UNIVERSIDADE FEDERAL DO PARANÁ

MATHEUS RENAN BATISTA

PREVISÃO DE DESEMPENHO ACADÊMICO UTILIZANDO MODELOS OCULTOS DE MARKOV

CURITIBA

2018

MATHEUS RENAN BATISTA

PREVISÃO DE DESEMPENHO ACADÊMICO UTILIZANDO MODELOS OCULTOS DE MARKOV

Trabalho de conclusão de curso apresentado ao curso de graduação de Engenharia Elétrica com Ênfase em Sistemas Eletrônicos Embarcados da Universidade Federal do Paraná.
Orientador: Dr. Carlos Marcelo Pedroso

CURITIBA
2018

TERMO DE APROVAÇÃO

MATHEUS RENAN BATISTA

PREVISÃO DE DESEMPENHO ACADÊMICO UTILIZANDO MODELOS OCULTOS DE MARKOV

Trabalho de Conclusão de Curso apresentado ao Curso de Engenharia Elétrica com Ênfase em Sistemas Eletrônicos Embarcados da Universidade Federal do Paraná como requisito à obtenção do título de Engenheiro Eletricista, pela seguinte banca avaliadora:

_____

Prof. Dr. Carlos Marcelo Pedroso
Orientador – Departamento de Engenharia Elétrica, UFPR

_____

Prof. Ph.D. Eduardo Gonçalves de Lima
Departamento de Engenharia Elétrica, UFPR

_____

Prof. Ph. D. John Jairo Villarejo Mayor
Departamento de Engenharia Elétrica, UFPR

Curitiba, 04 de dezembro de 2018

AGRADECIMENTOS

Agradeço primeiramente ao meu professor orientador Carlos Marcelo Pedroso, por ter me trazido a oportunidade de trabalhar neste projeto e por toda a tutela durante a elaboração do mesmo. Agradeço também aos outros professores pelo conhecimento compartilhado, pela dedicação em sala de aula, pelas conversas nos intervalos de aula e nos corredores da universidade.

Por fim, agraço à minha família pelo constante apoio durante essa trajetória e aos meus amigos por sempre me motivarem e correr atrás dos meus objetivos.

RESUMO

A qualidade do ensino é um assunto amplamente discutido dentro das universidades. Particularmente as Universidades Federais devem prezar pela excelência, seja nos cursos de graduação ou de pós graduação – e isto não é diferente na Universidade Federal do Paraná, em particular no curso de graduação em Engenharia Elétrica. Os principais agentes responsáveis pela administração dos cursos de graduação são o coordenador e o colegiado do curso, que tomam suas decisões utilizando dados fornecidos por vários sistemas computacionais que disponibilizam informações sobre matrículas, notas, frequência, entre outras. Uma parcela significativa do tempo do coordenador de curso é dedicada à resolução de problemas relacionados à matrículas de alunos, horários de disciplinas, equivalências em disciplinas, etc., que são atividades necessárias, mas tendem a desviar a atenção do coordenador da atuação em ações relacionadas com a qualidade de ensino e efetividade do aprendizado. O objetivo desse projeto é elaborar um sistema de suporte à orientação acadêmica, cuja a função é analisar os diversos dados de desempenho acadêmico disponíveis e, com isso, alertar os alunos que requerem orientação acadêmica ou então qualquer outro problema. O desenvolvimento do trabalho foi realizado com a aplicação de ferramentas estatísticas, como média, desvio padrão e distribuições de probabilidade. Para a previsão de desempenho acadêmico foram utilizadas técnicas de clusterização e aplicação de Modelos Ocultos de Markov. Foi utilizado o software R como ferramenta computacional para a clusterização e o treinamento/validação do Modelo Oculto de Markov. Os resultados obtidos servirão de base para orientação pedagógica para os estudantes, além de servir também para a administração do curso que pode mudar as condições de oferta de disciplina.

Palavras-chave: Mineração de dados. Desempenho acadêmico. Clusterização. Modelos Ocultos de Markov.

ABSTRACT

The quality of teaching is a subject widely discussed within universities. Particularly the Federal Universities should value excellence, either in undergraduate or postgraduate courses - and this is no different at the Federal University of Paraná, particularly in the undergraduate course in Electrical Engineering. The main agents responsible for administering the undergraduate courses are the coordinator and the collegiate of the course, who make their decisions using data provided by various computer systems that provide information on enrollments, grades, frequency, among others. A significant portion of the course coordinator's time is devoted to solving problems related to student enrollment, course schedules, discipline equivalencies, etc., which are necessary activities, but tend to divert the coordinator's attention from acting on related actions with the teaching quality and effectiveness of learning. The purpose of this project is to develop a system of support for academic guidance, whose function is to analyze the various academic performance data available and, with this, to alert students that require academic guidance or any other problem. The development of the work was performed with the application of statistical tools, such as mean, standard deviation and probability distributions. For the prediction of academic performance, we used clusters and application of Hidden Markov Models. The R software was used as a computational tool for clustering and the training / validation of the Hidden Markov Model. The results obtained will serve as a basis for pedagogical guidance for the students, as well as also serve for the administration of the course that can change the conditions of offer of discipline.

Keywords: Data Mining. Academic achievement. Clustering. Hidden Markov Models.

LISTA DE ILUSTRAÇÕES

LISTA DE ABREVIATURAS E SIGLAS

CCE	Centro de Computação Eletrônica
EDM	*Educational Data Mining*
Enade	Exame Nacional de Desempenho dos Estudantes
HMM	*Hidden Markov Model*
IES	Instituições de Ensino Superior
MEC	Ministério da Educação
UFPR	Universidade Federal do Paraná

SUMÁRIO

1 INTRODUÇÃO

A qualidade da educação superior pode ser definida como “uma questão de especificar objetivos de aprendizagem que valem a pena e capacitar os alunos a alcançá-los” (SPARKES, 1995). Essa qualidade pode ser mensurada de diversas maneiras, como por exemplo, a efetividade da conclusão do curso pelo aluno em relação ao tempo. Por outro lado, não há nada que signifique maior fracasso institucional do que se o aluno evadir do curso (LOBO, 2017).

Nas Instituições de Ensino Superior (IES) é utilizado, para o acompanhamento do processo de aprendizagem, o Exame Nacional de Desempenho dos Estudantes (Enade) promovido pelo Ministério da Educação (MEC). Os resultados deste exame servem como um fator de ranqueamento dos cursos, e os que não possuem uma boa avaliação, pode motivar visitas de comissões de avaliação e tendem a receber um número menor de estudantes. Tendo isto em vista, os responsáveis pelo ensino das instituições são desafiados a melhorar o nível de formação dos egressos dos cursos (PEDROSO, 2012).

Os sistemas computacionais presentes nos cursos permitem que o coordenador e a equipe responsável pela manutenção e evolução do sistema de ensino analisarem informações a respeito do desempenho dos estudantes. Com a implementação de um sistema capaz de realizar a previsão de desempenho, que analise as informações disponíveis em uma base de dados, afim de “aprender” com situações passadas, as ações que essa equipe poderia tomar seriam mais preventivas do que apenas reativas (PEDROSO, 2017).

Um assunto que vem despertando interesse na comunidade de pesquisa, é o uso de análises preditivas na área educacional, em especial, técnicas de mineração de dados. Previsão de desempenho de estudantes em um curso, em uma disciplina, perfis de estudantes com maior probabilidade de sucesso, são as principais aplicações dessas técnicas (PEDROSO, 2017).

A proposta deste projeto consiste na implementação de um modelo capaz de realizar previsões de desempenho acadêmico com técnicas de mineração de dados. O modelo será construído a partir da análise de uma base de dados com informações sobre os estudantes, como nota, frequência, quantidade de disciplinas cursadas (apro-

vadas e reprovadas) e carga horária efetiva. Os resultados poderão utilizados para reorganizar as condições de oferta de disciplinas do curso ou até mesmo acompanhar os estudantes com baixo desempenho. O sistema será elaborado utilizando as informações dos alunos do curso noturno de Engenharia Elétrica da UFPR.

1.1 OBJETIVOS

A elaboração do projeto ficou dividido em duas etapas, uma que ocorrerá no primeiro semestre de 2018 e a outra no segundo semestre de 2018. O detalhe das atividades e o tempo de cada uma, pode ser visto no Apêndice A. Dessa maneira será possível atingir o objetivo geral e os objetivos específicos descritos para este projeto.

1.1.1 Objetivo Geral

Identificar os estudantes com dificuldades pedagógicas utilizando como métricas notas, frequências, quantidade de disciplinas cursadas (aprovadas e reprovadas) e a carga horária efetiva.

1.1.2 Objetivos Específicos

- Levantamento dos trabalhos relacionados sobre o tema do projeto;
- importação das informações dos estudantes, que obtiveram sucesso ou fracasso (evasão) em uma base de dados;
- classificação dos alunos com técnicas de mineração de dados como clusterização hierárquica;
- modelagem do problema utilizando Modelos Ocultos de Markov;
- interpretação dos resultados obtidos;
- aplicação do modelo nos estudantes ativos no curso de engenharia elétrica e disponibilizar a informação em uma página web.

1.2 PROBLEMA E JUSTIFICATIVA

Atualmente, o curso de Engenharia Elétrica da UFPR oferta duas modalidades: uma no período matutino e vespertino com ênfase em Eletrotécnica e outra no período noturno com ênfase em Sistemas Eletrônicos Embarcados. Esses cursos estão estruturados em dois estágios. O primeiro acontece no inicio do curso, no qual é realizado a

formação de competências matemáticas seguido das competências fundamentais na área de eletricidade, magnetismo e eletrônica (PEDROSO, 2017).

Assim que é concluído o ciclo básico, o aluno pode escolher entre duas áreas de formação para curso diurno, enquanto que o noturno possuí uma formação específica. Para o primeiro, há possibilidade de escolher entre eletrotécnica ou eletrônica/telecomunicações. Já o segundo, na área de sistemas eletrônicos embarcados (PEDROSO, 2017).

Em virtude a Lei nº 12.711, de agosto de 2012, estabelece o sistema de cotas em instituições federais de ensino superior. Isso fez que a universidade recebesse estudantes com níveis diferentes relacionados à formação básica. O curso de Engenharia Elétrica não possui ferramentas de gestão para tratar os problemas devido a essa diversidade na formação básica dos alunos (PEDROSO, 2017).

Todavia, a diversidade na formação básica dos alunos não é a única questão a ser considerada. É necessário olhar também ao desempenho que os egressos do curso vem tendo no Enade. A nota do Enade é normalizada entre 1 (menor) e 5 (maior). Nos últimos 4 exames realizados pelos egressos do curso foram obtidas as notas 3, 3, 3 e 3, em 2008, 2011, 2014 e 2017, respectivamente. Isso significa que o grau de formação dos estudantes está de acordo com a média geral brasileira. Além da vocação da UFPR ser estabelecer referências de ensino de graduação e pós graduação, a missão é formar profissionais de alta qualidade (PEDROSO, 2017). Dessa maneira, conclui-se que é preciso identificar os problemas, para então repensar nas políticas pedagógicas e permitir a execução de novas medidas.

O Centro de Computação Eletrônica (CCE) da UFPR disponibiliza aos coordenadores de curso uma ferramenta computacional chamada SIE, a qual permite a impressão de relatórios sobre as informações dos alunos, como por exemplo, notas, frequências, reprovações, matrículas, entre outras (PEDROSO, 2017).

Hoje, o curso vem desenvolvendo uma ferramenta que utiliza estes relatórios para verificar as situações dos alunos. A identificação dos alunos que necessitam de orientação acadêmica ou a identificação dos alunos que não estão aptos a realizar estágio obrigado, são alguns exemplos do que esse sistema é capaz de fazer. Entretanto, não existe um algoritmo inteligente que permita a previsão de desempenho acadêmico, a fim de possibilitar a tomada de ações antes que o problema ocorra.

Um modelo capaz de descrever o comportamento dos alunos e então rea-

lizar uma previsão desempenho, deixaria essa ferramenta do curso mais completa, permitindo identificar previamente alunos com dificuldades. Na comunidade de pesquisa, existem vários pesquisadores explorando o uso de técnicas de mineração de dados para analisar o aprendizado de estudantes a fim de melhorar a qualidade dos procedimento educativos (GRIVOKOSTOPOULOU; PERIKOS; HATZILYGEROUDIS, 2014).

Considerada uma área emergente, a mineração de dados voltada pra educação provou ser bem eficaz do que as análises estatísticas educacionais. Isso devido ao enorme poder computacional e aos algoritmos disponíveis de mineração de dados. Os resultados do *Educational Data Mining* (EDM), permite a tomada de ações de acordo com o que foi compreendido, em consequência dos métodos utilizados para interpretar melhor os fatores que afetam o desempenho acadêmico (SACHIN; VIJAY, 2012).

O objetivo desse projeto é elaborar um modelo que descreve o comportamento dos estudantes no curso noturno de Engenharia Elétrica. A função é analisar os diversos dados disponíveis dos estudantes já conhecidos e, com isso, construir um modelo e analisar os alunos atuais do curso. Dessa maneira, será possível identificar os alunos que requerem orientação acadêmica ou os estudantes com uma alta probabilidade de evasão. O desenvolvimento do trabalho requisitará a aplicação de ferramentas de mineração de dados, como técnicas de clusterização e aplicação de Modelos Ocultos de Markov.

Os resultados obtidos servirão tanto para os estudantes acompanharem seu desempenho, quanto para a administração do curso, permitindo a tomada de ações prévias, para mitigar o problema. O sistema irá oferecer informações fazendo com que as ações da coordenação de curso relacionadas à qualidade do ensino sejam mais preventivas do que reativas.

1.3 ESTRUTURA DO TRABALHO

O trabalho se apresenta estruturado em cinco capítulos, os quais foram divididos conforme a seguir:

- O capítulo 1 é composto pela introdução, em que é comentado sobre qualidade de ensino dando o destaque para a mineração de dados voltado para educação. Neste capitulo está presente os objetivos a serem atingidos do trabalho, assim

como o problema e a justificativa.

- O capítulo 2 está exposta a metodologia a ser seguida, nesse é mostrado etapas de avaliação, aplicação do modelo, a validação do mesmo, comparação com os trabalhos relacionados e a implementação num cenário real.
- Está presente no capítulo 3 os principais conceitos das técnicas utilizadas, como a clusterização hierárquica, Cadeias de Markov e Modelos Ocultos de Markov. Além destes conceitos é apresentado os trabalhos relacionados.
- O quarto capítulo é a implementação do Modelo Oculto de Markov, no qual é exibido as etapas e atividades que foram realizadas juntamente com os resultados obtidos. Além da apresentação dos resultados, há comentários e discussões a cerca destes.
- Por fim, o último capítulo é realizado uma conclusão a cerca do projeto, destacando os principais resultados obtidos e enfatizando os objetivos atingidos. Além disso é apresentado os trabalhos futuros.

2 METODOLOGIA

A metodologia tem o objetivo de descrever os caminhos para alcançar um resultado. Em outras palavras, é a explicação mais detalhada dos passos que envolvem o desenvolvimento do projeto. Para o trabalho em questão, a metodologia pode ser dividida em cinco etapas: avaliação, aplicação do modelo, validação, comparação e implementação. Os detalhes das etapas será descrito nos tópicos na sequência.

2.1 AVALIAÇÃO

A primeira etapa do projeto está relacionada com a avaliação, que é o levantamento dos trabalhos relacionados. É entender o que comunidade acadêmica está realizando a cerca do tema e quais são os resultados obtidos. Para essa avaliação, será utilizado as principais revistas da área, que são: *IEEE Transactions on Education* e *Elsevier Computers & Education*.

2.2 APLICAÇÃO

A aplicação do modelo envolve as etapas que descrevem como o modelo será implementado. As etapas são: levantamento dos dados históricos; avaliação das informações relevantes; discretização dos dados; treinamento do modelo e validação dos resultados.

É necessário ter uma base de dados em que a trajetória seja conhecida, ou seja, ter os dados dos alunos que já completaram o ciclo de estudos, com sucesso ou fracasso. A partir disto, avaliar as informações relevantes e realizar uma discretização de dados para então aplicar em algoritmos de treinamento e treinar o modelo efetivamente. Após essa etapa é realizada a validação do modelo com o método *k-fold*.

Para a discretização dos dados será utilizada a linguagem de programação Java, enquanto que o treinamento do modelo e a validação dos resultados, será utilizado o software R. O primeiro foi escolhido, pois é um linguagem muito popular e ainda está tendo crescimento acentuado e como trata-se de manipulação de informação, a linguagem conta com algumas funções próprias facilitando a implementação. O software R foi escolhido, pois nele já existem bibliotecas para implementação das

técnicas que serão utilizadas, que são: clusterização hierárquica e Modelos Ocultos de Markov. Com essas bibliotecas prontas, não há necessidade da implementação dos algoritmos que compõe o treinamento e extração dos parâmetros importantes, com isso, será possível focar na parte da aplicação.

Cada uma das etapas, terá uma abordagem com mais detalhes no Capítulo 4, no qual está incluso discussões a respeito das técnicas utilizadas e a maneira que foram implementadas.

2.3 COMPARAÇÃO

A comparação do modelo com os trabalhos relacionados tem como objetivo encontrar as estratégias de utilização para previsão de desempenho acadêmico disponíveis. Com os parâmetros obtidos na etapa de validação será possível comparar com os resultados obtidos pelos modelos concorrentes.

2.4 IMPLEMENTAÇÃO

Por fim, aproveitar todo conhecimento adquirido e implementar o modelo nos alunos atuais do curso. Os resultados obtidos serão disponibilizados em uma página *web*, na qual alunos e professores terão acesso e poderão acompanhar o desempenho sempre que necessário.

3 CONCEITOS FUNDAMENTAIS

Este capítulo tem como objetivo descrever os principais conceitos utilizados para a elaboração do trabalho. apresenta os estudos realizados durante o planejamento e execução deste projeto. Noções daa técnica de clusterização utilizada, Cadeias de Markov e Modelos Ocultos de Markov serão descritos, assim como os trabalhos relacionados.

3.1 TRABALHOS RELACIONADOS

São diversas as aplicações de Modelos Ocultos de Markov e das técnicas de clusterização, embora não seja comum serem utilizados juntos, existem trabalhos que exploram essas técnicas e obtêm resultados interessantes. Nesta seção será comentado sobre as técnicas utilizadas para previsão de desempenho acadêmico e os principais resultados obtidos.

3.1.1 Previsão de desempenho de estudantes do Ensino Médio

(LUCKIN et al., 2007) propõe o uso de Modelos Ocultos de Markov em conjunto com um sistema de tutoria em matemática para modelar os traços de ação dos alunos, é utilizado uma variável oculta denotando seu nível de engajamento durante esse sistema de tutoria inteligente.

Os estudantes resolveram uma série de problemas matemáticos e tinham cinco opções de respostas, havendo a possibilidade de passar para a próxima questão sem responder. As respostas dos estudantes são classificadas da seguinte maneira: “Supôs”, “Pulou”, “Indeciso”, “Preciso” e “Certo” (LUCKIN et al., 2007). Com essa classificação foi gerada uma sequência utilizando Modelos Ocultos de Markov.

Com um total de aproximadamente 200 alunos, foram identificados dois grupos distintos, um grupo com melhor desempenho e outro com menor. O número de estudantes do primeiro grupo era menor que o segundo.

Os resultados da previsão apresentaram uma precisão de aproximadamente 50%. Mesmo não tendo uma precisão alta, as ações que podem ser tomadas, a fim de melhorar o ensino, são melhores que não conhecer nada dos alunos. Mesmo que o

modelo preveja um desempenho ruim em determinado aluno, as ações para melhorar ainda serão validas, por outro lado o problema é quando não há previsão que o aluno tenha um desempenho baixo (LUCKIN et al., 2007).

3.1.2 Prevendo a retenção de estudantes em cursos à distância

Um outro trabalho, utiliza Modelos Ocultos de Markov para modelar o comportamento do estudante ao longo do tempo em um curso à distância. Através das iterações com a plataforma de ensino, é possível modelar diversas características que podem ser utilizadas como indicativo de que o estudante irá seguir até o final do curso (BALAKRISHNAN; COETZEE, 2013).

Como existem muitos dados a serem explorados, para a modelagem do problema, o autor considera duas abordagens. Uma aplicando a técnica de clusterização para verificar em quantos grupos os dados irão se formar e com essa quantidade, é possível aferir o número de estados ocultos que o modelo possuirá para então treiná-lo. Uma outra abordagem, é que seriam treinados diversos modelos, onde cada um seria modelado com um único conjunto de dados e a escolha seria baseada com o modelo que apresentasse uma melhor taxa de acerto (BALAKRISHNAN; COETZEE, 2013).

Os dados foram escolhidos de um curso online da Universidade de Berkeley, chamado de “CS169.1x Software as a Service”, um curso de 6 semanas com 29.882 inscritos. O curso era divido em 6 módulos (um por semana) e conforme as semanas passavam o aluno poderia avançar para o módulo seguinte (BALAKRISHNAN; COETZEE, 2013). As informações coletadas para a análise do modelo eram:

- os cliques na tela que o usuário realizava durante os vídeos e materiais de estudo, cliques como *pause*, *play*, avançar o vídeo, passar de página, entre outros;
- as notas obtidas em cada módulo;
- interações com a comunidade, como comentários em *forums*, leitura de *posts*, etc.

Os resultados mostraram relações interessantes com os dados estudados. Quando os alunos não interagem com a comunidade, não visualizam tópicos e nem realizam comentários, a probabilidade de evasão fica maior que 35%. Enquanto os alunos que constantemente interagem no *forum*, tem uma probabilidade de evasão menor que 5%. Outra relação, é quando os estudantes não acompanham seu progresso

no curso, possuem uma probabilidade de evasão de, aproximadamente, 40%, enquanto os alunos que verificam seu progresso mais de duas vezes na semana, possuem uma probabilidade próxima a 20% (BALAKRISHNAN; COETZEE, 2013).

A modelagem com Modelos Ocultos de Markov, mostrou bons resultados para a identificação de comportamentos e padrões dos estudantes. Estes modelos não se restringem apenas à áreas acadêmicas, são amplamente utilizados para reconhecimento de falar e nas áreas biológicas, com sequenciamento genético, por exemplo.

3.1.3 Mineração de dados para previsão de desempenho

Há diversos trabalhos que aplicam técnicas para realizar a previsão de desempenho acadêmico. Na sequência será apresentado alguns trabalhos que utilizam técnicas de mineração de dados com este fim, assim como será comentado os resultados obtidos.

(GRIVOKOSTOPOULOU; PERIKOS; HATZILYGEROUDIS, 2014) propõe uma metodologia para analisar a aprendizagem dos alunos e extrair regras semânticas que podem ser usadas para prever o desempenho final do aluno no curso. Técnicas de árvores de decisão são utilizadas para realizar previsões sobre o desempenho final do estudante no curso.

De acordo com seu desempenho, o estudante pode classificado em uma das três classes: “aprovado”, “reprovado” e “em alerta”, sendo o último em que o aluno está quase reprovando. A classificação é realizada com a técnica de árvores de decisão e os algoritmos utilizados são *J48* e *CART*. Os resultados obtidos com o primeiro algoritmo tiveram uma taxa de acerto de 87,55%, enquanto que o segundo teve 81,34%, concluindo que ambos algoritmos tiveram um resultado satisfatório e apresentaram uma precisão alta (GRIVOKOSTOPOULOU; PERIKOS; HATZILYGEROUDIS, 2014).

(VERA; MORALES; SOTO, 2013) propõe a aplicação de técnicas de mineração de dados para prever falhas e abandono escolar. Utilizando diversos algoritmos com métodos de classificação de caixa branca, como regras de indução e árvores de decisão, foi possível determinar qual desses possui o melhor resultado.

As classificações ficaram limitadas em duas classes: “aprovação” e “reprovação” e os resultados foram validados com o método *k-fold* da validação cruzada. Os resultados obtidos utilizando o algoritmo *ADTree* tiveram uma precisão de 97,3%,

porém a taxa de acerto das predições da classe “reprovado” foi menor que 80%. Já os resultados do algoritmo *JRip* tiveram uma precisão um pouco menor que o anterior, em torno de 96%, mas a taxa de acerto nas predições negativas foi de aproximadamente 82% (VERA; MORALES; SOTO, 2013).

Todos os algoritmos utilizados tiveram uma alta precisão, quase sempre maior que 90%. De acordo com os atributos utilizados para a classificação, um algoritmo apresentava resultados superior ao outro. Entretanto, esse mesmo algoritmo pode apresentar resultados inferiores se utilizados atributos diferentes (VERA; MORALES; SOTO, 2013).

Os métodos relatados refletem o quanto a previsão de desempenho no ensino é de interesse da comunidade de pesquisa. Além de ser interessante também para as instituições de ensinos superiores em utilizar essas ferramentas de mineração de dados para o aprimoramento do ensino de graduação.

3.2 CLUSTERIZAÇÃO

Técnicas para particionar objetos em grupos otimamente homogêneos com base em medidas empíricas de similaridade entre esse objetivos receberam atenção crescente em diversos campos (JOHNSON, 1967). São diversas as ferramentas que permitem essa divisão dos dados em grupos e cada uma delas utiliza parâmetros diferentes que “realizam” o agrupamento. Para este projeto, a técnica de agrupamento a ser utilizada é a clusterização hierárquica.

Essa técnica de agrupamento consiste, a partir de um vetor das medições de similaridade, em construir um sistema hierárquico de agrupamento (JOHNSON, 1967). Esse sistema parte em que cada caso da base dados é um grupo, até a situação em que todos os casos da base dados pertencem a um mesmo grupo.

Em outras palavras, essa técnica separa cada observação em um grupo isolado, na sequência, examina todas as distâncias entre todas as amostras e emparelha as duas mais próximos para formar um novo grupo. Esta é uma operação simples, já que os métodos hierárquicos requerem uma matriz de distância e representam exatamente o objetivo do sistema (as distâncias entre as observações individuais). Portanto, encontrar o primeiro grupo a ser formado significa simplesmente procurar o menor número na matriz de distância e unir as duas observações que a distância corresponde a um novo

grupo (SPECTOR, 2011). Alguns dos métodos propostos são para levar a distância mínima entre uma observação e qualquer membro de um grupo para encontrar: a distância máxima; a distância média; ou usar algum tipo de medida que minimize as distâncias entre observações dentro do grupo. Cada um desses métodos revelará certos tipos de estrutura nos dados (SPECTOR, 2011).

Tradicionalmente, a clusterização hierárquica utiliza atalhos computacionais (ou algoritmos inteligentes) ao atualizar a matriz de distância para refletir novos grupos. Em particular, quando um novo grupo é formado e a matriz de distância é atualizada, todas as informações sobre os membros individuais do grupo são descartadas para tornar os cálculos mais rápidos (SPECTOR, 2011).

A aplicação desta técnica, deve atender as seguintes condições (JOHNSON, 1967):

1. a entrada deve consistir unicamente nas medidas de similaridade $n.(n-1)/2$ entre os objetos de estudo, sendo 'n' o número de objetos. Isto está em contraste com alguns métodos que requerem que cada objeto seja inicialmente representado como um ponto no espaço euclidiano;
2. os dados de entrada devem ser normalizados antes da aplicação da técnica. É comum encontrar conjunto de dados em escalas diferentes, como por exemplo, peso e altura, enquanto que o primeiro está numa faixa de valores entre, 0,5 e 2 m e o segundo entre 20 a 100 kg;
3. os grupos gerados pela técnica devem possuir um significado claro e objetivo. Há muitas técnicas de agrupamento que possuem excelentes algoritmos, mas enfrentam o problema de não ter um significado claro para o grupo.

3.3 CADEIAS DE MARKOV

Andrew A. Markov, em 1907, iniciou um estudo sobre um processo que o resultado futuro de um experimento depende apenas do resultado do experimento atual. Hoje, esse tipo de processo é conhecido como Cadeias de Markov.

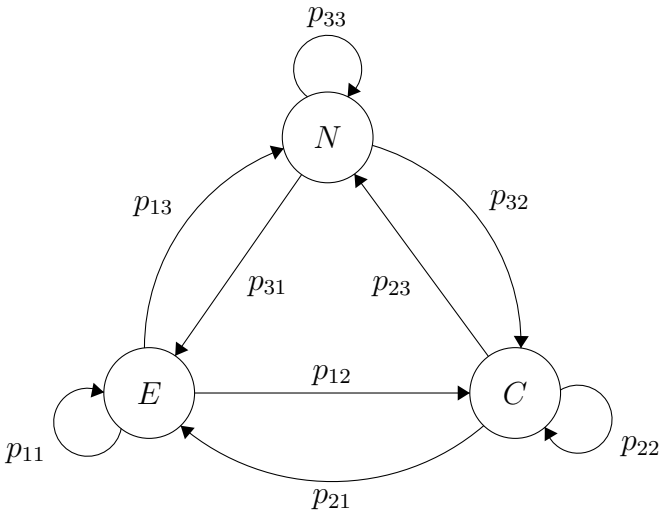
Uma Cadeia de Markov pode ser especificada da seguintes maneira: considerando um conjunto de estados $S = s_1, s_2, ..., s_r$. O processo inicia-se em um destes r estados e troca para um outro estado com uma probabilidade. Cada troca é chamada de passo. Por exemplo, se a cadeira encontra-se no estado s_i e alterna para o estado

s_j, a probabilidade é denotada como p_{ij} e chamada de probabilidade de transição. Essas probabilidades não dependem dos estados anteriores do processo (GRINSTEAD; SNELL, 2006).

Uma maneira de entender a definição de Cadeia de Markov, é analisar o comportamento de um sapo em uma vitória-régia. O sapo sempre está em cima da planta e nunca mergulha na água. De tempos em tempos o sapo da um salto e troca de planta, mas pode ser que ele apenas dê um pulo e permaneça na mesma. A analogia é feita não com os padrões de pulos do sapo, mas sim em qual planta ele se encontra. A posição da vitória-régia descreve o estado, enquanto que o sapo alternando de planta descreve os passos. Essas trocas acontecem de acordo com uma probabilidade de transição (HOWARD, 1960).

Para entender o equacionamento de Cadeias de Markov, é possível utilizar um exemplo do clima. Supondo que a cidade tenha apenas três tipos de clima (ensolarado, chuvoso e nublado) e o clima dure o dia inteiro. A previsão do tempo é baseada em tentar adivinhar qual será o clima do dia seguinte, com base no dia anterior (FOSLER-LUSSIER, 1998). Como visto anteriormente, um processo que o estado futuro depende apenas do estado atual, pode ser modelado com uma Cadeia de Markov.

FIGURA 3.1 – Diagrama de Transição de Estados



Fonte: Autor

Existem representações gráficas para descrever uma Cadeia de Markov. Quando o problema é modelado com poucos estados, é possível representá-lo por meio de uma máquina de estados como, por exemplo, a Figura 3.1 que representa a máquina de estados do clima da cidade. Dessa maneira, é notável que para sair de um clima

para outro, há uma probabilidade, por exemplo, a probabilidade do clima futuro ser *Chuvoso* $= C$, dado que o atual é *Ensolarado* $= E$, então a probabilidade é de p_{12}.

Além da representação gráfica, é possível representar as probabilidades de transição por meio de uma matriz, no qual as linhas representam o estado atual e as colunas o estado futuro. Um exemplo pode ser visto na equação 3.1.

$$\mathbf{P} = \begin{array}{c} \\ E \\ C \\ N \end{array} \begin{array}{c} \begin{array}{ccc} E & C & N \end{array} \\ \begin{pmatrix} p_{11} & p_{12} & p_{13} \\ p_{21} & p_{22} & p_{23} \\ p_{31} & p_{32} & p_{33} \end{pmatrix} \end{array} \tag{3.1}$$

Considerando esse modelo simplificado de previsão de tempo, é necessário saber quais foram os climas passados para saber o clima seguinte, ou seja, encontrar a maior probabilidade do clima de hoje baseado no clima de ontem, anteontem e assim por diante (FOSLER-LUSSIER, 1998). A equação 3.2 denota essa probabilidade.

$$P(w_n|w_{n-1}, w_{n-2}, ..., w_1) \tag{3.2}$$

Supondo que o clima dos três dias anteriores são: ensolarado, ensolarado e nublado, usando a equação 3.2, é possível encontrar a probabilidade de amanhã ser um dia chuvoso, por exemplo (FOSLER-LUSSIER, 1998). A equação 3.3 descreve o exemplo.

$$P(w_4 = C|w_3 = \textit{N}, w_2 = \textit{E}, w_1 = \textit{E}) = p_{11} \,.\, p_{13} \,.\, p_{32} \tag{3.3}$$

Outra análise importante é a probabilidade de se encontrar um estado no passo n futuro. Essa relação é dada pela equação 3.4 em que U representa a matriz coluna das probabilidades da sequência iniciar em um determinado estado.

$$P(w_n|w_1) = U \,.\, P^n \tag{3.4}$$

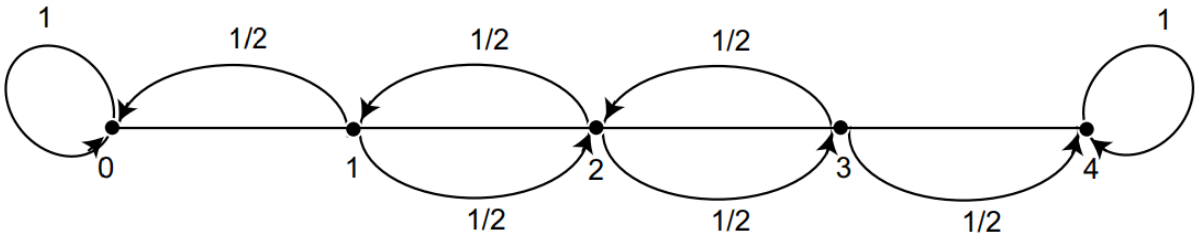
3.3.1 Cadeias de Markov Absorventes

Uma Cadeia de Markov Absorvente é definida se pelo menos um estado da cadeia possui uma probabilidade de transição para ele mesmo de 100% ($p_{ii} = 1$) e

que exista uma probabilidade de transição para esse estado a partir dos outros (não necessariamente em um passo) (GRINSTEAD; SNELL, 2006).

Para entender o equacionamento de uma Cadeia de Markov Absorvente, é possível observar um exemplo da Figura 3.2 sobre o andar do bêbado. Suponha que os estados representem esquinas e o homem segue andando até a próxima esquina. Imaginando que o rapaz chegue nas esquinas 1, 2 ou 3, ele pode tanto ir para a direita quanto para a esquerda com probabilidades iguais (50%). Ele segue até alcançar as esquinas 4 ou 0, que representam o bar e sua casa, respectivamente. Se ele chegar ou no bar ou em casa, ele permanece lá (GRINSTEAD; SNELL, 2006).

FIGURA 3.2 – Andar do bêbado



Fonte: (GRINSTEAD; SNELL, 2006).

Dessa maneira, é possível determinar que esta é uma Cadeia de Markov Absorvente, pois a partir de qualquer um dos estados (1, 2 ou 3) é possível chegar nos estados absorventes (GRINSTEAD; SNELL, 2006). A expressão 3.5 ilustra a matriz de transição dos estados.

$$\mathbf{P} = \begin{array}{c} \\ 0 \\ 1 \\ 2 \\ 3 \\ 4 \end{array} \begin{array}{c} \begin{array}{ccccc} 0 & 1 & 2 & 3 & 4 \end{array} \\ \begin{pmatrix} 1 & 0 & 0 & 0 & 0 \\ \frac{1}{2} & 0 & \frac{1}{2} & 0 & 0 \\ 0 & \frac{1}{2} & 0 & \frac{1}{2} & 0 \\ 0 & 0 & \frac{1}{2} & 0 & \frac{1}{2} \\ 0 & 0 & 0 & 0 & 1 \end{pmatrix} \end{array} \tag{3.5}$$

É natural que ao analisar esse problema, algumas questões surjam de imediato, dentre elas é possível destacar:

(a) Qual a probabilidade que um processo seja absorvido?

(b) Quantos passos em média para o processo ser absorvido?

(c) Quantos passos em média o processo estará nos estados transientes até ser absorvido?

Em princípio, as respostas para as perguntas irão depender em qual estado o processo se inicia e das probabilidades de transição (GRINSTEAD; SNELL, 2006).

3.3.1.1 Forma Canônica

Reorganizando a matriz da equação 3.5 para que os estados transientes apareçam primeiro e os estados absorventes fiquem por último, a matriz ganha uma forma que é chamada de canônica, e pode ser vista na expressão 3.6.

$$\mathbf{P} = \begin{array}{c} \\ 1 \\ 2 \\ 3 \\ 0 \\ 4 \end{array} \begin{array}{c} \begin{array}{ccccc} 1 & 2 & 3 & 0 & 4 \end{array} \\ \begin{pmatrix} 0 & \frac{1}{2} & 0 & \frac{1}{2} & 0 \\ \frac{1}{2} & 0 & \frac{1}{2} & 0 & 0 \\ 0 & \frac{1}{2} & 0 & 0 & \frac{1}{2} \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 \end{pmatrix} \end{array} \tag{3.6}$$

De modo geral, a forma canônica de uma Cadeia de Markov Aborvente pode ser escrita conforme a equação 3.7. Na qual a matriz Q representa as probabilidades de transição dos estados transientes, a matriz R representa as probabilidades de transição para os estados absorventes, a matriz I, uma matriz identidade e a matriz 0 representa uma matriz de zeros.

$$\mathbf{P} = \left(\begin{array}{c|c} Q & R \\ \hline 0 & I \end{array}\right) \tag{3.7}$$

Considerando t como o número de estados transientes e r como o número de estados absorventes, a matriz Q é uma matriz quadrada de tamanho $t \times t$, a matriz R possui tamanho $t \times r$, a matriz I com tamanho $r \times r$ e a matriz de zeros com dimensão de $r \times t$ (GRINSTEAD; SNELL, 2006).

3.3.1.2 Matriz Fundamental

Com a forma canônica de uma Cadeia de Markov Absorvente, é possível identificar os teoremas que podem ser usados para responder as perguntas levantadas no início desta seção. Chamada de matriz fundamental e definida na equação 3.8,

é possível identificar o número esperado de vezes que o processo encontra-se nos estados transientes, apenas olhando para ela (GRINSTEAD; SNELL, 2006).

$$\mathbf{N} = (\mathbf{I} - \mathbf{Q})^{-1} \tag{3.8}$$

Usando como exemplo anterior, do andar do bêbado na forma canônica da equação 3.6, é possível identificar as matrizes $\mathbf{Q}$ e $\mathbf{I}$ para então calcular a matriz $\mathbf{N}$.

$$\mathbf{Q} = \begin{array}{c} \\ 1 \\ 2 \\ 3 \end{array} \begin{array}{c} \begin{array}{ccc} 1 & 2 & 3 \end{array} \\ \begin{pmatrix} 0 & \frac{1}{2} & 0 \\ \frac{1}{2} & 0 & \frac{1}{2} \\ 0 & \frac{1}{2} & 0 \end{pmatrix} \end{array} \tag{3.9}$$

$$\mathbf{I} - \mathbf{Q} = \begin{array}{c} \\ 1 \\ 2 \\ 3 \end{array} \begin{array}{c} \begin{array}{ccc} 1 & 2 & 3 \end{array} \\ \begin{pmatrix} 1 & -\frac{1}{2} & 0 \\ -\frac{1}{2} & 1 & -\frac{1}{2} \\ 0 & -\frac{1}{2} & 1 \end{pmatrix} \end{array} \tag{3.10}$$

$$\mathbf{N} = (\mathbf{I} - \mathbf{Q})^{-1} = \begin{array}{c} \\ 1 \\ 2 \\ 3 \end{array} \begin{array}{c} \begin{array}{ccc} 1 & 2 & 3 \end{array} \\ \begin{pmatrix} \frac{3}{2} & 1 & \frac{1}{2} \\ 1 & 2 & 1 \\ \frac{1}{2} & 1 & \frac{3}{2} \end{pmatrix} \end{array} \tag{3.11}$$

Supondo que o processo inicia-se no estado 1, pela equação 3.11 é esperado que o número de vezes que a cadeia permaneça nos estados 1, 2 e 3 seja de $\frac{3}{2}$, 1 e $\frac{1}{2}$, respectivamente.

3.3.1.3 Passos até a absorção

A resposta da questão (b) pode ser encontrada com o próximo teorema, que diz respeito ao número médio de passos até o processo ser absorvido (GRINSTEAD; SNELL, 2006).

Para encontrar o número médio de passos que o processo transita entre os estados transientes, dado um estado inicial é utilizado a matriz fundamental. Somando esses passos médios, encontra-se o número esperado de vezes que o processo permanece nos estados transientes, iniciando em qualquer um dos estados transitórios,

ou seja, o tempo médio antes do processo ser absorvido. Dessa maneira, com a soma das entradas da i-ésima linha de $\mathbf{N}$ é possível obter o teorema definido na equação 3.12, no qual $\mathbf{c}$ é um vetor coluna com todos os valores igual a 1 (GRINSTEAD; SNELL, 2006).

$$\mathbf{t} = \mathbf{Nc} \tag{3.12}$$

Ainda utilizando o exemplo do andar do bêbado, o número médio de passos até a absorção pode ser visto na equação 3.14.

$$\mathbf{t} = \mathbf{Nc} = \begin{array}{c} \\ 1 \\ 2 \\ 3 \end{array} \begin{array}{c} \begin{array}{ccc} 1 & 2 & 3 \end{array} \\ \begin{pmatrix} \frac{3}{2} & 1 & \frac{1}{2} \\ 1 & 2 & 1 \\ \frac{1}{2} & 1 & \frac{3}{2} \end{pmatrix} \end{array} \quad \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix} \tag{3.13}$$

$$\mathbf{t} = \begin{array}{c} \\ 1 \\ 2 \\ 3 \end{array} \begin{pmatrix} 3 \\ 4 \\ 3 \end{pmatrix} \tag{3.14}$$

Supondo que o processo inicia-se no estado 1, é esperado que o número de vezes que a cadeia permaneça nos estados transientes seja igual a 3.

3.3.1.4 Probabilidade de Absorção

O último teorema está relacionado com as probabilidades que cada um dos estados transientes seja absorvido, ou seja, dado um estado s_j, a probabilidade que ele seja absorvido é de b_{ij}. Sendo $\mathbf{B}$ a matriz com as probabilidade b_{ij} de dimensão $t \times r$, essa pode ser definida como a multiplicação da matriz $\mathbf{N}$ pela matriz $\mathbf{R}$ de acordo com a equação 3.15 (GRINSTEAD; SNELL, 2006).

$$\mathbf{B} = \mathbf{NR} \tag{3.15}$$

Para encontrar a probabilidade em que o homem chegue em casa (estado 0) ou no bar (estado 4), basta multiplicar as matrizes $\mathbf{N}$ e $\mathbf{R}$ de acordo com a equação

3.17 (GRINSTEAD; SNELL, 2006).

$$\mathbf{B} = \mathbf{NR} = \begin{pmatrix} \frac{3}{2} & 1 & \frac{1}{2} \\ 1 & 2 & 1 \\ \frac{1}{2} & 1 & \frac{3}{2} \end{pmatrix} \cdot \begin{pmatrix} \frac{1}{2} & 0 \\ 0 & 0 \\ 0 & \frac{1}{2} \end{pmatrix} \tag{3.16}$$

$$\mathbf{B} = \begin{array}{c} \\ 1 \\ 2 \\ 3 \end{array} \begin{array}{c} \begin{array}{cc} 0 & 4 \end{array} \\ \begin{pmatrix} \frac{3}{4} & \frac{1}{4} \\ \frac{1}{2} & \frac{1}{2} \\ \frac{1}{4} & \frac{3}{4} \end{pmatrix} \end{array} \tag{3.17}$$

Sendo assim, a probabilidade que o homem chegue no bar, dado que o estado atual é o 1, é de 75% e para chegar em casa é de 25% (GRINSTEAD; SNELL, 2006).

3.4 MODELOS OCULTOS DE MARKOV

Modelos Ocultos de Markov ou, do inglês, *Hidden Markov Model* (HMM) são uma extensão das Cadeias de Markov. Esses modelos são caracterizados como um processo estocástico embutido com outro processo subjacente que não é observável (está oculto) e que só podem ser inferidos a partir de uma sequência de observações (RABINER, 1989).

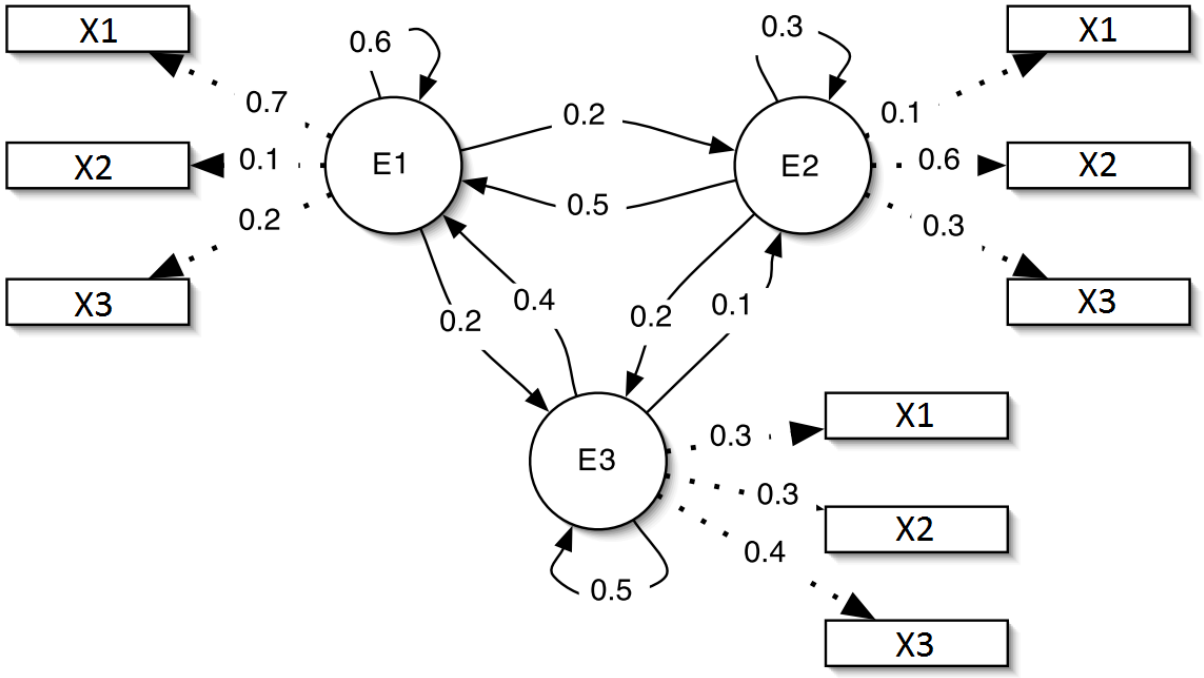
Essa mudança de uma Cadeia de Markov para um Modelo Oculto de Markov permite que todas as observações sejam caracterizadas a partir de cada um dos estados com uma probabilidade finita. Além disso, torna o modelo mais completo para representar e modelar diversas situações (BLUNSOM, 2004).

A grande diferença é que o estado atual desse modelo é desconhecido (por isso o nome, modelo oculto). Por outro lado, é possível calcular o estado mais provável a partir de uma sequência de observações (BLUNSOM, 2004).

Para entender uma aplicação de um Modelo Oculto de Markov, é possível ampliar o exemplo anterior, sobre a previsão do clima. Suponha que exista uma pessoa presa e incapaz de observar diretamente o clima atual. A única evidência dessa pessoa, é observar o vigia verificando se ele trouxe ou não um guarda-chuva. Conhecendo a probabilidade de um determinado clima de acordo com a observação da presença ou não do guarda-chuva, é possível determinar o estado atual do sistema (FOSLER-LUSSIER, 1998).

Um Modelo Oculto de Markov também pode ser representado por uma máquina de estados, da mesma maneira que uma Cadeia de Markov, porém é necessário incluir as observações. A Figura 3.3 é um exemplo da representação gráfica de um Modelo Oculto de Markov.

FIGURA 3.3 – Modelo oculto de Markov



Fonte: Adaptado de (BLUNSOM, 2004).

Esse modelo permite que todas as observações sejam emitidos de cada estado com uma probabilidade finita, representada na imagem pelos estados ocultos E1, E2 e E3. Como dito anteriormente, em um Modelo Oculto de Markov o estado atual é desconhecido. Cada estado possui uma probabilidade de ocorrência para cada evento observável, chamada de matriz de emissão e denotado na imagem por X1, X2 e X3. Por exemplo, no estado E1 a probabilidade do observável X1 é de 70%, X2 de 10%, e assim por diante.

Um Modelo Oculto de Markov é caracterizado pelo seguinte (RABINER, 1989):

- N sendo o número de estados do modelo, os estados individuais definidos como $S = S_1, S_2, ...S_N$ e o estado no tempo t de q_t;
- M sendo o número de símbolos distintos observados, sendo os símbolos denotados como $V = V_1, V_2, ...V_M$;
- a matriz de transição dos estados $A = \{a_{ij}\}$, onde $a_{ij} = P[q_{t+1} = S_j | q_t = S_i]$;

- a matriz de emissão dos símbolos no estado j, $B = \{b_j(k)\}$, onde $b_j(k) = P[V_k \textit{ no instante } t|q_t = S_j]$;
- a distribuição das probabilidades do estado inicial $\pi = \{\pi_i\}$ onde $\pi_i = P[q_1 = S_i]$.

Com os valores apropriados de N, M, A, B e π, o modelo poder ser usado como um gerador e retornar um sequência se observações $O = O_1, O_2, ..., O_T$ onde O_t é um dos símbolos de V, e T é o número de observações na sequência. Por conveniência, é possível representar um Modelo Oculto de Markov com uma notação mais compacta $\lambda = (A, B, \pi)$ (RABINER, 1989).

Dada a definição formal desse modelo, há três problemas básicos de interesse que devem ser resolvidos para seja possível aplicar esse modelo em cenários reais, são eles (RABINER, 1989):

1. dado o modelo λ e uma sequência de observações $O = O_1, O_2, ..., O_T$, qual a probabilidade das observações serem geradas pelo modelo, ou seja, $P\{O|\lambda\}$?
2. dado o modelo λ e uma sequência de observações $O = O_1, O_2, ..., O_T$, qual é a sequência de estados mais provável a partir das observações ?
3. dado o modelo λ e uma sequência de observações $O = O_1, O_2, ..., O_T$, como ajustar os parâmetros $\{A, B, \pi\}$ a fim de maximizar $P\{O|\lambda\}$?

O primeiro é conhecido como problema de avaliação. O segundo é chamado de problema de decodificação, que está relacionado ao reconhecimento contínuo e à segmentação. O último, é o problema de aprendizado que deve ser resolvido para encontrar os melhores parâmetros do modelo $\{A, B, \pi\}$ (WARAKAGODA, 1996).

3.4.1 Solução para o problema 1

É possível calcular as probabilidades de encontrar uma sequência de observações $O = O_1, O_2, ..., O_T$, a partir do modelo $\lambda = (A, B, \pi)$, apenas com argumentos probabilísticos simples. Entretanto esse cálculo envolve o número de operações na ordem de $O(N^T)$. Mesmo que o tamanho da sequência (T) seja moderado, esse valor ainda é muito grande (WARAKAGODA, 1996). Contudo, é necessário procurar por métodos mais otimizados para realizar esse cálculo.

O algoritmo chamado *forward algorithm* utiliza recursos que simplificam as operações, chegando a uma complexidade de $O(N^2T)$. O funcionamento do algoritmo é dado em duas etapas: primeiro é calculada a probabilidade de uma parte da sequência

(t observações) terminar em um determinado estado, ou seja, $P(S_t|O_{1:t})$; e a segunda parte é calculada a probabilidade do modelo gerar a sequência das observações restantes ($T-t$), ou seja, $P(O_{t+1:T}|S_t)$. Dessa maneira, com a multiplicação dos valores é possível determinar a probabilidade da sequência de observações serem geradas pelo modelo (WARAKAGODA, 1996).

3.4.2 Solução para o problema 2

Uma abordagem para encontrar a sequência mais provável dos estados, dado um sequência de observações $O = O_1, O_2, ..., O_T$ é utilizando um algoritmo comumente conhecido como algoritmo de Viterbi. Esse algoritmo calcula todos os caminhos possíveis para um determinada sequência e escolhe o caminho com uma maior probabilidade de ocorrência, ou seja, o caminho mais provável (WARAKAGODA, 1996). Em outras palavras, o algoritmo calcula a probabilidade de uma observação ser gerada por cada um dos estados, então o procedimento se repete para as próximas observações. Além disso, o algoritmo leva em conta as probabilidades de transição entre os estados. Dessa maneira, é retornada a sequência de estados mais provável que o modelo percorreu.

3.4.3 Solução para o problema 3

A partir do conjunto de observações fornecido (também chamado de conjunto de treinamento), o desafio é encontrar a melhor maneira de fazer com que os parâmetros do modelo represente da melhor forma a aplicação pretendida. O quão otimizado o modelo deve ficar, pode variar de aplicação para aplicação. Em outras palavras, existem critérios de otimização mais adequados para cada aplicação (WARAKAGODA, 1996).

Existem vários critérios de otimização que podem ser encontrados na literatura, porém o critério utilizado será o de Máxima Verossimilhança (MV) para o treinamento do modelo proposto neste projeto. Esse critério visa maximizar a probabilidade de gerar a sequência de observações O dado o modelo λ.

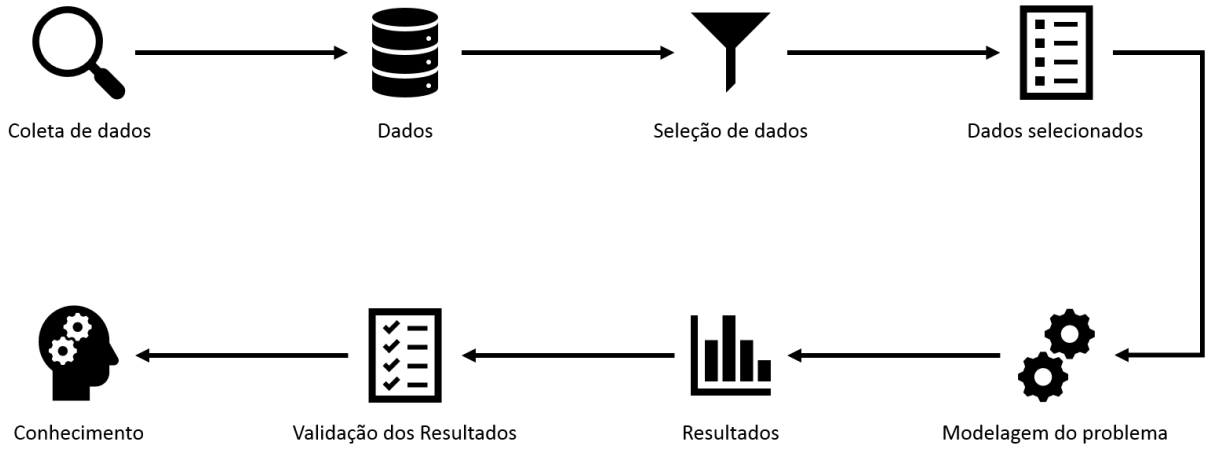
Chamado de *Baum-Welch algorithm* (ou como *forward-backward algorithm*) funciona de maneira similar com o algoritmo apresentado anteriormente, porém com uma etapa a mais. Esse pode ser descrito em três etapas: a primeira é encontrado $P(S_t|O_{1:t})$ da mesma maneira que o anterior; depois o algoritmo calcula $P(O_{t+1:T}|S_t)$, também explicado anteriormente; por fim, os valores são combinados para obter a

probabilidade da sequência ser gerada pelo modelo. Dessa maneira, os parâmetros do modelo são atualizados e essas etapas são repetidas iterativamente até atingir um nível desejado de convergência $\{A, B, \pi\}$ (WARAKAGODA, 1996).

4 USO DE HMM PARA PREVISÃO DE DESEMPENHO ACADÊMICO

A Figura 4.1 representa o diagrama geral do projeto, o qual tem como objetivo ilustrar as etapas que compõe a implementação do Modelo Oculto de Markov para previsão de desempenho acadêmico. Será feita um detalhamento de cada etapa na sequência, explicando do que se trata e os resultados obtidos.

FIGURA 4.1 – Diagrama geral do projeto



Fonte: Autor, 2018.

4.1 COLETA DE DADOS

A coleta de dados é a etapa inicial para elaboração do modelo, essa consiste na extração das informações da base de dados da UFPR. É a coleta dos dados dos alunos que possuem um desfecho, ou seja, os estudantes que concluíram o curso ou os estudantes que não obtiveram êxito e acabaram abandonando o curso.

É uma atividade minuciosa, pois é necessário saber quais alunos tiveram êxito e quais alunos não tiveram e então buscar as informações um a um dentro da base de dados disponibilizada pelo sistema da UFPR. Tendo isto em vista, foi escolhido um grupo seleto de alunos, em que essa coleta ficou restrita nos alunos do curso noturno dos anos de 2009 até 2017. No total, foram selecionados 224 estudantes, nos quais apenas 59 estudantes atingiram o sucesso (concluíram o curso), isso significa que mais de 70% dos estudantes não tiveram êxito (abandonaram ou cancelaram o curso).

Toda informação coletada é organizada em um arquivo texto de extensão .csv[1] (do inglês *Comma-Separated Values)*. Um exemplo da informação contida nessa base e organizada em um arquivo, é ilustrado pela Figura 4.2.

FIGURA 4.2 – Dados de entrada

```
MATR_ALUNO, NOME_ALUNO, ANO,  MEDIA_FINAL,PERIODO,      SITUACAO,           COD_ATIV_CURRIC, FREQUENCIA,  CH_TOTAL_CONSIDERADA
Aluno 1,              , 2012, 80          ,1o. Semestre, Aprovado,           TE100,           100,         60
Aluno 1,              , 2012, 35          ,1o. Semestre, Reprovado por nota, TE239,           76.66666667, 0
Aluno 1,              , 2012, 52          ,1o. Semestre, Aprovado,           TE240,           76.66666667, 60
Aluno 1,              , 2012, 70          ,1o. Semestre, Aprovado,           TE241,           91.66666667, 60
Aluno 1,              , 2012, 90          ,1o. Semestre, Aprovado,           TE243,           80,          60
Aluno 1,              , 2012, 80          ,2o. Semestre, Aprovado,           TE148,           100,         30
Aluno 1,              , 2012, 76          ,2o. Semestre, Aprovado,           TE202,           100,         30
Aluno 1,              , 2012, 74          ,2o. Semestre, Aprovado,           TE208,           86.66666667, 60
Aluno 1,              , 2012, 15          ,2o. Semestre, Reprovado por nota, TE244,           76.66666667, 0
Aluno 1,              , 2012, 88          ,2o. Semestre, Aprovado,           TE246,           80,          60
Aluno 1,              , 2012, 86          ,2o. Semestre, Aprovado,           TE247,           93.33333333, 30
Aluno 1,              , 2012, 28          ,2o. Semestre, Reprovado por nota, TE248,           80,          0
Aluno 1,              , 2013, 0           ,1o. Semestre, Reprov Conhecimento, TE239,          0,           0
```

Fonte: Autor, 2018.

Os dados estão estruturados da seguinte maneira: as colunas contém o tipo de informação e as linhas a informação propriamente dita. As linhas significam os resultados obtidos pelo aluno em apenas uma disciplina. Utilizando o exemplo da Figura 4.2, o Aluno 1 no primeiro semestre de 2012 cursou a disciplina TE100, na qual obteve uma média final de 80 com 100% de frequência, isso ocasionou a aprovação na disciplina e a contabilização de 60 horas (carga horária).

Todos esses dados são sanitizados, isso significa que as informações confidenciais dos estudantes (nome e matricula) foram removidas, dessa maneira não há como saber quais estudantes estão presentes na base de dados. É possível notar no exemplo anterior que o nome do aluno está vazio e a matricula está escrito “Aluno 1”.

Uma vez que esses dados forem coletados, a base de dados estará formada. Entretanto é necessário trabalhar essa base e selecionar apenas os dados que são úteis para treinamento do modelo.

4.2 SELEÇÃO DOS DADOS

Os dados coletados trazem diversas informações dos alunos e são disponibilizadas em uma estrutura de dado composta com as seguintes informações: matricula do aluno, nome do aluno, nome da disciplina que o aluno cursou, código da disciplina

[1] são arquivos de texto de formato regulamentado pelo RFC 4180, que faz uma ordenação separando os valores com vírgulas.

que o aluno cursou, período que a disciplina foi ofertada, média que o aluno obteve na disciplina, frequência do aluno nessa disciplina, situação do aluno na disciplina (aprovado ou reprovado), entre muitas outras.

Dentre todas essas informações dos estudantes, algumas são desconsideradas por não fornecerem nenhum conhecimento de interesse, pelo menos para o modelo que está sendo proposto. As informações que não apresentaram uma relevância podem ser vistos nos itens na sequência juntamente com os motivos.

1. Disciplinas que não fazem parte do currículo do curso. Existem alunos que cursaram disciplinas de outros cursos e até possuem nota e frequência, mas como não fazem parte da grade curricular do curso, foram desconsiderados.
2. Disciplinas cursadas nos semestres especiais (que são aqueles ofertados durantes o período de férias). Por se tratar de uma ocasião especial para ocorrência da oferta dessas disciplinas, esses aspectos foram desprezados.
3. Situação do aluno na disciplina diferente de “Aprovado”, “Reprovado por nota” ou “Reprovado por frequência”. O arquivo contempla diversas situações, como por exemplo “Aprov Aproveitamento”, “Dispensa de disciplina”, entre muitos outros. Assim como a anterior, são casos isolados dos estudantes e não foram contemplados;
4. Alunos que cancelaram a disciplina. Esses alunos não foram considerados nesse momento, porém serão analisados em uma estatística diferente, sendo possível analisar qual a taxa de desistência nas disciplinas.

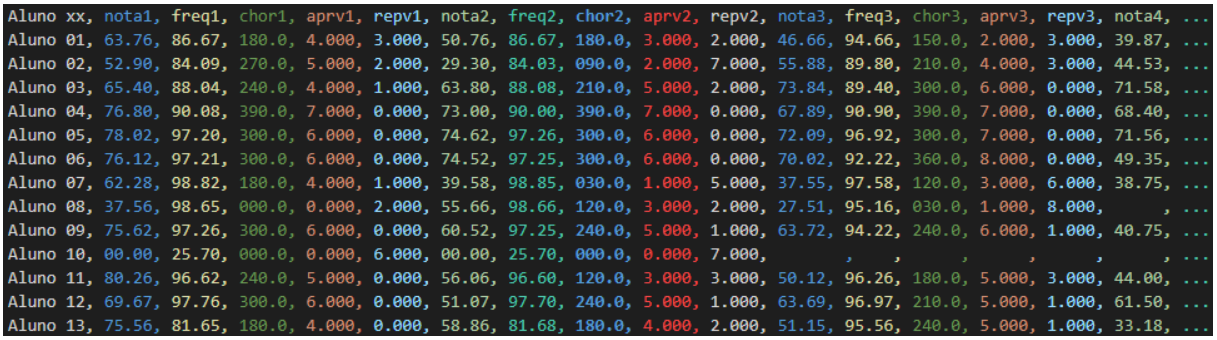
4.3 PRÉ PROCESSAMENTO E DISCRETIZAÇÃO DOS DADOS

O pré processamento consiste em transformar os dados de uma maneira que permita a aplicação das técnica necessárias. Por exemplo, a clusterização hierárquica utiliza como métrica de agrupamento a distância euclidiana entre cada as amostras. Contudo, é necessário preparar a base com valores numéricos - por isso é utilizada como métrica as notas, frequências, quantidade de disciplinas aprovadas e reprovadas e ainda a carga horária total do período.

A ideia é organizar as informações dos estudante em um espécie de série no tempo, isso significa organizar as informações existes ao longo dos períodos cursados pelos alunos. Por exemplo, dado que o primeiro semestre de um determinado aluno

foi realizado em 2010, então é necessário analisar todas as disciplinas cursadas no primeiro semestre desse ano, fazer uma média das notas e frequência, ver quantas disciplinas o aluno foi aprovado e em quantas disciplinas foi reprovado, além de somar toda carga horária adquirida com as disciplinas aprovadas.

FIGURA 4.3 – Exemplo dos dados pré processados



Aluno xx,	nota1,	freq1,	chor1,	aprv1,	repv1,	nota2,	freq2,	chor2,	aprv2,	repv2,	nota3,	freq3,	chor3,	aprv3,	repv3,	nota4,	...
Aluno 01,	63.76,	86.67,	180.0,	4.000,	3.000,	50.76,	86.67,	180.0,	3.000,	2.000,	46.66,	94.66,	150.0,	2.000,	3.000,	39.87,	...
Aluno 02,	52.90,	84.09,	270.0,	5.000,	2.000,	29.30,	84.03,	090.0,	2.000,	7.000,	55.88,	89.80,	210.0,	4.000,	3.000,	44.53,	...
Aluno 03,	65.40,	88.04,	240.0,	4.000,	1.000,	63.80,	88.08,	210.0,	5.000,	2.000,	73.84,	89.40,	300.0,	6.000,	0.000,	71.58,	...
Aluno 04,	76.80,	90.08,	390.0,	7.000,	0.000,	73.00,	90.00,	390.0,	7.000,	0.000,	67.89,	90.90,	390.0,	7.000,	0.000,	68.40,	...
Aluno 05,	78.02,	97.20,	300.0,	6.000,	0.000,	74.62,	97.26,	300.0,	6.000,	0.000,	72.09,	96.92,	300.0,	7.000,	0.000,	71.56,	...
Aluno 06,	76.12,	97.21,	300.0,	6.000,	0.000,	74.52,	97.25,	300.0,	6.000,	0.000,	70.02,	92.22,	360.0,	8.000,	0.000,	49.35,	...
Aluno 07,	62.28,	98.82,	180.0,	4.000,	1.000,	39.58,	98.85,	030.0,	1.000,	5.000,	37.55,	97.58,	120.0,	3.000,	6.000,	38.75,	...
Aluno 08,	37.56,	98.65,	000.0,	0.000,	2.000,	55.66,	98.66,	120.0,	3.000,	2.000,	27.51,	95.16,	030.0,	1.000,	8.000,	,	...
Aluno 09,	75.62,	97.26,	300.0,	6.000,	0.000,	60.52,	97.25,	240.0,	5.000,	1.000,	63.72,	94.22,	240.0,	6.000,	1.000,	40.75,	...
Aluno 10,	00.00,	25.70,	000.0,	0.000,	6.000,	00.00,	25.70,	000.0,	0.000,	7.000,	,	,	,	,	,	,	...
Aluno 11,	80.26,	96.62,	240.0,	5.000,	0.000,	56.06,	96.60,	120.0,	3.000,	3.000,	50.12,	96.26,	180.0,	5.000,	3.000,	44.00,	...
Aluno 12,	69.67,	97.76,	300.0,	6.000,	0.000,	51.07,	97.70,	240.0,	5.000,	1.000,	63.69,	96.97,	210.0,	5.000,	1.000,	61.50,	...
Aluno 13,	75.56,	81.65,	180.0,	4.000,	0.000,	58.86,	81.68,	180.0,	4.000,	2.000,	51.15,	95.56,	240.0,	5.000,	1.000,	33.18,	...

Fonte: Autor, 2018.

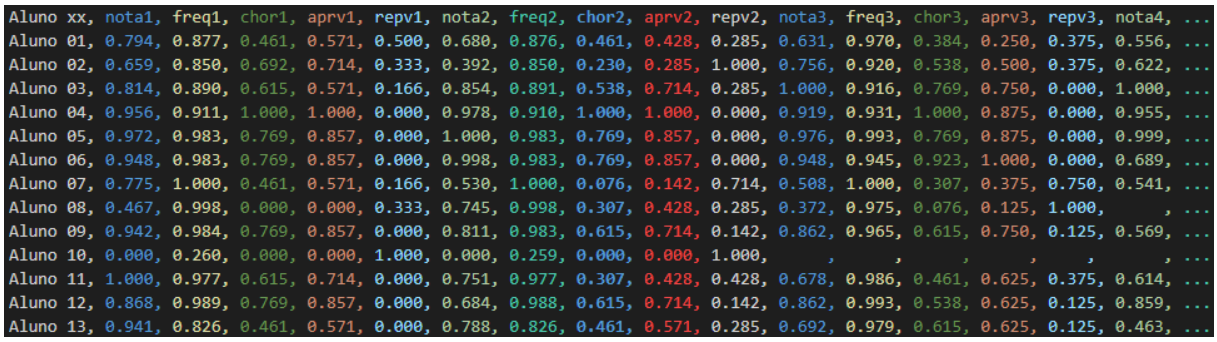
A Figura 4.3 representa o resultado parcial do pré-processamento dos dados, em que esses dados agora seguirão por dois caminhos distintos, um deles será realizado uma normalização destes valores (deixando-os num intervalo entre 0 a 1) e o outro seguirá por uma etapa de discretização. Nesse exemplo, é notável que cada linha representa um aluno com seu respectivo desempenho ao longo dos semestres, na qual o campo “nota1” representa a média das notas obtidas nas disciplinas cursadas em seu primeiro semestre. Os campos “freq”, “chor”, “aprv” e “repv” representam a média das frequências, a soma da carga horária equivalente, quantidade de disciplinas aprovadas e reprovadas, respectivamente e o número ao lado representa o semestre como, por exemplo, “freq2” representa a média das frequências no segundo semestre.

Ainda na Figura 4.3 é possível perceber alguns campos vazios como, por exemplo, do aluno 8, na qual o campo “nota4” não possui nenhum valor. Isso pode significar que o estudante, nesse semestre, trancou o curso ou realizou intercâmbio (mobilidade acadêmica) ou desistiu do curso. Na etapa de discretização esses campos serão preenchidos de acordo com o evento que o estudante estava naquele semestre.

Um dos caminhos é a normalização dos dados para aplicação das técnicas de clusterização hierárquica. Há diversos métodos de normalização dos dados, porém o método escolhido é o mais simples, que consiste em dividir todos os valores pelo maior valor dessa respectiva coluna. Dessa maneira, todos os dados ficarão em intervalos

entre 0 e 1. A Figura 4.4 é um exemplo da normalização dos dados do exemplo anterior.

FIGURA 4.4 – Exemplo dos dados normalizados



Fonte: Autor, 2018.

O outro caminho é a discretização dos dados, necessário para o treinamento do modelo com Modelos Ocultos de Markov. Sendo assim, o critério utilizado para discretizar a base pode ser visto na Tabela 1.

TABELA 1 – Critérios de discretização dos dados

Equivalente	Nota / Frequência	Carga Horária	Aprovações	Reprovações
A	de 90 a 100	de 360 a 660	de 10 a 20	0
B	de 70 a 90	de 240 a 360	de 7 a 9	de 0 a 3
C	de 50 a 70	de 90 a 240	de 4 a 6	de 4 a 6
D	de 0 a 50	de 0 a 90	de 0 a 3	de 7 a 9
E	0	0	0	de 10 a 20
M	Mobilidade	Mobilidade	Mobilidade	Mobilidade
T	Trancamento	Trancamento	Trancamento	Trancamento
F	Fracasso	Fracasso	Fracasso	Fracasso
S	Sucesso	Sucesso	Sucesso	Sucesso

O critério de discretização escolhido foi baseado em sistemas de ensino que utilizam letras para representar o desempenho dos estudantes. Entretanto, há algumas diferenças, pois a classificação é realizada de acordo com a média das notas, das frequência que um aluno obteve no semestre. Para a carga horária, aprovações e reprovações é realizado uma soma desses parâmetros para cada uma das disciplinas que o aluno cursou em um semestre, e então é realizada a classificação. O resultado da discretização pode ser visto na Figura 4.5.

FIGURA 4.5 – Exemplo dos dados discretizados

```
Aluno xx, nota1, freq1, chor1, aprv1, repv1, nota2, freq2, chor2, aprv2, repv2, nota3, freq3, chor3, aprv3, repv3, nota4, ...
Aluno 01,   C,     B,     C,     C,     B,     C,     B,     C,     D,     B,     D,     A,     C,     D,     B,     D,   ...
Aluno 02,   C,     B,     B,     C,     B,     D,     B,     C,     D,     B,     C,     B,     C,     C,     B,     D,   ...
Aluno 03,   C,     B,     B,     C,     B,     C,     B,     C,     B,     B,     B,     B,     B,     C,     A,     B,   ...
Aluno 04,   B,     A,     A,     B,     A,     B,     A,     C,     B,     A,     C,     A,     A,     B,     A,     C,   ...
Aluno 05,   B,     A,     B,     C,     A,     B,     A,     B,     B,     A,     B,     A,     B,     B,     A,     B,   ...
Aluno 06,   B,     A,     B,     C,     A,     B,     A,     B,     B,     A,     B,     A,     A,     B,     A,     D,   ...
Aluno 07,   C,     A,     C,     C,     B,     D,     A,     D,     D,     C,     D,     A,     C,     D,     C,     D,   ...
Aluno 08,   D,     A,     E,     E,     B,     C,     A,     C,     D,     B,     D,     A,     D,     D,     D,     T,   ...
Aluno 09,   B,     A,     B,     C,     A,     C,     A,     C,     B,     B,     C,     A,     C,     B,     B,     C,   ...
Aluno 10,   E,     D,     E,     E,     C,     E,     D,     E,     E,     B,     F,     F,     F,     F,     F,     F,   ...
Aluno 11,   B,     A,     B,     C,     A,     C,     A,     C,     D,     B,     C,     A,     C,     B,     B,     D,   ...
Aluno 12,   C,     A,     B,     C,     A,     C,     A,     C,     B,     B,     C,     A,     C,     B,     B,     C,   ...
Aluno 13,   B,     B,     C,     C,     A,     C,     B,     C,     B,     B,     C,     A,     C,     B,     B,     D,   ...
```

Fonte: Autor, 2018.

Nessa etapa é possível notar que os campos vazios mostrados anteriormente, agora estão preenchidos de acordo com a situação do estudante naquele momento. Por exemplo, o Aluno 08 no quarto período solicitou o trancamento do curso, enquanto que o Aluno 10 havia abandonado o curso já no terceiro período. Para melhor visualização desses cenários de trancamento, mobilidade, fracasso e sucesso foi selecionado os alunos dos exemplos, porém somente com as informações das notas desses estudantes ao longo dos semestres. O exemplo pode ser visto na Figura 4.6.

FIGURA 4.6 – Exemplo das situações

Aluno xx	nota1	nota2	nota3	nota4	nota5	nota6	nota7	nota8	nota9	nota10	nota11	nota12	nota13	nota14	nota15	nota16	nota17	nota18	nota19	nota20
Aluno 01	D	D	D	C	C	C	B	C	C	D	C	B	C	C	C	C	D	D	D	F
Aluno 02	D	B	D	D	D	B	C	D	C	C	C	C	C	D	T	C	B	C	C	F
Aluno 03	D	D	E	D	E	D	D	D	E	S	S	S	S	S	S	S	S	S	S	S
Aluno 04	E	E	E	D	E	D	E	C	E	T	M	M	M	E	M	D	D	E	D	E
Aluno 05	E	E	E	E	E	E	E	M	M	E	M	E	E	D	E	S	S	S	S	S
Aluno 06	E	E	E	D	D	D	D	D	E	E	D	E	C	C	E	S	S	S	S	S
Aluno 07	D	C	C	C	D	C	D	C	D	T	T	F	F	F	F	F	F	F	F	F
Aluno 08	D	D	B	T	T	T	D	D	D	F	F	F	F	F	F	F	F	F	F	F
Aluno 09	E	D	D	D	E	E	E	E	E	E	E	E	S	S	S	S	S	S	S	S
Aluno 10	C	B	F	F	F	F	F	F	F	F	F	F	F	F	F	F	F	F	F	F
Aluno 11	E	D	D	D	E	D	E	T	E	E	D	E	D	D	E	E	S	S	S	S
Aluno 12	E	D	D	D	E	E	E	E	E	M	M	M	M	T	E	M	S	S	S	S
Aluno 13	E	D	D	C	E	D	E	D	E	D	D	D	D	D	E	D	E	E	S	S

Fonte: Autor, 2018.

O exemplo ilustra os 13 alunos com seus resultados obtidos desde o primeiro período até o seu último período cursado. As situações de sucesso, falha, trancamento ou mobilidade acadêmica são representados pelas letras “T”, “S”, “F” e “M”, respectivamente. A partir do momento que o estudante é classificado como fracasso ou sucesso, o mesmo permanece assim até o final da sequência, visto que não há possibilidade do estudante retornar no curso - salva exceção em que o aluno realiza o vestibular novamente, mas nesse caso há um outro número de matricula, ou seja, será considerado um estudante diferente do inicial.

Lembrando que tanto a etapa de normalização quanto a etapa de discretização é realizada para todos os alunos de acordo com a quantidade de semestres cursadas de cada um.

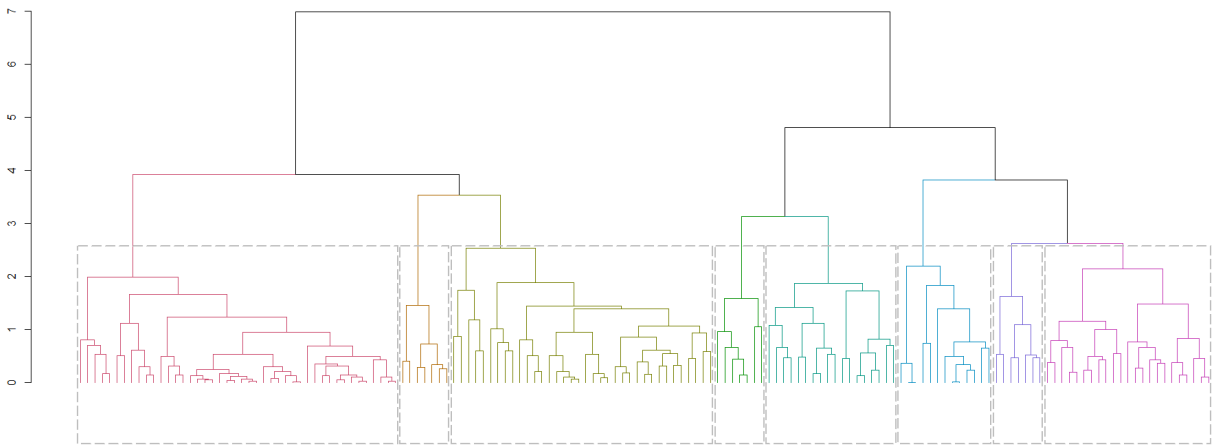
4.4 CLUSTERIZAÇÃO HIERÁRQUICA

A clusterização hierárquica é uma das etapas da modelagem do sistema, a qual tem como objetivo agrupar os dados similares a partir da distância euclidiana entre eles. O resultado dessa técnica será utilizado para determinar a quantidade de estados ocultos para treinamento do modelo.

Conforme explicado anteriormente, a aplicação dessa técnica é realizada com os dados normalizados e o código utilizado pode ser visto no Apêndice B. O procedimento utiliza as informações dos alunos de cada período, e em caso de um determinado estudante ainda não ter cursado o período 'x', ele é excluído e o processo é realizado normalmente. Isso significa que o número de estudantes nos primeiros procedimentos é maior que nos últimos.

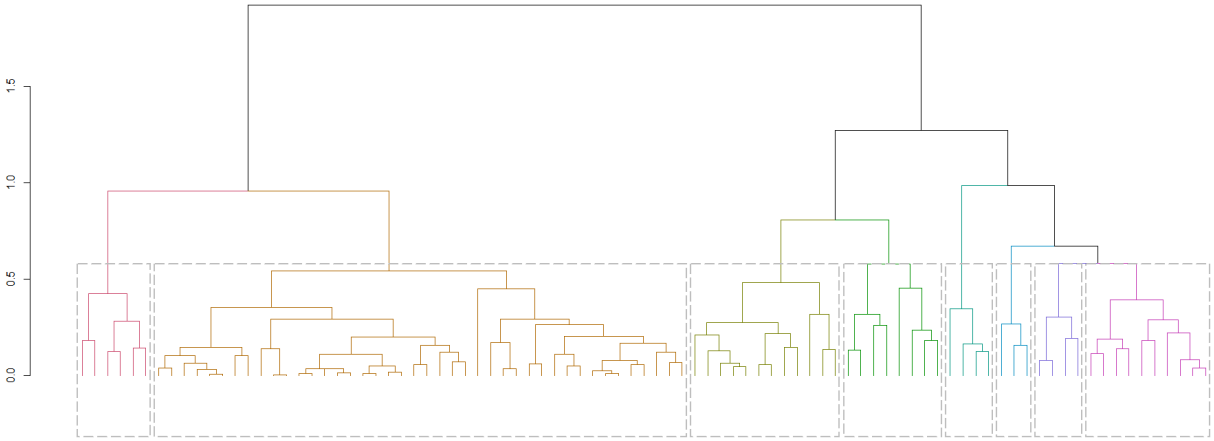
É possível analisar o resultado da técnica por meio de um artifício gráfico chamado dendrograma, para exemplificação há dois gráficos nas Figuras 4.7 e 4.8 que indicam o resultado da clusterização do 3º e do 8º semestre. Os resultados dos outros semestres pode ser visto no Apêndice C, nas Figuras C.1, C.2, C.3, C.4, C.5, C.6, C.7 e C.8.

FIGURA 4.7 – Dendrograma do 3º semestre



Fonte: Autor.

FIGURA 4.8 – Dendrograma do 8º semestre



Fonte: Autor.

Após a aplicação da técnica, os dendrogramas apresentados nas Figuras 4.7 e 4.8 indicam os alunos que estão agrupados ao longo do eixo x e a distância em que o grupo foi formado ao longo do eixo y. As distâncias ao longo do eixo x não possuem sentido em um dendrograma, pois todas as observações são igualmente espaçadas para facilitar a leitura. Os grupos mais representativos são aqueles com maior comprimento no eixo y.

Para interpretar esse tipo de gráfico, é necessário entender o que ele representa. As linhas verticais representam os grupos, enquanto que as linhas horizontais significam quais grupos estão se juntando. Por exemplo, na parte mais baixa do gráfico, os grupos são formados por apenas um estudante e subindo ao longo do eixo y, haverá um linha horizontal ligando duas linhas verticais, significando que um novo grupo foi formado a partir de dois outros grupos. O eixo y representa o quão próxima as observações eram quando foram mescladas em grupos.

Analisando os dendrogramas apresentados, identificam-se oito grupos distintos estando e para destacá-los foram desenhados retângulos pontilhados na imagem. Isso não quer dizer que olhar para soluções com mais ou menos grupos não teria sentido, mas os dados parecem sugerir oito grupos e esse número será utilizado como a quantidade de estados ocultos para o modelo.

Além da análise gráfica, o software R permite analisar em detalhes os grupos formados, ou seja, ver quais estudantes pertencem a quais grupos, assim como seus parâmetros de desempenho (nota, freq., etc.). Como esperado, um dos grupos é formado pelos estudantes com um desempenho alto e o um outro grupo é formado

por estudantes com desempenho muito baixo. O restante dos grupos é formado ou pelos estudantes possuírem notas e frequência parecidas ou por apresentarem uma quantidade similar de disciplinas aprovadas e reprovadas.

4.5 TREINAMENTO DO MODELO

Um Modelo Oculto de Markov é um artifício estatístico utilizado para modelar eventos. É uma extensão de uma Cadeia de Markov que pode ser caracterizada por determinar parâmetros desconhecidos através de parâmetros observáveis (conhecidos).

De acordo com as analises anteriores, o modelo foi construído considerando oito estados distintos para o treinamento e toda a base de dados discretizada. Após o treinamento, é necessário realizar uma análise para verificar como os conjuntos foram agrupados. O algoritmo para o treinamento do modelo pode ser visto no Apêndice D, no qual também foi desenvolvido utilizando o software R.

Para isso, o primeiro passo é observar o vetor das probabilidades dos estados iniciais, ou seja, o vetor coluna que indica as probabilidades da sequência iniciar em determinado estado. A equação 4.1 representa essas probabilidades.

$$\pi = \begin{matrix} \textbf{E1} & \textbf{E2} & \textbf{E3} & \textbf{E4} & \textbf{E5} & \textbf{E6} & \textbf{E7} & \textbf{E8} \end{matrix} \\ \begin{pmatrix} 0.247 & 0.018 & 0.118 & 0.163 & 0.000 & 0.455 & 0.000 & 0.000 \end{pmatrix} \tag{4.1}$$

A partir dessas probabilidades iniciais, é possível concluir que nenhum aluno pode iniciar a sequência com os estados 5, 7 e 8 (sendo representados na matriz como E5, E7 e E8, respectivamente), ficando limitado aos outros estados. Um possível significado para isso, é que os estados que não podem ser iniciados representam estados de sucesso ou falha e até mesmo um estado para mobilidade acadêmica, tendo em vista que nenhum aluno além de não poder atingir o sucesso no primeiro semestre esse aluno não pode realizar mobilidade acadêmica. Com relação os estudantes que desistiram já no primeiro semestre não foram considerados na modelagem.

Outra análise importante é a matriz com as probabilidades de transição entre os estados e as matrizes de emissão de cada um dos parâmetros utilizados na modelagem (notas, frequência, carga horária, quantidade de disciplinas aprovadas e reprovadas). Essas matrizes de probabilidades podem ser visualizadas nas equações 4.2, 4.3, 4.4,

4.5, 4.6 e 4.7 respectivamente.

$$A = \begin{array}{c} \\ \textbf{E1} \\ \textbf{E2} \\ \textbf{E3} \\ \textbf{E4} \\ \textbf{E5} \\ \textbf{E6} \\ \textbf{E7} \\ \textbf{E8} \end{array} \begin{array}{c} \begin{array}{cccccccc} \textbf{E1} & \textbf{E2} & \textbf{E3} & \textbf{E4} & \textbf{E5} & \textbf{E6} & \textbf{E7} & \textbf{E8} \end{array} \\ \begin{pmatrix} 0.353 & 0.107 & 0.316 & 0.012 & 0.000 & 0.152 & 0.003 & 0.058 \\ 0.057 & 0.502 & 0.156 & 0.014 & 0.005 & 0.031 & 0.000 & 0.234 \\ 0.129 & 0.105 & 0.428 & 0.000 & 0.000 & 0.029 & 0.006 & 0.303 \\ 0.017 & 0.003 & 0.000 & 0.702 & 0.020 & 0.182 & 0.076 & 0.000 \\ 0.000 & 0.077 & 0.000 & 0.385 & 0.538 & 0.000 & 0.000 & 0.000 \\ 0.178 & 0.056 & 0.071 & 0.194 & 0.003 & 0.451 & 0.042 & 0.004 \\ 0.000 & 0.000 & 0.000 & 0.000 & 0.000 & 0.000 & 1.000 & 0.000 \\ 0.000 & 0.000 & 0.000 & 0.000 & 0.000 & 0.000 & 0.000 & 1.000 \end{pmatrix} \end{array} \tag{4.2}$$

A matriz de transição tem o seguinte conceito: as colunas são os estados futuros e as linhas são o estado atual. Para verificar a probabilidade de transitar para o estado seguinte basta olhar para os valores, por exemplo, a probabilidade de transitar do estado 1 para o estado 6 é de 15,2% ou a probabilidade do estado 5 permanecer no mesmo estado é de 53,8%.

Essa matriz carrega informações importantes, pois os estados 7 e 8 são estados absorventes, ou seja, no momento que a sequência atingir um desses estados, ela permanecerá sempre nesse estado. Os outros estados transitam entre si, sendo notável uma alta probabilidade de permanecem no mesmo estado, como por exemplo, o estado 4, existe uma probabilidade dele permanecer nele mesmo de 70,2%.

$$B_1 = \begin{array}{c} \\ \textbf{E1} \\ \textbf{E2} \\ \textbf{E3} \\ \textbf{E4} \\ \textbf{E5} \\ \textbf{E6} \\ \textbf{E7} \\ \textbf{E8} \end{array} \begin{array}{c} \begin{array}{ccccccccc} \textbf{A} & \textbf{B} & \textbf{C} & \textbf{D} & \textbf{E} & \textbf{F} & \textbf{M} & \textbf{S} & \textbf{T} \end{array} \\ \begin{pmatrix} 0.000 & 0.000 & 0.109 & 0.888 & 0.003 & 0 & 0 & 0 & 0 \\ 0.000 & 0.000 & 0.000 & 0.000 & 0.000 & 0 & 0 & 0 & 1 \\ 0.000 & 0.000 & 0.000 & 0.440 & 0.560 & 0 & 0 & 0 & 0 \\ 0.101 & 0.834 & 0.064 & 0.000 & 0.000 & 0 & 0 & 0 & 0 \\ 0.000 & 0.000 & 0.000 & 0.000 & 0.000 & 0 & 1 & 0 & 0 \\ 0.043 & 0.298 & 0.615 & 0.044 & 0.000 & 0 & 0 & 0 & 0 \\ 0.000 & 0.000 & 0.000 & 0.000 & 0.000 & 0 & 0 & 1 & 0 \\ 0.000 & 0.000 & 0.000 & 0.000 & 0.000 & 1 & 0 & 0 & 0 \end{pmatrix} \end{array} \tag{4.3}$$

As matrizes de emissão carregam a informação do que realmente o modelo representa. Analisando a matriz de emissão das notas, da equação 4.3, é possível notar que o estado 4 tem uma probabilidade de 10,1% de gerar o descritor A, enquanto que

a probabilidade de gerar o descritor B é muito maior, em torno de 83,4%. Entretanto, somente a matriz de emissão das notas não representa muita coisa, é necessário analisar todas as outras, que estão descritas nas equações 4.4, 4.5, 4.6 e 4.7.

$$B_2 = \begin{array}{c} \\ \textbf{E1} \\ \textbf{E2} \\ \textbf{E3} \\ \textbf{E4} \\ \textbf{E5} \\ \textbf{E6} \\ \textbf{E7} \\ \textbf{E8} \end{array} \begin{array}{c} \begin{array}{ccccccccc} \textbf{A} & \textbf{B} & \textbf{C} & \textbf{D} & \textbf{E} & \textbf{F} & \textbf{M} & \textbf{S} & \textbf{T} \end{array} \\ \begin{pmatrix} 0.193 & 0.445 & 0.257 & 0.099 & 0.006 & 0 & 0 & 0 & 0 \\ 0.000 & 0.000 & 0.000 & 0.000 & 0.000 & 0 & 0 & 0 & 1 \\ 0.018 & 0.093 & 0.165 & 0.638 & 0.087 & 0 & 0 & 0 & 0 \\ 0.719 & 0.262 & 0.019 & 0.000 & 0.000 & 0 & 0 & 0 & 0 \\ 0.000 & 0.000 & 0.000 & 0.000 & 0.000 & 0 & 1 & 0 & 0 \\ 0.617 & 0.364 & 0.014 & 0.000 & 0.004 & 0 & 0 & 0 & 0 \\ 0.000 & 0.000 & 0.000 & 0.000 & 0.000 & 0 & 0 & 1 & 0 \\ 0.000 & 0.000 & 0.000 & 0.000 & 0.000 & 1 & 0 & 0 & 0 \end{pmatrix} \end{array} \tag{4.4}$$

$$B_3 = \begin{array}{c} \\ \textbf{E1} \\ \textbf{E2} \\ \textbf{E3} \\ \textbf{E4} \\ \textbf{E5} \\ \textbf{E6} \\ \textbf{E7} \\ \textbf{E8} \end{array} \begin{array}{c} \begin{array}{ccccccccc} \textbf{A} & \textbf{B} & \textbf{C} & \textbf{D} & \textbf{E} & \textbf{F} & \textbf{M} & \textbf{S} & \textbf{T} \end{array} \\ \begin{pmatrix} 0.000 & 0.016 & 0.427 & 0.547 & 0.010 & 0 & 0 & 0 & 0 \\ 0.000 & 0.000 & 0.000 & 0.000 & 0.000 & 0 & 0 & 0 & 1 \\ 0.000 & 0.000 & 0.000 & 0.000 & 1.000 & 0 & 0 & 0 & 0 \\ 0.325 & 0.675 & 0.000 & 0.000 & 0.000 & 0 & 0 & 0 & 0 \\ 0.000 & 0.000 & 0.000 & 0.000 & 0.000 & 0 & 1 & 0 & 0 \\ 0.004 & 0.432 & 0.511 & 0.054 & 0.000 & 0 & 0 & 0 & 0 \\ 0.000 & 0.000 & 0.000 & 0.000 & 0.000 & 0 & 0 & 1 & 0 \\ 0.000 & 0.000 & 0.000 & 0.000 & 0.000 & 1 & 0 & 0 & 0 \end{pmatrix} \end{array} \tag{4.5}$$

$$B_4 = \begin{array}{c} \\ \textbf{E1} \\ \textbf{E2} \\ \textbf{E3} \\ \textbf{E4} \\ \textbf{E5} \\ \textbf{E6} \\ \textbf{E7} \\ \textbf{E8} \end{array} \begin{array}{c} \begin{array}{ccccccccc} \textbf{A} & \textbf{B} & \textbf{C} & \textbf{D} & \textbf{E} & \textbf{F} & \textbf{M} & \textbf{S} & \textbf{T} \end{array} \\ \begin{pmatrix} 0.000 & 0.000 & 0.062 & 0.928 & 0.010 & 0 & 0 & 0 & 0 \\ 0.000 & 0.000 & 0.000 & 0.000 & 0.000 & 0 & 0 & 0 & 1 \\ 0.000 & 0.000 & 0.000 & 0.000 & 1.000 & 0 & 0 & 0 & 0 \\ 0.056 & 0.325 & 0.616 & 0.004 & 0.000 & 0 & 0 & 0 & 0 \\ 0.000 & 0.000 & 0.000 & 0.000 & 0.000 & 0 & 1 & 0 & 0 \\ 0.000 & 0.002 & 0.744 & 0.254 & 0.000 & 0 & 0 & 0 & 0 \\ 0.000 & 0.000 & 0.000 & 0.000 & 0.000 & 0 & 0 & 1 & 0 \\ 0.000 & 0.000 & 0.000 & 0.000 & 0.000 & 1 & 0 & 0 & 0 \end{pmatrix} \end{array} \tag{4.6}$$

$$B_5 = \begin{array}{c} \\ \textbf{E1} \\ \textbf{E2} \\ \textbf{E3} \\ \textbf{E4} \\ \textbf{E5} \\ \textbf{E6} \\ \textbf{E7} \\ \textbf{E8} \end{array} \begin{array}{c} \begin{array}{ccccccccc} \textbf{A} & \textbf{B} & \textbf{C} & \textbf{D} & \textbf{E} & \textbf{F} & \textbf{M} & \textbf{S} & \textbf{T} \end{array} \\ \begin{pmatrix} 0.000 & 0.456 & 0.493 & 0.035 & 0.016 & 0 & 0 & 0 & 0 \\ 0.000 & 0.000 & 0.000 & 0.000 & 0.000 & 0 & 0 & 0 & 1 \\ 0.000 & 0.377 & 0.557 & 0.066 & 0.000 & 0 & 0 & 0 & 0 \\ 0.860 & 0.140 & 0.000 & 0.000 & 0.000 & 0 & 0 & 0 & 0 \\ 0.000 & 0.000 & 0.000 & 0.000 & 0.000 & 0 & 1 & 0 & 0 \\ 0.314 & 0.686 & 0.000 & 0.000 & 0.000 & 0 & 0 & 0 & 0 \\ 0.000 & 0.000 & 0.000 & 0.000 & 0.000 & 0 & 0 & 1 & 0 \\ 0.000 & 0.000 & 0.000 & 0.000 & 0.000 & 1 & 0 & 0 & 0 \end{pmatrix} \end{array} \tag{4.7}$$

Utilizando o mesmo exemplo dado anteriormente, só que analisando todas as matrizes de emissão, é perceptível que o estado 4 representa os estudantes, em sua grande maioria, com boas notas. Isso significa que a média das suas notas estão entre 70 e 100. O mesmo ocorre em todos os outros parâmetros:

- a frequência desses estudantes está maior que 75%;
- a carga horária efetiva (contabilizada somente em caso de aprovação) maior que 240 horas;
- a quantidade de disciplinas aprovadas maior que 7 e
- a quantidade de disciplinas reprovadas menor que 3, com sua grande maioria sem reprovações.

Pode-se concluir que o estado 4 é um indicador de bom desempenho para os estudantes e apresenta uma probabilidade de 7,6% de transição para o estado de sucesso e 0% pro estado de fracasso. Os alunos classificados aqui, apresentam um desempenho excelente com boas chances de atingirem o sucesso. A explicação de cada um dos outros estados podem ser vistos na sequência.

A análise para os estado 2, 5, 7 e 8 é imediata, pois as matrizes de emissão indicam que todos os estudantes que trancaram o curso são classificados no estado 2. O mesmo ocorre para os estudantes que realizaram mobilidade acadêmica, que estão classificados no estado 5 e os estudantes que atingiram o sucesso ou fracassaram estão descritos nos estado 7 e 8, respectivamente.

O estado 3 é o estado mais crítico, pois verificando a matriz de transição dos estados, a probabilidade de transitar para o estado de sucesso é de 0,6%, enquanto para o estado de falha é de 30,3%. Analisando a matriz de emissão das disciplinas

aprovadas é possível chegar nessa conclusão também, pois essa matriz indica que o índice de aprovação dos estudantes é zero, ou seja, não há nenhuma aprovação no semestre. Para reforçar essa análise, basta olhar para as matrizes de emissão das notas e das reprovações, que apresentam uma média de notas menor que 5, sendo a grande maioria apresentando notas zero e a carga horária zerada também, o que era esperado, visto que não apresentou nenhuma aprovação.

O estado 1 apresenta um desempenho ruim dos estudantes, com a médias das notas menor que 5 em sua maioria, e em pouquíssimos casos maior que 5 mas nunca maior que 7, isso ocorre em aproximadamente 10% dos casos. A carga horária é menor que 240h, sendo um número bem baixo, visto que a carga horária dos períodos do curso é superior a esse valor. O índice de reprovação muito alto, de até 9 disciplinas por semestre, e claro que o índice de aprovação é baixo, sendo inferior que 4 disciplinas por semestre. Esses estado não é tão crítico quanto o estado 3, mas é ruim. Olhando para a matriz de transição, esse estado possui uma probabilidade de 0,3% para transitar para o estado de sucesso e 5,8% para transitar para o estado de falha.

O estado 6 é considerado um estado regular, visto que em sua grande maioria apresenta um desempenho satisfatório. A média das notas nesse estado, é maior que 5 o que resulta em um índice de aprovação maior que 4 disciplinas e uma carga horária maior que 240h. O índice de reprovação reforça a ideia desse estado ser considerado regular, apresentando um número menor que 3 reprovações por semestre e nunca maior que isso. Já as probabilidades de transição para os estados de sucesso ou fracasso estão em 4,2% e 0,4%, respectivamente.

É possível fichar os estados da seguinte maneira: o estado 1 é ruim; o estado 2 é do trancamento; estado 3 é péssimo; o estado 4 é excelente; o estado 5 é da mobilidade acadêmica; o estado 6 é regular; o estado 7 é do sucesso e o estado 8 é do fracasso. Esse fichamento será utilizado quando o modelo for aplicado nos estudantes atuais do curso, tentando tornar mais autoexplicativo quando o modelo classificar os estudantes em algum estado.

4.6 EXTRAÇÃO DAS INFORMAÇÕES DA MATRIZ DE TRANSIÇÃO

Mesmo sendo um modelo oculto de Markov é possível observar a matriz de transição de estados como uma cadeia de Markov absorvente e encontrar resultados

interessantes que serão utilizados para a validação do modelo. Os resultados de interesse são as respostas para as seguintes perguntas: (a) a probabilidade que um processo seja absorvido, (b) número médio de passos para o processo ser absorvido e (c) número médio de passos em que o processo permanece nos estados transientes antes de ser absorvido.

A primeira etapa é transcrever a matriz de transição da equação 4.2 na forma canônica, apresentado no Capítulo 3 pela equação 3.7 e então aplicar nas equações 3.8, 3.12 e 3.15, para encontrar a matriz fundamental, o número médio de passos até a absorção e a probabilidade de um estado ser absorvido, respectivamente. Os resultados obtidos podem ser vistos nas equações 4.8, 4.9 e 4.10.

O algoritmo utilizado para encontrar essas matrizes está presenta no Apêndice E e também foi utilizado o software R para esses cálculos.

A matriz fundamental apresenta resultados interessantes, visto que são poucos passos que o processo permanece nos estados de transição, em média 1 até 2 passos. Há casos em que o esse número é 4 ou 5, mas são casos atípicos. Utilizando a matriz da equação 4.8 supondo que o sistema inicia-se no estado 5 o número médio de passos que o sistema permanece nos estados 1, 2 e 3 é menor que 1, enquanto que nos estado 4, 5 e 6 é de, aproximadamente, 4, 2 e 2.

$$\mathbf{N} = \begin{array}{c} \\ \textbf{E1} \\ \textbf{E2} \\ \textbf{E3} \\ \textbf{E4} \\ \textbf{E5} \\ \textbf{E6} \end{array} \begin{array}{c} \begin{array}{cccccc} \textbf{E1} & \textbf{E2} & \textbf{E3} & \textbf{E4} & \textbf{E5} & \textbf{E6} \end{array} \\ \begin{pmatrix} 2.285 & 0.977 & 1.664 & 0.916 & 0.057 & 1.080 \\ 0.592 & 2.433 & 1.059 & 0.564 & 0.054 & 0.543 \\ 0.688 & 0.714 & 2.386 & 0.424 & 0.029 & 0.498 \\ 0.950 & 0.707 & 0.971 & 5.062 & 0.239 & 2.032 \\ 0.890 & 0.994 & 0.986 & 4.312 & 2.375 & 1.784 \\ 1.233 & 0.914 & 1.308 & 2.224 & 0.125 & 3.022 \end{pmatrix} \end{array} \tag{4.8}$$

Somando o número médio de passos das linhas da matriz fundamental, é possível encontrar o número médio de passos que o processo permaneça nos estados

transientes. Esse número pode ser visto na equação 4.9.

$$\mathbf{t} = \begin{matrix} \textbf{E1} \\ \textbf{E2} \\ \textbf{E3} \\ \textbf{E4} \\ \textbf{E5} \\ \textbf{E6} \end{matrix} \begin{pmatrix} 6.638 \\ 6.739 \\ 8.374 \\ 13.503 \\ 2.877 \\ 8.958 \end{pmatrix} \tag{4.9}$$

Supondo que o processo inicia-se no estado 4 o número médio de passos que o sistema permanece no sistema é de 14 (arredondado o valor para cima). Lembrando que esse é o melhor estado que um aluno pode ser classificado, faz sentido, visto que enquanto estudante estiver indo bem, ele permanece no sistema saindo apenas quando se forma (obtêm o sucesso). Outra conclusão a cerca dessa matriz, é que o estado regular (estado 6) está logo atrás no número médio de passos, seguido do pior estado (estado 3). O estado péssimo possui um número médio de passos até que alto, pois grande parte dos estudantes classificados nesse estado possuem uma alta probabilidade de trancar o curso (ir para o estado 2), antes de abandonar o curso, conforme visto na matriz de transição dos estados.

O resultado mais importante para a validação dos dados é a matriz com as probabilidades de absorção representada na equação 4.10.

$$\mathbf{B} = \begin{matrix} & \textbf{E7} & \textbf{E8} \\ \textbf{E1} \\ \textbf{E2} \\ \textbf{E3} \\ \textbf{E4} \\ \textbf{E5} \\ \textbf{E6} \end{matrix} \begin{pmatrix} 0.132 & 0.868 \\ 0.074 & 0.926 \\ 0.069 & 0.931 \\ 0.477 & 0.523 \\ 0.410 & 0.590 \\ 0.306 & 0.694 \end{pmatrix} \tag{4.10}$$

Esse resultado indica a probabilidade de um determinado estado terminar em um dos estados absorventes no fim do processo. É importante analisar os estados que possuem maior probabilidade de terminar em sucesso, que são os estados 4, 5 e 6, fichados como os estados excelente, mobilidade e regular, respectivamente. Por outro

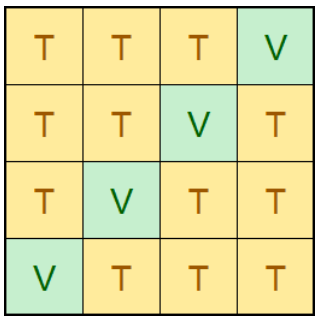
lado, os estados com maior probabilidade de terminar no fracasso são os estados 1, 2 e 3, fichados como ruim, trancamento e péssimo, respectivamente.

Analisar essa matriz assusta inicialmente, pois praticamente todos os estados possuem maior probabilidade de falha que sucesso o que faz sentido, visto que no treinamento do modelo quase que 70% dos estudantes presentes na base de dados falharam. Outra análise, é que essa matriz reforça a interpretação do modelo, no qual ressalta que o melhor estado é o estado 4 e o pior é o estado 3.

4.7 VALIDAÇÃO DOS RESULTADOS

O desenvolvimento de um modelo analítico envolve um processo de avaliação para quantificar e qualificar sua eficiência. Esse processo é conhecido como a validação dos resultados. Essa etapa visa levantar os parâmetros de eficiência do modelo e com esses, confirmar se o modelo é bom ou não. A técnica a ser utilizado é a Validação Cruzada (ou do inglês *Cross-Validation*), utilizando o método *k-fold*, o qual tem como objetivo dividir o conjunto de dados em k subgrupos de mesmo tamanho. Para o modelo, será definido $k = 4$, ou seja, a base será dividida em quarto porções iguais, na qual cada porção possuí 25% dos dados da base.

FIGURA 4.9 – Método de validação *k-fold* com $k = 4$



Fonte: Autor, 2018.

A Figura 4.9 apresenta uma matriz quadrada de ordem 4, na qual a letra “T” representa 25% da base de dados que será destinada ao treinamento do modelo e a letra “V” representa outros 25% da base que será destinada a validação desse modelo. Cada linha dessa matriz representa uma etapa de validação e as colunas representam parcelas de 25% da base de dados com sua respectiva destinação, se a parcela vai para o treinamento do modelo ou para validação.

Com o modelo construído com 75% da base de dados, os dados restantes são testados nesse modelo com a utilização do algoritmo de Viterbi, o qual é utilizado para encontrar a sequência mais provável de estados ocultos a partir de uma sequência dos eventos observados (notas, frequência, etc.). Em outras palavras, a entrada do algoritmo são esses dados pré processados e discretizados e a saída será o caminho mais provável que determinado estudante percorreu antes de entrar em um estado absorvente. O algoritmo de validação pode ser visto no Apêndice F.

A partir dos estados identificados pelo algoritmo será possível determinar os parâmetros, os quais compõe uma matriz conhecida como Matriz de Confusão e essa pode ser visualizada na Figura 4.10. É uma matriz composta por duas linhas e duas colunas descrevendo a quantidade de “Verdadeiro Positivos”, “Falso Positivos”, “Falso Negativos” e “Verdadeiro Negativos”.

FIGURA 4.10 – Matriz de Confusão

Fonte: (MNMK, 2013).

A explicação desses parâmetros e a maneira utilizada para quantificá-los podem ser vistos nos itens a seguir:

- Verdadeiro Positivos (TV): quando o modelo detecta algo como positivo e é verdadeiro. Será considerado quando um estudante for classificado no melhor estado (estado 4) pelo algoritmo e foi um aluno que se formou.
- Falso Positivos (FP): quando o modelo detecta algo como positivo e é falso. Será considerado quando um estudante for classificado no pior estado (estado 3) e esse aluno realmente não atingiu o sucesso, ou seja, o estudante abandonou o curso.

- Falso Negativos (FN): quando o modelo detecta algo como negativo e é falso. Será determinado quando um estudante for classificado no melhor estado (estado 4) e esse aluno também não obteve sucesso.
- Verdadeiro Negativos (VN): quando o modelo detecta algo como negativo e é verdadeiro. Será considerado quando um estudante for classificado no pior estado (estado 3) e esse aluno não fracassou, ou seja, teve êxito.

Os resultados encontrados podem ser visto na Figura 4.11 com o número total de cada um dos parâmetros.

FIGURA 4.11 – Matriz de confusão

Fonte: Autor, 2018.

Com a quantidade levantada de cada um desses parâmetros, será possível determinar a eficiência do modelo com algumas métricas. Essas métricas são conhecidas como: acurácia, sensibilidade, especificidade, eficiência e precisão e podem ser calculadas pelas Equações 4.11, 4.12, 4.13, 4.14, 4.15 e 4.16, respectivamente.

$$Acuracia = \frac{VP + VN}{VP + FN + FP + VN} = \frac{45 + 145}{45 + 23 + 11 + 145} = 0,8482 \tag{4.11}$$

A acurácia é a proporção de predições corretas, significando que em aproximadamente 85% dos casos o modelo acerta na predição. Esse parâmetro não leva em consideração se a predição é positiva ou negativa e pode gerar uma conclusão errada sobre o desempenho, pois essa métrica não divide entre quantos acertos foram positivos nem quantos acertos foram negativos.

$$Sensibilidade = \frac{VP}{VP + FN} = \frac{45}{45 + 11} = 0,8036 \tag{4.12}$$

Com um resultado de 80% a sensibilidade é a porcentagem de verdadeiros positivos, ou seja, é a capacidade que o sistema tem em prever os casos positivos.

$$\textit{Especificidade} = \frac{\textit{TN}}{\textit{TN} + \textit{FP}} = \frac{145}{145 + 23} = 0,8631 \tag{4.13}$$

Ao contrário da sensibilidade, a especificidade pode ser definida como a proporção de verdadeiros negativos, ou seja, a capacidade do sistema em predizer corretamente os casos negativos. O resultado obtido pelo modelo foi em torno de 86%.

Na pratica, a sensibilidade e a especificidade variam em direções opostas. Em outras palavras, quando um modelo é muito sensível a positivos, a tendência é que esse modelo gere muitos falsos positivos (FP) e o mesmo ocorre quando um modelo é sensível aos negativos, gerando muitos falsos negativos (FN). Dessa maneira, um cenário ideal seria ter 100% de sensibilidade e 100% de especificidade, circunstância que é raramente alcançada. A eficiência é a média aritmética entre a sensibilidade e a especificidade, essa relação pode ser vista na Equação 4.14.

$$\textit{Eficiencia} = \frac{\textit{Sensibilidade} + \textit{Especificidade}}{2} = 0,8333 \tag{4.14}$$

A última métrica, precisão é a quantidade de previsões correta, seja elas positivas ou negativos. Conhecida também pode Valor Preditivo Positivo (VPP), é a proporção de verdadeiros positivos (TP) a todas as predições positivas corretas. O mesmo acontece para as predições negativas, que é conhecida como Valor Preditivo Negativo (VPN). Esses valores podem ser encontrados por meio das Equações 4.15 e 4.16, respectivamente.

$$\textit{VPP} = \frac{\textit{TP}}{\textit{TP} + \textit{FP}} = \frac{45}{45 + 23} = 0,6618 \tag{4.15}$$

$$\textit{VPN} = \frac{\textit{TN}}{\textit{TN} + \textit{FN}} = \frac{145}{145 + 11} = 0,9295 \tag{4.16}$$

De todas as predições positivas que o modelo gerou, em 66% dos casos foi acertado e para as previsões negativas foi em torno de 93%. Isso significa que o modelo acerta bem mais os casos negativos do que os positivos e isso não é ruim, visto que o modelo identificando os casos negativos os gestores da educação do curso podem gerar alertas para o curso tomar providências evitando que esses alunos desistam. O

problema está em quando o modelo errar a predição, dizendo que um aluno irá obter sucesso, quando na verdade o aluno irá desistir.

Essas métricas são amplamente conhecidas na comunidade acadêmica e pertinentes na avaliação de modelos. Entretanto, há mais um critério a ser analisado: o número de semestres antes que o modelo consegue identificar que um estudante irá falhar. Esse fator é de extrema importância para validação do modelo, pois com o sistema detectando possíveis evasões semestres antes, há tempo de traçar ações para mitigar esse problema. Para encontrar esse valor, foi verificado em qual instante o modelo identifica um estudante com o estado 3 (o pior estado) e o momento que o estudante falhou. A média das diferenças entre esses valores será o parâmetro de interesse. Esse pode ser visto na equação 4.17.

$$\textit{Semestres antes} = 2,6904 \tag{4.17}$$

Em média, foram encontrados 2,6904 semestres antes que o modelo é capaz de prever. Esse resultado é muito bom, visto que há tempo para os gestores da educação da instituição tomarem ações para evitar com que esses estudantes venham a abandonar a faculdade.

4.8 APLICAÇÃO DO MODELO NOS ALUNOS ATIVOS DO CURSO

A aplicação do modelo nos alunos ativos do curso acaba repetindo algumas etapas anteriores como, por exemplo, a etapa de coleta e pré-processamento dos dados. Entretanto, a coleta dos alunos é muito mais simples, pois no sistema da UFPR já possuí a extração de um relatório com as informações de todos os estudantes ativos do curso, ou seja, são aqueles estudantes que estão matriculados em disciplinas ou em alguma situação que não seja de formação nem desistência.

A etapa do pré-processamento consiste também na organização dos dados em uma espécie de série no tempo, com os mesmo parâmetros estudados anteriormente: média das notas e frequência, soma da carga horária efetiva, quantidade de disciplinas aprovadas e reprovadas em um semestre. Todas essas informações discretizadas de acordo com a Tabela 1 apresentada anteriormente.

São um total de 752 estudantes, incluindo estudantes tanto do curso Diurno quanto estudantes do curso Noturno. Mesmo que o modelo tenha sido confeccionado

somente com as informações conhecidas de estudantes do curso Noturno, esse será aplicado para os estudantes do outro turno nesse momento. Um dos trabalho futuros será confeccionar um modelo exclusivo para esses alunos, que muito provavelmente, apresentará um comportamento diferente, isso será explicado com mais detalhes no Capítulo 5.

FIGURA 4.12 – Exemplo do caminho percorrido pelos alunos

Aluno xxx	T1	T2	T3	T4	T5	T6	T7	T8	T9	T10	T11	T12	T13	T14	T15	T16	T17
Aluno 014	State 6	State 6	State 6	State 1	State 6	State 1	State 1	State 1	State 1	State 6	State 6	State 6	State 2	State 4	State 6	State 4	State 7
Aluno 018	State 4	State 6	State 6	State 6	State 4	State 6	State 5	State 5	State 5	State 1	State 1	State 1	State 1	State 1	State 1	State 1	State 1
Aluno 019	State 6	State 1	State 6	State 6	State 1	State 3	State 1	State 1	State 2	State 2	State 6	State 6	State 6	State 1	State 1	State 3	State 8
Aluno 032	State 6	State 6	State 2	State 6	State 6	State 6	State 4	State 4	State 4	State 6	State 6	State 6	State 6	State 1	State 6	State 7	State 7
Aluno 036	State 1	State 1	State 1	State 3	State 3	State 1	State 6	State 6	State 1	State 6	State 6	State 6	State 6	State 1	State 1	State 3	State 8
Aluno 067	State 6	State 6	State 1	State 3	State 1	State 6	State 6	State 1	State 1	State 1	State 1	State 1	State 8	State 8	State 8	State 8	State 8
Aluno 368	State 1	State 1	State 6	State 6	State 1	State 6	State 7	State 7	State 7	State 7	State 7	State 7	State 7	State 7	State 7	State 7	State 7
Aluno 369	State 4	State 6	State 4	State 4	State 4	State 4	State 6	State 7	State 7	State 7	State 7	State 7	State 7	State 7	State 7	State 7	State 7
Aluno 426	State 4	State 4	State 4	State 4	State 4	State 7	State 7	State 7	State 7	State 7	State 7	State 7	State 7	State 7	State 7	State 7	State 7
Aluno 427	State 4	State 6	State 4	State 4	State 7	State 7	State 7	State 7	State 7	State 7	State 7	State 7	State 7	State 7	State 7	State 7	State 7
Aluno 428	State 4	State 4	State 4	State 4	State 4	State 7	State 7	State 7	State 7	State 7	State 7	State 7	State 7	State 7	State 7	State 7	State 7
Aluno 651	State 4	State 7	State 7	State 7	State 7	State 7	State 7	State 7	State 7	State 7	State 7	State 7	State 7	State 7	State 7	State 7	State 7
Aluno 652	State 1	State 3	State 8	State 8	State 8	State 8	State 8	State 8	State 8	State 8	State 8	State 8	State 8	State 8	State 8	State 8	State 8

Fonte: Autor, 2018.

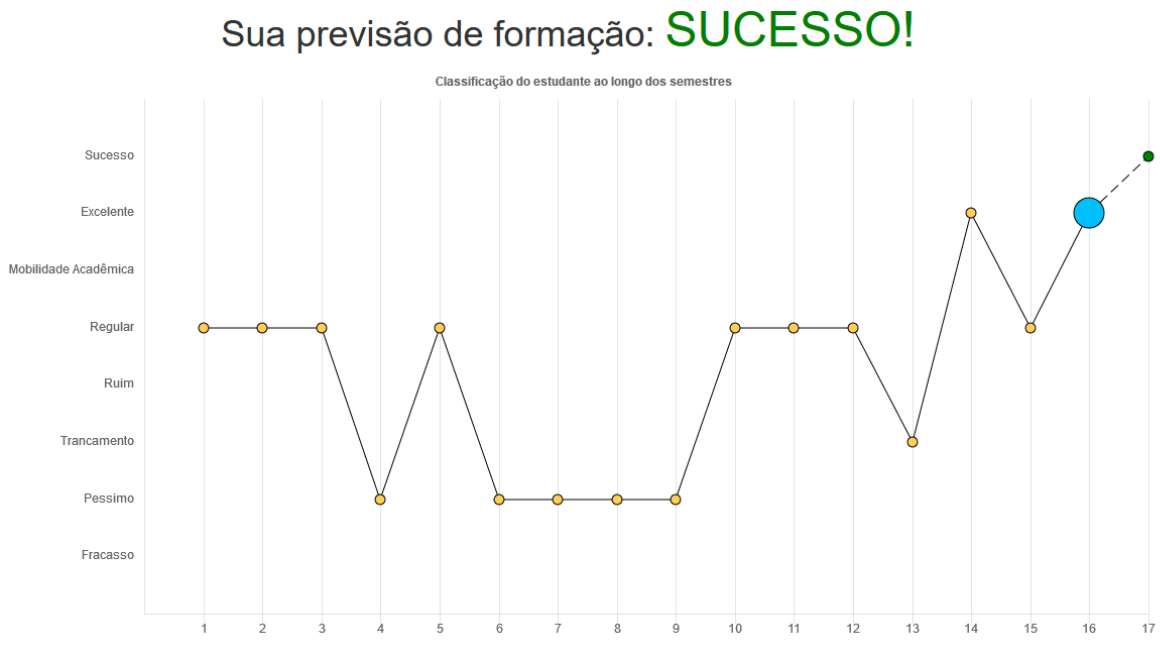
A Figura 4.12 ilustra a saída do algoritmo de Viterbi aplicado. Foram destacados os estados que representam o trancamento, a mobilidade acadêmica, o sucesso e o fracasso, que são os estados 2 (representado de amarelo), estado 5 (representado de azul), estado 7 (representado de verde) e estado 8 (representado de vermelho), respectivamente.

É interessante notar que o algoritmo previu os estados futuros terminando nos estados absorventes. Por exemplo, o Aluno 32 cursou até o 15º semestre e foi previsto que nos próximos semestres o estudante estaria no estado 7, que é o estado de sucesso. É possível aferir a quantidade de semestre que o algoritmo pode prever, entretanto é necessário analisar até o instante que o estudante for classificado em um dos estados absorventes (estado 7 ou 8).

Foi confeccionado uma página para cada estudante, onde cada um possui um caminho percorrido diferente, ou seja, cada aluno possui a sequência dos estados percorridos ao longo da sua vida acadêmica que o modelo identificou. Além dessa página possuir essa sequência, será mostrado a explicação de cada um dos estados destacando o estado atual do estudante. Essas páginas serão incluídas no sistema que esta em desenvolvimento no curso de Engenharia Elétrica, que pode ser visto na seguinte URL: http://eletrica.ufpr.br/situacaogrr. As Figuras 4.13, 4.14 e 4.15 são

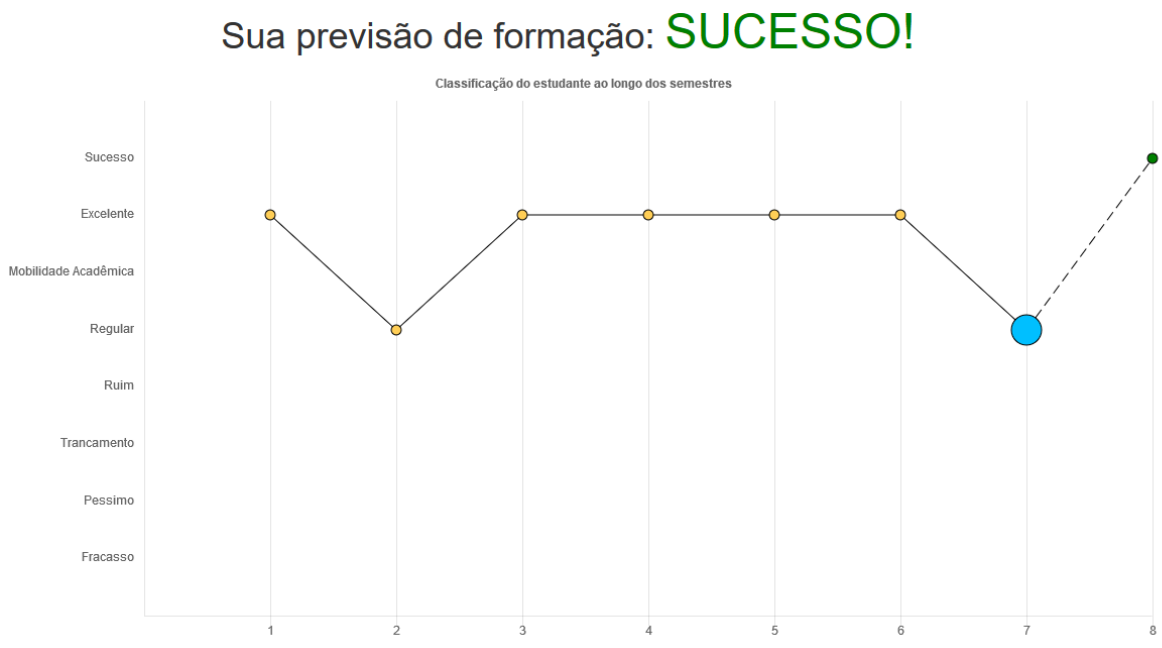
exemplos das sequências dos alunos 014, 368 e 019, enquanto que a Figura 4.16 é a explicação dos estados.

FIGURA 4.13 – Exemplo da página do aluno 014



Fonte: Autor, 2018.

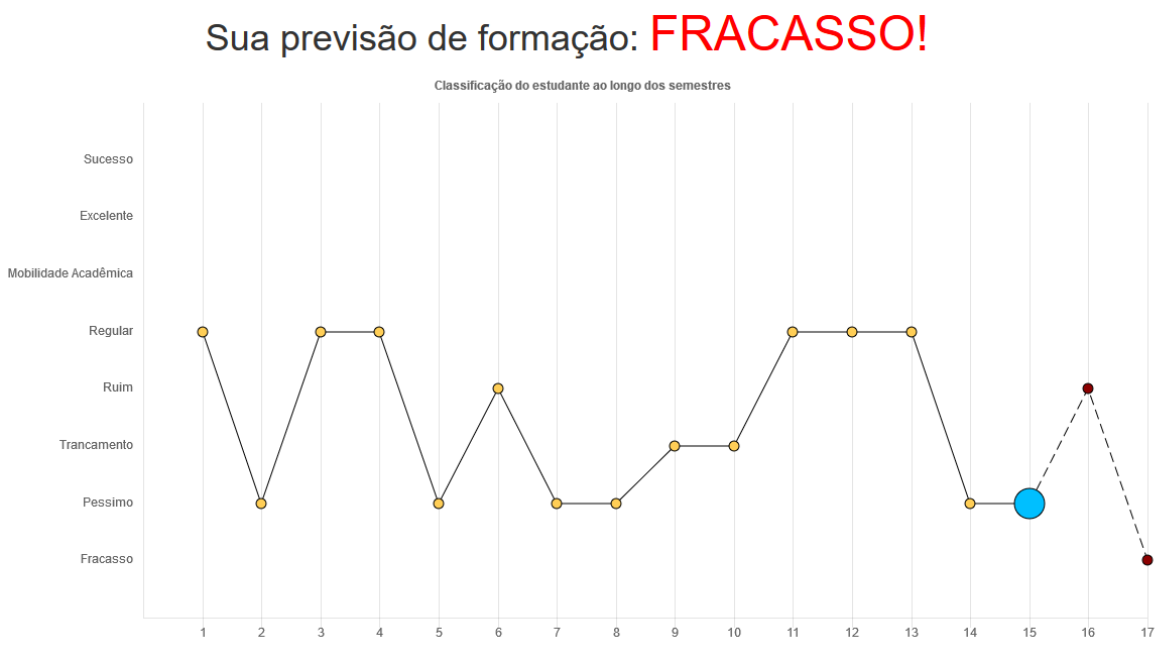
FIGURA 4.14 – Exemplo da página do aluno 368



Fonte: Autor, 2018.

Esses gráficos demonstram o caminho dos estados que o estudante percorreu em cada semestre, indicado pelas bolinhas pequena com o fundo amarelo. Além disso, é destacado o estado atual do aluno, o qual está representado pela bola grande com o fundo azul. A partir do estado atual, há uma linha pontilhada, a qual representa a previsão que o modelo realizou, levando ao fracasso ou ao sucesso.

FIGURA 4.15 – Exemplo da página do aluno 019

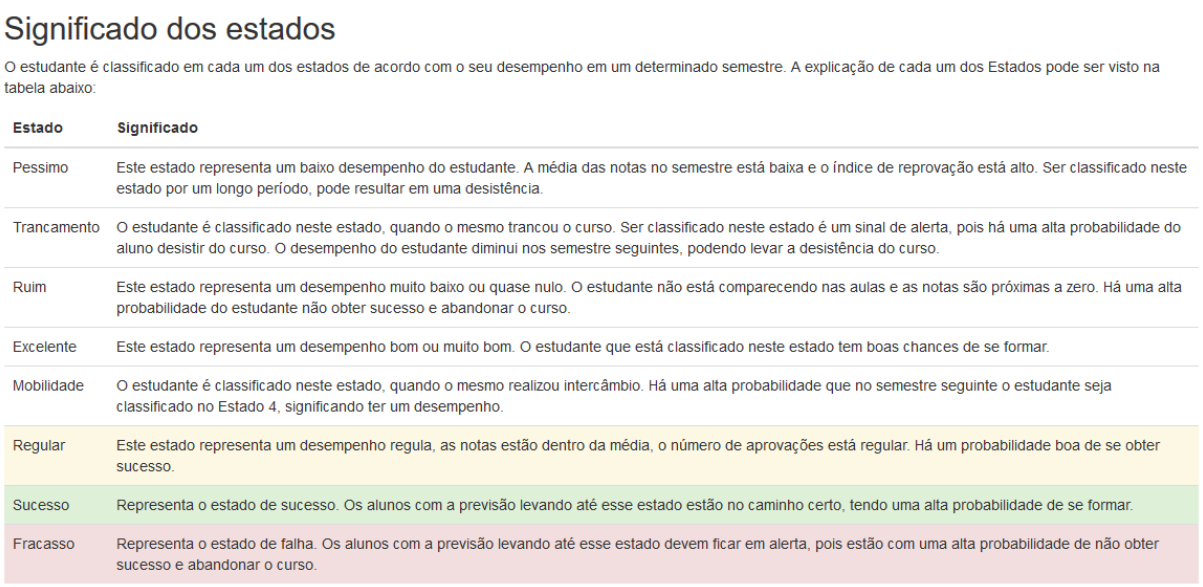


Fonte: Autor, 2018.

No exemplo da Figura 4.15 o modelo realizou a previsão do estudante entrar no estado ruim antes de falhar efetivamente. Isso é interessante notar, pois há tempo de tomar alguma ação corretiva para evitar com que esse estudante abandone o curso.

Na explicação dos estados é notável que um dos estados está pintado de verde e outro de vermelho. Conforme essa representação foi utilizada ao longo do projeto para ilustrar os estados de sucesso e fracasso, agora não é diferente. Também é eminente a cor amarela, diferente da cor que representava o estado de trancado, essa é usada para ressaltar qual o estado atual do estudante. Cada página possui esse destaque em amarelo de acordo com seu estado atual. Para o aluno 014 por exemplo, seu estado atual é o Excelente, portanto, na página desse aluno esse estado estaria com o preenchimento amarelo.

FIGURA 4.16 – Explicação dos estados na página



Significado dos estados

O estudante é classificado em cada um dos estados de acordo com o seu desempenho em um determinado semestre. A explicação de cada um dos Estados pode ser visto na tabela abaixo:

Estado	Significado
Pessimo	Este estado representa um baixo desempenho do estudante. A média das notas no semestre está baixa e o índice de reprovação está alto. Ser classificado neste estado por um longo período, pode resultar em uma desistência.
Trancamento	O estudante é classificado neste estado, quando o mesmo trancou o curso. Ser classificado neste estado é um sinal de alerta, pois há uma alta probabilidade do aluno desistir do curso. O desempenho do estudante diminui nos semestre seguintes, podendo levar a desistência do curso.
Ruim	Este estado representa um desempenho muito baixo ou quase nulo. O estudante não está comparecendo nas aulas e as notas são próximas a zero. Há uma alta probabilidade do estudante não obter sucesso e abandonar o curso.
Excelente	Este estado representa um desempenho bom ou muito bom. O estudante que está classificado neste estado tem boas chances de se formar.
Mobilidade	O estudante é classificado neste estado, quando o mesmo realizou intercâmbio. Há uma alta probabilidade que no semestre seguinte o estudante seja classificado no Estado 4, significando ter um desempenho.
Regular	Este estado representa um desempenho regula, as notas estão dentro da média, o número de aprovações está regular. Há um probabilidade boa de se obter sucesso.
Sucesso	Representa o estado de sucesso. Os alunos com a previsão levando até esse estado estão no caminho certo, tendo uma alta probabilidade de se formar.
Fracasso	Representa o estado de falha. Os alunos com a previsão levando até esse estado devem ficar em alerta, pois estão com uma alta probabilidade de não obter sucesso e abandonar o curso.

Fonte: Autor, 2018.

5 CONCLUSÃO

Uma área emergente dentro das comunidades de pesquisas é da mineração de dados voltada para a educação. Definida como uma área que busca o desenvolvimento de métodos e modelos para explorar os dados específicos de sistemas educacionais, na qual esses métodos são utilizados para compreender melhor os fatores que afetam o desempenho acadêmico de estudantes, permitindo a tomada de ações corretivas e/ou preventivas a partir do que for concluído pela análise.

Um grande desafio desse projeto foi a elaboração de uma boa base de dados, pois essa base tem que estar preparada para a aplicação de ferramentas de mineração de dados. Como cada ferramenta possuí entradas específicas, é imprescindível que a base esteja bem estruturada para atender essas situações. A própria base de dados permite análises interessantes dos alunos, como foi observado que todos os alunos que fizeram a mobilidade acadêmica, terminaram a sequência com sucesso, ou seja, concluíram o curso. A integração com o software R facilita esse tipo de análise, assim como facilita na implementação de ferramentar para mineração de dados.

O estudo das informações de desempenho acadêmico dos estudantes da Engenharia Elétrica da UFPR possibilitou que os aplicativos fossem desenvolvidos para classificar e identificar os alunos e ainda construir uma sequência temporal dos resultados obtidos em cada semestre. Utilizando as notas, frequência, carga horária, aprovações e reprovações foi possível encontrar padrões nos estudantes aplicando a técnica de clusterização hierárquica. Essa técnica permitiu observar a diversidade entre os estudantes e servir como parâmetro para modelagem do sistema. Foi identificado, em média, oito grupos de alunos por período e esse número foi utilizado na definição da quantidade de estados ocultos do modelo.

Com os estados definidos e o modelo devidamente treinado, foi possível obter resultados muito interessantes. Há estados que indicam um desempenho muito baixo dos alunos e que possuem uma alta probabilidade de evasão, enquanto os estados que indicam um desempenho bom possuem uma boa probabilidade para o sucesso. O treinamento do modelo organizou dois estados para algumas situações que os estudantes podem se encontrar, que são os casos de mobilidade acadêmica e trancamento. Isso possibilitou identificar que os estudantes que trancam o curso, tem alta probabilidade de

evadir, enquanto que os estudantes que realizaram intercâmbio tem melhores chances de sucesso.

A validação do modelo apresentou resultados satisfatórios, com uma precisão de acerto de aproximadamente 85%, sendo maior que os resultados apresentados pelos trabalhos relacionados. Além disso, o modelo é capaz de prever antecipadamente que um estudante venha a desistir, com uma média de 2 semestres antes de realmente ocorrer. O modelo apresentou uma taxa de acerto bem maior para os casos negativos do que os positivos e isso não é ruim, identificar um estudante que irá falhar e ele não falhar, não traz grandes preocupações. O problema está em quando o modelo identificar que um aluno irá obter sucesso, quando na realidade não acontece.

Os objetivos específicos propostos no trabalho foram atingidos e apresentaram resultados superiores ao esperado, possibilitando a aplicação nos alunos ativos do curso de Engenharia Elétrica da UFPR.

5.1 TRABALHOS FUTUROS

As próximas atividades são: treinamento do modelo com mais dados; criação de um modelo para o curso diurno; melhorar os gráficos apresentados nas páginas; adicionar mais informações nas páginas e incluir o desempenho em disciplinas específicas no modelo. Cada uma dessas atividades está descrita na sequência.

5.1.1 Treinamento do modelo com mais dados

Essa atividade será repetida, basicamente, em todo semestre. Cada semestre uma grande quantidade de alunos se formam e é possível levantar os estudantes que abandonaram ou cancelaram sua matricula no curso. Dessa maneira, uma grande quantidade de dados pode ser inserida no sistema e então realizar o treinamento de um novo modelo. Pode acontecer do modelo ter um comportamento complemente diferente do apresentado anteriormente, inclusive apresentar uma precisão maior que o apresentado neste trabalho.

5.1.2 Criação de um modelo para o curso diurno

Conforme explicado no Capítulo 4, os dados utilizados para confecção do modelo foi, em sua totalidade, alunos do curso noturno. Mesmo assim, esse modelo

foi aplicado para todos os alunos do curso de Engenharia Elétrica, dos dois turnos. A ideia então, é elaborar um modelo exclusivamente para os alunos do diurno, o qual acreditasse que terá um comportamento diferente.

O que foi citado anteriormente, sobre um novo treinamento por semestre, o mesmo ocorreria para esse modelo. Conforme novas turmas se formam, mais informações são coletadas e que podem ser utilizadas na construção do modelo.

5.1.3 Melhorar os gráficos apresentados nas páginas

Foi apresentado um gráfico com o caminho mais provável que determinado estudante percorreu durante sua vida acadêmica. Por se tratar de um sistema probabilístico é importante mostrar os valores das probabilidade do estudante ser classificado em cada um dos estados, pois como foi mostrado na validação dos resultados, o modelo possui erros. É importante que o estudante tenha essa visão e caso a previsão do estudante seja negativa, mostre qual a porcentagem do estudante ser classificado na falha e qual a porcentagem do estudante ser classificado no sucesso, por exemplo. 80% e 20% respectivamente.

5.1.4 Adicionar mais informações nas páginas

De acordo com os resultados apresentados no Capítulo 4, é muita informação que pode ser extraída desse modelo. É interessante adicionar essas informações para deixar ainda mais completo que uma “simples” previsão. Por exemplo, seria interessante adicionar as probabilidade de transição de cada um dos estados, assim o estudante consegue analisar o estado que tá e vê a probabilidade de alternar de um estado ruim para o bom.

5.1.5 Inclusão de desempenho em disciplinas específicas

Com a inclusão do desempenho de disciplinas específicas podem ser identificados “gargalos” ou indicadores mais fortes para previsão de sucesso ou fracasso. Por exemplo, descobrir se existe algum comportamento comum para os alunos que irão desistir que possa ser identificado já no 1º ou 2º período.

REFERÊNCIAS BIBLIOGRÁFICAS

BALAKRISHNAN, G.; COETZEE, D. Predicting student retention in massive open online courses using hidden markov models. *Electrical Engineering and Computer Sciences University of California at Berkeley*, 2013.

BLUNSOM, P. Hidden markov models. *Lecture notes, August*, Citeseer, v. 15, n. 18-19, p. 48, 2004.

FOSLER-LUSSIER, E. Markov models and hidden markov models: A brief tutorial. *International Computer Science Institute*, 1998.

GRINSTEAD, C. M.; SNELL, J. L. *Introduction to Probability*. AMS, 2006. GNU FDL PDF online. Disponível em: <http://www.dartmouth.edu/~chance/teaching_aids/books_articles/probability_book/book.html>.

GRIVOKOSTOPOULOU, F.; PERIKOS, I.; HATZILYGEROUDIS, I. Utilizing semantic web technologies and data mining techniques to analyze students learning and predict final performance. *Teaching, Assessment and Learning (TALE), 2014 International Conference on*, IEEE, p. 488–494, 2014.

HOWARD, R. A. *Dynamic Programming and Markov Processes*. Cambridge, MA: MIT Press, 1960.

JOHNSON, S. C. Hierarchical clustering schemes. *Psychometrika*, Springer, v. 32, n. 3, p. 241–254, 1967.

LOBO, R. L. A evasão no ensino superior brasileiro. *Instituto Lobo*, 2017.

LUCKIN, R. et al. Modeling learning patterns of students with a tutoring system using hidden markov models. *Artificial intelligence in education: Building technology rich learning contexts that work*, IOS Press, v. 158, p. 238, 2007.

MNMK. *Matriz Confusão*. 2013. Acesso em 24 nov. 2018. Disponível em: <http://developerdeveloper.blogspot.com/2013/11/matriz-confusao.html>.

PEDROSO, C. M. Um método para análise de qualidade de formação e planejamento do ensino de graduação. *XXXX Congresso Brasileiro de Ensino em Engenharia*, COBENGE, 2012.

____. Desenvolvimento de ferramentas inteligentes para avaliação de qualidade do processo de ensino. *Projeto de Pesquisa*, 2017.

RABINER, L. R. A tutorial on hidden markov models and selected applications in speech recognition. *Proceedings of the IEEE*, IEEE, v. 77, n. 2, p. 257–286, 1989.

SACHIN, R. B.; VIJAY, M. S. A survey and future vision of data mining in educational field. *Advanced Computing & Communication Technologies (ACCT), 2012 Second International Conference on*, IEEE, p. 96–100, 2012.

SPARKES, J. J. Quality in higher education. *Engineering Science and Education Journal*, THE INSTITUTION OF ELECTRICAL ENGINEERS, v. 4, p. 253–258, 1995.

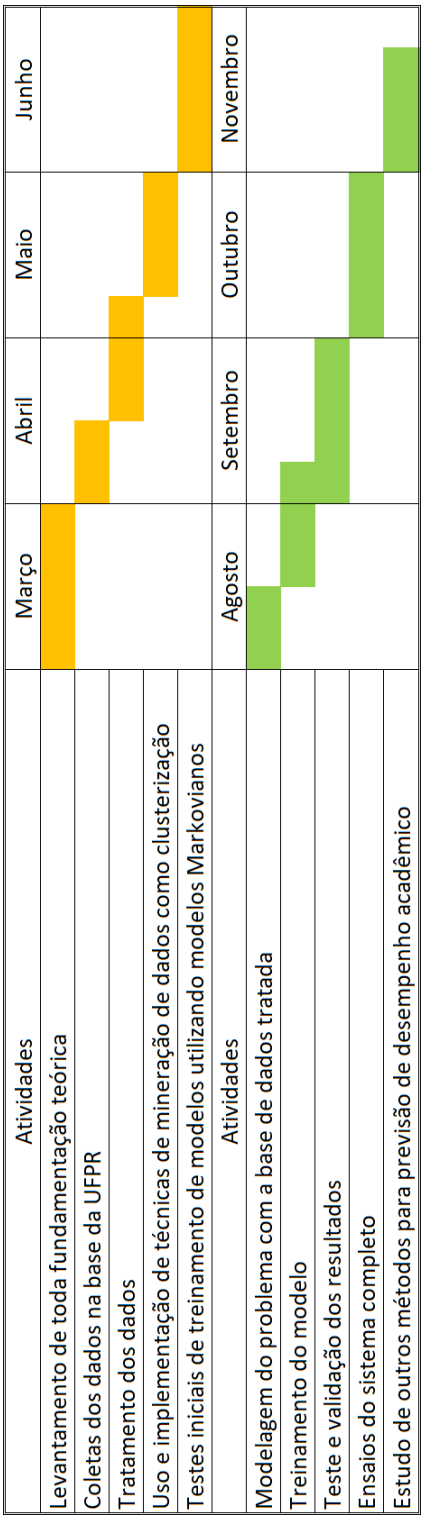
SPECTOR, P. *Cluster Analysis*. 2011. Acesso em 07 dez. 2017. Disponível em: <https://www.stat.berkeley.edu/~s133/Cluster2a.html>.

VERA, C. M.; MORALES, C. R.; SOTO, S. V. Predicting school failure and dropout by using data mining techniques. *IEEE Journal of Latin-American Learning Technologie*, v. 8, n. 1, p. 7–14, 2013.

WARAKAGODA, N. *Three basic problems of HMMs*. 1996. Acesso em 10 jun. 2018. Disponível em: <http://jedlik.phy.bme.hu/~gerjanos/HMM/node6.html>.

APÊNDICE A – CRONOGRAMA DE ATIVIDADES

FIGURA A.1 – Cronograma de atividades



Fonte: Autor.

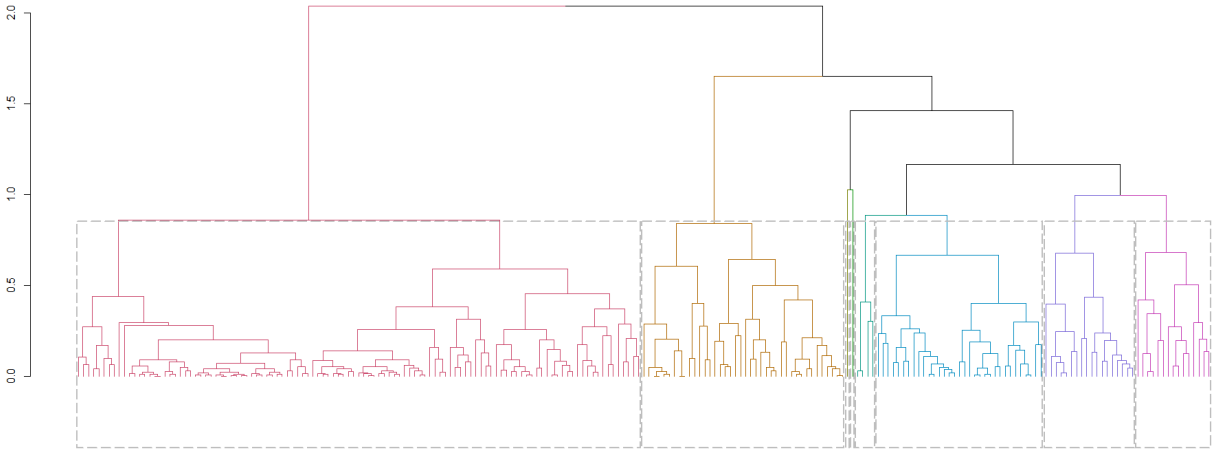
APÊNDICE B – CÓDIGO DA CLUSTERIZAÇÃO NO R

```
library(dendextend)

# Putting the path files into variables
semestre01_path <- "/Dados/semestre01.csv"
semestre02_path <- "/Dados/semestre02.csv"
semestre03_path <- "/Dados/semestre03.csv"
semestre04_path <- "/Dados/semestre04.csv"
semestre05_path <- "/Dados/semestre05.csv"
semestre06_path <- "/Dados/semestre06.csv"
semestre07_path <- "/Dados/semestre07.csv"
semestre08_path <- "/Dados/semestre08.csv"
semestre09_path <- "/Dados/semestre09.csv"
semestre10_path <- "/Dados/semestre10.csv"

# Putting the data files into variables
semestre01 <- read.table(semestre01_path, sep=",", row.names=1, header=TRUE, na.strings=".")
semestre02 <- read.table(semestre02_path, sep=",", row.names=1, header=TRUE, na.strings=".")
semestre03 <- read.table(semestre03_path, sep=",", row.names=1, header=TRUE, na.strings=".")
semestre04 <- read.table(semestre04_path, sep=",", row.names=1, header=TRUE, na.strings=".")
semestre05 <- read.table(semestre05_path, sep=",", row.names=1, header=TRUE, na.strings=".")
semestre06 <- read.table(semestre06_path, sep=",", row.names=1, header=TRUE, na.strings=".")
semestre07 <- read.table(semestre07_path, sep=",", row.names=1, header=TRUE, na.strings=".")
semestre08 <- read.table(semestre08_path, sep=",", row.names=1, header=TRUE, na.strings=".")
semestre09 <- read.table(semestre09_path, sep=",", row.names=1, header=TRUE, na.strings=".")
semestre10 <- read.table(semestre10_path, sep=",", row.names=1, header=TRUE, na.strings=".")

semestre01.dist <- dist(semestre01, method="euclidean")
semestre02.dist <- dist(semestre02, method="euclidean")
semestre03.dist <- dist(semestre03, method="euclidean")
semestre04.dist <- dist(semestre04, method="euclidean")
semestre05.dist <- dist(semestre05, method="euclidean")
semestre06.dist <- dist(semestre06, method="euclidean")
semestre07.dist <- dist(semestre07, method="euclidean")
semestre08.dist <- dist(semestre08, method="euclidean")
semestre09.dist <- dist(semestre09, method="euclidean")
semestre10.dist <- dist(semestre10, method="euclidean")

semestre01.hclust <- hclust(semestre01.dist, method = "complete")
semestre02.hclust <- hclust(semestre02.dist, method = "complete")
semestre03.hclust <- hclust(semestre03.dist, method = "complete")
semestre04.hclust <- hclust(semestre04.dist, method = "complete")
semestre05.hclust <- hclust(semestre05.dist, method = "complete")
semestre06.hclust <- hclust(semestre06.dist, method = "complete")
semestre07.hclust <- hclust(semestre07.dist, method = "complete")
semestre08.hclust <- hclust(semestre08.dist, method = "complete")
semestre09.hclust <- hclust(semestre09.dist, method = "complete")
semestre10.hclust <- hclust(semestre10.dist, method = "complete")

dend01 <- as.dendrogram(semestre01.hclust)
dend02 <- as.dendrogram(semestre02.hclust)
dend03 <- as.dendrogram(semestre03.hclust)
dend04 <- as.dendrogram(semestre04.hclust)
dend05 <- as.dendrogram(semestre05.hclust)
dend06 <- as.dendrogram(semestre06.hclust)
dend07 <- as.dendrogram(semestre07.hclust)
dend08 <- as.dendrogram(semestre08.hclust)
```

```
dend09 <- as.dendrogram(semestre09.hclust)
dend10 <- as.dendrogram(semestre10.hclust)

dend01 %>% set("labels", "") %>% set("branches_k_color", k = 8) %>% plot
dend01 %>% rect.dendrogram(k = 8, border = 8, lty = 5, lwd = 2)

dend02 %>% set("labels", "") %>% set("branches_k_color", k = 8) %>% plot
dend02 %>% rect.dendrogram(k = 8, border = 8, lty = 5, lwd = 2)

dend03 %>% set("labels", "") %>% set("branches_k_color", k = 8) %>% plot
dend03 %>% rect.dendrogram(k = 8, border = 8, lty = 5, lwd = 2)

dend04 %>% set("labels", "") %>% set("branches_k_color", k = 8) %>% plot
dend04 %>% rect.dendrogram(k = 8, border = 8, lty = 5, lwd = 2)

dend05 %>% set("labels", "") %>% set("branches_k_color", k = 8) %>% plot
dend05 %>% rect.dendrogram(k = 8, border = 8, lty = 5, lwd = 2)

dend06 %>% set("labels", "") %>% set("branches_k_color", k = 8) %>% plot
dend06 %>% rect.dendrogram(k = 8, border = 8, lty = 5, lwd = 2)

dend07 %>% set("labels", "") %>% set("branches_k_color", k = 8) %>% plot
dend07 %>% rect.dendrogram(k = 8, border = 8, lty = 5, lwd = 2)

dend08 %>% set("labels", "") %>% set("branches_k_color", k = 8) %>% plot
dend08 %>% rect.dendrogram(k = 8, border = 8, lty = 5, lwd = 2)

dend09 %>% set("labels", "") %>% set("branches_k_color", k = 8) %>% plot
dend09 %>% rect.dendrogram(k = 8, border = 8, lty = 5, lwd = 2)

dend10 %>% set("labels", "") %>% set("branches_k_color", k = 8) %>% plot
dend10 %>% rect.dendrogram(k = 8, border = 8, lty = 5, lwd = 2)
```

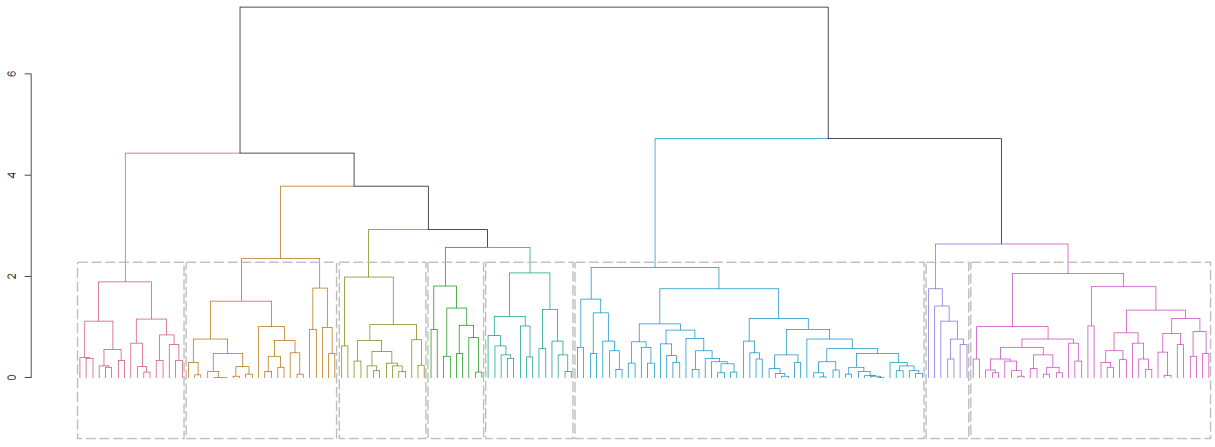
APÊNDICE C – DENDROGRAMA DE CADA SEMESTRE

FIGURA C.1 – Dendrograma do 1º semestre



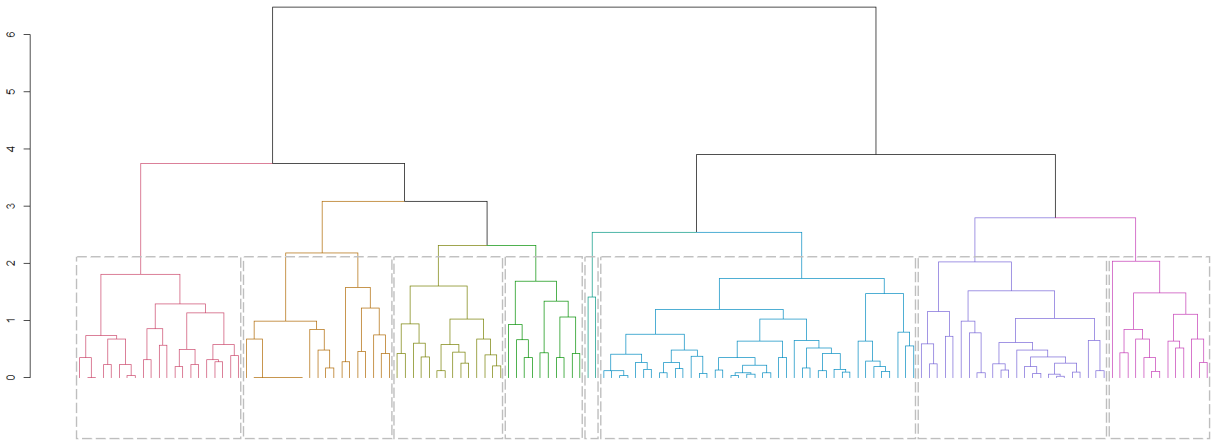
Fonte: Autor.

FIGURA C.2 – Dendrograma do 2º semestre



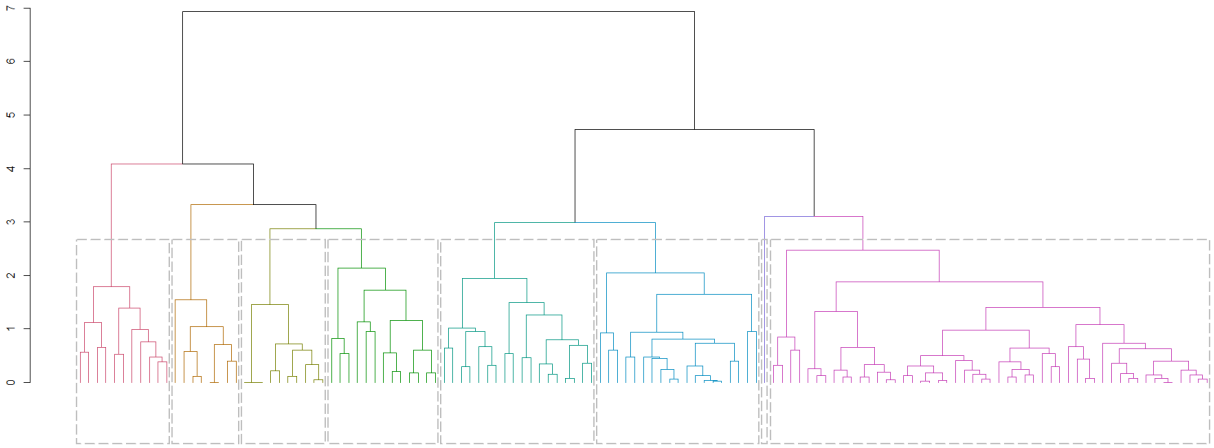
Fonte: Autor.

FIGURA C.3 – Dendrograma do 4º semestre



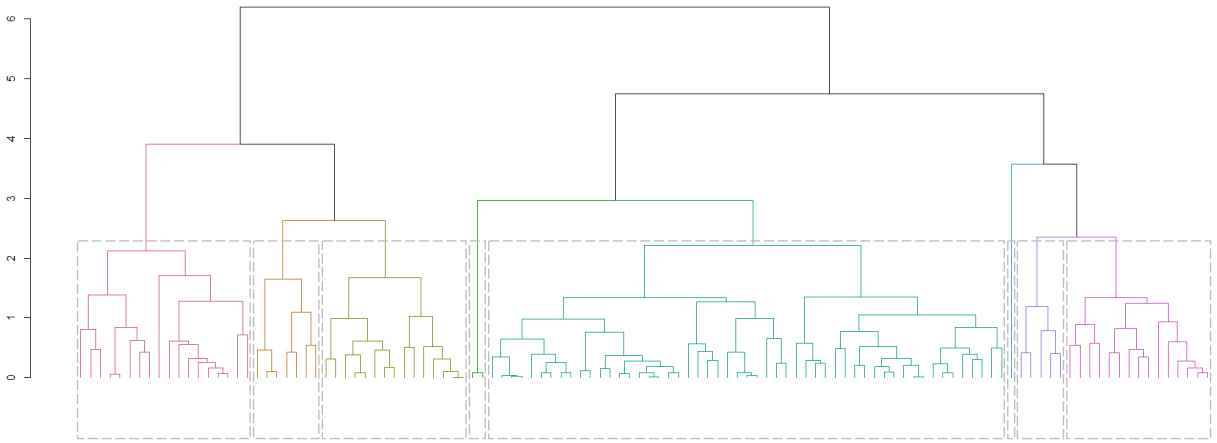
Fonte: Autor.

FIGURA C.4 – Dendrograma do 5º semestre



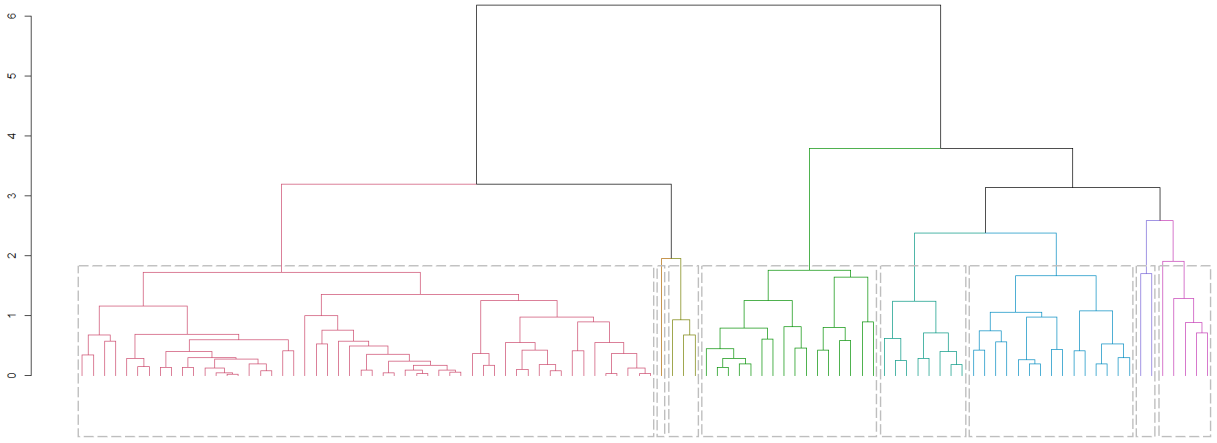
Fonte: Autor.

FIGURA C.5 – Dendrograma do 6º semestre



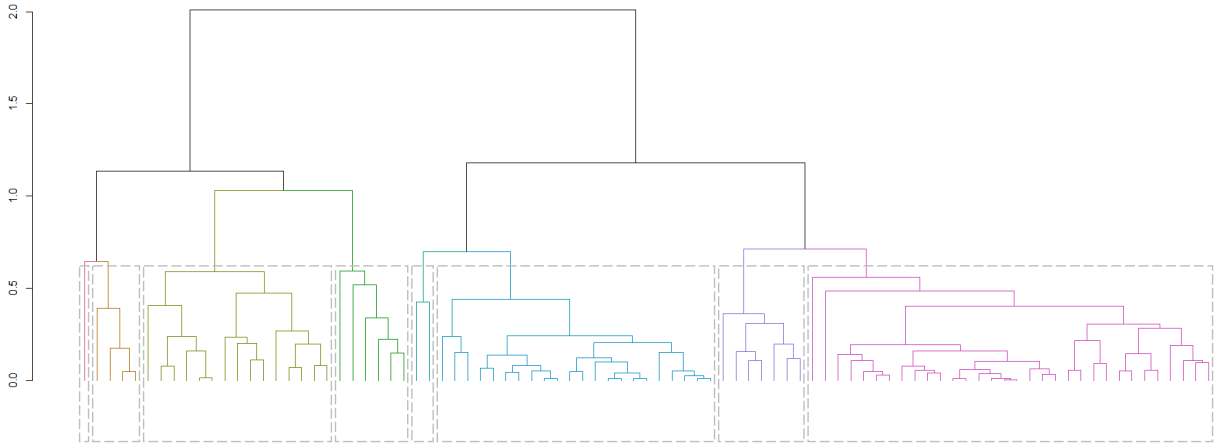
Fonte: Autor.

FIGURA C.6 – Dendrograma do 7º semestre



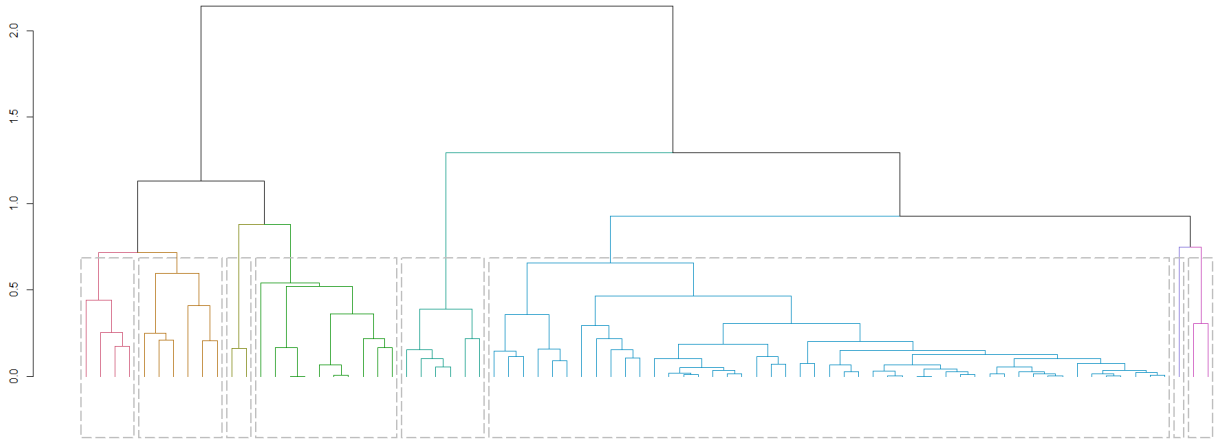
Fonte: Autor.

FIGURA C.7 – Dendrograma do 9º semestre



Fonte: Autor.

FIGURA C.8 – Dendrograma do 10º semestre



Fonte: Autor.

APÊNDICE D – CÓDIGO DO TREINAMENTO DA HMM NO R

```r
library(seqHMM)

# Putting the path files into variables
nota_path <- "/Dados/nota.csv"
freq_path <- "/Dados/freq.csv"
chor_path <- "/Dados/chor.csv"
aprv_path <- "/Dados/aprv.csv"
repv_path <- "/Dados/repv.csv"

# Putting the data files into variables
nota <- read.table(nota_path, sep=",", header=TRUE, row.names=1)
freq <- read.table(freq_path, sep=",", header=TRUE, row.names=1)
chor <- read.table(chor_path, sep=",", header=TRUE, row.names=1)
aprv <- read.table(aprv_path, sep=",", header=TRUE, row.names=1)
repv <- read.table(repv_path, sep=",", header=TRUE, row.names=1)

# Building sequence objects
nota.seq <- seqdef(nota)
freq.seq <- seqdef(freq)
chor.seq <- seqdef(chor)
aprv.seq <- seqdef(aprv)
repv.seq <- seqdef(repv)

# ========= 8 Estados na HMM ================================================
B_nota <- matrix(NA, nrow = 8, ncol = 9)
B_nota[1,] <- seqstatf(nota.seq[, 1:3])[, 2]
B_nota[2,] <- seqstatf(nota.seq[, 4:6])[, 2]
B_nota[3,] <- seqstatf(nota.seq[, 7:9])[, 2]
B_nota[4,] <- seqstatf(nota.seq[, 10:12])[, 2]
B_nota[5,] <- seqstatf(nota.seq[, 13:15])[, 2]
B_nota[6,] <- seqstatf(nota.seq[, 16:17])[, 2]
B_nota[7,] <- seqstatf(nota.seq[, 18:19])[, 2]
B_nota[8,] <- seqstatf(nota.seq[, 20:21])[, 2]
B_nota <- B_nota / rowSums(B_nota)
B_freq <- matrix(NA, nrow = 8, ncol = 9)
B_freq[1,] <- seqstatf(freq.seq[, 1:3])[, 2]
B_freq[2,] <- seqstatf(freq.seq[, 4:6])[, 2]
B_freq[3,] <- seqstatf(freq.seq[, 7:9])[, 2]
B_freq[4,] <- seqstatf(freq.seq[, 10:12])[, 2]
B_freq[5,] <- seqstatf(freq.seq[, 13:15])[, 2]
B_freq[6,] <- seqstatf(freq.seq[, 16:17])[, 2]
B_freq[7,] <- seqstatf(freq.seq[, 18:19])[, 2]
B_freq[8,] <- seqstatf(freq.seq[, 20:21])[, 2]
B_freq <- B_freq / rowSums(B_freq)
B_chor <- matrix(NA, nrow = 8, ncol = 9)
B_chor[1,] <- seqstatf(chor.seq[, 1:3])[, 2]
B_chor[2,] <- seqstatf(chor.seq[, 4:6])[, 2]
B_chor[3,] <- seqstatf(chor.seq[, 7:9])[, 2]
B_chor[4,] <- seqstatf(chor.seq[, 10:12])[, 2]
B_chor[5,] <- seqstatf(chor.seq[, 13:15])[, 2]
B_chor[6,] <- seqstatf(chor.seq[, 16:17])[, 2]
B_chor[7,] <- seqstatf(chor.seq[, 18:19])[, 2]
B_chor[8,] <- seqstatf(chor.seq[, 20:21])[, 2]
B_chor <- B_chor / rowSums(B_chor)
B_aprv <- matrix(NA, nrow = 8, ncol = 9)
B_aprv[1,] <- seqstatf(aprv.seq[, 1:3])[, 2]
```

```r
B_aprv[2,] <- seqstatf(aprv.seq[, 4:6])[, 2]
B_aprv[3,] <- seqstatf(aprv.seq[, 7:9])[, 2]
B_aprv[4,] <- seqstatf(aprv.seq[, 10:12])[, 2]
B_aprv[5,] <- seqstatf(aprv.seq[, 13:15])[, 2]
B_aprv[6,] <- seqstatf(aprv.seq[, 16:17])[, 2]
B_aprv[7,] <- seqstatf(aprv.seq[, 18:19])[, 2]
B_aprv[8,] <- seqstatf(aprv.seq[, 20:21])[, 2]
B_aprv <- B_aprv / rowSums(B_aprv)
B_repv <- matrix(NA, nrow = 8, ncol = 9)
B_repv[1,] <- seqstatf(repv.seq[, 1:3])[, 2]
B_repv[2,] <- seqstatf(repv.seq[, 4:6])[, 2]
B_repv[3,] <- seqstatf(repv.seq[, 7:9])[, 2]
B_repv[4,] <- seqstatf(repv.seq[, 10:12])[, 2]
B_repv[5,] <- seqstatf(repv.seq[, 13:15])[, 2]
B_repv[6,] <- seqstatf(repv.seq[, 16:17])[, 2]
B_repv[7,] <- seqstatf(repv.seq[, 18:19])[, 2]
B_repv[8,] <- seqstatf(repv.seq[, 20:21])[, 2]
B_repv <- B_repv / rowSums(B_repv)

# Initial values for transition matrix
A <- matrix(c(1/8, 1/8, 1/8, 1/8, 1/8, 1/8, 1/8, 1/8,
  1/8, 1/8, 1/8, 1/8, 1/8, 1/8, 1/8, 1/8,
  1/8, 1/8, 1/8, 1/8, 1/8, 1/8, 1/8, 1/8,
  1/8, 1/8, 1/8, 1/8, 1/8, 1/8, 1/8, 1/8,
  1/8, 1/8, 1/8, 1/8, 1/8, 1/8, 1/8, 1/8,
  1/8, 1/8, 1/8, 1/8, 1/8, 1/8, 1/8, 1/8,
  1/8, 1/8, 1/8, 1/8, 1/8, 1/8, 1/8, 1/8,
  1/8, 1/8, 1/8, 1/8, 1/8, 1/8, 1/8, 1/8), nrow = 8, ncol = 8, byrow = TRUE)

# Initial values for initial state probabilities
initial_probs <- c(1/8, 1/8, 1/8, 1/8, 1/8, 1/8, 1/8, 1/8)

# Building the hidden Markov model with initial parameter values
bhmm <- build_hmm(
  observations = list(nota.seq, freq.seq, chor.seq, aprv.seq, repv.seq),
  initial_probs = initial_probs, transition_probs = A,
  emission_probs = list(B_nota, B_freq, B_chor, B_aprv, B_repv),
  choranne1_names = c("Notas", "Frequência", "Carga Horária", "Aprovações", "Reprovações")
)

# Fitting the hmm (using only the default MLSL algorithm)
hmm <- fit_model(bhmm, control_em=list(restart=list(times=10, transition=FALSE)))

hmm_path <- "/hmm.rds"

saveRDS(hmm, hmm_path)
```

APÊNDICE E – CÓDIGO DO CALCULO DOS PARÂMETROS DA CADEIA DE MARKOV

```
library(seqHMM)
library(markovchain)
library(matlab)

# Putting the path files into variables
hmm_path <-  "/hmm.rds"

/# Reading the hmm
hmm <- readRDS(hmm_path)

mca <- new("markovchain", transitionMatrix=hmm$model$transition_probs)

extractMatrices <- function(mcObj) {
        require(matlab)
        mcObj <- canonicForm(object = mcObj)
        #get the indices of transient and absorbing
        transIdx <- which(states(mcObj) %in% transientStates(mcObj))
        absIdx <- which(states(mcObj) %in% absorbingStates(mcObj))
        #get the Q, R and I matrices
        Q <- as.matrix(mcObj@transitionMatrix[transIdx,transIdx])
        R <- as.matrix(mcObj@transitionMatrix[transIdx,absIdx])
        I <- as.matrix(mcObj@transitionMatrix[absIdx, absIdx])
        #get the fundamental matrix
        N <- solve(eye(size(Q)) - Q)
        #computing final absorbion probabilities
        NR <- N %*% R
        #return
        out <- list(
        canonicalForm = mcObj,
        Q = Q,
        R = R,
        I = I,
        N=N,
        NR=NR
        )
        return(out)
}

emca <- extractMatrices(mcObj = mca)
```

APÊNDICE F – CÓDIGO DA VALIDAÇÃO DA HMM NO R

```r
library(seqHMM)

a.nota <- read.table("nota/a.nota.txt", sep="\t", header=TRUE, row.names=1)
b.nota <- read.table("nota/b.nota.txt", sep="\t", header=TRUE, row.names=1)
c.nota <- read.table("nota/c.nota.txt", sep="\t", header=TRUE, row.names=1)
d.nota <- read.table("nota/d.nota.txt", sep="\t", header=TRUE, row.names=1)

a.freq <- read.table("freq/a.freq.txt", sep="\t", header=TRUE, row.names=1)
b.freq <- read.table("freq/b.freq.txt", sep="\t", header=TRUE, row.names=1)
c.freq <- read.table("freq/c.freq.txt", sep="\t", header=TRUE, row.names=1)
d.freq <- read.table("freq/d.freq.txt", sep="\t", header=TRUE, row.names=1)

a.chor <- read.table("chor/a.chor.txt", sep="\t", header=TRUE, row.names=1)
b.chor <- read.table("chor/b.chor.txt", sep="\t", header=TRUE, row.names=1)
c.chor <- read.table("chor/c.chor.txt", sep="\t", header=TRUE, row.names=1)
d.chor <- read.table("chor/d.chor.txt", sep="\t", header=TRUE, row.names=1)

a.aprv <- read.table("aprv/a.aprv.txt", sep="\t", header=TRUE, row.names=1)
b.aprv <- read.table("aprv/b.aprv.txt", sep="\t", header=TRUE, row.names=1)
c.aprv <- read.table("aprv/c.aprv.txt", sep="\t", header=TRUE, row.names=1)
d.aprv <- read.table("aprv/d.aprv.txt", sep="\t", header=TRUE, row.names=1)

a.rprv <- read.table("rprv/a.rprv.txt", sep="\t", header=TRUE, row.names=1)
b.rprv <- read.table("rprv/b.rprv.txt", sep="\t", header=TRUE, row.names=1)
c.rprv <- read.table("rprv/c.rprv.txt", sep="\t", header=TRUE, row.names=1)
d.rprv <- read.table("rprv/d.rprv.txt", sep="\t", header=TRUE, row.names=1)

# Junta 'a', 'b', 'c' e 'd' para treinamento completo
nota <- rbind(a.nota, b.nota, c.nota, d.nota)
freq <- rbind(a.freq, b.freq, c.freq, d.freq)
chor <- rbind(a.chor, b.chor, c.chor, d.chor)
aprv <- rbind(a.aprv, b.aprv, c.aprv, d.aprv)
rprv <- rbind(a.rprv, b.rprv, c.rprv, d.rprv)

nota.seq <- seqdef(nota)
freq.seq <- seqdef(freq)
chor.seq <- seqdef(chor)
aprv.seq <- seqdef(aprv)
rprv.seq <- seqdef(rprv)

# Junta 'a', 'b' e 'c' e valida 'd'
t1.nota <- rbind(a.nota, b.nota, c.nota)
t1.freq <- rbind(a.freq, b.freq, c.freq)
t1.chor <- rbind(a.chor, b.chor, c.chor)
t1.aprv <- rbind(a.aprv, b.aprv, c.aprv)
t1.rprv <- rbind(a.rprv, b.rprv, c.rprv)

t1.nota.seq <- seqdef(t1.nota)
t1.freq.seq <- seqdef(t1.freq)
t1.chor.seq <- seqdef(t1.chor)
t1.aprv.seq <- seqdef(t1.aprv)
t1.rprv.seq <- seqdef(t1.rprv)

v1.nota.seq <- seqdef(d.nota)
v1.freq.seq <- seqdef(d.freq)
v1.chor.seq <- seqdef(d.chor)
```

```
v1.aprv.seq <- seqdef(d.aprv)
v1.rprv.seq <- seqdef(d.rprv)

# Junta 'a', 'b' e 'd' e valida 'c'
t2.nota <- rbind(a.nota, b.nota, d.nota)
t2.freq <- rbind(a.freq, b.freq, d.freq)
t2.chor <- rbind(a.chor, b.chor, d.chor)
t2.aprv <- rbind(a.aprv, b.aprv, d.aprv)
t2.rprv <- rbind(a.rprv, b.rprv, d.rprv)

t2.nota.seq <- seqdef(t2.nota)
t2.freq.seq <- seqdef(t2.freq)
t2.chor.seq <- seqdef(t2.chor)
t2.aprv.seq <- seqdef(t2.aprv)
t2.rprv.seq <- seqdef(t2.rprv)

v2.nota.seq <- seqdef(c.nota)
v2.freq.seq <- seqdef(c.freq)
v2.chor.seq <- seqdef(c.chor)
v2.aprv.seq <- seqdef(c.aprv)
v2.rprv.seq <- seqdef(c.rprv)

# Junta 'a', 'c' e 'd' e valida 'b'
t3.nota <- rbind(a.nota, c.nota, d.nota)
t3.freq <- rbind(a.freq, c.freq, d.freq)
t3.chor <- rbind(a.chor, c.chor, d.chor)
t3.aprv <- rbind(a.aprv, c.aprv, d.aprv)
t3.rprv <- rbind(a.rprv, c.rprv, d.rprv)

t3.nota.seq <- seqdef(t3.nota)
t3.freq.seq <- seqdef(t3.freq)
t3.chor.seq <- seqdef(t3.chor)
t3.aprv.seq <- seqdef(t3.aprv)
t3.rprv.seq <- seqdef(t3.rprv)

v3.nota.seq <- seqdef(b.nota)
v3.freq.seq <- seqdef(b.freq)
v3.chor.seq <- seqdef(b.chor)
v3.aprv.seq <- seqdef(b.aprv)
v3.rprv.seq <- seqdef(b.rprv)

# Junta 'b', 'c' e 'd' e valida 'a'
t4.nota <- rbind(b.nota, c.nota, d.nota)
t4.freq <- rbind(b.freq, c.freq, d.freq)
t4.chor <- rbind(b.chor, c.chor, d.chor)
t4.aprv <- rbind(b.aprv, c.aprv, d.aprv)
t4.rprv <- rbind(b.rprv, c.rprv, d.rprv)

t4.nota.seq <- seqdef(t4.nota)
t4.freq.seq <- seqdef(t4.freq)
t4.chor.seq <- seqdef(t4.chor)
t4.aprv.seq <- seqdef(t4.aprv)
t4.rprv.seq <- seqdef(t4.rprv)

v4.nota.seq <- seqdef(a.nota)
v4.freq.seq <- seqdef(a.freq)
v4.chor.seq <- seqdef(a.chor)
v4.aprv.seq <- seqdef(a.aprv)
v4.rprv.seq <- seqdef(a.rprv)

B_nota8 <- matrix(NA, nrow = 8, ncol = 9)
B_nota8[1,] <- seqstatf(nota.seq[, 1:3])[, 2]
```

```
B_nota8[2,] <- seqstatf(nota.seq[, 4:6])[, 2]
B_nota8[3,] <- seqstatf(nota.seq[, 7:9])[, 2]
B_nota8[4,] <- seqstatf(nota.seq[, 10:12])[, 2]
B_nota8[5,] <- seqstatf(nota.seq[, 13:15])[, 2]
B_nota8[6,] <- seqstatf(nota.seq[, 16:17])[, 2]
B_nota8[7,] <- seqstatf(nota.seq[, 18:19])[, 2]
B_nota8[8,] <- seqstatf(nota.seq[, 20:21])[, 2]
B_nota8 <- B_nota8 / rowSums(B_nota8)

B_freq8 <- matrix(NA, nrow = 8, ncol = 9)
B_freq8[1,] <- seqstatf(freq.seq[, 1:3])[, 2]
B_freq8[2,] <- seqstatf(freq.seq[, 4:6])[, 2]
B_freq8[3,] <- seqstatf(freq.seq[, 7:9])[, 2]
B_freq8[4,] <- seqstatf(freq.seq[, 10:12])[, 2]
B_freq8[5,] <- seqstatf(freq.seq[, 13:15])[, 2]
B_freq8[6,] <- seqstatf(freq.seq[, 16:17])[, 2]
B_freq8[7,] <- seqstatf(freq.seq[, 18:19])[, 2]
B_freq8[8,] <- seqstatf(freq.seq[, 20:21])[, 2]
B_freq8 <- B_freq8 / rowSums(B_freq8)

B_chor8 <- matrix(NA, nrow = 8, ncol = 9)
B_chor8[1,] <- seqstatf(chor.seq[, 1:3])[, 2]
B_chor8[2,] <- seqstatf(chor.seq[, 4:6])[, 2]
B_chor8[3,] <- seqstatf(chor.seq[, 7:9])[, 2]
B_chor8[4,] <- seqstatf(chor.seq[, 10:12])[, 2]
B_chor8[5,] <- seqstatf(chor.seq[, 13:15])[, 2]
B_chor8[6,] <- seqstatf(chor.seq[, 16:17])[, 2]
B_chor8[7,] <- seqstatf(chor.seq[, 18:19])[, 2]
B_chor8[8,] <- seqstatf(chor.seq[, 20:21])[, 2]
B_chor8 <- B_chor8 / rowSums(B_chor8)

B_aprv8 <- matrix(NA, nrow = 8, ncol = 9)
B_aprv8[1,] <- seqstatf(aprv.seq[, 1:3])[, 2]
B_aprv8[2,] <- seqstatf(aprv.seq[, 4:6])[, 2]
B_aprv8[3,] <- seqstatf(aprv.seq[, 7:9])[, 2]
B_aprv8[4,] <- seqstatf(aprv.seq[, 10:12])[, 2]
B_aprv8[5,] <- seqstatf(aprv.seq[, 13:15])[, 2]
B_aprv8[6,] <- seqstatf(aprv.seq[, 16:17])[, 2]
B_aprv8[7,] <- seqstatf(aprv.seq[, 18:19])[, 2]
B_aprv8[8,] <- seqstatf(aprv.seq[, 20:21])[, 2]
B_aprv8 <- B_aprv8 / rowSums(B_aprv8)

B_rprv8 <- matrix(NA, nrow = 8, ncol = 9)
B_rprv8[1,] <- seqstatf(rprv.seq[, 1:3])[, 2]
B_rprv8[2,] <- seqstatf(rprv.seq[, 4:6])[, 2]
B_rprv8[3,] <- seqstatf(rprv.seq[, 7:9])[, 2]
B_rprv8[4,] <- seqstatf(rprv.seq[, 10:12])[, 2]
B_rprv8[5,] <- seqstatf(rprv.seq[, 13:15])[, 2]
B_rprv8[6,] <- seqstatf(rprv.seq[, 16:17])[, 2]
B_rprv8[7,] <- seqstatf(rprv.seq[, 18:19])[, 2]
B_rprv8[8,] <- seqstatf(rprv.seq[, 20:21])[, 2]
B_rprv8 <- B_rprv8 / rowSums(B_rprv8)

A8 <- matrix(c(1/8, 1/8, 1/8, 1/8, 1/8, 1/8, 1/8, 1/8,
                          1/8, 1/8, 1/8, 1/8, 1/8, 1/8, 1/8, 1/8,
                          1/8, 1/8, 1/8, 1/8, 1/8, 1/8, 1/8, 1/8,
                          1/8, 1/8, 1/8, 1/8, 1/8, 1/8, 1/8, 1/8,
                          1/8, 1/8, 1/8, 1/8, 1/8, 1/8, 1/8, 1/8,
                          1/8, 1/8, 1/8, 1/8, 1/8, 1/8, 1/8, 1/8,
                          1/8, 1/8, 1/8, 1/8, 1/8, 1/8, 1/8, 1/8), nrow = 8, ncol = 8, byrow = TRUE)

```

```
initial_probs8 <- c(1/8, 1/8, 1/8, 1/8, 1/8, 1/8, 1/8, 1/8)

bhmm <- build_hmm(
  observations = list(nota.seq, freq.seq, chor.seq, aprv.seq, rprv.seq),
  initial_probs = initial_probs8, transition_probs = A8,
  emission_probs = list(B_nota8, B_freq8, B_chor8, B_aprv8, B_rprv8),
  choranel_names = c("Notas", "Frequencia", "Carga Horaria", "Aprovacoes", "Reprovacoes")
)

t1.bhmm <- build_hmm(
  observations = list(t1.nota.seq, t1.freq.seq, t1.chor.seq, t1.aprv.seq, t1.rprv.seq),
  initial_probs = initial_probs8, transition_probs = A8,
  emission_probs = list(B_nota8, B_freq8, B_chor8, B_aprv8, B_rprv8),
  choranel_names = c("Notas", "Frequencia", "Carga Horaria", "Aprovacoes", "Reprovacoes")
)

t2.bhmm <- build_hmm(
  observations = list(t2.nota.seq, t2.freq.seq, t2.chor.seq, t2.aprv.seq, t2.rprv.seq),
  initial_probs = initial_probs8, transition_probs = A8,
  emission_probs = list(B_nota8, B_freq8, B_chor8, B_aprv8, B_rprv8),
  choranel_names = c("Notas", "Frequencia", "Carga Horaria", "Aprovacoes", "Reprovacoes")
)

t3.bhmm <- build_hmm(
  observations = list(t3.nota.seq, t3.freq.seq, t3.chor.seq, t3.aprv.seq, t3.rprv.seq),
  initial_probs = initial_probs8, transition_probs = A8,
  emission_probs = list(B_nota8, B_freq8, B_chor8, B_aprv8, B_rprv8),
  choranel_names = c("Notas", "Frequencia", "Carga Horaria", "Aprovacoes", "Reprovacoes")
)

t4.bhmm <- build_hmm(
  observations = list(t4.nota.seq, t4.freq.seq, t4.chor.seq, t4.aprv.seq, t4.rprv.seq),
  initial_probs = initial_probs8, transition_probs = A8,
  emission_probs = list(B_nota8, B_freq8, B_chor8, B_aprv8, B_rprv8),
  choranel_names = c("Notas", "Frequencia", "Carga Horaria", "Aprovacoes", "Reprovacoes")
)

hmm <- fit_model(t1.bhmm, control_em=list(restart=list(times=20, transition=FALSE)))
t1.hmm <- fit_model(t1.bhmm, control_em=list(restart=list(times=20, transition=FALSE)))
t2.hmm <- fit_model(t2.bhmm, control_em=list(restart=list(times=20, transition=FALSE)))
t3.hmm <- fit_model(t3.bhmm, control_em=list(restart=list(times=20, transition=FALSE)))
t4.hmm <- fit_model(t4.bhmm, control_em=list(restart=list(times=20, transition=FALSE)))

# Validating the hmm
v1.bhmm <- build_hmm(
  observations = list(v1.nota.seq, v1.freq.seq, v1.chor.seq, v1.aprv.seq, v1.rprv.seq),
  initial_probs = t1.hmm$model$initial_probs, transition_probs = t1.hmm$model$transition_probs,
  emission_probs = list(t1.hmm$model$emission_probs$'Channel 1', t1.hmm$model$emission_probs$'Channel 2',
        t1.hmm$model$emission_probs$'Channel 3', t1.hmm$model$emission_probs$'Channel 4',
        t1.hmm$model$emission_probs$'Channel 5'),
  choranel_names = c("Notas", "Frequencia", "Carga Horaria", "Aprovacoes", "Reprovacoes")
)

v2.bhmm <- build_hmm(
  observations = list(v2.nota.seq, v2.freq.seq, v2.chor.seq, v2.aprv.seq, v2.rprv.seq),
  initial_probs = t2.hmm$model$initial_probs, transition_probs = t2.hmm$model$transition_probs,
  emission_probs = list(t2.hmm$model$emission_probs$'Channel 1', t2.hmm$model$emission_probs$'Channel 2',
        t2.hmm$model$emission_probs$'Channel 3', t2.hmm$model$emission_probs$'Channel 4',
        t2.hmm$model$emission_probs$'Channel 5'),
  choranel_names = c("Notas", "Frequencia", "Carga Horaria", "Aprovacoes", "Reprovacoes")
)

```

```
v3.bhmm <- build_hmm(
  observations = list(v3.nota.seq, v3.freq.seq, v3.chor.seq, v3.aprv.seq, v3.rprv.seq),
  initial_probs = t3.hmm$model$initial_probs, transition_probs = t3.hmm$model$transition_probs,
  emission_probs = list(t3.hmm$model$emission_probs$'Channel 1', t3.hmm$model$emission_probs$'Channel 2',
        t3.hmm$model$emission_probs$'Channel 3', t3.hmm$model$emission_probs$'Channel 4',
        t3.hmm$model$emission_probs$'Channel 5'),
  chorannel_names = c("Notas", "Frequencia", "Carga Horaria", "Aprovacoes", "Reprovacoes")
)

v4.bhmm <- build_hmm(
  observations = list(v4.nota.seq, v4.freq.seq, v4.chor.seq, v4.aprv.seq, v4.rprv.seq),
  initial_probs = t4.hmm$model$initial_probs, transition_probs = t4.hmm$model$transition_probs,
  emission_probs = list(t4.hmm$model$emission_probs$'Channel 1', t4.hmm$model$emission_probs$'Channel 2',
        t4.hmm$model$emission_probs$'Channel 3', t4.hmm$model$emission_probs$'Channel 4',
        t4.hmm$model$emission_probs$'Channel 5'),
  chorannel_names = c("Notas", "Frequencia", "Carga Horaria", "Aprovacoes", "Reprovacoes")
)

v1.h_paths = hidden_paths(v1.bhmm)
v2.h_paths = hidden_paths(v2.bhmm)
v3.h_paths = hidden_paths(v3.bhmm)
v4.h_paths = hidden_paths(v4.bhmm)
```

APÊNDICE G – CÓDIGO DO CAMINHO MAIS PROVÁVEL DOS ALUNOS ATIVOS NO R

```
library(seqHMM)

# Putting the path files into variables
nota_path <- "/Dados/repv.csv"
hmm_path <-  "/hmm.rds"
h_paths_path <- "/h_paths.csv"

hmm <- readRDS(hmm_path)

# Initial probs of the model
initial_probs <- hmm$model$initial_probs

# Transition probs of the model
transition_probs <- hmm$model$transition_probs

# Emission probs of the model

ep.nota <- hmm$model$emission_probs$'Channel 1'
ep.freq <- hmm$model$emission_probs$'Channel 2'
ep.chor <- hmm$model$emission_probs$'Channel 3'
ep.aprv <- hmm$model$emission_probs$'Channel 4'
ep.repv <- hmm$model$emission_probs$'Channel 5'

# Putting the data files into variables
nota <- read.table(nota_path, sep=",", row.names=1, na.strings=".",
        stringsAsFactors=FALSE, colClasses = c("character"))
freq <- read.table(freq_path, sep=",", row.names=1, na.strings=".",
        stringsAsFactors=FALSE, colClasses = c("character"))
chor <- read.table(chor_path, sep=",", row.names=1, na.strings=".",
        stringsAsFactors=FALSE, colClasses = c("character"))
aprv <- read.table(aprv_path, sep=",", row.names=1, na.strings=".",
        stringsAsFactors=FALSE, colClasses = c("character"))
repv <- read.table(repv_path, sep=",", row.names=1, na.strings=".",
        stringsAsFactors=FALSE, colClasses = c("character"))

# Building sequence objects
nota.seq <- seqdef(nota, gaps="DEL")
freq.seq <- seqdef(freq, gaps="DEL")
chor.seq <- seqdef(chor, gaps="DEL")
aprv.seq <- seqdef(aprv, gaps="DEL")
repv.seq <- seqdef(repv, gaps="DEL")

bhmm <- build_hmm(observations = list(nota.seq, freq.seq, chor.seq, aprv.seq, repv.seq),
                                      transition_probs = transition_probs, initial_probs = initial_probs,
                                      emission_probs = list(ep.nota, ep.freq, ep.chor, ep.aprv, ep.repv)
)

h_paths = hidden_paths(bhmm)

write.csv(h_paths, file = h_paths_path)
```