



UNIVERSIDADE FEDERAL RURAL DE PERNAMBUCO
PRÓ-REITORIA DE PÓS-GRADUAÇÃO
PROGRAMA DE PÓS-GRADUAÇÃO EM INFORMÁTICA APLICADA

JULIANA FERREIRA DOS SANTOS

ANÁLISE DO IMPACTO DA EVASÃO E
RETENÇÃO NO ENSINO SUPERIOR UTILIZANDO
CADEIAS DE MARKOV ABSORVENTES

RECIFE – PE

2022

JULIANA FERREIRA DOS SANTOS

**ANÁLISE DO IMPACTO DA EVASÃO E
RETENÇÃO NO ENSINO SUPERIOR UTILIZANDO
CADEIAS DE MARKOV ABSORVENTES**

Dissertação submetida ao Programa de Pós-Graduação em Informática Aplicada da Universidade Federal Rural de Pernambuco, como parte dos requisitos para obtenção do grau de Mestre em Informática Aplicada.

ORIENTADOR: Gabriel Alves de Albuquerque Júnior

RECIFE – PE

2022

Dados Internacionais de Catalogação na Publicação
Universidade Federal Rural de Pernambuco
Sistema Integrado de Bibliotecas
Gerada automaticamente, mediante os dados fornecidos pelo(a) autor(a)

- S237a Santos, Juliana Ferreira dos
Análise do Impacto da Evasão e Retenção no Ensino Superior Utilizando Cadeias de Markov Absorventes / Juliana Ferreira dos Santos. - 2022.
76 f. : il.
- Orientador: Gabriel Alves de Albuquerque Junior.
Inclui referências e apêndice(s).
- Dissertação (Mestrado) - Universidade Federal Rural de Pernambuco, Programa de Pós-Graduação em Informática Aplicada, Recife, 2022.
1. Análise de sobrevivência. 2. Cadeia de markov. 3. Ensino superior. 4. Evasão. 5. Retenção. I. Junior, Gabriel Alves de Albuquerque, orient. II. Título

JULIANA FERREIRA DOS SANTOS

**ANÁLISE DO IMPACTO DA EVASÃO E
RETENÇÃO NO ENSINO SUPERIOR UTILIZANDO
CADEIAS DE MARKOV ABSORVENTES**

Dissertação submetida ao Programa de Pós-Graduação em Informática Aplicada da Universidade Federal Rural de Pernambuco, como parte dos requisitos para obtenção do grau de Mestre em Informática Aplicada.

Aprovada em: 28 de 07 de 2022.

BANCA EXAMINADORA

Gabriel Alves de Albuquerque Júnior (Orientador)
Universidade Federal Rural de Pernambuco - UFRPE
Departamento de Estatística e Informática – DEINFO

Ermeson Carneiro de Andrade
Universidade Federal Rural de Pernambuco - UFRPE
Departamento de Computação - DC

Jean Carlos Teixeira de Araujo
Universidade Federal do Agreste de Pernambuco - UFAPE
Departamento de Inovação Tecnológica e Empreendedorismo

Dedico este trabalho a todos que me ajudaram a produzi-lo.

Agradecimentos

Agradeço primeiramente aos meus pais, em especial a minha mãe Rejane, que sempre me incentivou e se dedicou pela minha educação. Agradeço a todos meus amigos e colegas de curso, pelo companheirismo e apoio durante o mestrado.

Agradeço também aos meus professores pela contribuição na minha formação acadêmica, em especial ao meu orientador Gabriel Alves, pelo incentivo, paciência, conhecimento, orientação e empenho dedicado à elaboração deste trabalho.

Resumo

A evasão e a retenção são problemas recorrentes em cursos de graduação, capazes de causar ociosidade de recursos ou gerar perda de receita nas universidades. A análise de sobrevivência permite verificar a influência de eventos, como a evasão, conclusão e vínculo estudantil, de acordo com o tempo e a probabilidade de ocorrência. Ela permite buscar soluções para resolver esses problemas com antecedência e evitar prejuízos às universidades. Este trabalho tem como objetivo analisar a evasão, conclusão e desvinculação de estudantes de graduação, através da proposta um modelo de Cadeia de Markov para realizar a análise de sobrevivência dos estudantes ao longo de cursos de graduação. A Cadeia de Markov absorvente proposta neste trabalho simula o progresso acadêmico, através de estados que representam os semestres em que os estudantes estão vinculados ao curso, com a inclusão de estados retidos. Nos estudos de caso, com dados de graduação de uma universidade pública brasileira, as análises realizadas identificaram diferenças no comportamento da evasão para cursos das áreas de agrárias, computação e saúde. Além disso, foi analisado como o semestre em que o estudante ficou retido impacta nas probabilidades de evasão, conclusão e desvinculação. Nas análises considerando as categorias curso, gênero e cor/raça, identificou-se como a diferença de comportamento dos estudantes destas categorias influenciam na evasão, conclusão e desvinculação. Também foram realizados experimentos que mostraram que para diminuir a taxa de evasão geral, o controle da retenção nos primeiros períodos possui um impacto maior que o controle da própria evasão.

Palavras-chave: Análise de Sobrevivência. Cadeia de Markov. Ensino Superior. Evasão. Retenção.

Abstract

Dropout and retention are recurring problems in undergraduate courses, capable of causing idle resources or generating revenue loss in universities. Survival analysis allows verifying the influence of events, such as dropout, completion and student bond, according to the time and probability of occurrence. It allows finding solutions to solve these problems in advance and avoid loss to universities. This work aims to analyze the dropout, completion and detachment of undergraduate students, through the proposal of a Markov Chain model to perform the survival analysis of students throughout undergraduate courses. The absorbing Markov Chain proposed in this work simulates academic progress, through states that represent the semesters in which students are bound to the course, with the inclusion of retained states. In the case studies, with graduation data from a Brazilian public university, the analyses identified differences in the behavior of dropout for courses in the areas of agrarian, computing and health. In addition, it was analyzed how the semester in which the student was retained impacts the probabilities of dropout, completion and detachment. In the analyzes considering the categories course, gender and race, it was identified how the difference in the behavior of students in these categories influence dropout, completion and detachment. Experiments were also carried out that showed that to decrease the overall dropout rate, controlling retention in the first semesters has a greater impact than controlling dropout itself.

Keywords: Evasion. Higher Education. Markov Chain. Retention. Survival Analysis.

Lista de Figuras

Figura 1 – Cadeia de Markov disponível em (BOUMI; VELA, 2020).	23
Figura 2 – Modelo de Cadeia de Markov.	31
Figura 3 – Modelo de Cadeia de Markov Absorvente.	32
Figura 4 – Exemplo de Análise de Sobrevivência.	34
Figura 5 – Metodologia.	36
Figura 6 – Modelo de Cadeia de Markov para Representar o Progresso Acadêmico.	41
Figura 7 – Análise de Sobrevivência da Evasão, Conclusão e Desvinculação e Probabilidades por Semestre para Concluídos, Evadidos e Vinculados.	45
Figura 8 – Análise de Sobrevivência ao Diminuir a Retenção em 25%, 50% e 75% no Primeiro Semestre.	46
Figura 9 – Análise de Sobrevivência da Evasão, Conclusão e Desvinculação para uma Redução de 50% na Evasão ou Retenção nos Quatro Primeiros Semestres.	48
Figura 10 – Gráfico de Pareto para a Categoria Curso.	50
Figura 11 – Análise de Sobrevivência do Modelo em Relação aos Dados para o Curso 1.	51
Figura 12 – Análise de Sobrevivência do Modelo em Relação aos Dados para o Curso 2.	51
Figura 13 – Análise de Sobrevivência para a Categoria Curso.	52
Figura 14 – Gráfico de Pareto para a Categoria Sexo.	54
Figura 15 – Análise de Sobrevivência do Modelo em Relação aos Dados para o Gênero Feminino.	54
Figura 16 – Análise de Sobrevivência do Modelo em Relação aos Dados para o Gênero Masculino.	55
Figura 17 – Análise de Sobrevivência para a Categoria Gênero.	56
Figura 18 – Gráfico de Pareto para a Categoria Cor/Raça.	58
Figura 19 – Análise de Sobrevivência do Modelo em Relação aos Dados para a Cor Parda.	59
Figura 20 – Análise de Sobrevivência do Modelo em Relação aos Dados para a Cor Branca.	59

Figura 21 – Análise de Sobrevivência do Modelo em Relação aos Dados para a Cor Preta.	60
Figura 22 – Análise de Sobrevivência para a Categoria Cor/Raça.	61
Figura 23 – Análise de Sobrevivência da Evasão para os Quatro Primeiros Semestres dos Cursos.	63
Figura 24 – Violin Plot da Evasão para os Quatro Primeiros Semestres dos Cursos.	64
Figura 25 – Análise de Sobrevivência por semestre de retenção.	65

Lista de tabelas

Tabela 1 – Trabalhos Relacionados.	25
Tabela 2 – Matriz de Transição da Cadeia de Markov.	31
Tabela 3 – Matriz de Transição da Cadeia de Markov Absorvente.	32
Tabela 4 – Resultados das Probabilidades da Redução da Retenção no Primeiro Semestre.	46
Tabela 5 – Resultados dos Tempos da Redução da Retenção no Primeiro Semestre.	46
Tabela 6 – Resultados das Probabilidades dos Experimentos para os Quatro Primeiros Semestres de Curso.	48
Tabela 7 – Resultados dos Tempos dos Experimentos para os Quatro Primeiros Semestres de Curso.	49
Tabela 8 – Intervalo de Confiança da Categoria Curso.	50
Tabela 9 – Resultados das Probabilidades e dos Tempos para a Categoria Curso. .	53
Tabela 10 – Intervalo de Confiança da Categoria Gênero.	53
Tabela 11 – Resultados das Probabilidades e dos Tempos para a Categoria Gênero.	56
Tabela 12 – Intervalo de Confiança da Categoria Cor/Raça.	57
Tabela 13 – Resultados das Probabilidades e dos Tempos para a Categoria Cor/Raça.	61
Tabela 14 – Intervalo de Confiança dos Cursos para Evasão nos Primeiros Quatro Semestres.	62

Lista de Siglas

CTMC	<i>Continuous-time Markov chain</i>
DTMC	<i>Discrete-time Markov chain</i>
IES	Instituições de Ensino Superior
IFES	Instituições Federais de Ensino Superior
INEP	Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira
MEC	Ministério da Educação
PNE	Plano Nacional de Educação
SISU	Sistema de Seleção Unificada
TCU	Tribunal de Contas da União
TSG	Taxa de Sucesso na Graduação
UFRPE	Universidade Federal Rural de Pernambuco

Sumário

1	Introdução	14
1.1	Motivação	17
1.2	Objetivos	18
1.3	Organização do Trabalho	19
2	Trabalhos Relacionados	20
2.1	Mineração de Dados	20
2.2	Modelagem e Simulação	21
2.3	Análise de Sobrevivência	24
2.4	Considerações Finais	25
3	Fundamentação Teórica	27
3.1	Conceitos Educacionais	27
3.2	Cadeias de Markov	28
3.2.1	Cadeias de Markov Absorventes	31
3.3	Análise de Sobrevivência	33
3.4	Considerações Finais	35
4	Metodologia	36
4.1	Entendimento do Problema	36
4.2	Implementação do Modelo	37
4.3	Definição de Parâmetros de Entrada	37
4.4	Execução do Modelo e Extração de Métricas	38
4.5	Análise dos Resultados	38
4.6	Considerações Finais	39
5	Modelo Proposto	40
5.1	Considerações Finais	42
6	Estudo de Caso	43
6.1	Estudo de Caso 1: Impacto da Retenção e Evasão para Cursos de Computação	43
6.1.1	Análise de Sobrevivência nos Cursos de Computação	44
6.1.2	Impacto da Retenção no Primeiro Semestre	45
6.1.3	Impacto da Retenção e Evasão nos Quatro Primeiros Semestres	47

6.2	Estudo de Caso 2: Análise de Sobrevivência Considerando Características dos Cursos e Estudantes	49
6.2.1	Curso	49
6.2.2	Gênero	53
6.2.3	Cor/Raça	57
6.3	Estudo de Caso 3: Evasão e Retenção para Cursos das Áreas de Agrárias, Computação e Saúde	62
6.3.1	Análise da Evasão nos Quatro Primeiros Períodos	62
6.3.2	Análise da Evasão, Conclusão e Desvinculação por Semestre de Retenção	64
6.4	Considerações Finais	65
7	Conclusão	67
7.1	Trabalhos Futuros	69
	Referências	70
	APÊNDICES	74
	APÊNDICE A Código da Cadeia de Markov	75

1 Introdução

A evasão no ensino superior é um fenômeno estudado há décadas e pode ser definida como a interrupção do progresso estudantil ao longo dos semestres (XAVIER; MENESES, 2020). Segundo a pesquisa do estado da arte apresentada em (MACIEL et al., 2019), a maior parte dos trabalhos buscam identificar elementos que contribuem para a evasão.

A evasão pode ser observada com relação a um único curso, a uma instituição ou mesmo ao sistema de ensino. Assim, um estudante trocar de curso em uma mesma instituição, pode ser visto como um “evadido” para o curso, mas não para a instituição e muito menos para o sistema de ensino como um todo. Neste trabalho, observamos a evasão sob a ótica do curso.

Trabalhos quantitativos e qualitativos realizados anteriormente, mostram que a evasão é um problema multifatorial (SILVA; SANTOS, 2017; SACCARO et al., 2019; CAMPOS, 2016; TIGRINHO, 2008), que possui causas emocionais, financeiras ou mesmo problemas relacionados ao curso, como a sua estrutura curricular ou seu sistema de ingresso. Do ponto de vista quantitativo, muitos trabalhos mostram que a retenção e a evasão ocorrem de forma mais intensa nos primeiros períodos dos cursos (XAVIER; MENESES, 2020; BOUMI; VELA, 2020). Neste trabalho, a análise da evasão ocorre de forma quantitativa, não sendo o objetivo investigar as suas causas.

A retenção é um assunto menos abordado, porém ela pode indicar um maior risco de evasão (SANTOS et al., 2021). Neste trabalho, a retenção é definida como quando o estudante é reprovado em uma quantidade de disciplinas maior que a quantidade padrão de um semestre ou caso ele tenha trancado em algum semestre, o que normalmente o levaria a não conseguir se formar no tempo ideal do curso.

A evasão e retenção intensa nos primeiros semestres pode ocasionar em uma dificuldade de preencher as vagas remanescentes dos estudantes evadidos. Portanto, é importante verificar como a retenção e evasão nos primeiros semestres do curso pode influenciar na evasão geral.

A conclusão e a desvinculação também são aspectos importantes para se avaliar em um curso. Neste estudo, a desvinculação representa o momento em que o estudante perde o vínculo com o curso, através da conclusão ou evasão. Esse conceito permite observar a probabilidade e tempo que o estudante leva para sair do curso, independente da forma

como o estudante perdeu o vínculo.

Uma forma de avaliar a evasão com maior precisão é considerar uma análise por grupos de estudantes, o que permite observar a relevância de determinadas características que contribuem para a evasão e visa fornecer dados para a criação de políticas direcionadas. Isso auxilia na identificação de alunos com maior probabilidade de evasão. Estudos que utilizam Mineração de Dados, Classificadores e Aprendizagem de Máquina (BRITO et al., 2020; CARRANO et al., 2019) servem como base para identificar quais atributos e características podem ser aplicados em modelos para uma maior precisão de suas análises. São identificados que atributos como curso, gênero, cor/raça, idade e cidade de residência, são alguns dos atributos que mais influenciam na evasão. Isso demonstra a importância de avaliar o comportamento de evasão com base nesses grupos de estudantes.

A evasão e a retenção são fatores que podem influenciar no progresso acadêmico e na chance de sucesso durante a graduação. Esses fatores podem se tornar um problema e impactar na gestão e nos recursos das universidades. Por isso, é necessário identificar o quanto esses fatores podem causar de impacto na probabilidade e no tempo de conclusão dos alunos. O uso de dados educacionais para prever o comportamento de estudantes tem sido amplamente difundido (MACIEL et al., 2019). Esses dados podem ser aplicados em modelos que permitem transformar os dados em métricas com o objetivo de contribuir para um melhor planejamento de cursos de graduação.

O uso de mineração de dados, técnicas de modelagem e aprendizagem de máquina permite avaliar o problema da evasão com precisão e prever de forma eficiente o comportamento dos estudantes que geram a retenção e a evasão (AMORIM et al., 2008). Diante disso, o uso de Cadeias de Markov é útil para analisar e gerar informações quantitativas sobre o progresso dos estudantes, considerando que permite simular eventos temporais discretos (CROSSEN, 2017).

A aplicação de modelos, como as Cadeias de Markov, permitem gerar métricas que auxiliam na tomada de decisão. Segundo (BREZAVŠČEK et al., 2017), a modelagem por Cadeias de Markov podem oferecer informações referentes ao progresso estudantil através de seus estados, além da possibilidade de prever o impacto de possíveis alterações no progresso acadêmico dos estudantes.

A modelagem por Cadeias de Markov é amplamente utilizada na literatura. Ela pode ser aplicada na área de educação (BREZAVŠČEK et al., 2017), expansão urbana

(ARSANJANI et al., 2013), mercado financeiro (BÄUERLE; RIEDER, 2011), área de saúde (DAI et al., 2018) e processamento de imagens (WANG et al., 2010).

As Cadeias de Markov são utilizadas para representar um sistema, através de um modelo estocástico a partir do qual podem ser realizadas análises e aferidas métricas probabilísticas (SHESKIN, 2016). Elas são formadas por estados e suas transições. Nelas é possível avaliar a ocorrência de eventos e seu progresso ao longo de um tempo. A modelagem por Cadeia de Markov é útil para modelar o progresso acadêmico, pois descreve sistemas onde a probabilidade do próximo estado depende apenas do estado atual, de modo similar ao representar os semestres necessários para a conclusão de um curso.

Nas Cadeias de Markov absorventes, cada estado pode chegar a um estado de absorção, que representa um estado final da cadeia, onde não é possível sair dele. Neste trabalho, os estados de absorção são a evasão e a conclusão. Assim, é possível aferir métricas como o tempo médio até que um evento ocorra, ou a probabilidade do sistema estar em um determinado estado e em um tempo t .

O modelo de Cadeia de Markov pode ser usado para prever informações sobre um estudante em qualquer momento em que ele esteja no curso. O que é uma vantagem, pois é capaz analisar o comportamento geral de um curso ou mesmo realizar previsões individualizadas. Por exemplo, é possível identificar a chance de evasão de um estudante retido no 4^o semestre de um curso de computação. Neste trabalho será utilizada uma Cadeia de Markov absorvente de tempo discreto, cuja vantagem é a possibilidade de aferir métricas, criar um modelo generalizado e realizar experimentos através dos parâmetros do modelo. Como resultado do modelo de Cadeia de Markov é possível extrair um vetor com os estados percorrido, além de métricas relacionadas ao tempo e probabilidade. Essas informações são aplicadas na Análise de Sobrevivência, permitindo gerar gráficos através das informações de tempo até os estudantes atingirem os estados de evasão ou conclusão.

A Análise de Sobrevivência é uma estatística não paramétrica usada para estimar a função de sobrevivência a partir de dados de eventos (KLEINBAUM; MITCHEL, 2010). Essa análise avalia o tempo até a ocorrência de eventos. Nela também pode ser analisada a probabilidade de ocorrência de um evento dentro de um intervalo de tempo, isso auxilia na observação de características que influenciam na probabilidade de acordo com o tempo.

A Análise de Sobrevivência é uma técnica amplamente utilizada em diversos contextos, como a área de saúde (HAGAR et al., 2014), educação (SACCARO et al.,

2019), financeira (MOORE; CHRISTIN, 2013) e social (INOUE et al., 2013), pois ela permite analisar a sobrevivência de um determinado evento de acordo com seu tempo de ocorrência.

Analisar os resultados obtidos através da Cadeia de Markov é importante para avaliar o impacto da evasão. Através dos dados coletados ao simular o progresso dos estudantes por meio da Cadeia de Markov, é possível utilizar Análise de Sobrevivência para avaliar o tempo até a ocorrência de eventos, como a evasão e a conclusão, observando a probabilidade e duração média até o evento. De forma, a avaliar diferentes situações e identificar características que podem propiciar um maior risco de evasão. Isso possibilita criar estratégias ou oferecer auxílio a esses estudantes, para aumentar a probabilidade de graduação e evitar desperdício de recursos das universidades.

Neste trabalho, o modelo de Cadeia de Markov absorvente proposto é capaz de representar o progresso dos estudantes de um curso de graduação. O modelo permite representar o progresso por semestres e o momento em que o estudante passa a ser considerado “retido”. Os parâmetros do modelo são definidos considerando as probabilidades de retenção, evasão e conclusão por semestre. Este modelo permite aferir métricas como o tempo e a probabilidade de um estudante evadir ou concluir o curso. Portanto, pode-se analisar a evasão nos primeiros semestres ou analisar a evasão, conclusão e desvinculação por semestre em que o estudante ficou retido.

As Cadeias de Markov têm a vantagem de possibilitar a realização de experimentos alterando-se os parâmetros do modelo conforme grupos de estudantes e análise a ser realizada. Portanto, pode-se observar como as mudanças nas probabilidades de retenção ou evasão nos primeiros semestres podem impactar na probabilidade de evasão, conclusão e desvinculação. Além da realização de experimentos estratificados por curso, gênero ou cor/raça.

1.1 Motivação

A evasão é considerado um problema de âmbito internacional, que atinge também o Brasil, onde o índice de estudantes que evadiram em 2017 foi de 27,5% (INEP, 2020). Esse problema afeta tanto Instituições de Ensino Superior (IES) públicas, como as IES privadas pois pode implicar em perda de receita ou subutilização de recursos (SILVA; SANTOS,

2017).

O ensino superior tem um papel importante no desenvolvimento de um país. A criação do REUNI, Programa de Apoio a Planos de Reestruturação e Expansão das Universidades Federais, tem em seus planos um aumento na oferta de cursos superiores e uma diminuição na taxa de evasão. Porém, as mudanças no processo de entrada nas universidades ao longo dos anos podem ter contribuído para o impacto sobre a retenção e evasão, como a ampla utilização do Sistema de Seleção Unificada (SISU) como sistema de entrada nas universidades. As diretrizes do Plano Nacional de Educação (PNE) para melhorar a educação superior estabelece a meta de elevar a taxa de matrícula no ensino superior, através da ampliação da oferta de vagas por meio de políticas de acesso, que incluem políticas de inclusão e de assistência estudantil para expandir o acesso e permanência na educação superior dos estudantes (EDUCAÇÃO, 2020).

Porém, o crescimento da oferta de vagas não necessariamente influenciou no aumento na quantidade de concluintes nos cursos de graduação, ainda que exista a proposta de elevar gradualmente a taxa de conclusão média dos cursos de graduação presenciais nas universidades públicas para 90%. Segundo o (INEP, 2020), apesar de um aumento na quantidade de ingressantes com a expansão do ensino superior, não houve um aumento proporcional equivalente dos estudantes concluintes nos últimos anos. Cerca de 23,3% das vagas em instituições federais estão remanescentes, isso ocorre principalmente por causa da evasão. A ampliação do acesso à educação superior para ser bem sucedida necessita, além do aumento no número de vagas em instituições de ensino superior, a redução da taxa de evasão e a ocupação das vagas remanescentes.

1.2 Objetivos

O objetivo deste trabalho é analisar o tempo e a probabilidade de evasão, conclusão e desvinculação de estudantes de graduação, através de um modelo de Cadeia de Markov absorvente. O modelo é capaz de representar o progresso dos estudantes ao longo dos anos em que estão matriculados em um curso, considerando também o momento em que o estudante passa a ser considerado “retido”. Uma vantagem da utilização de Cadeias de Markov, é a possibilidade de realizar experimentos, alterando-se os parâmetros do modelo. Assim, é possível observar como mudanças na probabilidade de retenção ou evasão

nos primeiros semestres podem impactar na probabilidade de evasão geral. Além disso, associado à Análise de Sobrevivência é possível observar o comportamento dos eventos evasão, conclusão e desvinculação de acordo com o semestre que o estudante está vinculado a um curso.

Para atingir o objetivo geral foram definidos os seguintes objetivos específicos:

- Propor um modelo de Cadeia de Markov para representar o progresso acadêmico de estudantes de graduação, em que os estados representam os semestres em que os estudantes estão progredindo no curso até a evasão ou conclusão, considerando a representação de estudantes retidos;
- Utilizar Análise de Sobrevivência para avaliar como diferentes aspectos influenciam no tempo e na probabilidade de evasão, de conclusão e de desvinculação de estudantes;
- Realizar estudos de caso com dados de uma universidade pública federal, para avaliar o modelo proposto;
- Criar os parâmetros do modelo conforme grupos de estudantes, para previsões mais precisa e individualizadas conforme suas características;
- Realizar análises com diferentes probabilidades de retenção e evasão nos primeiros períodos.

1.3 Organização do Trabalho

Este trabalho está organizado em seis capítulos: o Capítulo 2 apresenta os trabalhos relacionados a esta pesquisa; o Capítulo 3 apresenta a fundamentação teórica, onde descreve os principais conceitos que guiam este trabalho; o Capítulo 4 apresenta a metodologia, através da descrição das etapas do trabalho; o Capítulo 5 apresenta o modelo proposto; o Capítulo 6 descreve os estudos de caso; e o Capítulo 7 apresenta as contribuições obtidas neste trabalho e os trabalhos futuros.

2 Trabalhos Relacionados

Neste capítulo é apresentada a revisão da literatura. O uso de técnicas de modelagem, análise de sobrevivência, mineração de dados, aprendizagem de máquina e pesquisas quantitativas e qualitativas, permitem analisar fatores que influenciam na retenção, evasão, conclusão e vínculo estudantil.

2.1 Mineração de Dados

Nos trabalhos (LIMA; COUTINHO, 2019) e (RODRIGUES et al., 2015), os autores realizam uma análise descritiva sobre a evasão e o desempenho acadêmico em cursos da área de tecnologia, através de dados de notas, frequências, reprovações e vínculo, com o intuito de investigar o comportamento de turmas ao longo do tempo, para encontrar causas que levam os estudantes a evadir do curso. Além de analisar o problema da evasão na perspectiva dos alunos. Os resultados apontaram problemas na estrutura curricular e um alto percentual de estudantes em atraso no curso. Neste trabalho, o atraso dos estudantes nos cursos será analisado no formato de retenção, onde o progresso dos estudantes retidos na modelagem será representado de forma diferente dos estudantes não retidos.

Técnicas de mineração de dados podem ser aplicadas para aferir indicadores e características que contribuem no entendimento da evasão. Em (BRITO et al., 2020), o algoritmo Random Forest é utilizado para identificar os atributos mais relevantes para a evasão. Os resultados apontaram o ano de admissão do aluno, cor/raça, gênero e estado civil como as características mais importantes, com o classificador apresentando uma acurácia média de 70%. Neste trabalho, serão utilizados curso, gênero, cor/raça e área de conhecimento para verificar como impactam na evasão, conclusão e desvinculação estudantil.

Nos estudos (JESUS et al., 2021; RIGO et al., 2014; TEODORO; KAPPEL, 2020), são aplicadas técnicas de aprendizagem de máquina, como Naive Bayes, K-Nearest Neighbors, Árvores de Decisão, Random Forest, Redes Neurais e Regressão Logística para prever a evasão dos estudantes. Os resultados apontam um desempenho entre 75% e 95% na predição de estudantes em risco de evasão no primeiro semestre. Isso permite, conforme visto em (BELTRAN et al., 2019; TEODORO; KAPPEL, 2020), monitorar o risco de

evasão dos estudantes e oferecer auxílio aos que apresentam maior risco de evasão. Neste trabalho, serão realizadas análises para prever o comportamento da evasão, conclusão e desvinculação ao alterar parâmetros de retenção e evasão nos primeiros semestres de curso.

2.2 Modelagem e Simulação

Simular a progressão acadêmica através de eventos discretos é muito comum na literatura. Nos estudos (JUHÁSOVA et al., 2016; BALOGH et al., 2013; BALOGH; KUCHÁRIK, 2019; TAJ; KUMARAVEL, 2019) são aplicadas Redes de Petri para modelar o currículo de programas de estudo de graduação. As modelagens são realizadas considerando os semestres e as disciplinas cursadas, assim é possível representar o fluxo de aprendizagem individual de um aluno. Como resultado, os modelos permitem melhor planejamento dos cursos e gerenciamento do processo de aprendizagem do aluno para otimizar a chance do estudante concluir o curso. Neste trabalho, será realizada a modelagem do progresso acadêmico por meio dos semestres em que os estudantes estão vinculados, pois permite representar a progressão dos estudantes até a conclusão ou evasão.

O uso de Cadeias de Markov para representar o progresso acadêmico é amplamente utilizado na literatura. Pode ser utilizado para avaliar desde a probabilidade de conclusão do ensino médio (ALVES et al., 2020) até prever a quantidade de estudantes que irão se matricular na pós-graduação (ADAM, 2015). Porém, a abordagem mais utilizada é a avaliação da evasão e conclusão em cursos de graduação. Neste trabalho, será analisada a evasão, conclusão e desvinculação dos estudantes, uma vez que essas métricas permitem avaliar e melhorar o planejamento de cursos de graduação.

Ao modelar o progresso acadêmico por meio de Cadeias de Markov, existe a necessidade de definição dos estados. O trabalho (HLAVATÝ; DÖMEOVÁ, 2014) considera cada disciplina do curso como um estado da Cadeia de Markov, enquanto em (AL-AWADHI; KONSOWA, 2007; ELEDUM; IDRIS, 2019; BAIRAGI; KAKATY, 2017) os estados representam os anos ou semestres do curso necessários para se formar. Os resultados permitem estimar o tempo médio que os alunos passam em cada semestre, o número médio de semestres necessários para concluir o curso e a probabilidade de evasão ou conclusão do estudante. Neste trabalho, os estados foram definidos pelos semestres de vínculo do estudante e seu estado de retenção.

Em (BREZAVŠČEK et al., 2017) o objetivo é desenvolver um modelo estocástico para estimar e monitorar continuamente vários indicadores de qualidade e eficácia de um programa de estudos. Um modelo de Cadeia de Markov é aplicado para investigar o padrão de matrícula dos alunos, seu desempenho e progresso acadêmico em uma instituição de ensino superior. A progressão dos alunos foi estimada considerando a duração prevista do estudo. Os dados obtidos através do modelo permitem estimar a progressão dos alunos entre os diferentes estágios, o tempo esperado que um aluno passa em um determinado estágio, a duração prevista para concluir o curso, a probabilidade de graduação ou desistência, e prever a futura matrícula dos estudantes.

Em (ARAÚJO, 2019), é apresentado um modelo de Cadeias de Markov absorventes para simular a trajetória acadêmica de estudantes de graduação de uma universidade. O modelo leva em consideração a quantidade de reprovações para representar os estados, além dos estados trancado, formado e evadido. Como resultado, é possível observar e simular a taxa de evasão dos estudantes, através de alterações nas probabilidades de transição do modelo.

Dentre os estudos citados acima, (BREZAVŠČEK et al., 2017) e (ARAÚJO, 2019) têm a representação do estado trancado, onde o aluno pode transitar para este estado a partir de qualquer estado que representa o progresso estudantil. Porém, ao seguir a matriz de transição, o estudante poderia sair do estado trancado para voltar em um estado diferente ao qual se encontrava antes de atingir o estado trancado. Portanto, no presente trabalho o trancamento está incluído, porém de forma implícita, em virtude do progresso acadêmico aqui simulado ser representado pelos semestres em que o estudante está no curso, o que inclui o tempo em que ele passa trancado.

O estudo (BOUMI; VELA, 2020) tem como objetivo avaliar a chance de graduação de uma universidade como um todo, bem como para grupos específicos. Simulações foram realizadas considerando aleatoriamente grupos de alunos de diferentes tamanhos de amostra utilizando Cadeia de Markov. No modelo, representado pela Figura 1, os alunos são classificados conforme quatro anos de curso e a transição entre os estados ocorre com base em seus créditos acadêmicos. Além disso, o modelo usa técnicas para reduzir o desvio padrão amostral da taxa de graduação, entre elas a Cadeia de Markov multinível. Dessa forma, foram estimadas as probabilidades de graduação para grupos com diferentes tamanhos de amostra e obteve-se uma média de 68.65%.

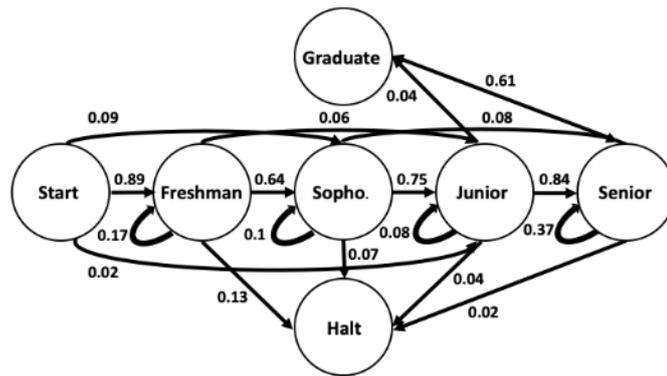


Figura 1 – Cadeia de Markov disponível em (BOUMI; VELA, 2020).

Apesar de (BREZAVŠČEK et al., 2017), (ARAÚJO, 2019) e (BOUMI; VELA, 2020) representarem o progresso acadêmico com Cadeia de Markov, eles não demonstram a criação de estados específicos para o progresso dos alunos retidos. Dessa forma, todos os estudantes têm o progresso acadêmico representado pelos mesmos estados e a retenção é representada por *self-loop*. Isso não demonstra o real comportamento dos estudantes retidos e pode fazer com que os dados que o modelo gera sejam irrealistas em relação ao tempo e probabilidade do estudante para a conclusão ou evasão. Portanto, no presente trabalho, o progresso estudantil é definido em relação à quantidade de semestres em que o estudante está vinculado ao curso. Isso permite criar progressos alternativos entre os alunos retidos e não retidos.

Analisar o progresso estudantil levando em consideração diferentes grupos de estudantes é uma prática que também pode ser utilizada em Cadeias de Markov. O trabalho (ALAWADHI; KONSOWA, 2010) tem como objetivo utilizar Cadeia de Markov para estimar o desempenho e evasão dos alunos. Como resultado, é produzido um comparativo entre diferentes áreas de conhecimento da universidade. Isso permite uma maior precisão e detalhamento dos resultados gerados pela Cadeia de Markov. Neste trabalho, serão realizadas análises considerando grupos de estudantes divididos por curso, área de conhecimento, gênero e cor/raça.

As Cadeias de Markov permitem a modelagem e análise de sistemas de eventos discretos. Assim, elas são capazes de representar de forma eficaz o progresso de alunos em cursos de graduação. Essa técnica destaca-se entre as demais utilizadas para avaliar a evasão, pois permitem criar um modelo generalizado capaz de prever o comportamento de estudantes em relação à probabilidade e o tempo, com a possibilidade de ser aplicado em diversos contextos sem necessidade de alteração do modelo, apenas de seus parâmetros, o

que permite também a realização de experimentos. Estes modelos permitem, por exemplo, observar o comportamento da taxa de evasão global ao mudar a retenção ou evasão em um período específico ou avaliar e prever diferentes comportamentos baseados no efeito da manipulação de parâmetros conforme as especificidades de cada grupo de estudantes. Desta forma, esta pesquisa pretende realizar uma modelagem do progresso dos estudantes de graduação utilizando Cadeia de Markov com a criação de estados específicos para representar os alunos retidos, o que ainda não foi feito nos trabalhos anteriores.

2.3 Análise de Sobrevivência

O uso de Análise de Sobrevivência permite avaliar a probabilidade de eventos em relação ao tempo. Neste contexto, os trabalhos (SACCARO et al., 2019), (CAMPOS, 2016) e (COSTA et al., 2018) têm o objetivo de analisar variáveis que influenciam no tempo de permanência e probabilidade de conclusão e evasão para cursos em instituições públicas e privadas. Foram utilizadas técnicas de Análise de Sobrevivência, tais como Kaplan-Meier, teste Log-rank e modelos de regressão semi-paramétricos, levando em consideração o número de semestres do curso, gênero, idade, forma de ingresso, turno, desempenho acadêmico (média das notas) e a existência de reprovação e trancamento entre os estudantes. Como resultado, percebe-se que a evasão é maior nos primeiros períodos e em instituições privadas.

A probabilidade de evasão maior durante o primeiro ano do curso foi um fator recorrente durante os trabalhos analisados (CAMPOS, 2016; SACCARO et al., 2019; COSTA et al., 2018). Ela demonstra, entre outros fatores, um efeito relacionado ao Sistema de Seleção Unificada (SISU), onde estudantes que entram na universidade por meio deste apresentam uma probabilidade de evadir maior em comparação a alunos que entram por outros meios, como avaliação seriada (SACCARO et al., 2019). Essa observação demonstra a importância de análises para avaliar como a diminuição da evasão e retenção nos primeiros anos pode impactar na evasão do curso como um todo. Neste trabalho, a Análise de Sobrevivência será utilizada para analisar as probabilidades dos eventos de evasão, conclusão e desvinculação conforme os semestres de curso.

A Tabela 1 apresenta um quadro comparativo entre os principais trabalhos relacionados analisados e este trabalho. As linhas representam os trabalhos analisados e as

colunas as características desses trabalhos. Os fatores avaliados referem-se aos seguintes aspectos: (a) Modelagem; (b) Análise de Sobrevivência; (c) Análise Quantitativa; (d) Análise Qualitativa; e (e) Análise Estratificada.

Tabela 1 – Trabalhos Relacionados.

Trabalho / Características	a	b	c	d	e
(LIMA; COUTINHO, 2019)				✓	
(RODRIGUES et al., 2015)				✓	
(BRITO et al., 2020)			✓		✓
(JESUS et al., 2021)			✓		
(RIGO et al., 2014)			✓		
(TEODORO; KAPPEL, 2020)			✓		
(JUHÁSOVA et al., 2016)	✓		✓		
(BALOGH et al., 2013)	✓		✓		
(BALOGH; KUCHÁRIK, 2019)	✓		✓		
(TAJ; KUMARAVEL, 2019)	✓		✓		
(BREZAVŠČEK et al., 2017)	✓		✓		
(ARAÚJO, 2019)	✓		✓		
(BOUMI; VELA, 2020)	✓		✓		✓
(ALAWADHI; KONSOWA, 2010)	✓		✓		✓
(SACCARO et al., 2019)		✓	✓		✓
(CAMPOS, 2016)		✓	✓		
(COSTA et al., 2018)		✓	✓		
Este Trabalho	✓	✓	✓		✓

2.4 Considerações Finais

Neste capítulo, foram apresentados os trabalhos relacionados que contribuíram como embasamento teórico e científico para esta dissertação. Os trabalhos relacionados

apresentados incluíram estudos que analisam o comportamento de estudantes em relação à evasão ou conclusão, através de técnicas de mineração de dados, modelagem e simulação ou Análise de Sobrevivência. No próximo capítulo será apresentada a fundamentação teórica.

3 Fundamentação Teórica

Neste capítulo é apresentada a fundamentação teórica, relacionada a este trabalho, que permite auxiliar na compreensão deste estudo. Serão abordados os conceitos educacionais, a Cadeia de Markov e a Análise de Sobrevivência.

3.1 Conceitos Educacionais

Diversos conceitos educacionais são utilizados para estudar aspectos da educação superior. Entre esses conceitos, as situações acadêmicas são de grande relevância e podem divergir conforme o estudo. As principais situações acadêmicas são o vínculo, retenção, evasão e conclusão. Elas representam o estado que o estudante está no curso e permitem manter uma organização dos estudantes.

O estudante com vínculo corresponde àquele com matrícula ativa no curso, seja “cursando” ou “matrícula trancada” (MEC, 2020). Este estudante pode ser classificado ainda como retido, além de progredir para uma situação de desvinculado, através da evasão ou conclusão.

Neste estudo, o estudante retido é determinado como quando o estudante é reprovado em uma quantidade de disciplinas maior que a quantidade de um período ou caso ele tenha trancado em algum semestre, o que o levaria a não conseguir se formar em um tempo válido, porém deve ser levado em consideração que existe a possibilidade do estudante fazer menos disciplinas do que o previsto por período, sem isso o definir como retido.

A evasão é definida como o estudante que perdeu o vínculo com o curso por qualquer motivo que não seja a conclusão. Isso inclui não efetuar a matrícula no prazo pré-estabelecido, transferência interna ou externa, desistência, reopção ou jubramento. Para calcular a evasão o Ministério da Educação (MEC) adota a Equação 3.1. Nela são utilizados dados de percentual de evasão (E), número de ingressantes (Ni), número de diplomados (Nd) e número de retidos (Nr) (MEC, 2020).

$$E = [(Ni - Nd - Nr)/Ni]x100 \quad (3.1)$$

A conclusão representa o estudante que finalizou o curso com sucesso. Isto significa

concluir todos os requisitos curriculares, por meio da conclusão da quantidade de créditos necessários e apresentação de um trabalho final. Ou seja, o estudante que adquiriu o título de graduado.

Calcular a quantidade de concluintes é um aspecto essencial para o acompanhamento de um curso. As Instituições Federais de Ensino Superior (IFES), precisam apresentar o indicador da Taxa de Sucesso na Graduação (TSG), conforme definido pelo Tribunal de Contas da União (TCU), por meio da decisão nº 408/2002 – TCU, para serem incluídos no Relatório de Gestão das IFES. Esse indicador é calculado pela relação entre o número de ingressantes (Ni) e número de diplomados (Nd), conforme a Equação 3.2 (TCU, 2004).

$$TSG = Nd/Ni \quad (3.2)$$

Por fim, a desvinculação ocorre quando o estudante sai do curso, através da conclusão ou evasão. Esse conceito permite observar a probabilidade e tempo que o estudante leva para sair do curso, independente de como ocorra.

3.2 Cadeias de Markov

As Cadeias de Markov permitem avaliar a ocorrência de eventos e seu progresso ao longo de um tempo. Elas são formadas por um conjunto finito de estados e suas transições. Esses estados podem ser definidos por $S = 0, 1, 2, \dots, N$, enquanto as transições são definidas conforme um intervalo de tempo t . As Cadeias de Markov são utilizadas para representar um sistema através de um processo estocástico, a partir do qual podem ser realizados experimentos e aferidas métricas probabilísticas (SHESKIN, 2016). Neste trabalho, foi utilizada uma Cadeia de Markov Absorvente de tempo discreto.

Um processo estocástico pode ser definido como um conjunto de variáveis aleatórias $\{X_t : t \in T\}$, onde X_t representa uma variável aleatória em um tempo $t \in T$, se $T \subseteq \mathbb{R}_+ = [0, \infty)$. O conjunto de todos os valores possíveis de X_t é conhecido como o espaço de estados S do processo estocástico (BOLCH et al., 2006).

O processo estocástico pode ser caracterizado por uma função de distribuição para um determinado conjunto de variáveis aleatórias $X_{t_1}, X_{t_2}, \dots, X_{t_n}$, vetor de parâmetros $t = (t_1, t_2, \dots, t_n) \in \mathbb{R}^n$, e vetor de estados $s = (s_1, s_2, \dots, s_n) \in \mathbb{R}^n$, onde $t_1 < t_2 < \dots < t_n$.

Um processo estocástico $X_t : t \in T$ constitui um processo de Cadeia de Markov, se para todo $0 = t_0 < t_1 < \dots < t_n < t_{n+1}$ e todo $s_i \in S$ a condição da função de distribuição $X_{t_{n+1}}$ depende apenas do último valor de X_{t_n} e não dos valores anteriores $X_{t_0}, X_{t_1}, \dots, X_{t_{n-1}}$. Portanto, a Equação 3.3 define um processo estocástico (BOLCH et al., 2006)..

$$P(X_{t_{n+1}} \leq s_{n+1} | X_{t_n} = s_n, X_{t_{n-1}} = s_{n-1}, \dots, X_{t_0} = s_0) = P(X_{t_{n+1}} \leq s_{n+1} | X_{t_n} = s_n). \quad (3.3)$$

A definição de um processo de Cadeia de Markov pode ser adotada em parâmetros discretos e contínuos. Por isso, as Cadeias de Markov podem ser classificadas em *Continuous-time Markov chain* (CTMC) ou *Discrete-time Markov chain* (DTMC). As CTMC se diferem das DTMC pois as transições de estado podem ocorrer em instantes arbitrários de tempo e não apenas em tempo discreto, como é o caso das DTMC. Portanto, o processo estocástico das CTMC fica definido como na Equação 3.4 (BOLCH et al., 2006).

$$P(X_{t_{n+1}} = s_{n+1} | X_{t_n} = s_n, X_{t_{n-1}} = s_{n-1}, \dots, X_{t_0} = s_0) = P(X_{t_{n+1}} = s_{n+1} | X_{t_n} = s_n). \quad (3.4)$$

As Cadeias de Markov de parâmetro discreto apresentam um conjunto de estados discreto S e um conjunto de tempos discretos T . Onde as propriedades da Cadeia de Markov podem ser descritas dado um processo estocástico $\{X_0, X_1, \dots, X_{n+1}, \dots\}$ nos tempos $0, 1, \dots, n + 1$, se a propriedade vale para todo $n \in \mathbb{N}_0$ e todo $s_i \in S$, como descrito na Equação 3.5 (BOLCH et al., 2006).

$$P(X_{n+1} = s_{n+1} | X_n = s_n, X_{n-1} = s_{n-1}, \dots, X_0 = s_0) = P(X_{n+1} = s_{n+1} | X_n = s_n). \quad (3.5)$$

As DTMC evoluem ao longo do tempo, dado um estado inicial de acordo com as probabilidades de transição. Essas probabilidades determinam o estado seguinte e ocorrem observando apenas o estado atual, sem considerar os estados anteriores. Isso ocorre através da transição de um estado s_n no tempo n , para um estado s_{n+1} no tempo $(n + 1)$. A Equação 3.6 representa a transição do estado i para o estado j no tempo n (BOLCH et al., 2006).

$$p_{ij}^{(1)}(n) = P(X_{n+1} = s_{n+1} = j | X_n = s_n = i). \quad (3.6)$$

As probabilidades de transição acontecem em n passos, através da transição do estado i no tempo k , para o estado j no tempo l . Essas transições podem ser formadas de transições intermediárias, do estado i no tempo k para um estado intermediário h no tempo m e então do estado h no tempo m para o estado j no tempo l , onde $k < m < l$ e $n = l - k$. Essas transições levam a equação Chapman-Kolmogorov, definida pela Equação 3.7 (BOLCH et al., 2006).

$$p_{ij}^{(n)}(k, l) = \sum_{h \in S} p_{ih}^{(m-k)}(k, m) p_{hj}^{(l-m)}(m, l), 0 \leq k < m < l \quad (3.7)$$

As Cadeias de Markov podem ser definidas de acordo com o uso do tempo em suas transições. Nas Cadeias de Markov de Tempo Discreto (DTMC) as transições sempre levam um tempo fixo e são baseadas em probabilidades. Essas cadeias também podem ser definidas como estacionárias, ou seja, quando as probabilidades de transição são constantes em relação ao tempo. Dessa forma, os estados de transição $v = (v_0, v_1, \dots, v_i, \dots)$ seguem as probabilidades de transição $P = [p_{ij}]$, onde essa relação pode ser expressa pela Equação 3.8 (BOLCH et al., 2006).

$$v = vP, \sum_{i \in S} v_i = 1 \quad (3.8)$$

Uma Cadeia de Markov pode ser formada por diversos tipos de estados, dependendo de sua aplicação. O estado recorrente ocorre quando ele pode ser visitado várias vezes. No estado transiente, os estados podem progredir entre diferentes estados conforme as probabilidades de transição. Por fim, o estado de absorção, é o último estado disponível para onde a transição pode ocorrer, ou seja, quando um evento se encontra nesse estado ele não pode sair mais dele.

As probabilidades de transição entre os estados formam a matriz de transição. As matrizes de transição são formadas por vetores bidimensionais com n linhas e n colunas, sendo n a quantidade de estados da Cadeia de Markov, onde cada linha i representa o estado atual no tempo t , enquanto cada coluna j representa o estado seguinte, no tempo

$t + 1$. A matriz de transição é expressa pela Equação 3.9.

$$P = [p_{ij}]_{n \times n} = \begin{bmatrix} p_{1,1} & p_{1,2} & \cdots & p_{1,n} \\ p_{2,1} & p_{2,2} & \cdots & p_{2,n} \\ \vdots & \vdots & \ddots & \vdots \\ p_{n,1} & p_{n,2} & \cdots & p_{n,n} \end{bmatrix} \quad (3.9)$$

A Figura 2 apresenta um exemplo de Cadeia de Markov, composta pelos estados $E1$ e $E2$, sendo ambos classificados como recorrentes e transientes. As transições nesses estados podem ocorrer de um estado para o outro ou para o mesmo estado anterior. A Tabela 2 representa um exemplo da matriz de transição.

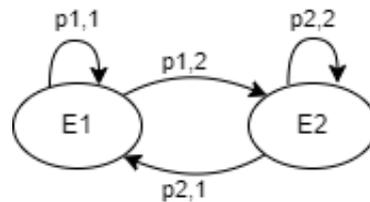


Figura 2 – Modelo de Cadeia de Markov.

Tabela 2 – Matriz de Transição da Cadeia de Markov.

	E1	E2
E1	0,5	0,5
E2	0,5	0,5

3.2.1 Cadeias de Markov Absorventes

Uma Cadeia de Markov é definida como absorvente quando as probabilidades de transição permitem os estados chegarem a estados de absorção, onde não é mais possível realizar transições. As Cadeias de Markov absorventes podem ser descritas como a equação apresentada em 3.10, onde Q é a matriz de transição entre os estados transientes, R é a

matriz de transição dos estados transientes para os estados absorventes, 0 é a matriz nula e I é a matriz identidade.

$$P = \begin{bmatrix} Q & R \\ 0 & I \end{bmatrix} \quad (3.10)$$

No caso de Cadeias de Markov absorventes, é possível aferir métricas como o tempo médio até que um evento ocorra, ou a probabilidade do sistema estar em um determinado estado e em um tempo t , por meio da Matriz Fundamental N , conforme a equação 3.11.

$$N = (I - Q)^{-1} \quad (3.11)$$

A Figura 3 apresenta um exemplo de Cadeia de Markov absorvente, adaptado do exemplo de Cadeia de Markov mostrado anteriormente. O estado $E1$ representa o estado inicial e os estados $E1$ e $E2$ são classificados como recorrentes e transientes. Observa-se também que o estado $E2$ permite que a Cadeia de Markov prossiga por mais de um caminho, realizando a transição para o estado $E1$ ou $E3$. O estado $E3$ representa o estado de absorção, que demonstra um estado final da cadeia. A matriz de transição desse exemplo é descrita conforme a Tabela 3.

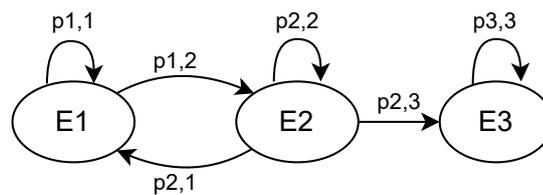


Figura 3 – Modelo de Cadeia de Markov Absorvente.

Tabela 3 – Matriz de Transição da Cadeia de Markov Absorvente.

	E1	E2	E3
E1	0,5	0,5	0
E2	0,33	0,33	0,33
E3	0	0	1

3.3 Análise de Sobrevivência

A Análise de Sobrevivência é uma técnica estatística utilizada principalmente na área de saúde. Porém, outras áreas de conhecimento como educação, financeira e sociologia também vêm utilizando essa técnica com resultados relevantes (JUNIOR et al., 2012). A Análise de Sobrevivência permite avaliar o tempo até a ocorrência de eventos, como a evasão e a conclusão de estudantes, observando a probabilidade e duração média até o evento (KLEINBAUM; MITCHEL, 2010).

A Análise de Sobrevivência é definida pela função de sobrevivência, que leva em consideração um tempo t até a ocorrência de um determinado evento e . O tempo de sobrevivência é uma variável t , que permite apenas valores positivos. Esse valor de t pode ser determinado através de uma função de sobrevivência. A função de sobrevivência representa o tempo de sobrevivência conforme uma distribuição de probabilidade. Ela é demonstrada na Equação 3.12, onde $S(t)$ é a função de sobrevivência, P é a probabilidade, t é o tempo e T é uma variável aleatória positiva, que indica o tempo de sobrevivência de um evento.

$$S(t) = P(T > t). \quad (3.12)$$

Na Figura 4, é representado um exemplo de Análise de Sobrevivência. Nela é demonstrada a sobrevivência de um evento num intervalo de tempo de 0 até 20, onde a cada uma unidade de tempo o evento ocorre e sua probabilidade de sobrevivência diminui.

A Análise de Sobrevivência apresenta diversas técnicas úteis para avaliar o tempo até a ocorrência de um determinado evento. Uma dessas técnicas é a censura, ela permite terminar a avaliação de um evento antes da ocorrência do seu evento de interesse. Dessa forma, permite uma análise parcial do tempo de sobrevivência do evento analisado. Isso acontece quando as informações relevantes para um estudo estão dentro de um determinado intervalo de tempo.

Existem três tipos de censura. A censura à direita de tipo I ocorre quando é determinado um tempo específico para o término do estudo, então todas as amostras em que não ocorreram o evento de interesse são censuradas conforme esse tempo preestabelecido. A censura à direita de tipo II ocorre quando acontece o evento de interesse em um número preestabelecido de amostras. Por fim, a censura aleatória acontece quando por qualquer

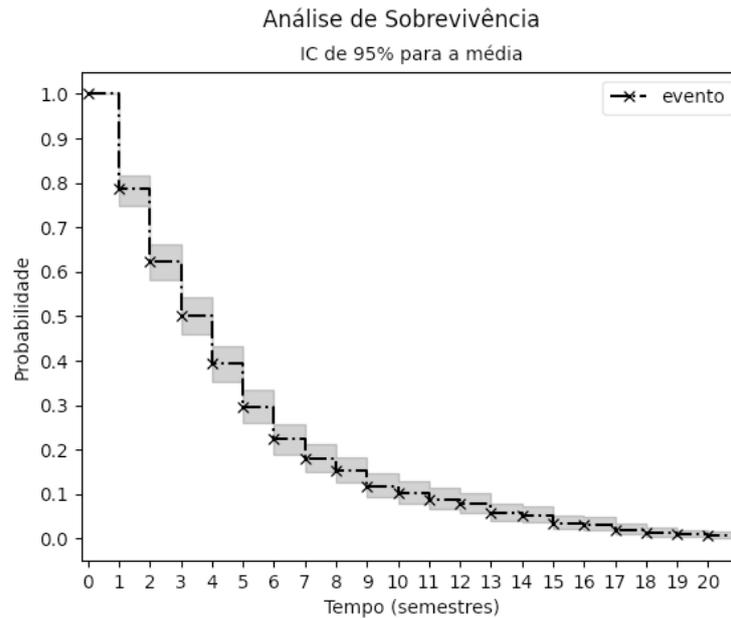


Figura 4 – Exemplo de Análise de Sobrevivência.

motivo aleatório uma amostra é removida durante o decorrer do estudo, como quando os participantes abandonam o estudo, o dado é perdido durante o estudo ou chega ao término antes do fim do estudo (COLOSIMO; GIOLO, 2006). Neste trabalho, será utilizada a censura à direita de tipo I, quando o estudante atinge o tempo limite de vinte semestres para concluir o curso.

Na Análise de Sobrevivência diversas técnicas estatísticas podem ser utilizadas para analisar dados em relação ao seu tempo de sobrevivência. A Análise de Sobrevivência de Kaplan-Meier, a mais comumente encontrada em diversos estudos, é uma estatística não paramétrica usada para estimar a função de sobrevivência a partir de dados de eventos (KAPLAN; MEIER, 1958). Ela apresenta resultados com precisão, visto que não apresenta vícios em grandes amostras e para amostras pequenas se mostra superior a outras técnicas. As técnicas não paramétricas são mais flexíveis, já que têm poucas restrições quanto a distribuições de tempo da população de onde foram produzidas as amostras. Porém, apresentam a limitação de não permitir avaliar o comportamento de muitas variáveis ao mesmo tempo.

Portanto, com base nas características apresentadas pela Análise de Sobrevivência, ela demonstra ser um bom complemento com as Cadeias de Markov para avaliar a evasão, conclusão e desvinculação no ensino superior.

3.4 Considerações Finais

Neste capítulo, foi apresentada a fundamentação teórica para a apresentação de conceitos relacionados à educação, a Cadeia de Markov e a Análise de Sobrevivência. O detalhamento desses temas auxiliam na compreensão do modelo e dos estudos de caso apresentados neste trabalho. O próximo capítulo apresentará a metodologia abordada.

4 Metodologia

A metodologia proposta neste estudo pretende elaborar um planejamento para realizar análises do tempo e probabilidade de evasão, conclusão e desvinculação de estudantes de graduação. Para isso, foi produzido um modelo de Cadeia de Markov absorvente e aplicado dados de uma universidade pública federal. Os dados aplicados foram utilizados como parâmetro para o modelo, que permite simular o comportamento dos estudantes.

A metodologia aplicada neste trabalho é baseada na metodologia para modelagem e simulação proposta por (OBAIDAT; BOUDRIGA, 2010; OBAIDAT; PAPADIMITRIOU, 2012). O processo é composto por cinco etapas que vão desde o entendimento do problema até a análise dos resultados, conforme representado na Figura 5. Essas etapas permitem guiar a criação de um modelo de Cadeia de Markov que irá gerar métricas para realizar análises de sobrevivência de estudantes de graduação.

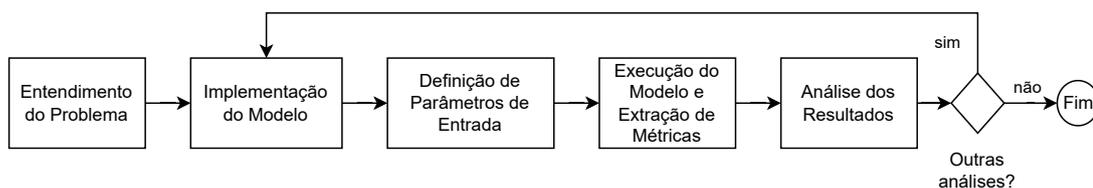


Figura 5 – Metodologia.

4.1 Entendimento do Problema

A etapa de entendimento do problema tem como objetivo entender um problema e planejar um modelo necessário para solucioná-lo. Neste trabalho, esta etapa compreendeu em entender o funcionamento da educação superior, onde foram identificados os problemas de evasão e retenção nos estudantes, principalmente nos primeiros semestres de curso. Para analisar os problemas, foi definido que os principais eventos a serem observados deveriam ser a evasão, conclusão e desvinculação, já que isso permite identificar a influência da evasão e retenção por semestre de curso e no resultado final dos estudantes.

Durante a pesquisa de trabalhos relacionados foram identificadas técnicas que permitiam uma boa análise da evasão, conclusão e desvinculação, onde se destacaram

a modelagem por Cadeia de Markov e a Análise de Sobrevivência, visto que permitem representar o progresso acadêmico e realizar análises de eventos por tempo. Também foram identificadas limitações, principalmente em relação à falta de experimentos para os primeiros semestres de cursos e a ausência de modelos que abordam a retenção como um fator para o progresso dos estudantes.

No planejamento para a implementação do modelo, os estados foram definidos com base no progresso acadêmico por semestre em que o estudante está vinculado ao curso, e se o estudante possui retenção ou não. Para a definição do intervalo de tempo a ser utilizado, o semestre foi considerado o período de tempo ideal, pois é ao final de cada semestre que o estudante pode ter alguma alteração no seu estado de vínculo com o curso.

4.2 Implementação do Modelo

O objetivo da etapa de implementação do modelo é a construção de um modelo capaz de representar o progresso acadêmico dos estudantes. Seguindo o planejamento apresentado na etapa anterior, foi definido um modelo de Cadeia de Markov absorvente, onde o progresso dos estudantes ocorre de acordo com o tempo de vínculo em um curso de graduação, levando em consideração a retenção. Para isso, foram definidos seus estados, suas transições e as probabilidades de transição. O modelo proposto é detalhado no Capítulo 5.

4.3 Definição de Parâmetros de Entrada

O objetivo desta etapa é definir os parâmetros de entrada necessários para o modelo. Esses parâmetros são os pesos do modelo, calculados a partir dos dados selecionados referentes ao progresso dos estudantes, onde é considerada a probabilidade de evadir, concluir ou progredir para o próximo semestre, tendo em vista também a possibilidade de retenção.

No modelo é necessário limitar a quantidade de semestres que serão representados. Essa limitação é calculada por meio do Princípio de Pareto, pois através dele é possível identificar até que tempo ocorre 95% das ocorrências dos eventos em relação a frequência.

Por meio do processamento de parâmetros do modelo existe a possibilidade de

realização de experimentos, através de variações das probabilidades de retenção e evasão, que podem ser passadas como parâmetros em diferentes semestres, ou incluindo parâmetros referentes a grupos de estudantes específicos, como curso, gênero ou cor/raça. Esses parâmetros, conforme são processados, permitem simular e avaliar diferentes aspectos do progresso dos estudantes durante um curso, o que possibilita também uma generalização do modelo para ser aplicado em diferentes contextos.

Os dados selecionados e analisados serão utilizados como parâmetros de entrada para o modelo, através das probabilidades de transição atribuídas por meio das probabilidades de retenção, evasão, conclusão e progresso dos estudantes de acordo com a duração de vínculo.

Para realizar as análises é necessário selecionar os dados, separá-los em grupos e calcular as probabilidades que serão utilizadas como parâmetros no modelo. Para cada grupo de estudantes analisados, é calculada a quantidade de estudantes que estavam matriculados, retidos, evadidos ou concluídos em cada semestre. Dessa forma, ao dividir pela quantidade total de estudantes em cada semestre, se obteve as probabilidades de progresso necessárias para cada estado da Cadeia de Markov.

4.4 Execução do Modelo e Extração de Métricas

Nesta etapa ocorre a execução do modelo e extração de suas métricas. O modelo é simulado conforme os parâmetros de entrada definidos para realizar os experimentos planejados. Neste trabalho, o modelo de Cadeia de Markov foi executado para realizar três estudos de caso, eles são: uma análise do impacto da retenção e evasão em cursos de computação; uma análise considerando características dos cursos e de estudantes; e uma análise da evasão e retenção para cursos das áreas de agrárias, computação e saúde. A execução do modelo é detalhada no Capítulo 6.

4.5 Análise dos Resultados

Na etapa de análise dos resultados, as métricas geradas pela execução do modelo são analisadas e aplicadas na Análise de Sobrevivência, com o objetivo de avaliar se estes resultados estão condizentes com a realidade e verificar como influenciam na evasão,

conclusão e desvinculação geral dos estudantes. A análise dos resultados é detalhada no Capítulo 6.

Os resultados da Cadeia de Markov são avaliados por meio de testes de hipótese e intervalo de confiança, para verificar se as previsões estão condizentes com os dados selecionados dos estudantes contemplados neste estudo, além de permitir comparar diferentes grupos de estudantes.

Os resultados do modelo são analisados pela Análise de Sobrevida para verificar o tempo e a probabilidade de evasão, conclusão e desvinculação dos estudantes de acordo com a quantidade de semestres em que estão matriculados em um curso. Dessa forma, os resultados obtidos na execução do modelo podem ser analisados e comparados.

As etapas descritas no processo de modelagem permitem organizar este trabalho por meio do entendimento do problema, implementação do modelo, definição de parâmetros de entrada, execução do modelo e a análise dos resultados, de modo a garantir os resultados a serem apresentados.

4.6 Considerações Finais

Neste capítulo, foi apresentada a metodologia para produzir um modelo de Cadeia de Markov absorvente e entregar as análises propostas. Foram descritas as etapas necessárias para criar um modelo capaz de gerar métricas relacionadas ao tempo e probabilidade de evasão, conclusão e desvinculação no ensino superior. No próximo capítulo será apresentado o modelo proposto.

5 Modelo Proposto

Neste capítulo será definido o modelo proposto neste trabalho. Isso inclui apresentar como foram definidos os parâmetros para a Cadeia de Markov e como será utilizada a Análise de Sobrevivência. O modelo descrito pode ser utilizado para representar o progresso acadêmico de qualquer curso de graduação, alterando apenas as probabilidades de transição dos estados. A vantagem de utilizar este modelo consiste na possibilidade de aferir métricas, criar um modelo generalizado e realizar experimentos através dos parâmetros do modelo.

O modelo proposto neste trabalho é definido por meio de uma Cadeia de Markov Absorvente de tempo discreto. Os estados do modelo foram definidos considerando os semestres em que o estudante está vinculado ao curso, com o objetivo de simular o progresso acadêmico. O modelo foi elaborado por meio da linguagem de programação Python, o código para a Cadeia de Markov está disponível no Apêndice A. A Cadeia de Markov foi produzida nessa linguagem pois permite a integração com outros trabalhos também realizados na linguagem Python.

O modelo utiliza de situações acadêmicas para determinar os estados do modelo. Entre as situações acadêmicas definidas como estados neste estudo, o estado retido é determinado como quando o estudante é reprovado em uma quantidade de disciplinas maior que a quantidade de um período ou caso ele tenha trancado em algum semestre. O estado evadido indica que o estudante que abandonou o curso, enquanto o estado concluído, representa o momento em que o estudante finaliza o curso com sucesso. Por fim, o estudante que se evadiu ou concluiu o curso é considerado como desvinculado.

O modelo é representado por uma versão simplificada na Figura 6. Ele é composto por estados de transição representados por Sn , onde n é a duração em períodos do vínculo e SxR demonstra o vínculo quando o estudante é considerado retido, onde x é a duração em períodos do vínculo, que pode ser superior aos estados que não apresentam retenção. Essa diferença entre os estados retidos e não retidos é necessária, visto que esses estudantes apresentam comportamentos distintos na progressão durante os períodos do curso de graduação. O *loop* em SxR é utilizado para definir um estado de limite de vínculo, onde os estados seguintes são representados dentro desse mesmo estado, quando o comportamento entre esses estados é muito semelhante e a quantidade de estudantes que chegam nesses estados é muito pequena. Entre os estados de absorção, o estado C representa o estudante

concluído, e o estado E representa o estudante evadido.

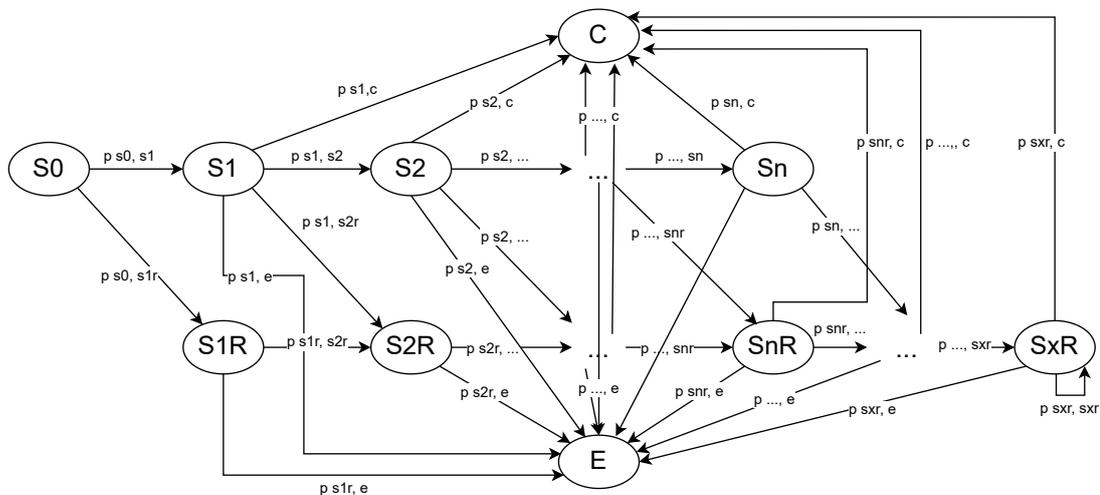


Figura 6 – Modelo de Cadeia de Markov para Representar o Progresso Acadêmico.

O modelo apresenta uma quantidade extra de semestres para os estudantes retidos representada por x , porém é necessário limitar essa quantidade de semestres para evitar distorções. Para isso, aplica-se uma censura, definida com o auxílio do Princípio de Pareto. Ele leva em consideração a frequência das ocorrências dos eventos, permitindo priorizá-los até um determinado tempo. Dessa forma, com uma confiança de 95% determina-se o estado de *self-loop* para cada experimento a ser realizado.

A transição pelos estados do modelo ocorre a partir do estado inicial representado pelo primeiro semestre do curso, onde o aluno pode realizar uma transição para o semestre seguinte, retido ou não, ou para o estado de evasão. A partir dos estados que representam o período padrão de conclusão do curso, também é possível atingir o estado de conclusão ou continuar progredindo pelos estados dos semestres seguintes.

As probabilidades de transição serão utilizadas para gerar a matriz de transição. Essas probabilidades que servem como peso do modelo proposto foram extraídas a partir de dados de estudantes coletados. Para isso, foram consideradas as probabilidades de retenção, evasão e conclusão de cada semestre por período ingressante, onde em cada estado existe uma probabilidade do conjunto de estudantes analisados de ir para o semestre seguinte sem retenção, ir para o semestre seguinte com retenção, evadir ou concluir.

Para experimentos, os parâmetros utilizados são a retenção e a evasão. Variações nas probabilidades desses parâmetros podem ser passadas em diferentes semestres. Esses parâmetros, conforme são alterados, permitem simular e avaliar diferentes aspectos do

progresso dos estudantes durante um curso. O processamento dos parâmetros para gerar as probabilidades da matriz de transição permite ao modelo de Cadeia de Markov se adaptar conforme as especificidades de grupos de estudantes considerando diversas categorias. Isso implica na possibilidade de avaliar e prever diferentes comportamentos baseados no efeito dessas possíveis alterações.

As simulações são realizadas de forma a representar a trajetória de 10000 estudantes, valor esse suficiente para que as métricas aferidas entrem em estabilidade, desde o primeiro semestre do curso até sua conclusão ou evasão, podendo passar por estados que refletem o comportamento de estudantes retidos. Ao final da simulação são geradas métricas referentes ao tempo e probabilidade de retenção, evasão, conclusão e desvinculação, que são validadas conforme o intervalo de confiança dos dados e teste Kolmogorov-Smirnov. Esse teste, além da validação dos dados, também é utilizado para avaliar a semelhança entre determinados grupos de estudantes. Isso permite analisar quais grupos apresentam a mesma distribuição considerando evasão, conclusão e desvinculação. Uma vez que, nesse teste, a hipótese nula é de que as amostras possuem a mesma distribuição. Portanto, se o p-valor for inferior a 0,05, a hipótese nula é rejeitada.

Como resultado do modelo, é gerado um vetor com os estados pelos quais cada uma das 10000 simulações passaram, assim é possível extrair informações relacionadas ao tempo e probabilidade desses estudantes. Esse vetor é aplicado na Análise de Sobrevida, sendo possível gerar seus gráficos através das informações de tempo até os estudantes atingirem os estados de evasão ou conclusão. Por fim, a Análise de Sobrevida é utilizada para avaliar a probabilidade em relação ao tempo de ocorrência dos eventos, adotando um intervalo de confiança de 95%.

5.1 Considerações Finais

Neste capítulo, foi apresentado o modelo de Cadeia de Markov absorvente que será utilizado nos estudos de caso. O modelo permite a análise de experimentos através de seus parâmetros, podendo observar diversos cenários. No próximo capítulo serão apresentados os estudos de caso.

6 Estudo de Caso

Neste capítulo, a partir das análises realizadas, verifica-se o comportamento dos eventos evasão, conclusão e desvinculação, relacionando a probabilidade de ocorrer esses eventos com o tempo de vínculo dos estudantes. Por meio dos resultados do modelo de Cadeia de Markov é realizada a Análise de Sobrevivência. Portanto, é possível realizar experimentos alterando-se parâmetros do modelo e observando o impacto sobre métricas relacionadas à evasão, à conclusão e ao vínculo dos estudantes. Os experimentos propostos neste trabalho foram validados pelos dados coletados.

Serão apresentados três estudos de casos, que utilizaram dados coletados dos estudantes da UFRPE, para definir as probabilidades de transição do modelo. Na Seção 6.1 é analisado o impacto ao reduzir a retenção e evasão nos primeiros períodos em relação a evasão, conclusão e desvinculação geral. Na Seção 6.2 é realizada a Análise de Sobrevivência considerando características, como curso, gênero e cor/raça. Na Seção 6.3 é analisada a evasão para os quatro primeiros semestres de cursos das áreas de agrárias, computação e saúde, além da evasão, conclusão e desvinculação por semestre em que o estudante ficou retido.

6.1 Estudo de Caso 1: Impacto da Retenção e Evasão para Cursos de Computação

No modelo de Cadeia de Markov proposto no Capítulo 5, o progresso acadêmico do estudante é representado por semestres, através dos estados e da atribuição de probabilidades de transições entre os estados. Neste primeiro estudo de caso são analisados dois cursos da área de computação para avaliar a evasão, conclusão e desvinculação dos estudantes, além de realizar análises para verificar os resultados ao diminuir a evasão ou a retenção durante os primeiros semestres de curso.

No modelo, a *matriz de transição*¹ é produzida com os dados coletados dos dois cursos da área de computação e considera as probabilidades dos estudantes progredirem de um semestre para o próximo semestre sem retenção, para o próximo semestre com retenção, para a evasão ou para a conclusão. O processamento das probabilidades na matriz de transição para a evasão e a retenção é o que permite os experimentos que serão

apresentados.

Nas simulações o tempo de observação se limitou até o vigésimo semestre. Esse tempo foi definido pois é o tempo limite máximo para a conclusão dos cursos, o que permite analisar todo o tempo em que o estudante está vinculado ao curso. O resultado das simulações permite coletar métricas relacionadas ao tempo e a probabilidade. Esses resultados aplicados na Análise de Sobrevida permitem analisar o comportamento a evasão, conclusão e desvinculação não apenas ao final do curso, mas durante todo o progresso dos estudantes.

6.1.1 Análise de Sobrevida nos Cursos de Computação

A Figura 7(a) apresenta a Análise de Sobrevida dos eventos evasão, conclusão e desvinculação, representando o progresso estudantil até o vigésimo semestre. Nesta análise, que será utilizada como cenário de referência, observa-se que a maior parte dos estudantes se desvinculam antes mesmo de algum deles concluir o curso. No quinto semestre, cerca de 50% dos estudantes não estão mais vinculados ao curso e após o décimo sexto período menos de 10% ainda mantém o vínculo. Em relação à conclusão, essa começa a ocorrer a partir do oitavo semestre, enquanto no décimo sexto semestre cerca de 80% dos estudantes que irão graduar, já o terão feito. Além disso, a maior probabilidade de evasão ocorre nos primeiros semestres, com aproximadamente 63% dos estudantes que irão evadir, evadindo até o final do quarto semestre, valor esse muito superior por semestre se comparado com os períodos seguintes.

Na Figura 7(b), é possível analisar as probabilidades de evasão, conclusão e vínculo em relação ao total de estudantes. Até o décimo quinto semestre 20% do total de estudantes já estão graduados, com acréscimo de apenas 4% até o vigésimo semestre. Além disso, 45% dos estudantes evadem até o final do quarto semestre. Dessa forma, propor formas de diminuir a evasão nos primeiros semestres pode impactar positivamente em todo o progresso estudantil.

¹ <https://github.com/JulianaFerreira/educacao/blob/6c3e469711d275a5967fed80c642585d29417a33/matrizes-de-transicao/comp-ate-2013.csv>

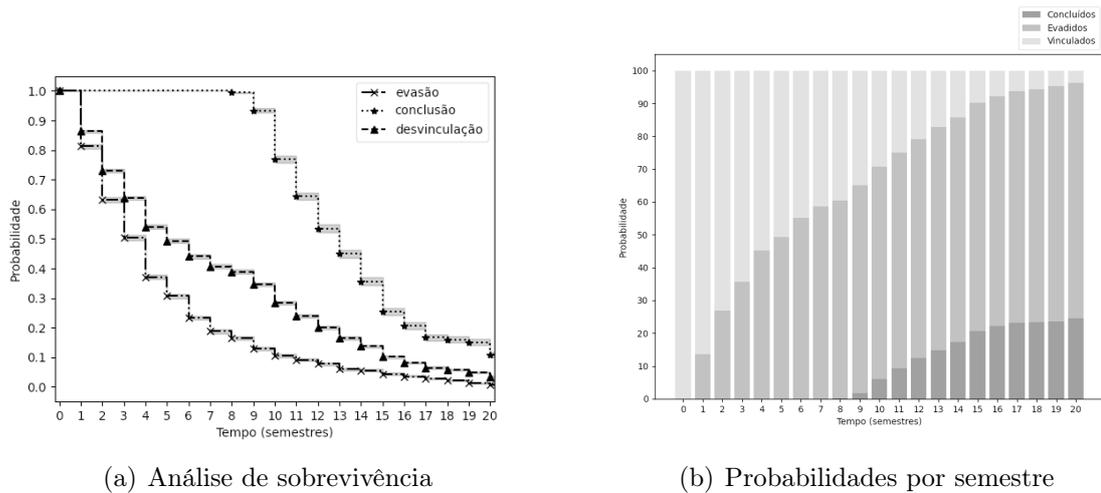


Figura 7 – Análise de Sobrevivência da Evasão, Conclusão e Desvinculação e Probabilidades por Semestre para Concluídos, Evadidos e Vinculados.

6.1.2 Impacto da Retenção no Primeiro Semestre

O modelo proposto permite o processamento dos parâmetros para avaliar as consequências da evasão e retenção no progresso de todo o curso. Ao verificar a evasão na Análise de Sobrevivência do cenário de referência, observa-se que no primeiro período a probabilidade de evasão já representa 20% do total de estudantes evadidos. Considerando esse dado, é importante analisar o impacto ao diminuir a retenção apenas no primeiro semestre na evasão geral e por período. Nas Figuras 8(a), 8(b) e 8(c), é analisada a diminuição da retenção em 25%, 50% e 75% apenas no primeiro período de curso, considerando a evasão, conclusão e desvinculação. Os resultados apresentam uma redução da probabilidade de evasão e desvinculação por período, entre os períodos de 2 até 13, ao diminuir a retenção. Porém, a conclusão não apresenta diferença, já que uma alteração da retenção apenas no primeiro semestre não apresenta influência o suficiente na conclusão.

Conforme a Tabela 4, observa-se que a evasão geral diminuiu em um pouco mais de 1% para cada 25% de redução da evasão no primeiro período. De acordo com a Tabela 5, os tempos de retenção, evasão e desvinculação aumentam um pouco a cada diminuição da retenção no primeiro período. Enquanto para a conclusão, observa-se um pequeno aumento.

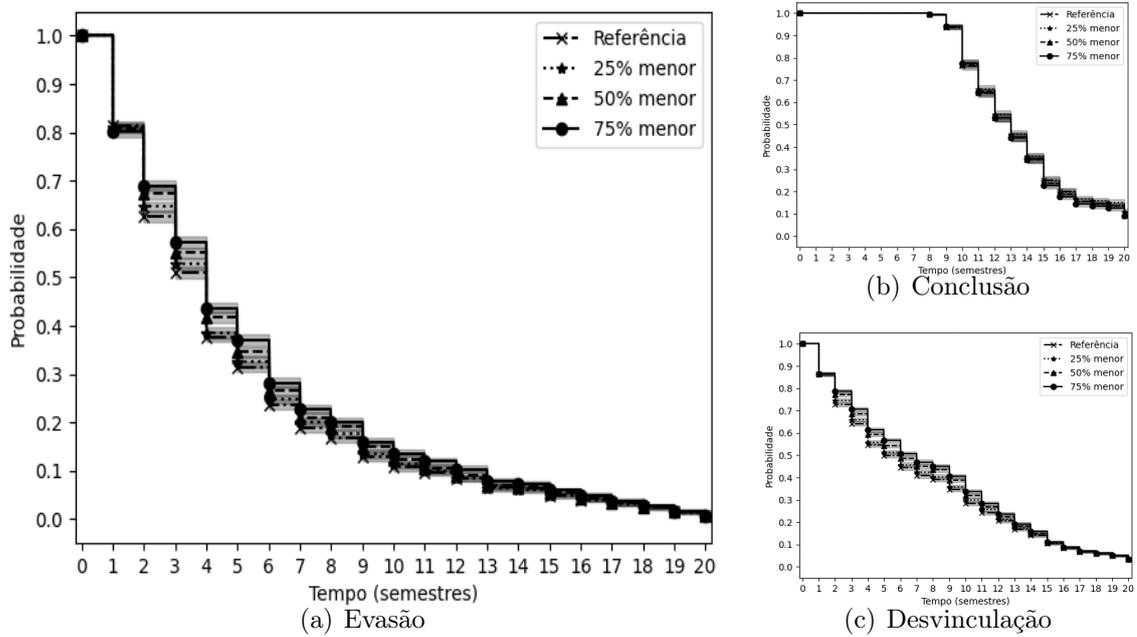


Figura 8 – Análise de Sobrevivência ao Diminuir a Retenção em 25%, 50% e 75% no Primeiro Semestre.

Tabela 4 – Resultados das Probabilidades da Redução da Retenção no Primeiro Semestre.

	Probabilidade (%)			
	Referência	25% menor	50% menor	75% menor
Retenção	70.35	72.91	76.16	79.89
Evasão	71.60	71.14	70.57	70.36
Conclusão	24.35	25.21	25.34	25.15
Desvinculação	95.95	96.35	95.91	95.51

Tabela 5 – Resultados dos Tempos da Redução da Retenção no Primeiro Semestre.

	Tempo (semestres)			
	Referência	25% menor	50% menor	75% menor
Retenção	3.26	3.21	3.16	3.10
Evasão	4.68	4.83	5.15	5.24
Conclusão	12.94	12.80	12.89	12.84
Desvinculação	6.78	6.91	7.20	7.24

6.1.3 Impacto da Retenção e Evasão nos Quatro Primeiros Semestres

Considerando que durante os quatro primeiros semestres a probabilidade de evadir é maior, no experimento a seguir, a proposta é alterar a probabilidade de evasão para 50% em relação ao cenário de referência nesses semestres. Pela Figura 9(a), é possível constatar uma diminuição da probabilidade de evasão nos semestres iniciais que impacta em cerca de 7% na evasão final, além de apenas 40% dos estudantes terem evadido até o quarto período. Conseqüentemente, essas alterações influenciam na probabilidade final da graduação e no vínculo dos estudantes conforme a probabilidade de evasão. Isso constata que diminuir em 50% a evasão apenas nos quatro primeiro períodos já apresenta um impacto positivo suficiente para modificar a evasão total. Portanto, isso permite planejar ações direcionadas para atender esse grupo de estudantes.

A retenção também pode ser considerada um fator importante na influência no vínculo do estudante. No experimento apresentado a seguir, a proposta é alterar a probabilidade de retenção para 50% do cenário de referência nos quatro primeiros períodos. Pela Figura 9(b), observa-se uma influência maior da retenção na evasão, graduação e vínculo se comparado com o experimento anterior da probabilidade de evasão. Isso acontece devido ao comportamento dos alunos retidos, que apresentam maior probabilidade de evasão. Apesar da evasão ainda se manter elevada nos dois primeiros períodos de curso, ao longo dos semestres seguintes ela diminui e apresenta uma redução final de aproximadamente 12%. Isso representa uma diferença de 5% em comparação com o experimento anterior.

Uma alternativa aos experimentos anteriores é alterar a probabilidade de ambos, retenção e evasão, para 50% do cenário de referência nos quatro primeiros períodos. Dessa forma, conforme observado na Figura 9(c), isso implica em uma diminuição ainda maior da probabilidade de evasão por semestre, com uma menor probabilidade de desvinculação nos primeiros períodos.

Os resultados disponíveis na Tabela 6, demonstram como a alteração das probabilidades apenas nos primeiros quatro períodos pode influenciar nas probabilidades finais de retenção, evasão, graduação e vínculo. Ao diminuir a evasão, houve um aumento da retenção, enquanto em todos os experimentos existiu uma maior probabilidade de graduação e diminuição da evasão. Porém, essas alterações tiveram pouco impacto na probabilidade de vínculo.

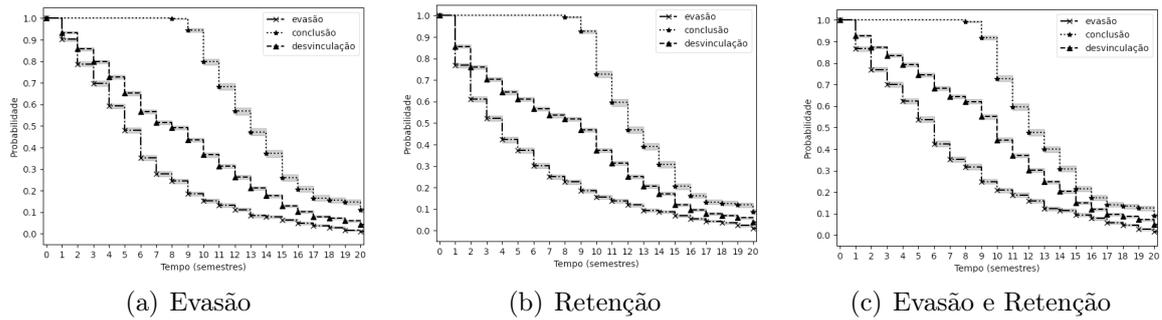


Figura 9 – Análise de Sobrevivência da Evasão, Conclusão e Desvinculação para uma Redução de 50% na Evasão ou Retenção nos Quatro Primeiros Semestres.

Tabela 6 – Resultados das Probabilidades dos Experimentos para os Quatro Primeiros Semestres de Curso.

Probabilidade (%)				
	Referência	50% da Evasão	50% da Retenção	50% da Retenção e Evasão
Retenção	70,80	79,40	60,50	68,80
Evasão	72,40	65,60	60,20	53,40
Graduação	24,10	29,40	34,80	40,60
Desvinculação	96,50	95,00	95,00	94,00

Na Tabela 7, observa-se os tempos por semestre ao realizar os experimentos nos quatro primeiros períodos. Constata-se, uma diminuição do tempo para graduação ao reduzir a retenção, assim como um aumento do tempo para desvinculação ao diminuir tanto a retenção, quanto a evasão. Dessa forma, é possível criar um planejamento melhor elaborado para organizar vagas extras nas turmas, com base na previsão de como esses fatores podem influenciar no tempo de vínculo do estudante.

As simulações realizadas para analisar a evasão e retenção nos primeiros períodos comprovam a importância de ações nos primeiros períodos de curso. Percebe-se que as ações nos estudantes até o quarto semestre de curso têm maior impacto, em um menor intervalo de tempo para diminuir a evasão final dos estudantes. Além de contribuir ao diminuir a retenção e permitir melhor gerenciamento das vagas remanescentes.

Tabela 7 – Resultados dos Tempos dos Experimentos para os Quatro Primeiros Semestres de Curso.

	Tempo (semestres)			
	Referência	50% da Evasão	50% da Retenção	50% da Retenção e Evasão
Retenção	3,20	3,15	4,72	4,63
Evasão	4,74	6,06	5,18	6,63
Graduação	13,00	13,07	12,53	12,62
Desvinculação	6,81	8,23	7,87	9,22

6.2 Estudo de Caso 2: Análise de Sobrevivência Considerando Características dos Cursos e Estudantes

Neste estudo de caso são analisadas características dos cursos e estudantes. A análise dessa forma permite maior precisão para a previsão dos resultados, dado que existem padrões em grupos que evidenciam comportamentos específicos. As matrizes de transição foram agrupadas conforme categorias, onde os experimentos utilizam os parâmetros conforme os grupos de estudantes selecionados. Neste estudo de caso, serão analisadas as categorias curso, gênero e cor/raça, observando as probabilidades e tempos de retenção, conclusão, evasão e desvinculação.

6.2.1 Curso

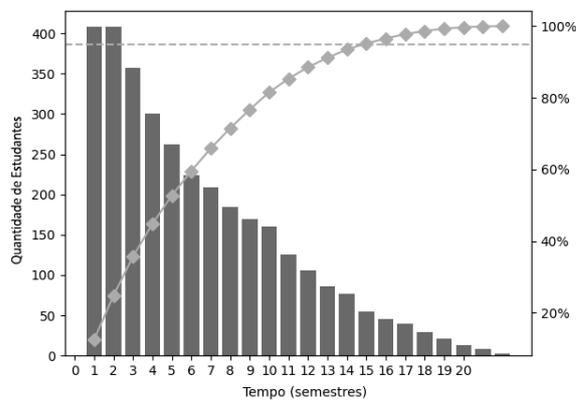
A primeira categoria a ser analisada é o curso, nela é verificado como cada curso se comporta individualmente. Considerando os dados, foram calculados os intervalos de confiança para cada curso, conforme pode ser observado na Tabela 8.

O primeiro passo para realizar a análise foi identificar qual o melhor tempo de censura para cada curso, isso permite priorizar os eventos em até um determinado tempo. Para isso, foi utilizado o Princípio de Pareto, conforme as Figuras 10(a) e 10(b), considerando 95% de confiança, onde o tempo de censura para o Curso 1 foi definido no tempo 13 e Curso 2 no tempo 15.

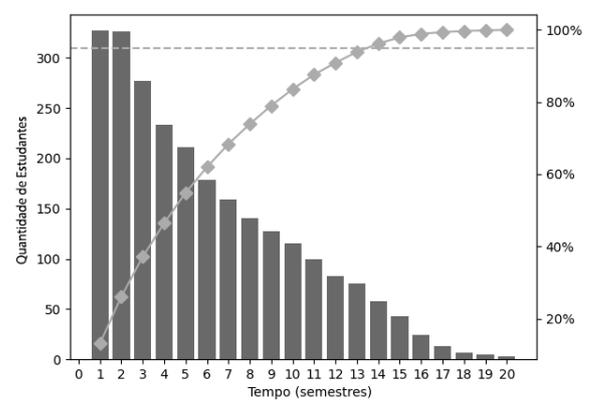
Para validação do modelo, além do intervalo de confiança, observou-se o

Tabela 8 – Intervalo de Confiança da Categoria Curso.

	Tempo (semestres)			
	Curso 1		Curso 2	
	média	95% IC	média	95% IC
Evasão	4,9271	[4,4681;5,3861]	4,2629	[3,8088;4,7170]
Conclusão	12,9285	[12,2634;13,5936]	12,5238	[12,0046;13,0429]
Desvinculação	6,6683	[6,1601;7,1766]	6,4588	[5,9176;7,0000]



(a) Curso 1



(b) Curso 2

Figura 10 – Gráfico de Pareto para a Categoria Curso.

comportamento da Análise de Sobrevivência do modelo em relação aos dados para a probabilidade e tempo. Nas Figuras 11(a) e 12(a), observa-se que tanto para o Curso 1 quanto para o Curso 2 a Análise de Sobrevivência da evasão apresenta um valor muito similar, enquanto para a conclusão e desvinculação a probabilidade em relação ao tempo se comporta de forma equivalente até o tempo 15, a partir desse tempo existe uma diferença provocada pelo tempo de censura.

O teste de hipótese Kolmogorov-Smirnov foi realizado para identificar se a distribuição do modelo é equivalente aos dados. Para o Curso 1, o p-valor da evasão, conclusão e desvinculação foi, respectivamente, 0,9994, 0,9556 e 0,9974. Enquanto, para o Curso 2, o p-valor da evasão, conclusão e desvinculação foi, respectivamente, 0,9999, 0,5809 e 0,9693. Em todos os testes, a hipótese nula de que a distribuição das duas amostras são equivalentes foi aceita, apesar da conclusão apresentar uma diferença maior do que a evasão e a desvinculação no Curso 2.

As simulações para os dois cursos analisados são realizadas individualmente, através

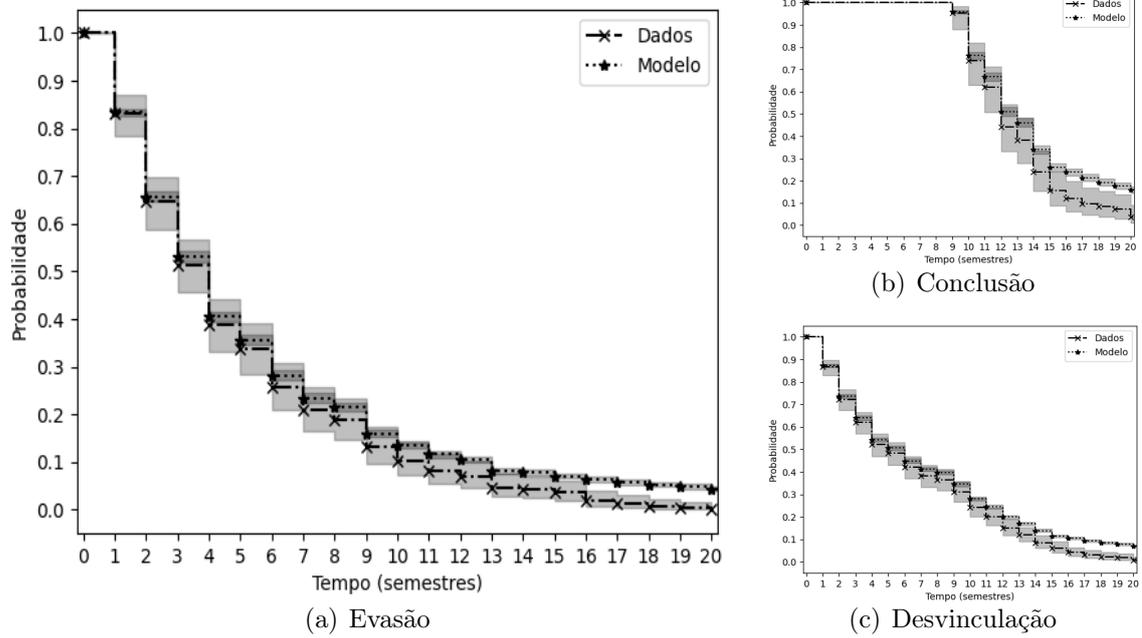


Figura 11 – Análise de Sobrevivência do Modelo em Relação aos Dados para o Curso 1.

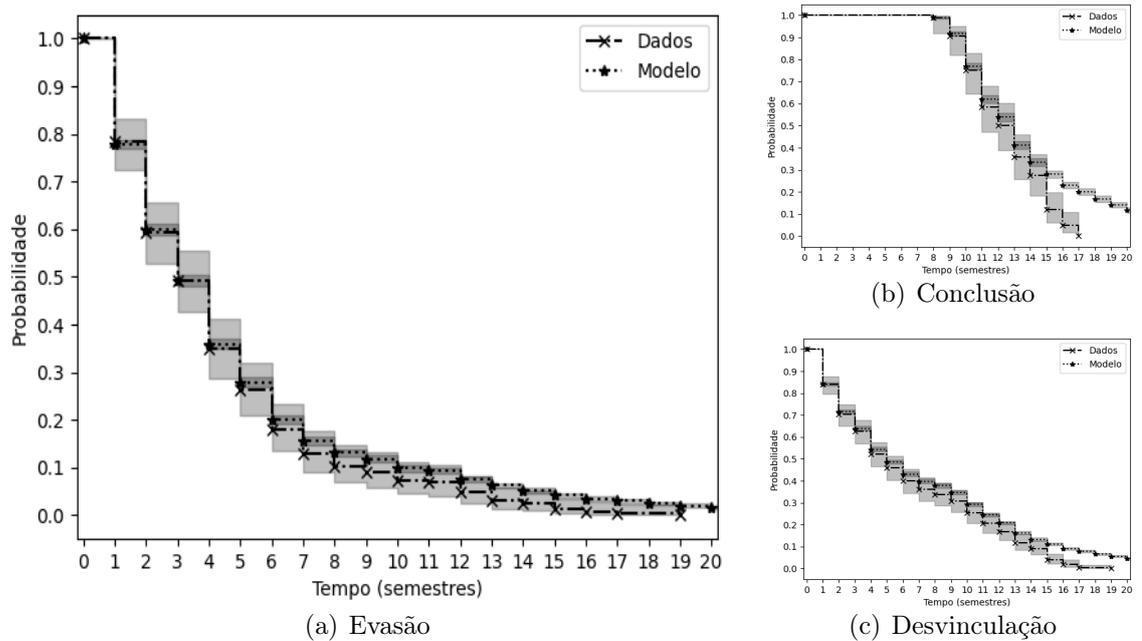


Figura 12 – Análise de Sobrevivência do Modelo em Relação aos Dados para o Curso 2.

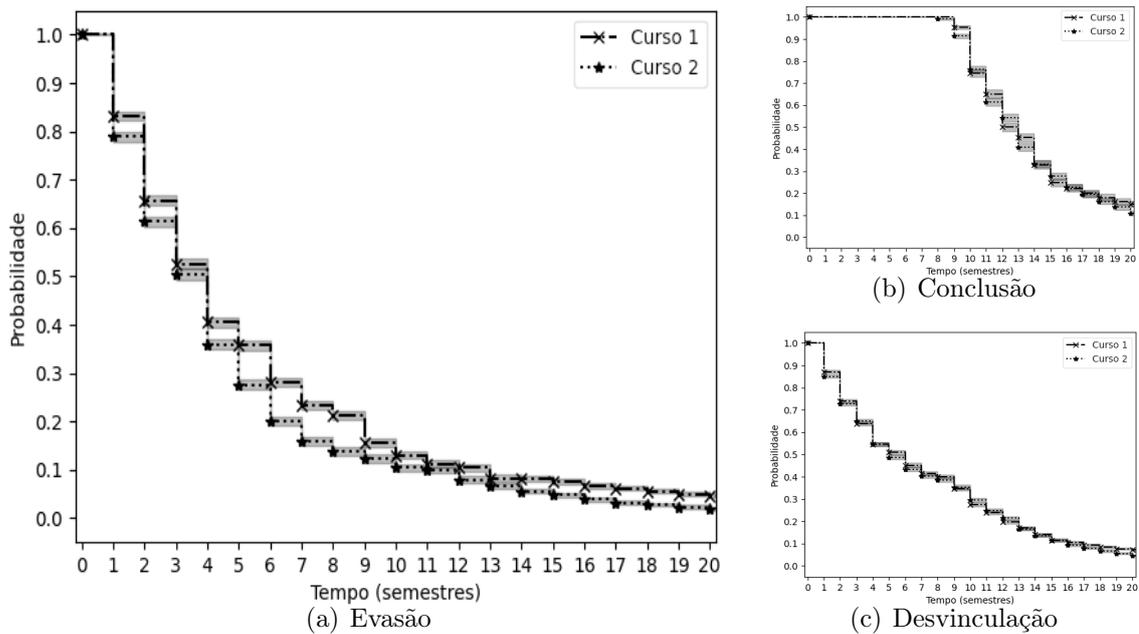


Figura 13 – Análise de Sobrevivência para a Categoria Curso.

da aplicação de suas *matrizes de transição*² no modelo. Então, o resultado do modelo é aplicado na Análise de Sobrevivência para verificar a diferença de comportamento das probabilidades de evasão, conclusão e desvinculação de acordo com o tempo em que os estudantes estão vinculados aos cursos. Ao observar as Figuras 13(a), 13(b) e 13(c), é evidente que o Curso 2 tem evasão maior por semestre, principalmente nos primeiros períodos, situação que se inverte depois do período 10. O que indica que uma diminuição da evasão nos primeiros semestres daria um impacto maior para o Curso 2. Apesar disso, a probabilidade em relação ao tempo por semestre para a conclusão apresenta comportamento semelhante nos dois cursos.

Em relação as probabilidades finais de retenção, evasão, conclusão e desvinculação, ao verificar na Tabela 9, observa-se que a probabilidade final de retenção do Curso 2 é menor do que do Curso 1, o que influencia na probabilidade de conclusão, onde o Curso 2 é maior do que o Curso 1. Apesar disso, o tempo para conclusão do curso é maior no Curso 2, enquanto a desvinculação é muito semelhante, por consequência do tempo maior para retenção, o que faz com que o tempo para evasão seja menor.

Ao realizar o teste de hipótese Kolmogorov-Smirnov para comparar os cursos, o resultado do p-valor para a evasão é de 1,0, para a conclusão é de 0,9972 e para a

² <https://github.com/JulianaFerreira/educacao/blob/6c3e469711d275a5967fed80c642585d29417a33/matrizes-de-transicao/curso1-ate-2013.csv>

² <https://github.com/JulianaFerreira/educacao/blob/6c3e469711d275a5967fed80c642585d29417a33/matrizes-de-transicao/curso2-ate-2013.csv>

Tabela 9 – Resultados das Probabilidades e dos Tempos para a Categoria Curso.

	Probabilidade (%)		Tempo (semestres)	
	Curso 1	Curso 2	Curso 1	Curso 2
Retenção	73,11	66,11	3,16	3,45
Evasão	72,40	69,40	4,79	4,25
Conclusão	20,63	25,62	12,52	12,76
Desvinculação	93,03	95,02	6,51	6,55

desvinculação é de 0,0008. Os resultados demonstram que os dois cursos apresentam uma distribuição equivalente para a evasão e conclusão, porém diferente para a desvinculação.

6.2.2 Gênero

A segunda categoria analisada é gênero, onde é avaliada a diferença de comportamento entre o gênero feminino e masculino. Considerando os dados, foram calculados os intervalos de confiança para cada gênero, conforme pode ser observado na Tabela 10. Ao utilizar o Princípio de Pareto, conforme as Figuras 14(a) e 14(b), os tempos de censura foram definidos para ambos os gêneros no tempo 14 utilizando um intervalo de confiança de 95%.

Tabela 10 – Intervalo de Confiança da Categoria Gênero.

	Tempo (semestres)			
	Feminino		Masculino	
	média	95% IC	média	95% IC
Evasão	4,2795	[3,4589;5,1002]	4,7142	[4,3580;5,0704]
Conclusão	12,3783	[11,5416;13,2150]	12,8244	[12,3381;13,3107]
Desvinculação	6,5846	[5,6906;7,4785]	6,5716	[6,1643;6,9789]

Para validação do modelo além do intervalo de confiança, foram observados através da Análise de Sobrevivência como se comporta o modelo em relação à probabilidade e tempo nos dados. Nas Figuras 15(a), 15(b) e 15(c), observa-se que a Análise de Sobrevivência para a evasão, conclusão e desvinculação do gênero Feminino apresenta comportamento

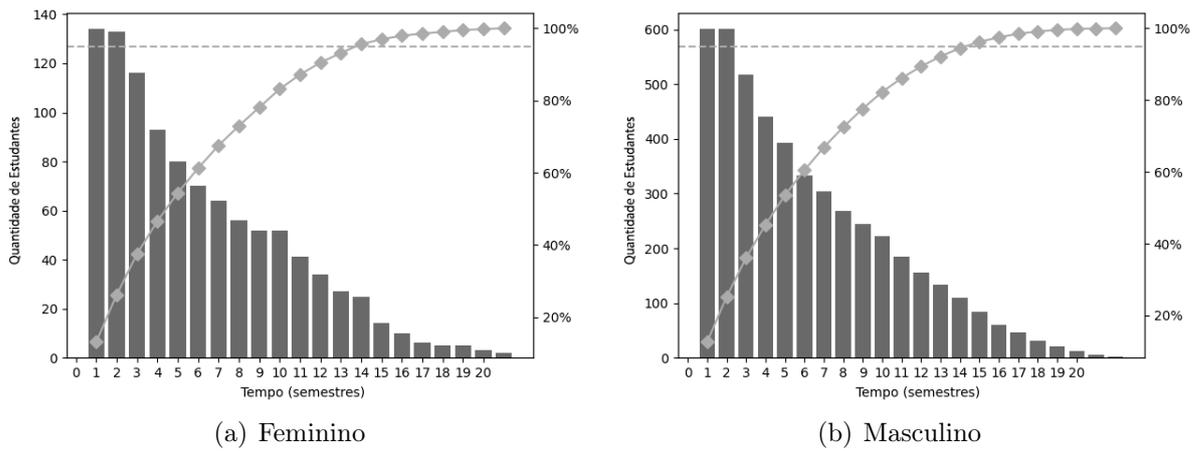


Figura 14 – Gráfico de Pareto para a Categoria Sexo.

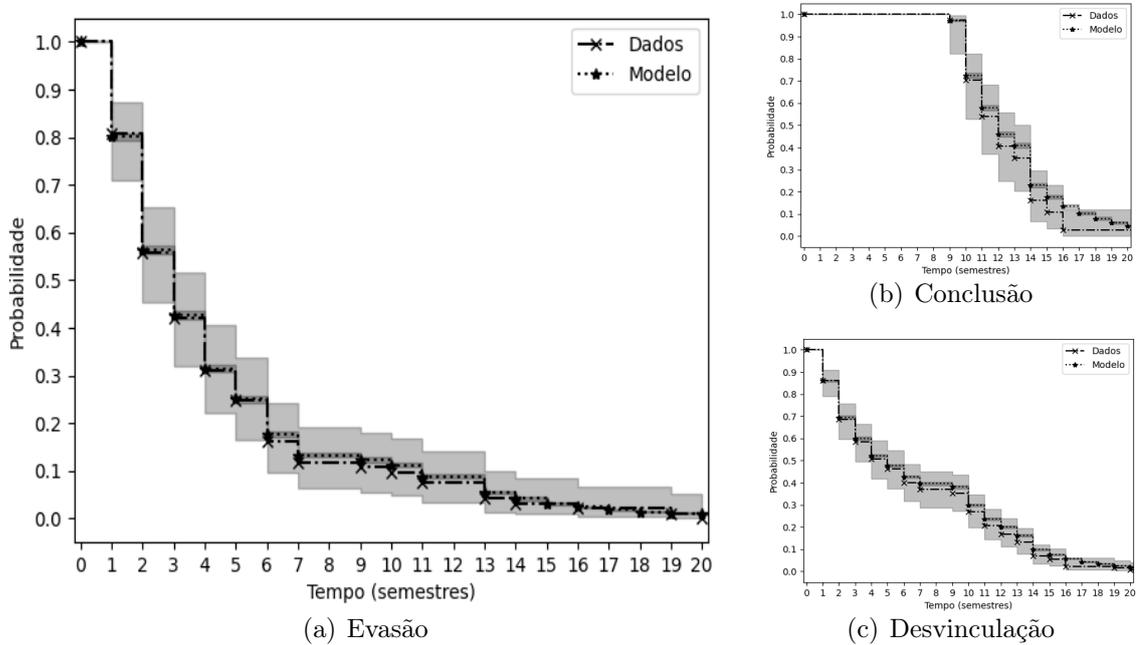


Figura 15 – Análise de Sobrevivência do Modelo em Relação aos Dados para o Gênero Feminino.

dentro do intervalo de confiança de 95% durante todo o tempo analisado.

Na análise do gênero Masculino, disponível nas Figuras 16(a), 16(b) e 16(c), a evasão apresenta comportamento dentro do intervalo de confiança durante todo o tempo analisado. Porém, ao analisar a desvinculação e conclusão observa-se que o modelo é equivalente aos dados apenas até o tempo 12.

O teste de hipótese Kolmogorov-Smirnov foi realizado para identificar se a distribuição do modelo é equivalente aos dados. Para o gênero Feminino, o p-valor da evasão, conclusão e desvinculação foi, respectivamente, 0,9999, 0,9934 e 0,9999. Enquanto, para o gênero Masculino, o p-valor da evasão, conclusão e desvinculação foi, respectivamente,

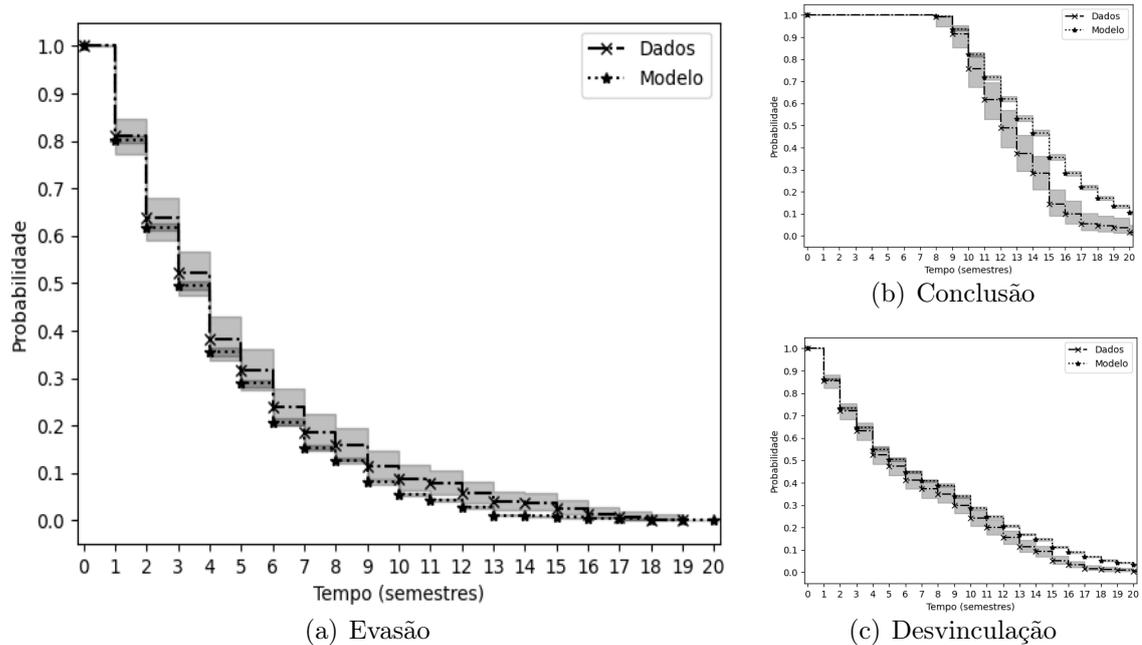


Figura 16 – Análise de Sobrevivência do Modelo em Relação aos Dados para o Gênero Masculino.

0,7957, 0,1710 e 0,8159. Em todos os testes a hipótese nula de que a distribuição das duas amostras são equivalentes foi aceita, apesar da conclusão apresentar uma diferença maior do que a evasão e a desvinculação para o gênero masculino.

As simulações dos gêneros são realizadas, através da aplicação de suas *matrizes de transição*³ no modelo. Então, o resultado do modelo para cada gênero é aplicado na Análise de Sobrevivência para verificar a diferença de comportamento das probabilidades de evasão, conclusão e desvinculação de acordo com o tempo em que os estudantes estão vinculados aos cursos. Nos experimento para categoria gênero, a Análise de Sobrevivência disponível nas Figuras 17(a), 17(b) e 17(c), apresenta a probabilidade em relação ao tempo para evasão, conclusão e desvinculação. Em relação a probabilidade de evasão, o gênero feminino é maior nos primeiros períodos, porém a partir do nono semestres o gênero masculino agora apresenta maior probabilidade de evasão. Em relação a conclusão, o gênero feminino apresenta uma probabilidade de conclusão maior por semestre, o que resulta em uma probabilidade maior de conclusão final e menos tempo necessário para concluir o curso.

Ao analisar a Tabela 11, as probabilidades finais demonstram que o gênero

³ <https://github.com/JulianaFerreira/educacao/blob/6c3e469711d275a5967fed80c642585d29417a33/matrizes-de-transicao/comp-sexo-f.csv>

³ <https://github.com/JulianaFerreira/educacao/blob/6c3e469711d275a5967fed80c642585d29417a33/matrizes-de-transicao/comp-sexo-m.csv>

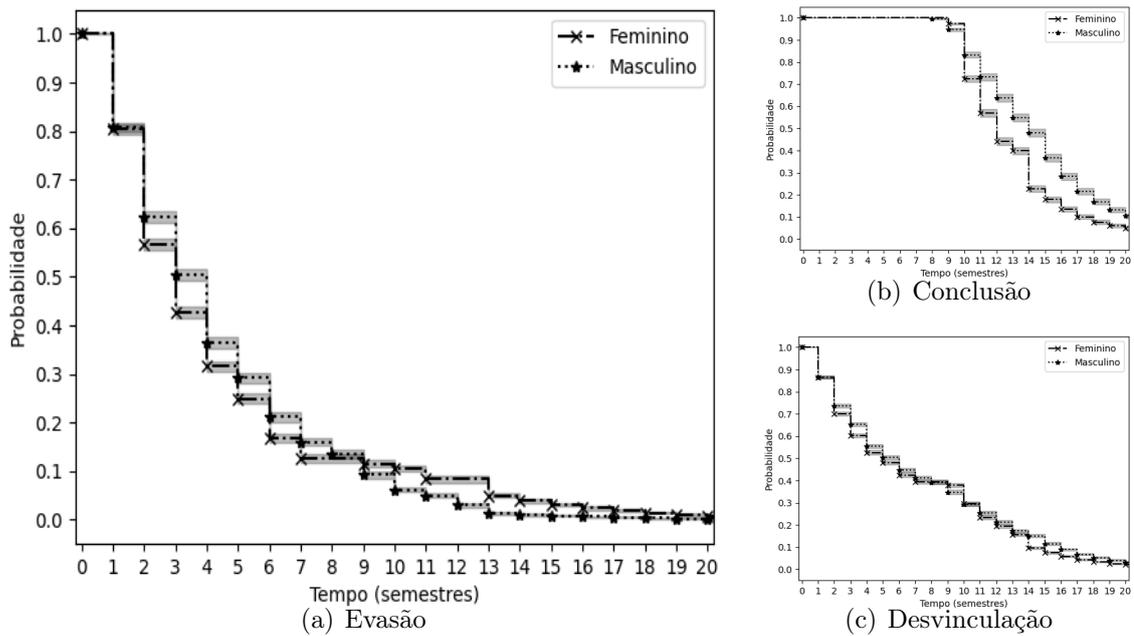


Figura 17 – Análise de Sobrevivência para a Categoria Gênero.

masculino apresenta maior chance de retenção e evasão, enquanto o gênero feminino maior probabilidade de conclusão. Em relação ao tempo, o gênero feminino apresenta menor tempo em todos os aspectos observados.

Tabela 11 – Resultados das Probabilidades e dos Tempos para a Categoria Gênero.

	Probabilidade (%)		Tempo (semestres)	
	F	M	F	M
Retenção	59,03	73,36	3,09	3,27
Evasão	68,19	69,40	4,17	4,30
Conclusão	29,84	27,29	12,55	13,69
Desvinculação	98,03	96,69	6,72	6,95

Ao realizar o teste de hipótese Kolmogorov-Smirnov para comparar os gêneros, o resultado do p-valor para a evasão é de 1,0, para a conclusão é de 0,9999 e para a desvinculação é inferior a 0,05. Os resultados demonstram que os dois cursos apresentam uma distribuição equivalente para a evasão e conclusão, porém diferente para a desvinculação.

6.2.3 Cor/Raça

A categoria cor/raça, avalia a diferença de comportamento entre as três raças com maior quantidade de estudantes. Considerando os dados, foram calculados os intervalos de confiança para cada cor/raça, conforme pode ser observado na Tabela 12, não foi possível criar um intervalo de confiança com a cor preta, pois existem apenas dezoito estudantes no curso que se auto identificaram com essa categoria, quantidade insuficiente para gerar um intervalo de confiança.

Tabela 12 – Intervalo de Confiança da Categoria Cor/Raça.

	Tempo (semestres)			
	Branca		Parda	
	média	95% IC	média	95% IC
Evasão	5,3166	[4,3889;6,2444]	4,7213	[4,0605;5,3820]
Conclusão	12,6153	[11,3845;13,8462]	11,4838	[10,8637;12,1039]
Desvinculação	7,5232	[6,4938;8,5526]	6,0915	[5,3985;6,7844]

Ao traçar o Princípio de Pareto, conforme as Figuras 18(a), 18(b) e 18(c), com 95% de confiança, identificou-se o melhor tempo de censura para cada cor/raça. A cor branca apresenta o tempo de censura em 15, parda em 13 e preta em 12. Porém a cor preta não converge para o intervalo de confiança, já que a quantidade de amostra é muito inferior ao necessário para criar parâmetros confiáveis para o modelo.

Para validação do modelo, além do intervalo de confiança, foram observados através da Análise de Sobrevivência como se comporta o modelo em relação à probabilidade e tempo nos dados. Nas Figuras 19(a), 19(b) e 19(c) observa-se que para a cor Parda, a Análise de Sobrevivência da desvinculação e evasão apresentam comportamento dentro do intervalo de confiança. Porém para a conclusão os dados apresentam dados apenas até o tempo 15, isso acontece por conta de uma limitação dos dados.

Nas Figuras 20(a), 20(b) e 20(c), observa-se que para a cor Branca, a Análise de Sobrevivência da evasão apresenta comportamento semelhante ao modelo em relação aos dados. Para a conclusão e desvinculação, os dados estão dentro do intervalo de confiança até o tempo 13.

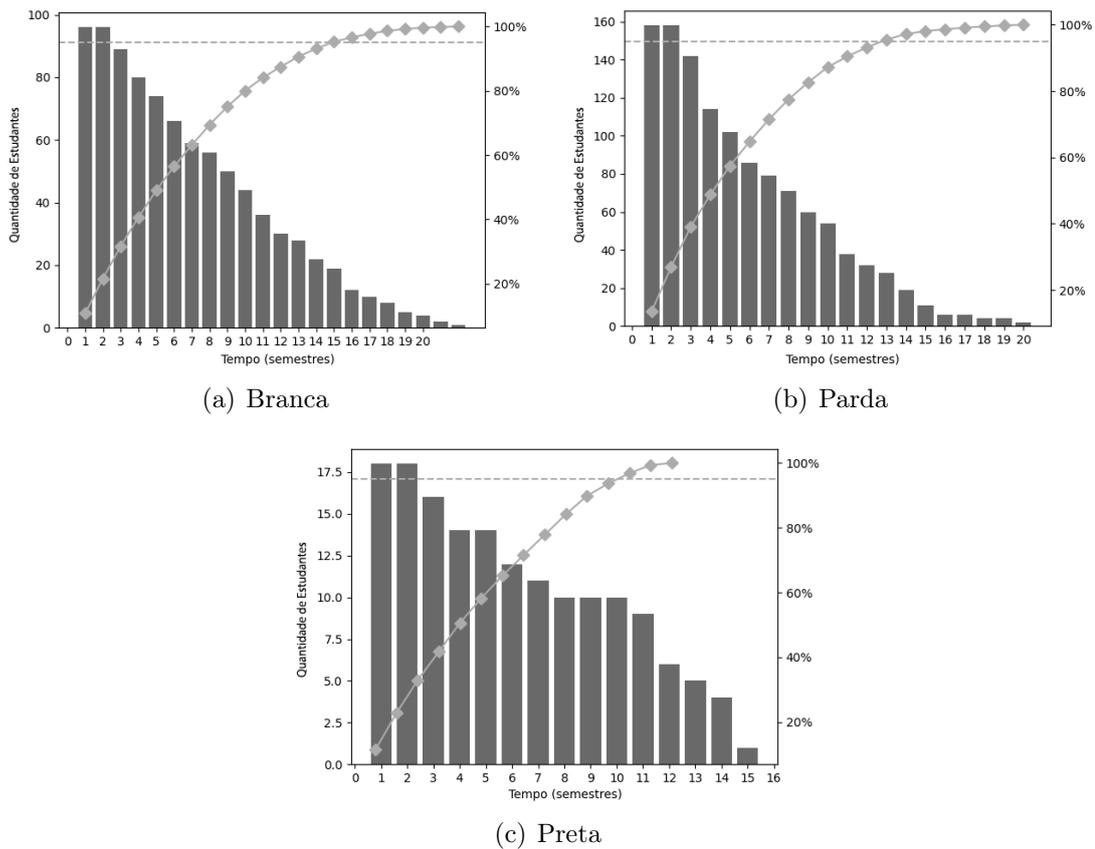
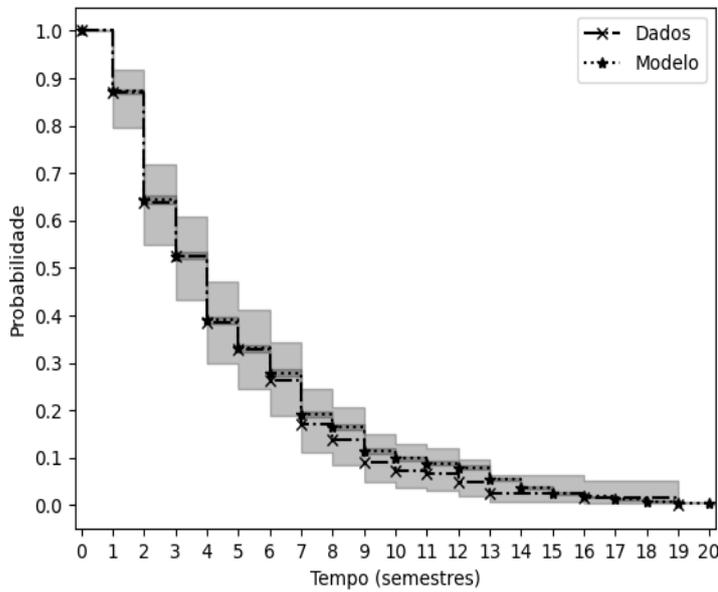


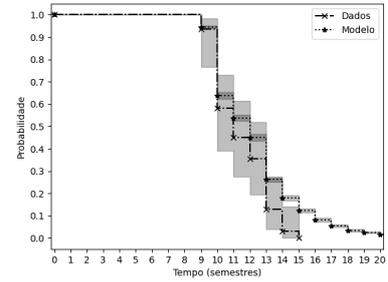
Figura 18 – Gráfico de Pareto para a Categoria Cor/Raça.

Nas Figuras 21(a), 21(b) e 21(c), observa-se que para a cor Preta, a Análise de Sobrevivência da evasão, conclusão e desvinculação apresentam comportamento dentro do intervalo de confiança, exceto pela evasão após o tempo 12, influenciado pelo tempo de censura, o que também impacta na desvinculação. Também observa-se que o intervalo de confiança desse grupo é maior que os outros, isso acontece devido a pequena quantidade de amostra disponível, o que implica em dados muito específicos e que podem não prever com tanta precisão.

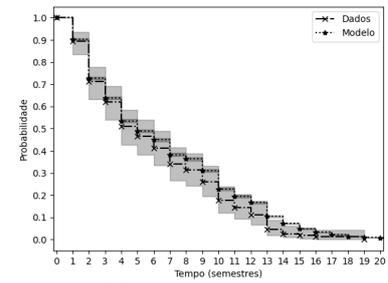
O teste de hipótese Kolmogorov-Smirnov foi realizado para identificar se a distribuição do modelo é equivalente aos dados. Para a cor Branca, o p-valor da evasão, conclusão e desvinculação foi, respectivamente, 0,9999, 0,9964 e 0,9999. Na cor Parda, o p-valor da evasão, conclusão e desvinculação foi, respectivamente, 0,9999, 0,6198 e 0,8117. Na cor Preta, o p-valor da evasão, conclusão e desvinculação foi, respectivamente, 0,5571, 0,9619 e 0,9742. Em todos os testes de hipótese, a hipótese nula de que a distribuição das duas amostras são equivalentes foi aceita, apesar da conclusão apresentar uma diferença maior para a cor parda e a evasão maior diferença para a cor preta.



(a) Evasão

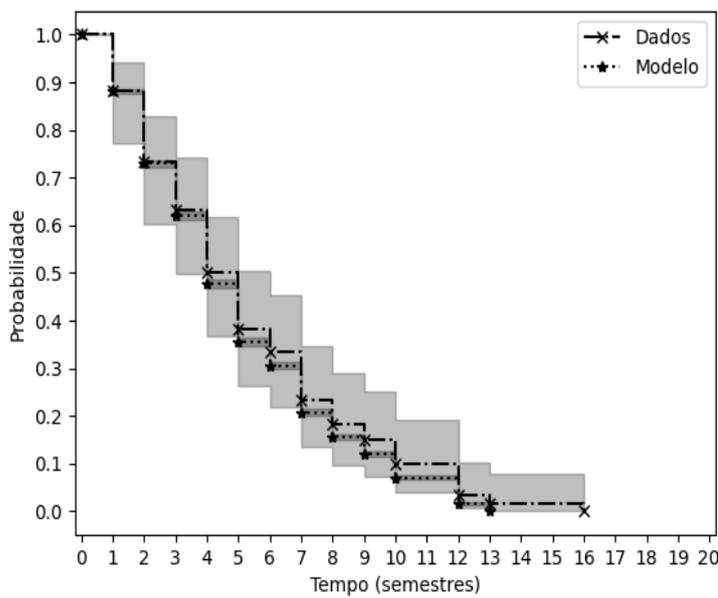


(b) Conclusão

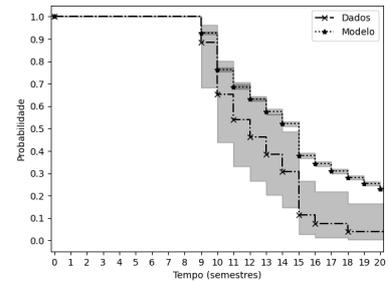


(c) Desvinculação

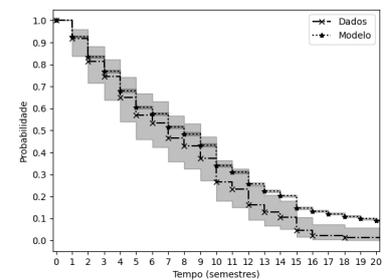
Figura 19 – Análise de Sobrevivência do Modelo em Relação aos Dados para a Cor Parda.



(a) Evasão



(b) Conclusão



(c) Desvinculação

Figura 20 – Análise de Sobrevivência do Modelo em Relação aos Dados para a Cor Branca.

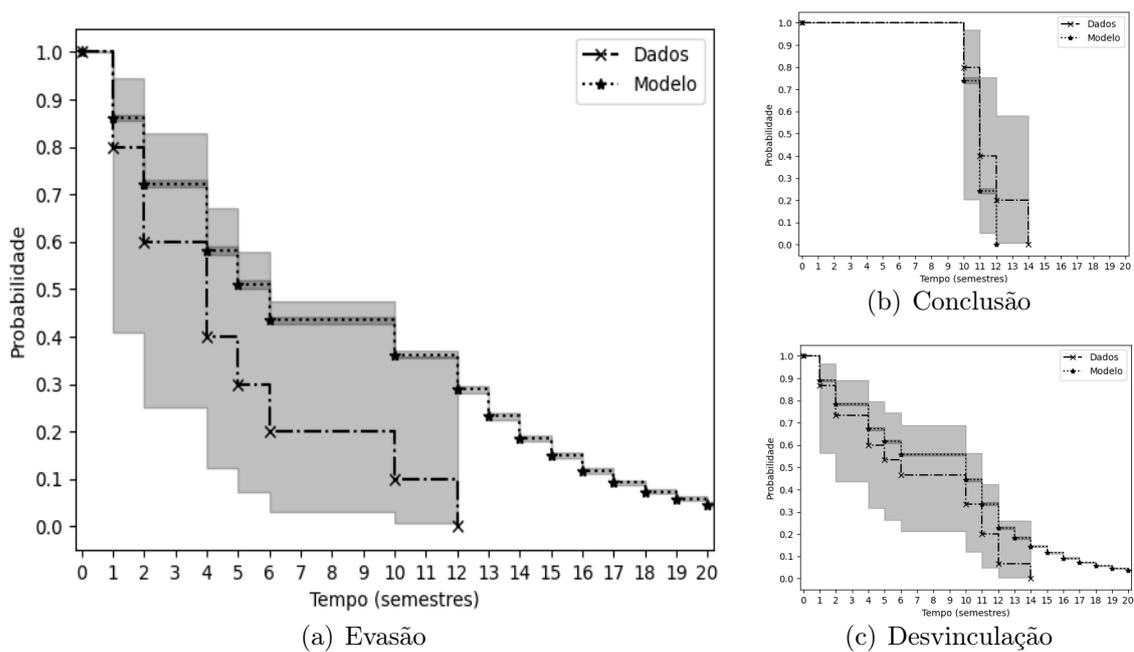


Figura 21 – Análise de Sobrevivência do Modelo em Relação aos Dados para a Cor Preta.

Simulações são realizadas para cada cor/raça, através da aplicação de suas *matrizes de transição*⁴ no modelo. Então, o resultado do modelo é aplicado na Análise de Sobrevivência para verificar a diferença de cada cor/raça nas probabilidades de evasão, conclusão e desvinculação de acordo com o tempo em que os estudantes estão vinculados aos cursos. Os resultados disponíveis, nas Figuras 22(a), 22(b) e 22(c), para a categoria cor/raça demonstram diferenças significativas entre estes grupos de estudantes. A cor Preta apresenta menor probabilidade de conclusão entre o nono e décimo primeiro período, porém após esse intervalo de tempo sua chance de conclusão aumenta, alcançando a probabilidade de conclusão por período da cor parda. Em contrapartida, a cor branca apresenta melhores probabilidades de conclusão até o décimo segundo semestres, então essa chance reduz de forma significativa por período.

A Tabela 13, destaca que a cor parda apresenta maior probabilidade de evasão e menor tempo para retenção. No entanto, a cor branca demora mais para concluir. A cor preta apresenta a maior probabilidade de conclusão, apesar de demorar mais tempo para concluir o curso.

⁴ <https://github.com/JulianaFerreira/educacao/blob/6c3e469711d275a5967fed80c642585d29417a33/matrizes-de-transicao/comp-cor-branca.csv>

⁴ <https://github.com/JulianaFerreira/educacao/blob/6c3e469711d275a5967fed80c642585d29417a33/matrizes-de-transicao/comp-cor-parda.csv>

⁴ <https://github.com/JulianaFerreira/educacao/blob/6c3e469711d275a5967fed80c642585d29417a33/matrizes-de-transicao/comp-cor-preta.csv>

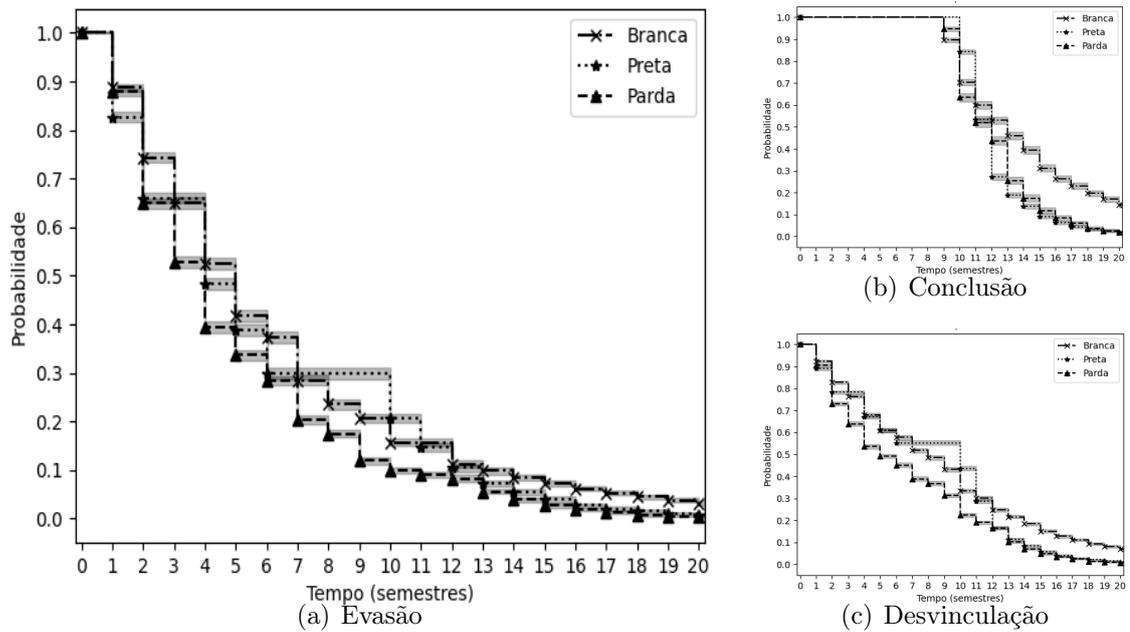


Figura 22 – Análise de Sobrevivência para a Categoria Cor/Raça.

Tabela 13 – Resultados das Probabilidades e dos Tempos para a Categoria Cor/Raça.

	Probabilidade (%)			Tempo (semestres)		
	Branca	Preta	Parda	Branca	Preta	Parda
Retenção	75,34	72,02	72,63	3,42	3,54	2,84
Evasão	60,49	55,97	79,77	4,95	4,71	5,66
Graduação	27,76	40,89	20,23	12,39	13,14	11,52
Desvinculação	88,25	96,86	100,00	7,29	8,27	6,84

As análises realizadas para verificar as características dos cursos e estudantes demonstram uma diferença significativa entre os critérios analisados. Percebe-se que ao analisar os cursos há uma diferença principalmente na evasão por semestre. Ao analisar o gênero, a diferença é notável na evasão e na conclusão por semestre. Na análise da cor/raça, todos os grupos apresentam diferenças na evasão, conclusão e desvinculação. Portanto, confirma o que foi demonstrado no estudo (BRITO et al., 2020), que evidencia que o gênero e a cor/raça são atributos importantes para avaliar a evasão estudantil.

6.3 Estudo de Caso 3: Evasão e Retenção para Cursos das Áreas de Agrárias, Computação e Saúde

No estudo de caso a seguir, é realizada uma análise da evasão para os quatro primeiros semestres de três cursos e uma análise da evasão, conclusão e desvinculação por semestre em que o estudante ficou retido. Os cursos selecionados para as análises são um da área de ciências agrárias, um de computação e um de saúde, onde as matrizes de transição do modelo foram definidas conforme essas áreas.

6.3.1 Análise da Evasão nos Quatro Primeiros Períodos

Na análise a seguir, é avaliada a evasão para os quatro primeiros semestres de três cursos. Esses semestres foram selecionados pois são os que apresentam maior probabilidade de evasão. Primeiramente, considerando os dados, foram calculados os intervalos de confiança para cada curso, conforme observa-se na Tabela 14. Nela o tempo para evasão dos três cursos estão dentro do intervalo de confiança um do outro, o que valida que esses cursos apresentam uma evasão até o quarto semestre equivalente entre si. Observa-se também que os intervalos de confiança dos dados e do modelo estão no mesmo intervalo, o que demonstra a confiabilidade do modelo.

Tabela 14 – Intervalo de Confiança dos Cursos para Evasão nos Primeiros Quatro Semestres.

	Tempo (semestres)			
	Dados		Modelo	
	média	95% IC	média	95% IC
Curso de Agrárias	2,2864	[2,1254;2,4474]	2,2986	[2,2624;2,3348]
Curso de Computação	2,4907	[2,2929;2,6884]	2,5249	[2,4868;2,5631]
Curso de Saúde	2,3035	[1,9768;2,6302]	2,3608	[2,2962;2,4253]

As simulações para os três cursos analisados são realizadas através da aplicação de suas *matrizes de transição*⁵ no modelo. Então, o resultado do modelo é aplicado na Análise de Sobrevivência, disponível na Figura 23, para verificar a diferença de comportamento das probabilidades de evasão, conclusão e desvinculação apenas para os quatro primeiros

semestres dos cursos. Observa-se que o Curso de Agrárias apresenta a maior probabilidade de evasão por semestre entre os cursos apresentados. Ao final dos quatro semestres, o Curso de Agrárias já apresenta uma evasão de 70% dos estudantes, enquanto nos outros cursos a evasão é em torno de 50%. Em relação a evasão geral, até o quarto período o Curso de Agrárias teve uma evasão de 38,63%, o Curso de Computação uma evasão de 27,93% e o Curso de Saúde uma evasão de 13,89%, o que representa uma diferença significativa entre todos os cursos, diferente do apresentado quando observamos a probabilidade de evasão apenas entre os evadidos por semestre.

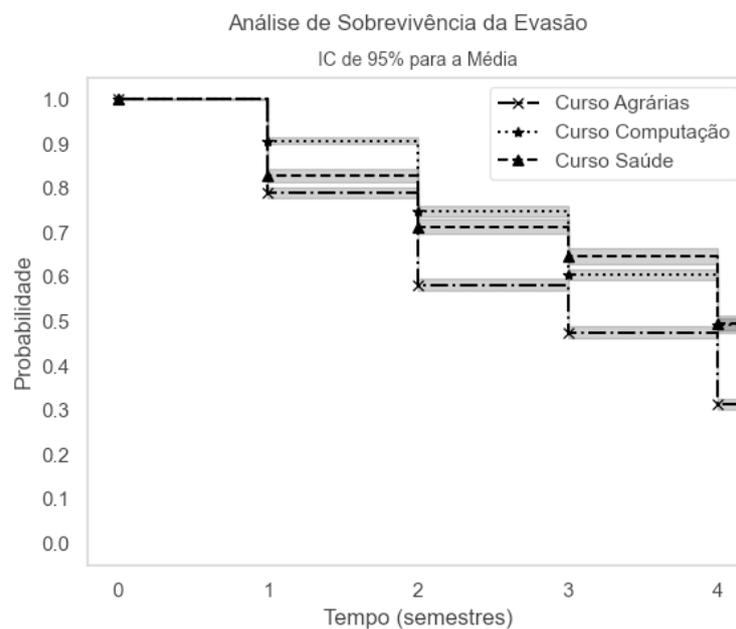


Figura 23 – Análise de Sobrevivência da Evasão para os Quatro Primeiros Semestres dos Cursos.

Na Figura 24, observa-se a densidade da evasão nos quatro primeiros semestres dos cursos, onde acontece a maior parte da evasão nos cursos apresentados. No curso de agrárias e saúde, a evasão é maior nos primeiros dois semestres e diminui no terceiro, diferente do que ocorre no curso de computação, onde a evasão se mantém estável pelos quatro primeiros semestres do curso.

Ao realizar o teste de hipótese Kolmogorov-Smirnov, para avaliar a evasão nos quatro primeiros semestres de curso, o resultado do p-valor foi inferior a 0,05 ao comparar

⁵ <https://github.com/JulianaFerreira/educacao/blob/6c3e469711d275a5967fed80c642585d29417a33/matrizes-de-transicao/agrarias.csv>

⁵ <https://github.com/JulianaFerreira/educacao/blob/6c3e469711d275a5967fed80c642585d29417a33/matrizes-de-transicao/saude.csv>

⁵ <https://github.com/JulianaFerreira/educacao/blob/6c3e469711d275a5967fed80c642585d29417a33/matrizes-de-transicao/curso2.csv>

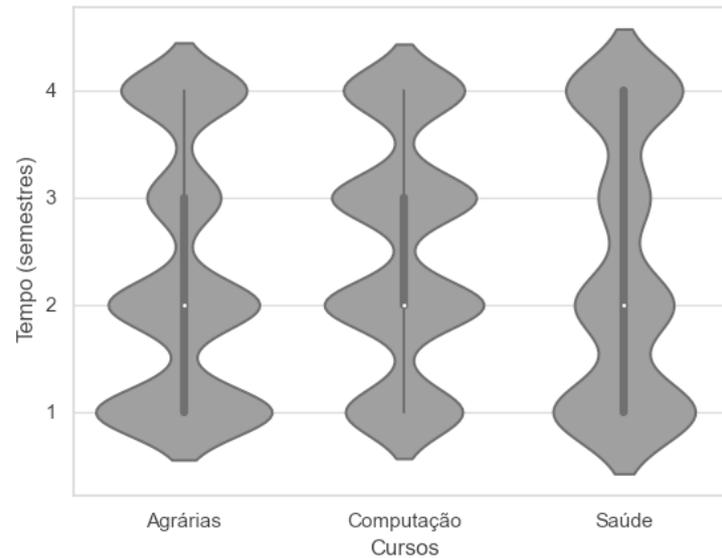


Figura 24 – Violin Plot da Evasão para os Quatro Primeiros Semestres dos Cursos.

todos os cursos, ou seja, nenhum dos três cursos aceitam a hipótese nula de que suas distribuições são equivalentes. Pois, apesar do tempo de evasão estar no intervalo de confiança, a distribuição da probabilidade por semestres dos estudantes apresenta diferença significativa.

6.3.2 Análise da Evasão, Conclusão e Desvinculação por Semestre de Retenção

Na análise a seguir, observa-se a evasão, conclusão e desvinculação por semestre em que o estudante entrou em retenção através da aplicação de uma *matriz de transição*⁶ no modelo. Serão analisados os estudantes sem retenção e os retidos no período 1, 2, 4, 6 e do período 8 em diante. Essa análise permite avaliar o comportamento dos estudantes conforme o período em que ficou retido e avaliar o seu impacto na evasão, conclusão e desvinculação.

Por meio da Análise de Sobrevivência, disponível na Figura 25(a), observa-se a evasão dos estudantes por semestre de retenção. Os estudantes sem retenção apresentam 80% de evasão já no primeiro período, com uma evasão menor por semestre até o nono período. Os estudantes retidos nos semestres 1 e 2 apresentam um comportamento similar

⁶ <https://github.com/JulianaFerreira/educacao/blob/6c3e469711d275a5967fed80c642585d29417a33/matrices-de-transicao/agrarias-comp-saude.csv>

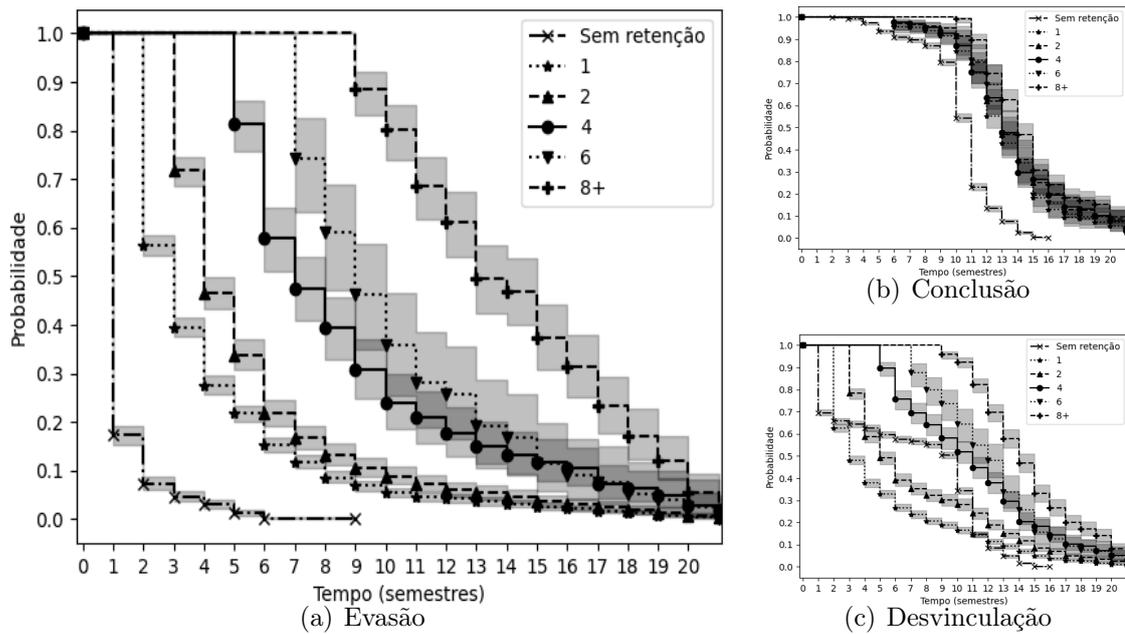


Figura 25 – Análise de Sobrevivência por semestre de retenção.

após o décimo semestre, com uma retenção nos primeiros semestres inferior à apresentada pelos estudantes não retidos. Os estudantes retidos nos semestres 4 e 6 apresentam um comportamento semelhante entre si, com uma evasão por semestre inferior ao apresentada pelos estudantes que retêm em semestres anteriores. Os estudantes retidos no semestre 8 em diante apresentam uma evasão equilibrada por semestre. Na Figura 25(b), observa-se a conclusão dos estudantes por semestre de retenção. Os estudantes sem retenção apresentam 90% de conclusão até o período 13. Nos estudantes com retenção, independente do semestre, a conclusão ocorre de forma similar, com 90% dos estudantes concluídos apenas em torno do período 18. Na Figura 25(c), observa-se a desvinculação dos estudantes por semestre de retenção, ela apresenta o momento em que o estudante perde o vínculo com o curso, tanto por evasão ou por conclusão.

Os resultados demonstram que o semestre em que o estudante ficou retido impacta na evasão por semestre, apesar de apresentar pouca influência na conclusão, onde a diferença fica evidente apenas entre os estudantes não retidos.

6.4 Considerações Finais

Neste capítulo, foram apresentados os estudos de caso. Nos estudos, os experimentos analisaram como o comportamento da retenção e evasão nos primeiros semestres influenciam

na evasão, conclusão e desvinculação geral. Além disso, também foi observado o comportamento dos estudantes considerando as categorias curso, gênero e cor/raça. Ademais, também foi realizada uma análise da evasão e retenção para cursos das áreas de agrárias, computação e saúde. No próximo capítulo será apresentada a conclusão deste trabalho.

7 Conclusão

A evasão e a retenção são problemas que afetam as instituições de ensino superior. Esses problemas podem causar diversos prejuízos para as universidades, como ociosidade de recursos ou perda de receita. Por isso, existe a necessidade de analisar seus impactos, para entender melhor esses problemas e determinar estratégias mais eficientes para combatê-los.

Este trabalho teve como objetivo realizar uma análise da retenção, evasão, conclusão e desvinculação, através de um modelo de Cadeia de Markov, proposto para simular o progresso dos estudantes durante um curso de graduação. O modelo é capaz de realizar experimentos e análises que seriam complexas sem ele. As análises realizadas foram representadas através de Análise de Sobrevivência e permitiram verificar o comportamento da evasão, conclusão e desvinculação, relacionando a probabilidade de ocorrer esses eventos com o tempo de vínculo dos estudantes.

Nos trabalhos relacionados foram identificadas as principais técnicas de modelagem utilizadas para analisar a evasão estudantil. O uso de Cadeias de Markov e Análise de Sobrevivência mostrou-se eficiente para representar o progresso acadêmico e identificar fatores que influenciam na evasão, conclusão e desvinculação.

Os dados de três cursos da UFRPE foram coletados e analisados, para serem utilizados como parâmetros para o modelo de Cadeia de Markov. Esse modelo permite generalização com alteração de parâmetros, isso possibilita aplicá-lo em diferentes contextos, além de permitir prever o comportamento dos estudantes ao reduzir a retenção e evasão nos primeiros semestres de curso, onde geralmente se concentra a maior probabilidade de evasão.

Dessa forma, foi proposto neste trabalho, uma Cadeia de Markov absorvente capaz de representar o progresso dos estudantes pelos semestres em que estão vinculados ao curso. Através dos resultados obtidos pelo modelo, verificou-se com Análise de Sobrevivência a evasão, conclusão e desvinculação dos estudantes de graduação.

Nos estudos de caso, foram avaliados: o comportamento de evasão, conclusão e desvinculação de dois cursos da área de computação; como a evasão e a retenção nos primeiros períodos de curso influenciam na evasão e conclusão dos cursos; grupos de estudantes por curso, gênero e cor/raça; a análise da evasão nos quatro primeiros semestres de três cursos das áreas de agrária, computação e saúde; e a análise da evasão, conclusão e

desvinculação por semestre em que o estudante ficou retido para três cursos das áreas de agrária, computação e saúde.

No primeiro estudo de caso, os resultados mostraram que ao analisar o impacto de diminuir a retenção apenas no primeiro semestre, existe um impacto somente em alguns períodos iniciais da evasão, o que representa uma diferença inferior a 5% na evasão final. Porém, ao analisar o impacto da diminuição da evasão e retenção nos quatro primeiros semestres, observou-se uma diminuição da evasão final em torno de 19%. Além de ser possível perceber que a retenção tem um impacto maior do que a evasão nas probabilidades finais gerais de evasão, conclusão e desvinculação dos estudantes, com um impacto menor no tempo. Dessa forma, adotar políticas que foquem na diminuição da retenção nesses semestres de curso contribui para uma redução significativa da evasão final. Portanto, permite criar um planejamento para direcionar esforços aos estudantes retidos nos primeiros períodos.

Ao analisar as categorias curso, gênero e cor/raça, os resultados demonstraram o comportamento de cada um desses grupos de estudantes individualmente. Isso permitiu identificar os grupos de estudantes com maior probabilidade de evasão, o que permite criar estratégias direcionadas, de forma a melhorar o desempenho acadêmico destes estudantes. Para a categoria curso, o Curso 1 apresentou maior evasão e retenção. No gênero, o masculino apresentou maior evasão e retenção. Por fim, para cor/raça, a cor branca apresentou maior retenção e a cor parda maior evasão.

No estudo de caso que observou o comportamento de cursos da área de agrárias, computação e saúde, ao analisar a evasão nos quatro primeiros semestres de curso, observou-se que o curso de saúde apresenta uma evasão por semestre maior do que os cursos de agrárias e computação, que apresentam comportamento semelhante. Apesar do curso de computação apresentar uma evasão mais equilibrada por semestres, ao contrário dos cursos de agrárias e saúde, onde a evasão diminui no terceiro período.

Quando analisada a evasão por semestre de retenção, pode-se observar que os estudantes sem retenção apresentam grande evasão no primeiro período e concluem o curso mais rápido. Também foi possível verificar que os estudantes retidos no primeiro e segundo semestre apresentam comportamento semelhante na evasão, o mesmo ocorre com os retidos no quarto e sexto semestre.

As principais contribuições apresentadas neste trabalhos foram a proposta de um

modelo de Cadeia de Markov absorvente para modelar o progresso acadêmico por semestres, a realização de testes e simulações de diferentes cenários apresentados nos estudos de caso e as análises apresentadas através da Análise de Sobrevivência para avaliar a evasão, conclusão e desvinculação. Apesar das contribuições apresentadas, houveram limitações relacionadas a pequena quantidade de dados utilizados para o estudo de caso de cor/raça.

7.1 Trabalhos Futuros

Como trabalho futuro, pode-se utilizar o modelo proposto para analisar dados de mais cursos da mesma universidade. Ademais, também é possível analisar os dados considerando outros agrupamentos de estudantes, para verificar como outras categorias se comportam considerando as probabilidades e tempos de evasão, conclusão e desvinculação. Isso permite aos gestores maior controle de quais estudantes precisam de maior auxílio para concluir o curso.

Os estudos de caso deste trabalho utilizaram dados de uma universidade específica, mas o modelo de Cadeia de Markov apresentado permite generalizações. Isso possibilita considerar outros parâmetros, além dos apresentados neste trabalho, o que permite identificar como esses influenciam nas probabilidades finais de evasão e conclusão. Além disso, também é possível aplicar o modelo para outras instituições de ensino superior ou considerar ações de combate a evasão para verificar seu impacto.

Outra possibilidade, é considerar diferentes configurações de estados para representar o progresso dos estudantes. Pode-se alterar o que determina um estudante retido ou considerar outro aspecto de grande impacto na evasão estudantil que crie um caminho alternativo na progressão dos estudantes, assim como a retenção.

Referências

- ADAM, R. Y. An application of markov modeling to the student flow in higher education in sudan. **International Journal of Science and Research**, v. 4, n. 2, p. 49–54, 2015.
- AL-AWADHI, S. A.; KONSOWA, M. An application of absorbing markov analysis to the student ow in an academic institution. **Kuwait J. Sci. Eng**, v. 34, n. 2A, p. 77–89, 2007.
- ALAWADHI, S.; KONSOWA, M. Markov chain analysis and student academic progress: An empirical comparative study. **Journal of Modern Applied Statistical Methods**, v. 9, n. 2, p. 26, 2010.
- ALVES, F. G.; MARTINS, J. V. G.; FRANCO, H. J. R.; PEREIRA, L. L. C. Determinação de indicadores educacionais do instituto federal do sudeste de minas gerais via cadeias absorventes de markov. **Revista de Matemática**, v. 2, p. 1–16, 2020.
- AMORIM, M. J.; BARONE, D.; MANSUR, A. U. Técnicas de aprendizado de máquina aplicadas na previsao de evasao acadêmica. In: **Brazilian Symposium on Computers in Education (Simpósio Brasileiro de Informática na Educação-SBIE)**. [S.l.: s.n.], 2008. v. 1, n. 1, p. 666–674.
- ARAÚJO, D. F. d. **Análise e avaliação da trajetória de estudantes de graduação baseadas em modelagem por cadeias de Markov**. Tese (Doutorado) — Programa de Pós-Graduação em Biometria e Estatística Aplicada, 2019. Departamento de Estatística e Informática.
- ARSANJANI, J. J.; HELBICH, M.; KAINZ, W.; BOLOORANI, A. D. Integration of logistic regression, markov chain and cellular automata models to simulate urban expansion. **International Journal of Applied Earth Observation and Geoinformation**, Elsevier, v. 21, p. 265–275, 2013.
- BAIRAGI, A.; KAKATY, S. C. A stochastic process approach to analyse students performance in higher education institutions. **Int. J. Stat. Syst**, v. 12, p. 323–342, 2017.
- BALOGH, Z.; KUCHÁRIK, M. Predicting student grades based on their usage of lms moodle using petri nets. **Applied Sciences**, Multidisciplinary Digital Publishing Institute, v. 9, n. 20, p. 4211, 2019.
- BALOGH, Z.; TURČÁNI, M.; MAGDIN, M. The possibilities of using petri nets for realization of a universal model of educational process. In: IEEE. **2013 IEEE 14th International Conference on Information Reuse & Integration (IRI)**. [S.l.], 2013. p. 162–169.
- BÄUERLE, N.; RIEDER, U. **Markov decision processes with applications to finance**. [S.l.]: Springer Science & Business Media, 2011.
- BELTRAN, C. A. R.; XAVIER-JÚNIOR, J. C.; BARRETO, C. A.; NETO, C. O. Plataforma de aprendizado de maquina para detecção e monitoramento de alunos com risco de evasão. In: **Brazilian Symposium on Computers in Education (Simpósio Brasileiro de Informática na Educação-SBIE)**. [S.l.: s.n.], 2019. v. 30, p. 1591.

BOLCH, G.; GREINER, S.; MEER, H. D.; TRIVEDI, K. S. **Queueing networks and Markov chains: modeling and performance evaluation with computer science applications**. [S.l.]: John Wiley & Sons, 2006.

BOUMI, S.; VELA, A. E. Improving graduation rate estimates using regularly updating multi-level absorbing markov chains. **Education Sciences**, Multidisciplinary Digital Publishing Institute, v. 10, n. 12, p. 377, 2020.

BREZAVŠČEK, A.; BACH, M. P.; BAGGIA, A. Markov analysis of students' performance and academic progress in higher education. **Organizacija**, Sciendo, v. 50, n. 2, p. 83–95, 2017.

BRITO, B. C. P. D.; MELLO, R. F. L. D.; ALVES, G. Identificação de atributos relevantes na evasão no ensino superior público brasileiro. In: SBC. **Anais do XXXI Simpósio Brasileiro de Informática na Educação**. [S.l.], 2020. p. 1032–1041.

CAMPOS, J. D. d. S. **Fatores explicativos para a evasão no Ensino Superior através da análise de sobrevivência: o caso da UFPE**. Dissertação (Mestrado) — Universidade Federal de Pernambuco, 2016.

CARRANO, D.; ALBERGARIA, E. T. D.; INFANTE, C.; ROCHA, L. Combinando técnicas de mineração de dados para melhorar a detecção de indicadores de evasão universitária. In: **Brazilian Symposium on Computers in Education (Simpósio Brasileiro de Informática na Educação-SBIE)**. [S.l.: s.n.], 2019. v. 30, p. 1321.

COLOSIMO, E. A.; GIOLO, S. R. **Análise de sobrevivência aplicada**. [S.l.]: Editora Blucher, 2006.

COSTA, F. J. d.; BISPO, M. d. S.; PEREIRA, R. d. C. d. F. Dropout and retention of undergraduate students in management: a study at a brazilian federal university. **RAUSP Management Journal**, SciELO Brasil, v. 53, n. 1, p. 74–85, 2018.

CROSSEN, D. M. Student retention performance using absorbing markov chains. In: **Organizational Productivity and Performance Measurements Using Predictive Modeling and Analytics**. [S.l.]: IGI Global, 2017. p. 293–323.

DAI, H.; ZHAO, G.; LIN, M.; WU, J.; ZHENG, G. A novel estimation method for the state of health of lithium-ion battery using prior knowledge-based neural network and markov chain. **IEEE Transactions on Industrial Electronics**, IEEE, v. 66, n. 10, p. 7706–7716, 2018.

EDUCAÇÃO, C. N. de. **Plano Nacional de Educação (PNE)**. MEC, 2020. Acesso em: 19/04/2021. Disponível em: <https://pne.mec.gov.br/images/pdf/pne_conhecendo_20_metas.pdf>.

ELEDUM, H.; IDRIS, E. I. M. An undergraduate student flow model: Semester system in university of tabuk (ksa). **Int. J. Stat. Appl. Math**, v. 4, p. 11–19, 2019.

HAGAR, Y.; ALBERS, D.; PIVOVAROV, R.; CHASE, H.; DUKIC, V.; ELHADAD, N. Survival analysis with electronic health record data: Experiments with chronic kidney disease. **Statistical Analysis and Data Mining: The ASA Data Science Journal**, Wiley Online Library, v. 7, n. 5, p. 385–403, 2014.

HLAVATÝ, R.; DÖMEOVÁ, L. Students' progress throughout examination process as a markov chain. **International Education Studies**, ERIC, v. 7, n. 12, p. 20–29, 2014.

INEP. **Sinopse Estatística da Educação Superior 2019**. [S.l.]: INEP, 2020. Disponível em: <<http://portal.inep.gov.br/basica-censo-escolar-sinopse-sinopse>>. Acesso em: 19/04/2021.

INOUE, S.; YORIFUJI, T.; TAKAO, S.; DOI, H.; KAWACHI, I. Social cohesion and mortality: a survival analysis of older adults in japan. **American journal of public health**, American Public Health Association, v. 103, n. 12, p. e60–e66, 2013.

JESUS, H. O. D.; RODRIGUEZ, L. C.; JUNIOR, A. d. O. C. Predição de evasão escolar na licenciatura em computação. **Revista Brasileira de Informática na Educação**, v. 29, p. 255–272, 2021.

JUHÁSOVA, A.; KAZLOV, I.; JUHÁS, G.; MOLNÁR, L. How to model curricula and learnflows by petri nets-a survey. In: IEEE. **2016 International Conference on Emerging eLearning Technologies and Applications (ICETA)**. [S.l.], 2016. p. 147–152.

JUNIOR, P. L.; SILVEIRA, F. L. d.; OSTERMANN, F. Análise de sobrevivência aplicada ao estudo do fluxo escolar nos cursos de graduação em física: um exemplo de uma universidade brasileira. **Revista brasileira de ensino de física**, SciELO Brasil, v. 34, 2012.

KAPLAN, E. L.; MEIER, P. Nonparametric estimation from incomplete observations. **Journal of the American statistical association**, Taylor & Francis, v. 53, n. 282, p. 457–481, 1958.

KLEINBAUM, D. G.; MITCHEL, K. **Survival analysis**. [S.l.]: Springer, 2010. v. 3.

LIMA, E.; COUTINHO, E. Uma análise sobre o desempenho de alunos de graduação em disciplinas iniciais de programação. In: **Brazilian Symposium on Computers in Education (Simpósio Brasileiro de Informática na Educação-SBIE)**. [S.l.: s.n.], 2019. v. 30, p. 1876.

MACIEL, C. E.; CUNHA, M.; LIMA, T. d. S. A produção científica sobre permanência e evasão na educação superior no brasil. **Educação e Pesquisa**, SciELO Brasil, v. 45, 2019.

MEC. **Indicadores de Fluxo da Educação Superior**. MEC, 2020. Acesso em: 05/04/2021. Disponível em: <https://download.inep.gov.br/informacoes_estatisticas/indicadores_educacionais/2017/metodologia_indicadores_trajetoria_curso.pdf>.

MOORE, T.; CHRISTIN, N. Beware the middleman: Empirical analysis of bitcoin-exchange risk. In: SPRINGER. **International Conference on Financial Cryptography and Data Security**. [S.l.], 2013. p. 25–33.

OBAIDAT, M. S.; BOUDRIGA, N. A. **Fundamentals of performance evaluation of computer and telecommunication systems**. [S.l.]: John Wiley & Sons, 2010.

OBAIDAT, M. S.; PAPADIMITRIOU, G. I. **Applied system simulation: methodologies and applications**. [S.l.]: Springer Science & Business Media, 2012.

- RIGO, S. J.; CAMBRUZZI, W.; BARBOSA, J. L.; CAZELLA, S. C. Aplicações de mineração de dados educacionais e learning analytics com foco na evasão escolar: oportunidades e desafios. **Revista Brasileira de Informática na Educação**, v. 22, p. 132, 2014.
- RODRIGUES, F. S.; BRACKMANN, C. P.; BARONE, D. A. C. Estudo da evasão no curso de ciência da computação da ufrgs. **Revista Brasileira de Informática na Educação**, v. 23, 2015.
- SACCARO, A.; FRANÇA, M. T. A.; JACINTO, P. d. A. Fatores associados à evasão no ensino superior brasileiro: um estudo de análise de sobrevivência para os cursos das áreas de ciência, matemática e computação e de engenharia, produção e construção em instituições públicas e privadas. **Estudos Econômicos (São Paulo)**, SciELO Brasil, v. 49, n. 2, p. 337–373, 2019.
- SANTOS, J.; SOUSA, J. D.; MELLO, R.; CRISTINO, C.; ALVES, G. Um modelo para análise do impacto da retenção e evasão no ensino superior utilizando cadeias de markov absorventes. In: **Anais do XXXII Simpósio Brasileiro de Informática na Educação**. Porto Alegre, RS, Brasil: SBC, 2021. p. 813–823.
- SHESKIN, T. J. **Markov chains and decision processes for engineers and managers**. [S.l.]: CRC Press, 2016.
- SILVA, A. M. D.; SANTOS, B. C. S. Eficácia de políticas de acesso ao ensino superior privado na contenção da evasão. **Avaliação: Revista da Avaliação da Educação Superior**, Universidade de Sorocaba, v. 22, n. 3, p. 741–757, 2017.
- TAJ, S. M.; KUMARAVEL, A. Colored petri nets for prediction of student knowledge level. In: AIP PUBLISHING LLC. **AIP Conference Proceedings**. [S.l.], 2019. v. 2177, n. 1, p. 020045.
- TCU. **Orientações Para o Cálculo dos Indicadores de Gestão**. 2004. Acesso em: 05/03/2022. Disponível em: <<http://portal.mec.gov.br/setec/arquivos/pdf/indicadores.pdf>>.
- TEODORO, L. D. A.; KAPPEL, M. A. A. Aplicação de técnicas de aprendizado de máquina para predição de risco de evasão escolar em instituições públicas de ensino superior no brasil. **Revista Brasileira de Informática na Educação**, v. 28, p. 838–863, 2020.
- TIGRINHO, L. M. V. Evasão escolar nas instituições de ensino superior. **Revista Gestão Universitária**, v. 173, n. 01-14, p. 15–21, 2008.
- WANG, W.; DONG, J.; TAN, T. Image tampering detection based on stationary distribution of markov chain. In: IEEE. **2010 IEEE International Conference on Image Processing**. [S.l.], 2010. p. 2101–2104.
- XAVIER, M.; MENESES, J. A literature review on the definitions of dropout in online higher education. In: **EDEN Conference Proceedings**. [S.l.: s.n.], 2020. p. 73–80.

APÊNDICES

APÊNDICE A – Código da Cadeia de Markov

```
1 import numpy as np
2 import pandas as pd
3
4
5 class Estudante:
6
7     def __init__(self, estudante_id, matriz):
8         self.estudante_id = estudante_id
9         self.df = pd.read_csv(matriz, index_col=0)
10        self.estados = self.df.columns
11        self.estado = self.estados[0]
12        self.tamanho = 1
13        self.markov = self.markov()
14        self.transicoes = [self.estados[0]]
15
16    def pegar_prox_estado(self):
17        return next(self.markov)
18
19    def markov(self):
20
21        while self.estado != 'G' and self.estado != 'E':
22
23            proximo_estado = np.random.choice(self.estados,
24                                               1, p=self.df.loc[f'{self.estado}'])[0]
25
26            if proximo_estado == 'G':
27                self.estado = 'G'
28                self.transicoes.append(self.estado)
29                self.tamanho += 1
30                yield self.estado
```

```
30
31     elif proximo_estado == 'E':
32         self.estado = 'E'
33         self.transicoes.append(self.estado)
34         self.tamanho += 1
35         yield self.estado
36
37     else:
38         self.estado = proximo_estado
39         self.transicoes.append(self.estado)
40         self.tamanho += 1
41         yield self.estado
```