Artificial.

ABSTRACT

QUEIROGA, Emanuel Marques. **Generating predictive models for at-risk students in distance technical courses using data mining techniques**. 2017. 92 f. Dissertação (Mestrado em Ciência da Computação) – Programa de Pós-Graduação em Computação, Centro de Desenvolvimento Tecnológico, Universidade Federal de Pelotas, Pelotas, 2017.

Evasion is considered one of the main problems related to e-learning (EAD). In this teaching modality, the interaction between students and teachers is usually mediated by a Virtual Learning Environment (AVA), where the actions taken by students and teachers during the teaching-learning process are recorded in their interaction logs. The large volume of data generated by these interactions allows the use of mining techniques to analyze student data. This work applies data mining and machine learning techniques to logs of students' interactions of distance technical courses within AVA's in order to identify students at risk of evasion, using as main input variables for the prediction models Only the count of these interactions and varied attributes of them. Data from logs in the AVA (interaction counts and students' final situation) of four EAD technical courses were used. Student interactions were counted separately as follows: number of daily interactions, sum of weekly interactions, weekly mean of interactions, standard deviation and final situation. Two different scenarios were evaluated: 1) Generation of prediction models with training and test using data from the course itself and from cross validation and 2) Training of the models with data from three courses and test of the models with data from the course remaining. In the first scenario, prediction models of prediction of ACG evasion up to 84% were obtained before the tenth week of course, reaching 95% until week 52. In the second scenario, most of the prediction models present results of up to 80% In the first ten weeks of the course reaching 98% before the middle of the course. One of the models has achieved an ACG of up to 95% since the first few weeks. In the direct comparison with the statistical model, both techniques showed close results in the first weeks. However, from the tenth week, the models generated through data mining showed a significant growth in the GCA, while the statistical model remained stable. The contribution of this work is the generation of models able to early predict dropout students..

Keywords: Data mining, Prediction, Dropout, e-learning, artificial intelligence.

LISTA DE FIGURAS

Figura 1 Matriculas em cursos a Distância 2015 - Adaptado de CENSO (2015) 17
Figura 2 Quantitativos de estudantes EAD - Adaptado de CENSO (2015) . . 18
Figura 3 Descoberta do conhecimento - KDD. Adaptado de FAYYAD; PIATETSKY-SHAPIRO; SMYTH (1996) 23
Figura 4 Descoberta do conhecimento - KDD. Adaptado de HAN; PEI; KAMBER (2011) 26
Figura 5 Formulá teorema de Bayes 30
Figura 6 Conjunto de dados 32
Figura 7 Exemplo de árvore 32
Figura 8 Formúla da transformada lógica 33
Figura 9 Modelo de uma rede neural - Adaptado de SEGATTO; COURY (2008) 34
Figura 10 Partes de um dos arquivos ARFF utilizado no projeto 37

Figura 11 Fluxo de dados e sequência dos projetos 48

Figura 12 Cenário 1 56
Figura 13 Cenário 2 57
Figura 14 Resultado Cenário 1 Experimento final Curso 3 VP 58
Figura 15 Resultado Cenário 1 Experimento final Curso 3 VN 59
Figura 16 Resultado Cenário 1 Experimento final Curso 1 VN 60
Figura 17 Resultado Cenário 1 Experimento final Curso 1 VP 61
Figura 18 Resultado Cenário 2 Experimento final Curso 3 VP 62
Figura 19 Resultado Cenário 2 Experimento final Curso 3 VN 63
Figura 20 Resultado Cenário 2 Experimento final Curso 1 VP 63
Figura 21 Resultado Cenário 2 Experimento final Curso 1 VN 64
Figura 22 Resultado curso 1 - Verdadeiros Negativos Cenário 2 65
Figura 23 Árvore de decisão curso 1 - Semana 25 67
Figura 24 Árvore de decisão curso 1 - Semana 50 68
Figura 25 Árvore de decisão curso 1 - Semana 100 69

LISTA DE TABELAS

Tabela 1 Quadro de evolução da Mineração de Dados - Adaptado de VASCONCELOS; CARVALHO (2004) 20

Tabela 2 Comparativo entre trabalhos relacionados - Parte 1 44
Tabela 3 Comparativo entre trabalhos relacionados - Parte 2 45
Tabela 4 Comparativo entre trabalhos relacionados - Parte 3 46

Tabela 5 Quantitivo de dados utilizados 52
Tabela 6 Modelo de log do Moodle 53
Tabela 7 Variáveis Utilizadas 54

Tabela 8 Comparativo entre Experimentos do Autor 70
Tabela 9 Comparativo com os Resultados Obtidos pelo Estado da Arte . . . 72

LISTA DE ABREVIATURAS E SIGLAS

ABED	Associação Brasileira de Educação a Distância
ABNT	Associação Brasileira de Normas Técnicas
ARFF	Attribute Relation File Format
AVA	Ambiente Virtual de Aprendizagem
CaVG	Campus Visconde da Graça
CSV	Comma-separated values
EAD	Educação a Distância
EDM	Mineração de Dados educacionais
IDH	Índice de Desenvolvimento Humano
IFSul	Instituto Federal Sul-rio-grandense
LDB	Lei de Diretrizes e Bases da Educação
KDD	Descoberta de conhecimento em Bases de Dados(Knowledge Discovery in Databases)
MD	Mineração de Dados
MEC	Ministério da Educação
Moodle	Modular Object-Oriented Dynamic Learning Environment
PNUD	Plano das Nações Unidas para o Desenvolvimento
UFPel	Universidade Federal de Pelotas
VP	Verdadeiros Positivos
VN	Verdadeiros Negativos
WEKA	Waikato Environment for Knowledge Analysis

SUMÁRIO

1 INTRODUÇÃO 12
1.1 Objetivos e metas 14
1.2 Estrutura do texto 15

2 REFERENCIAL TEÓRICO 16
2.1 Educação a Distância 16
2.2 Evasão Escolar 18
2.3 Mineração de Dados 20
2.3.1 Seleção dos dados 23
2.3.2 Pré-processamento e limpeza dos dados 24
2.3.3 Mineração de dados educacionais 26
2.3.4 Aprendizagem de Máquina 29
2.3.5 Ferramentas para mineração de dados 35

3 TRABALHOS RELACIONADOS 39
3.1 Descrição dos trabalhos relacionados 39
3.2 Análise dos trabalhos relacionados 47

4 METODOLOGIA E EXPERIMENTOS 50
4.1 Contexto 50
4.2 Metodologia 51
4.2.1 Coleta 51
4.2.2 Pré-processamento dos dados 52
4.2.3 Modelo Estatístico descritivo 54
4.2.4 Geração e avalição dos modelos de predição 55
4.2.5 Configuração do experimento final 56
4.3 Resultados encontrados 57
4.3.1 Resultados Cenário 1 58
4.3.2 Resultados Cenário 2 61
4.3.3 Árvores de Decisão 65

5 DISCUSSÃO DOS RESULTADOS 70
5.1 Comparação com os Trabalhos Relacionados 71

6 CONSIDERAÇÕES FINAIS 74

REFERÊNCIAS 76

ANEXO A RESULTADOS EXPERIMENTO FINAL 82

ANEXO B CÓDIGO EM JAVA DE UTILIZAÇÃO DA BIBLIOTECA DO WEKA 90

1 INTRODUÇÃO

No atual contexto social tornou-se indispensável a busca por conhecimentos e qualificação das pessoas, de forma que nos últimos anos o governo brasileiro através do Ministério da Educação e entidades de fomento têm feito uma série de investimentos em programas de educação buscando a qualificação da mão de obra produtiva no país.

Considerando que o Brasil é um país de grandes dimensões, diversas cidades estão afastadas dos grandes centros universitários e acabam ficando isoladas de programas de graduação e cursos técnicos profissionalizantes. Desta forma, uma das alternativas adotadas pelo governo federal para a expansão do acesso a educação foi a utilização da modalidade à distância (Educação a Distância - EAD), que tem como um de seus objetivos levar o ensino a estas localidades, geralmente utilizando Ambientes Virtuais de Aprendizagem (AVAs) (DELANO; CORRêA, 2013).

O AVA é o "local virtual" onde os cursos na modalidade a distância, ou semipresenciais, normalmente acontecem. São ambientes que utilizam plataformas especialmente planejadas para abrigar cursos. Uma das plataformas mais utilizadas no país é o Modular Object-Oriented Dynamic Learning Environment (Moodle[1]).

No Moodle existem diversas áreas para apresentação de conteúdos em diversos formatos, atividades de verificação da aprendizagem e espaços para interação síncrona, por meio de chats, e assíncrona, através de fóruns de discussão. Tratam-se de recursos que permitem a interação dos estudantes entre si e com a equipe de tutores e professores. A organização do ambiente virtual permite ao aluno um acompanhamento organizado e sistematizado daquilo que é estudado a cada semana. A recuperação da informação e dos conteúdos estudados também é um dos benefícios proporcionados por cursos a distância que utilizam AVAs (SEGUNDO; RAMOS, 2005).

Um dos principais desafios da EAD é obter a diminuição do índice de evasão, que conforme o Censo EAD (CensoEAD, 2013), foi de 18,6% em 2010, 20,5% em 2011, 11,74% em 2012 e 16,94% em 2013 nos cursos autorizados pelo Ministério da Educação (MEC). Num contexto onde em 2013 haviam 5754 cursos autorizados pelo

[1] https://Moodle.org/

MEC e a taxa de matriculas anual foi de 882.843, temos em torno de 149.553 alunos evadidos.

BARROSO; FALCãO (2004) agrupam as condições desencadeantes para evasão em 3 classificações, i) econômica - impossibilidade de permanecer no curso por questões socioeconômicas; ii) vocacional – o aluno não se identifica com o curso; iii) institucional – abandono por fracasso nas disciplinas iniciais, inadequação aos métodos de estudo, dificuldades de relacionamento com colegas ou com membros da instituição.

Segundo MANHãES et al. (2011), a detecção precoce de grupos de alunos com risco de evasão é uma condição importante para reduzir o problema da evasão, uma vez que um tratamento mais adequado pode ser ofertado a esses alunos. Ainda segundo MANHãES et al. (2011), atualmente o processo de identificação desse grupo de alunos é manual, subjetivo, empírico e sujeito a falhas, pois depende primordialmente da experiência acadêmica e do envolvimento dos docentes. Geralmente, estes desempenham inúmeras atividades, portanto é difícil acompanhar e reconhecer as necessidades de cada aluno e identificar aqueles alunos que apresentam risco de evasão.

Com o grande volume de dados gerados pelos ambientes virtuais de aprendizagem, o trabalho de descoberta do conhecimento através da análise dessas informações sem uma ferramenta adequada se torna mais complexo, trabalhoso e dispendioso. Assim, tentando minimizar este problema e os apontados por MANHãES et al. (2011), a mineração de dados surge como uma alternativa no tratamento e para descoberta de conhecimento nessas bases.

A mineração de dados é a disciplina que estuda a descoberta de novas informações a partir da análise de grandes quantidades de dados, tendo como objetivo identificar relações e padrões nos dados e, assim, produzir novas informações. Estas informações podem propiciar a descoberta de novas regras ou padrões associados ao comportamento e assim possibilitar a predição de situações (BAKER; YACEF, 2009).

Como exemplo, poderíamos aplicar a mineração de dados nas informações de venda de um mercado em um determinado período, assim podendo identificar a relação entre os produtos e reorganizar o estoque para que os produtos com a relação mais próxima de venda fiquem em lugares estratégicos ou até mesmo planejar o estoque de determinados produtos em determinados períodos.

A mineração de dados também pode ser aplicada em problemas mais complexos, tais como: a predição de condições climáticas, a identificação de faces, a alocação de banda de um servidor, a predição de características do solo e a utilização de dados educacionais para descoberta de padrões e predição de comportamento.

A mineração de dados educacionais (Educational Data Mining - EDM) é uma área

de pesquisa recente e que tem como principal objetivo o desenvolvimento e aplicação de técnicas de mineração de dados na exploração de conjuntos de dados coletados em ambientes educacionais.

Atualmente a EDM vem se estabelecendo como uma forte e consolidada linha de pesquisa que possui grande potencial para melhorar a qualidade do ensino (BAKER; ISOTANI; CARVALHO, 2011). Essa área pode ajudar as instituições a criarem modelos de predição que tenham condições de avaliar as chances de um determinado acadêmico evadir.

A aplicação da EDM pode possibilitar o tratamento diferenciado entre os alunos, dedicando formas de auxílio diferenciadas a um determinado aluno que esteja com uma probabilidade maior de evasão.

1.1 Objetivos e metas

Este trabalho tem como objetivo estudar e aplicar as técnicas de mineração de dados e aprendizagem de máquina em dados disponíveis da EAD do Instituto Federal Sul-rio-grandense (IFSUL), propondo um modelo de predição para evasão de alunos baseado somente na contagem de interações e suas variações, assim possibilitando o emprego em diferentes domínios de aplicação.

O objetivo geral se desdobra nas seguintes metas específicas, a serem contempladas neste trabalho.

- Realizar um levantamento das pesquisas disponíveis na área de mineração de dados educacionais, principalmente para a predição de estudantes em risco de evasão e/ou reprovação.

- Documentar as principais teorias e conceitos aplicados na mineração de dados educacionais e principalmente nos dados de interações de alunos com ambientes de aprendizagem, bem como os algoritmos que apresentam os resultados mais satisfatórios na área e suas características.

- Gerar e testar modelos de predição para identificação de estudantes de cursos técnicos a distância em risco de evasão, utilizando somente contagem de interações e diferentes variações desta.

Considerando os aspectos apresentados, o foco deste trabalho é utilizar apenas contagem de interações ao longo do tempo, uma vez que esta é uma métrica facilmente generalizável para outras plataformas e abordagens de ensino, em contraste com abordagens que são extremamente específicas (e.g. utilizando tipos de interações que não necessariamente existem em todas plataformas ou são utilizadas em todas execuções de cursos).

1.2 Estrutura do texto

Essa dissertação esta constituída por este e mais seis capítulos, conforme estrutura descrita a seguir.

No capítulo 2 são apresentados os referenciais teóricos sistematizados que foram utilizados no desenvolvimento dessa dissertação, considerando o tema central de pesquisa.

No capítulo 3 são apresentados diversos trabalhos na área de mineração de dados educacionais enfocados no problema da predição de estudantes em risco, assim como as principais técnicas aplicadas por esses trabalhos e uma breve discussão sobre as diferenças entre os mesmos e os resultados obtidos.

No capítulo 4 são apresentados a metodologia e os experimentos desenvolvidos no decorrer desta dissertação. Para isto, é apresentado o contexto dos dados que foram utilizados para os experimentos, os processos utilizados para mineração de dados e a geração dos modelos propriamente dita.

No capítulo 5 são apresentados os resultados encontrados nos experimentos deste trabalho.

No capítulo 6 é feita a discussão dos resultados obtidos nesta dissertação e sua comparação com o estado da arte na predição de evasão.

O capítulo 7 apresenta as considerações finais bem como propostas para trabalhos futuros.

2 REFERENCIAL TEÓRICO

Este capítulo apresenta a sistematização de conceitos que foi feita considerando o tema central de pesquisa desta dissertação. Nesse sentido, são apresentados os conceitos sobre Educação a Distância, Evasão Escolar, Mineração de Dados, Mineração de Dados Educacionais (EDM) e ainda Aprendizagem de Máquina.

2.1 Educação a Distância

A educação a distância(EAD) é uma modalidade de ensino onde discentes e docentes estão separados fisicamente, ou seja, não estão no mesmo local físico. Atualmente a forma de mediação entre os discentes e docentes mais utilizada são os ambientes virtuais como o Moodle (LITTO; FORMIGA, 2011).

Historicamente a educação a distância faz parte da formação profissional e cultural de milhões de pessoas que não possam frequentar uma instituição de ensino presencial, por motivos como por exemplo, a dificuldade de acesso aos grandes centros profissionalizantes e/ou universitários. Entretanto, somente nas últimas décadas passou a fazer parte das atenções pedagógicas (MAIA; MATTAR, 2008).

Após as grandes guerras novas iniciativas de ensino se tornaram uma necessidade pelo aumento da demanda social por educação. Com o aperfeiçoamento das técnicas de transporte e comunicação como os serviços de correios, telégrafos, radiofônicos e até mesmo telefônico podendo ser aplicados diretamente na comunicação e informação não somente dos exércitos, mas também da sociedade, cresceu a demanda por uma forma de educação que pode-se levar às pessoas, afastadas dos grandes centros, educação de qualidade. Baseado em um modelo de ensino por correspondência, que surgiu na antiga União Soviética em 1922, a França criou seu serviço de ensino postal para atender estudantes deslocados pelo êxodo das guerras e o mesmo, em dois anos, passou a atender 350 mil usuários (JUNIOR, 2013).

Com a evolução dos meios de comunicação o ensino a distância utilizou integradamente o áudio e o videocassete, as transmissões de rádio e televisão, o videotexto, o computador e, com o advento da internet, a tecnologia de multimeios, que combinando

textos, sons, imagens, assim como mecanismos de geração de caminhos alternativos de aprendizagem (hipertextos, diferentes linguagens) e instrumentos para fixação de aprendizagem com feedback imediato (programas tutoriais informatizados) JUNIOR (2013).

Como sugere FARIA (2004), no século XXI necessitam as instituições de ensino estarem preparadas para interagir com uma geração mais atualizada e mais informada, porque os modernos meios de comunicação, liderados pela Internet, permitem o acesso instantâneo à informação e os alunos têm mais facilidade para buscar conhecimento por meio da tecnologia colocada à sua disposição. Os procedimentos didáticos, nesta nova realidade, devem privilegiar a construção coletiva dos conhecimentos, mediados pela tecnologia, na qual o professor é um partícipe pró-ativo que intermedeia e orienta esta construção.

Na Figura 1 constantes no Censo Brasileiro de educação a distância CENSO (2015), é possível analisar os números de matriculas registradas em cursos de educação a distância, nos diferentes formatos disponíveis no Brasil atualmente.

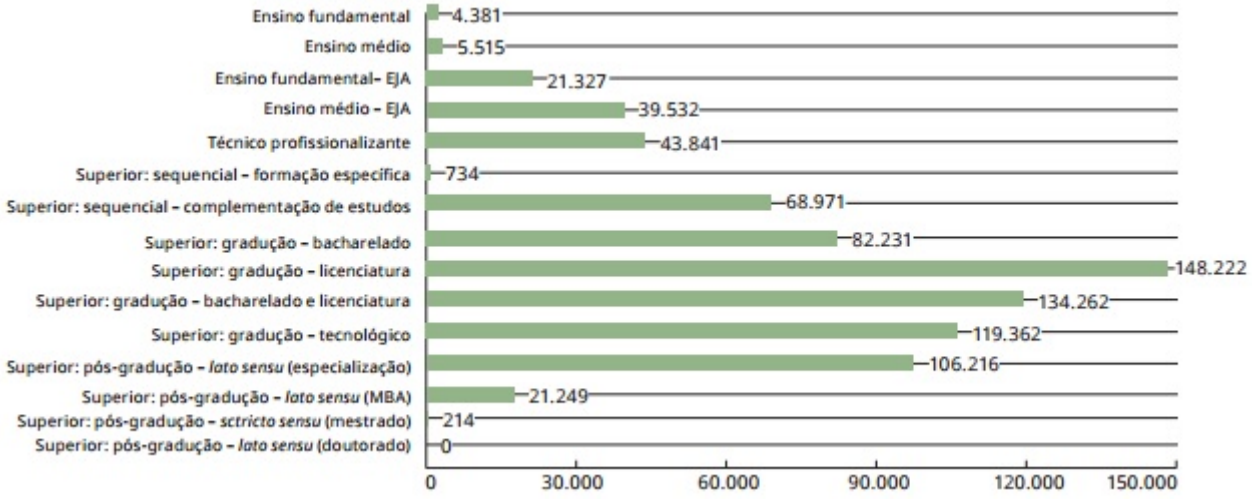

Figura 1: Matriculas em cursos a Distância 2015 - Adaptado de CENSO (2015)

A Figura 2, apresenta os quantitativos de estudantes que em 2015 estavam cursando algum tipo de curso a distância homologado pelo Ministério da Educação.

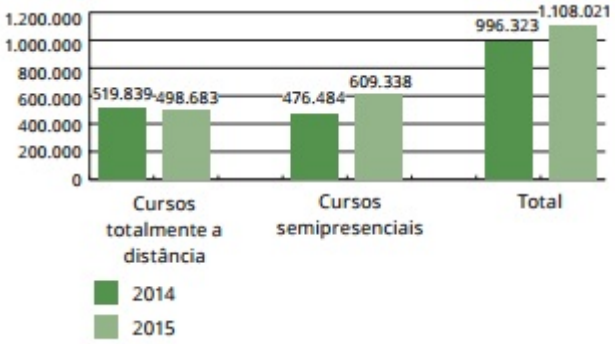

Figura 2: Quantitativos de estudantes EAD - Adaptado de CENSO (2015)

Segundo a Associação Brasileira de Educação a Distância (ABED), os três principais problemas enfrentados pela Educação a Distância no Brasil são a evasão, a resistência dos próprios educadores e a dificuldade de adaptação dos estudantes a essa modalidade de ensino CENSO (2015).

Como esse trabalho tem como um de seus objetivos a geração de modelos de predição que auxiliem na identificação de estudantes em risco de evasão, esse problema será abordado na próxima seção desse trabalho.

2.2 Evasão Escolar

A evasão está entre os temas que tem grande relevância no debate sobre educação no cenário das políticas públicas, sendo um dos temas mais discutidos na educação. Assim a evasão é um grande desafio para as escolas, pais e para o sistema educacional, sendo ela um dos principais problemas enfrentados pela educação no Brasil.

O conceito de evasão escolar definido por EYNG et al. (2013) é quando o aluno deixa de frequentar a aula, caracterizando o abandono da escola durante o ano letivo. Ele ainda define a evasão como um processo gradativo, que muitas vezes vai ocorrendo no andamento do curso e não é notado. Para isso ele usa o seguinte relato: ”No começo eram faltas esporádicas. Depois, passaram a ser semanais. A professora foi se acostumando ao silêncio quando chamava o nome do aluno. Até que um dia ele não apareceu mais.”

Segundo a legislação brasileira vigente, a responsabilidade da evasão na educação é tanto do Estado quanto da família, desta forma tendo esses dois agentes o dever da orientação sócioeducacional de crianças e adolescentes. A Lei de Diretrizes e Bases da Educação (LDB) é clara quanto a isso como podemos ver abaixo(QUEIROZ, 2001).

> Art. 2º. A educação, dever da família e do Estado, inspirada nos princípios de liberdade e nos ideais de solidariedade humana, tem por

> finalidade o pleno desenvolvimento do educando, seu preparo para o exercício da cidadania e sua qualificação para o trabalho.

O governo brasileiro, ciente deste problema, vem estudando a evasão há alguns anos e definiu que a evasão é definida em três eixos(MANHãES et al., 2011):

- Evasão de curso - quando o estudante desliga-se do curso de graduação em situações diversas: abandono (deixa de se matricular), desistência (oficial), 3 transferência (mudança de curso) ou exclusão por norma institucional;
- Evasão da instituição - quando o estudante desliga-se da instituição na qual está matriculado;
- Evasão do sistema - quanto o estudante abandona de forma definitiva ou temporária o ensino superior.

Hoje, mesmo com o déficit nacional de mão de obra especializada e as universidades oferecendo a cada ano um crescente número de vagas, o número de formados reduz a cada ano(BRASIL, 2013). Somando a isso há ainda o fator de que a ocupação de uma vaga em uma instituição pública de ensino seguida do abandono tornou-se um problema generalizado, independente da instituição, gerando perdas pessoais, sociais e financeiras(MANHãES et al., 2011).

Desta forma o problema da evasão deixou de ser um problema pessoal do estudante e passa para a ser um problema de estado, que precisa ser combatido para não comprometer ainda mais a mão de obra produtiva do país em um futuro próximo.

Alguns autores como BARROSO; FALCãO (2004) que avaliam o problema da evasão, definem os fatores que desencadeiam a evasão escolar em três principais agrupamentos:

- econômica - impossibilidade de permanecer no curso por questões socioeconômicas;
- vocacional – o aluno não se identifica com o curso;
- institucional – abandono por fracasso nas disciplinas iniciais, inadequação aos métodos de estudo, dificuldades de relacionamento com colegas ou com membros da instituição.

Apesar de existirem muitos trabalhos sobre a evasão, poucos são voltados para o ensino técnico. Da mesma forma, existe uma gama muito grande de relatórios com os números da evasão no ensino superior, mas praticamente nada sobre o ensino técnico. Valendo ressaltar que no Brasil o ensino técnico é enquadrado como ensino médio e fica clara uma separação entre técnico e médio nos dados divulgados pelo Governo Brasileiro.

Em 2012 um relatório desenvolvido pelo Programa das Nações Unidas para o Desenvolvimento (PNUD), classifica o Brasil como a terceira maior taxa de evasão no ensino médio, entre os 100 países com maior Índice de Desenvolvimento Humano (IDH), chegando a alcançar taxas de 24,3%(DE OLIVEIRA, 2017).

Assim, a evasão se tornou um problema de grandes proporções na educação brasileira e a criação de novas ferramentas que possam fornecer uma forma de auxílio aos educadores nessa tarefa se torna de suma importância. Desta forma, este trabalho através da mineração de dados educacionais tem como uma de suas metas a geração de modelos de predição para identificação de estudantes de cursos técnicos a distância que apresentem risco de evasão.

2.3 Mineração de Dados

Com o surgimento dos sistemas de informação o armazenamento de dados se tornou uma das prioridades das instituições. Para isso foram criadas bases de dados que cresceram de forma demasiadamente rápida, como podemos notar com maior evidencia nas últimas décadas, seja pela redução no preço de aquisição dos equipamentos ou pela maior utilização da comunidade em geral (CIOS; PEDRYCZ; SWINIARSKI, 1998).

Tabela 1: Quadro de evolução da Mineração de Dados - Adaptado de VASCONCELOS; CARVALHO (2004)

Etapa Evolucionária	Questão Comercial	Tecnologias Disponíveis	Fornecedores de Produtos	Características
Coleção de dados(1960s)	“Qual foi minha receita total nos últimos cinco anos?”	Computadores, fitas e discos	IBM, CDC	Retrospectiva, distribuição de dados estática
Acesso a dados(1980s)	“Quais foram as vendas unitárias de São Paulo em março?”	Bancos de dados relacionais(RDBMS), Structured Query Language (SQL),ODBC	Oracle, Sybase, Informix, IBM, Microsoft	Retrospectiva, distribuição de dados dinâmica a nível de registros
Data Warehousinge Suporte à Decisão(1990s)	“Quais foram as vendas unitárias de São Paulo em março? Avalie também Campinas.”	On-Line Analytical Processing(OLAP)	Pilot, Comshare, Arbor, Cognos, Microstrategy	Retrospectiva, distribuição dinâmica de dados múltiplos níveis
Mineração de Dados	“Qual a previsão para as vendas de Campinas no próximo mês? Porquê?”	Algoritmos avançados, computadores multiprocessados, banco de dados massivos.	Pilot, Lockheed, IBM, SGI, e outras (novas empresas)	Prospectiva, distribuição de informação ativa.

Devido a essa melhora na tecnologia da informação e o crescimento da Internet, as organizações são capazes de coletar e armazenar enormes quantidades de dados.

Para o armazenamento foram desenvolvidas diferentes estruturas de armazenamento para as novas demandas que foram surgindo. Estas podem ser desde simples base de dados contendo o estoque de um super-mercado até a indexação de grandes motores de busca como o Google.

Entre essas bases de dados podemos citar algumas que têm um enorme tamanho como descritas por (BRAMER, 2013):

- Os satélites de observação da NASA geram cerca de um terabyte de dados por dia;
- O projeto Genoma armazena milhares de bytes para cada uma das bilhões de bases genéticas;
- Instituições mantêm repositórios com milhares de transações dos seus clientes;

Percebeu-se gradualmente que os dados não são iguais a informação, que os dados devem ser analisados e extraídos. Assim surgiu uma pergunta crucial ”Com o volume de dados armazenados crescendo diariamente, o que fazer com os dados armazenados?”(CAMILO; SILVA, 2009).

Mesmo com profissionais treinados para analisar e interpretar os dados, os aumentos na quantidade de dados, tipo de dados, e dimensões de análise, têm dificultado estas ações. A computação tem ido além do armazenamento, transmissão e processamento. Os dados precisam ser convertidos em informação e conhecimento para apoiar a tomada de decisão(PASTA, 2011).

Fazer com que os dados armazenados, seja em um grande Data Center ou em pequenos servidores, se transformem de simples códigos sem sentido aparente em uma série de informações úteis é um dos principais desafios. Esse processo de descoberta pode fazer com que uma empresa simplesmente perca sua competitividade ou uma instituição de ensino deixe de formar inúmeros alunos em um ano.

Com a interessante tarefa de descoberta de conhecimento nas bases de dados, tendo em vista que as técnicas tradicionais de exploração de dados não são mais adequadas para tratar a grande maioria dos repositórios, foi proposta, no final da década de 80, a Mineração de Dados, do inglês Data Mining. Assim a questão levantada anteriormente pode ser respondida(CAMILO; SILVA, 2009).

Desta forma, a mineração de dados com suas tarefas e técnicas representa a fase principal Descoberta de Conhecimento em Bases de Dados, do inglês para Knowledge Discovery in Databases (KDD).

Alguns dos casos onde a mineração de dados pode auxiliar os processos de tomada de decisão de forma satisfatória são citados por BRAMER (2013), OLSON; DELEN (2008), WITTEN et al. (2016) e CAMILO; SILVA (2009):

- Retenção de clientes: identificação de perfis para determinados produtos, venda cruzada;
- Bancos: identificar padrões para auxiliar no gerenciamento de relacionamento com o cliente;
- Cartão de Crédito: identificar segmentos de mercado, identificar padrões de rotatividade;
- Cobrança: detecção de fraudes;
- Telemarketing: acesso facilitado aos dados do cliente;
- Eleitoral: identificação de um perfil para possíveis votantes;
- Medicina: indicação de diagnósticos mais precisos;
- Segurança: na detecção de atividades terroristas e criminais;
- Auxílio em pesquisas biométricas;
- RH: identificação de competências em currículos [9];
- Tomada de Decisão: filtrar as informações relevantes, fornecer indicadores de probabilidade.
- Comércio: Melhorar a disposição de seus produtos nas prateleiras, através do padrão de consumo de seus clientes;
- Marketing: direcionar o envio de mensagens promocionais, obtendo melhores retornos;

Tradicionalmente o modelo de transformação dos dados em informação segundo FAYYAD; PIATETSKY-SHAPIRO; SMYTH (1996), consiste no processamento manual de todas essas informações por especialistas que produzem relatórios para análise. Assim ele cita que com a sobre-carga de dados gerada pela era da informação esse processo de descoberta do conhecimento manual se tornou impraticável, tanto pelo tempo demandado quanto pela dificuldade da tarefa, e a KDD surge como uma tentativa de solucionar este problema.

A definição de KDD amplamente utilizada na área é dada por FAYYAD; PIATETSKY-SHAPIRO; SMYTH (1996):

> KDD é um processo, não trivial, de extração de informações implícitas, previamente desconhecidas e potencialmente úteis, a partir dos dados armazenados em um banco de dados".

Pode se dizer que o processo é não trivial por termos técnicas de busca ou inferência envolvidas, ou seja, não é apenas um processo de computação direta. Os padrões descobertos tem de ser válidos com algum grau de certeza, novos (para o sistema e de preferência também para o usuário), potencialmente úteis (trazer algum benefício) e compreensíveis (se não imediatamente então depois da interpretação).

A geração do conhecimento se dá através de uma sequência de etapas que executadas de forma correta podem resultar na geração de conhecimento útil. Estas etapas podem ser basicamente resumidas em 5: seleção dos dados a serem utilizados; preparação para a utilização através de um tratamento prévio (pré-processamento); transformação para um formato adequado; o processamento do conjunto de dados por algoritmos especialistas (mineração de dados) e a análise dos resultados obtidos(FAYYAD; PIATETSKY-SHAPIRO; SMYTH, 1996).

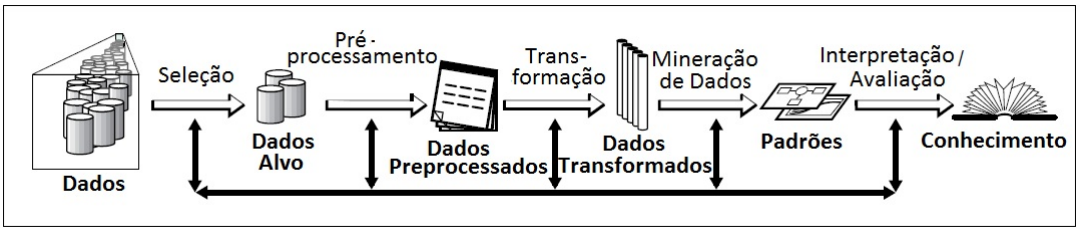

Figura 3: Descoberta do conhecimento - KDD. Adaptado de FAYYAD; PIATETSKY-SHAPIRO; SMYTH (1996)

2.3.1 Seleção dos dados

A etapa de seleção dos dados possui um impacto significativo sobre a qualidade do resultado final e é a primeira no processo de descobrimento de informação, uma vez que nesta fase é escolhido o conjunto de dados contendo todas as possíveis variáveis também conhecidas como características ou atributos e registros também chamados de casos ou observações que farão parte da análise. Normalmente essa escolha dos dados fica a critério de um especialista do domínio, ou seja, alguém que realmente entende do assunto em questão(PRASS, 2004).

Segundo PRASS (2004), a seleção é um processo complexo por tratar de dados que podem vir de diversas fontes, que podem ser desde planilhas até sistemas legados. Assim esses dados podem possuir diversos formatos, sendo comum que ocorra a necessidade de um software específico para o tratamento destes dados, tamanha a peculiaridade envolvida na aplicação.

2.3.2 Pré-processamento e limpeza dos dados

O Pré-processamento e limpeza dos dados é a etapa onde são efetuadas tarefas como a verificação e eliminação das redundâncias e inconsistências das bases, recuperação dos dados incompletos ou outliers que em estatística são valores atípicos, é uma observação que apresenta um grande afastamento das demais da série.

DUNKEL et al. (1997) cita como um dos problemas do pré-processamento é a identificação dos dados inapropriados dentro do conjunto disponível. Assim, tomar a atitude de classificar um dado com ruim é de grande complexidade e o auxílio de um especialista do domínio é fundamental, pois na maioria dos casos apenas alguém que realmente entende do assunto é capaz de dizer se um dado é um outlier ou um erro de digitação.

Segundo HAN; PEI; KAMBER (2011) as técnicas de pré-processamento podem ser divididas em:

- Limpeza dos dados (data cleaning);
- Integração de dados (data integration);
- Transformação de dados (data transformation);
- Redução de dados (data reduction);

2.3.2.1 Limpeza dos dados

A etapa de limpeza dos dados, ou data cleaning, é uma etapa de investigação visando detectar dados que estejam duplicados, incorretos e/ou faltantes(missing values). Para corrigir isso HAN; PEI; KAMBER (2011) sugerem como soluções seguir as seguintes etapas:

- Ignorar os registros;
- Completar manualmente os valores faltantes;
- Substituir por uma constante global;
- Uso da média para preencher os valores faltantes;
- Uso do valor mais provável, que pode ser predito com auxílio de uma regressão, árvores de decisão, entre outras.

Os outliers são outro problema que é tratado nesta etapa do pré-processamento. Estes valores podem ser detectados através da análise de agrupamento, onde os valores similares formam grupos, destacando-se os outliers SCHMITT (2005).

As mesmas soluções anteriormente citadas para o tratamento dos dados faltantes podem ser utilizadas para os outiliers com o cuidado para a exclusão, esta só deve ser efetuada quando o dado representar algum erro, seja de medida ou algo similar. Assim não comprometendo as etapas seguintes (SCHMITT, 2005).

2.3.2.2 Integração dos dados

Como dito anteriormente, o processo de mineração é complexo e diversas vezes há a necessidade de tratar dados vindos de diversas fontes diferentes. Para isto é necessária a etapa de integração dos dados, seja manualmente ou automatizadamente através de um sistema. Esta etapa pode acabar por ser extremamente demorada dependendo dos dados, alguns autores como FERNANDEZ (2010) chegam a estimar que dependendo da base essa etapa pode consumir até mesmo 70% do tempo do projeto.

2.3.2.3 Transformação

Antes da fase onde acontece a efetiva mineração dos dados na KDD, os dados necessitam ser armazenados e formatados adequadamente para que os algoritmos possam ser aplicados.

Em alguns casos, dependendo do algoritmo utilizado é necessária a transformação do tipo de dados, pois existem algoritmos que trabalham apenas com valores numéricos e outros apenas com valores categóricos. Assim, pode ser necessária a conversão dos dados de um tipo para outro.

Segundo PRASS (2004), nesta fase, se necessário, é possível obter dados faltantes através da transformação ou combinação de outros, são os chamados “dados derivados”. Um exemplo de um dado que pode ser calculado a partir de outro é a idade de um indivíduo, que pode ser encontrada a partir de sua data de nascimento. Outro exemplo é o valor total de um financiamento que pode ser calculado a partir da multiplicação do número de parcelas pelo valor da parcela.

Quando a mineração de dados envolve a utilização de algoritmos de redes neurais, autores como HAN; PEI; KAMBER (2011) aconselham a utilização da transformação min-max, pois esta pode melhorar a eficiência dos algoritmos. A transformação min-max é definida por: seja uma determinada variável A, com valores $A_1, A_2, A_3, A_4, ..., A_n$. Sendo o valor mínimo representado por min_A e o valor máximo representado por max_A e deseja-se transformar os valores em $A_1, A_2, A_3, A_4, ..., A_n$ para valores em um intervalo [a,b], então os valores $A_1, A_2, A_3, A_4, ..., A_n$ são dados pela equação:

$$A_i' = \frac{A_i - \min_A}{\max_A - \min_A}.(b-a) + a \qquad \text{i} = (1,2,3, ..., \text{n})$$

Figura 4: Descoberta do conhecimento - KDD. Adaptado de HAN; PEI; KAMBER (2011)

2.3.2.4 Redução dos dados

Apesar de ser de difícil quantificação, muitas as vezes as bases de dados possuem dados que não apresentam grande relevância para a mineração. A redução da dimensionalidade destes dados pode diminuir o custo computacional da tarefa de mineração. Mesmo com o tempo de processamento diretamente ligado a variáveis como o tamanho da base, a tarefa a ser realizada e o algoritmo a ser utilizado, tende-se em uma base otimizada a conseguir um melhor custo benefício na variável tempo de processamento/resultado final (PRASS, 2004).

Diversos autores divergem sobre os processos utilizados para a redução da base, mas podemos dizer que uma das teorias mais aceitas é a de HAN; PEI; KAMBER (2011) que utiliza a técnica de componentes principais (JOHNSON; WICHERN et al., 2002). Está técnica consiste na identificação dos componentes de maior impacto na base através de uso de estatística e seu isolamento, assim excluindo os componentes de pouco ou nenhum impacto.

2.3.3 Mineração de dados educacionais

Diversos trabalhos procuram modelar comportamentos de alunos de EAD com a intenção de realizar previsões sobre estes, e utilizando uma variedade de técnicas e bases de dados, em geral com resultados satisfatórios. Estas pesquisas geralmente se diferenciam nas técnicas utilizadas e no seu objetivo de previsão. Assim, alguns destes trabalhos são voltados para a predição das notas dos alunos nas avaliações de uma determinada disciplina ou até mesmo no curso todo, enquanto outros trabalhos visam à predição da situação de evasão do aluno demonstrando se o mesmo está ou não em situação de risco de evasão.

Segundo ROMERO; VENTURA (2013), a maioria das técnicas tradicionais de mineração de dados incluindo, mas não se limitando a classificação, agrupamento, e técnicas de análise de associação já foram aplicadas com êxito no domínio da educação.

Grande parte das técnicas seguem a taxonomia proposta por BAKER; ISOTANI; CARVALHO (2011) e aceita por grande parte dos pesquisadores da área como Romero, Ventura e Jayaprakash entre outros, da seguinte forma:

- Predição

  - Classificação
  - Regressão
  - Estimação de densidade
- Agrupamento
- Mineração de relações
  - Mineração de regras de associação
  - Mineração de correlações
  - Mineração de padrões sequenciais
  - Mineração de causas
- Destilação dos dados
- Descobertas com modelos

Segundo BAKER; ISOTANI; CARVALHO (2011), na predição através da análise e fusão das informações contidas nos dados são criados modelos que infiram características e informações sobre esses dados, estas são chamadas de variáveis preditivas(predicted variables). Para isto é necessário que se tenha uma quantidade relativa de dados e que haja uma codificação manual da identificação de uma ou mais variáveis.

Entre os 3 tipos de predição citados por BAKER; ISOTANI; CARVALHO (2011), os mais utilizados são classificação e regressão, enquanto que estimação de densidade dificilmente é utilizado devido a dependência estatística dos dados. Os outros dois tipos de predição variam conforme a variável a ser prevista. Quando esta variável é numérica são utilizados geralmente algoritmos de regressão linear e redes neurais. Enquanto que com variáveis binárias ou categóricas são utilizados algoritmos de classificação como árvores de decisão.

O agrupamento tem por objetivo a identificação de grupos de dados que apresentem semelhanças entre si em alguma variável ou aspecto, assim auxiliando na tarefa de descoberta de novas informações. Geralmente os dados são agrupados utilizando alguma medida de distância que decide a semelhança entre os dados, uma vez feita esta etapa, os dados podem voltar a ser analisados pois podem ser gerados reagrupamentos a partir dos anteriores (ROMERO; VENTURA, 2013).

Na EDM, o agrupamento pode ser utilizado para agrupar alunos, interações ou até mesmo materiais. Assim esta etapa passa tanto pela descoberta da relação das

variáveis como pela tentativa de aprendizado de quais variáveis e valores tem um maior impacto sobre alguma outra variável que geralmente é a que se busca prever. No conceito taxonômico de BAKER; YACEF (2009) existem 4 tipos de mineração diferentes usadas para identificar relações, sendo eles:

- Regras de associação;
- Correlações;
- Sequências;
- Causas;

A mineração por regras de associação tem como premissa básica a busca por variáveis que tenham seu valor associado ao valor de outras variáveis. Para isso utiliza-se se-então (if-then), assim quando uma variável assume um determina valor podemos inferir o valor da segunda variável.

Assim as regras de associação representam padrões existentes em transações armazenadas. Utilizando o exemplo de VASCONCELOS; CARVALHO (2004), a partir de uma base de dados, na qual registram-se os itens adquiridos por clientes, uma estratégia de mineração, com o uso de regras de associação, poderia gerar a seguinte regra: cinto, bolsa ! sapato, a qual indica que o cliente que compra cinto e bolsa, com um determinado grau de confiança, compra também sapato. Este grau de certeza de uma regra é definido por dois índices: o fator de suporte e o fator de confiança.

Na mineração de correlações, o objetivo é encontrar correlações lineares entre as variáveis. BAKER; ISOTANI; CARVALHO (2011) cita como exemplo um conjunto de dados educacionais que após onde busca-se identificar a nota de um aluno a partir do tempo gasto na aula por esse em tarefas não relacionadas às dadas pelo professor.

Em mineração de sequências, busca-se encontrar uma associação temporal entre os eventos e o impacto deles nas variáveis. Como exemplo BAKER; ISOTANI; CARVALHO (2011) cita a trajetória dos atos e ações de um determinado aluno e o resultado de sua aprendizagem.

Na mineração de causas, a principal idéia é que se busquem identificar eventos que ocasionem outros eventos. Isso se dá através da análise de padrões de covariância. Como exemplo podemos citar o comportamento de um aluno em aula, muitas vezes seu comportamento inadequado está ligado diretamente à sua dificuldade de aprendizagem. Assim um resultado ruim em determinadas tarefas pode representar na verdade um problema de aprendizagem (BAKER; ISOTANI; CARVALHO, 2011).

A destilação dos dados busca facilitar a compreensão de dados complexos e suas características. Essa etapa possibilita que os dados sejam analisados e assim as pessoas os compreendam e identifiquem padrões sobre eles, auxiliando na tomada

de decisões. Um exemplo seria a curva de aprendizagem de um aluno, que são representações matemáticas do desempenho de um estudante quando submetido a tarefas de ensino (ARGOTE, 1999). À medida que as repetições são efetuadas, o estudante demanda menos tempo para o aprendizado, seja pela familiaridade adquirida com os meios, seja pela adaptação às ferramentas utilizadas ou pela descoberta de "atalhos" para realização da tarefa (DAR-EL, 2013).

Segundo autores como Bevitt (BEVITT et al., 2015), o melhor período para se fornecer ao aluno um feedback de seu possível desempenho seria com duas semanas a partir do início do semestre. Assim restando ao aluno mais tempo para uma mudança em seu comportamento e uma possível melhora em seu desempenho acadêmico.

2.3.4 Aprendizagem de Máquina

Nesta seção será abordado o conceito de aprendizagem de máquina e os algoritmos utilizados nesta dissertação.

Desde de a invenção das máquinas, a humanidade vem buscando formas para que elas imitem o comportamento humano em determinadas tarefas, seja por bonecos falantes ou até mesmo robôs que pratiquem esportes. Entretanto, com o surgimento dos computadores e o aumento da capacidade tanto de armazenamento quanto de processamento surgiu a área chamada de aprendizagem de máquina.

A aprendizagem de máquina é uma sub-área da Ciência da Computação, tendo como objetivo de pesquisa o desenvolvimento de técnicas e sistemas computacionais que adquiram conhecimento sobre determinados dados de forma autônoma. Esses sistemas são algoritmos desenvolvidos para a predição de situações que possam ocorrer em determinados problemas a partir dos fatos ocorridos anteriormente em situações parecidas. Cada um destes algoritmos possui determinadas características que podem possibilitar sua classificação quanto à linguagem de descrição, modo, paradigma e forma de aprendizagem(MONARD; BARANAUSKAS, 2003).

A aprendizagem geralmente é dívida nas seguintes classificações:

- Aprendizagem
  - Aprendizagem Supervisionada
    - Classificação
    - Regressão
  - Aprendizagem Não supervisionada
    - Algoritmos de agrupamento

Em aprendizagem supervisionada existe um "professor" que verifica as saídas dos algoritmos com o padrão de entradas atuais, assim fazendo alterações no algoritmo

para que as respostas fiquem o mais corretas possível. Ou seja, observa-se alguns pares de exemplos de entrada e saída, de forma a aprender uma função que mapeia a entrada para a saída. Este tipo de aprendizagem é utilizado em algoritmos como redes neúrais por exemplo, onde cada ajuste de peso das sinapses são voltados para ajustar a saída.

Em aprendizagem não supervisionada, não existe o ”professor”, pois nem sempre os conjuntos de dados de entrada tem a chamada saída. Assim os algoritmos vão lendo a entrada e tentando descobrir as saídas identificando os padrões sozinhos. Um dos exemplo disto são os algoritmos de clusterização que buscam descobrir similaridades e diferenças entre os padrões existentes, assim como derivar conclusões úteis a respeito deles (HASTIE; TIBSHIRANI; FRIEDMAN, 2009).

A seguir apresentaremos os algoritmos que foram utilizados na pesquisa desenvolvida nesse trabalho.

2.3.4.1 Algoritmos Bayesianos

Baseado no teorema de Bayes, os algoritmos bayesianos em suas diversas formas são modelos probabilísticos que buscam calcular a probabilidade de uma determinada variável pertencer a uma determinada classe. Este método é conhecido como predição estatística e ela faz parte da aprendizagem supervisionada(MARQUES; DUTRA, 2002).

Os algoritmos baseanos são baseados no teorema de Bayes, onde se indica que é possível calcular a probabilidade de um determinado evento ocorrer dada a probabilidade de um outro evento que já tenha ocorrido. Assim temos: Probabilidade(A dado B) = Probabilidade(A e B)/Probabilidade(A)

$$P(A \mid B) = \frac{P(A) \times P(B \mid A)}{P(A) \times P(B \mid A) + P(\overline{A}) \times P(B \mid \overline{A})}$$

Figura 5: Formulá teorema de Bayes

Alguns experimentos demonstram que os algoritmos Bayesianos conhecidos como Naive Bayes podem obter resultados similares às Árvores de Decisão e às Redes Neurais, entretanto tem um custo computacional menor(ZHANG, 2004). Isso é uma de suas características e geralmente associada a sua simplicidade mesmo assim guardando um alto pode preditivo, assim sendo um dos algoritmos mais utilizados (CAMILO; SILVA, 2009).

Os algoritmos bayesianos tem como principal vantagem o fato de que agentes podem tomar decisões racionais mesmo quando não existe informação suficiente para se provar que uma ação funcionará(CHARNIAK, 1991).

2.3.4.2 Árvores de decisão - J48 e Random Forest

Árvores de decisão são algoritmos de classificação supervisionada, pois neles é necessário saber quais são as classes de cada registro do conjunto de treinamento. Neste trabalho utilizamos dois algoritmos diferentes dessa técnica, J48 e Random Forest(Florestas aleatórias).

Esse tipo de algoritmo gera uma estrutura de árvore que que classifica as amostras desconhecidas. Para isso, utiliza os dados dos conjuntos de treinamento, criando uma árvore e a partir desta, classificando as amostras desconhecidas sem necessariamente testar todos os valores dos seus atributos(MICHALSKI; CARBONELL; MITCHELL, 2013).

Para este tipo de algoritmo é necessário que ele sempre defina quais são os elementos desta árvore. Assim podemos fazer a analogia a uma árvore normal e ver os seus nós conectados às ramificações.

Desta forma, existem basicamente três tipos de nós: o nó raiz, que inicia a árvore, os nós comuns que dividem um determinado atributo e geram ramificações e os nós folha que contém as informações de classificação do algoritmo. Já as ramificações possuem todos os valores possíveis do atributo indicado no nó para facilitar a compreensão e interpretação(QUINLAN, 1986).

Com a árvore montada cada nó tem a tarefa de testar um atributo dos novos nós. Desta forma segundo PICHILIANI (2008), podemos dizer que o atributo que melhor classificar os dados deve ser escolhido como um nó da árvore. Para facilitar a compreensão, é comum colocar os valores das probabilidades de cada classe dentro do nó.

A classificação de cada novo elemento da árvore é feita percorrendo os ramos e nós da árvore de acordo com os valores dos atributos da amostra desconhecida. Este algoritmo permite uma análise mais detalhada levando em consideração cada valor de todos os atributos (PICHILIANI, 2008).

Na Figura 6 temos um conjunto de dados simples demonstrando se o dia é apto ou não a prática de tênis(RUSSELL; NORVIG, 2010):

Exemplos de Treino

Dia	Aspecto	Temp.	Humidade	Vento	Jogar Ténis
D1	Sol	Quente	Elevada	Fraco	Não
D2	Sol	Quente	Elevada	Forte	Não
D3	Nuvens	Quente	Elevada	Fraco	Sim
D4	Chuva	Ameno	Elevada	Fraco	Sim
D5	Chuva	Fresco	Normal	Fraco	Sim
D6	Chuva	Fresco	Normal	Forte	Não
D7	Nuvens	Fresco	Normal	Fraco	Sim
D8	Sol	Ameno	Elevada	Fraco	Não
D9	Sol	Fresco	Normal	Fraco	Sim
D10	Chuva	Ameno	Normal	Forte	Sim
D11	Sol	Ameno	Normal	Forte	Sim
D12	Nuvens	Ameno	Elevada	Forte	Sim
D13	Nuvens	Quente	Normal	Fraco	Sim
D14	Chuva	Ameno	Elevada	Forte	Não

Figura 6: Conjunto de dados

Na Figura 7 abaixo podemos ver a árvore que seria gerada a partir do dados do conjunto de treinamento anterior.

Figura 7: Exemplo de árvore

O algoritmo J48, baseado em árvores de decisões nada mais é que uma versão do algoritmo C4.5 desenvolvida em Java e utilizada dentro da biblioteca do Weka. O C4.5 é um algoritmo utilizado para criar uma árvore de decisão e foi desenvolvido por QUINLAN (1993) C4.5 é uma extensão do algoritmo anterior de Quinlan o ID3. As árvores de decisão geradas pelo algoritmo C4.5 podem ser utilizadas para classificação e são portanto conhecidas como classificadores estatísticos.

Utilizando a abordagem gulosa para induzir árvores de decisão para posterior classificação. 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 (MARTINS; MARQUES; COSTA, 2009).

Ainda segundo MARTINS; MARQUES; COSTA (2009), uma vez que é 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. Este algoritmo é escolhido para comparar a percentagem de acerto com outros algoritmos(QUINLAN, 1993).

O algoritmo Random Forest é uma técnica de classificação desenvolvida por BREIMAN (2001). Nela diferentemente das árvores de decisão padrão, os dados são divídivos em diversos subconjuntos menores onde cada um deles são criadas amostragens chamadas de bootstrap. Esse técnica chamada de bootstrap é utilizada para garantir que geralmente 1/3 dos exemplos sejam usados para testar as árvores após sua construção (HO, 1995).

Com os subconjuntos separados, então é desenvolvida uma árvore de decisão para cada um deles. Então a floresta aleatória é a coleção dessas árvores dos subconjuntos(HO, 1995).

2.3.4.3 Simple Logistic

O Simple Logistic é um classificador que utiliza modelos de regressão logística linear simples. Uma regressão logística simples é uma regressão logística de apenas um parâmetro (AGRESTI; KATERI, 2011).

Regressão logística vem do fato de que a regressão linear também pode ser usada para executar a classificação do problema, apenas transformando o alvo categórico com valores contínuos. A idéia de regressão logística é fazer com que a regressão linear produza probabilidades. Podemos dizer que é melhor predizer probabilidades de classe em vez de prever as classes em si. Assim a Regressão logística estima probabilidades das classe diretamente utilizando a transformada lógica (AGRESTI; KATERI, 2011).

Para a regressão linear nós temos uma soma linear. Na regressão logística, esta soma linear é incorporada na fórmula conhecida como transformada lógica. Essa transformada é um modelo de curva com formato de S que tem valores entre 0 e 1 (AGRESTI; KATERI, 2011).

$$Pr(Y = 1 \mid X) = P(X) = \frac{E^{B_0 + B_1 . X}}{1 + e^{B_0 + B_1 . X}}$$

Figura 8: Formúla da transformada lógica

2.3.4.4 Redes Neurais

As redes neurais artificiais surgiram em 1943 como uma tentativa de criar um modelo matemático que imitasse o comportamento de um neurônio biológico (MCCUL-

LOCH; PITTS, 1943). Elas podem ser definidas como técnicas computacionais que apresentam um modelo matemático inspirado na estrutura neural de organismos inteligentes e que adquirem conhecimento através da experiência (CARVALHO, 2009).

As redes neurais artificiais basicamente são formadas por um conjunto de terminais de entrada, também conhecidos como camada de entrada, que repassam a informação para as camadas intermediárias onde ocorre o processamento e uma camada de saída, que é onde saem as informações processadas.

Essas redes podem ser compostas por várias camadas de processamento de simples funcionamento. Cada camada é conectada com a próxima e esta associada a um peso, assim cada neurônio processa somente os dados que recebe em sua entrada e repassa o resultado para a camada seguinte.

Assim as Redes neurais artificias geralmente são apresentadas como sistemas de neurônios interconectados que podem computar valores de entradas (CARVALHO, 2009).

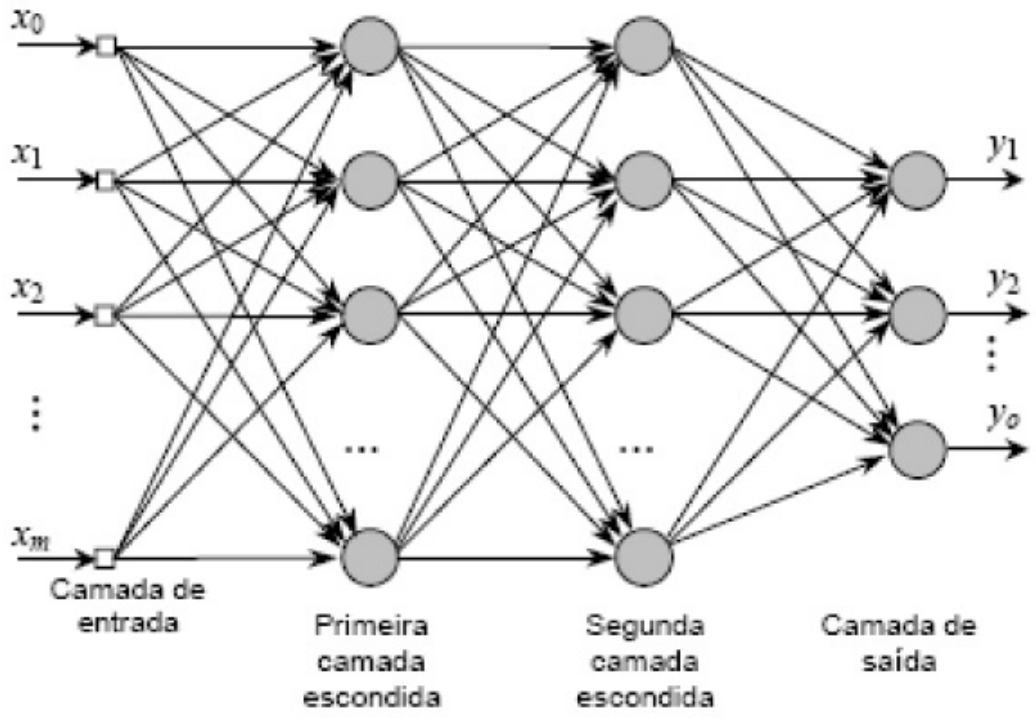

Figura 9: Modelo de uma rede neural - Adaptado de SEGATTO; COURY (2008)

O tamanho das redes neurais pode variar de acordo com a tarefa a que ela é utilizada, assim variando de uma rede de um neurônio até centenas ou milhares (CARVALHO, 2009).

Para este trabalho foram utilizadas redes neurais de Perceptron Multi-Camadas (Multi Layer Perceptron - MLP). Este modelo de rede foi criado buscando sanar alguns problemas que existiam nas redes de uma única camada por RUMELHART; HINTON; WILLIAMS (1985).

Neste novo modelo de rede foi feita a inclusão de camadas intermediárias de neurônios e de um algoritmo de aprendizagem por retro-propagação (back-propagation). Nesse modelo de rede todos os neurônios de uma camada estão ligados a todos os neurônios das camadas anterior e posterior. Assim foi possível um treinamento eficiente, pois cada camada tem uma função específica.

2.3.5 Ferramentas para mineração de dados

Hoje a mineração de dados é mais importante do que nunca para as instituições, sejam elas grandes ou pequenas, se querem alavancar seus produtos é necessário a mineração de dados para lhes fornecer um subsídio para as tomadas de decisões buscando uma vantagem competitiva.

Com a adoção de técnicas bem estabelecidas, ferramentas e recorrendo à ajuda de especialistas em mineração de dados, as ações baseadas em evidências de dados e análises avançadas têm melhores chances de aumentar a otimização de seus produtos e facilitar o crescimento de seu negócio, seja ele a venda de produtos ou um instituição de ensino que quer minimizar os efeitos da evasão em seus cursos (GUPTA, 2014).

Atualmente as ferramentas de mineração de dados disponíveis na Web são abundantes. Sendo que a maioria dos fornecedores oferece uma demonstração, freeware ou ambos para ajudá-lo a determinar qual ferramenta para mineração é melhor para os seus dados. Podemos destacar algumas como RapidMiner, RProgramming, Orange, Knime, NLTK, Oracle Data Mining, SAS, Cognos e o Weka.

Neste trabalho, optou-se pela utilização do WEKA (Waikato Environment for Knowledge Analysis), na versão 3.9.0, que é desenvolvida por universitários da Universidade de Waikato, na Nova Zelândia. A escolha deste software se motivou por alguns fatores como facilidade de utilização, obtenção direta da página do desenvolvedor sem custos, número considerável de algoritmos disponíveis para utilização junto com a possibilidade de alteração dos parâmetros de execução dos mesmos e a possibilidade de fácil comparação entre os algoritmos testados.

O Weka é uma ferramenta baseada em Java e de código aberto, sob a GNU General Public License, e que possui dois tipos de utilização primários, um pela sua interface gráfica e outra em pacote de classes em Java. Ele teve sua primeira versão desenvolvida em 1999 principalmente para analisar dados agrícolas(HALL et al., 2009).

Alguns dos principais atributos do Weka são o pré-processamento de dados, a visualização e análise preditiva, além de técnicas de modelagem, clustering, agrupamento, associação, regressão e classificação.

Na utilização da interface gráfica é possível a interação com os dados, entretanto uma das restrições dessa utilização é o limite de tamanho das bases, o que não permiti o uso de quantidades elevadas de dados, e não tem suporte a utilização de threads

por exemplo. Já na utilização na classe em Java, esta pode ser importada em um projeto de aplicação e customizada para o uso conforme a vontade de desenvolvedor, o que é uma grande vantagem em relação por exemplo ao RapidMiner.

Neste projeto foram utilizadas os dois tipos de implementação do Weka. Para os testes iniciais foi utilizado o aplicativo do próprio Weka e sua interface visual. Entretanto conforme os testes foram avançando e a base de dados foi aumentando substancialmente, optamos pelo desenvolvimento de um software próprio na tentantiva de automatizar algumas etapas do projeto, possibilidade de uma maior customização dos testes, suporte ao processamento paralelo e assim consequentemente agilidade na geração dos modelos, suporte a alta quantidade de dados do projeto e a integração com um banco de dados externo (BOUCKAERT et al., 2010).

Segundo HALL et al. (2009), o Weka em seu aplicativo pode ser utilizado de diversas formas, em função do mesmo possuir quatro diferentes interfaces implementadas acessíveis através de sua tela inicial, que são elas:

1. Explorer: Nesta interface são aplicadas as tarefas e técnicas de MD sobre a base de dados;

2. Experimenter: Esta interface é útil para a aplicação de um ou mais técnicas de classificação sobre uma grande base de dados e em seguida fazer comparações estatísticas sobre elas;

3. Knowledge-flow: Esta é considerada a interface que mais apresenta o funcionamento da ferramenta, uma vez que tem sua representação de forma gráfica;

4. Simple client: Esta interface oferece um local para inserção de comandos. Mesmo possuindo uma aparência considerada simples, é nela que realiza qualquer operação suportada pelo WEKA.

Os formatos de arquivos que o Weka trabalha são somente 2, CSV(Comma-separated values) ou ARFF(Attribute Relation File Format). Segundo REPICI (2010), Comma-separated values também conhecido como CSV, são arquivos texto de formato regulamentado pelo RFC 4180 que faz uma ordenação de bytes ou um formato de terminador de linha. Enquanto que ARFF é um formato de arquivo próprio do Weka onde estão contidas informações como: a definição do domínio dos atributos e as instâncias, que representam os dados que serão trabalhados(HALL et al., 2009).

Neste projeto foram utilizados somente arquivos ARFF, isso se deu por uma opção de desenvolvimento do projeto. Uma das propriedades deste tipo de arquivo que pesou na seleção de trabalharmos somente com Attribute Relation File Format, foi o fato de o Weka ter menos problemas na interpretação desse arquivo.

Os arquivos ARFF são compostos por uma estrutura de três partes conforme definida por PASTA (2011), sendo eles:

1. Cabeçalho;
2. Declaração dos atributos;
3. Seção de dados.

A declaração dos atributos é feita no cabeçalho, sendo um atributo por linha. Os tipos de atributos a que o Weka tem suporte sãoPASTA (2011):

1. Numérico: trabalhando com números reais e decimais;
2. Inteiros: números sem decimais;
3. Datas;
4. String: com ressalva para a substituição de espaços em branco por underline;
5. Enumerados: nos quais os tipos são previamente definidos pelo usuário, Ex.: Sexo Masc, Fem.

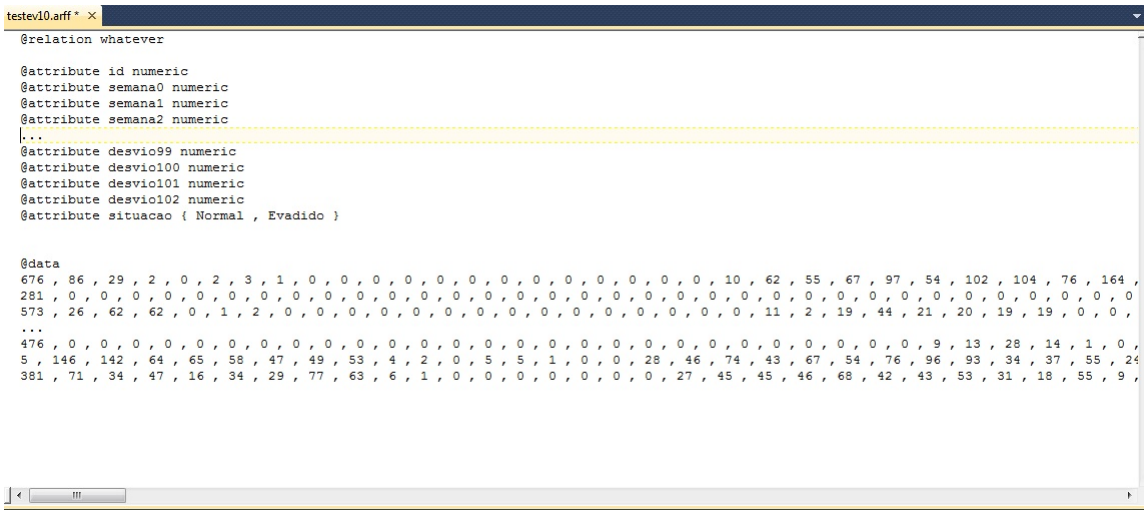

Figura 10: Partes de um dos arquivos ARFF utilizado no projeto

No código listado no Anexo B, podemos ver uma parte da utilização do Weka via seu pacote de classes para Java. Este trecho de código foi retirado de uma das aplicações iniciais utilizando o algoritmo Bayesnet, uma base de dados em ARFF, utiliza classificadores e fez parte dos experimentos onde o treino e o teste das bases de dados eram efetuados juntos utilizando validação cruzada.

3 TRABALHOS RELACIONADOS

Agora apresentaremos algumas das pesquisas de relevância na área de mineração de dados educacionais e os resultados obtidos pelas mesmas.

3.1 Descrição dos trabalhos relacionados

No experimento de JAYAPRAKASH et al. (2014), busca-se criar um sistema de alerta de risco quanto ao desempenho do aluno, a fim de diminuir as taxas de evasão e retenção escolares, fornecendo ao aluno um feedback atualizado de seu possível rendimento escolar. Para isto ele utiliza dados demográficos como sexo e idade, interações dos alunos com o ambiente virtual de aprendizagem, desempenho acadêmico anterior, tempo na universidade, tempo online no ambiente virtual, dados do teste de aptidão escolar (SAT Verbal e Matemático), entre outros. Assim analisando dados de 9938 alunos e aplicando árvores de decisão com o algoritmo J48, redes Bayesianas com o Naive Bayes, Máquinas de suporte Vetorial com o SVM/SMO e regressão logística. Na tarefa de predição todos algoritmos apresentaram resultados muito próximos, tendo o classificador de regressão logística apresentado resultados ligeiramente maiores que os outros 3 com 94,20% de acurácia geral e 66,70% de precisão na predição de alunos em risco de evasão.

LYKOURENTZOU et al. (2009), propõe um sistema de predição de alunos em situação de risco de evasão que combine os resultados da aplicação de 3 algoritmos diferentes, sendo eles, Redes Neurais, Máquinas de suporte vetorial e sequência mínima de otimização (SVM/SMO) e conjunto probabilístico simplificado Fuzzy ART-MAP (PESFAM). Em sua pesquisa o autor utiliza dados demográficos invariantes no decorrer do curso como sexo e residência, além de dados acadêmicos como performance e nível escolar, e dados variantes como número de interações com o ambiente virtual, notas e até mesmo a data da entrega dos trabalhos. Com a aplicação dos algoritmos são criados 3 esquemas diferentes buscando a predição da evasão, onde no primeiro um estudante é considerado evadido se pelo menos uma técnica classificou este estudante como tal, no segundo o estudante é considerado evadido se pelo

menos duas técnicas indicam essa situação e no terceiro e último necessita que as 3 técnicas classifiquem o aluno como evadido para que este seja assim classificado. Os resultados obtidos variam de 73% a 94%, sendo que os mais satisfatórios foram obtidos pelo esquema 1 que chegou a atingir 94%.

YADAV; PAL (2012) em seu experimento utilizam árvores de decisão com os algoritmos C4.5, ID3 e CART para a predição final da situação de alunos quanto a evasão. Para tal, ele utiliza os dados do curso de engenharia da Veer Bahadur Singh Purvanchal University de Jaunpur, Índia (VBS). Apesar de não disponibilizar um quantitativo dos dados, o autor cita que é feita a análise e processamento dos dados disponíveis da década de 90 até 2010 e utiliza variáveis demográficas como sexo, renda familiar, escolaridade e ocupação dos pais, e também os resultados escolares dos alunos no primeiro ano do curso. O autor considera em seus resultados não o valor da acurácia geral e sim o valor da classificação dos alunos em risco de evasão, assim nos testes realizados tanto o algoritmo ID3 quanto o CART obtém valores de 62,22% enquanto o C4.5 67,77%.

A pesquisa de MANHãES et al. (2011), tem como objetivo a aplicação de técnica de mineração de dados para identificação precoce de alunos em risco de evasão nos cursos de graduação em Engenharia presencial da Universidade Federal do Rio de Janeiro. São utilizados dados sobre o desempenho dos alunos em duas disciplinas do primeiro semestre do curso, e aplicados 10 diferentes algoritmos para geração dos modelos dentro da ferramenta WEKA. Sendo eles, OneR e JRip (baseados em aprendizagem de regras), DecisionTable (tabela de decisão), SimpleCart, J48 e RandomForest(árvores de decisão) e SimpleLogistic (regressão logística). Tendo a acurácia média variando entre 75% e 80%.

Outro trabalho relacionado que apresenta resultados interessantes é o de DETONI; ARAUJO; CECHINEL (2014), que busca utilizar unicamente contagem de interações para predição de reprovação de alunos em disciplinas de EAD. Para isto são utilizados dois cursos oferecidos pela Universidade Federal de Pelotas(UFPel), Licenciatura em Educação do Campo (CLEC) e Licenciatura em Pedagogia(CLPD). Em sua pesquisa o autor opta por extrair as interações dos alunos, tutores e professores e agrupá-las de forma semanal. A partir das interações ainda são calculadas a média, mediana, média da diferença (média da diferença entre a semana i e a semana i+1 .), razão com professores (razão entre o total de interações do aluno e dos professores.), razão com tutores (razão entre o total de interações do aluno e dos tutores.) e fator de empenho (razão entre as interações da semana do aluno e a média de interações da turma naquela semana.). Após isto o autor aplica os algoritmos Redes Bayesianas, Redes Neurais, J48 e Random Forest, obtendo resultados de até 67% de acurácia na predição do desempenho do aluno.

Por sua vez CAMBRUZZI; RIGO; BARBOSA (2015), propõe um sistema de apren-

dizagem educacional que combine uma série de ferramentas complementares para visualização dos dados, previsão de abandono escolar e apoio a ações pedagógicas e análise textual, entre outros, para diminuição da evasão em cursos a distância. Para isto as ferramentas complementares são adicionadas em uma camada chamada de MultiTrail, que é responsável pela leitura dos dados, processamento e posterior envio dos resultados para o cliente. Na etapa de predição são utilizadas Redes Neurais para geração do modelo de predição que são aplicados a dados dos alunos da Universidade do Vale do Rio dos Sinos (UNISINOS), assim obtendo segundo o autor uma precisão de 87% na predição de alunos em risco de evasão e conseguindo uma diminuição de até 11% nesta evasão após a implantação do sistema.

YUKSELTURK (2014) busca prever a evasão de alunos de cursos a distância utilizando dados coletados a partir de um questionário sobre dados demográficos como, sexo, idade, nível de escolaridade, ocupação, disponibilidade, entre outros, que é aplicado aos alunos dos cursos do Programa de Certificação em Tecnologias de Informação da Kirikkale University (Turquia). Para classificação dos alunos foram utilizados 4 algoritmos, sendo eles: K-Nearest Neighbour (k-NN), Árvore de Decisão (DT), Naive Bayes (NB) e Redes Neurais (RN). Os resultados mais significativos foram obtidos com o K-NN com a taxa de 87% de acerto para alunos em risco de evasão.

DEKKER; PECHENIZKIY (2009) ressalta a importância da predição precoce da evasão de alunos em instituições de ensino superior em seu estudo de caso, buscando assim a previsão da possibilidade de abandono do curso de alunos que acabaram o primeiro semestre do mesmo. Para tanto, ele utiliza como base dados dos alunos do curso de Engenharia Elétrica da Universidade de Tecnologia de Eindhoven na Holanda recolhidos no período de 2000 a 2009 e que são compostos por testes de aptidão prévios, desempenho acadêmico, realizações precedentes da faculdade, dados demográficos, entre outros. Ainda segundo Dekker, pode-se obter taxas de acerto entre 75 e 80% utilizando os algoritmos de árvores de decisão disponíveis na ferramenta WEKA.

ROMERO et al. (2008) apresentam uma ferramenta integrada ao Moodle para a configuração e execução de técnica de mineração de dados tentando assim tornar mais fácil a utilização da EDM para os educadores na tarefa de predição de desempenho acadêmico. Desta forma, esta ferramenta foi integrada com o Moodle da Universidade de Córdoba na Espanha. Para os testes foram utilizados dados de 438 alunos de 7 cursos diferentes, sendo que os alunos são previamente classificados de acordo com seu desempenho acadêmico em 4 estados: reprovado (quando menor que 5), aceito (se for maior ou igual a 5 e menor que 7), bom (se for maior ou igual a 7 e menor que 9), e excelente (se for maior ou igual a 9). Os resultados obtidos neste experimento alcançam valores de até 65% utilizando algoritmos como o CART e C4.5.

Uma outra abordagem é proposta por BOYER; VEERAMACHANENI (2015), que

busca construir modelos preditivos de fácil generalização que possam prever em tempo real o comportamento dos alunos de cursos on-line com o objetivo de prever a probabilidade do aluno abandonar o curso. Para isto o autor utiliza os dados dos alunos de um determinado curso oferecido pelo Massachusetts Institute of Technology(MIT), estes dados estão disponíveis no ambiente virtual de aprendizagem e entre eles estão tempo total on-line, número de submissões, percentual das submissões certas, número de problemas corrigidos, entre outros. O autor propõe a separação dos dados por semanas e que nestes conjuntos de dados sejam utilizadas técnicas de aprendizagem de máquina gerando modelos. Estes por sua vez são aplicados às próximas turmas do curso. Nos experimentos realizados, os resultados obtidos variam entre 50% e 90% utilizando Naive Bayes, Transductive learning e Logistic Regression. Um dos principais destaques deste trabalho é sua proposta de utilizar somente dados que podem ser facilmente generalizados, entretanto, nem todos os ambientes virtuais contam com os dados utilizados o que pode acabar dificultando essa generalização.

A pesquisa de HALAWA; GREENE; MITCHELL (2014) propõe um preditor de alunos em risco de evasão que antecipe essa situação em 14 dias a partir dos dados das interações dos alunos com o ambiente. Desta forma, segundo o autor, é possível a reversão da situação de evasão. Para alcançar o objetivo são aplicados os algoritmos, que o autor não esclarece quais são, sobre os dados de se o aluno assistiu todas as vídeo aulas, se ignorou algum determinado material ou atividade, o atraso do aluno no acompanhamento do material(se ele está em dia com as aulas ou tem aulas atrasadas) e o desempenho nas atividades. A partir disso, busca-se classificar os alunos em 3 bandeiras, verde que representa os alunos que estão em baixo risco de evasão, amarela alunos que apresentam risco de evasão moderado e vermelho para os alunos que estão em situação de alto risco de evasão. Com isso são alcançados resultados entre 40% e 50% na predição com duas semanas de antecedência.

COHEN (2017) em sua pesquisa busca identificar mudanças na atividade dos alunos durante o período do curso, assim tentando a identificação precoce de alunos em risco de evasão antes que eles realmente abandonem o curso. Para isso, é proposta uma metodologia que faz a análise mensal dos dados das atividades dos estudantes a partir dos arquivos de logs de duas disciplinas disponibilizadas do ambiente virtual de aprendizagem. Este estudo de caso utiliza dados de uma universidade de Israel, em cursos presenciais com algumas disciplinas sendo disponibilizadas no ambiente virtual. Como resultados foram obtidos até 66% de acurácia na precisão de alunos em situação de evasão.

A proposta de BURGOS et al. (2017) é a utilização de mineração de dados e a criação de modelos de predição utilizando regressão logística linear para prever o risco de evasão de alunos. São utilizados dados de 104 alunos de diversos cursos de pequena duração (20 semanas) na modalidade a distância. Os resultados demons-

trados pelo mesmo apresentam valores de até 100% de acurácia geral já na quarta semana do curso. Segundo o autor, a aplicação desta técnica junto a um plano de tutorial diminuiu em 14% o abandono escolar nos cursos em que foram aplicadas.

Já KANTORSKI et al. (2016) propõe uma metodologia visando a predição da evasão em cursos superiores presenciais, aplicada sobre dados de 791 estudantes de um curso de uma universidade brasileira. Como dados de entrada, é proposta a utilização de dados demográficos como sexo, idade, estado civil, formação básica entre outros e também dados sobre participação em programas de auxílio estudantil, como, se o estudante é detentor de bolsa e utiliza acompanhamento psicossocial. São aplicados 4 algoritmos com precisão de até 73%.

Tabela 2: Comparativo entre trabalhos relacionados - Parte 1

Trabalho	Objetivo	Fonte dos Dados	Informações Utilizadas	Modelos Utilizados	Tipo de ensino	Melhores Taxas de sucesso (%)	Pontos fortes do trabalho	Limitações	Abordagem Validada em diferentes contextos
Jayaprakash (JAYAPRAKASH et al., 2014)	Evasão e retenção	Sistema acadêmico e Ambiente de auxilio ao ensino SAKAI	Dados demográficos, performance acadêmica, interações com ambiente virtual de 9938 alunos	Regressão lógica SVM/SMO Naive Bayes J48	Superior - Presencial	Variando de 62,50 a 84,82 para predição de evasão. Melhor modelo = 84,82	Resultados obtidos e apoio da comunidade acadêmica.	Alguns alunos mesmo com o auxilio acabam por não melhorar suas práticas.	Sim, testes efetuados em outras universidade de características similares.
Lykourentzou (LYKOURENTZOU et al., 2009)	Evasão e retenção	Sistema acadêmico e Ambiente virtual de ensino Moodle	Dados demográficos, nível educacional, performance acadêmica e interações com ambiente virtual de 193 alunos	Redes Neurais SVM/SMO PESFAM Combinação 1 Combinação 2 Combinação 3	Especialização	Variando de 73 a 94. Na predição de evasão. Melhor modelo = 94	Resultados obtidos com o método de combinação dos algoritmos	Dinâmica dos cursos.	Não difícil replicação dos dados.
Yavad (YADAV; PAL, 2012)	Desempenho acadêmico	Sistema acadêmico	Dados demográficos e de desempenho acadêmico	Árvores de decisão com C4.5, ID3 e CART	Superior - Presencial	Variando de 62,22 a 67,77	Resultados obtidos provam que pode ser possível a predição do desempenho em alunos ingressantes nos cursos	Desempenho com grande não foi superior a com pequena quantidade de dados como esperado	Não, restrito aos dados utilizados nos testes.

Tabela 3: Comparativo entre trabalhos relacionados - Parte 2

Manhaes(MANH ˜ aES et al., 2011)	Evasão	Sistema acadêmico	Desempenho acadêmico	OneR, JRip, DecisionTable, SimpleCart, J48, RandomForest, SimpleLogistic, Regressão lógica linear.	Superior – Presencial	Variando entre 75% e 80%.	Abordagem inovadora que visa a predição com grande antecedência e resultados mostrou-se viável.	Desbalanceamento da base de dados dificulta influência diretamente na acurácia geral.	Não, aplicado somente aos dados dos testes.
Douglas ´ Detoni(DETONI; ARAUJO; CECHINEL, 2014)	Reprovação	Ambiente virtual Moodle	Desempenho acadêmico	Redes Bayesianas, Redes Neurais, J48 e Random Forest	Superior - a distância	Variando entre 56 e 67 Melhor modelo Redes Bayesianas com = 67	Predição utilizando somente interações provou-se viável.	Menor precisão que em trabalhos que utilizam mais dados	Sim, modelos aplicados a duas bases de dados distintas.
Cambruzzi(CAMBRUZZI; RIGO; BARBOSA, 2015)	Evasão, Analise textual e desempenho	Sistema acadêmico e ambiente virtual Moodle	Não especificado	Redes Neurais	Superior presencial e a distância	87 no predição de alunos em risco de evasão	Diminuição de 11% na taxa de evasão nos cursos aplicados	Dificuldade na predição com certa antecedência da situação do aluno.	Aplicado a dados de diferentes turmas e cursos da instituição.
Yukselturk(YUKSELTURK, 2014)	Evasão	Questionário aplicado aos alunos	Dados demográficos	KNN Árvores de decisão Naive Bayes Redes Neurais	Especialização a distância	Melhor resultado com o KNN = 87	Resultados obtidos	Dados dos questionários nem sempre podem bater com a realidade, assim dificultando o trabalho de predição	Somente a dados de um curso.

Tabela 4: Comparativo entre trabalhos relacionados - Parte 3

Dekker (DEKKER; PECHENIZKIY; VLEESHOUWERS, 2009)	Evasão	Diversas	Teste de aptidão, desempenho acadêmico, dados demográficos e etc	Arvores de decisão	Superior - Presencial	Variando de 75 a 80 Melhor resultado 80	Abordagem baseada em diversos dados, tentando assim facilitar o processo de predição	Abordagem não supre os alunos que se evadem ainda no primeiro semestre	Não - Particularidade dos dados impede a generalização
Romero e Ventura (ROMERO et al., 2008)	Desempenho acadêmico	Sistema acadêmico e Ambiente Virtual Moodle	Desempenho e Interações	Diversos com destaque para Cart e C4.5	Superior-Presencial	Melhor resultado com Cart e C4.5 = 65	Integração com o Moodle acarretando em uma fácil utilização	Resultados	Sim - Pode ser utilizado em outro contexto presencial
Cohen	Evasão	Sistema acadêmico e Ambiente Virtual	Análise das atividades virtuais dos alunos no ambiente virtual	Não especificado	Superior Presencial	Até 66%	Abordagem inovadora que visa a predição com grande antecedência e resultados	Modelo restrito ao ensino presencial	Não - Difícil aplicação em outros contexto devido a especificidades dos dados
Burgos	Evasão	Sistema acadêmico e Ambiente Virtual	Análise das atividades dos alunos	Regreção logistica linear	Cursos de pequena duração (20 semanas)	Até 100%	Resultados muito interessantes e aplicação prática com bons resultados	Cursos de baixa duração e baixa quantidade de dados utilizados	Sim - desde que aplicados a dados de cursos parecidos
Kantorski	Evasão	Sistema acadêmico	Análise de dados demográficos e participação em programas de auxilio.	Cart, J48 e Naive Bayes	Cursos Superiores Presenciais	Até 73%	Abordagem interessante pela utilização de dados como participação em programas de auxílio.	Especificidade dos dados no geral	Não - Difícil aplicação em outros contexto devido a especificidades dos dados

3.2 Análise dos trabalhos relacionados

Como podemos ver, existem diversos trabalhos relacionados à área, entretanto eles se diferenciam quanto ao tipo de dados que utilizam, tipo de ensino utilizado no estudo, as fontes de dados utilizadas e aos objetivos. Além disto, mesmo trabalhos com foco muito próximo ainda apresentam resultados muito diferentes e podem ser ou não de fácil generalização.

A classificação inicial dos trabalhos se dá pelo seu objetivo, que geralmente varia entre predição de evasão e retenção e/ou predição de desempenho escolar. Os trabalhos voltados para a predição da evasão buscam prever a situação final do aluno no curso, classificando estes como em risco ou normal. Enquanto que os trabalhos que buscam a predição do desempenho escolar, diferenciam-se quanto a classificação por disciplina como ROMERO et al. (2008) ou semestre CAMBRUZZI; RIGO; BARBOSA (2015).

No conceito de generalização em mineração de dados exposto por BRAGA (2005), um modelo muito complexo e que leve em consideração muitos dados particulares pode se ajustar bem aos dados de treinamento e, no entanto, não ter um bom desempenho para outros dados. Ainda vale ressaltar que nem sempre os dados específicos utilizados para geração dos modelos vão estar disponíveis para os testes em outros locais.

Em relação aos dados utilizados podemos notar que entre as pesquisas analisadas pelo menos 5 utilizam dados demográficos em sua composição, enquanto 4 utilizam dados das interações dos alunos com ambientes virtuais e uma utiliza somente o desempenho escolar. Apenas uma pesquisa utiliza somente os dados de interações como base.

A utilização de dados específicos pode acabar por dificultar o processo de generalização, tendo em vista que variáveis demográficas não são simples de serem repetidas em situações ou até mesmo localidades diferentes. Como exemplo podemos citar testes que são aplicados em um determinado local e podem ou não ser aplicados em outros como o teste SAT (aplicado por algumas universidades dos Estados Unidos), ainda o grau de escolaridade pode ser diferente quando os cursos são diferentes e variáveis como sexo e idade podem ter impactos diferentes, dependendo do ambiente.

Nos trabalhos analisados geralmente os autores optam por seguir o fluxo dado na Figura 11 em suas pesquisas. Assim, esse é o fluxo seguido na metodologia desenvolvida nesse projeto.

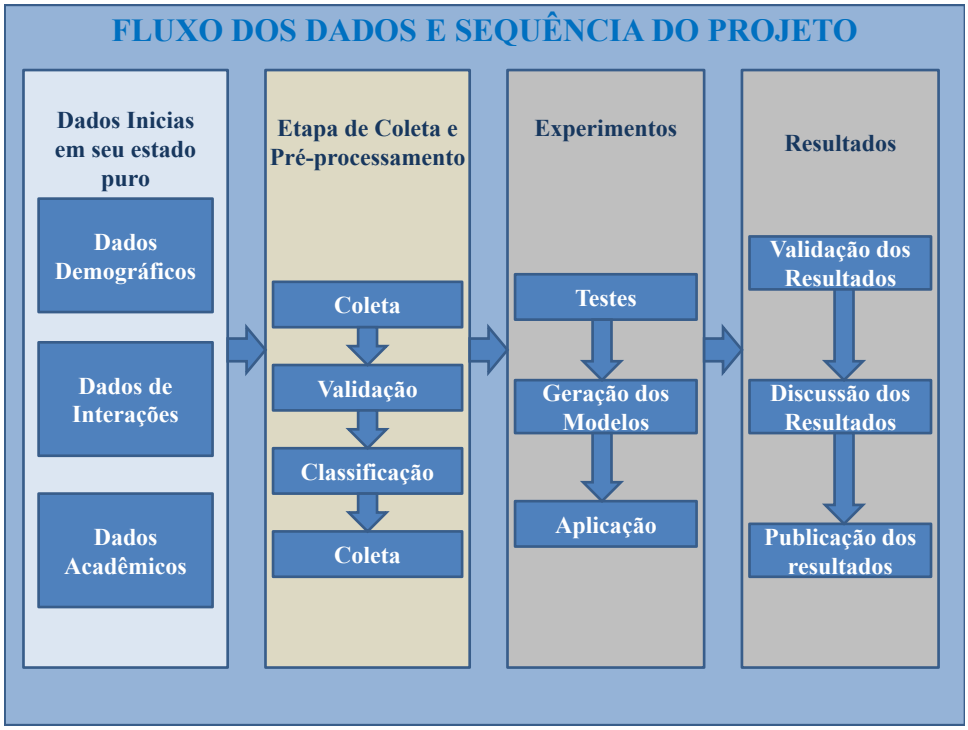

Figura 11: Fluxo de dados e sequência dos projetos

Podemos usar como exemplo o trabalho de LYKOURENTZOU et al. (2009) que apesar dos resultados expressivos apresentados, dificilmente poderia ser aplicado em outros locais fora da sua instituição, pois leva em conta dados muito específicos o que pode tornar mais difícil a generalização dos modelos. Estes dados podem ser tanto os dados demográficos quanto até mesmo dados específicos dos cursos como sua duração, onde nesta pesquisa citada o autor análisa cursos com duração de apenas 8 aulas.

A pesquisa de JAYAPRAKASH et al. (2014) é outro exemplo que podemos analisar quanto a generalização, pois mesmo outras universidades dos EUA teriam problemas para utilizar seus modelos devido a grande especificidade dos dados utilizados. Dados como os retirados do SAKAI ou questões demográficas podem não estar disponíveis em outros locais, o que inviabiliza a utilização dos modelos em outros ambientes.

Entretanto trabalhos como o de DETONI; ARAUJO; CECHINEL (2014) acredita-se que possam ser de fácil generalização, pois trabalham somente com as interações geradas no ambiente virtual. Assim facilitando a generalização de seus modelos e estes podendo ser aplicados em outras instituições que desejem utilizar a EDM como forma de auxílio na predição de desempenho de alunos.

Ainda temos um grande problema quanto à ocultação de técnicas e resultados específicos, pois muitos autores como CAMBRUZZI; RIGO; BARBOSA (2015) não demonstram algumas partes interessantes de suas pesquisas. Neste caso, por exemplo, o autor não revela especificamente os dados utilizados e sim somente de onde

estes dados são retirados. Outro problema na mesma linha é a falta de resultados específicos, como em caso de evasão as taxas de verdadeiro positivo.

A etapa de pré-processamento dos dados é de grande importância para o andamento dos projetos e os diferentes autores acabam diferindo na técnica utilizada mesmo quando trabalham com tipo de dados iguais como contagem de interações. DETONI; ARAUJO; CECHINEL (2014) por exemplo, utiliza em sua pesquisa contagem de interações semanais separando as interações dos alunos, tutores e professores. Enquanto que LYKOURENTZOU et al. (2009) utiliza contagem de interações por aula do curso tendo em vista a breve duração dos cursos avaliados. JAYAPRAKASH et al. (2014) utiliza a totalização das interações no decorrer do primeiro semestre do curso, totalizando-as no final sem distinção da sua data, mas levando em conta a data de entrega de trabalhos ou tarefas no ambiente virtual.

Nos resultados, as diversas pesquisas tendem a ter valores de acurácia bem diferentes. Na questão da evasão, os valores variaram entre 62,50% e 94%. Enquanto que na predição do desempenho acadêmico os valores variaram entre 56% e 67,77%. Entretanto, estes valores podem trazer uma distorção nos fatos, pois nem sempre os autores divulgam os resultados mais importantes na predição, como a taxa de verdadeiros positivos.

Desta forma, fica aparente uma lacuna na área de predição de evasão, pois mesmo com diversos modelos estes costumam levar em conta dados muitos específicos e que nem sempre estão disponíveis em diferentes contexto de educação a distância. Neste sentido, este trabalho diferencia-se dos relacionados por apresentar uma proposta de metodologia para geração de modelos de predição de alunos em risco de evasão, que utiliza a contagem de interações e suas variações o que pode permitir contemplar diferentes domínios de aplicação.

4 METODOLOGIA E EXPERIMENTOS

Nesse capítulo é apresentado o contexto dos dados utilizados nos experimentos, a metodologia proposta e seu desenvolvimento, bem como os experimentos realizados. No andamento da pesquisa foram alcançados resultados iniciais que auxiliaram a delinear a metodologia utilizada nesta dissertação, sendo estes publicados em conferências como, QUEIROGA; CECHINEL; ARAÚJO (2015) e QUEIROGA et al. (2016).

4.1 Contexto

Para o desenvolvimento deste projeto foram utilizados dados de quatro cursos técnicos na modalidade a distância(EAD) do Instituto Federal Sul-riograndense (IF-Sul) campus Visconde de Graça (CaVG). Estes cursos são ministrados em 18 polos espalhados pelo interior do estado do Rio Grande do Sul e funcionam com atividades semanais, que são postadas no ambiente pelo professor, com os alunos tendo uma semana para o desenvolvimento destas com auxílio dos tutores.

Cada curso tem um tempo de realização máximo de 103 semanas com carga horária total de 1215 horas divididas nas disciplinas do curso dentro do período de 24 meses, contando com 3 intervalos também chamados de férias, sendo que a situação final do aluno é determinada pelo seu resultado nas avaliações.

O prazo máximo para integralização do curso é de quatro (4) anos, podendo o aluno repetir somente uma vez cada disciplina e por consequência o ano. Ele ainda tem a opção de levar até 2 disciplinas como dependência para o próximo ano e cursá-las de forma concomitante às outras disciplinas do curso. Para aprovação o aluno deverá ter média igual ou superior a seis em cada uma das disciplinas da matriz curricular. Considera-se evadido o aluno que passe um período de 365 dias sem interações com o ambiente virtual ou não efetue sua rematrícula anual, sendo desligado do curso.

Assim, o aluno pode assumir 2 estados diferentes no final das atividades, aprovado ou reprovado, entretanto este estudo tem como objetivo a predição dos alunos que entrem em situação de evasão no decorrer do curso. Para tal define-se que o aluno será considerado evadido caso abandone, não efetue as atividades no decorrer do

curso e também sua matrícula no semestre seguinte.

Como este trabalho propõe modelos que possam ser de fácil generalização e que assim acredita-se que possam ser aplicados em outros cursos do IFSul ou até mesmo em outras instituições de ensino que utilizem o modelo da Rede e-TEC, se optou por utilizar as contagens diárias e semanais de interações dos alunos com o ambiente virtual.

4.2 Metodologia

A metodologia seguida para o desenvolvimento desse trabalho está dividida nas seguintes etapas: coleta de dados, pré-processamento dos dados, geração e avaliação dos modelos de predição.

4.2.1 Coleta

Esta etapa se iniciou com o recolhimento dos dados brutos das interações dos estudantes com o Moodle. Este processo precisou ser efetuado manualmente por uma condição técnica do servidor onde os cursos são disponibilizados, assim foram feitos os downloads dos logs de interações de cada uma das disciplinas separadamente e não a conexão direta com o banco de dados dos cursos.

Após a coleta, os dados foram identificados e separados por curso, sendo este um processo que necessita ser efetuado manualmente demandando um considerável período de tempo. A partir deste momento iniciou-se uma das etapas cruciais para um bom desenvolvimento do projeto que é a validação dos dados obtidos, onde são comparados os dados da situação dos alunos junto ao ambiente virtual e ao sistema acadêmico da instituição. Este também acaba sendo um trabalho manual, pois são dois sistemas separados e não integrados e que pode demandou um certo tempo pelo volume de dados envolvidos e pela possibilidade de haver alguma inconsistência.

Neste momento a situação dos alunos nas disciplinas ainda é tratada à parte, pois não são disponibilizados nos logs de ação do Moodle com os outros dados e sim em uma tabela disponível por disciplina e turma. Desta forma, estes dados são salvos a parte dos log's para junção em etapa posterior.

Os quantitativos dos dados coletados para utilização neste projeto são encontrados na Tabela 5.

Tabela 5: Quantitivo de dados utilizados

Cursos	Quant. Logs	Nº de alunos	Evadidos	Concluintes
Curso 1	682.773	407	212	195
Curso 2	1.033.910	729	301	428
Curso 3	933.221	615	246	369
Curso 4	1.051.012	752	354	398
Totais	3.700.916	2503	1113	1390

4.2.2 Pré-processamento dos dados

Com a total disponibilidade dos dados e sua validação iniciou-se o pré-processamento, etapa onde os dados foram limpos e classificados a partir de sua relevância para a utilização nos testes. Isso se dá por um processo manual a partir dos dados que tenham relevância e possam impactar no processo de predição da evasão. Os nomes dos alunos foram modificados por padrão atendendo as normas da instituição cedente, que solicita que informações que possam identificar os alunos sejam anonimizadas.

Como é possível observar na Tabela 5 foram coletados 3.700.916 logs de interações dos alunos, para o pré-processamento deste volume considerável de dados foi desenvolvido um sistema em Java que tem como objetivo automatizar esta etapa de pré-processamento dos dados , assim como também a etapa de geração e avaliação dos modelos de predição. Este sistema, além de outras funcionalidades, efetua a leitura dos arquivos de logs exportados do ambiente virtual e grava as informações em um banco de dados próprio. Na Tabela 6 é exposto a forma e o conteúdo dos dados exportados do ambiente virtual antes de passarem pelo pré-processamento.

Após os dados serem inseridos no banco de dados do sistema iniciou-se a separação dos dados por dia e semana, em um primeiro momento foi montado o calendário do curso a partir das datas de início, férias e término, resultando nas 103 semanas letivas de que os cursos são compostos.

Na sequência foi feita uma busca pela quantidade de interações que cada um dos alunos do curso efetuou por dia de curso e a soma das interações na semana efetuadas em cada uma das 103 semanas. Todos estes passos foram automatizados no sistema de pré-processamento, salientando-se que este sistema passou por uma longa etapa de desenvolvimento e aperfeiçoamento visando que todas as etapas de pré-processamento e geração dos modelos sejam efetuadas. Isto tem como objetivo aumentar a agilidade, a fácil generalização e diminuir as chances de erro no pré-processamento.

No final desta etapa obtivemos, conforme a Tabela 7 apresenta, um esquema no banco de dados para cada um dos cursos onde constam o id do aluno, 103 campos com as interações semanais dos alunos, 721 com o dia, 103 com a média, 103 com

Tabela 6: Modelo de log do Moodle

Dado	Descrição	Exemplo
Curso	Turma e curso ao qual o aluno esta matriculado.	Informática Aplicada - 2013/2 Administração
Hora	Data e hora que foi efetuada a ação que gerou este determinado log.	2012 abril 8 10:39
Endereço IP	O endereço IP ao qual o computador utilizado para o acesso ao Moodle tinha no exato momento do acesso.	187.86.133.66
Nome Completo	Nome cadastrado do aluno no ambiente virtual.	José Alves da Cunha
Ação	Tipo da ação geradora do log que foi efetuada pelo usuário.	Course view (http://Moodle.cavg.ifsul.edu.br /course/view.php?id=75)
Atividade	Se refere ao local/link onde foi efetuada a ação	Download - LibreOffice (BrOffice)

a mediana, 103 com o desvio padrão semanal, o total de interações e a situação final do aluno no curso.

Tabela 7: Variáveis Utilizadas

Variável	Descrição
Interações diárias (1 até 721 dias)	Contagem de interações diárias
Interações Semanais (1 até 103 semanas)	Contagem das interações do estudante na semana
Média semanal (1 até 103 semanas)	Média das contagens dos estudantes na semana
Mediana semanal (1 até 103)	Mediana do conjunto de interações na semana
Desvio padrão semanal (1 até 103)	Desvio padrão do conjunto de interações na semana
Id	Id do estudante
Situação final no curso	Situação final do estudante no curso

4.2.3 Modelo Estatístico descritivo

Com o objetivo de poder comparar os modelos de predição gerados por meio de aprendizagem de máquina com algum modelo inicial de predição, decidiu-se por criar um modelo baseado em informações obtidas unicamente por meio de estatística descritiva.

Este modelo se baseia na média e desvio padrão semanal das turmas. Por exemplo, um aluno que está com um desvio padrão N acima da média semanal da turma se evade, isso pode representar que isto seja uma tendência. Da mesma forma, um aluno com um desvio padrão N abaixo da média está mais propenso a se evadir.

Como definição de média aritmética podemos utilizar que, a média é definida como o valor que mostra para onde se concentram os dados de uma distribuição como o ponto de equilíbrio das frequências em um histograma. Média também é interpretada como um valor significativo de uma lista de números(BRASIL; NATUREZA, 2002).

O desvio padrão (dp) é o resultado positivo da raiz quadrada da variância. Na prática, o desvio padrão indica qual é o “erro” se quiséssemos substituir um dos valores coletados pelo valor da média.

O modelo de predição criado funciona então da seguinte maneira: considera acadêmicos que possuam a quantidade de interações duas vezes abaixo ou acima do desvio padrão da turma como evadido, e considera acadêmicos dentro da faixa de -2 a 2 como aprovados.

Para ser considerado fora do desvio padrão foi definido um valor abaixo de -2 do desvio ou acima de +2. Assim o aluno que estivesse fora dessa faixa era considerado em risco de evasão.

Este modelo estatístico faz parte apenas dos experimentos finais deste trabalho.

4.2.4 Geração e avalição dos modelos de predição

Com as bases de dados validadas e separadas podemos iniciar a implementação dos algoritmos que serão testados. A definição pelos algoritmos que serão implementados passa pela primeira etapa onde acontece a descoberta do estado da arte na mineração de dados educacionais principalmente com ênfase na predição de situação e comportamento de alunos em ambientes virtuais.

No andamento deste projeto diversos experimentos foram realizados antes do delineamento completo da metodologia que seria implementada, sendo que dois desses experimentos demonstraram maior relevância e serão brevemente relatados aqui, sendo eles: **Experimento 1** - Utilização de dados das disciplinas iniciais de 1 (um) único curso; e **Experimento 2** - Utilização de dados de todas as disciplinas de 2 (dois) cursos.

O **Experimento 1** tinha como objetivo geral verificar a possibilidade da descoberta de informações úteis nas bases de dados disponíveis. Para isto, foi utilizada uma adaptação da técnica proposta por MANHãES et al. (2011), que consiste na aplicação de mineração de dados sobre os dados das disciplinas iniciais do curso.

Nesta etapa do estudo, optou-se pela utilização dos dados de duas turmas de polos diferentes, mas que tiveram o inicio de suas aulas na mesma data em 2013. A turma do primeiro polo é composta por 50 alunos, enquanto que a do segundo tem 34 alunos. Foram utilizadas as interações de duas disciplinas cursadas pelos alunos do curso, tendo respectivamente carga horária de 45h e 60h. Estas disciplinas são as primeiras interações dos alunos com a EAD e o Moodle (QUEIROGA; CECHINEL; ARAÚJO, 2015).

Os resultados mais significativos deste experimento demonstraram que a quantidade de dados nas fases de treinamento poderia apresentar um impacto maior na acurácia que o polo a que o aluno estava matriculado. Assim no terceiro teste foram obtidos valores de até 79,79%.

O **Experimento 2** tinha como objetivo verificar os resultados que poderiam ser obtidos utilizando dados de dois cursos com a metodologia inicialmente proposta, além de testar a aplicabilidade da contagem de interações semanais em bases de dados maiores e a generalização dos modelos gerados.

Para isto foram coletados os dados de dois cursos técnicos na modalidade a distância do Instituto Federal Sul-riograndense (IFSul), gerando um total de 1.716.683 logs de interações dos alunos. Para o seu pré-processamento foi utilizada a primeira versão do sistema em Java que estava sendo desenvolvido.

Neste teste, foram gerados 103 modelos diferentes treinados com os dados do curso 1 e aplicados diretamente nos dados do curso 2. Isso tem como objetivo testar a aplicabilidade de modelos de predição de fácil generalização em outros cursos. Estes modelos necessitaram de uma maior quantidade de dados para obterem resul-

tados mais satisfatórios, entretanto, a partir da semana 11 esses resultados já atingem 63,12% de acurácia e na semana 26 já obtém 91,26% de acerto. Ainda foram alcançados percentuais de verdadeiros positivos de até 98,67%, resultado obtido utilizando o algoritmo J48 na semana 51 do curso.

4.2.5 Configuração do experimento final

Após o desenvolvimento dos experimentos anteriores e a publicação dos mesmos, surgiram algumas dúvidas. Entre elas estavam se a inserção de novos dados além da contagem semanal de interações traria algum ganho nos resultados preditórios.

Buscando esclarecer algumas destas dúvidas foram inseridos os dados das interações diários de cada aluno e as médias de interação da turma por semana. Tendo como objetivo a diminuição da granularidade dos dados e verificar o impacto dessas novas informações nos valores percentuais de acurácia atingidos pelos algoritmos.

Além dos dados já citados também foi inserido o modelo estatístico descritivo. Este modelo, como citado anteriormente, tem como base as médias e desvio padrão das interações diárias e semanais dos alunos.

Com a considerável quantidade de dados disponíveis foi possível a execução de diferentes testes, assim foram criados dois cenários diferentes de aplicação onde em cada um deles o treinamento e a aplicação dos modelos eram distintos.

No primeiro cenário, como demonstra a Figura 12, cada um dos quatro cursos é testado separadamente utilizando o método de validação cruzada com 10 folds. Assim os modelos são gerados utilizando 9 subconjuntos diferentes e o teste é feito em 1 subconjunto, esse processo é repetido 10 vezes e a acurácia se dá pela média dos 10 testes.

Figura 12: Cenário 1

No segundo cenário, como demonstra a Figura 13, é gerada uma base com os dados de três cursos para a etapa de treinamento dos algoritmos utilizados e a aplicação

dos modelos resultantes é efetuada no curso restante. Assim simulando minimamente a generalização na aplicação dos modelos de predição criados neste trabalho.

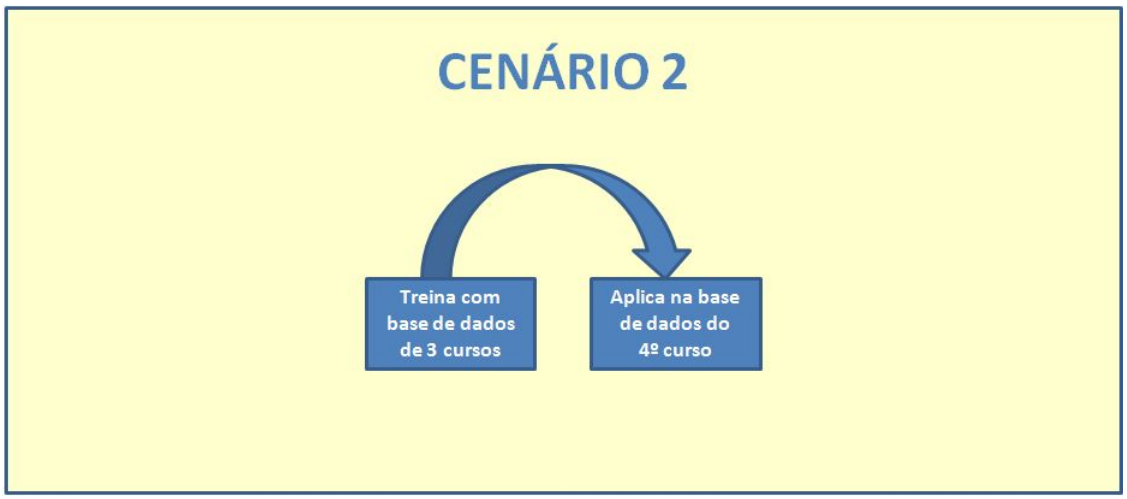

Figura 13: Cenário 2

Para a execução desta implementação optou-se pela utilização de um hardware comum ao mercado atual tentando assim simular o ambiente que o sistema seria utilizado por um usuário normal. Este hardware era composto por um processador intel i5 4670 de quarta geração, 8 gb de memória ram e sistema operacional Microsoft Windows 7.

Com as características do hardware utilizado o tempo de processamento acabou por ser variável, tanto pelo volume de dados dos experimentos quanto pelas características próprias de cada um dos algoritmos. Por exemplo o algoritmo Multilayer Perceptron teve um tempo de execução médio nos experimentos de 30 minutos em comparação ao Bayes Net onde o tempo médio foi de 30 segundos.

Cada um dos experimentos feitos neste trabalho consiste na geração de um modelo por algoritmo para cada semana do curso. Desta forma, como os cursos tem 103 semanas são criados 103 modelos por algoritmo. Assim para a semana 1 de um curso, são criados 5 modelos (um para cada algoritmo). No total temos 515 modelos aplicados em cada um dos cursos por cenário.

4.3 Resultados encontrados

Neste capítulo serão apresentados os resultados obtidos nos dois cenários diferentes. A acurácia dos resultados é medida utilizando o percentual de Verdadeiros Positivos (VP), que consiste na taxa de acertos em prever se um aluno irá se evadir, e no percentual de Verdadeiros Negativos (VN), que consiste taxa de acertos em prever que um aluno irá finalizar o curso com êxito.

4.3.1 Resultados Cenário 1

No cenário 1 os 515 modelos são gerados e aplicados na mesma base de dados do curso utilizando validação cruzada como dito anteriormente. Alguns dos resultados obtidos neste cenário são analisados a seguir, os demais podem ser observados no Anexo A desta dissertação. Nas Figuras 14 e 15, são expostos respectivamente os resultados de Verdadeiros Positivos (VP) e Verdadeiros Negativos (VN), encontrados no curso 3 desse cenário.

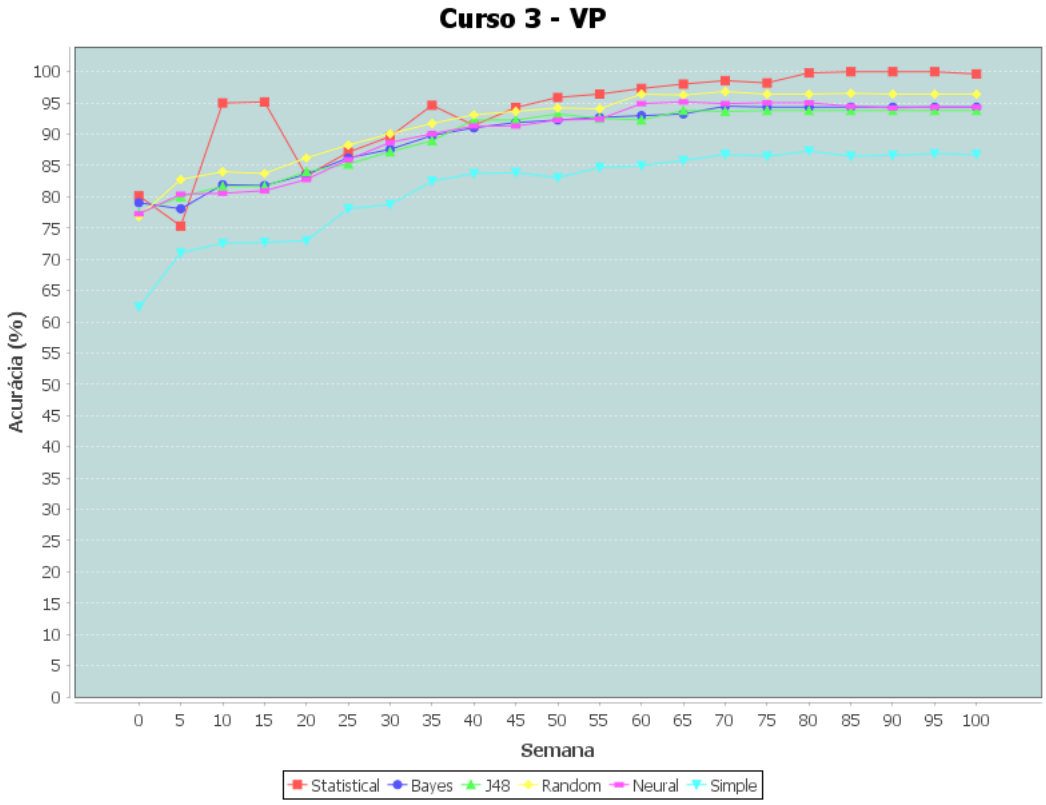

Figura 14: Resultado Cenário 1 Experimento final Curso 3 VP

Como podemos observar na Figura 14, em geral, os diferentes algoritmos apresentam desempenhos semelhantes para classificação de estudantes em risco de evasão. Destacam-se os resultados obtidos pelo modelo com base nas estatísticas descritivas e do Random Forest, tendo este último apresentado a maior taxa de acerto entre os algoritmos que utilizam aprendizagem de máquina.

Nesse curso desde a primeira semana as taxas de acerto foram superiores a 75% em quase todos os algoritmos, excetuando-se os resultados obtidos pelo Simple Logistic. Com o passar das semanas as taxas de acerto se elevam para 87% antes da semana 25 que é o fim do primeiro semestre do curso. No segundo semestre já são obtidos resultados próximos a 94% antes do final do mesmo. Assim é possível dizer com quase 95% de exatidão se um aluno irá terminar o curso antes do final do primeiro ano do mesmo. Nos dois últimos semestres do curso os resultados ultrapassam

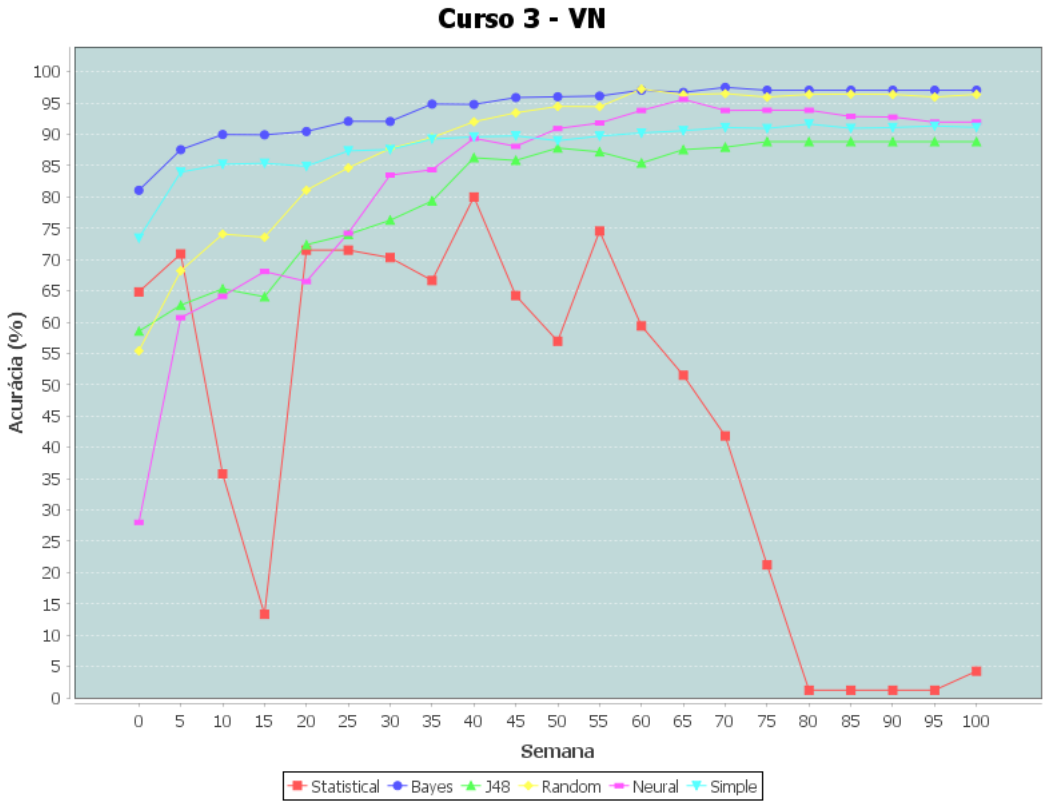

Figura 15: Resultado Cenário 1 Experimento final Curso 3 VN

os 95% com o Random Forest, chegando até 96%. Os demais algoritmos tem seus resultados próximos aos 95%.

Ainda na Figura 14, podemos observar o resultado obtido pelo modelo estatístico. Esse experimento especificamente foi o único onde este modelo obteve resultados próximos, e em alguns momentos até mesmo maior, aos dos algoritmos de aprendizagem de máquina. Entre a semana 5 e a 20, este modelo chega a obter taxas de até 95% e no último semestre resultados de até 100% de acerto.

É importante notar que, embora as taxas de VP do modelo com base em estatísticas descritivas sejam altas, este modelo não é capaz de prever o sucesso dos alunos em um nível satisfatório. De fato, esses experimentos para esse curso específico (cenário 1 para o Curso 3) são os únicos em que tais modelos apresentaram resultados semelhantes em comparação aos obtidos por modelos treinados de aprendizado de máquinas.

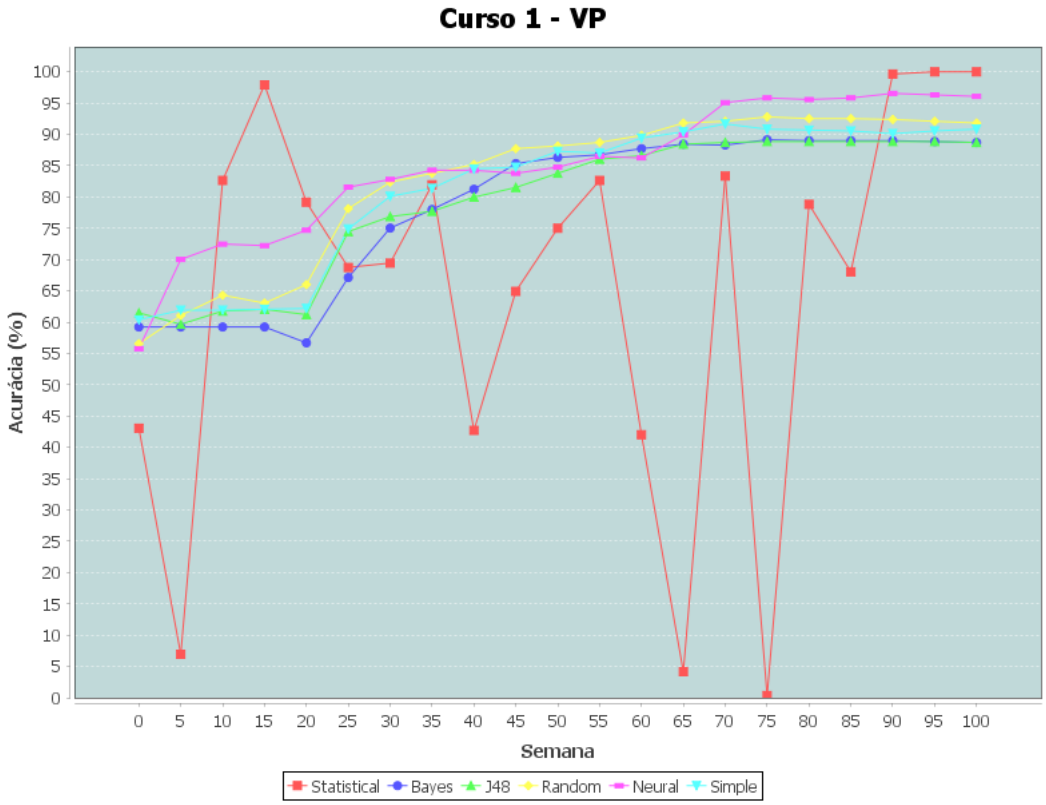

Figura 16: Resultado Cenário 1 Experimento final Curso 1 VN

Os desempenhos instáveis dos modelos com base em estatísticas descritivas podem ser confirmados, observando as Figuras 16 e 17, que apresenta os resultados obtidos para o Curso 1. Como se pode notar, os modelos do Curso 1 neste primeiro cenário apresentaram menor performances em comparação com o Curso 3. No entanto, ainda é possível ver que as performances em termos de VP e TN podem ser consideradas de alguma forma satisfatórias, pois permitem classificar ambas as categorias de alunos (evasão e sucesso) com 81% de precisão antes do final do primeiro semestre (semana 25). Para este curso, Multilayer Perceptron apresentou os melhores desempenhos, seguido de Random Forest e Simple Logistic.

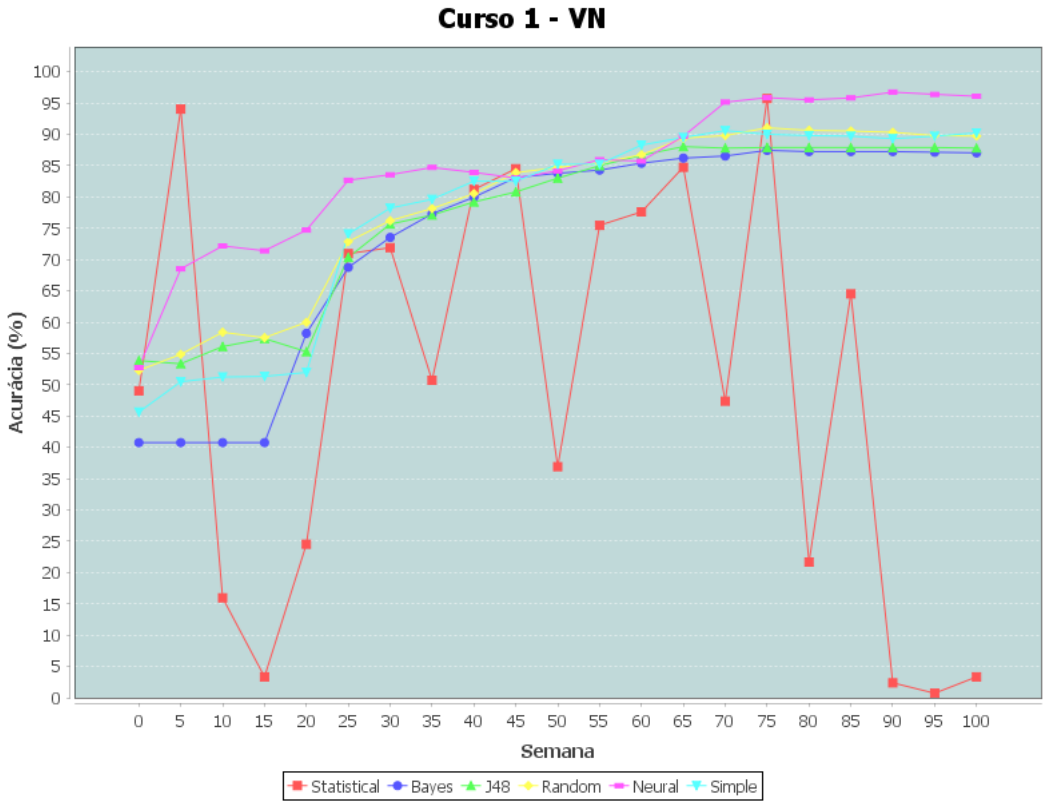

Figura 17: Resultado Cenário 1 Experimento final Curso 1 VP

4.3.2 Resultados Cenário 2

O cenário 2 consiste na simulação da generalização dos modelos, para isto é gerado um modelo por semana utilizando a base de dados de 3 cursos e este é aplicado na base de dados do 4º curso. Esta tarefa é feita 4 vezes, assim todos os cursos em algum momento ficam somente para a etapa de teste.

Nas Figuras 18, 19, 20 e 21, são expostos respectivamente os resultados de Verdadeiros Positivos (VP) e Verdadeiros Negativos (VN), encontrados utilizando como teste os cursos 3 e 1 respectivamente.

A Figura 18 apresenta os resultados obtidos quanto a Verdadeiros Positivos (VP) na predição de alunos em risco de evasão, para isto foram utilizados os dados dos cursos 1, 2 e 4 para geração e treinamento dos modelos de predição e consequente aplicação no curso 3. Nessa figura podemos observar que desde a primeira semana de curso todos algoritmos testados obtém taxas de acurácia na predição de alunos em risco de evasão superiores a 77%.

Nesse experimento, destaca-se o algoritmo Random Forest tendo ele obtido já na primeira semana valores acima de 95% alcançando 99% a partir da quinta semana de curso e mantendo-se próximo a esse valor até o final do curso, tendo em alguns momentos inclusive alcançado o valor de 100% de acerto. O algoritmo J48 também merece destaque neste experimento. Desde as primeiras semanas já é obtido 84%

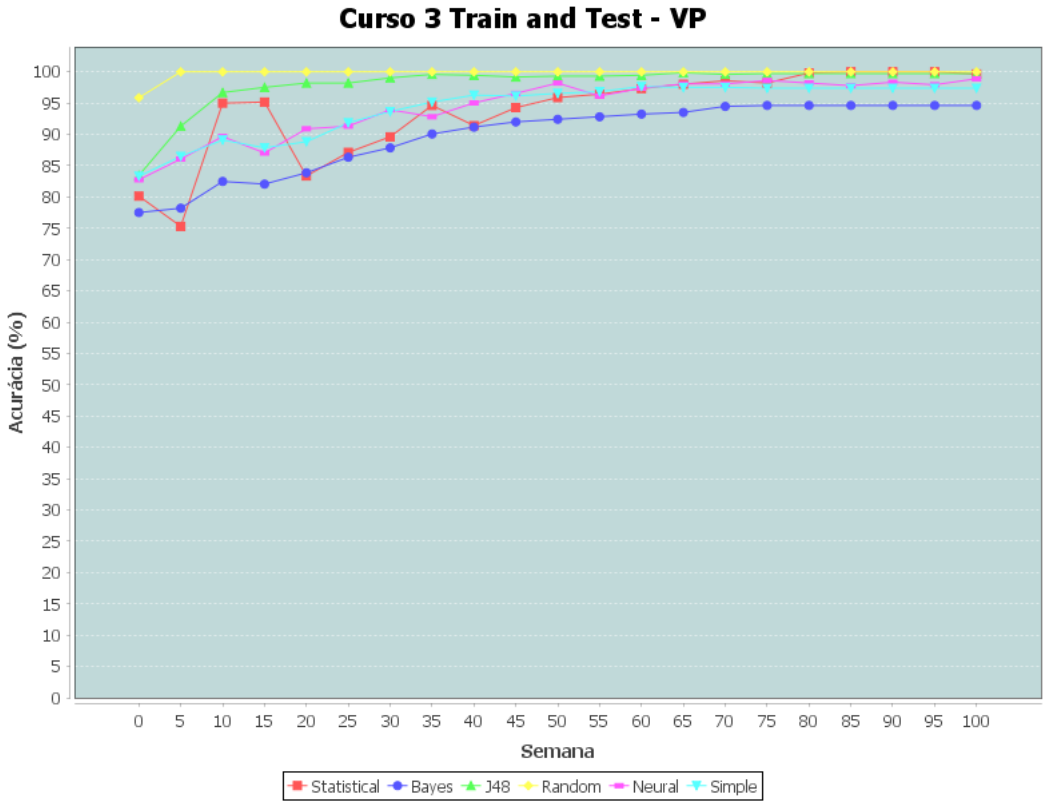

Figura 18: Resultado Cenário 2 Experimento final Curso 3 VP

de acurácia, com esse valor crescendo gradativamente até alcançar 97% no fim do primeiro semestre do curso (semana 25).

Na Figura 19 é possível observar os resultados obtidos quanto a Verdadeiros Negativos (VN), ou seja, os estudantes que são previstos como sem risco de evasão e realmente concluem o curso. Para esta etapa de predição foram utilizados os dados dos cursos 1, 2 e 4 na geração e treinamento dos modelos de predição e a consequente aplicação no curso 3.

Neste experimento novamente o algoritmo Random Forest obteve as melhores taxas de acurácia. Nos testes feitos com esse algoritmo, desde a primeira semana já são atingidos resultados de 97%, alcançando 99% na semana 5 e posteriores 100% até o final do curso.

Os demais algoritmos utilizados no projeto desenvolvido nessa dissertação obtêm resultados muito próximos a partir da 30 semana. Destacando-se novamente o J48 como o algoritmo que obteve os segundos melhores resultados nos testes, sendo possível notar que a partir da quinta semana ele já obtêm valores próximos a 95% e que crescem até o final do curso.

Como pode ser visto na Figura 20, no experimento utilizando o curso 1 para a aplicação dos modelos os resultados nas primeiras semanas do curso ficam entre 65% com as Redes Neurais e 84% com o Simple Logistic. Esses percentuais se mantêm praticamente estáveis até a 25 semana de curso, onde começam a ter um

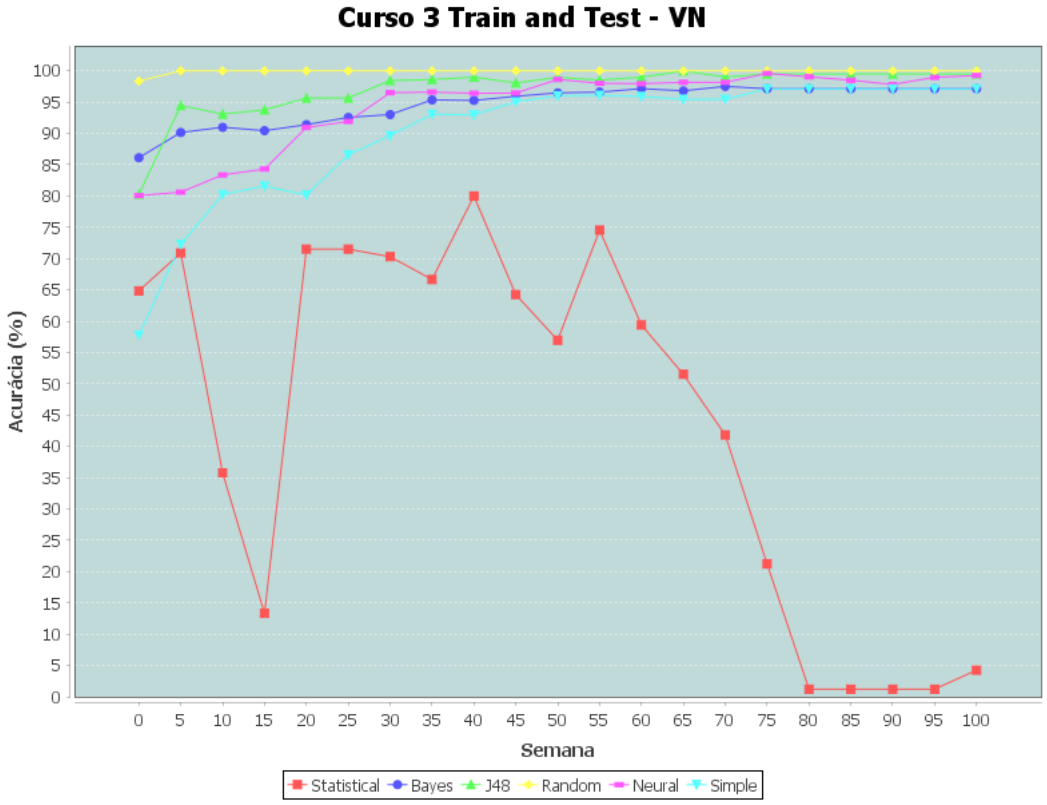

Figura 19: Resultado Cenário 2 Experimento final Curso 3 VN

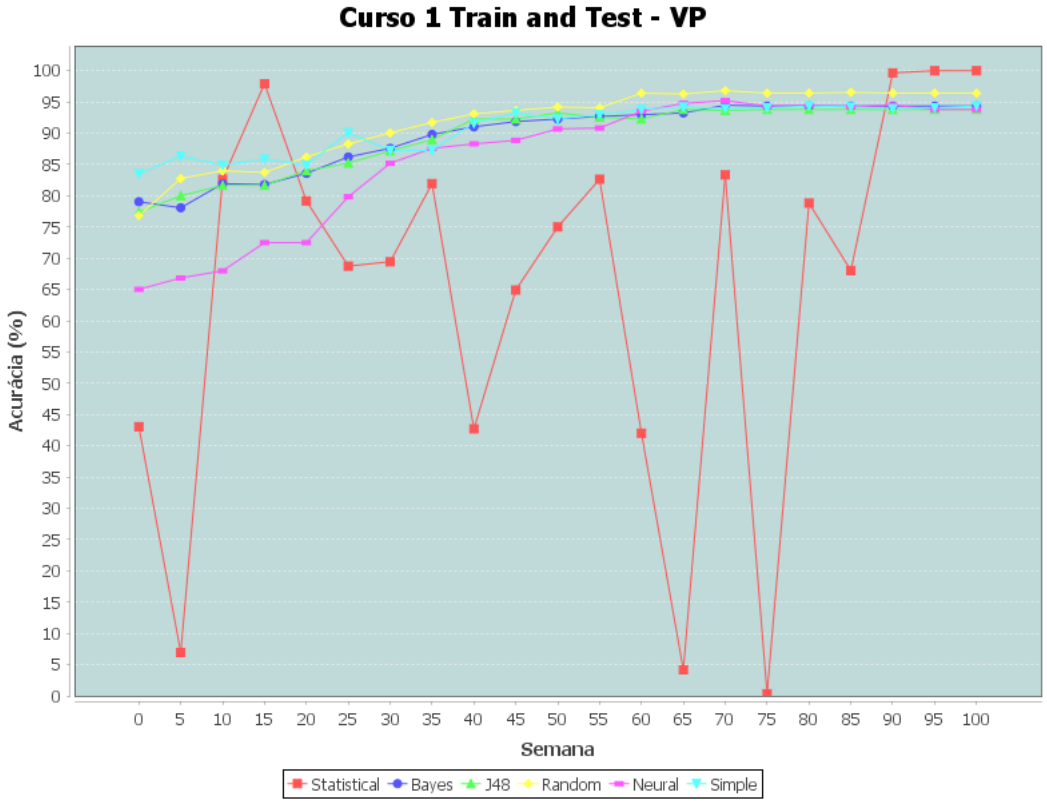

Figura 20: Resultado Cenário 2 Experimento final Curso 1 VP

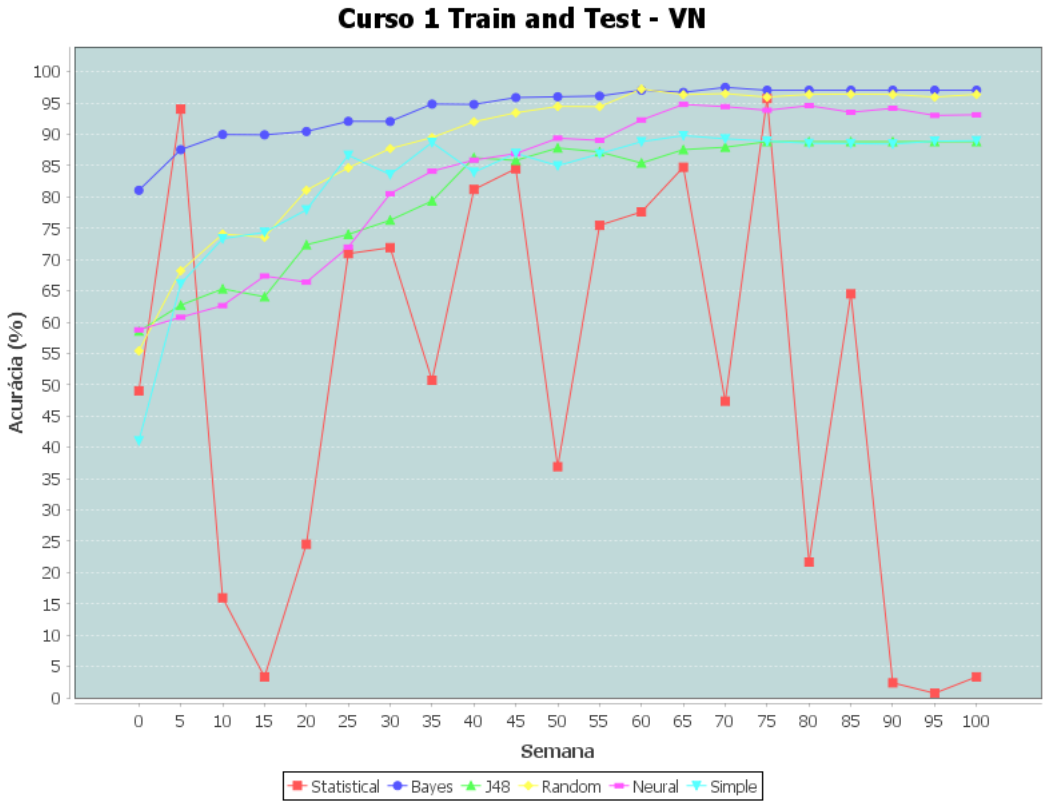

Figura 21: Resultado Cenário 2 Experimento final Curso 1 VN

crescimento chegando a 94% na semana 50 com Random Forest.

Neste experimento novamente o algoritmo Random Forest obteve, na maior parte do curso, as melhores taxas de acurácia chegando a 96% na semana 60 e mantendo-se assim até o fim do curso. Já o modelo estatístico em contraste com os resultados apresentados no curso anterior chega em alguns momentos a obter resultados próximos aos algoritmos de aprendizagem de máquina, e em alguns momento até mesmo superiores como 98% na semana 15.

Já utilizando a métrica de Verdadeiros Negativos (VN), como pode ser visto na Figura 21 o algoritmo BayesNet obtêm resultados próximos a 81% desde a primeira semana de curso. Esses resultados vão se elevando no decorrer do curso com valores de 90% na semana 10 e de pelo menos 95% partir da semana 35.

Os demais algoritmos neste experimento só apresentam resultados interessantes a partir da semana 20, onde obtêm valores próximos a 80%. Novamente em alguns momentos o modelo estatístico alcança resultados próximos aos dos algoritmos de aprendizagem de máquina.

Neste experimento é possível notar que os modelos de aprendizagem de máquinas são mais estáveis do que os modelos baseados em estatísticas descritivas, justificando assim o esforço computacional necessário para gerar os modelos. Ainda podemos notar que os diferentes algoritmos testados neste trabalho, tendem a obter resultados diferentes quando são alterados tanto o conjunto de treinamento quanto o

de teste.

Como pode ser observado nos experimentos apresentados nesse cenário e nos constantes no Anexo B dessa dissertação, é possível obter resultados interessantes tanto na predição de alunos em risco de evasão quanto em estudantes que tendem a terminar os cursos sem maiores problemas no que se refere a evasão.

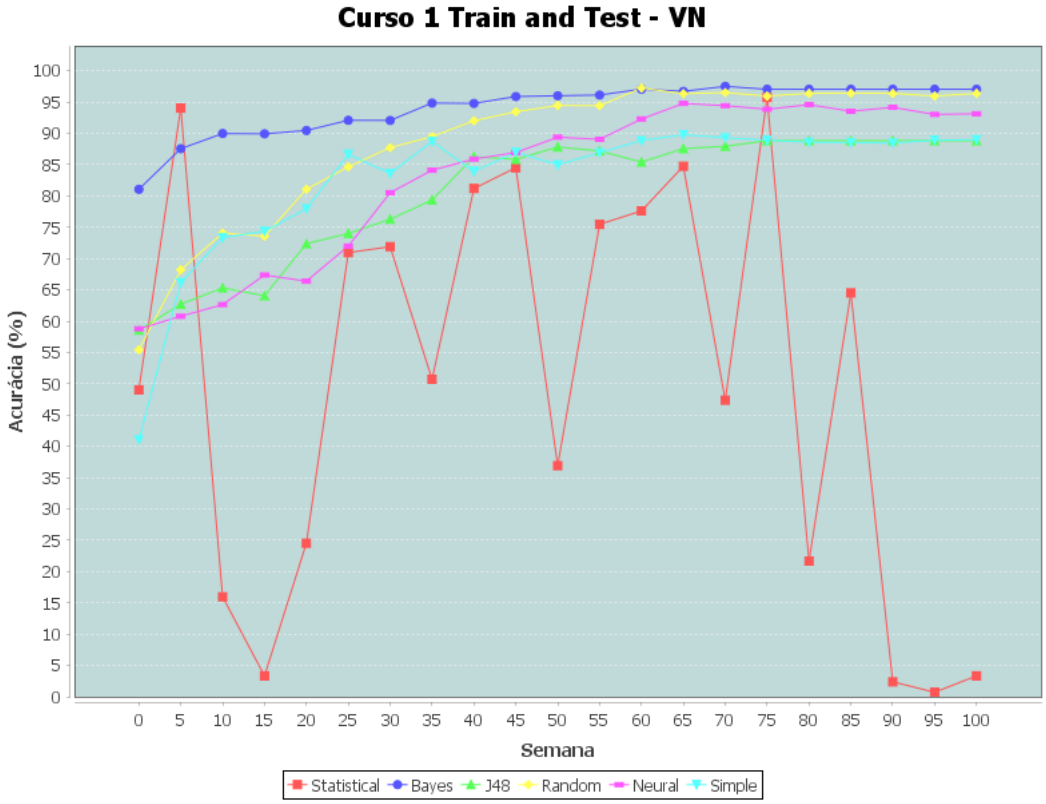

Figura 22: Resultado curso 1 - Verdadeiros Negativos Cenário 2

4.3.3 Árvores de Decisão

Nessão seção serão apresentadas algumas das árvores de decisão geradas pelo algoritmo J48. Estas árvores foram geradas com os dados até a semana do fim de cada semestre do curso 1, assim temos na Figura 23 a árvore da semana 25 do curso, na Figura 24 a árvore gerada na semana 50 e na Figura 25 a árvore da semana 100.

A geração e avaliação da árvores de decisão neste projeto tem como objetivo buscar identificar uma variável que se sobressaia e tenha um alto impacto na evasão. Sendo essa identificação possível, essa variável ou conjunto delas, pode fornecer de forma clara um auxílio na identificação de algum dos problemas que causa a evasão em seus cursos e auxíliar na tentativa de reversão do estado de evasão dos estudantes.

Na semana 25 apresenta na Figura 23 temos a árvore de decisão gerada no fim do primeiro semestre de curso, essa árvore apresenta como variável principal a média

da semana 25. Esta variável ainda se repete pelo menos mais duas vezes no decorrer da árvore, assim podemos entender que para esse tempo de curso está é a variável que apresenta o maior peso entre as que foram utilizadas.

As árvores geradas para entre as semanas 50 e 75 são iguais, nesse momento do curso a variável que apresenta um maior peso é o desvio padrão da semana 45. Outra variável chave nessas árvores é a média da semana 2 do curso, onde mesmo com a continuação do mesmo ela ainda apresenta um alto valor para a predição.

A árvore gerada para a semana 100 de curso apresenta como váriavel principal o desvio padrão da semana 58 de curso. Esta árvore foi a menor das geradas nesse experimento e a que levou em consideração um maior número de desvio semanais.

Após a análise das árvores geradas para o curso 1, não foi encontrado um padrão claro nas árvores analisadas a partir dos modelos gerados. Assim uma variável ou um conjunto destas, que se sobressaia e possa fornecer de forma clara uma forma de identificação do risco de evasão nos estudantes sem uma análise maior não foi identificada. Entretanto, foi possível notar como o aumento da granularidade impacta nas árvores de decisão, com diversas variáveis que foram geradas a partir desse processo sendo utilizadas.

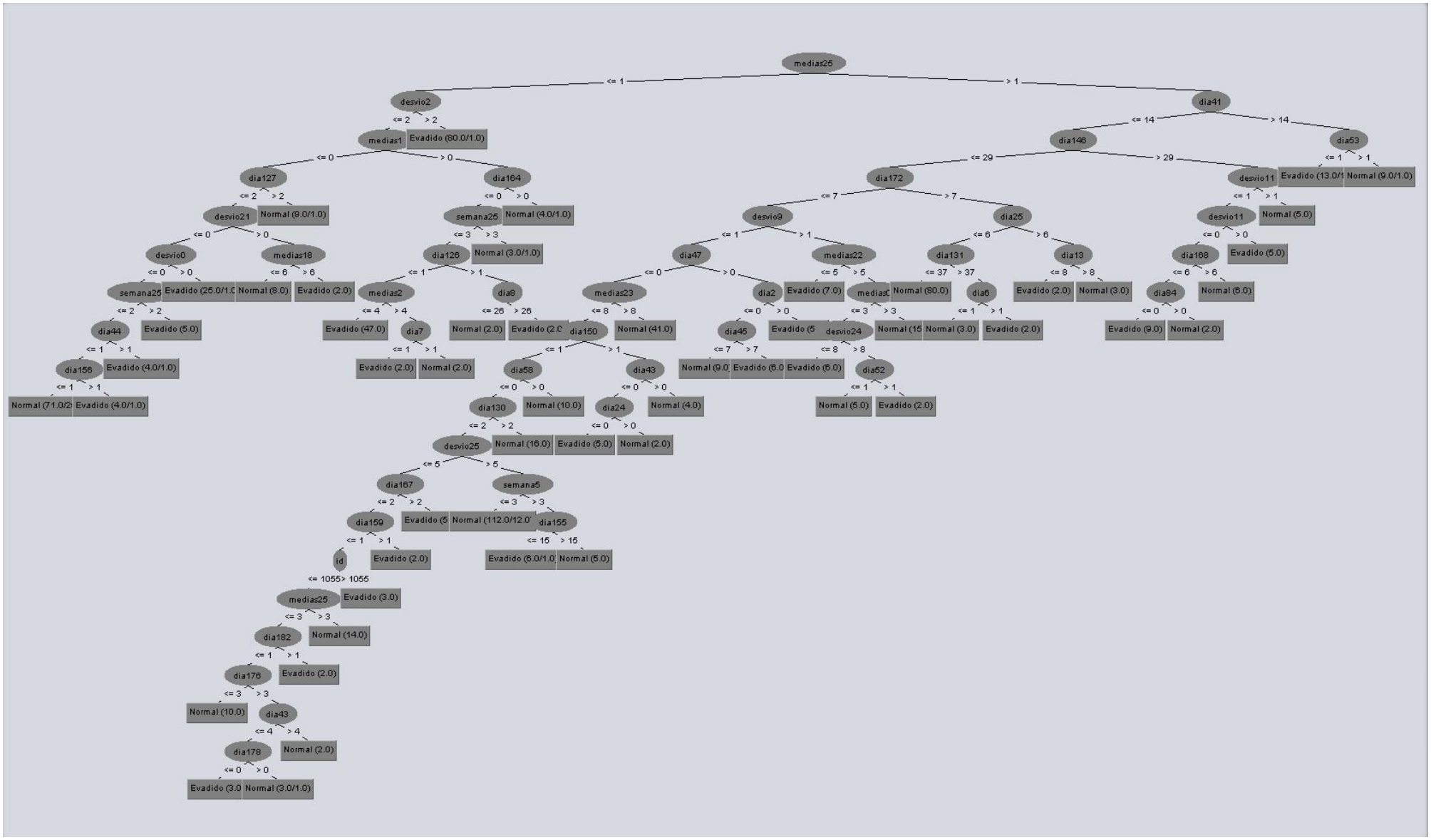

Figura 23: Árvore de decisão curso 1 - Semana 25

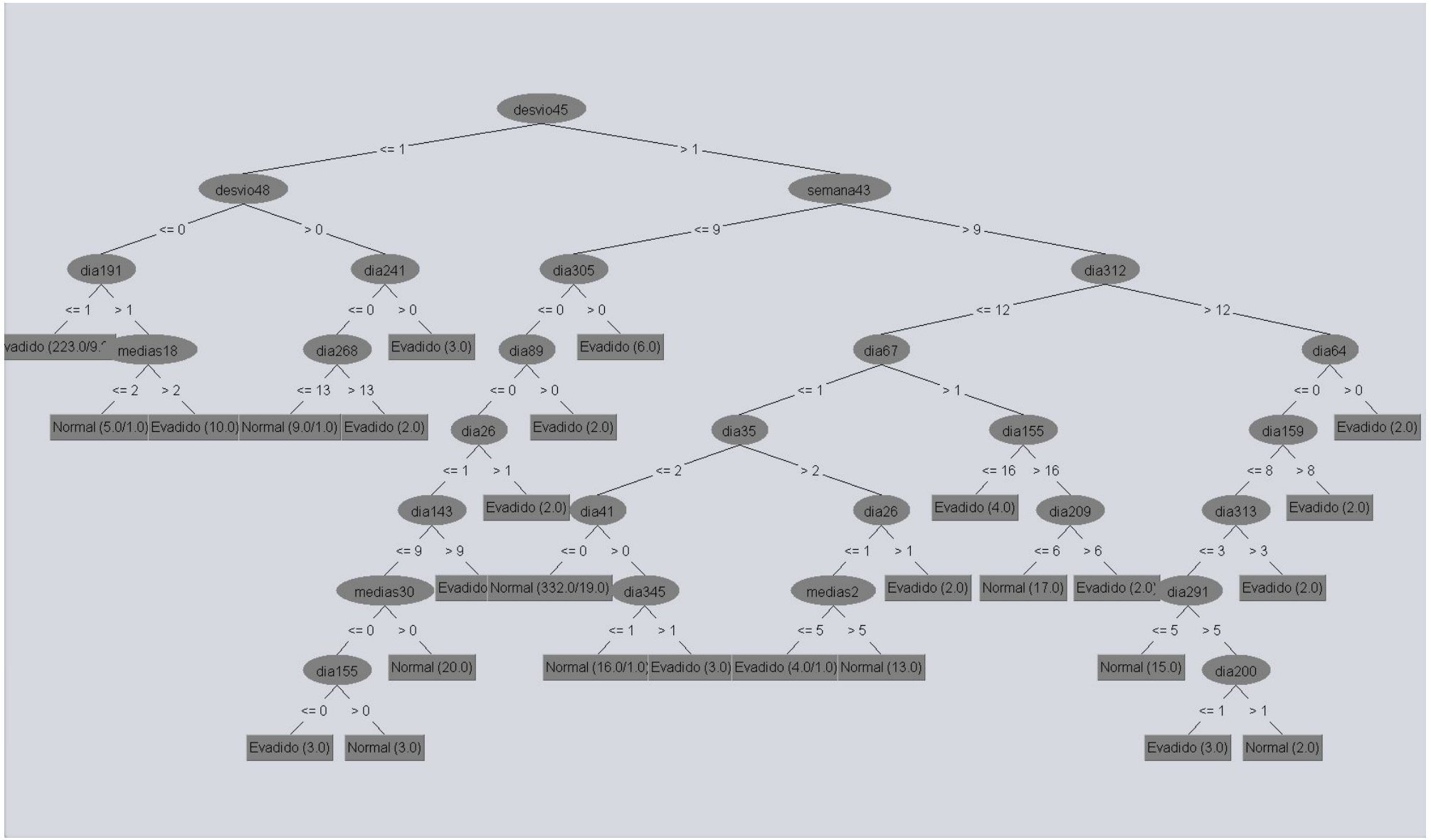

Figura 24: Árvore de decisão curso 1 - Semana 50

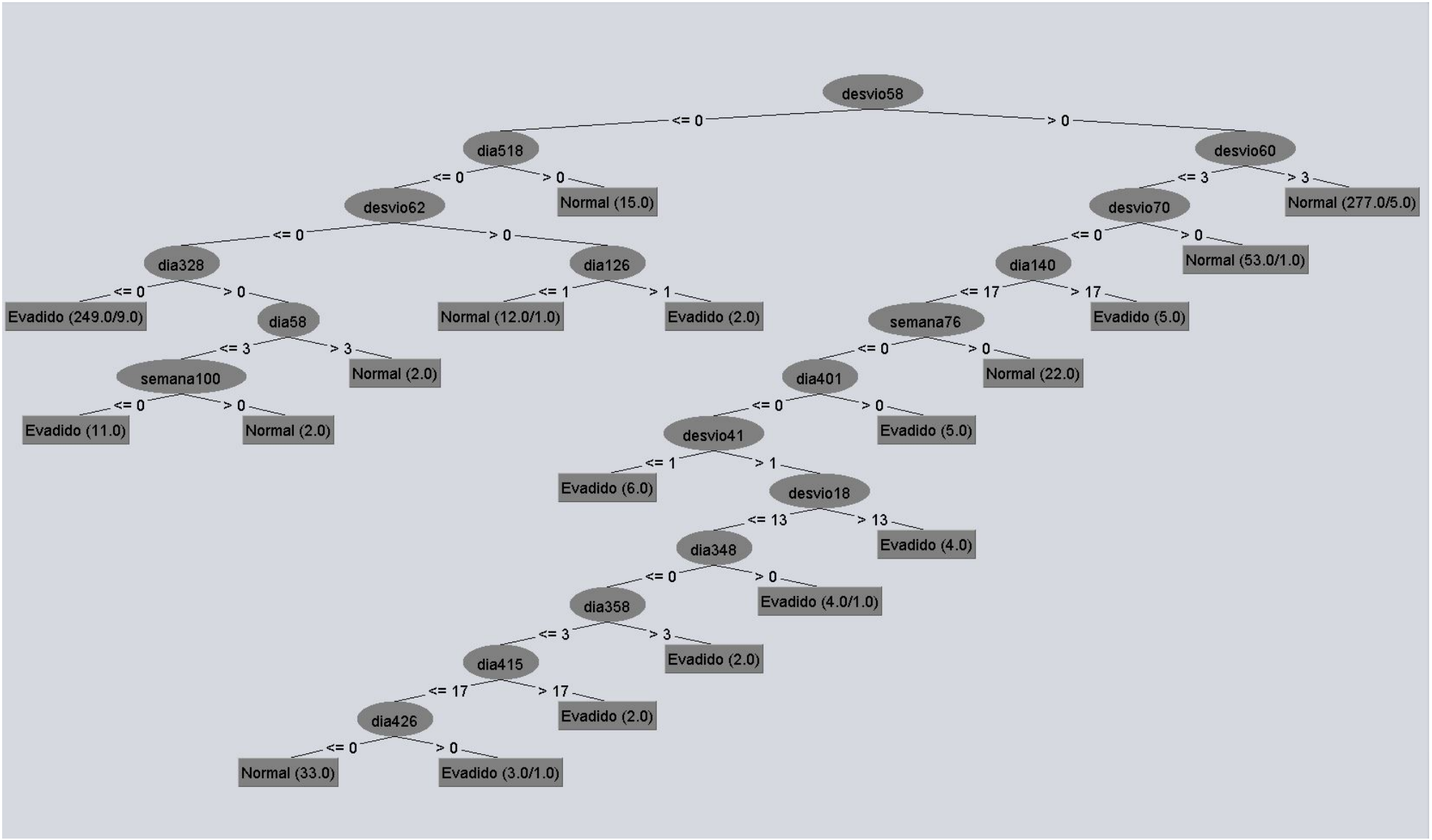

Figura 25: Árvore de decisão curso 1 - Semana 100

5 DISCUSSÃO DOS RESULTADOS

Analisando os diferentes resultados obtidos tanto de verdadeiros positivos quanto de verdadeiros negativos e tendo em vista que o objetivo do trabalho é a predição precoce de alunos em risco de evasão, podemos notar que o balanceamento dos dados é de suma importância para a aprendizagem dos algoritmos. Assim quanto mais balanceado os dados de treinamento forem, melhores os resultados obtidos nos testes tendem a ser.

As diferentes abordagens utilizadas desde os experimentos iniciais deste trabalho, demonstram que variáveis como quantidade de dados, balanceamento do conjunto de treinamento e granularidade dos dados, podem trazer um grande impacto nas taxas de acurácia.

Na Tabela 8, podemos ver a diferença entre os melhores resultados obtidos nos experimentos. Ficando assim nítido que a inserção de variáveis de granularidade, como o dia da interação, e a maior quantidade de dados envolvidos nas etapas de treinamento pode resultar em um diferença considerável nas etapas de predição.

Tabela 8: Comparativo entre Experimentos do Autor

Experimento	Semana 1		Semana 25		Semana 50		Semana 75		Semana 100	
	VP	VN	VP	VN	VP	VN	VP	VN	VP	VN
Experimento Inicial	58	68	82	81	93	93	94	94	97	97
Experimento Final Cenário 1	80	81	84	92	93	95	97	97	97	97
Experimento Final Cenário 2	95	97	100	100	100	100	100	100	100	100

Como é possível observar, desde as primeiras semanas de cursos são apresentados resultados satisfatórios na predição de estudantes em risco de evasão. No comparativo direto entre os experimentos deste trabalho, podemos notar o impacto que a quantidade de dados na etapa de geração dos modelos tem na predição.

No primeiro experimento são utilizados somente dados de um curso e contagem de interações semanais para a geração dos modelos, enquanto que no segundo são

utilizados dados de 3 cursos e além da contagem semanal a diária. Isso acaba por gerar modelos mais completos e que apresentam resultados significativamente maiores na predição.

5.1 Comparação com os Trabalhos Relacionados

Como é possível notar no capítulo 3 (Trabalhos Relacionados), diversas abordagens são aplicadas na predição de desempenho e evasão de alunos. Onde podem variar desde os dados utilizados, técnicas e até mesmo os tipos de cursos onde são extraídos os dados para os testes.

Entretanto podemos notar que as diversas técnicas aplicadas são de difícil generalização, o que acaba por dificultar a aplicação em outros cursos. Esta dificuldade pode-se dar pelos tipos de dados utilizados nem sempre estarem disponíveis em outros ambientes.Desta forma, seria uma tarefa extremamente complexa conseguir replicar os testes efetuados pelos autores em outros ambientes que não tem as mesmas características.

Na Tabela 9, é apresentado um comparativo entre os melhores resultados encontrados nos trabalhos citados anteriormente no capítulo 3 e o alcançado nesta dissertação.

A tarefa de comparação com os trabalhos relacionados é de difícil execução, tendo em vista que cada trabalho utiliza técnicas e dados diferentes, ademais, muitos deles trabalham com tipos de ensino e métricas para medição dos resultados diferentes. Desta forma, optou-se por normatizar os resultados na taxa de acerto de alunos em risco de evasão.

Como podemos notar na Tabela 9 se traçarmos um comparativo simples entre os resultados encontrados, os apresentados neste projeto demonstram a aplicabilidade da técnica proposta ficando entre os que obtiveram as melhores taxas.

Esta diferenciação nos resultados encontrados pode se dar por diversos fatores, entre os quais vale destacar a técnica utilizada, a quantidade de dados utilizada para geração dos modelos e a granularidade dos dados.

Na quantidade dos dados utilizados neste trabalho destaca-se que entre os demais trabalhos analisados que utilizam a contagem de interações, a maioria aplica suas técnicas a uma quantidade de dados que não passa da casa de 10 mil. No comparativo com esse trabalho, onde são utilizados os quantitativos de dados apresentados na Tabela 5, essas quantidades podem ser consideradas baixas.

Além disto a granularidade dos dados pode ser considerada menor, pois geralmente utilizam somente a contagem semanal ou mensal. Enquanto, este trabalho utiliza a contagem de interações diárias dos estudantes.

Esta granularidade é de suma importância na tarefa de predição como fica exposto

Tabela 9: Comparativo com os Resultados Obtidos pelo Estado da Arte

Autor	Tipo de Ensino	Melhores Resultados Verdadeiros Positivos (VP%)	Aplicação em Diferentes Contextos
JAYAPRAKASH et al. (2014)	Superior Presencial	84	Não
LYKOURENTZOU et al. (2009)	Especialização	94	Não
MANHãES et al. (2011)	Superior Presencial	80	Sim
CAMBRUZZI; RIGO; BARBOSA (2015)	Superior Presencial e a Distância	87	Não
YUKSELTURK (2014)	Especialização a Distância	87	Não
DEKKER; PECHENIZKIY (2009)	Superior Presencial	80	Não
ROMERO; VENTURA (2013)	Superior Presencial	65	Sim
COHEN (2017)	Superior Presencial	66	Não
BURGOS et al. (2017)	Cursos de Pequena Duração	100	Sim
KANTORSKI et al. (2016)	Superior Presencial	73	Não
QUEIROGA	Técnico a Distância	100	Sim

no comparativo entre os experimentos deste trabalho, onde no primeiro utilizamos somente a contagem de interações semanais e no outro onde já é utilizada a contagem de interações diária. No primeiro os resultados apresentados ficam muito próximos aos dos outros trabalhos na área.

Outro fator que pode ser impactante na predição é o tempo de curso, muitos dos trabalhos na área utilizam cursos de baixa duração (às vezes até mesmo 8 semanas) o que pode facilitar a predição da evasão.

Assim apesar dos resultados obtidos, a maioria das pesquisas acabam por se restringir a um determinado ambiente e geralmente com uma quantidade de dados relativamente pequena. Em contra partida a proposta do autor deste trabalho, apresenta

um ambiente com a utilização de uma maior granularidade e quantitativos de dados utilizados e isso se replica em modelos mais precisos.

Desta forma, entendemos que a geração de modelos de fácil generalização levando em conta somente dados das interações dos alunos é aplicável e pode apresentar resultados próximos ou até mesmo maiores que as outras técnicas propostas. Entretanto, ela pode se tornar vantajosa, pois os dados utilizados estão disponíveis na maioria das instituições de ensino que utilizem os ambientes virtuais de aprendizagem.

6 CONSIDERAÇÕES FINAIS

Podemos notar que diversas abordagens são aplicadas na predição do risco de evasão de estudantes, com variação dos dados utilizados, técnicas e até mesmo os tipos de cursos de onde os mesmos são extraídos para os testes.

Essas diferenças acabam por dificultar a aplicação das diversas técnicas em outros cursos. Desta forma, seria uma tarefa extremamente complexa conseguir replicar os testes efetuados pelos autores em outros ambientes que não têm as mesmas características.

Um dos aspectos que chama a atenção nas pesquisas relacionadas avaliadas é que apesar de existirem muitos trabalhos sobre a evasão, poucos são voltados para o ensino técnico. Assim, esse trabalho busca auxiliar essa etapa importante no ensino e, consequentemente, na formação de mão de obra.

A proposta principal desse trabalho que é a geração e o teste de modelos de predição para identificação de estudantes de cursos técnicos a distância em risco de evasão, utilizando somente contagem de interações e diferentes variações desta, apresentou resultados satisfatórios na comparação aos encontrados por outras pesquisas na área.

Como acreditamos que os modelos gerados nesse projeto sejam de mais fácil generalização e aplicação a outros contextos, seria interessante a aplicação dos mesmos em dados de outras instituições de ensino. Fica essa aplicação como um trabalho futuro.

Em cada um dos cenários testados a variação entre o algoritmo que obteve os melhores resultados é grande. Assim a aplicação de uma metodologia de votação que utilize a combinação dos algoritmos por exemplo, poderia melhorar os resultados ainda mais. Além disto, a criação de um modulo de integração direta do software com a base de dados do Moodle ou de um plugin para o próprio Moodle pode facilitar a utilização dos modelos gerados nessa dissertação. Desta forma, ficando estas duas tarefas como possíveis trabalhos futuros.

O objetivo principal deste trabalho foi estudar e aplicar as técnicas de mineração de dados e aprendizagem de máquina em dados disponíveis da EAD do Instituto Federal

Sul-rio-grandense (IFSUL), propondo um modelo de predição para evasão de alunos baseado somente contagem de interações e suas variações, assim possibilitando o emprego em diferentes domínios de aplicação. Concluímos que os resultados obtidos comprovam a aplicabilidade da técnica proposta.

REFERÊNCIAS

AGRESTI, A.; KATERI, M. **Categorical data analysis**. [S.l.]: Springer, 2011.

ARGOTE, L. **Organizational Learning**: Creating, Retaining, and Transferring Knowledge. 1st.ed. Norwell, MA, USA: Kluwer Academic Publishers, 1999.

BAKER, R.; ISOTANI, S.; CARVALHO, A. Mineração de Dados Educacionais: Oportunidades para o Brasil. **Revista Brasileira de Informática na Educação**, [S.l.], v.19, n.02, p.03, 2011.

BAKER, R. S. J. D.; YACEF, K. The State of Educational Data Mining in 2009 : A Review and Future Visions. **Journal of Educational Data Mining**, [S.l.], v.1, n.1, p.3–16, 2009.

BARROSO, M. F.; FALCãO, E. B. Evasão universitária: O caso do Instituto de Física da UFRJ. **Encontro Nacional de Pesquisa em Ensino de Física**, [S.l.], v.9, p.1–14, 2004.

BEVITT, D.; BALDWIN, C.; CALVERT, J.; BEVITT, D.; BALDWIN, C.; CALVERT, J. Intervening Early : Attendance and Performance Monitoring as a Trigger for First Year Support in the Biosciences Trigger for First Year Support in the Biosciences. **Bioscience Education**, [S.l.], v.7860, n.December, 2015.

BOUCKAERT, R. R.; FRANK, E.; HALL, M.; KIRKBY, R.; REUTEMANN, P.; SEEWALD, A.; SCUSE, D. WEKA manual for version 3-7-3. **The university of WAIKATO**, [S.l.], 2010.

BOYER, S.; VEERAMACHANENI, K. Transfer Learning for Predictive Models in Massive Open Online Courses. **Artificial Intelligence**, [S.l.], p.1–12, 2015.

BRAGA, L. **Introdução à Mineração de Dados - 2a edição**: Edição ampliada e revisada. [S.l.]: e-papers, 2005.

BRAMER, M. **Undergraduate Topics in Computer Science. Principles of Data Mining**. [S.l.]: Springer, London, 2013.

BRASIL, C. d. N.; NATUREZA, C. da. Matemática e suas Tecnologias. **PCN+ Ensino Médio: orientações educacionais complementares aos Parâmetros Curriculares Nacionais. Brasília**, [S.l.], 2002.

BRASIL, M. **Censo Escolar. 2012, INEP, Brasília, INEP, 2012**.

BREIMAN, L. Random forests. **Machine learning**, [S.l.], v.45, n.1, p.5–32, 2001.

BURGOS, C.; CAMPANARIO, M. L.; PEÑA, D. de la; LARA, J. A.; LIZCANO, D.; MARTÍNEZ, M. A. Data mining for modeling students' performance: A tutoring action plan to prevent academic dropout. **Computers & Electrical Engineering**, [S.l.], v.0, p.1–16, 2017.

CAMBRUZZI, W.; RIGO, S. J.; BARBOSA, J. L. V. Dropout Prediction and Reduction in Distance Education Courses with the Learning Analytics Multitrail Approach. **Journal of Universal Computer Science**, [S.l.], v.21, n.1, p.23–47, 2015.

CAMILO, C. O.; SILVA, J. C. d. Mineração de dados: Conceitos, tarefas, métodos e ferramentas. **Universidade Federal de Goiás (UFC)**, [S.l.], p.1–29, 2009.

CARVALHO, A. P. d. L. F. Algoritmos genéticos. **Instituto de Ciências**, [S.l.], 2009.

CENSO, E. BR 2013-Relatório Analítico da Aprendizagem a Distância no Brasil. **Acesso em**, [S.l.], v.16, n.08, 2015.

CHARNIAK, E. Bayesian networks without tears. **AI magazine**, [S.l.], v.12, n.4, p.50, 1991.

CIOS, K. J.; PEDRYCZ, W.; SWINIARSKI, R. W. Data mining and knowledge discovery. In: **Data Mining Methods for Knowledge Discovery**. [S.l.]: Springer, 1998. p.1–26.

COHEN, A. Analysis of student activity in web-supported courses as a tool for predicting dropout. **Educational Technology Research and Development**, [S.l.], 2017.

DAR-EL, E. **HUMAN LEARNING**: From Learning Curves to Learning Organizations. [S.l.]: Springer US, 2013. (International Series in Operations Research & Management Science).

DE OLIVEIRA, G. B. Uma discussão sobre o conceito de desenvolvimento. **Revista da FAE**, [S.l.], v.5, n.2, 2017.

DEKKER, G.; PECHENIZKIY, M. Predicting students drop out: A case study. **EDM'09 - Educational Data Mining 2009: 2nd International Conference on Educational Data Mining**, [S.l.], p.41–50, 2009.

DELANO, R.; CORRêA, D. S. Redes na Educação a Distância: Uma Análise Estrutural do Sistema UAB em Minas Gerais. **Revista PRETEXTO.**, [S.l.], 2013.

DETONI, D.; ARAUJO, R. M.; CECHINEL, C. Predição de Reprovação de Alunos de Educação a Distância Utilizando Contagem de Interações. In: SIMPóSIO BRASILEIRO DE INFORMáTICA NA EDUCAçãO, 2014. **Anais...** [S.l.: s.n.], 2014. v.25, n.1, p.896–905.

DUNKEL, B.; SOPARKAR, N.; SZARO, J.; UTHURUSAMY, R. Systems for KDD: From concepts to practice. **Future Generation Computer Systems**, [S.l.], v.13, n.2-3, p.231–242, 1997.

EYNG, A. M.; GISI, M.; ENS, R.; PACIEVITCH, T. Diversidade e padronização nas políticas educacionais: configurações da convivência escolar. **Ensaio: Avaliação e Políticas Públicas em Educação**, [S.l.], v.21, n.81, p.773–800, 2013.

FARIA, E. T. O professor e as novas tecnologias. **Ser professor**, [S.l.], v.5, 2004.

FAYYAD, U.; PIATETSKY-SHAPIRO, G.; SMYTH, P. From data mining to knowledge discovery in databases. **AI magazine**, [S.l.], v.17, n.3, p.37, 1996.

FERNANDEZ, G. **Data mining using SAS applications**. [S.l.]: CRC press, 2010.

GUPTA, G. **Introduction to data mining with case studies**. [S.l.]: PHI Learning Pvt. Ltd., 2014.

HALAWA, S.; GREENE, D.; MITCHELL, J. Dropout Prediction in MOOCs using Learner Activity Features. **European MOOC Summit, EMOOCs**, [S.l.], v.37, n.March, p.1–10, 2014.

HALL, M.; FRANK, E.; HOLMES, G.; PFAHRINGER, B.; REUTEMANN, P.; WITTEN, I. H. The WEKA data mining software: an update. **ACM SIGKDD explorations newsletter**, [S.l.], v.11, n.1, p.10–18, 2009.

HAN, J.; PEI, J.; KAMBER, M. **Data mining**: concepts and techniques. [S.l.]: Elsevier, 2011.

HASTIE, T.; TIBSHIRANI, R.; FRIEDMAN, J. **Unsupervised learning**. [S.l.]: Springer, 2009.

HO, T. K. Random decision forests. In: DOCUMENT ANALYSIS AND RECOGNITION, 1995., PROCEEDINGS OF THE THIRD INTERNATIONAL CONFERENCE ON, 1995. **Anais...** [S.l.: s.n.], 1995. v.1, p.278–282.

JAYAPRAKASH, S. M.; MOODY, E. W.; LAURIA, E. J. M.; REGAN, J. R.; BARON, J. D. Early Alert of Academically At-Risk Students: An Open Source Analytics Initiative. **Journal of Learning Analytics**, [S.l.], v.1, n.1, p.6–47, 2014.

JOHNSON, R. A.; WICHERN, D. W. et al. **Applied multivariate statistical analysis**. [S.l.]: Prentice hall Upper Saddle River, NJ, 2002. v.5, n.8.

JUNIOR, L. C. D. M. **A Autoeducação e o Século 21**. [S.l.]: Litteris, 2013.

KANTORSKI, G.; FLORES, E. G.; SCHMITT, J.; HOFFMANN, I.; BARBOSA, F. Predição da Evasão em Cursos de Graduação em Instituições Públicas. **Brazilian Symposium on Computers in Education (Simpósio Brasileiro de Informática na Educação - SBIE)**, [S.l.], v.27, n.1, p.906, 2016.

LITTO, F. M.; FORMIGA, M. **Educação a distância**: o estado da arte. [S.l.]: Pearson, 2011.

LYKOURENTZOU, I.; GIANNOUKOS, I.; NIKOLOPOULOS, V.; MPARDIS, G.; LOUMOS, V. Dropout prediction in e-learning courses through the combination of machine learning techniques. **Computers & Education**, [S.l.], v.53, n.3, p.950–965, 2009.

MAIA, C.; MATTAR, J. **ABC da EaD**: a educação a distância hoje. [S.l.]: Pearson Prentice Hall, 2008.

MANHãES, L. M. B.; CRUZ, S. d.; COSTA, R. J. M.; ZAVALETA, J.; ZIMBRãO, G. Previsão de Estudantes com Risco de Evasão Utilizando Técnicas de Mineração de Dados. **Anais do XXII SBIE-XVII WIE, Aracaju**, [S.l.], 2011.

MARQUES, R. L.; DUTRA, I. Redes Bayesianas: o que são, para que servem, algoritmos e exemplos de aplicações. **Coppe Sistemas–Universidade Federal do Rio de Janeiro, Rio de Janeiro, Brasil**, [S.l.], 2002.

MARTINS, A. C.; MARQUES, J. M.; COSTA, P. D. Estudo comparativo de três algoritmos de machine learning na classificação de dados electrocardiográficos. **Trabalho (Mestrado em Informática Médica)–Universidade do Porto, Porto**, [S.l.], 2009.

MCCULLOCH, W. S.; PITTS, W. A logical calculus of the ideas immanent in nervous activity. **The bulletin of mathematical biophysics**, [S.l.], v.5, n.4, p.115–133, 1943.

MICHALSKI, R. S.; CARBONELL, J. G.; MITCHELL, T. M. **Machine learning**: An artificial intelligence approach. [S.l.]: Springer Science & Business Media, 2013.

MONARD, M. C.; BARANAUSKAS, J. A. Conceitos sobre Aprendizado de Máquina. **Sistemas inteligentes: fundamentos e aplicações**, [S.l.], p.39–56, 2003.

OLSON, D. L.; DELEN, D. **Advanced data mining techniques**. [S.l.]: Springer Science & Business Media, 2008.

PASTA, A. **plicação da técnica de Data Mining na base de dados do ambiente de gestão educacional**: Um estudo de caso de uma instituição de ensino superior de Blumenau-SC. [S.l.]: Universidade do Vale do Itajaí, São José, 2011.

PICHILIANI, M. Data mining na prática: Árvores de Decisão. **Disponıvel em: http://imasters. com. br/artigo/5130/sql-server/data-mining-na-pratica-arvores-de-decisao**, [S.l.], 2008.

PRASS, F. S. **Estudo comparativo entre algoritmos de análise de agrupamentos em data mining. 2004. 71 f**. 2004. Tese (Doutorado em Ciência da Computação) — Dissertação (Mestrado em Mestrado em Ciência da Computação)-Área de Concentração de Sistemas de Computação, Universidade Federal de Santa Catarina, Florianópolis.

QUEIROGA, E.; CECHINEL, C.; ARAÚJO, R. Um Estudo do Uso de Contagem de Interações Semanais para Predição Precoce de Evasão em Educação a Distância. In: WORKSHOPS DO CONGRESSO BRASILEIRO DE INFORMÁTICA NA EDUCAÇÃO, 2015. **Anais...** [S.l.: s.n.], 2015. v.4, n.1, p.1074.

QUEIROGA, E.; CECHINEL, C.; ARAÚJO, R.; COSTA BRETANHA, G. da. Generating models to predict at-risk students in technical e-learning courses. In: LEARNING OBJECTS AND TECHNOLOGY (LACLO), LATIN AMERICAN CONFERENCE ON, 2016. **Anais...** [S.l.: s.n.], 2016. p.1–8.

QUEIROZ, L. D. Um estudo sobre a evasão escolar: para se pensar na inclusão escolar. **Associação Nacional de Pós-Graduação e Pesquisa em Educação (Anpad). Disponível em www. anped. org. br/reunioes/25/lucileidedomingosqueirozt13. rtf. Acesso em**, [S.l.], v.3, 2001.

QUINLAN, J. R. Induction of decision trees. **Machine learning**, [S.l.], v.1, n.1, p.81–106, 1986.

QUINLAN, J. R. **C4. 5**: programs for machine learning. [S.l.]: Elsevier, 1993.

REPICI, J. The comma separated value (csv) file format. **Creativyst Inc**, [S.l.], 2010.

ROMERO, C.; VENTURA, S.; ESPEJO, P. G.; HERVÁS, C. Data mining algorithms to classify students. **Educational Data Mining 2008 - 1st International Conference on Educational Data Mining, Proceedings**, [S.l.], p.8–17, 2008.

ROMERO; VENTURA. Data mining in education. **Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery**, [S.l.], v.3, n.1, p.12–27, 2013.

RUMELHART, D. E.; HINTON, G. E.; WILLIAMS, R. J. **Learning internal representations by error propagation**. [S.l.]: DTIC Document, 1985.

RUSSELL, S.; NORVIG, P. **Artificial Intelligence**: A Modern Approach. [S.l.]: Prentice Hall, 2010. (Prentice Hall series in artificial intelligence).

SCHMITT, J. **Pré-processamento para a mineração de dados**: uso da análise de componentes principais com escalonamento ótimo. 2005. Tese (Doutorado em Ciência da Computação) — Dissertação (Mestrado em Mestrado em Ciência da Computação)-Área de Concentração de Sistemas de Computação, Universidade Federal de Santa Catarina, Florianópolis.

SEGATTO, Ê. C.; COURY, D. V. Redes neurais aplicadas a relés diferenciais para transformadores de potência. **Sba: Controle & Automação Sociedade Brasileira de Automatica**, [S.l.], v.19, n.1, p.93–106, 2008.

SEGUNDO, F. R.; RAMOS, D. K. Soluções baseadas no uso de software livre: alternativas de suporte tecnológico à educação presencial e a distância. **Anais do 12 Congresso Internacional de Educação a Distância**, [S.l.], v.12, p.18–22, 2005.

VASCONCELOS, L. M. R. de; CARVALHO, C. L. de. Aplicação de regras de associação para mineração de dados na web. **Instituto de Informática da Universidade Federal de Goiás**, [S.l.], 2004.

WITTEN, I. H.; FRANK, E.; HALL, M. A.; PAL, C. J. **Data Mining**: Practical machine learning tools and techniques. [S.l.]: Morgan Kaufmann, 2016.

YADAV, S. K.; PAL, S. Data Mining : A Prediction for Performance Improvement of Engineering Students using Classification. **World of Computer Science and Information Technology Journal WCSIT**, [S.l.], v.2, n.2, p.51–56, 2012.

YUKSELTURK, E. Predicting Dropout Student : an Application of Data Mining Methods in an Online Education Program. **Computers & Education**, [S.l.], v.17, n.1, p.118–133, 2014.

ZHANG, H. The optimality of naive Bayes. **AA**, [S.l.], v.1, n.2, p.3, 2004.

ANEXO A RESULTADOS EXPERIMENTO FINAL

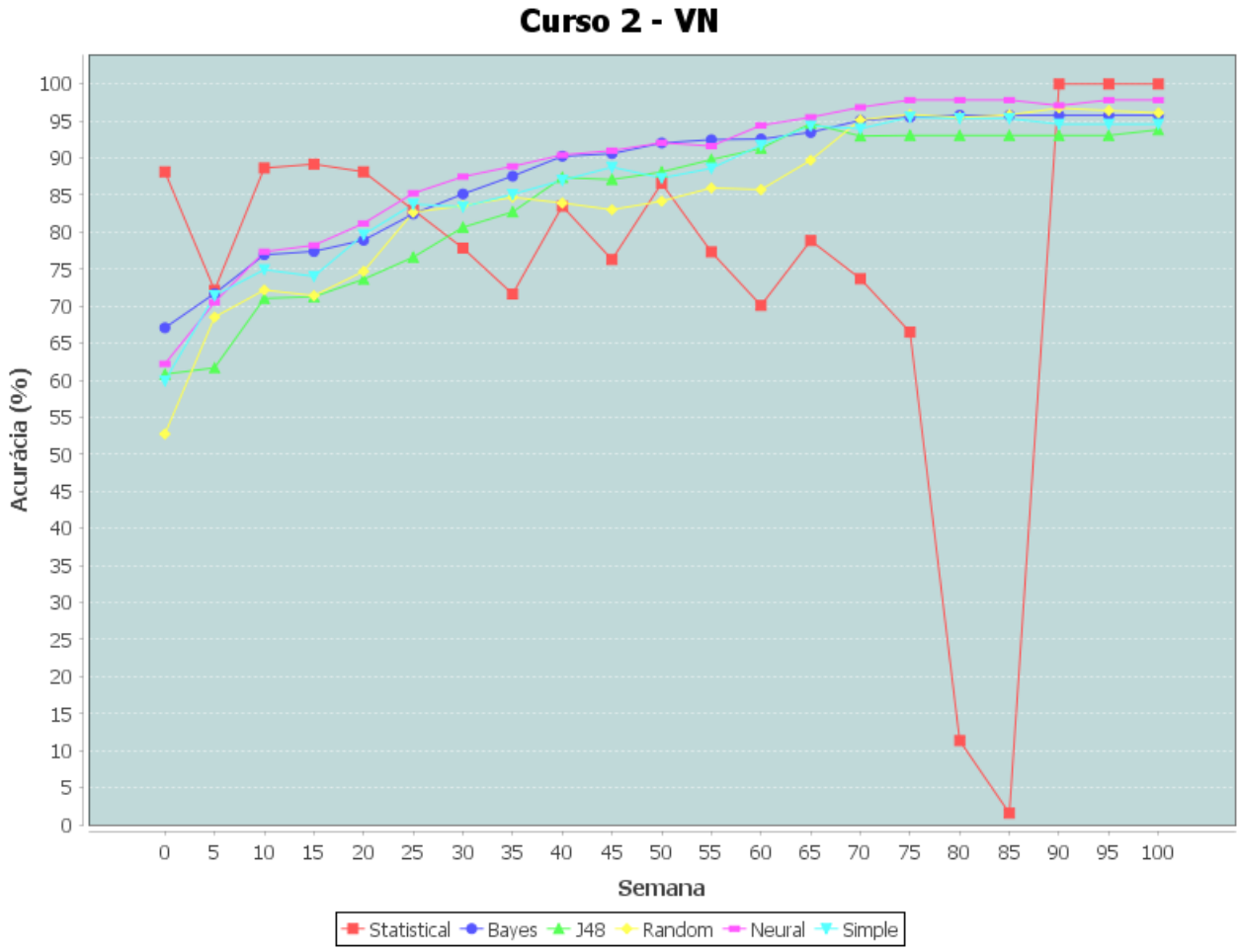

Figura 26: Resultado Cenário 1 Experimento final Curso 2 VN

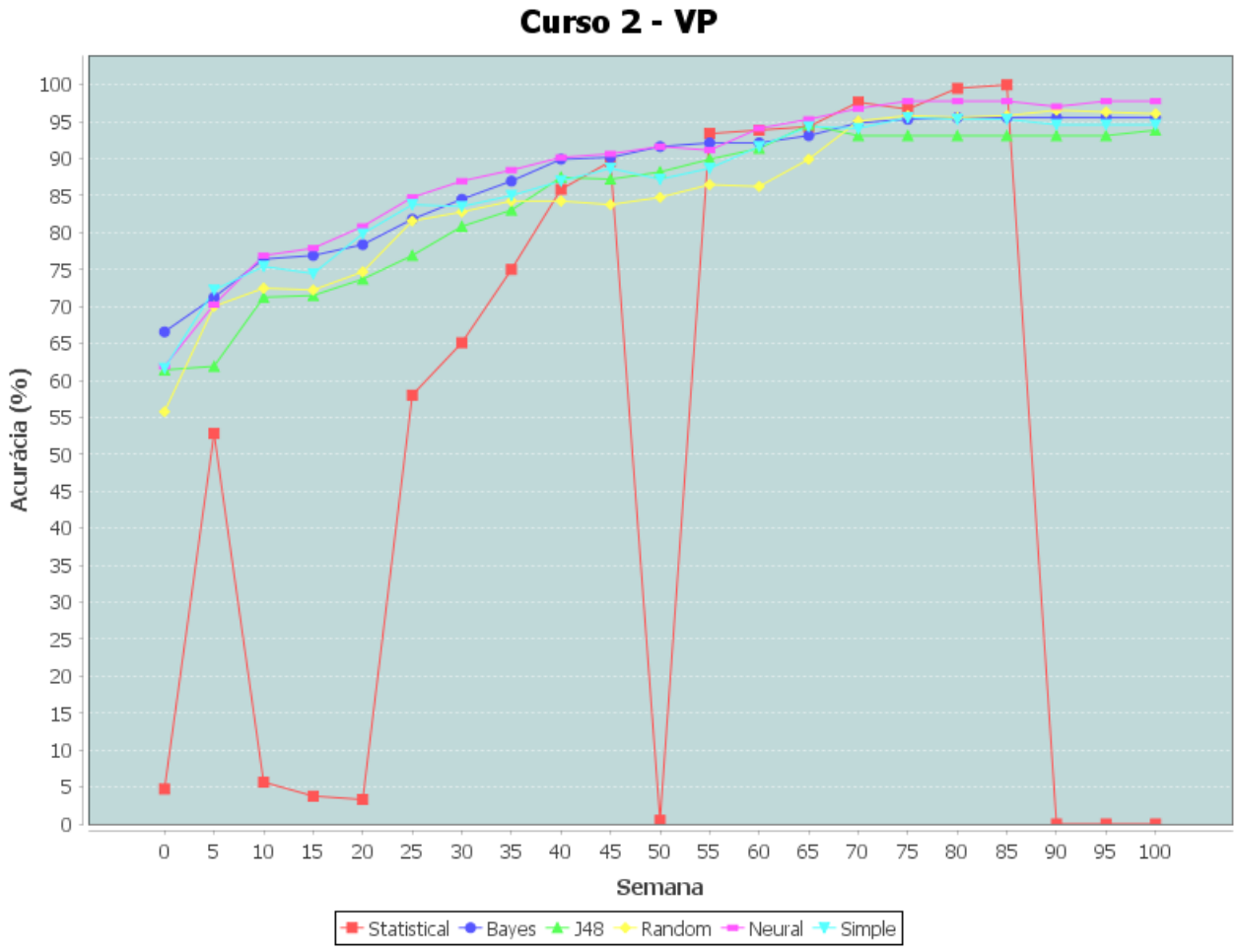

Figura 27: Resultado Cenário 1 Experimento final Curso 2 VP

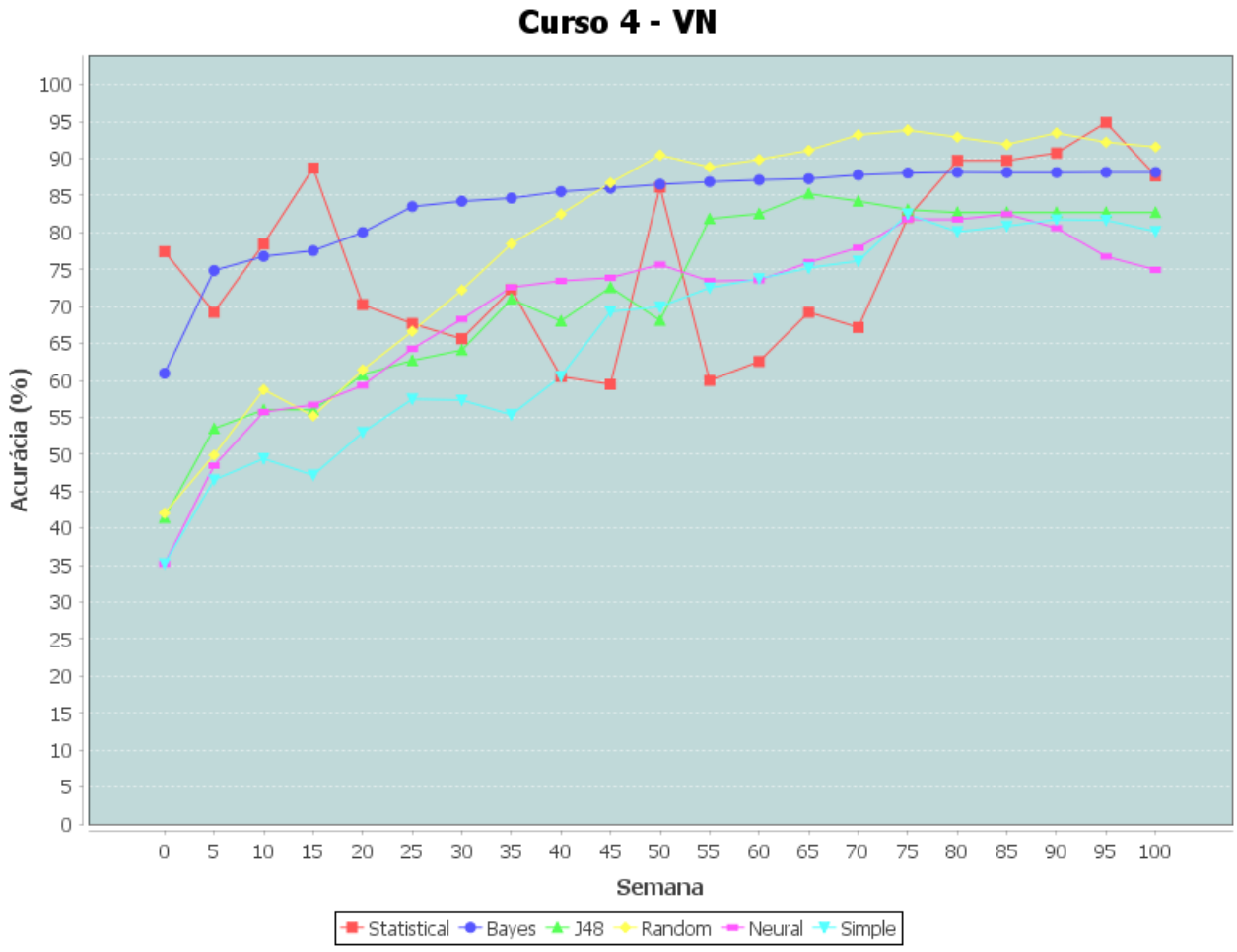

Figura 28: Resultado Cenário 1 Experimento final Curso 4 VN

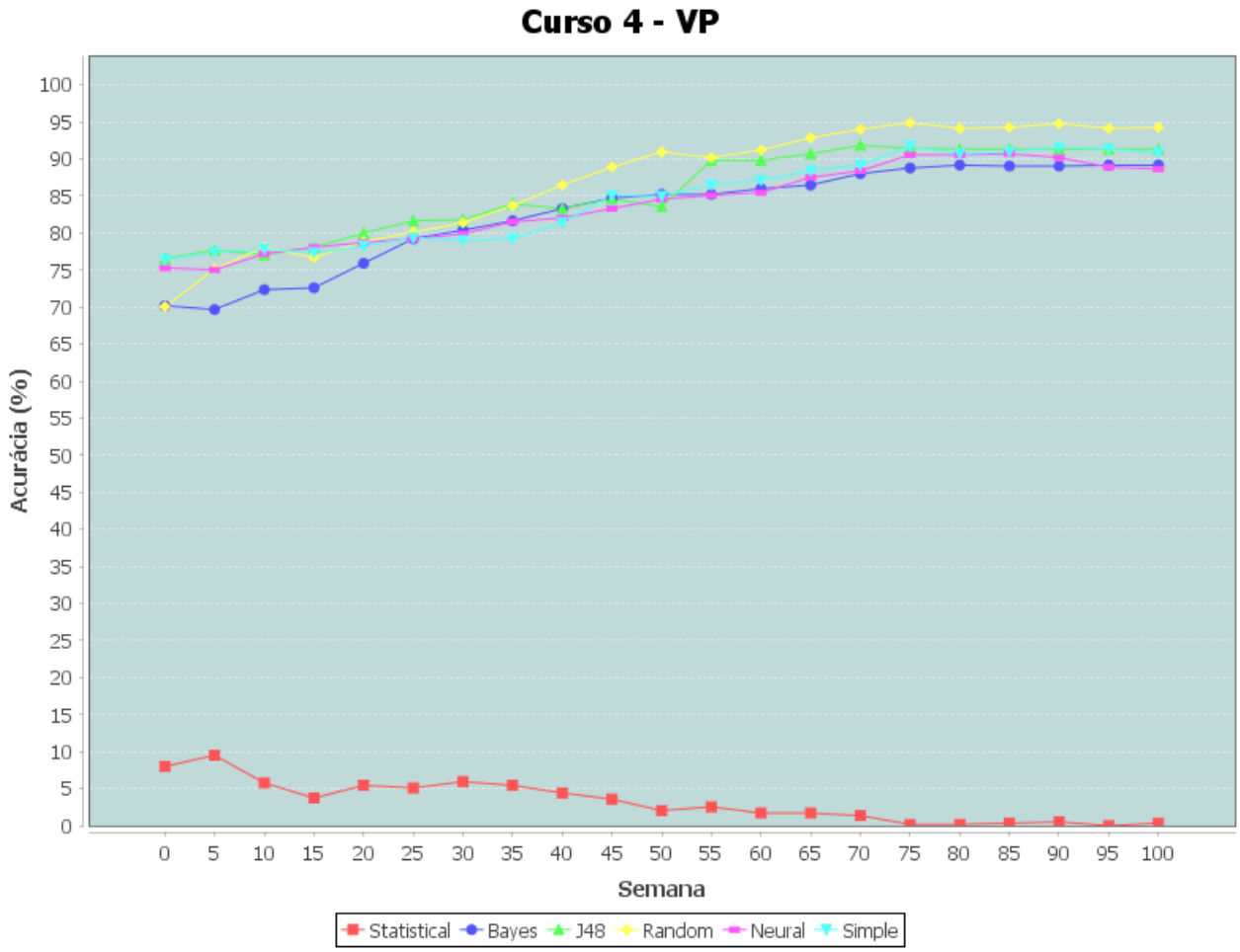

Figura 29: Resultado Cenário 1 Experimento final Curso 4 VP

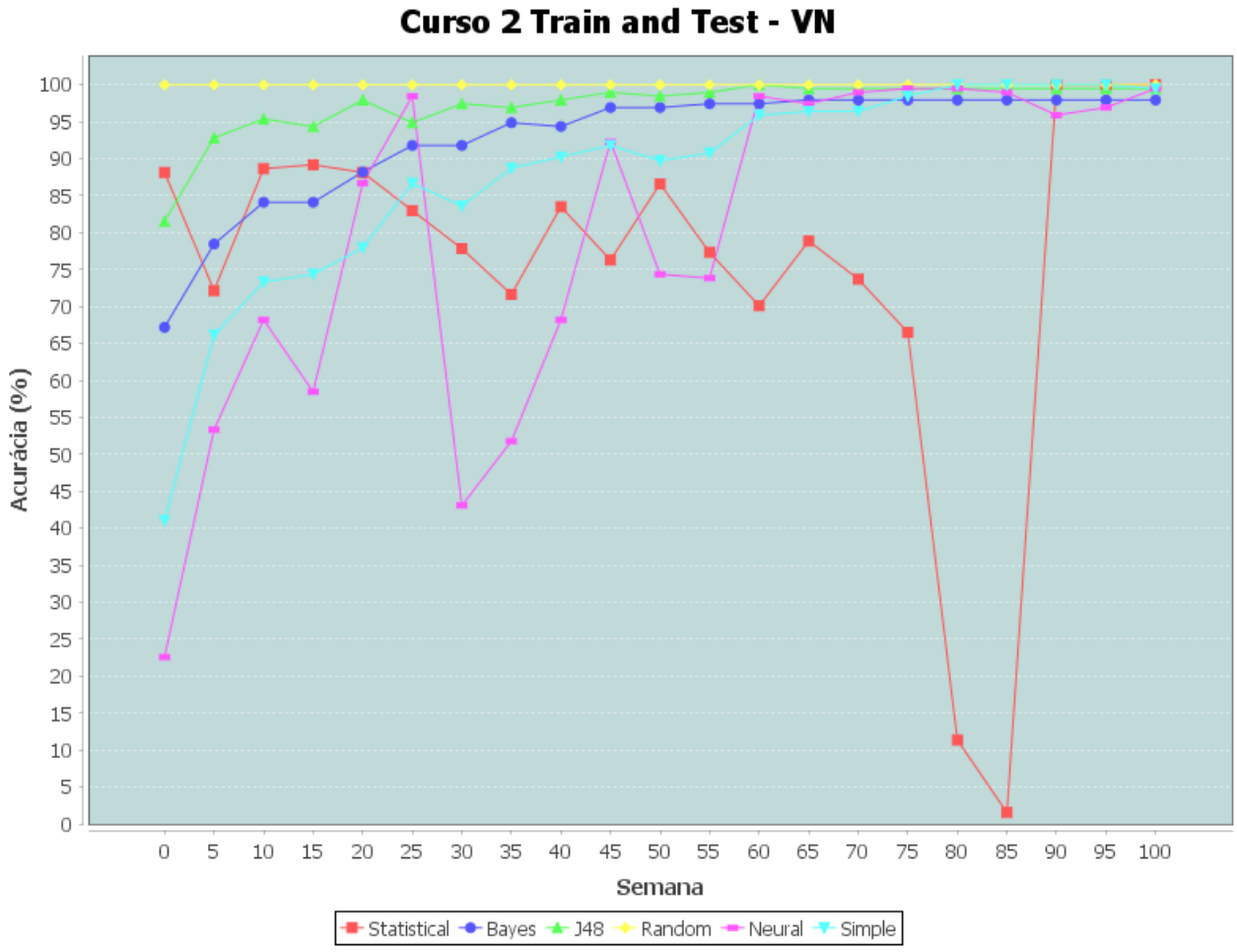

Figura 30: Resultado Cenário 2 Experimento final Curso 2 VN

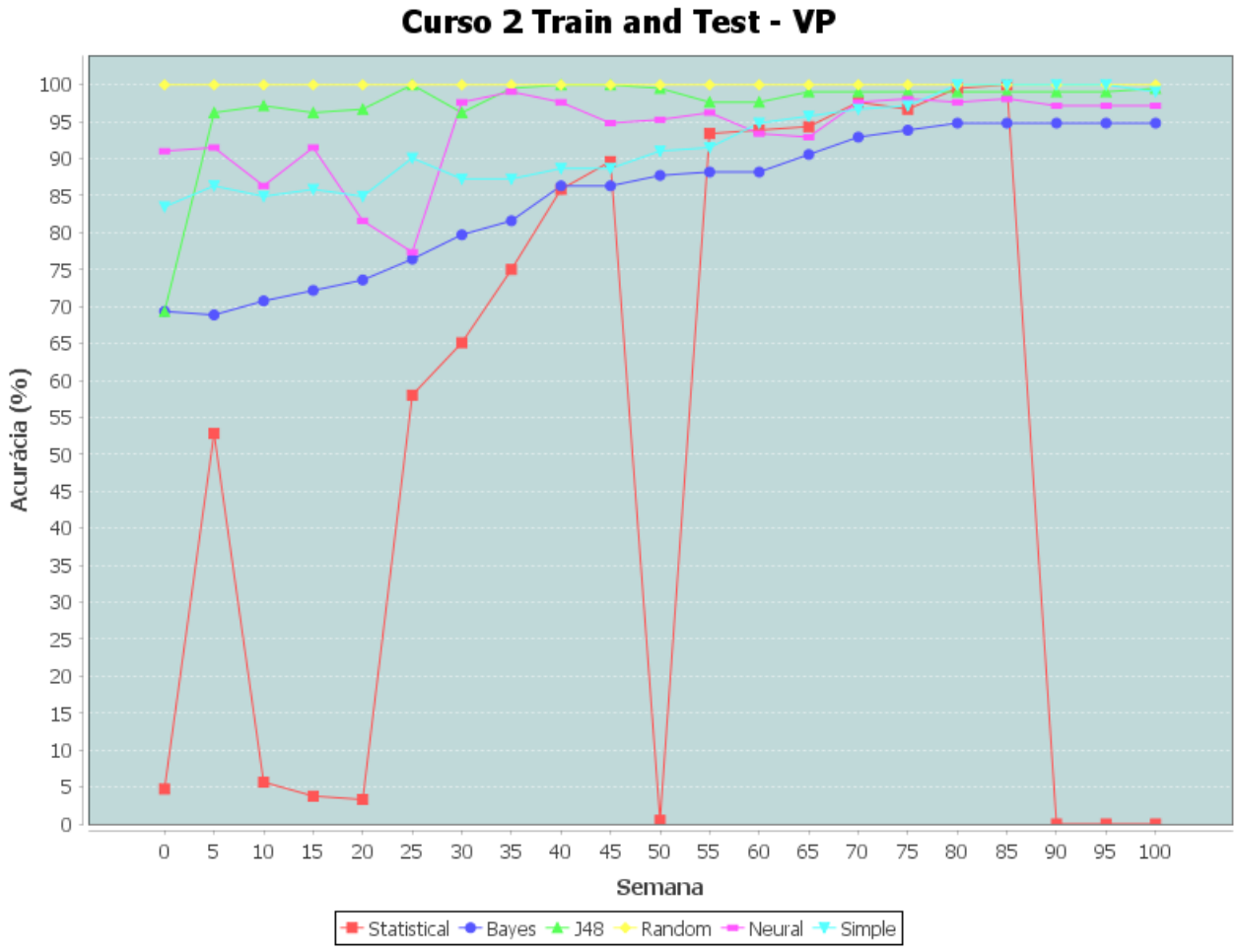

Figura 31: Resultado Cenário 2 Experimento final Curso 2 VP

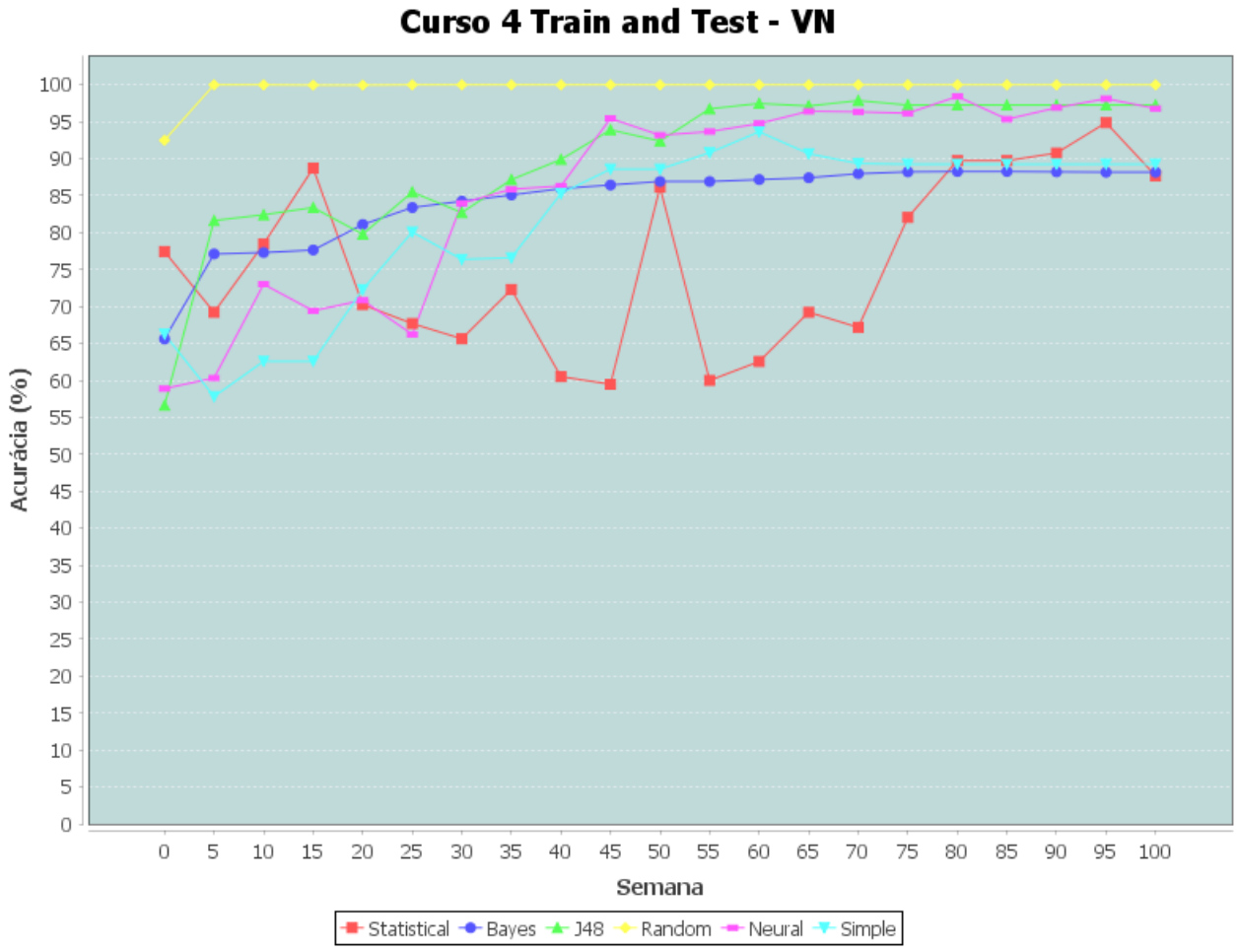

Figura 32: Resultado Cenário 2 Experimento final Curso 4 VN

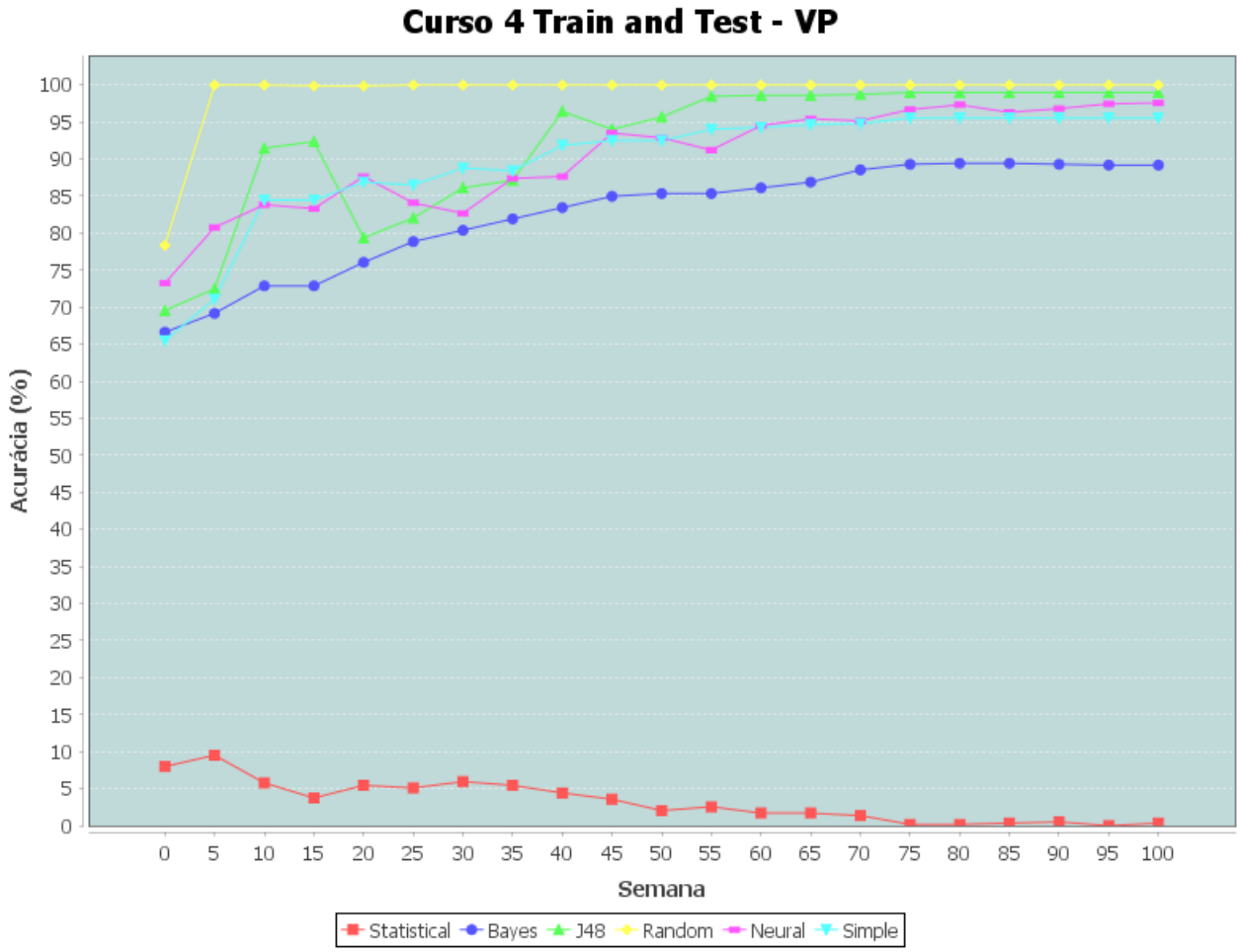

Figura 33: Resultado Cenário 2 Experimento final Curso 4 VP

ANEXO B CÓDIGO EM JAVA DE UTILIZAÇÃO DA BIBLIOTECA DO WEKA

```
1  /**
2  * comentario
3  import tratadados.*;
4  import weka.core.Instances;
5  import weka.experiment.InstanceQuery;
6  import java.io.BufferedReader;
7  import java.io.FileNotFoundException;
8  import java.io.FileReader;
9  import java.io.IOException;
10 import weka.classifiers.bayes.BayesNet;
11 import weka.filters.unsupervised.attribute.Remove;
12 import weka.classifiers.meta.FilteredClassifier;
13 import weka.classifiers.Evaluation;
14 import java.util.Random;
15 import tratadados.DAO.teste;
16
17
18 public class ABayestrainetest{
19
20     public ABayestrainetestnewversion() throws
          FileNotFoundException, IOException, Exception {
21         teste banco = new teste();
22         String algoritmo = "bayes";
23         BufferedReader reader = new BufferedReader(
24                 new FileReader("C:\\Moodledatas\\adm.arff"));
25         Instances data = new Instances(reader);
26         reader.close();
27
28         BufferedReader reader2 = new BufferedReader(
```

```
29                  new FileReader("C:\\Moodledatas\\bio.arff"));
30          Instances datatest = new Instances(reader2);
31          reader2.close();
32
33          // setting class attribute
34          data.setClassIndex(data.numAttributes() - 1);
35          datatest.setClassIndex(datatest.numAttributes() - 1);
36
37          Instances train = data;        // from somewhere
38          Instances test = datatest;     // from somewhere
39          // filter
40          for (int i = 103; i >= 1; i--) {
41
42              int temp = i;
43              System.out.println("Resultado da Semana:" + (temp -
                    1));
44              Remove rm = new Remove();
45
46              // classifier
47              BayesNet alg = new BayesNet();
48              // meta-classifier
49              FilteredClassifier fc = new FilteredClassifier();
50              rm.setAttributeIndices(coluna);  // remove 1st
                    attribute
51              fc.setFilter(rm);
52              fc.setClassifier(alg);
53              // train and make predictions
54              fc.buildClassifier(train);
55              fc.buildClassifier(test);
56              Evaluation eval = new Evaluation(train);
57              eval.evaluateModel(alg, test);
58
59              System.out.println(eval.weightedFalseNegativeRate());
60              System.out.println(eval.toSummaryString("\nResults\n
                    ======\n", false));
61              System.out.println(eval.toMatrixString());
62
63              int tamanho_teste = test.numInstances();
64
65              int conta_i = i - 1;
66
```

```
67              banco.salvaResultados(conta_i, eval.
                    weightedTruePositiveRate(), 1, algoritmo);
68              banco.salvaResultados(conta_i, eval.
                    weightedTrueNegativeRate(), 2, algoritmo);
69              banco.salvaResultados(conta_i, eval.
                    weightedFalsePositiveRate(), 3, algoritmo);
70              banco.salvaResultados(conta_i, eval.
                    weightedFalseNegativeRate(), 4, algoritmo);
71          }
72      }
73  }
```