UNIVERSIDADE FEDERAL DO RIO GRANDE DO NORTE
CENTRO DE CIÊNCIAS EXATAS E DA TERRA
DEPARTAMENTO DE INFORMÁTICA E MATEMÁTICA APLICADA
BACHARELADO EM ENGENHARIA DE SOFTWARE

Identificação de Perfis e Padrões de Participação dos Estudantes de Cursos a Distância na UFRN por meio de Mineração de Dados

Thaisa Ramos de Almeida

Natal-RN
Junho de 2015

Thaisa Ramos de Almeida

Identificação de Perfis e Padrões de Participação dos Estudantes de Cursos a Distância na UFRN por meio de Mineração de Dados

Monografia de Graduação apresentada ao Departamento de Informática e Matemática Aplicada do Centro de Ciências Exatas e da Terra da Universidade Federal do Rio Grande do Norte como requisito parcial para a obtenção do grau de bacharel em Engenharia de Software.

Orientadora

Profª Drª Márcia Jacyntha Nunes Rodrigues Lucena

Co-orientadora

Profª Dra Apuena Vieira Gomes

Universidade Federal do Rio Grande do Norte – UFRN

Departamento de Informática e Matemática Aplicada – DIMAp

Natal-RN

Junho de 2015

Monografia de Graduação sob o título Identificação de Perfis e Padrões de Participação dos Estudantes de Cursos a Distância na UFRN por meio de Mineração de Dados apresentada por Thaisa Ramos de Almeida e aceita pelo Departamento de Informática e Matemática Aplicada do Centro de Ciências Exatas e da Terra da Universidade Federal do Rio Grande do Norte, sendo aprovada por todos os membros da banca examinadora abaixo especificada:

Profª. Drª. Márcia Jacyntha Nunes Rodrigues Lucena

Orientadora

DIMAp - Departamento de Informática e Matemática Aplicada

UFRN –Universidade Federal do Rio Grande do Norte

Profª. Drª. Apuena Vieira Gomes

Co-orientadora

IMD - Instituto Metrópole Digital

UFRN - Universidade Federal do Rio Grande do Norte

Prof.º Dr. Ricardo Alexsandro de Medeiros Valentim

DEB - Departamento de Engenharia Biomédica

UFRN - Universidade Federal do Rio Grande do Norte

Ms. Giovani Ângelo Silva da Nóbrega

DEE – Departamento de Engenharia Elétrica

UFRN - Universidade Federal do Rio Grande do Norte

Natal-RN, 17 de junho de 2015

Agradecimentos

Agradeço a Deus, por estar sempre comigo me dando força e perseverança para alcançar meus sonhos.

Agradeço a minha mãe, Wilsa Ramos por ser uma mulher excepcional que sempre me ensino tanto e me apoiou nas horas difíceis, você é uma grande fonte de inspiração para mim. Obrigada por suas contribuições valiosas neste estudo.

Agradeço ao meu pai, Vilmar por todo o apoio e confiança, fazendo sempre tudo o que estava ao seu alcance para pode me ajudar e nunca deixando de acreditar no meu potencial.

Agradeço ao meu irmão, Igor por acreditar em mim e nunca me deixar desistir.

Agradeço ao Giovani Nóbrega, por ter compartilhado todo o seu conhecimento e por ter me ensinado tanto, você foi uma peça chave nesse trabalho.

Agradeço ao Prof. Dr. Ricardo Valentim pela oportunidade e confiança na elaboração deste estudo.

Agradeço a todos os meus amigos e familiares que sempre me incentivaram tanto e torceram por mim, mesmo alguns estando distantes. Um abraço especial para minhas avós Maysa e Ana.

Agradeço a toda a equipe da SEDIS, por terem compartilhado seus conhecimentos em educação a distância e terem esclarecido todas as minhas dúvidas.

Agradeço a minha orientadora Márcia Jacyntha, pelo suporte, auxilio nas correções e incentivo.

Agradeço a minha co-orientadora, Apuena Vieira pelas dicas e contribuições para este trabalho.

“Se a educação sozinha não transforma a sociedade, sem ela tampouco a sociedade muda.”

Paulo Freire

Identificação de Perfis e Padrões de Participação dos Estudantes de Cursos a Distância na UFRN por Meio de Mineração de Dados

Autor: Thaisa Ramos de Almeida

Orientador(a): Profª Drª Márcia Jacyntha Nunes Rodrigues Lucena

RESUMO

O interesse pelo tema mineração de dados educacionais surgiu da preocupação de ter informação sobre o desempenho do aluno durante a disciplina em cursos online. O *Moodle* possui vários tipos de relatórios, porém, não oferece dados precisos sobre as atividades e o perfil de desempenho do aluno para tomada de decisões por parte do professor. A pesquisa teve como objetivo aplicar a metodologia de mineração de dados para o levantamento dos perfis e padrão de participação de alunos em disciplina de curso superior a distância resultando na predição das chances de aprovação de cada aluno. Para tanto, desenvolveu um projeto de aplicação baseado em uma adaptação do KDD para mineração de dados educacionais. Os dados analisados foram extraídos de uma disciplina de curso superior a distância do banco de dados da SEDIS, com 497 (quatrocentos e noventa e sete) estudantes matriculados. O resultado obtido foi o agrupamento de 4 clusters que designam os perfis e tipos de participação dos estudantes, a saber: ativos, medianos, inconstantes e ausentes. O estudo é um contributo da área da computação para a educação, pois possibilita ao professor ter acesso aos dados de desempenho dos alunos no decorrer da disciplina podendo assim tomar medidas preventivas que evitem a reprovação ou trancamento da disciplina por parte dos alunos.

Palavras chaves: Mineração de Dados Educacionais; KDD; Perfis e Padrões de Participação; K-Means; Moodle.

Identification Profile and Participation Patterns from UFRN Students in Distance Courses Using Data Mining

Author: Thaisa Ramos de Almeida
Advisor: Profª Drª Márcia Jacyntha Nunes Rodrigues Lucena

ABSTRACT

The Interest in data mining theme emerged from the concern of having information about the student's performance during the course in online learning. Moodle has several types of reports, however it provides no data about the students activities and profile of students performance for teachers decision-making. The research aimed to apply a data mining methodology to identify the profiles and participation patterns of students in distance education courses, resulting in predicting the chances of approval for each student. It was developed an applied project based on an adaptation of KDD for educational data mining in Moodle platform. The analyzed data was extracted from a subject of online courses from the SEDIS database with 497 enrolled students. The result was the formation of four clusters designating the profiles and variety of student participation, namely: active, median, inconsistent and absent students. This study is a contribution of computer area for education because it allows the teacher to have access to student performance profile during the subject and then take some preventive measures to avoid disapproval or locking of subject by the students.

Keywords: Educational Data Mining; KDD; Profiles and Participation Patterns; K-Means; Moodle.

Lista de figuras

Lista de tabelas

Lista de abreviaturas e siglas

AVA – Ambiente Virtual de Aprendizagem

EDM – Educational Data Mining

KDD – Knowledge Discovery in Databases

LMS – Learning Management Systems

PCA – Principal Component Analysis

SEDIS – Secretaria de Educação a Distância da Universidade Federal do Rio Grande do Norte

TIC – Tecnologias de Informação e Comunicação

UAB – Universidade Aberta do Brasil

UFRN – Universidade Federal do Rio Grande do Norte

Sumário

1 Introdução

1.1 Contexto

O número de cursos a distância tem crescido, exponencialmente, e como resultado aumenta a demanda por melhoria dos processos de ensino e aprendizagem mediados pelas novas Tecnologias de Informação e Comunicação (TIC). Conforme os dados do Censo de Educação Superior de 2012, entre 2009 e 2012, a educação a distância teve um aumento superior a 275.000 matrículas o que corresponde a um crescimento de 32,9% neste período, correspondendo a uma taxa de aproximadamente 11% ao ano.

Entretanto, para os pesquisadores da área de educação superior a distância (NISTOR e NEUBAUER, 2010; FINNEGAN et al.2008; LYKOURENTZOU, 2009), um dos grandes problemas têm sido as altas taxas de evasão que são maiores que as do ensino presencial. Neste cenário, tem surgido métodos que permitem identificar de forma preventiva os alunos em situação de risco de reprovação e de evasão (NISTOR e NEUBAUER, 2010; LYKOURENTZOU, 2009). Entre esses procedimentos, destaca a Mineração de Dados Educacionais (do inglês: Educational Data Mining - EDM) que tem por objetivo desenvolver ou adaptar métodos para compreender melhor os dados oriundos de ambientes educacionais, e com isso oferecer melhores qualidade do processo de ensino a distância (COSTA et. al. 2012). Entre outras possibilidades, esses métodos visam mapear os perfis e padrões de participação podendo auxiliar os professores e gestores na melhoria dos processos de ensino e aprendizagem.

A educação a distância tem utilizado *software*s do tipo *Learning Management Systems* (LMS) para o desenvolvimento e realização dos seus cursos. Para Barros e Carvalho (2011, p.214), “os LMS são espaços eletrônicos construídos para permitir a veiculação e interação de conhecimentos e usuários”. O *Moodle* (*Modular Object-Oriented Dynamic Learning Environment*) é um dos LMS mais usados no ensino superior em várias Universidades em outros países. No Brasil, ele foi adotado pelas

Universidades que fazem parte do Sistema Universidade Aberta do Brasil (UAB). A Universidade Federal do Rio Grande do Norte (UFRN), participante da UAB, tem utilizado o *Moodle* para a oferta dos seguintes cursos a distância: Física, Matemática, Química, Geografia, Ciências Biológicas, Administração Pública, Educação Física, Pedagogia e Letras. Neste contexto, justifica-se a opção pela aplicação do projeto utilizando os dados do *Moodle* visto que o estudo foi realizado na UFRN.

Segundo Ramos e Medeiros (2010, p.54), o *Moodle* é:

> Um software livre de código aberto distribuído gratuitamente, que possibilita o trabalho colaborativo entre os participantes em um mesmo ambiente de aprendizagem mediante o uso da internet. O termo técnico software livre significa que o usuário pode modificar, usar e distribuir o programa de acordo com suas necessidades didáticas e de conteúdo. Entretanto, as modificações/alterações de melhorias que o usuário realiza no sistema devem retornar para a comunidade sem nenhum custo. O Moodle oferece aos alunos e professores características semelhantes a uma sala de autoria de material didático e a administração da sala virtual, além do monitoramento de atividades virtuais.

O Moodle tem capacidade de armazenar uma grande quantidade de dados sobre as ações dos alunos no ambiente, porém, essas informações são muito extensas o que torna o processo de visualização e entendimento delas muito complexo.

Neste sentido, o estudo proposto visa a identificação de perfis e padrões de participação por meio da mineração de dados educacionais por se tratar de uma metodologia relevante para a melhoria da qualidade da gestão educacional em cursos superiores a distância.

1.2 Motivação

A necessidade pedagógica de acompanhar o desempenho do aluno durante uma disciplina é cada vez maior. Apesar do *Moodle* disponibilizar algumas ferramentas, não é possível ter durante a disciplina um perfil real do desempenho do estudante em termos de participação no ambiente.

O *Moodle* possui um relatório que mostra as ações do aluno na

disciplina de forma pontual. Entretanto, não oferece dados de todas as atividades dos estudantes de forma comparativa com outros alunos, mostrando poucas informações detalhadas sobre o perfil de desempenho dos alunos, principalmente, daqueles que estão em risco de reprovação. Devido a isso, existe a necessidade de proporcionar aos gestores educacionais e professores uma metodologia para análise do perfil e do tipo de participação dos estudantes que possa predizer o sucesso ou o fracasso dos alunos.

1.3 Objetivos

Este estudo tem como objetivo principal aplicar a metodologia de mineração de dados para o levantamento dos perfis e padrões de participação de alunos em disciplina de curso superior a distância resultando na predição das chances de aprovação de cada aluno.

1.3.1 Objetivos Específicos

1. Realizar um estudo sobre a estrutura do banco de dados do *Moodle*.
2. Desenvolver uma análise preliminar dos dados coletados referente as atividades e recursos utilizados no ambiente virtual do *Moodle* da SEDIS.
3. Selecionar uma disciplina e as tabelas do banco referentes a esta disciplina.
4. Criar uma tabela de sumarização dos dados dos alunos da disciplina selecionada e verificar quais ações no ambiente têm mais correlação com a nota final.
5. Aplicar técnicas de limpeza de dados e transformação, tipo normalização Min-Max e redução de dados com PCA.
6. Utilizar a técnica de clusterização K-Means para gerar os grupos de alunos com ações similares no ambiente virtual e validar o número de clusters utilizados.
7. Interpretar e analisar os dados encontrados.

1.4 Metodologia

Esta monografia trata de um projeto de aplicação de uma metodologia de mineração de dados para o levantamento dos perfis e padrões de participação de alunos em disciplina de curso superior a distância na Universidade Federal do Rio Grande do Norte. Espera-se que a partir dos resultados e de estudos complementares possa ser construído um modelo de predição dos resultados de desempenho, que ofereça ao professor e ao tutor formas de visualização dos tipos de participação dos alunos para identificar aqueles que estão em risco de reprovação e propor ações de remediação ou resgate.

Os dados de análise foram extraídos da base de dados do *Moodle* utilizado pela Secretaria de Educação a Distancia (SEDIS) da UFRN para os cursos de graduação a distância. A utilização do banco de dados da SEDIS foi concedida mediante a formalização da solicitação de autorização para o uso de dados, conforme consta no anexo A.

1.5 Organização do trabalho

Este trabalho está organizado em cinco capítulos, dos quais o primeiro é a introdução onde foram abordadas motivação, objetivo e breve sumário da metodologia da pesquisa. O segundo é a revisão de literatura, o terceiro é a descrição da aplicação da metodologia de mineração de dados, o quarto é a interpretação e análise dos dados e o quinto a conclusão e trabalhos futuros.

2 Referencial Teórico

Nessa seção são apresentados os principais assuntos relacionados a pesquisa, que são: ambientes virtuais de aprendizagem, perfis e padrões de participação dos alunos em cursos a distância, mineração de dados, mineração de dados educacionais e as técnicas de mineração de dados utilizadas.

2.1 Ambientes Virtuais de Aprendizagem

Desde duas décadas, foram criados sistemas eletrônicos para a oferta de cursos a distância baseado na web. Esses sistemas são *software* projetado para oferecer recursos e ferramentas para a construção e gestão de salas de aula virtuais e têm como característica o gerenciamento de integrantes, relatório de acesso e atividades, promoção da interação entre os participantes de forma síncrona ou assíncronas, publicação de conteúdos e repositório de vários objetos de aprendizagem em formatos tipo vídeo, áudio, etc. (BARROS e CARVALHO, 2011).

O Moodle é um software tipo LMS, por isso, ele possibilita a criação de ambientes virtuais de aprendizagem (AVA) para serem usados como sala de aula virtual em cursos a distância, e também, em apoio a cursos presenciais.

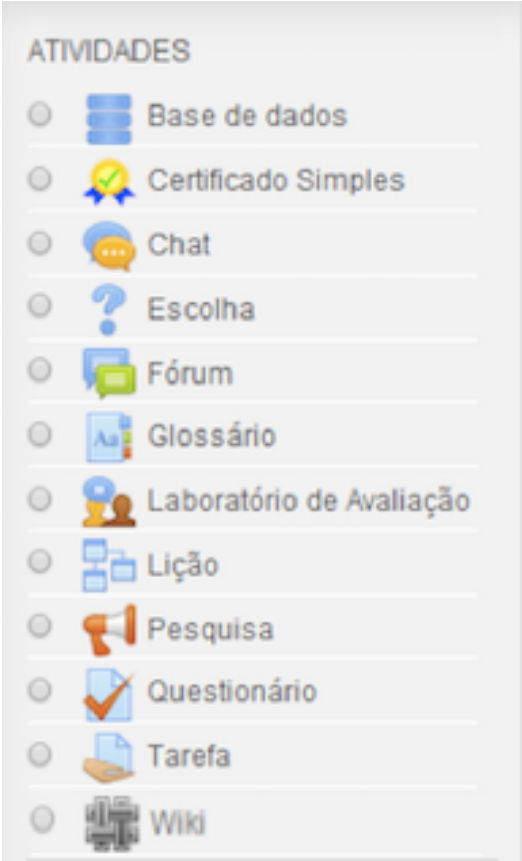
Para os autores Barros e Carvalho (2011) os AVA se diferem de outros ambientes web por terem uma dinâmica própria que atendem bem ao processo de ensino e aprendizagem e ao cumprimento de metas para o aluno.

O *Moodle* possui vários tipos de ferramentas para auxiliar na eficiência de um curso virtual, divididas em dois grupos:

1. **Atividades**: Corresponde a tudo que é criado no ambiente com o intuito de obter uma ação e interação do aluno com o objeto criado, normalmente as atividades são criadas pelo professor, tutor ou responsável pelo ambiente. A Figura 1 ilustra algumas da atividades

que o *Moodle* possui.

Figura 1 - Atividades do Moodle



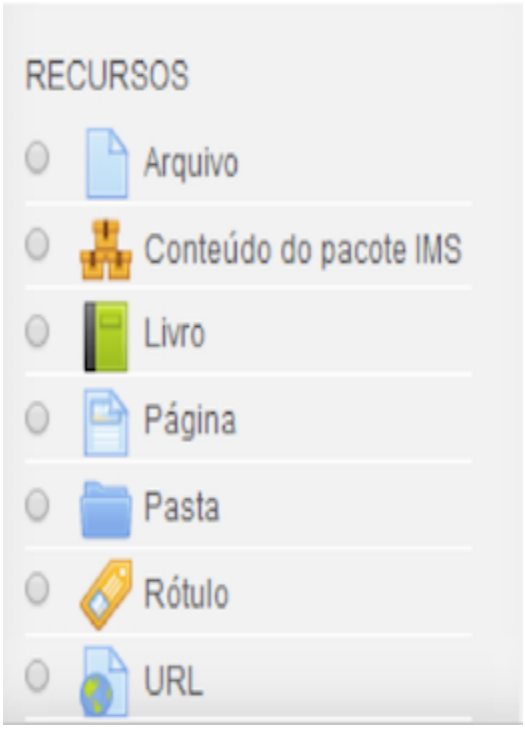
Fonte: autoria própria

A seguir é apresentado uma breve descrição das atividades mais utilizadas:

- *Fórum*: permite discussões assíncronas sobre um determinado tema. O aluno pode realizar uma postagem em uma discussão já aberta ou criar uma nova discussão.
- *Assign*: é uma atividade em que se deve enviar um arquivo ou texto, que será submetido para a avaliação do professor.
- *Quiz*: consiste de um conjunto de questões elaboradas pelo professor, as questões podem ser de múltipla escolha, verdadeiro ou falso, respostas curtas, numéricas, entre outras.
- *Chat*: permite conversação em tempo real entre os seus participantes, professores e tutores,

2. **Recursos**: são informações e conteúdos disponibilizados pelos professores para os alunos. A Figura 2 ilustra alguns dos recursos que o *Moodle* possui.

Figura 2 - Recursos do *Moodle*



Fonte: autoria própria

Os recursos mais utilizados são explicados a seguir:

- *Resource*: é um arquivo podendo ser de diferentes formatos, disponível para visualização ou download.
- *Label*: são textos, imagens ou vídeos que podem ser utilizados como divisão de uma semana ou tópico.
- *Folder*: é uma pasta que contém vários arquivos como texto, imagem, vídeos e etc.
- *URL*: é um link de uma página fora do ambiente.
- *Page*: é uma página que contém informações ou instruções.

Outra função importante do *Moodle* é a atribuição de papéis aos usuários do sistema. Os papéis de usuários podem ser: gerente,

criador de cursos, professor, moderador, estudante, visitante, usuário autenticado, usuário autenticado na página inicial, tutor a distância, tutor presencial, secretário, coordenador de polo, professores.

2.2 Perfis e Padrões de Participação dos Alunos em Cursos a Distância

O Estudo visa identificar os perfis e padrões de participação dos alunos em cursos a distância. Na análise do tipo de perfil, os autores têm usado o termo sobre diversos prismas, mas pode-se dizer que ele envolve a ideia de pertencimento a uma comunidade (JALDEMARK et al., 2006) e de cumprimento de atividades (HRASTINSKI, 2008, 2009). As atividades online são diversas e tudo que envolve a análise da realização dessas atividades está envolvendo a construção de teorias sobre o tema. Na literatura (HRASTINSKI, 2009), a participação tem sido aferida pela quantidade de logs e de acesso aos recursos do AVA. Os logs e acessos representam para o estudante, as escolhas, as necessidades e a motivação para os estudos. Nesse sentido: "a participação envolve tudo que nós fazemos e sentimos quando tomamos parte em uma experiência significativa". (HRASTINSKI, 2009, p.81). Sendo assim, para o autor, a participação online é um processo complexo que inclui fazer, falar, pensar, sentir e perceber.

O autor realizou a revisão da literatura e identificou seis níveis de participação, são eles:

Nível 1 - Participação pelo acesso ao ambiente virtual, em que a participação é igualada ao número de vezes que o aluno acessa o ambiente.

Nível 2 - Participação na forma de escrita, em que o aluno posta muitas mensagens ou várias palavras em um fórum.

Nível 3 - Participação por meio de mensagens de qualidade, o aluno contribui com postes no fórum de alta qualidade, caracterizando uma participação mais ativa do que os outros.

Nível 4 - Participação na forma de escrita e leitura, alunos que visualizam e postam mensagens nos fóruns são mais ativos do que os que

não realizam essas ações.

Nível 5 - Participação quando o estudante percebe que suas contribuições escritas são importantes para o grupo, portanto, ele passa a participar de forma mais ativa que os demais.

Nível 6 - Participação em termos de participar de um diálogo gratificador em um fórum, desta forma, o aluno se sente parte do diálogo.

Outros autores, como Nistor e Neubauer (2010) também estudaram o tema. Para eles, é importante entender e identificar os diferentes tipos de participação no AVA para evitar a evasão. A evasão em cursos online é o calcanhar de Aquiles para os gestores educacionais.

Nistor e Neubauer (2010) dividem participação em dois tipos, quantitativa e qualitativa, porém, só fizeram uso da análise quantitativa, que corresponde ao número de ações feitas, frequência, tamanho das mensagens postadas em fóruns. Os dados da análise quantitativa podem expressar a participação passiva, em que os estudantes visualizam todas as atividades, porém não produzem conteúdo, não participam, sendo denominados de *lurkers* ou de observadores. Nos estudos dos autores, eles encontraram quatro tipo de participação:

1. Alunos altamente comprometidos: consiste de estudantes ativos que participam em todas as modalidades.
2. Alunos de locais mais isolados: apresentam um esforço de aprendizagem mínimo.
3. Alunos oriundos do curso de Ciência da Educação: apresentam um desempenho muito padronizado.
4. Estudantes evadidos: abandonaram o curso.

Rodrigues et al (2013) também realizaram um estudo para identificação de padrões de participação dos estudantes, eles encontraram os seguintes padrões:

1. Cursista ativo: cursista presente que participa durante todo o curso.
2. Cursista mediano: cursista que participa medianamente e demonstra querer a aprovação.

3. Cursista intermediário: possuem uma participação mais ativo do meio para o final do curso e podem ter mais chance de aprovação do que os estudantes ativos desde o começo do curso.
4. Cursista passivo: não possui uma participação regular, podendo até participar em termos quantitativo mais sem muita qualidade, esses cursistas podem ou não ser aprovados.
5. Cursista ausente: quase não tem interação com o ambiente, não demonstra compromisso social com a turma e não tem interesse de obter a nota mínima para aprovação.

Os mesmos autores concluíram que as atividades de visualização de fórum e de entrega de atividades (tarefa) estão altamente relacionadas com a aprovação do estudante. O estudo de Rodrigues et al (2013) demonstrou que existem muitas formas de classificar os tipos e padrões de participação.

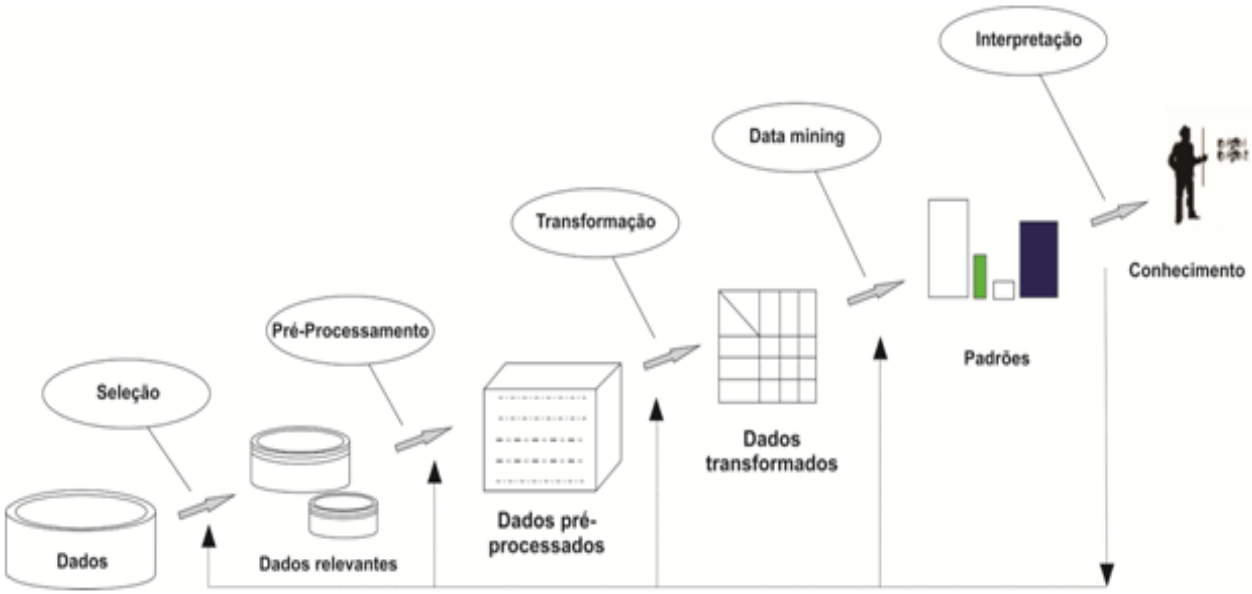
2.3 Mineração de Dados

A mineração de dados faz parte da principal etapa de um processo amplo conhecido como Descoberta de Conhecimento em Bancos de Dados (Knowledge Discovery in Databases, KDD). A mineração de dados é definida segundo Carvalho (2001) como:

> O uso de técnicas automáticas de exploração de grandes quantidades de dados de forma a descobrir novos padrões e relações que, devido ao volume de dados, não seriam facilmente descobertos a olho nu pelo ser humano

A definição de KDD é dada justamente por aqueles que o criaram, “processo, não trivial, de extração de informações implícitas, previamente desconhecidas e potencialmente úteis, a partir dos dados armazenados em um banco de dados”. (FAYYAD, 1996). Visando melhor entendimento do KDD é preciso entender que o significado de não trivial quer dizer que necessita de técnicas de busca ou inferência e o significado de potencialmente útil quer dizer que devem trazer benefícios aos usuários. O processo de KDD pode ser visualizado na Figura 3.

Figura 3 - Ciclo do processo KDD.



Fonte: Adaptação de (FAYYAD, 1996, Brito, 2012).

Conforme Fig. 3, o KDD contém as seguintes fases:

1. Seleção do dados: seleciona-se o conjunto de dados que serão analisados.
2. Pré-processamento: realiza uma limpeza dos dados para remover dados inconsistentes, dados redundantes e completar dados ausentes, fazendo com que os dados não atrapalhem as próximas etapas.
3. Transformação: transforma e formata os dados de forma adequada, utilizando-se de métodos de redução de dados e de transformação para diminuir o número de variáveis e normalizar os atributos e, melhorar a performance dos algoritmos de mineração.
4. Mineração de dados: nessa etapa escolhe os melhores métodos para a busca e extração de padrões.
5. Interpretação e avaliação: identifica os padrões e avalia todos os resultados obtidos para tomadas de decisões.

O processo KDD deve seguir as fases na ordem descrita, porém, a qualquer momento pode-se sentir a necessidade de retornar a uma ou mais etapas para obter um melhor resultado. Nesse estudo o processo KDD foi reiniciado varias vezes para conseguir um melhor resultado condizente com os objetivos do nosso estudo.

2.4 Mineração de Dados Educacionais

A Mineração de Dados Educacionais (Educational Data Mining - EDM) analisa dados procedentes de ambientes educacionais. Segundo Baker (2009) é uma área em crescimento e contínuo desenvolvimento que tem enorme potencial de melhorar o processo de ensino e aprendizagem dos alunos.

Os dados coletados de ambientes educacionais possuem várias informações importantes que podem compreender o comportamento e aprendizagem dos alunos e, com isso, auxiliar os docentes a criar ambientes mais eficazes para o processo de aprendizagem.

De acordo com Baker, Isotani, Carvalho (2011) nem todos os algoritmos e ferramentas de mineração de dados podem ser utilizados para análise de dados educacionais devido à falta de independência estatística dos dados oriundos de ambientes educacionais. Os mesmos autores propuseram uma taxonomia das técnicas que deve ser usada em mineração de dados educacionais. As técnicas são:

- Predição (*Prediction*): prediz aspectos específicos dos dados.
    - Classificação (*Classification*)
    - Regressão (*Regression*)
    - Estimação de Densidade (*Density Estimation*)
- Agrupamento (*Clustering*): o objetivo principal é formar grupos que possuem semelhanças. Estes grupos não são conhecidos inicialmente.
- Mineração de relações (*Relationship Mining*): descobrir possíveis relações entre variáveis em conjunto de dados.
    - Mineração de Regras de Associação (*Association Rule Mining*)
    - Mineração de Correlações (*Correlation Mining*)
    - Mineração de Padrões Sequenciais (*Sequential Pattern Mining*)
    - Mineração de Causas (*Causal Mining*)
- Destilação dos dados para facilitar decisões humanas (*Distillation of Data for Human Judgment*): apresenta dados complexos de uma forma de fácil compreensão.

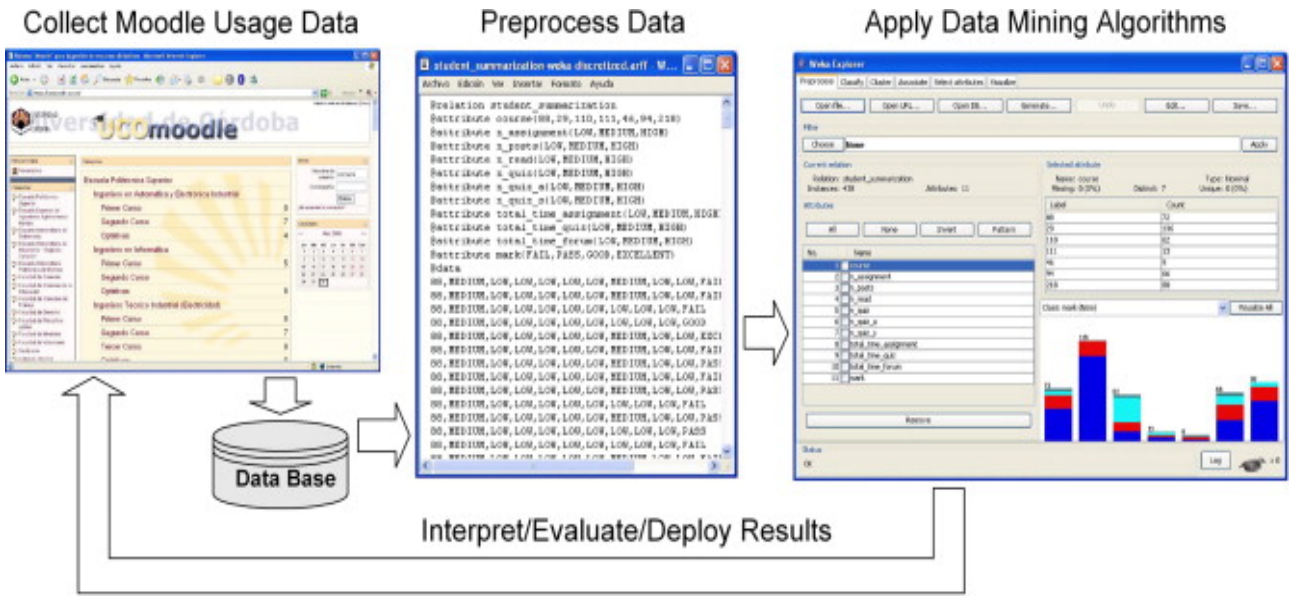
- Descoberta por modelos (*Discovery with Models*): utiliza-se de modelos existentes.

Romero, Ventura e Garcia (2008) explicam que o processo de mineração de dados educacionais em *e-learning* consiste dos quatro seguintes passos:

1. Coleta de dados: armazenar as informações das interações dos estudantes no ambiente em um banco de dados.
2. Pré-processamento dos dados: é feito uma limpeza e transformação dos dados em formatos apropriados para a utilização da mineração de dados no próximo passo.
3. Aplicar mineração de dados: Algoritmos de mineração são utilizados para encontrar os padrões desejados.
4. Interpretar, avaliar e implementar os resultados: os resultados são encontrados, interpretados e utilizados para tomada de decisões futuras que auxiliem os estudantes, professores, tutores e administradores e melhore o processo de ensino e aprendizagem.

Romero e Ventura (2007) mencionam que a aplicação de mineração em ambientes virtuais de aprendizagem é um ciclo interativo.

Figura 4 - Processo KDD aplicado a EDM



Fonte: Romero, Ventura e Garcia (2008)

A Fig. 4 ilustra os passos descritos do processo KDD aplicado a EDM.

2.4.1 Trabalho Relacionados

Diversos estudos têm sido feitos na área de mineração de dados educacionais, são eles:

1. Senechal (2013) realizou um pré-processamento de dados de informações extraídos do *Moodle* para identificar perfis de aprendizagem e com base nos perfis detectar maneiras de auxiliar o processo de ensino e aprendizagem dos alunos. O estudo de Senechal (2013) assim como este estudo, utilizou-se da técnica de agrupamento para gerar os perfis de aprendizagem, porém, foi utilizado a técnica rede neural de Kohonen, diferentemente, deste estudo que utilizou o K-Means.
2. Marques (2014) realizou um estudo para identificar estudantes com risco de evasão ou reprovação por meio de técnicas de mineração de dados educacionais, para caracterizar o perfil de acesso dos alunos. O estudo analisou as interações do estudante no ambiente e suas características sociais. Uma grande diferença do estudo de Marques (2014) para este é a técnica de mineração utilizada, o dele utilizou a predição, diferentemente deste estudo que utilizou agrupamento.
3. Conti (2011) realizou uma análise dos prazos e submissão de atividades no *Moodle* para identificar padrões referentes as postagens das tarefas, e com isso auxiliar na tomada de decisão sobre postagens enviadas próximo da data limite a fim de evitar a reprovação ou evasão do aluno, como também integrar o processo de KDD ao *Moodle*. O estudo de Conti (2011) utilizou-se da técnica de agrupamento e classificação na etapa de mineração de dados.

Dentre os trabalhos correlatos, não foi possível encontrar um trabalho que identificasse os tipos de perfis e padrões de participação dos alunos que pudesse predizer as chances de aprovação de cada grupo de acordo com o tipo de participação no ambiente.

2.5 Técnicas de Mineração de Dados Utilizadas

As técnicas de mineração de dados utilizadas para o estudo em questão, foram: mineração de correlação, normalização, redução de dimensionalidade, agrupamento e determinação do número ótimo de clusters.

A seguir, descrevemos cada uma das técnicas.

2.5.1 Mineração de correlação

Esta técnica tem por objetivo encontrar correlações entre diferentes atributos. Para atingir esse objetivo foi utilizado correlação de Pearson e Spearman.

O coeficiente de correlação de Pearson foi descrito por Karl Pearson em 1897 de acordo com SCHULTZ e SCHULTZ (1992), ele mede o quanto dois conjunto de dados estão linearmente correlacionados, os valores do coeficiente variam de -1 ate +1, em que -1 indica perfeita correlação negativa, +1 perfeita correlação positiva e 0 significa que não a correlação. Para Cohen (1988), valores entre 0,10 e 0,29 podem ser considerados pequenos, entre 0,30 e 0,49 podem ser considerados como médios e, valores entre 0,50 e 1 podem ser compreendidos como grandes. O importante é entender que quanto mais próximo de 1 mais forte é a correlação.

O coeficiente de correlação de postos de Spearman foi introduzido por Spearman (1904), ele reflete a intensidade e o sentido das relações monótonas entre dois conjuntos de variáveis. O valor do coeficiente é interpretado da mesma forma que o coeficiente de Pearson, ou seja, varia de -1 ate +1.

Quando o coeficiente de Pearson e de Spearman apresentarem valores parecidos, é uma relação linear, quando Spearman for maior que Pearson é uma relação não linear monótona, quando Pearson for maior, provavelmente ocorre a presença de elementos fora da curva normal (*outliers*).

Os coeficientes de Pearson e Spearman são bastante utilizados na área de EDM, um exemplo seria o estudo de Peña-Ayala (2014) que realizou um

teste e encontrou uma correlação entre a nota que o aluno tira na prova de matemática no vestibular e a nota do seu primeiro teste de programação na disciplina analisada. Outro exemplo é o estudo de Rajadhyax e Shirwaikar (2012) que utiliza Pearson para medir o quanto duas disciplinas estão correlacionadas. Já o estudo de Mödritscher, Andergassen e Neumann (2013) utiliza a correlação de Pearson para analisar as influências das atividades do aluno na nota final.

2.5.2 Normalização

Conforme Visalakshi e Thangavel (2009), normalização dos dados é um dos procedimentos de pré-processamento realizados antes da aplicação de técnicas de mineração de dados em que os atributos são ajustados para um intervalo especifico como -1.0 ate 1.0 ou 0.0 até 1.0. Existem alguns métodos de normalização como, por exemplo, normalização Min-Max, Z-Score e Escalonamento decimal. Os mesmo autores realizaram um estudo em que aplicaram os 3 diferentes tipos de normalização antes de utilizar o algoritmo K-Means e encontraram que o melhor método de normalização depende da natureza dos dados que serão normalizados. Este estudo escolheu a normalização Min-Max, pois os dados são limitados e não possui muita variação de mínimo para máximo. De acordo com Peña-Ayala (2014, p.52) “na educação o método mais utilizado é a normalização Min-Max”.

A normalização Min-Max aplica uma transformação linear nos dados em que o maior valor será 1 e o menor será 0, os valores são calculados da seguinte forma:

$$\mathrm{v}' = \frac{v - minA}{maxA - minA}$$

O valor v' representa o resultado da normalização, o v é o valor que deseja normalizar e o minA e maxA representam o menor valor e maior valor encontrado nos dados.

Segundo Visalakshi e Thangavel (2009) a normalização é um passo essencial antes de utilizar algoritmos de agrupamento, especialmente, os que utilizam distância euclidiana, pois isso os torna muito sensível a grandes

escalas dos atributos e pode ocorrer de um atributo sobrepor o outro. Esses mesmos autores realizaram um estudo em que a utilização de normalização dos dados, feita antes de utilizar o algoritmo K-Means, resultou em uma melhor qualidade dos *clusters* gerados.

2.5.3 Redução de dimensionalidade

Redução da dimensionalidade é um método que reduz a dimensionalidade de um conjunto de dados procurando manter as características do conjunto sem afetar o resultado final. Segundo Fodor (2012) a redução de dimensionalidade é importante na análise de cluster, pois reduz os dados dimensionais elevados e os custos computacionais, como também provê aos usuários uma visão mais clara e uma visualização dos dados de interesse, levando a um melhor aproveitamento dos algoritmos de mineração de dados.

O nosso estudo escolheu o método de análise de componentes principais (*principal component analysis* - PCA) para redução. O PCA é indicado para conjuntos de medidas correlacionadas linearmente, que assim podem ser reduzidas a poucas variáveis sintéticas, denominadas componentes principais (PIELOU, 1984; MANLY, 1994).

Medeiros e Costa (2009, p. 2) explicam que o PCA:

> Aplica uma transformação linear sobre um conjunto n-dimensional de dados de entrada e encontra um novo sistema de coordenadas de forma que a projeção de maior variância possível do conjunto de dados de entrada coincida com o primeiro eixo desse novo sistema (denominado primeiro componente principal), a de segunda maior variância, com o segundo eixo e assim sucessivamente, para os n novos eixos. O PCA pode ser usado para obter uma redução de uma dimensão original n para uma dimensão reduzida m(m<n) selecionando-se os m primeiros componentes principais de um determinado conjunto de dados e ignorando os demais.

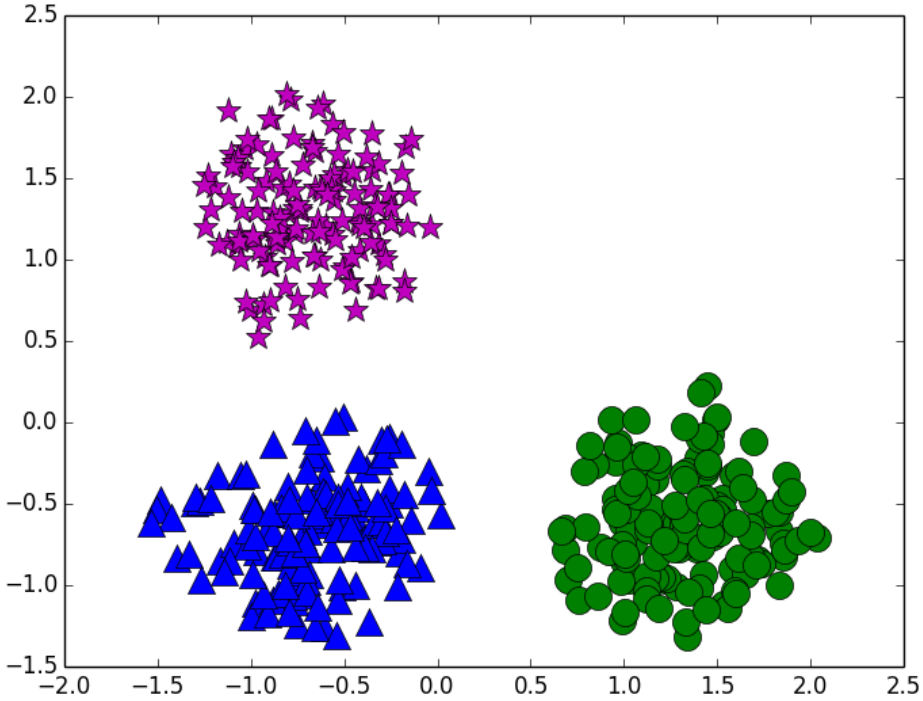
2.5.4 Agrupamento

Agrupamento é uma técnica que forma grupos (clusters) que possuem características semelhantes. De acordo com Tan, Steinbach e Kumar (2006),

quanto maior a semelhança ou homogeneidade dentro de um grupo e maior a diferença de um grupo para outro, melhor será o agrupamento.

A técnica de agrupamento ou clusterização é um método não supervisionado, nesse tipo de abordagem os dados utilizados não são rotulados. Um exemplo de agrupamento é ilustrado na Fig. 5.

Figura 5 - Exemplo de uma aplicação de uma técnica de agrupamento



Fonte: autoria própria

Na Fig. 5, é possível observar a formação de 4 clusters e a distância que cada um se encontra do outro, cada clusters é representado por uma cor e um formato diferente.

A clusterização tem sido muito utilizada na área de mineração de dados educacionais. Romero, Ventura, Garcia (2008) fizeram uma revisão da literatura e encontraram algumas investigações que sido feito com agrupamento, são eles: identificar alunos com características similares de aprendizado promovendo a aprendizagem colaborativa baseada em cada grupo, agrupamento de estudantes com propósito de acompanha-los de

acordo com suas competências e características, agrupamento de usuários de acordo com tempo de navegação de cada sessão.

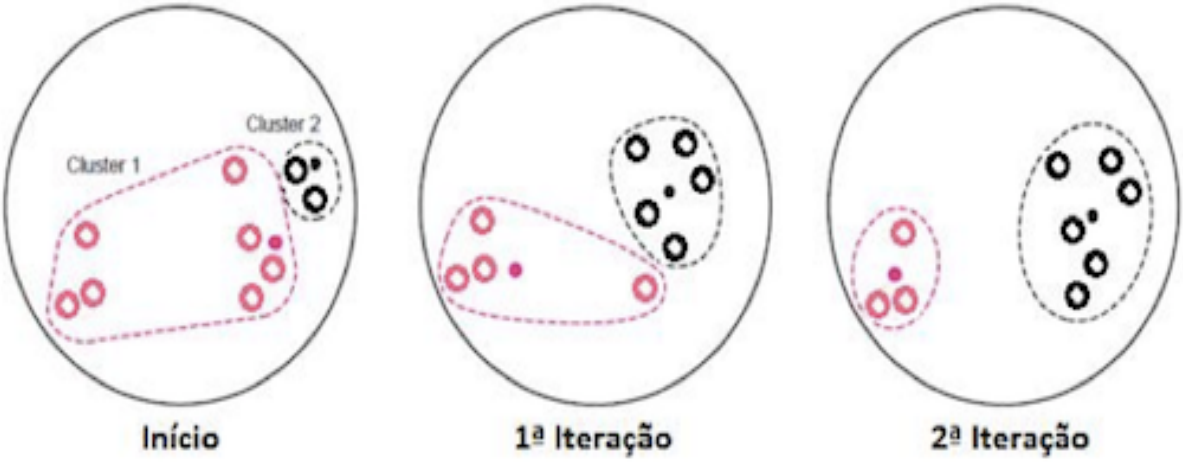
Os algoritmos de clusterização podem ser hierárquicos e não hierárquicos (particionamento). Antonenko, Toy e Niederhauser (2012) explicam que o método hierárquico começam com cada caso um cluster separado e em seguida vai combinando sequencialmente os clusters para construir uma hierarquia de clusters aninhados e a cada interação o número de clusters é reduzido. Este método é útil quando não se sabe o número certo de agrupamentos no banco de dados. Já no método não hierárquico o número de clusters (K) é previamente definido, um exemplo é o algoritmo K-Means que foi utilizado no nosso estudo.

Segundo Jain (2008) o algoritmo K-Means é dos algoritmos de clusterização mais simples e popular, ele foi publicado a primeira vez em 1955, a razão da sua popularidade é pelo fato de ser um algoritmo de fácil implementação, simples e eficiente. O algoritmo funciona da seguinte maneira:

1. Escolhe o número de clusters K.
2. Escolhe K centros dos clusters (pode ser aleatoriamente, forçar os centros a serem distantes um do outro, entre outros)
3. Atribui cada instância a um clusters com base na menor distância euclidiana do atributo ao centros dos clusters.
4. Recalcula o centro dos clusters fazendo a média (centróides) de todos as instâncias associadas a ele no passo anterior.
5. Volta para passo 3 até que os centróides permaneçam estáveis, ou o número de interações alcance o limite.

A Figura 6 mostra um exemplo dos passos do algoritmo K-Means. No início, os centros do clusters são definidos, representados pelos pontos fechados e as instâncias são associadas ao centro mais próximo, na 1ª iteração o centro é recalculado e as instâncias são realocadas para o centro de menor distância euclidiana, na 2ª iteração novamente centros e instâncias são realocados e o algoritmo é finalizado.

Figura 6 - Demonstração do K-Means.



Fonte: Grassi et al. (2013)

2.5.4 Determinação do número ótimo de clusters

Existem varias técnicas para determina o melhor número de clusters (K) para os dados. Essas técnicas têm por objetivo encontrar o melhor número K que faz com que as instâncias de um mesmo clusters sejam o mais similar possível e distante de outros clusters gerados.

O nosso estudo escolheu o método coeficiente de silhueta para determinação do K, de acordo com Medeiros (2014) esse método é uma dos mais utilizados para validação do número de cluster. Esta técnica foi proposta por Rousseeuw (1987) e é calculado da seguinte maneira:

$$S_i = \frac{(b_i - a_i)}{\max(b_i,\ a_i)}$$

Para cada instância i é calculado o coeficiente de silhueta (S_i), em que b_i representa a média da distância mínima da instância i para os demais clusters e a_i representa a média da distância da instância i aos demais atributos do clusters que ela pertence. O coeficiente pode variar de -1 ate +1

sendo que quanto mais próximo de +1 significa que a instância está bem localizada no cluster, próximo de 0 indica que a instância está muito perto de outro cluster e quanto mais próximo de -1 significa que a instância pode ter sido atribuída ao cluster errado.

Para calcular a média de todas as instâncias dos clusters gerados utiliza-se a seguinte fórmula:

$$S_{t=} \frac{\sum_{i=1}^{i=n} S_i}{n}$$

O S_t representa a média do coeficiente de silhueta, que é calculado pela soma de todos os coeficientes de silhueta de cada instância *i* dividido por *n,* que representa o número total de instâncias.

De acordo com Smith (2008) para os valores do coeficiente de silhueta entre 0.71 até 1.0 pode ser considerado que uma forte estrutura foi formada, valores entre 0.51 até 0.70 é considerado que uma estrutura razoável foi formada, 0.26 até 0.50 foi formada uma estrutura fraca e abaixo de 0.25 nenhuma estrutura substancial foi encontrada.

A revisão de literatura mostra a importância da técnica de mineração de dados para o trabalho de gestores e professores na compreensão dos tipos e padrões de comportamento dos estudantes, permitindo antecipar e prevenir situações de fracasso nos estudos ou evasão. Na próxima etapa, é descrito a metodologia do estudo.

3 Aplicação da Metodologia de Mineração de Dados Educacionais

Esta monografia tem como objetivo principal aplicar a metodologia de mineração de dados para o levantamento dos perfis e padrões de participação de alunos em disciplina de curso superior a distância resultando na predição das chances de aprovação de cada aluno.

3.1 Descrição da Metodologia de Mineração de Dados

Os dados para análise foram extraídos da base de dados do *Moodle* utilizado pela SEDIS para os cursos de graduação a distância.

Na escolha do processo utilizado para aplicação da metodologia, optou-se pela adaptação do KDD para EDM tendo em vista as suas funcionalidades e a congruência com os objetivos da pesquisa.

A adaptação do processo KDD seguiu as seguintes etapas: análise preliminar dos dados do *Moodle* da SEDIS, pré-processamento, aplicar mineração de dados e interpretar e analisar os dados. A etapa de coleta de dados não foi utilizada pois os dados já haviam sido coletados, essa etapa foi substituída pela de análise preliminar dos dados e a fase de implementação dos resultados não foi utilizada e é proposta como trabalhos futuros.

Para tanto, as atividades desenvolvidas seguiram os seguintes passos:

Etapa 1 - Análise preliminar dos dados do *Moodle* da SEDIS.

Foi realizada uma análise preliminar dos dados do *Moodle* para o planejamento das demais etapas e foi divida em quatro fases:

Fase 1. Estudo das tabelas importantes do *Moodle* da SEDIS

Nessa fase foi feito uma análise de todas as tabelas do banco de dados do *Moodle* para entender sua organização e poder extrair as informações necessárias

Fase 2. Análise das atividades e recursos mais utilizados

Foi verificado quais atividades e recursos são mais utilizados pelos professores da universidade, para auxiliar nos atributos que seriam selecionados para formar os perfis de participação dos alunos.

Fase 3. Seleção da turma Teste

Foi selecionado uma turma para servir de teste na realização de todos os procedimentos. Essa turma foi nomeada turma teste.

Fase 4. Seleção das tabelas e dos atributos

Foram selecionadas todas as tabelas do banco que continham as informações necessárias para extrair as informações da turma teste.

Etapa 2 - Pré-processamento

Conforme explicado no referencial teórico o pré-processamento envolveu uma limpeza e transformação dos dados em formatos apropriados para a utilização da mineração na etapa seguinte. Essa etapa foi dividida em 5 fases:

Fase 1. Criação da tabela de sumarização

Foi criado uma tabela de sumarização para armazenar todos os dados de cada aluno na turma teste.

Fase 2. Limpeza dos dados

Foi realizado uma limpeza de dados inconsistentes e foi feita a padronização das unidades.

Fase 3. Seleção dos atributos mais significativos

Foram feitos testes de correlações entre a nota final e cada atributo, para assim poder selecionar os que tinham maior impacto na nota e descartar os que não a influenciaram.

Fase 4. Normalização

Nessa fase foi realizada a normalização Min-Max para cada atributo selecionado.

Fase 5. Redução dos dados

Foi aplicado o PCA para reduzir a dimensionalidade dos dados, os dados possuíam 9 dimensões e foram reduzidos para 2.

Etapa 3 – Aplicar Mineração de dados

Essa etapa foi dividida em 2 fases:

Fase 1. K-Means

Nessa fase foi utilizado o algoritmo K-Means para agrupar os alunos em 4 (quatro) grupos, conforme suas ações no ambiente.

Fase 2. Teste de silhueta

Foi feito um teste de silhueta para verificar o melhor número de cluster para o algoritmo K-Means.

Etapa 4 - Interpretar e Analisar os Dados

Foi feito uma análise e uma interpretação de todos os dados gerados e foi descrito os padrões encontrados.

3.2 Aplicação da Abordagem

Nesta parte, foi apresentada a descrição de cada etapa e os produtos/atividades realizadas.

Etapa 1 - Análise preliminar dos dados do *Moodle* da SEDIS.

Essa etapa foi divida em quatro fases:

Fase 1: Estudo das tabelas importantes do *Moodle*

Foram analisadas todas as 324 tabelas do banco de dados do *Moodle* da UFRN e observado como o banco estava estruturado e organizado para a partir disso extrair as informações necessárias.

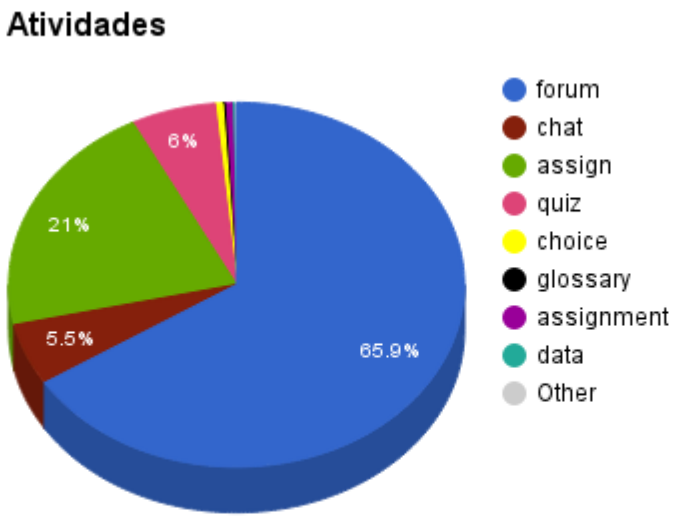
É importante ressaltar que a coleta dos dados no *Moodle* ocorre a medida que os alunos vão interagindo com o sistema, informações referente as ações dos usuários vão sendo armazenadas em uma tabela de log que grava o *id* do usuário, o horário e data da ação, o endereço *IP*, o modulo que foi acessado e a sua ação (adicionar, deletar, visualizar, etc).

A base de dados utilizada possui informações coletadas desde 2006 até 2013 de todas as turmas criadas durante esse período.

Fase 2: Análise das atividades e recursos mais utilizados

Foi realizado uma análise preliminar do banco de dados para todas as disciplinas online, do cursos a distância da UFRN e foi identificado quais as atividades que são mais utilizados pelos professores. Por exemplo, observa-se na Figura 7 que o *forum* (fórum) é mais utilizado com 65,9% (sessenta e cinco, e nove), seguido do *assign* (tarefa) com 21% (vinte e um) e do *quiz* (questionário) com 6% (seis), sendo que o *chat* (bate papo) aparece em 4ª lugar com 5,5% (cinco e cinco).

Figura 7 - Distribuição de tipos de atividades nas disciplinas do AVA - UAB/UFRN.

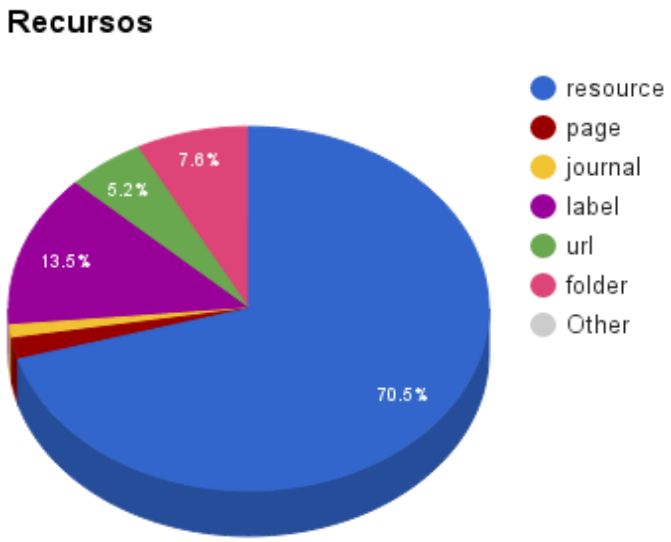


Fonte: autoria própria

Foi também realizado uma análise do banco de dados para identificar quais os recursos mais utilizados pelos professores. Por exemplo, observa-se na Fig. 8 que o *resource* (arquivo) é mais utilizado

com 70,5% (setenta, e cinco), seguido do *label* (rótulo) com 13,5% (treze, e cinco) e do *folder* (pasta) com 7,6% (sete, e seis), sendo que a URL aparece em 4º lugar com 5,2% (cinco, e dois) e *page* (página) em 5º lugar com 1,9% (um, e nove).

Figura 8 - Distribuição de tipos de recursos nas disciplinas do AVA - UAB/UFRN.



Fonte: autoria própria

Fase 3: Seleção da turma teste

A seleção da turma teste seguiu os seguintes critérios:

- A disciplina do *Moodle* que continha maior quantidade de ferramentas em uso;
- Número de alunos superior a 100;
- Disponibilidade da nota final do aluno.

A turma tinha 497 estudantes matriculados e foi ofertada no semestre 2013.1, o nome da disciplina não será mencionado por

questões de sigilo.

A turma selecionada deveria possibilitar a validação da metodologia permitindo que a partir dos resultados encontrados fossem realizadas correções e ajustes para reaplicação em novos estudos.

Fase 4: Seleção das tabelas e dos atributos

Após a análise, foram escolhidas as tabelas que possuíam os atributos necessários. A Tab. 1 mostra uma síntese das tabelas usadas e quais informações elas contêm.

Tabela. 1 - Tabelas do banco de dados utilizadas

Tabela do banco	**Informação**
mdlacademico_log	Todos os dados das ações do usuário no sistema.
mdlacademico_course_modules	As atividades e recursos que a disciplina possui.
mdlacademico_grade_grades	Nota dos alunos nas disciplinas e nos módulos.
mdlacademico_role_assignments	Qual o papel do usuário no sistema em um determinado contexto.

Fonte: autoria própria

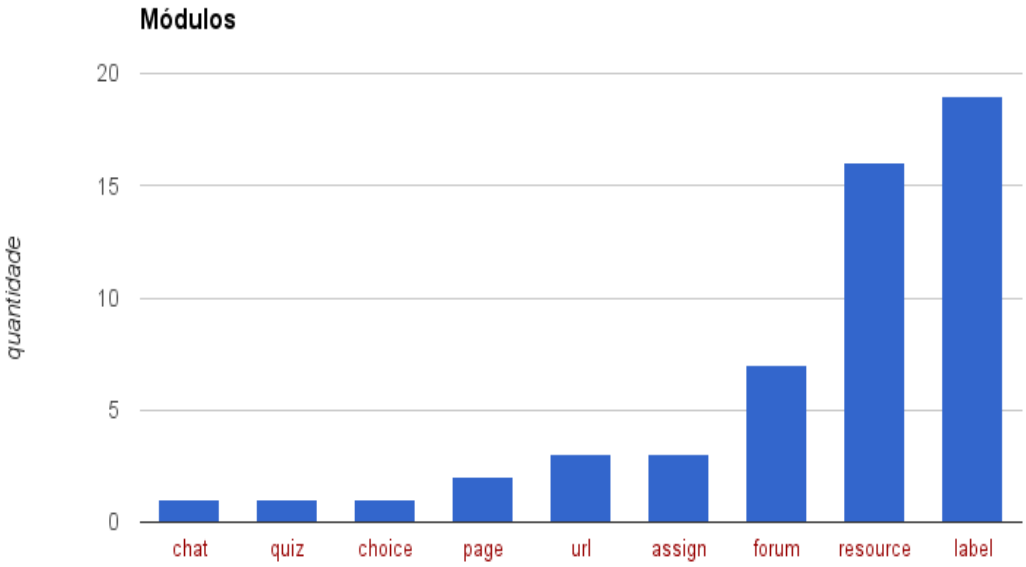
A tabela mdlacademico_log, possui todas as ações do usuário na disciplina teste, e com isso ela foi utilizada para capturar as informações quantitativas das ações do aluno. Os campos utilizados dessa tabela foram: *userid*, *course*, *module* e *action*.

Na tabela mdlacademico_course_modules foi possível verificar quais os módulos e a quantidade de vezes que estão sendo utilizados

na disciplina teste. Os campos utilizados nessa tabela foram: *course*, *module* e *instance*.

Observa-se que a Figura 9 demonstra o quantitativo de uso dos recursos e atividades do *Moodle* por parte do docente referente a turma teste.

Figura 9 - Módulos mais utilizados na turma teste



Fonte: autoria própria

Como demonstra a Figura 9, o *label* é o recurso mais utilizado com 19 inserções, seguido do *resource* com 16 e em seguida o *forum* com 7 inserções, *url* e *assign* com 3, *page* com 2 e *chat*, *quiz* e *choice* foram usados uma única vez.

Na tabela mdlacademico_grade_grades fica armazenado a nota obtida por cada aluno em cada atividade avaliativa de cada disciplina e, o somatório dessas atividades. Vale ressaltar que, esse somatório é referente somente as atividades apuradas na plataforma. Entretanto, a nota final do estudante refere-se as atividades do *Moodle* mais a nota

da prova presencial, sendo que a nota da prova presencial foi extraída de uma planilha complementar. Os atributos utilizados dessa tabela foram: *itemid*, *userid* e *finalgrade*.

A tabela mdlacademico_role_assignments contém informações sobre qual o papel do usuário em um determinado contexto. Essa tabela foi utilizada para identificar na turma teste todos os usuários que possuíam papel de estudante. Os campos utilizados dessa tabela foram: *roleid*, *contextid* e *userid* que refletem, respectivamente o identificador do papel que no caso é estudante, o determinado contexto que é a disciplina teste e o identificador do usuário.

Etapa 2 – Pré-processamento

Essa etapa foi dividida em 5 fases, as fases de 1 a 3 utilizaram a ferramenta pgAdmin3 (http://www.pgadmin.org/), e a de 4 a 5 foi implementada em Python utilizando uma biblioteca de aprendizado de máquina (Machine Learning) Scikit-Learn (http://scikit-learn.org/).

Fase 1: Criação da tabela de sumarização

Devido ao fato das informações estarem espalhadas em várias tabelas foi necessário criar uma tabela de sumarização que contém todas as informações necessárias. Cada atributo da tabela de sumarização com sua respectiva descrição é mostrado na Tabela 2, essas informações são referentes a cada aluno na turma teste.

Tabela 2 - Tabela de sumarização

Atributo	**Descrição**
disciplina	Id da turma
n_cliques	Número total de cliques na disciplina
total_quiz	Porcentagem de *quiz* realizado
total_assign	Porcentagem de *assign* realizado
total_resource	Porcentagem de *resource* visualizado
total_page	Porcentagem de *page* visualizado
total_url	Porcentagem de *url* visualizado
total_forum	Porcentagem de *forum* visualizado
poste_forum	Número total de postes realizados em fóruns
disc_forum	Número de discussões abertas em fóruns
total_chat	Porcentagem de chat visualizado
conversa_chat	Número de conversas em chats.
primeiro_acesso	Quanto tempo, após a turma inicializar, o aluno acessou.
nota_atividades	Nota final obtida de todas as atividades.
nota_final	Nota final, contendo as atividades e a prova presencial.

Fonte: autoria própria

Fase 2: Limpeza dos dados

Os atributos total_quiz, total_assign, total_resource, total_page, total_url, total_forum e total_chat, seguiram os seguintes padrões de limpeza de dados:

1. Limpeza de dados inconsistente: confere o *id* do módulo em que a informação esta sendo capturada e verifica na tabela mdlacademico_course_modules se este *id* realmente existe.
2. Padronização das unidades: transforma o atributo em

porcentagem, dividindo o número de ocorrências do atributo encontrado na tabela de log pelo número total de ocorrências do módulo e o resultado é multiplicado por 100. Por exemplo, um aluno realizou 2 *assign* e a disciplina tem um total de 4 *assign*, com isso o aluno fica com o atributo total_assign no valor de: 2 / 4 * 100, ou seja, 50% (cinquenta).

Fase 3: Seleção dos atributos significativos

Foi utilizada a correlação de Pearson e Spearman entre cada atributo selecionado e a nota final, com a finalidade de identificar quais atributos tiveram maior impacto na nota e, partir disso, remover os atributos que não tinham muita correlação.

Somente os atributos médios (coeficiente entre 0,30 e 0,50) e grandes (coeficiente entre 0,50 e 1) identificados na correlação de Pearson ou de Spearman foram escolhidos. A Tabela 3 ilustra os atributos selecionados. Nela pode ser observado que todos os atributos são médios ou grandes, com exceção do atributo total_chat na correlação de Spearman. Entretanto como ele é médio para correlação de Pearson ele também foi selecionado. Também é possível observar que, o atributo total_assign é o que possui maior correlação com a nota, isso é explicado pelo fato dele compor 30% da nota final.

Tabela 3 - Atributos selecionados

Atributo	**Atributo**	**Pearson**	**Spearman**
notafinal	n_cliques	0.336	0.355
notafinal	total_quiz	0.542	0.425
notafinal	total_assign	0.814	0.585
notafinal	total_resource	0.501	0.338
notafinal	total_page	0.489	0.397
notafinal	total_url	0.572	0.470
notafinal	total_forum	0.614	0.383
notafinal	poste_forum	0.349	0.461
notafinal	total_chat	0.345	0.254

Fonte: autoria própria

A Tabela 4 mostra os atributos descartados, nela é possível verificar que disc_forum, conversa_chat e primeiro_acesso possuem uma correlação pequena, por isso não foram escolhidos.

Tabela 4 - Atributos descartados

Atributo	**Atributo**	**Pearson**	**Spearman**
notafinal	disc_forum	0.130	0.077
notafinal	conversa_chat	0.117	0.143
notafinal	primeiro_acesso	0.148	-0.262

Fonte: autoria própria

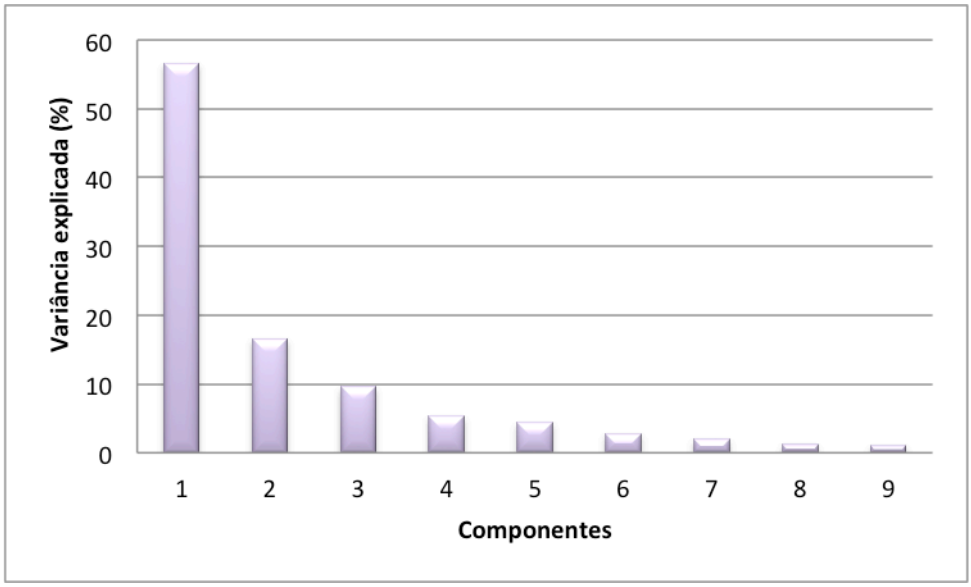
Fase 4: Normalização dos dados

Foi utilizada a normalização Min-Max em todos os atributos selecionados. Essa normalização foi a mais adequada devido a natureza dos dados.

Fase 5: Redução dos dados

Para reduzir a dimensionalidade dos dados foi utilizado o PCA. A tabela de atributos selecionados antes do PCA possuía 9 dimensões (n_cliques, total_quiz, total_assign, total_resource, total_page, total_url, total_forum, poste_forum e total_chat). Essa tabela foi reduzida para 2 dimensões para poder obter uma maior precisão e um melhor aproveitamento do algoritmo de clusterização utilizado na etapa de mineração de dados. A redução para 2 foi escolhida após ter sido calculado a percentagem de variância explicada para cada componente principal, o que pode ser visualizado na Figura 10.

Figura 10 – Variância explicada de cada componente



Fonte: autoria própria

Na Fig. 10 é possível observar que o componente 1 e 2 possuem a maior proporção da variância, somando os dois da um total de 73,3%. Devido a isso é possível reduzir para 2 dimensões sem perdas significativas.

Etapa 3 – Mineração de dados

Essa etapa foi implementada em Python utilizando a biblioteca *Scikit-Learn* e foi dividida em 2 fases:

Fase 1: Utilização do algoritmo *K-Means*

Nessa fase foi utilizado o algoritmo *K-Means* para gerar os clusters de alunos, o algoritmo recebeu como entrada os seguintes atributos:

- Os dados gerados após o uso do PCA;
- O número de clusters, que no caso foi escolhido quatro;
- A forma de inicialização dos centros dos clusters. Para isso foi escolhida a técnica k-means++, que força os centros a serem distantes um dos outros.

Fase 2: Teste de silhueta

O teste de silhueta buscou encontrar o melhor número K para execução do K-Means, o teste foi feito com os valores de K entre 2 até 7, sendo comprovado que o melhor valor para o K-Means foi 4. A Fig. 11 mostra o valor do coeficiente de silhueta para os valores de K.

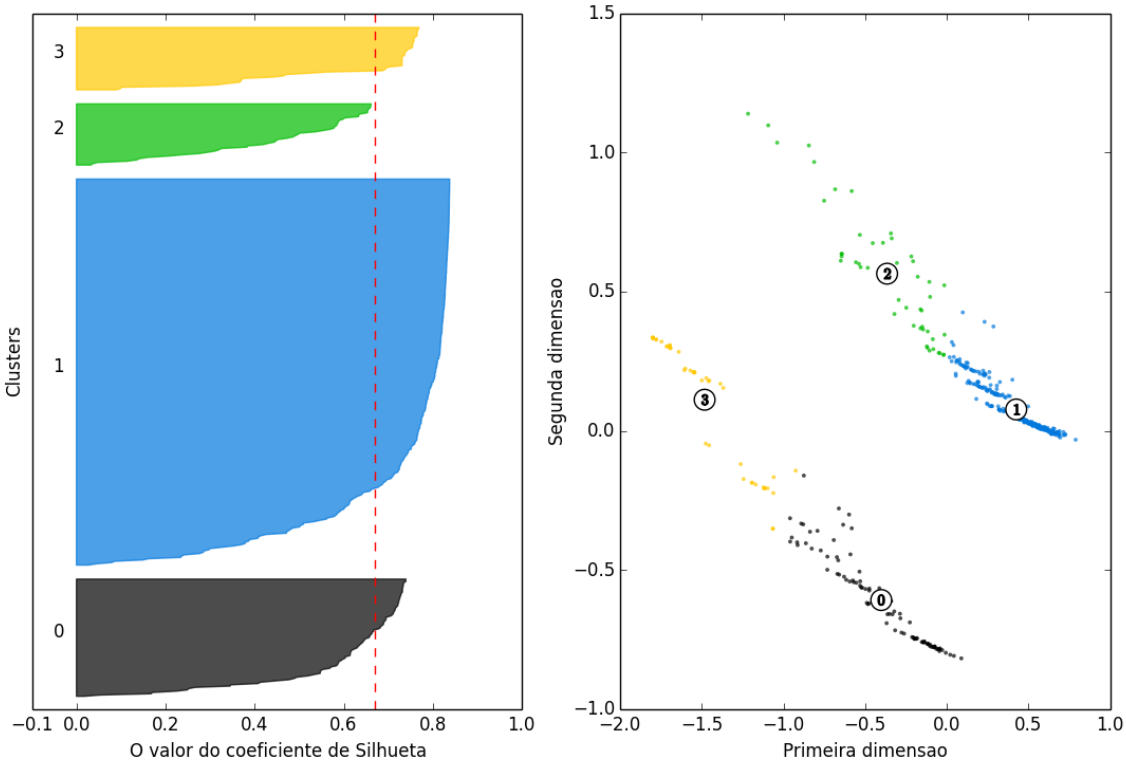
Figura 11 - Coeficiente de silhueta para diferentes K

Fonte: autoria própria

É possível verificar na Fig. 11 que o maior número do coeficiente de silhueta é 0,669 para K=4, seguido do valor 0,665 quando k igual 3.

A Figura 12 ilustra o resultado do teste para o K=4.

Figura 12 - Análise de Silhueta para o K-Means com 4 Clusters



Fonte: autoria própria

Na Fig. 12 pode ser verificado o valor do coeficiente de silhueta para cada cluster e a média do coeficiente que foi 0.669, representada pela linha vermelha tracejada, de acordo com Smith (2008). Este valor pode significar que uma estrutura razoável foi formada. No segundo quadro da figura o eixo x representa a primeira dimensão gerada do PCA e o eixo y a segunda dimensão, nesse quadro pode ser observado a divisão dos clusters, a distância que cada um se encontra do outro e os centróides, representados pelos números dentro do círculo.

No Anexo B, encontra-se a análise do coeficiente de silhueta para os valores de K entre 2 até 3 e 5 até 7. Nele é possível observar que para K igual a 2, 3, 5, 6 e 7 existem clusters com valores do coeficiente de silhueta negativo, o que torna o K=4 apresentando na Fig.12 a melhor escolha.

4 Interpretação e Análise dos Dados

Nessa etapa foi analisado cada um dos clusters gerados, e pode ser identificado os perfis e padrões de participação no AVA para cada grupo. Para melhor entendimento dos clusters é apresentada a Tabela 5, que demonstra a média e o desvio padrão de cada atributo encontrado em cada grupo.

Os valores do *quiz, assign, resource, page, url, chat e forum* foram normalizados com a normalização Min-Max, cliques e postes não foram normalizados.

Tabela 5 - Média e desvio padrão dos atributos por cluster

	quiz	***assign***	***resource***	***page***	***url***	***chat***	***forum***	**cliques**	**postes**
Cluster 1	0,31± 0,46	0,71± 0,22	0,19± 0,25	0,42± 0,38	0,15± 0,17	0,19± 0,13	0,17± 0,16	32,7± 29,3	0,2± 0,5
Cluster 2	0,61± 0,48	0,59± 0,38	0,30± 0,2	0,78± 0,26	0,36± 0,25	100± 0,0	0,55± 0,22	140,0± 70,6	1,5± 2,1
Cluster 3	0,94± 0,22	0,92± 0,19	0,41± 0,24	0,90± 0,21	0,65± 0,36	0,0± 0,0	0,64± 0,23	192,3± 102	2,9± 2,8
Cluster 4	0,93± 0,08	0,95± 0,20	0,51± 0,21	97,0± 0,16	0,87± 0,22	100± 0,0	0,85± 0,14	424± 248	5,2± 5,1

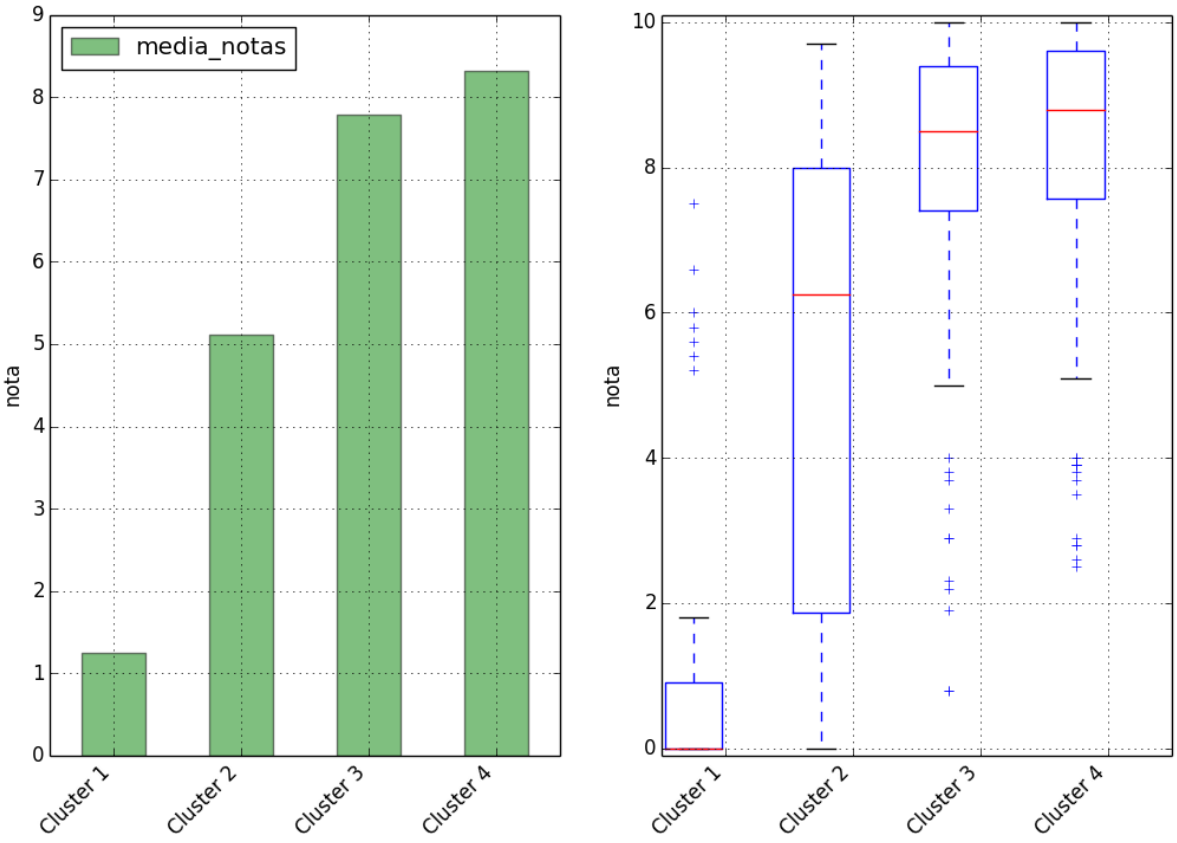
Fonte: autoria própria

É possível verificar na Tabela 5 a diferença de um cluster para o outro, sendo o cluster 4 o que possui as melhores médias. Este grupo tem um desvio padrão alto para o atributo cliques, devido ao fato de existirem alunos nele com uma quantidade extremamente alta de cliques e alguns com uma quantidade baixa. Após esse vem o cluster 3 com médias um pouco inferiores e um desvio padrão também alto para o atributo cliques. Em seguida vem o grupo 2 com médias mais baixas que o 3 e apenas a média do chat superior, esse grupo possui um desvio padrão alto para o atributo *quiz* e *assign.* E por último vem o cluster 1 com as piores médias e desvio padrão

alto para os atributos *quiz* e *page*. É importante ressaltar que o desvio padrão deu valores alto para alguns atributos devido ao fato desses atributos terem uma variação muito grande em seus valores, como por exemplo, o *quiz* possui apenas valores 0 ou 100, o que acaba elevando muito o valor do desvio padrão.

A Figura 13 ilustra as notas finais de cada cluster, vale ressaltar que essa nota é composta por 4 pontos de atividades sendo 3 pontos referente aos 3 *assign* e 1 ponto do *quiz,* mais 6 pontos da prova presencial.

Figura 13 - Distribuição das notas por clusters



Fonte: autoria própria

No primeiro quadro da Figura 13 é demonstrado a média das notas finais de cada cluster, é possível observa que a média das notas aumenta de um cluster para o outro, sendo que o primeiro grupo é o que possui a pior média e o ultimo grupo a melhor.

No segundo quadro da Fig. 13, as notas finais são representas por um gráfico de caixas. Nesse tipo de gráfico as caixas (retângulos azuis) contém 50% dos dados, o limite superior da caixa indica o percentil de 75% dos dados e o limite inferior da caixa indica o percentil de 25%, a linha vermelha indica a mediana, os extremos do gráfico indicam os valores mínimo e máximo, porém, podem existir os pontos fora do extremo, que são valores discrepantes (outliers) representados pelo (+).

Pode-se concluir que, após analisar o gráfico caixa, o cluster 1 possui notas muito baixas e alguns poucos alunos com notas mais altas, sua mediana é 0. Já o cluster 2 tem uma mediana de 6,2 e uma dispersão enorme, visto que sua caixa se estende dos valores de 2 até 8. O cluster 3 tem uma mediana de 8,4 e a maioria dos alunos possuem nota alta com exceção de alguns poucos *outliers* com notas baixo de 4. O cluster 4 tem comportamento parecido com o cluster 3, com uma mediana de 8,5 e sua dispersão não é muito grande, podendo ser observado que todos os alunos retirando-se os *outliers* obtiveram aprovação.

A Figura 14, mostra a preditibilidade de cada cluster na obtenção de aprovação por parte dos alunos na disciplina.

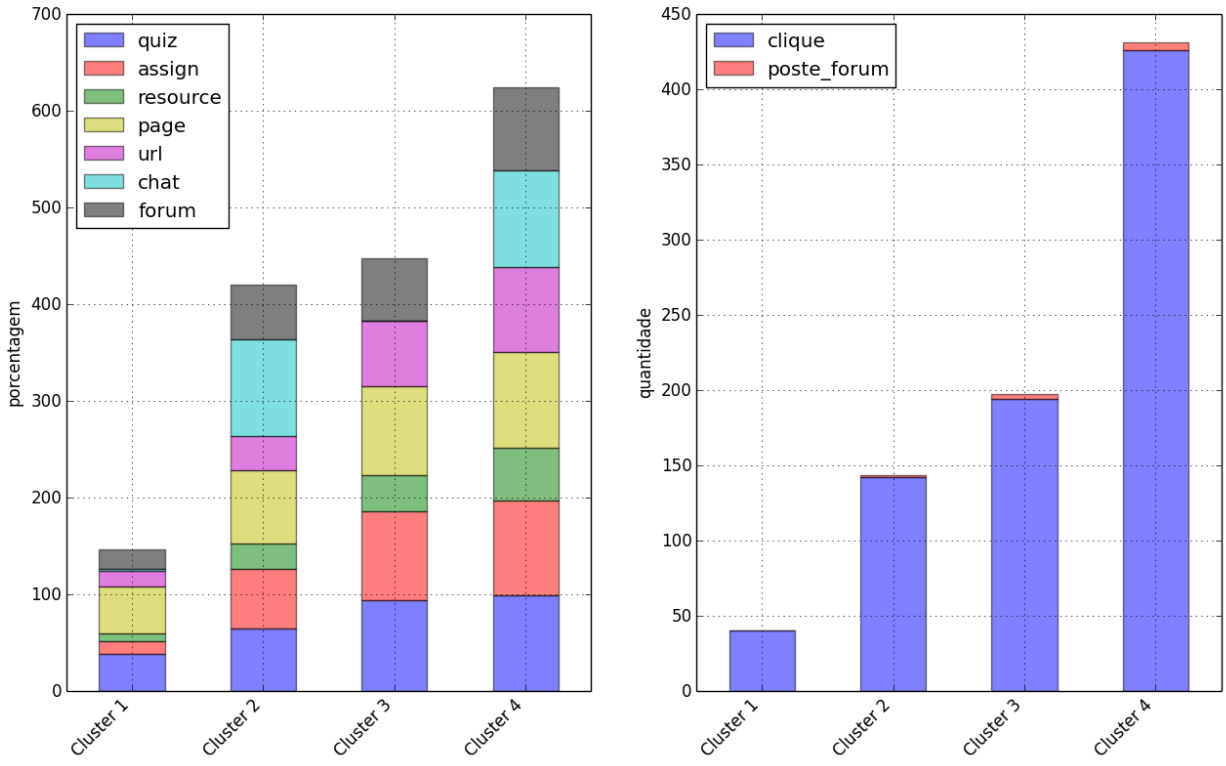
Figura 14 - Preditibilidade de aprovação de cada cluster

Fonte: autoria própria.

Pode ser observado na Figura 14 que o cluster 1 tem as menores chances de aprovação com 15,7% (quinze, e sete), seguido do cluster 2 com 55,8% (cinquenta e cinco, e oito) de chance e, logo, em seguida, o cluster 3 com 86,1% (oitenta e seis, e um). Por último, o cluster 4 com 95%(noventa e cinco), sendo este o grupo que possui as maiores chances de obter um sucesso na disciplina teste.

A Figura 15, assim como a Tabela 5 ilustra a média de cada atributo em cada cluster.

Figura 15 - Média dos atributos selecionados por cada cluster



Fonte: autoria própria

No primeiro quadro da Figura 15, o eixo x representa cada um dos 4 cluster. No eixo y encontra-se a média de cada um dos atributos selecionados na pesquisa. As cores da legenda representam os atributos que formam cada cluster. Os atributos foram calculados em porcentagem como já explicado na etapa de limpeza de dados.

No segundo quadro o eixo x representa cada um dos 4 clusters e o eixo y representa a quantidade de cliques e postes no fórum.

Em síntese, após a análise dos dados, podemos apontar a existência de 4 tipos de perfis e padrões de participação no AVA, exemplificados a seguir:

Cluster 4: estudantes ativos, são alunos que possuem uma alta participação na plataforma, realizam quase 100% das atividades, visualizam um grande número de recursos, possuem grande quantidade de cliques no ambiente e vários postes em fóruns. Os estudantes desse grupo possuem

95% de chance de aprovação. Os alunos desse grupo que não obtiveram aprovação foram os que não realizaram a prova final.

Cluster 3: estudantes medianos, são cursistas que possuem uma participação moderada no ambiente, realizando boa parte das atividades, visualizando parcialmente os recursos, acessando a plataforma regularmente e realizando alguns postes em fóruns. Esses estudantes não visualizam o chat, mas, possuem um bom rendimento no final, eles possuem 86,1% de chance de aprovação. Esse grupo mostra que a participação no *chat* não está relacionada a aprovação e que nem sempre é necessário ter um acesso excessivo a plataforma para obter uma nota boa, já que esse grupo obteve uma média de cliques de 192 o que equivale a menos da metade de cliques do grupo 4.

Cluster 2: estudantes inconstantes, são alunos irregulares que de vez em quando acessam o ambiente, realizam um pouco mais da metade das atividades, acessam pouco os recursos e realizam poucos postes em fóruns. São estudantes que visualizam o chat e podem ou não serem aprovados, suas chances são de 55,8% de aprovação.

Cluster 1: estudantes ausentes, são cursistas pouco ativos que praticamente não acessam a plataforma, realizam bem menos do que a metade das atividades, quase não visualizam os recursos disponíveis e têm uma quantidade baixíssima de postes em fóruns. As suas chances de aprovação são de 15,7%. Os alunos que obtiveram aprovação nesse grupo foram estudantes que realizaram a prova final que valia 6 pontos e obtiveram uma nota boa e com isso conseguiram a aprovação.

Os perfis encontrados em comparação com os achados na literatura (HRASTINSKI, 2009; NISTOR e NEUBAUER, 2010; RODRIGUES et al, 2013) apresentam similaridades nos seguintes aspectos:

- A maior participação encontrada foi no cluster 4 sendo esse grupo é o que possui maior média de acesso a plataforma, o que assimila com o nível 1 do estudo de Hrastinski (2009) em que participação é igualada ao número de vezes que o aluno acessa o ambiente. Esse cluster

também condiz com o grupo 1 do estudo de Nistor e Neubauer (2010) que afirma que esses alunos participam em todas as modalidades e enquadra no grupo 1 do estudo de Rodrigues et al (2013) que alega que esses estudantes participam durante todo o curso.

- O nível 4 do estudo de Hrastinski (2009) afirma que alunos que visualizam e postam muitas mensagens no fórum são mais ativos que os outros, isso foi encontrado no Cluster 4.
- O cluster 3 possui uma participação mediana o que assimila ao grupo 2 do estudo de Rodrigues et al (2013) que são alunos medianos que procuram a aprovação.
- O cluster 2 é formado por alunos inconstantes o que condiz com o grupo 4 do estudo de Rodrigues et al (2013) em que os alunos não possuem uma participação regular.
- O cluster 1 é formado por alunos pouco ativos o que assimila ao grupo 5 do estudo de Rodrigues et al (2013) que é formado por alunos com pouca interação no ambiente e também condiz com o grupo 4 do estudo de Nistor e Neubauer (2010) que afirma que esse grupo é formado por estudantes evadidos.

5 Conclusão e Trabalhos Futuros

Essa pesquisa buscou ampliar a visão sobre os perfis de participação dos estudantes no ambiente virtual de aprendizagem a partir de uma organização didático pedagógica planejada e executada pelo professor. O objetivo principal foi aplicar a metodologia de mineração de dados para o levantamento dos de perfis e padrões de participação de alunos em disciplina de curso superior a distância na Universidade Federal do Rio Grande do Norte.

Neste sentido, pode-se afirmar que o objetivo foi cumprido. O estudo aplicou a metodologia de mineração de dados segundo o KDD o que permitiu identificar os perfis e padrões de participação dos alunos em cursos a distância, com base nos registros de interação do *Moodle* na disciplina selecionada. Ficou evidenciado que a alta participação no ambiente está relacionada com a aprovação sendo possível verificar que o aluno do grupo que apresentou o melhor perfil de participação teve maiores chances de obter sucesso na disciplina, apesar, de também termos encontrado alunos com baixa participação, que obtiveram aprovação.

O estudo contribui para a área de educação porque possibilita ao professor ter acesso aos dados de desempenho dos alunos no decorrer da disciplina. Para tanto, a metodologia aplicada deverá ser utilizada como forma de monitoramento da turma durante a oferta da disciplina porque pode facilitar a identificação dos estudantes que se encontram em situação de risco no processo de aprendizagem. Inclusive, pode ser usado para medidas preventivas que evitem altas taxas de reprovação ou trancamento da disciplina por parte dos alunos que sentem que irão reprovar. E também deve incentivar os professores a utilizar mais os recursos e atividades que o *Moodle* disponibiliza. Desta forma, o estudo demonstra o potencial do uso da metodologia de mineração de dados para gerar antecipadamente relatórios técnicos e acadêmicos válidos para gestores e professores que pretendem atuar na prevenção e na diminuição da evasão em cursos online.

A metodologia aplicada foi de extrema importância para geração dos resultados, sendo possível observar que todas as etapas do KDD têm sua relevância e devem ser seguidas na ordem, porém, sempre que necessário deve-se retornar a uma ou mais etapa para realizar reajustes. A inclusão da verificação do número ótimo K para o K-Means na etapa de mineração tem grande importância, pois comprova o número de clusters utilizado, este tipo de verificação não foi encontrada em outros estudados que utilizaram o K-Means na área de EDM.

A maioria das dificuldades encontradas no estudo se referem a estrutura do banco de dados, a falta de informações consistentes e desagregadas. Por exemplo, disciplinas que as notas finais não estavam armazenadas no *Moodle* e, sim, em outro sistema utilizado pela Universidade, que não permite o acesso e, consequentemente, inviabiliza a mineração de dados de determinadas disciplinas.

Recomenda-se que a metodologia seja testada em outros cursos, comparando, por exemplo, os tipos de recursos e atividades mais usados por diferentes professores e os tipos de recursos e atividades que propiciam maior participação e aprovação. Também, considera-se importante que seja realizada pesquisa qualitativa sobre as percepções e níveis de satisfação dos alunos frente ao uso das atividades e recursos do *Moodle*. E por último, recomenda-se que seja feita a implementação da metodologia do estudo na plataforma do *Moodle*.

Referências

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, 1 (1), 3-17, 2009. Disponivel em: <http://educationaldatamining.org/JEDM/index.php/JEDM/article/view/8/2>

BAKER, R. S. J. d.; ISOTANI, S.; CARVALHO, A. de. Mineração de dados educacionais: Oportunidades para o Brasil. Revista Brasileira de Informática na Educação, 2011.

BARROS, M. G.; CARVALHO, A. B. G. As concepções de interatividade nos ambientes virtuais de aprendizagem. Campina Grande: EDUEPB , 2011.. Disponivel em: <http://books.scielo.org/id/6pdyn/pdf/sousa-9788578791247-09.pdf>

BRASIL. MEC/INEP. Censo da educação superior 2012. Resumo técnico. – Brasília: Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira, 2014. Disponível em: http://download.inep.gov.br/download/superior/censo/2012/resumo_tecnico_censo_educacao_superior_2012.pdf. Acesso em: 02 de junho de 2015.

Brito, Marcelo. Aspectos teóricos da mineração de dados e aplicação das regras de classificação para apoiar o comércio, 2012. Disponível em: <http://www.devmedia.com.br/aspectos-teoricos-da.-mineracao-de-dados-e-aplicacao-das-regras-de-classificacao-para-apoiar-o-comercio/25429#comentariosArtigo>

C. ROMERO, S. VENTURA. Educational data mining: A survey from 1995 to 2005. Expert Systems with Applications, 33 (1), pg. 135–146, 2007.

CARVALHO, L. A. V. Datamining: a Mineração de Dados no Marketing, Medicina, Economia, Engenharia e Administração. São Paulo: Érica, 2001.

COHEN, Jacob. Statistical power analysis for the behavioral sciences. Hillsdale, NJ, Erlbaum 1988.

CONTI, F. Mineração de dados no moodle: análise de prazos de entrega de atividades. 2011. Dissertação de mestrado.

COSTA, E. , BAKER, R. S. J. D., AMORIN, L. MAGALHOES, J., MARINHO, T. Mineração de Dados Educacionais: Conceitos, Técnicas, Ferramentas e Aplicações. 2012. Disponível em: < http://www.br-ie.org/pub/index.php/pie/article/view/2341/2096>

FAYYAD, U. M., Piatetsky-Shapiro, G., and Smyth, P. Ad- vances in knowledge discovery and data mining. In Fayyad, U. M., Piatetsky-Shapiro, G., Smyth, P., and Uthurusamy, R., editors, A, chapter From data mining to knowledge discovery: an overview, pages 1–34. American Association for Artificial Intelligence, Menlo Park, CA, USA, 1996.

FINNEGAN, C.; MORRIS, L. V.; LEE, K. Differences by course discipline on student behavior, persistence, and achievement in online courses of undergraduate general education. Journal of College Student Retention: Research, Theory and Practice, 10 (1), p. 39-54, 2008.

FODOR, I.K: “A Survey of Dimension Reduction Techniques. LLNL Technical Report, UCRL-ID-148494”, pp.1-18, 2002.

HRASTINSKI, S. A theory of online learning as online participation. Computers & Education, v. 52, p. 78-82, 2009.

HRASTINSKI, S. What is online learner participation? A literature review. Computers & Education, v. 51, p. 1755–1765, 2008.

JAIN, K.A. Data Clustering: 50 Years Beyond K-Means. 2008. Disponível em <http://www.cs.utah.edu/~piyush/teaching/kmeans50yrs.pdf>

JALDEMARK, J., LINDBERG, J. O., & OLOFSSON, A. D. Sharing the distance or a distance shared: Social and individual aspects of participation in ICT-supported distance-based teacher education. In Chaib, M. & Svensson, A. K. (Eds.). ICT in teacher education: Challenging prospects. Jönköping: Jönköping University Press. 2006. (pp. 142–160).

MANLY, B.F.J. Multivariate statistical methods: a primer. 2nd ed., London, Chapman & Hall, 1994.

MARQUES, J. L. Q. Mineração de dados educacionais: um estudo de caso utilizando o ambiente virtual SENAI. Trabalho de conclusão de curso. 2014.

MEDEIROS, C. A. Extração de conhecimento em bases de dados espaciais: Algoritmo CHSMST+. 2014, Dissertação de mestrado.

MEDEIROS, C. J. F., COSTA, J. A. F. Uma Comparação de Métodos de Redução de Dimensionalidade Utilizando Índices de Preservação da Topologia. Pg.2, 2009.

Mödritscher, F., Andergassen, M., & Neumann, G. Dependencies between ELearning Usage Patterns and Learning Results. In Proceedings of the 13th International Conference on Knowledge Management and Knowledge Technologies (pp. 1–8). Graz, Austria: ACM Press, (2013). Disponivel em: < http://nm.wu-wien.ac.at/research/publications/b1004.pdf>

N. Karthikeyani Visalakshi and K. Thangavel. Impact of Normalization in Distributed K-Means Clustering. International Journal of Soft Computing, 4: 168-172, 2009 Disponivel em: <http://www.medwelljournals.com/fulltext/?doi=ijscomp.2009.168.172>

NISTOR, N., & NEUBAUER, K. From participation to dropout: Quantitative participation patterns in online university courses. Computers & Education, v. 55(2), p. 663-672, 2010.

PEÑA-AYALA, A. Education data mining: Applications and trends (p. 52). Newyork: Springer, 2014.

PIELOU, E.C. The interpretation of ecological data; a primer on classification and ordination. New York, Wiley, 1984.

Rajadhyax, N. Shirwaikar, R. Data Mining on Educational Domain, 2012. Disponivel em: <http://arxiv.org/pdf/1207.1535.pdf>

RAMOS, W. M.; MEDEIROS, L. A Universidade Aberta do Brasil: desafios da construção do ensino e aprendizagem em ambientes virtuais. In: Amaralina Souza; Leda Fiorintini; Maria Alexandra Rodrigues. (Org.). Comunidade de Trabalho e Aprendizagem Em rede (CTAR). 2ed.Brasilia: Editora UnB, 2009, v. 1, p. 3-260.

RODRIGUES, Daniel Röhe Salomon da Rosa; RAMOS, W. M.; MENDES TAVARES, C. Padrões de participação e estilos de Aprendizagem em cursos massivos online. In: II Congresso Ibero Americano de Estilos de Aprendizagem, Tecnologias e Inovações na Educação II CIEATIE, 2013, Brasília. Anais do Estilos de Aprendizagem, Tecnologias e Inovações na Educação, Brasília: Universidade de Brasília, 2013. v. 1.

ROMERO, C.; VENTURA, S.; GARCIA, H. Data mining in course management systems: Moodle case study and tutorial. Computers e Education, v. 51, n. 1, pg. 368 – 384, 2008. Disponível em: <http://www.sciencedirect.com/science/article/pii/S0360131507000590>.

SCHULTZ, Duane P.; SCHULTZ, Sydney Ellen. História da psicologia moderna. 16. ed. São Paulo: Cultrix, 1992.

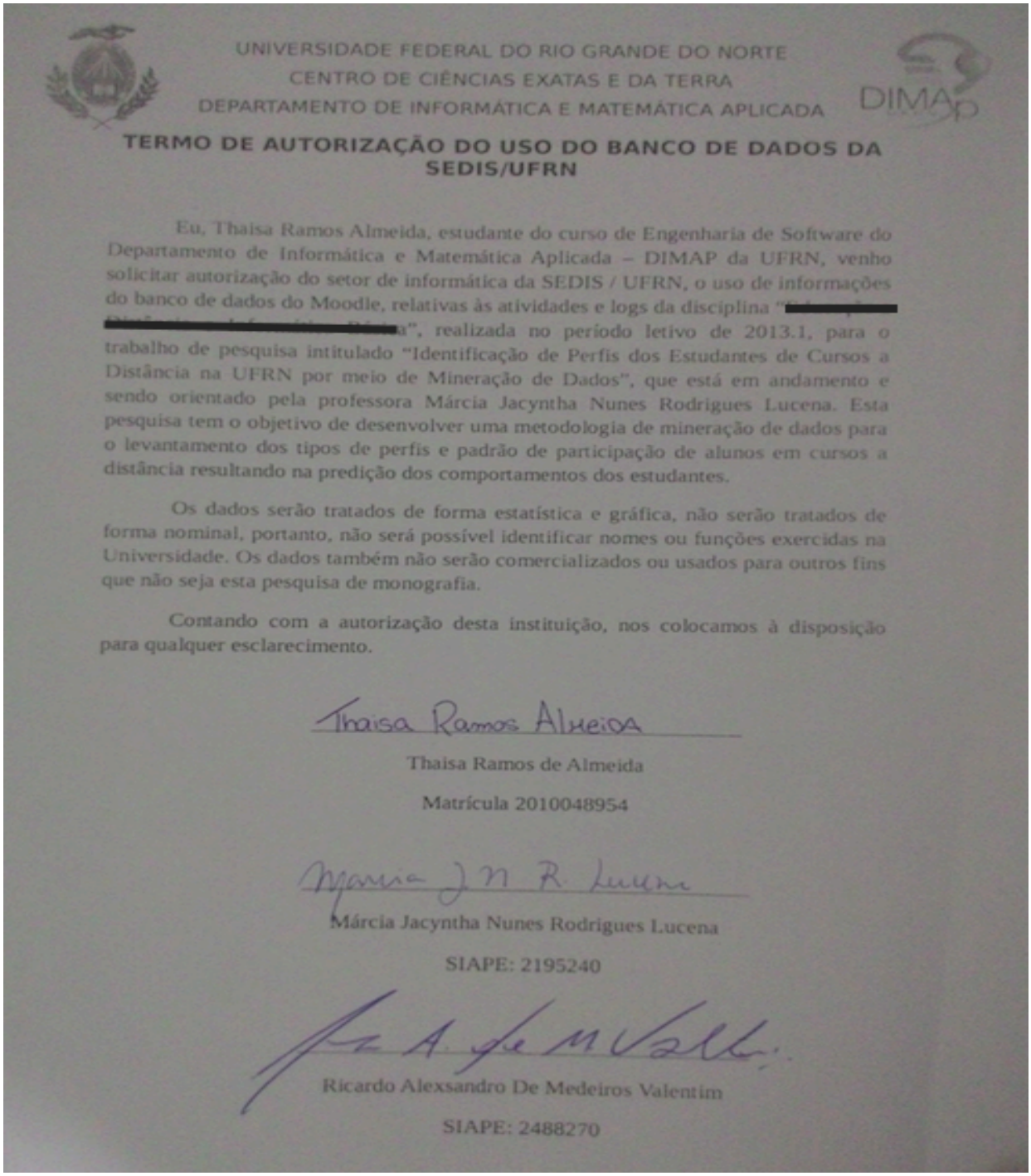
SENECHAL, A. C. L., Análise e pré-processamento de dados utilizando técnicas de mineração de dados educacionais para o Moodle. 2013. Trabalho de conclusão de curso.

SMITH, B. L. Automated Identification of Traffic Patterns. University of Virginia Center for Transportation Studies. 2008.

SPEARMAN, C. The proof and measurement of association between two things. Amer. J. Psychol, 15, 72-101, 1904.

TAN, P.-N., STEINBACH, M., and KUMAR, V. Introduction to Data Mining. Addison Wesley,2006.

ANEXO A – Termo de Autorização do Banco de Dados



UNIVERSIDADE FEDERAL DO RIO GRANDE DO NORTE
CENTRO DE CIÊNCIAS EXATAS E DA TERRA
DEPARTAMENTO DE INFORMÁTICA E MATEMÁTICA APLICADA

TERMO DE AUTORIZAÇÃO DO USO DO BANCO DE DADOS DA SEDIS/UFRN

Eu, Thaisa Ramos Almeida, estudante do curso de Engenharia de Software do Departamento de Informática e Matemática Aplicada – DIMAP da UFRN, venho solicitar autorização do setor de informática da SEDIS / UFRN, o uso de informações do banco de dados do Moodle, relativas às atividades e logs da disciplina "[redacted]", realizada no período letivo de 2013.1, para o trabalho de pesquisa intitulado "Identificação de Perfis dos Estudantes de Cursos a Distância na UFRN por meio de Mineração de Dados", que está em andamento e sendo orientado pela professora Márcia Jacyntha Nunes Rodrigues Lucena. Esta pesquisa tem o objetivo de desenvolver uma metodologia de mineração de dados para o levantamento dos tipos de perfis e padrão de participação de alunos em cursos a distância resultando na predição dos comportamentos dos estudantes.

Os dados serão tratados de forma estatística e gráfica, não serão tratados de forma nominal, portanto, não será possível identificar nomes ou funções exercidas na Universidade. Os dados também não serão comercializados ou usados para outros fins que não seja esta pesquisa de monografia.

Contando com a autorização desta instituição, nos colocamos à disposição para qualquer esclarecimento.

Thaisa Ramos Almeida

Thaisa Ramos de Almeida

Matrícula 2010048954

Márcia Jacyntha Nunes Rodrigues Lucena

SIAPE: 2195240

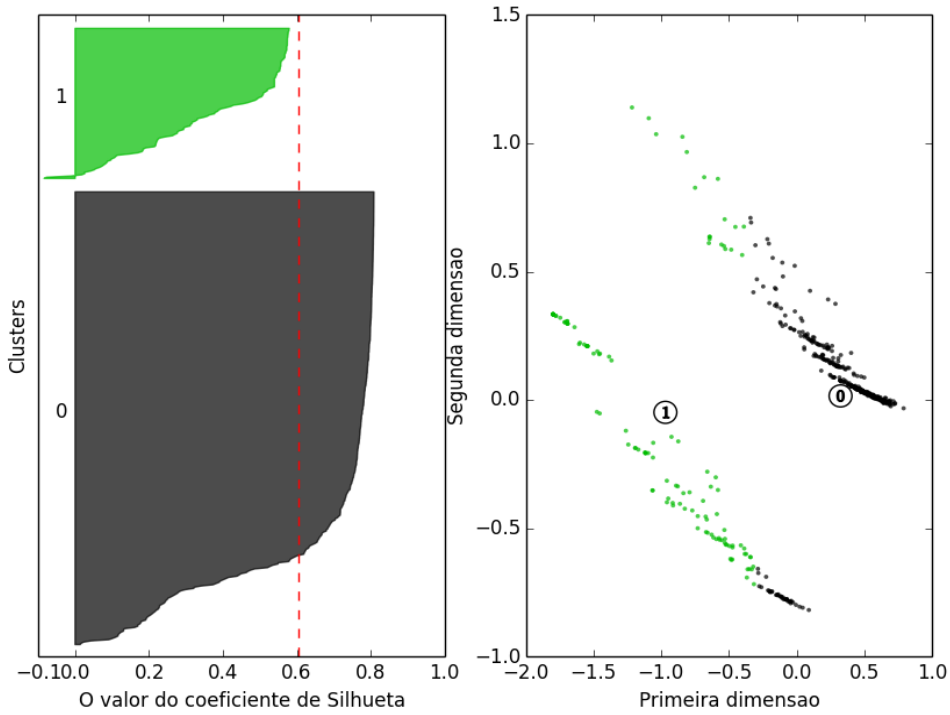
Ricardo Alexsandro De Medeiros Valentim

SIAPE: 2488270

ANEXO B – Análise do Coeficiente de Silhueta

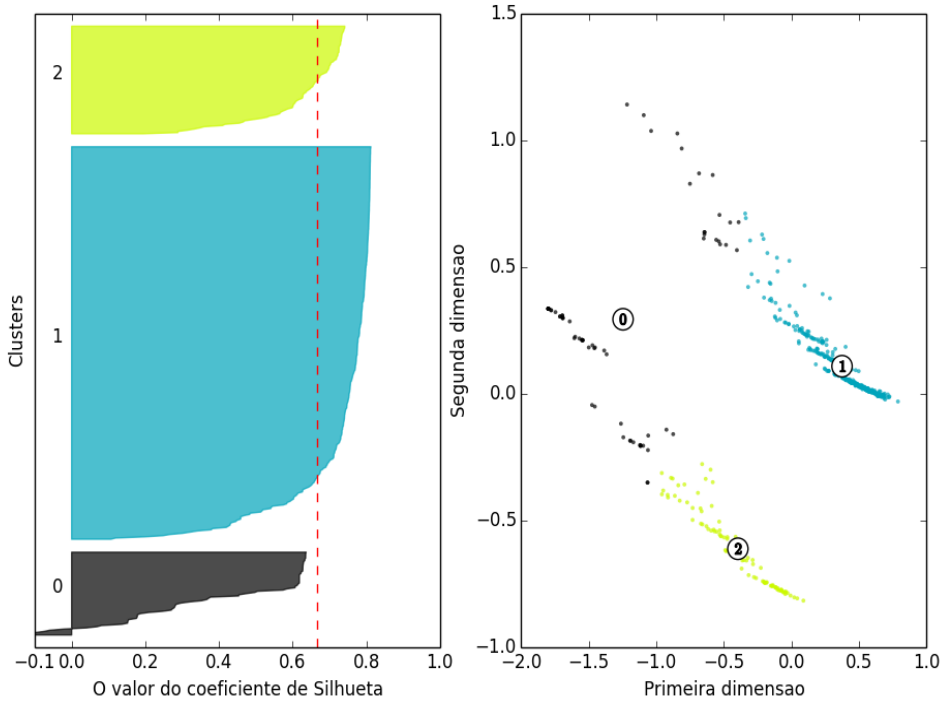
O Anexo B apresenta a análise do coeficiente de silhueta, a divisão de cada cluster e o centróides para os valores de K igual a 2, 3, 5, 6 e 7.

Figura 16 - K = 2



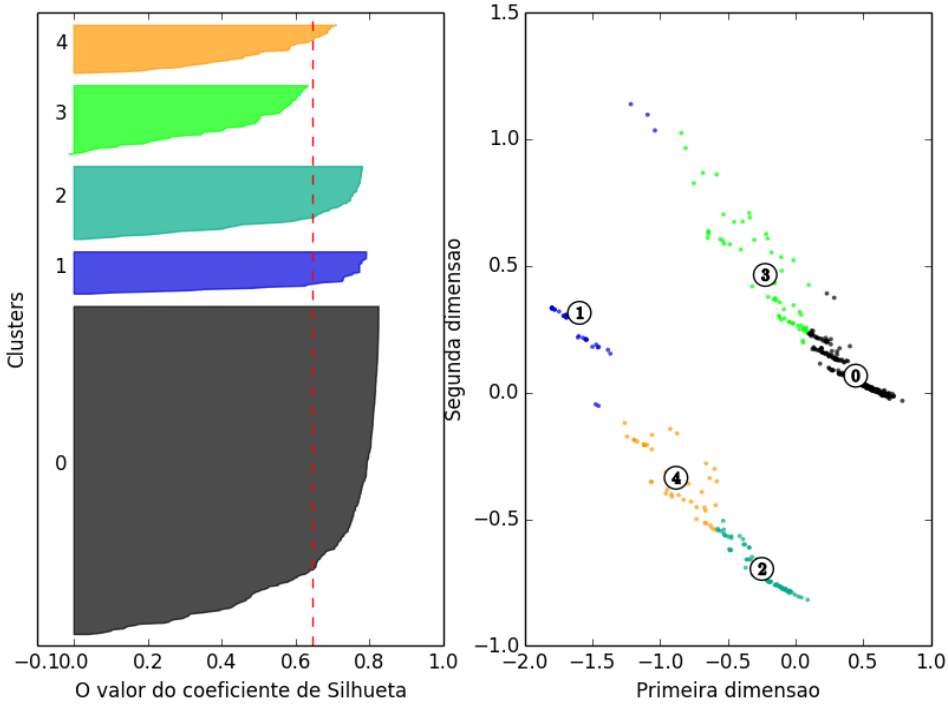
Fonte: autoria própria

Figura 17 - k = 3



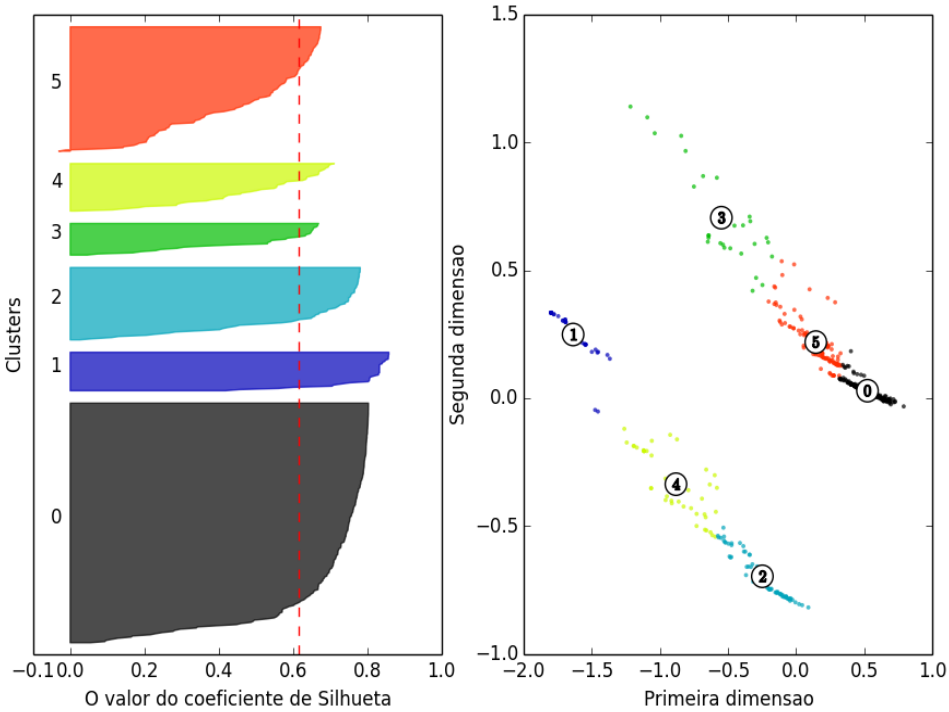
Fonte: autoria própria

Figura 18 - K = 5



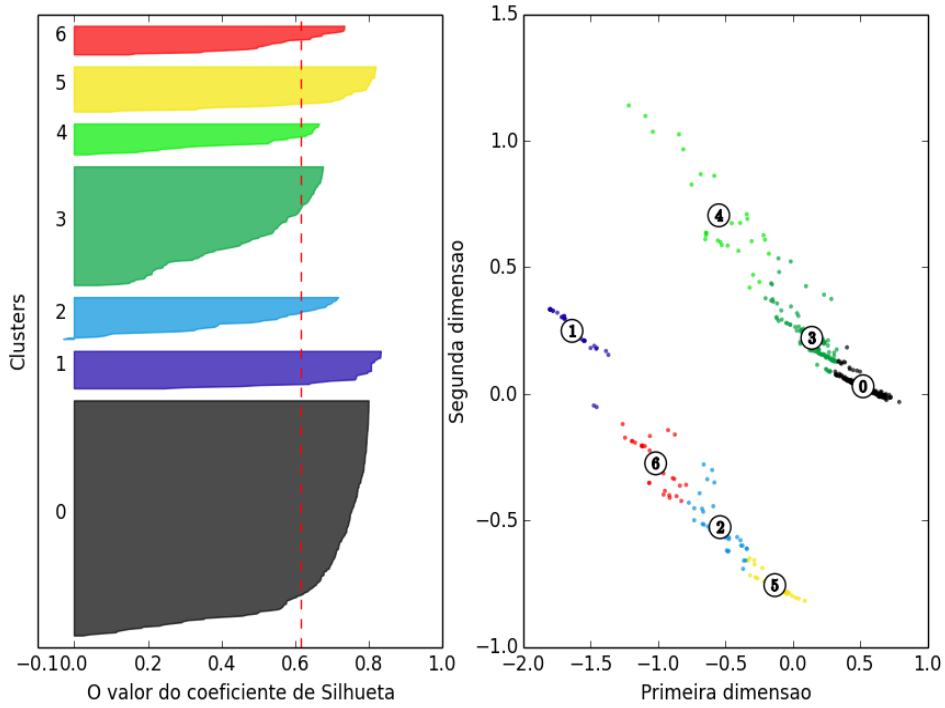
Fonte: autoria própria

Figura 19 - k = 6



Fonte: autoria própria

Figura 20 - k = 7



Fonte: autoria própria