



MINISTÉRIO DA EDUCAÇÃO
INSTITUTO FEDERAL DE EDUCAÇÃO, CIÊNCIA E TECNOLOGIA DE BRASÍLIA

SIMULAÇÃO ESTOCÁSTICA APLICADA À GESTÃO ESCOLAR

FABIO NOGUEIRA CARLUCCI

BRASÍLIA/DF
2026

FABIO NOGUEIRA CARLUCCI

SIMULAÇÃO ESTOCÁSTICA APLICADA À GESTÃO ESCOLAR

Tese apresentada ao Instituto Federal de Educação,
Ciência e Tecnologia de Brasília – IFB, para
progressão funcional à Classe E (Professor Titular) no
Magistério do Ensino Básico, Técnico e Tecnológico.

BRASÍLIA/DF
2026

C285s Carlucci, Fabio Nogueira.
Simulação estocástica aplicada à gestão escolar. / Fabio Nogueira Carlucci.
– Brasília, 2026.
99 f. : il. color. ; 30 cm.

Tese Acadêmica Inédita (Professor Titular) – Instituto Federal de Educação,
Ciência e Tecnologia de Brasília, *campus* Brasília, 2026.

1. Cadeia de Markov. 2. Simulação de Monte Carlo. 3. Matriz de transição
estocástica. 4. Gestão escolar. 5. Medida de probabilidade. I. Instituto Federal
de Brasília. II. Título.

CDU 37.07


MINISTÉRIO DA EDUCAÇÃO

Instituto Federal de Educação, Ciência e Tecnologia de Brasília

PARECER CIRCUNSTANCIADO DE AVALIAÇÃO DE PROMOÇÃO À CLASSE DE PROFESSOR TITULAR.
DADOS DO CANDIDATO:

Nome Completo	Fabio Nogueira Carlucci
Nº do Processo	23508.000961.2026-56
Matrícula SIAPE	1323601
E-mail de contato	fabio.carlucci@ifb.edu.br
Telefones para Contato	(61) 99111-1551

BANCA DE AVALIAÇÃO:

NOME DOS AVALIADORES	SIAPE	FUNÇÃO
Cristiano Pereira da Silva (IFB)	1920257	Interno - Presidente
Jean Carlos da Silva Galdino (IFRN)	1349806	Externo – (Sigla da instituição)
Arlyson Alves do Nascimento (IFAL)	1705985	Externo – (Sigla da instituição)
Paulo Roberto Vilarim (IFMS)	1845753	Externo – (Sigla da instituição)

PARECER
descrição das atividades

- A banca iniciou as **13:10** do dia **25 de setembro** do **2026** e finalizou às **15:15**
- Defesa realizada de forma virtual pelo link:
https://calendar.google.com/calendar/finvxf?tot=AZ3PcR8waNSi5H4quEpYNH0WxlURgOUeF4KfXzmlmAgVMo2SeBj9hnnw5RZgpn6f180s6pJD2L8Wv1A-3ITx94RDih_zGwIZKRqEJC-dxiv2YHn3FgiV
- Foi dada a palavra ao candidato que fez a apresentação da tese inédita por 60 min (1h).
- Os membros da banca avaliadora fizeram perguntas e considerações que foram respondidas satisfatoriamente pelo candidato.
- O presidente da banca fez a leitura do parecer com o resultado da sessão pública.

Considerando:

- O processo e os documentos comprobatórios avaliados;
- A realização de sessão pública de defesa de tese inédita;
- Os pareceres individuais de cada membro da banca de avaliação.

Todos os avaliadores, Arlyson Alves do Nascimento (IFAL), Jean Carlos da Silva Galdino (IFRN), Paulo Roberto Vilarim (IFMS) e Cristiano Pereira da Silva (IFB) apresentaram parecer favorável e votaram pela **aprovação** da promoção do candidato à Classe de Professor Titular.

LOCAL E DATA: Brasília, 25 de setembro de 2026



Documento assinado digitalmente

CRISTIANO PEREIRA DA SILVA

Data: 25/09/2026 16:23:03-0300

 Verifique em <https://validar.it.gov.br>

Cristiano Pereira da Silva
Presidente da Banca



INSTITUTO FEDERAL
Brasília
Campus Brasília

SGAN 610, Módulos D, E, F e G
Asa Norte – Brasília-DF, CEP 70.830-450
(61) 2193-8128 | ifb@edu.br

Dedicatória

Dedico esta tese para minha esposa, meus filhos e netos que sempre me apoiaram em meus estudos e aos colegas professores nos quais procuro me espelhar.

Agradecimentos

Agradeço a todos os servidores do Instituto Federal de Educação, Ciência e Tecnologia de Brasília pelo excelente ambiente de trabalho, propiciando e incentivando o ensino, a pesquisa e a extensão.

Resumo

Na presente tese, após revisitar o sistema educacional brasileiro e a norma ABNT NBR ISO 21001 (ISO 21001:2018), estabelecendo a tomada de decisão baseada em evidências, como um dos pilares da gestão acadêmica, foi estabelecida essa análise com base em dados de avaliações, evasão e desempenho para as quais foram definidos os Indicadores Chave de Performance (KPIs). Posto isto, com foco no Ensino Médio Integrado, mas que poderia ser o Ensino Fundamental ou Ensino Médio ou, ainda, Ensino Superior com as devidas adaptações e, com a premissa de que a complexidade da tomada de decisões é diretamente proporcional à quantidade de informações disponíveis para fundamentá-las, incorporamos o estudo de fenômenos que passam de um estado inicial para uma sequência de estados, cujas transições ocorrem segundo algum nível de incerteza, caracterizada por uma probabilidade, aliando às boas práticas como escolher as melhores ações. Foi realizada uma revisão a respeito da modelagem de Markov e cadeias de Markov, abordando os conceitos de estados estacionários e a evolução do sistema ao longo do tempo, seja por meio de diagrama de árvores ou por meio da álgebra matricial, com a introdução dos conceitos de matriz de transição estocástica. Foi abordado, também, o conceito de estado absorvente, que pode estar associado a falha ou término do sistema. A partir desses conceitos foi introduzida a matriz fundamental, que obtém o número de intervalos de tempo necessário para o sistema entrar no estado absorvente. Por meio da álgebra das matrizes de transições estocásticas, foi apresentado um exemplo de aplicação. A contribuição desta tese foi introduzir algoritmos e códigos, para realização de simulações estocásticas em diversas situações, que poderão ser úteis para a tomada de decisão da gestão escolar, visando aumentar a eficiência dos cursos em oferta e diminuir a latência entre os dados oficiais e a tomada de decisão visando auxiliar a gestão escolar. Naturalmente, foram realizados testes diagnósticos e introduzidas a matriz de confusão, a curva ROC e o índice de Gini, para interpretação dos resultados e analisar seu impacto. Ao final, foram sugeridas ações e as respectivas consequências, baseadas nas simulações realizadas, para que a gestão escolar, por meio desses algoritmos, possa aumentar a eficiência dos cursos com os atuais recursos.

Palavras-chave: gestão escolar, cadeias de Markov, matriz fundamental, simulação estocástica, testes diagnósticos.

Abstract

In this thesis, after revisiting the Brazilian educational system and the ABNT NBR ISO 21001 standard (ISO 21001:2018)—establishing evidence-based decision-making as a pillar of academic management—we conducted an analysis based on data regarding assessments, dropout rates, and performance, for which Key Performance Indicators (KPIs) were defined. Focusing on Integrated High School (though the approach could be adapted for elementary, standard high school, or higher education), and proceeding from the premise that decision-making complexity is directly proportional to the volume of available supporting information, we incorporated the study of phenomena that transition from an initial state through a sequence of states—transitions occurring with a degree of uncertainty characterized by probability—while combining this with best practices for selecting the optimal courses of action. A review of Markov modeling and Markov chains was conducted, covering concepts such as steady states and system evolution over time—analyzed via tree diagrams or matrix algebra—and introducing the concept of the stochastic transition matrix. The concept of an absorbing state, which may correspond to system failure or termination, was also addressed. Building on these concepts, the fundamental matrix was introduced to determine the number of time intervals required for the system to reach the absorbing state. An application example was presented using stochastic transition matrix algebra. The thesis contributes by introducing algorithms and code for performing stochastic simulations across various scenarios; these tools can aid school administrators in decision-making by increasing the efficiency of offered courses and reducing the time lag between the availability of official data and the implementation of management decisions. Naturally, diagnostic tests were performed, and the confusion matrix, ROC curve, and Gini index were introduced to interpret the results and analyze their impact. Finally, based on the simulations, actions and their potential consequences were proposed to enable school administrators to use these algorithms to enhance course efficiency with existing resources.

Keywords: school management, Markov chains, fundamental matrix, stochastic simulation, diagnostic tests.

Lista de ilustrações

Figura 1: Evolução do índice de Satisfação Geral.....	14
Figura 2: Taxa de evasão por Período Letivo.....	15
Figura 3: Diagrama de Pareto.....	15
Figura 4: Taxa de aprovação em componentes críticos.....	16
Figura 5: Exemplo de cadeia de Markov, associada à Tabela 1	23
Figura 6: Diagrama de árvore da cadeia de Markov associada à Tabela 1.....	24
Figura 7: Progresso estudantil no ensino médio	33
Figura 8: Partição da matriz P de transições do sistema.....	36
Figura 9: Número de formados com e sem jubilamento.....	56
Figura 10: Análise de sensibilidade para testar taxas de evasão	58
Figura 11: Curva ROC e a área sob essa curva.....	66
Figura 12: Dados desbalanceados.....	77
Figura 13: Dados balanceados undersampling	79
Figura 14: Dados balanceados oversampling.....	80
Figura 15: 500 valores da distribuição normal $Z \sim N(0,1)$	87
Figura 16: Distribuições normais com mesma média.....	88
Figura 17: 500 valores da distribuição t de Student.....	88
Figura 18: 500 valores da distribuição de quiquadrado $\chi^2(5)$	89
Figura 19: 500 valores da distribuição F.....	89
Figura 20: Curva logística no intervalo $[-5,5]$	94
Figura 21: Curva ROC, área sob a curva (AUC) e índice de Gini no R.....	96

Lista de Tabelas

Tabela 1 – Estrutura para os Auditores	16
Tabela 2 – Probabilidades de ocorrer chuva (A) ou seca (B)	22
Tabela 3 – Probabilidades dos estados A e B.....	24
Tabela 4 – Probabilidades $P(A)$ e $P(B)$ segundo o tempo e os estados	23
Tabela 5 – Probabilidades de pertencer a um dos grupos de pai para filho.....	30
Tabela 6 – Probabilidades de pertencer a um dos grupos de pai para neto.....	30
Tabela 7 – Simulação do número de meninos em três nascimentos e 100 amostras.....	43
Tabela 8 – Simulação de resultados da matriz de transição.....	49
Tabela 9 – Exemplo com dados simulados – Método exato.....	52
Tabela 10 – Comparativa do destino final do Ingressante.....	60
Tabela 11 - Matriz de transição (P).....	61
Tabela 12 – Resultados possíveis de um teste de hipóteses.....	62
Tabela 13 – Possibilidades de classificação de eventos, segundo o <i>cutoff</i>	63
Tabela 14 – Testes diagnósticos segundo o índice de Gini.....	67
Tabela 15 – Resumo sobre a curva Roc e o índice de Gini.....	68
Tabela 16 – Sugestões para a Gestão Escolar.....	68
Tabela 17 – Funções correspondentes às principais distribuições de probabilidade.....	90
Tabela 18 – Matriz de confusão para cutoff igual a 0,5	96
Tabela 19 – Banco de dados – maximvaero.xlsx.....	97

Sumário

INTRODUÇÃO.....	12
CAPÍTULO I – MODELAGEM DE MARKOV	20
1.1 FUNDAMENTOS.....	20
1.2 MATRIZ DE TRANSIÇÃO ESTOCÁSTICA.....	25
1.2.1 Metodologia.....	25
1.2.2 Probabilidades limites dos estados.....	23
1.3 ESTADOS ABSORVENTES.....	27
1.4 CADEIAS DE MARKOV DE TEMPO CONTÍNUO.....	28
CAPÍTULO II – ESTUDO DE CASO: Ensino Médio Integrado.....	31
2.1 INTRODUÇÃO.....	31
2.2 OBJETIVOS E METODOLOGIA.....	32
2.2.1 Objetivos.....	32
2.2.2 Metodologia.....	33
2.2.3 Hipóteses do modelo.....	34
2.4. ANÁLISE POR MEIO DE MATRIZES DE TRANSIÇÕES ESTOCÁSTICAS.....	37
CAPÍTULO III – MONTE CARLO via CADEIAS DE MARKOV.....	41
3.1 INTRODUÇÃO.....	41
3.1.1 Teorema - Lei Fraca dos Grandes Números.....	42
3.1.2 Teorema do Limite Central.....	42
3.1.3 Programa 1- Simulação do número de meninos em 100 nascimentos... ..	43
3.2 SIMULAÇÃO ESTOCÁSTICA DE 10.000 ALUNOS DO EMI	44
3.2.1 Programa 2 - Simulação estocásticas de 10.000 alunos	45
3.2.2 Considerações sobre esse algoritmo.....	46
3.2.3 Programa 3 - Simulação de 10.000 alunos com código rápido e cursos seriados genéricos.....	46
3.2.4 Programa 4 - Simulação de 100.000 alunos com código mais rápido e cursos seriados genéricos.....	47
3.2.5 Considerações sobre esse algoritmo.....	48
3.2.6 Definição do modelo – Matriz de transição.....	48
3.2.7 Programa 5 - Simulação de 10.000 alunos.....	49
3.2.8 Discussão dos resultados desse último algoritmo.....	50
3.3 SIMULAÇÃO COM A ÁLGEBRA DAS MATRIZES.....	50
3.3.1 Tópicos de álgebra das matrizes.....	50
3.3.2 Implementação no R (Cálculo Exato)	51
3.3.3 PROGRAMA 6 - Matrizes de Transição de Probabilidades.....	51
3.4 INTRODUÇÃO DA REGRA DE JUBILAMENTO.....	52
3.4.1 Modelagem dos estados.....	53
3.4.2 Implementação no R (Método Exato com Jubilamento)	53
3.4.3 PROGRAMA 7 - Matriz de transição de probabilidades – Jubilamento....	53
3.4.4 Comparação dos resultados.....	54
3.4.5 Preparação dos dados.....	54
3.4.6 PROGRAMA 8 - Comparação dos resultados com ou sem jubilamento...	55
3.4.7 Análise do Gráfico.....	56
3.5 ANÁLISE DE SENSIBILIDADE.....	57
3.5.1 PROGRAMA 9 - Análise de sensibilidade.....	57
3.5.2 Visualização do Impacto.....	58
3.5.3 Resultados da análise.....	59
3.5.4 Resumo comparativo.....	60
3.6.4 Análise dos Indicadores.....	60

CAPÍTULO IV - TESTES DIAGNÓSTICOS.....	61
4.1 VALIDAÇÃO E DIAGNÓSTICOS.....	61
4.1.1 Definição dos estados.....	62
4.2 DIGNÓSTICO – CURVA ROC – ÍNDICE DE GINI.....	62
4.2.1 A curva ROC.....	62
4.2.2 Resultado esperado.....	64
4.2.3 Resumo do diagnóstico.....	65
4.5 SIMULAÇÕES DE CADEIAS DE MARKOV.....	65
4.5.1 Simulação de 1.000 alunos com taxas pré-fixadas.....	65
4.5.2 Simulação dos Resultados	66
4.5.3 Visualização da Curva ROC.....	66
4.5.4 Análise dos resultados.....	67
5. CONCLUSÕES E CONSIDERAÇÕES FINAIS.....	69
REFERÊNCIAS.....	71
APÊNDICE 1: Fundamentos de Probabilidade.....	74
APÊNDICE 2: Balanceamento de dados e códigos do RMarkdown para auxiliar na tomada de decisão baseada em evidências.....	76
APÊNDICE 3: Simulações de Variáveis Aleatórias.....	85
APÊNDICE 4: Construção da curva ROC, AUC e GINI no R.....	91

INTRODUÇÃO

“This interpretation of quantum mechanics was developed over about two years. Some people didn't like it very much because it involved talking about probabilities, not the certainties of classical mechanics. This, however, is characteristic of quantum mechanics. In general, calculations in quantum mechanics tell us the probability of certain variables having certain values”.¹

Quantum Mechanics: Probabilities Not Certainties.
Paul Adrien Maurice Dirac, 1975
1933 Nobel Prize in Physics

Muitos processos que ocorrem na natureza ou na sociedade podem ser estudados como se o fenômeno em estudo passasse de um estado inicial, para uma sequência de estados, onde a transição de cada estado para o seguinte ocorresse segundo determinada probabilidade².

A complexidade da tomada de decisões é diretamente proporcional à quantidade de informações disponíveis para fundamentá-las. Considerando essa premissa e o tipo de fenômeno acima citado, esta tese contribui como uma ferramenta para a gestão escolar, pelo menos em primeira aproximação, tomar suas decisões, aliando a teoria às boas práticas de como escolher suas ações, com o intuito de otimizar seus resultados. Dessa forma essa pesquisa caracteriza-se como exploratória e quantitativa.

Começaremos revisitando a Lei de Diretrizes e Bases da Educação Nacional (Brasil, 1996)³ que é o marco legal que organiza o sistema educacional brasileiro, fundamentada na Constituição Federal, e na qual há definição dos princípios, níveis, modalidades e carga horária mínima da educação básica e superior, com foco na formação para cidadania e trabalho.

A abrangência dessa Lei cobre desde a educação infantil até o ensino superior, incluindo educação de jovens e adultos, educação especial, profissional e indígena. Sua estrutura divide a educação escolar em Educação Básica (infantil, fundamental, médio) e Educação Superior. A matrícula é obrigatória para crianças e adolescentes de 4 a 17 anos. O calendário define no mínimo 800

¹ DIRAC, Paul Adrien Maurice. **Directions in Physics**: Lectures Delivered During a Visit to Australia and New Zealand. New York: John Wiley & Sons, 1978.

² Boldrini et al., 1986, p.14.

³ BRASIL. **Lei nº 9.394, de 20 de dezembro de 1996**. Estabelece as diretrizes e bases da educação nacional. Disponível em: planalto.gov.br. Acesso em: 4 mai. 2026.

horas anuais, distribuídas em pelo menos 200 dias de trabalho escolar efetivo (para fundamental e médio), tendo como princípios a igualdade de acesso e permanência, liberdade de aprender/ensinar, pluralismo de ideias, gestão democrática do ensino público, valorização do profissional da educação e os currículos devem ter base nacional comum, complementada por parte diversificada, incluindo estudos de história afro-brasileira e indígena. A Lei das Diretrizes e Bases da Educação Nacional também estipula a avaliação contínua e cumulativa do rendimento escolar, exigindo frequência mínima de 75% para aprovação.

Por outro lado, a norma ABNT NBR ISO 21001 (ISO 21001:2018) é a primeira norma internacional de sistemas de gestão estruturada especificamente para Organizações Educacionais. Ela adota a mesma Estrutura de Alto Nível da ISO 9001:2015, facilitando a integração, mas substitui o foco genérico em "clientes e produtos" por conceitos próprios do ambiente de ensino, como foco em alunos, outros beneficiários e serviços de aprendizagem. A seguir estão os pilares centrais e a estrutura dos requisitos auditáveis da norma aplicados às faculdades e escolas básicas:

- 1. Foco nos alunos e outros beneficiários:** Atender e superar as necessidades de aprendizado e desenvolvimento.
- 2. Liderança visionária:** Envolvimento da direção acadêmica para criar um ambiente de engajamento.
- 3. Engajamento das pessoas:** Valorização e capacitação constante do corpo docente e técnico-administrativo.
- 4. Abordagem de processo:** Compreender as atividades da instituição como fluxos interdependentes (ex: Secretaria → Coordenação → Sala de Aula).
- 5. Melhoria:** Foco contínuo na evolução dos métodos e instalações.
- 6. Tomada de decisão baseada em evidências:** Análises com base em dados de avaliações, evasão e desempenho.
- 7. Gestão de relacionamentos:** Sinergia com parceiros, empresas conveniadas, MEC e comunidade.
- 8. Responsabilidade social:** Prover serviços educacionais sustentáveis e acessíveis
- 9. Acessibilidade e equidade:** Atender à diversidade de alunos, garantindo inclusão e igualdade de oportunidades.
- 10. Conduta ética na educação:** Transparência, integridade e eliminação de conflitos de interesse.
- 11. Segurança e proteção de dados:** Garantir o sigilo dos prontuários e registros acadêmicos (alinhado à LGPD).

Pelo item 6, para medir e evidenciar o princípio da “Tomada de decisão baseada em evidências” em uma instituição de ensino, é preciso transformar os dados brutos de avaliações, evasão e desempenho em Indicadores **Chave de**

Performance (KPIs). A seguir está a estrutura de como monitorar, medir e gerar evidências para cada um desses três pilares:

1. Avaliações (Satisfação e Clima): Este pilar mede a percepção dos alunos, responsáveis e professores sobre a qualidade do serviço prestado.

Métricas e Indicadores Chave de Performance (KPIs):

- **Índice de Satisfação Geral (ISG):** Nota média obtida nas pesquisas de clima escolar ou nos relatórios da CPA (Comissão Própria de Avaliação).
- **Net Promoter Score (NPS) Educacional:** Mede o quanto alunos ou pais recomendariam a instituição ("Em uma escala de 0 a 10, o quanto você recomendaria nossa instituição?").
- **Taxa de Resposta:** Percentual de alunos/pais que efetivamente respondem às pesquisas (mostra o engajamento da comunidade)

Como comprovar a tomada de decisão (Evidência): Apresentar a Ata de Análise Crítica ou o Plano de Ação gerado após a pesquisa. *Exemplo:* Se a nota do laboratório de informática foi baixa, a evidência será o cronograma aprovado de investimento em novos computadores ou licenças de software.

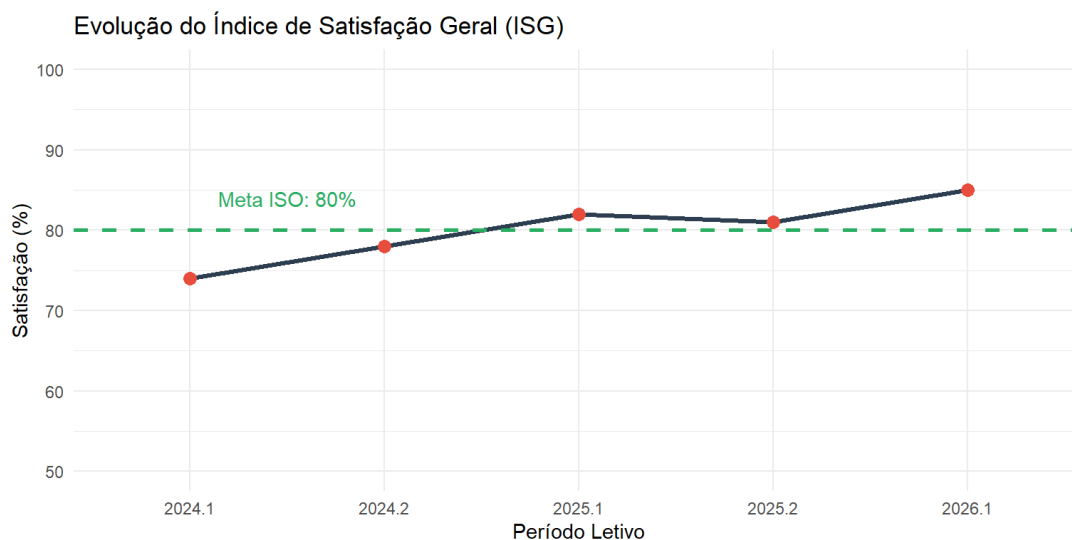


Figura 1 – ISG – Elaboração do autor – vide apêndice 2

2. Evasão Escolar / Acadêmica: Mede a capacidade da instituição de reter seus alunos e garantir a sustentabilidade do negócio e do ciclo educacional.

- **Indicadores Chave de Performance (KPIs):**
 - **Taxa de Evasão Mensal/Semestral:** Percentual de alunos que evadiram (trancamento, desistência ou transferência) em relação ao total de matriculados
 - **Índice de Rotatividade (Churn Rate):** Quantidade de alunos perdidos em um período específico.

- **Motivos de Evasão (Pareto):** Gráfico que categoriza as causas da saída (financeira, mudança de cidade, insatisfação pedagógica, dificuldade de aprendizado).

Como comprovar a tomada de decisão (Evidência): Demonstrar um Comitê de Retenção ou Relatórios do Sistema de Gestão Escolar. *Exemplo:* Dados apontaram evasão alta por motivos financeiros no 3º semestre; a decisão evidenciada foi a criação de um programa interno de parcelamento ou bolsas de estudo para reter esses estudantes.

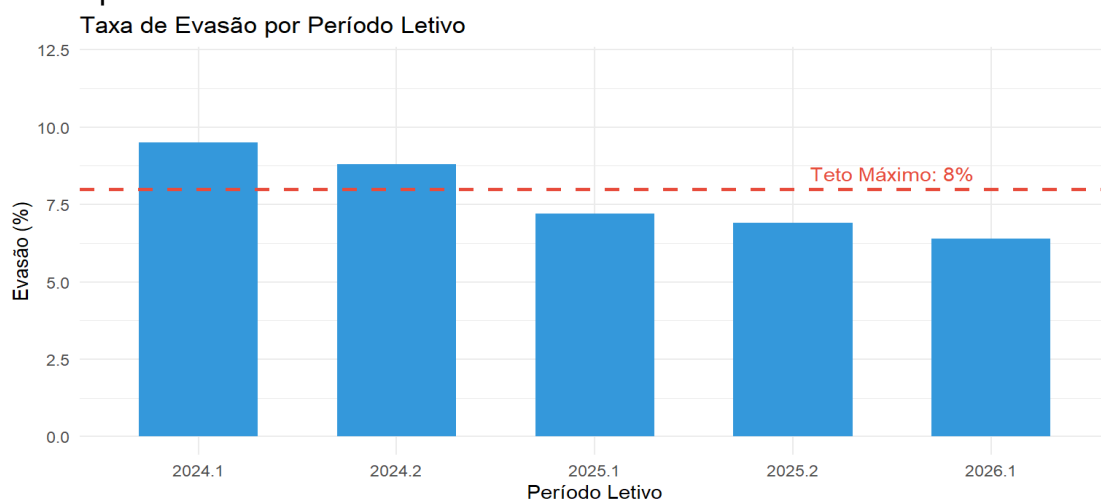


Figura 2 – TEPL – Elaboração do autor – 2026 - vide apêndice 2

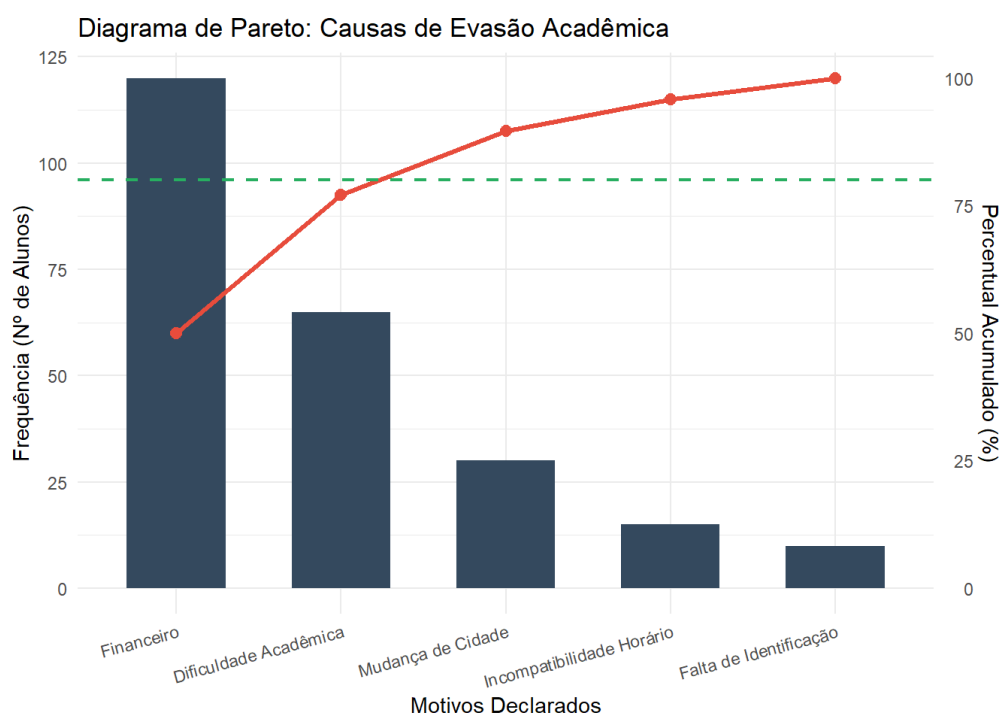


Figura 3 – Causas da evasão (Pareto) - Elaboração do autor – 2026 - vide apêndice 2

3. Desempenho (Acadêmico e Pedagógico): Mede a eficácia real do processo de ensino-aprendizagem.

- **Indicadores Chave de Performance (KPIs):**

- **Taxa de Aprovação por Disciplina/Série:** Percentual de alunos aprovados sem necessidade de recuperação final ou dependência (DP).
- **Índice de Alunos em Recuperação/Tutorias:** Volume de estudantes que necessitaram de suporte pedagógico extra.
- **Desempenho em Exames Externos:** Notas obtidas em simulados do Enem/vestibulares (Educação Básica) ou conceitos obtidos no Enade/MEC (Educação Superior).

Como comprovar a tomada de decisão (Evidência): Apresentar planos de intervenção pedagógica. *Exemplo:* O indicador apontou que 40% da turma reprovou em Cálculo I ou Matemática Básica; a evidência de decisão baseada em dados será a alteração da ementa, a inclusão de oficinas extras de nivelamento no início do semestre ou a troca da metodologia de ensino pelo colegiado.

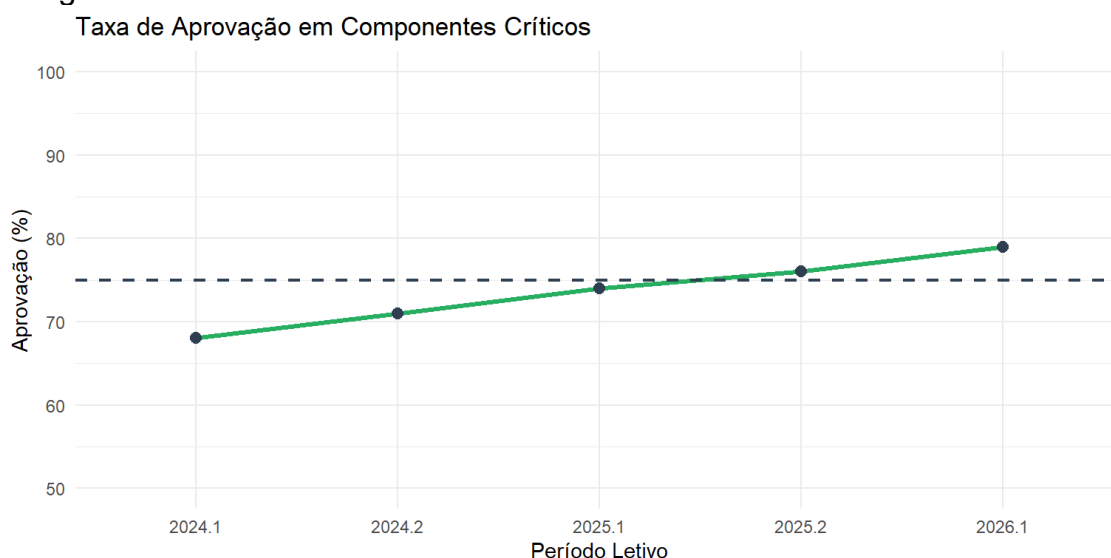


Figura 4 – TACC – Elaboração do autor – 2026 - vide apêndice 2

Tabela 1 - Estrutura para os Auditores

Para que o indicador seja aceito na auditoria de certificação, cada um deles deve possuir uma **Ficha de Indicador** contendo:

Campo	Exemplo Prático
Nome do Indicador	Taxa de Evasão de Alunos da Graduação
Objetivo	Manter a sustentabilidade financeira e acadêmica dos cursos
Fórmula de Cálculo	$(\text{Número de evadidos no semestre} / \text{Total de matriculados no início}) \times 100$
Frequência de Medição	Semestral
Meta	Menor que 8% ao ano
Responsável	Coordenador de Curso ou Diretor de Ensino

Nesta tese, o exemplo utilizado para apresentar as metodologias da álgebra linear e da simulação estocástica será a evolução do Ensino Médio Integrado (EMI) que funciona como um curso único em que o estudante cursa, simultaneamente e na mesma instituição, as disciplinas do ensino médio regular (Formação Geral Básica) e as disciplinas técnicas de um curso profissionalizante.

Ao final dos três anos, o aluno recebe um único diploma que atesta tanto a conclusão da educação básica quanto a habilitação profissional técnica. Diferente de outras modalidades, o estudante possui apenas uma matrícula para ambos os núcleos de ensino (geral e técnico).

Atualizada pela Lei 14.945/2024 (Brasil, 2024), a carga horária mínima da formação geral básica será de 2.100 (duas mil e cem) horas, admitindo-se que até 300 (trezentas) horas da carga horária da formação geral básica sejam destinadas ao aprofundamento de estudos de conteúdos da Base Nacional Comum Curricular diretamente relacionados à formação técnica profissional oferecida. Esse curso deverá totalizar, no mínimo, 3.000 horas ao longo dos três anos”.

Manutenção da possibilidade de arranjo curricular que garanta a constituição de um núcleo de integração/articulação entre a formação geral básica (BNCC e parte diversificada) e a formação profissional.

O itinerário de formação técnica e profissional (FTP) é um dos caminhos que o aluno pode escolher, focando na preparação para o mercado de trabalho por meio de vivências práticas e convênios com instituições credenciadas. Ao final, o diploma adquirido pelo aluno tem validade nacional e permite que o formado siga tanto para o mercado de trabalho como técnico quanto para o ensino superior.

Para a gestão escolar de uma Instituição na qual o EMI (ou outro curso) é ofertado, uma questão que se faz necessária a respeito desse curso é a que trata de como medir ou avaliar sua eficiência?

A resposta, em uma primeira aproximação, é ao final do curso computar quantos alunos, que entraram no sistema, se formaram e quantos alunos saíram do sistema, antes de se formarem ao longo dos três anos e, se esses números trazem alguma informação relevante, a ponto de sugerir medidas saneadoras para que a gestão escolar possa promover alguma melhora nesse índice. Nesse

sentido, usando as técnicas adequadas, uma das contribuições desta tese é responder à questão acima formulada e, ainda, sugerir atitudes e métodos adequados para otimizar a eficiência do Ensino Médio Integrado (ou do Ensino Fundamental ou Ensino Médio ou Ensino Superior), usando os recursos que já estão disponíveis, caso seja possível.

No Capítulo I, inicialmente, para compreensão das técnicas utilizadas para esse estudo, foi realizada revisão de uma das técnicas da Pesquisa Operacional, que é a Modelagem de Markov discreta, por meio de descrições e exemplos, para que se compreenda sua utilização.

No Capítulo II, por meio de Cadeias de Markov e Álgebra Linear foi apresentada metodologia já usada para analisar a eficiência de um curso de Ensino Médio Integrado de três anos.

A eficiência a que esta tese se refere é a diminuição da evasão, seja por decisão do estudante, seja por jubramento, tendo como consequência educacional o aumento do percentual de alunos que se formam no curso, em relação ao número de alunos que entram no curso, se possível sem aumento de custos.

Dessa forma, foi analisada a eficiência, de um curso de Ensino Médio Integrado de alguma rede de ensino, cujos dados relativos ao número de alunos foram obtidos do censo escolar de 2022 a 2024, com recortes específicos.

No Capítulo III, foi apresentado o Método de Monte Carlo, que abrange uma grande gama de métodos, para realização de simulações das cadeias de Markov, usando um software adequado, cujos códigos foram apresentados, podendo ser reproduzidos e utilizados. Nessas simulações, com grande número de alunos para que a gestão escolar possa medir ou avaliar a eficiência do curso de Ensino Médio Integrado estudado no capítulo anterior, em situações hipotéticas diversas, foram realizadas para a sugestão de medidas visando aumentar a eficiência não somente desse tipo de curso, mas de todos cursos da rede de ensino a qual o curso pertence, procurando manter os recursos disponíveis e, também, diminuir a latência entre os dados publicados e os dados que precisam ser utilizados para a tomada de decisão.

No Capítulo IV, foram apresentados métodos para os testes diagnósticos dos resultados das simulações do capítulo anterior, por meio da matriz de confusão, da curva ROC, da interpretação dos resultados e do impacto do índice

de Gini, visando a adequação das medidas que a gestão escolar poderá adotar tendo em vista a eficiência dos cursos em oferta.

No apêndice 1, apresentam-se alguns fundamentos da teoria de probabilidade com os conceitos de álgebra, σ -álgebra, espaços de medida, medida, espaço de probabilidade, medida de probabilidade, variáveis aleatórias e funções mensuráveis.

No apêndice 2, apresentam-se códigos de balanceamento de dados e códigos do RMarkdown para auxiliar na tomada de decisão baseada em evidências, apresentadas na Introdução.

No apêndice 3, apresentam-se conceitos de simulações de variáveis aleatórias e diversos exemplos no R.

No apêndice 4, há detalhamento da construção da curva ROC e do índice de Gini por meio de um exemplo.

CAPÍTULO 1 – MODELAGEM DE MARKOV

1.1 FUNDAMENTOS

It might seem superfluous to emphasize the word "probability" in a book dealing with the subject. Yet on the one hand, one used to hear such specious utterance as "probability is just a chapter of measure theory"; on the other hand, many still use probability as a front for certain types of analysis such combinatorial, functional and whatnot.⁴

Um primeiro conceito no estudo de probabilidade é de conjunto não vazio, constituído de todos os possíveis resultados de um determinado experimento que é chamado de *espaço amostral* desse experimento, indicado por Ω .

O espaço amostral será dito enumerável se for finito ou infinito enumerável (ou contável), isto é, se há uma correspondência biunívoca entre seus elementos e o conjunto dos números inteiros. Por outro lado, será infinito não enumerável (ou não contável) se for infinito e não houver correspondência biunívoca entre seus elementos e o conjunto dos números inteiros.

Por exemplo, no experimento que consiste no lançamento de dois dados idênticos, cada um com seis faces numeradas de 1 a 6, o espaço amostral é o conjunto $\Omega = \{(x, y): 1 \leq x \leq 6 \text{ e } 1 \leq y \leq 6\}$. Portanto esse espaço amostral é finito e, assim, enumerável.

Por outro lado, considere o experimento que consiste em observar o tempo (t) de reação a um determinado estímulo aplicado em certa pessoa. O espaço amostral é o conjunto dos números reais positivos, ou seja,

$$\Omega = \{t \in \mathbf{R}: t \geq 0\} = [0, \infty).$$

Por exemplo, a aplicação de um choque elétrico em uma pessoa, visando estimular determinado nervo (ciátio, por exemplo) durante um intervalo de 20 minutos. Nesse caso, o espaço amostral desse experimento é

$$\Omega = \{t \in \mathbf{R}: 0 \leq t \leq 20\} = [0, 20].$$

⁴ Chung, Kay Lay., 1968, Preface, p. v

Dessa forma, não há correspondência biunívoca entre os elementos desse conjunto e o conjunto dos números inteiros, logo é um espaço amostral infinito e não enumerável.

Por outro lado, dado um experimento, qualquer subconjunto A de seu espaço amostral $\Omega \neq \emptyset$ é chamado de *evento* ou *ocorrência* desse experimento. Nesse caso, se o resultado do experimento pertence ao conjunto A , dizemos que o evento A ocorreu.

Naturalmente, sendo A e B subconjuntos de Ω , podemos realizar com esses eventos as operações usuais da álgebra dos conjuntos, ou seja, inclusão, união, intersecção, complementação e produto cartesiano.

Uma família de subconjuntos de Ω é uma σ -álgebra (ou campo de Borel) denotada por $\mathcal{B}$, se satisfizer as propriedades seguintes:

- i) $\emptyset \in \mathcal{B}$, isto é, o conjunto vazio pertence ao conjunto $\mathcal{B}$;
- ii) se $A \in \mathcal{B}$, então $A^c \in \mathcal{B}$, ou seja, $\mathcal{B}$ é fechado por complementação;
- iii) se $A_1, A_2, A_3 \dots \in \mathcal{B}$, então $\bigcup_{i=1}^{\infty} A_i \in \mathcal{B}$, isto é, $\mathcal{B}$ é fechado por uniões enumeráveis.

Por exemplo, se $\Omega = \{1,2,3\}$, o conjunto das partes de Ω é uma σ -álgebra de subconjuntos de Ω .

De fato:

$$\mathcal{B} = \{\emptyset, \{1\}, \{2\}, \{3\}, \{1,2\}, \{1,3\}, \{2,3\}, \{1,2,3\}\}$$

satisfaz as propriedades (i), (ii) e (iii) da definição acima.

Considerando-se um espaço amostral $\Omega \neq \emptyset$ e uma σ - álgebra associada $\mathcal{B}$, uma medida de probabilidade⁵ P é uma função, cujo domínio é $\mathcal{B}$ que satisfaz:

- a) $P(A) \geq 0$, para todo $A \in \mathcal{B}$;
- b) $P(\Omega) = 1$;
- c) Se $A_1, A_2, \dots \in \mathcal{B}$ forem disjuntos dois a dois, então

$$P(\bigcup_{i=1}^{\infty} A_i) = \sum_{i=1}^{\infty} P(A_i).$$

Sejam E um experimento não determinístico e $\Omega \neq \emptyset$ o espaço amostral associado a esse experimento. Uma *variável aleatória* X é uma função que

⁵ Vide Apêndice 1, para uma definição, usada na teoria da medida.

associa a cada resultado do espaço amostral ($s \in \Omega$) um número real $X(s)$, mediante uma lei de formação, isto é, $X: \Omega \rightarrow \mathbf{R}$ tal que $s \rightarrow X(s)$.

Um *processo estocástico* é uma coleção de variáveis aleatórias. Isto é, para cada $t \in T$, onde T é um conjunto de índices, $X(t)$ é uma variável aleatória. Considere um processo estocástico $\{X_n | n = 0, 1, 2, \dots\}$ que assume um número finito ou infinito enumerável de possíveis valores. A menos de aviso em contrário, daqui em diante esse conjunto de possíveis valores desse processo será denotado pelo conjunto dos inteiros não negativos $\{0, 1, 2, 3, \dots\}$.

Se $X_n = i$, diz-se que o processo está no estado i no tempo n . Vamos supor que, sempre que o processo estiver no estado i , existe uma probabilidade fixa p_{ij} de que ele passe para o estado j em seguida. Se, além disso, tivermos:

$P[X_{n+1} = j | X_n = i, X_{n-1} = i_{n-1}, \dots, X_1 = i_1, X_0 = i_0] = P[X_{n+1} = j | X_n = i] = p_{ij}$ para todos os estados $i_0, i_1, \dots, i_{n-1}, i, j$ e $n \geq 0$, esse processo estocástico é chamado de *Cadeia de Markov*.

Nesse processo, dado um conjunto de estados, a probabilidade de migrar de um estado para outro é constante e independe do tempo passado ou futuro.

Por exemplo⁶, suponhamos que em determinada região, observa-se que se chover durante um ano, a probabilidade de que chova no ano seguinte é 0,5, e que a probabilidade de não chover no ano seguinte é a mesma. Por outro lado, se houver seca em um ano a probabilidade de que faça seca no ano seguinte é 0,75 e de fazer chuva é 0,25, de acordo com a Tabela 2, a seguir:

Tabela 2 – Probabilidades de ocorrer chuva (A) ou seca (B)

	A	B
A	0,50	0,50
B	0,25	0,75

Fonte: Elaboração do autor

Dessa forma, na Tabela 2, tem-se um sistema no qual dois estados estão identificados por A (chuva) e B (seca), assim como suas probabilidades de permanecer ou deixar cada estado em um intervalo de tempo (Δt), de modo que

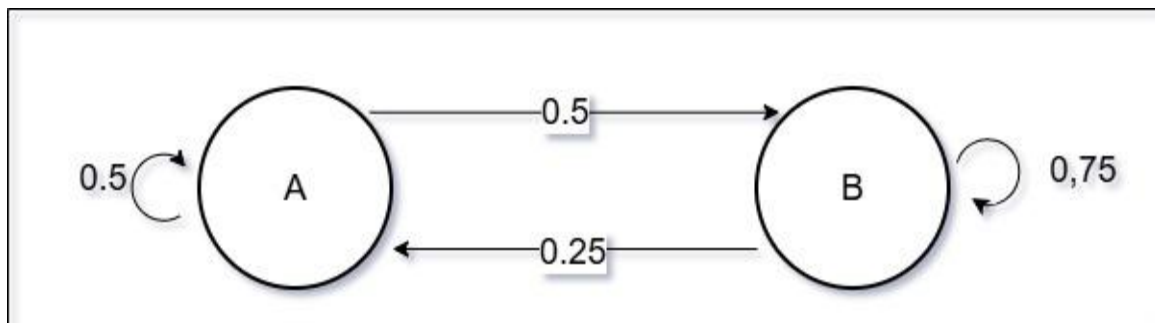
⁶ Boldrini et al, 1986, p.19.

estando em A, o sistema pode permanecer em A com probabilidade 0,5 ou pode migrar para B com a mesma probabilidade.

Analogamente, a probabilidade de o sistema permanecer em B é igual a 0,75 ou a probabilidade de B migrar para A é 0,25. Tais probabilidades são assumidas constantes para todos os intervalos de tempo (Δt) no futuro.

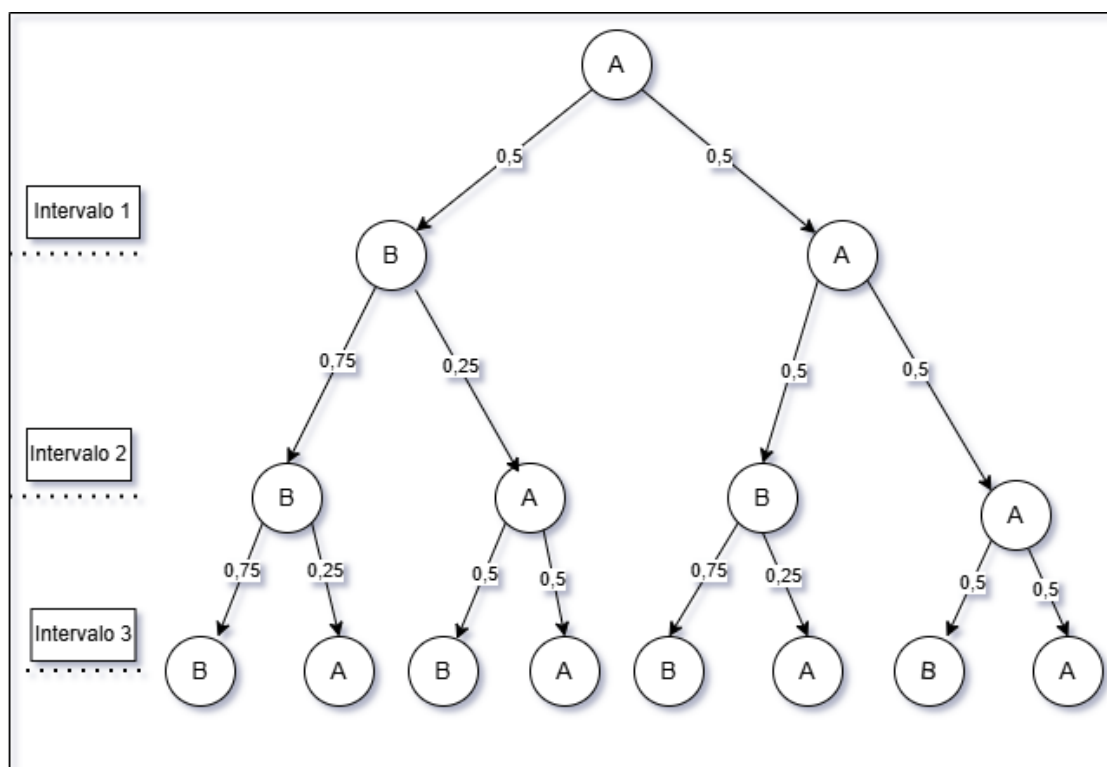
Ou seja, esse sistema é estacionário, pois dado um conjunto de estados a probabilidade de migrar de um estado para outro é constante e independe do tempo passado ou futuro. Tem-se então uma Cadeia de Markov, visto que o sistema é estacionário e o movimento entre os estados ocorre em passos discretos. Na Figura 5, apresentada a seguir, tem-se uma ilustração desse processo.

Figura 5: Exemplo de cadeia de Markov, associada à Tabela 2



Fonte: Elaboração do autor.

Nessa cadeia de Markov, em cada estado, a soma das probabilidades de residir em um estado ou migrar para outro(s) estado(s) deve ser igual a 1 (unidade). A evolução desse sistema no tempo pode ser determinada por um diagrama de árvore com início, por exemplo em A, como na figura 6 a seguir.

Figura 6: Diagrama de árvore da cadeia de Markov associada à Tabela 2

Fonte: Elaboração do autor.

Dessa árvore, pode-se concluir que, começando em A, depois de três intervalos de tempo, a probabilidade de o sistema estar no estado A é:

$$\frac{1}{2} \times \frac{3}{4} \times \frac{1}{4} + \frac{1}{2} \times \frac{1}{4} \times \frac{1}{2} + \frac{1}{2} \times \frac{1}{2} \times \frac{1}{4} + \frac{1}{2} \times \frac{1}{2} \times \frac{1}{2} = \frac{11}{32} = 0,344$$

Continuando a árvore até o intervalo de tempo 6, tem-se a Tabela 3, apresentada a seguir.

Tabela 3 – Probabilidades dos estados A e B

Intervalo de tempo	Iniciando em A		Iniciando em B	
	P(A)	P(B)	P(A)	P(B)
1	0,500	0,500	0,250	0,750
2	0,375	0,625	0,312	0,688
3	0,344	0,656	0,328	0,672
4	0,336	0,664	0,332	0,668
5	0,332	0,668	0,332	0,668
6	0,333	0,667	0,333	0,667

Fonte: Elaboração do autor.

Pode-se observar que as probabilidades de transição tendem para uma única distribuição de probabilidades (um estado estacionário). Além disso, que embora no período transitório as probabilidades são muito dependentes do valor inicial, os valores limites das probabilidades são totalmente independentes da condição inicial.

Na Tabela 3, observe que nos intervalos de tempo de 1 a 5, tem-se os *estados transitórios* e no intervalo de tempo 6, tem-se o *estado estacionário*. O sistema da Figura 5 é um exemplo de sistema ergódico, pois independentemente de se iniciar em A ou em B, no estado estacionário as probabilidades são as mesmas e, a partir desse estado, as probabilidades permanecem as mesmas.

Nesse exemplo, se o estado inicial for o estado A ou o estado B, o período transitório será diferente, mas os limites das probabilidades independem dos estados iniciais. Um sistema (ou processo) no qual os limites dos valores das probabilidades independem das condições iniciais é chamado de sistema (ou processo) ergódico.

Naturalmente, o diagrama de árvore torna-se inviável se a cadeia possuir um número muito grande de estados e/ou de intervalos de tempo. Nesse caso, utiliza-se outra metodologia de análise conhecida como *matriz de transição estocástica* que passaremos a estudar e seguir.

1.2 MATRIZ DE TRANSIÇÃO ESTOCÁSTICA

1.2.1 Metodologia

Com esse método é possível obter de maneira eficiente:

- a) a probabilidade absoluta do sistema ocupar os vários estados, após um dado número de transições;
- b) a probabilidade de encontrar a probabilidade limite dos estados.

Inicialmente, é necessário deduzir uma matriz que representa as probabilidades de se efetuarem as transições entre os estados do sistema em um dado intervalo de tempo, nas seguintes condições:

- i) P_{ij} é a probabilidade de transição para a estado j , depois de um intervalo de tempo, dado que o sistema estava inicialmente no estado i ;
- ii) $P_{i1} + P_{i2} + \dots + P_{in} = \sum_{j=1}^n P_{ij} = 1$, ou seja, a soma de qualquer linha dessa matriz deverá ser igual a 1 (unidade), onde

$$P = \begin{bmatrix} P_{11} & P_{12} & \dots & P_{1n} \\ P_{21} & P_{22} & \dots & P_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ P_{n1} & P_{n2} & \dots & P_{nn} \end{bmatrix}$$

No exemplo da Figura 1, a matriz de transição estocástica é

$$P = \begin{bmatrix} 1/2 & 1/2 \\ 1/4 & 3/4 \end{bmatrix}$$

ou seja, $P_{AA} = 0,5 = \frac{1}{2}$; $P_{AB} = 0,5 = \frac{1}{2}$; $P_{BA} = 0,25 = \frac{1}{4}$; $P_{BB} = 0,75 = \frac{3}{4}$

Sendo $E(0) = [1 \ 0]$ a condição inicial, que no exemplo começa no estado A, tem-se:

$$E(1) = E(0) \times P = [1 \ 0] \begin{bmatrix} 1/2 & 1/2 \\ 1/4 & 3/4 \end{bmatrix} = [1/2 \ 1/2]$$

$$E(2) = E(1) \times P = E(0)PP = E(0)P^2 = [1 \ 0] \begin{bmatrix} 3/8 & 5/8 \\ 5/16 & 11/16 \end{bmatrix} = [3/8 \ 5/8]$$

$$E(3) = E(2) \times P = E(0)P^2 \times P = E(0)P^3 = [1 \ 0] \begin{bmatrix} 11/32 & 21/32 \\ 21/64 & 43/64 \end{bmatrix} = \left[\frac{11}{32} \ \frac{21}{32} \right]$$

Pela Tabela 3, segue que esse resultado apresenta as probabilidades de transição do sistema, resumidos na Tabela 4 a seguir.

Tabela 4 – Probabilidades P(A) e P(B) segundo o tempo e os estados

Intervalo de tempo	Iniciando em A		
	Transições	P(A)	P(B)
1	E(1)	$\frac{1}{2}=0,500$	$\frac{1}{2}=0,500$
2	E(2)	$\frac{3}{8}=0,375$	$\frac{5}{8}=0,625$
3	E(3)	$\frac{11}{32}=0,344$	$\frac{21}{32}=0,656$

Fonte: Elaboração do autor

De modo geral, por indução, segue:

$$E(r) = E(0) \times P^r, r = 1, 2, 3, \dots$$

1.2.2 Probabilidades limites dos estados

Os valores limite ou estacionários das probabilidades de um sistema (processo) ergódico podem ser avaliados da seguinte forma:

$$E(1) = E(0) \times P$$

$$E(2) = E(1) \times P = E(0) \times P \times P = E(0) \times P^2$$

$$E(3) = E(2) \times P = E(0) \times P \times P \times P = E(0) \times P^3$$

.....

$$E(n) = E(n-1) \times P,$$

para n suficientemente grande, teremos

$$E(n) = E(n) \times P,$$

pois no limite, $E(n-1) = E(n)$, ou seja, a probabilidade não se altera.

Dessa forma, o princípio fundamental dessa técnica é que tendo sido determinada as probabilidades do estado limite, pela multiplicação de matrizes, qualquer multiplicação adicional não altera os valores limites das probabilidades.

Por exemplo, considerando-se um sistema de dois estados e sendo sua probabilidade limite $E(n) = [P_1 \ P_2]$, teremos:

$$[P_1 \ P_2] \begin{bmatrix} P_{11} & P_{12} \\ P_{21} & P_{22} \end{bmatrix} = [P_1 \ P_2].$$

Note que esse resultado independe da condição inicial, pois esse sistema é ergódico. Usando o fato de que $P_1 + P_2 = 1$ e efetuando a multiplicação acima, teremos o sistema

$$\begin{cases} (P_{11} - 1) \times P_1 + P_{21} \times P_2 = 0 \\ P_1 + P_2 = 1 \end{cases}$$

que é um sistema de duas equações a duas incógnitas, que na forma matricial é

$$\begin{bmatrix} P_1 & P_2 \end{bmatrix} \begin{bmatrix} P_{11} - 1 & 1 \\ P_{21} & 1 \end{bmatrix} = \begin{bmatrix} 0 & 1 \end{bmatrix}$$

Com relação ao sistema inicial do exemplo da Figura 5, teremos:

$$\begin{cases} (\frac{1}{2} - 1) \times P_1 + 1/4 \times P_2 = 0 \\ P_1 + P_2 = 1 \end{cases} \Rightarrow \begin{cases} \left(-\frac{1}{2}\right) \times P_1 + \left(\frac{1}{4}\right) \times P_2 = 0 \\ P_1 + P_2 = 1 \end{cases} \Rightarrow \begin{cases} P_1 = \frac{1}{3} \\ P_2 = \frac{2}{3} \end{cases}$$

Portanto,

$$\lim_{n \rightarrow \infty} E(n) = [1/3 \quad 2/3] \cong [0,333 \quad 0,667].$$

1.3 ESTADOS ABSORVENTES

Alguns estados do sistema podem ser absorventes, isto é, uma vez atingidos não é possível deixá-los até que se reinicie o sistema. Esses estados podem estar associados a falhas do sistema ou ao término do processo.

Se $\mathbf{P}$ é uma matriz de transição estocástica, uma matriz $\mathbf{Q}$ truncada pode ser obtida eliminando-se uma linha e uma coluna da matriz $\mathbf{P}$ associadas ao estado absorvente.

Por exemplo, se o estado 3 é absorvente, obtém-se a matriz $\mathbf{Q}$ eliminando-se a terceira linha e a terceira coluna de $\mathbf{P}$. Assim, se a matriz $\mathbf{P}$ tem ordem $n \times n$, a matriz $\mathbf{Q}$ terá ordem $(n - 1) \times (n - 1)$.

Por outro lado, sendo $\mathbf{X}$ uma variável aleatória que pode assumir os valores $(x_1, x_2, \dots, x_n)$ com probabilidades respectivas $(p_1, p_2, \dots, p_n)$, o valor esperado de $\mathbf{X}$ é dado por

$$E(\mathbf{X}) = \sum_{i=1}^n x_i p_i$$

que pode ser aplicado para uma matriz de múltiplas probabilidades $\mathbf{Q}$.

Sendo N o número esperado de intervalos de tempo:

$$N = 1 \times I + 1 \times Q + 1 \times Q^2 + \dots + 1 \times Q^{n-1}$$

De fato:

Sendo I é a matriz identidade de mesma ordem de Q nas condições iniciais do sistema temos:

$$N = 1 \times I$$

Decorrido mais 1 (um) intervalo de tempo, as probabilidades do sistema estão na matriz Q e tem-se

$$N = 1 \times I + 1 \times Q$$

e assim sucessivamente, por indução finita até obterem-se n intervalos de tempo.

Espera-se que após n intervalos de tempo, o sistema entre no estado absorvente. Por outro lado, sendo 0 a matriz nula e I a matriz identidade ambas de mesma ordem da matriz Q , tem-se

$$\lim_{n \rightarrow \infty} Q^n = 0 \Rightarrow \lim_{n \rightarrow \infty} [I - Q^n] = I \Rightarrow$$

portanto, para $n \rightarrow \infty$, teremos:

$$[I - Q][I + Q + Q^2 + \dots + Q^{n-1}] = I \Rightarrow$$

logo,

$$I + Q + Q^2 + \dots + Q^{n-1} = I \times [I - Q]^{-1} \Rightarrow$$

$$N = [I - Q]^{-1}.$$

Portanto, N é o número de intervalos de tempo decorridos até o sistema entrar no estado absorvente. Essa matriz N é chamada de *matriz de Leontief*⁷ ou *matriz fundamental*.

⁷ Wassily Wassilyovitch Leontief foi um economista russo, naturalizado estadunidense. Prêmio Nobel de Economia em 1973.

No exemplo da Figura 5, se considerarmos o estado B como absorvente, retirando-se a segunda linha e segunda coluna da matriz de transições estocástica $P = \begin{bmatrix} 1/2 & 1/2 \\ 1/4 & 3/4 \end{bmatrix}$ obtém-se

$$Q = [1/2] \Rightarrow N = [I - Q]^{-1} = \left[1 - \frac{1}{2}\right]^{-1} = 2.$$

Assim, esse sistema gastará em média 2 intervalos de tempo no estado A, antes de entrar no estado absorvente B.

Exemplo 2⁸. Numa pesquisa, procurou-se estabelecer correlação entre os níveis de escolaridade de pais e filhos. Sendo a letra F para os que concluíram o ensino fundamental, M para o ensino médio e S para o ensino superior, a probabilidade de um filho pertencer a um desses grupos, dependendo do grupo em que o pai está é dada pela Tabela a seguir.

Tabela 5 – Probabilidades de pertencer a um dos grupos de pai para filho

Pai p/Filho		F	M	S
$P =$	F	0,6667	0,3333	0,0000
	M	0,3333	0,3333	0,3333
	S	0,0000	0,3333	0,6667

Qual a probabilidade de um neto de um indivíduo que realizou o ensino médio ser do ensino superior?

Solução: Para o neto, teremos duas transições, a partir do pai, ou seja,

Tabela 6 – Probabilidades de pertencer a um dos grupos de pai para neto

pai p/Neto		F	M	S
$P^2 =$	F	0,5556	0,3333	0,1111
	M	0,3333	0,3333	0,3333
	S	0,1111	0,3333	0,5556

Resposta: $0,3333 = 1/3$

⁸ Boldrini et al, 1986, p.27

CAPÍTULO II – ESTUDO DE CASO: Ensino Médio Integrado

2.1 INTRODUÇÃO

Art. 36-A. Sem prejuízo do disposto na Seção IV deste Capítulo, o ensino médio, atendida a formação geral do educando, poderá prepará-lo para o exercício de profissões técnicas

Parágrafo único. A preparação geral para o trabalho e, facultativamente, a habilitação profissional poderá ser desenvolvida nos próprios estabelecimentos de ensino médio ou em cooperação com instituições especializadas em educação profissional.

Art. 36-B. A educação profissional técnica de nível médio será desenvolvida nas seguintes formas: I – articulada com o ensino médio; II – subsequente, em cursos destinados a quem já tenha concluído o ensino médio.⁹

Esse capítulo consiste no estudo da eficiência de um Ensino Médio Propedêutico, Curso Técnico Integrado ou Ensino Médio Normal/Magistério.

Naturalmente, uma das possíveis medidas do progresso estudantil no sistema da educação básica é o tempo esperado de permanência dos estudantes no sistema ou, seja, o percentual de alunos que completam o curso no período esperado.

Em particular, no ensino médio integrado a relevância desse tema consiste em verificar se o tempo esperado de permanência dos estudantes no sistema está próximo de três anos.

Na hipótese de ser maior do que três anos, haverá custo adicional ao sistema no curto prazo, pois isto significará aumento do número de alunos repetentes ou defasados e, conseqüentemente, aumento do número de professores, de turmas e outros insumos inerentes ao sistema.

Por outro lado, se o tempo esperado de permanência no sistema for menor do que três anos, esse resultado não significará alto desempenho do sistema, mas sim será um indicador do aumento da saída do sistema pelos alunos, que no longo prazo, também acarretará custo adicional, pois esses estudantes irão compor o contingente de alunos do ensino de jovens e adultos, devido às exigências atuais do mercado de trabalho.

⁹ BRASIL. **Lei nº 9.394, de 20 de dezembro de 1996**. Estabelece as diretrizes e bases da educação nacional. Disponível em: planalto.gov.br. Acesso em: 4 mai. 2026.

Portanto, o ideal é que o tempo de permanência dos alunos no sistema seja de três anos, pois isto não acarretaria custos adicionais. Este tempo ideal poderá significar economia para o sistema, e conseqüentemente para o governo, uma vez que não alteraria o valor esperado dos insumos.

Neste estudo, para realizar essa medida de tempo esperado de permanência dos estudantes no sistema, será utilizado o Ensino Médio Integrado, admitindo-se que o mesmo é constituído por três séries, de maneira que os alunos aprovados na 3ª série serão considerados os concluintes desta etapa.

As três séries em estudo, de 1ª a 3ª séries, mais a conclusão (F) e a evasão¹⁰ (E) serão considerados como estados e o conjunto destes estados formam um sistema, que será chamado sistema do ensino médio integrado, sendo que os estados conclusão (F) e evasão (E) serão considerados estados absorventes, e que se comporta da seguinte maneira:

- i) O aluno completa as três séries do ensino médio integrado, até concluir o ensino médio integrado de três anos;
- ii) O aluno sofre uma ou mais repetências durante o percurso, fazendo com que ele demore mais de três anos para concluir o ensino médio;
- iii) O aluno abandona o sistema (evade) em uma das três séries e, neste caso, será considerado que o mesmo não voltará mais ao sistema (hipótese simplificadora).

Este capítulo é composto desta introdução e mais duas seções. Na seção dois, apresentam-se os objetivos desse estudo e a metodologia usada para atingi-los. Os resultados e conclusões finais foram alocados na seção três.

2.2 OBJETIVOS E METODOLOGIA

2.2.1 Objetivos

- avaliar o tempo médio (anos) que o aluno permanece em uma determinada série;
- avaliar o tempo médio de permanência dos alunos no sistema;

¹⁰ Nesse texto, o termo **evasão** não significa mero abandono, mas sim saída deste sistema, podendo ser por transferência ou alguma outra conveniência familiar ou do aluno, caso seja menor de idade.

- obter as probabilidades de conclusão do ensino médio integrado e de saída do sistema antes de concluir o curso, dado o estado inicial de ingresso no sistema.

O tempo médio de permanência dos alunos no sistema tem como base um modelo de proposta para o progresso e a retenção no ensino médio integrado da 1ª à 3ª séries, apresentado a seguir na Figura 3, seguindo o modelo apresentado por Pereira e Silva (1986, p. 51).

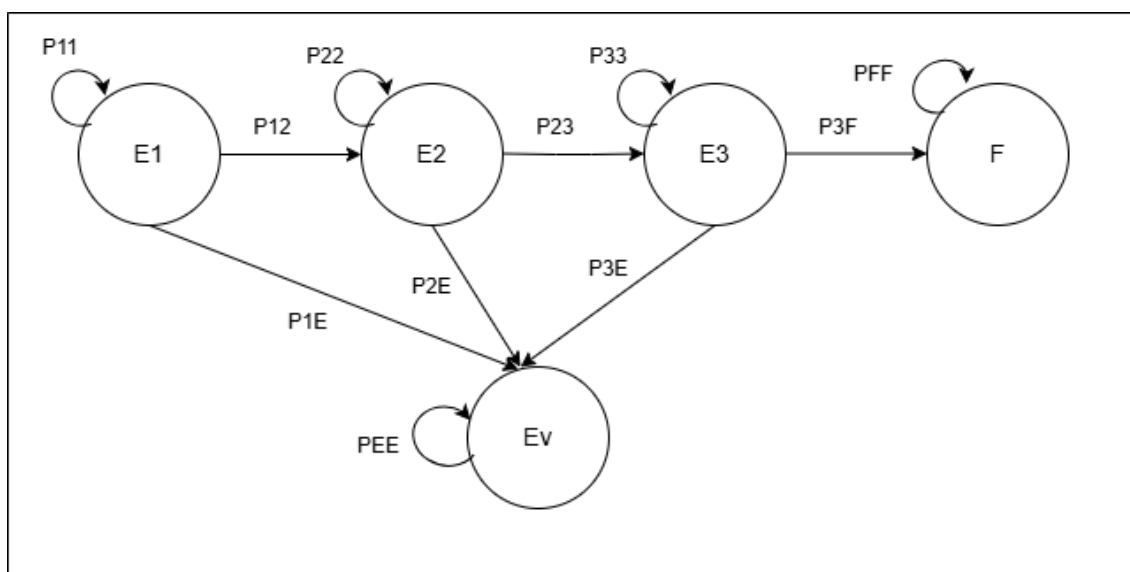
2.2.2 Metodologia

Os estados E1, E2 e E3 representam respectivamente a 1ª, 2ª e 3ª séries e, sucessivamente, o estado F representa os alunos que concluíram o ensino médio e o estado Ev representa os alunos que saíram do sistema antes de concluir o ensino médio (evasão do sistema).

Para o aluno cursar a série $j = i + 1$, deverá ter cursado a série i e ter sido nela aprovado.

Para estudo desse fenômeno, será adotado o modelo matemático discreto de um processo de Markov, que constitui uma subclasse dos processos estocásticos e que pode ser assim caracterizado, pela Figura 3, a seguir.

Figura 7 – Progresso estudantil no ensino médio integrado



Fonte: Elaboração do Autor

O estudante ingressa no ensino médio integrado na primeira série e é aprovado para a segunda série com probabilidade P_{12} , ou repete e cursa novamente a primeira série com probabilidade P_{11} ou, ainda, sai do sistema (evade) com probabilidade P_{1E} .

Se for aprovado para a segunda série, passa para a terceira com probabilidade P_{23} , ou repete a segunda série com probabilidade P_{22} ou, ainda, sai do sistema (evade) com probabilidade P_{2E} .

Se for aprovado para a terceira série, o aluno sai do sistema (evade) com probabilidade P_{3E} , ou repete a terceira série com probabilidade P_{33} , ou ainda, o aluno se forma com probabilidade P_{3F} .

Notações:

- P_{ii} é a proporção (probabilidade) de estudantes no estado i , no período t , que permanecem no estado i no período $t + 1$ (repetentes);
- P_{ij} é a proporção (probabilidade) de estudantes no estado i , no período t , que passam para o estado $j = i + 1$, no período $t + 1$ (aprovados);
- P_{EE} é a proporção (probabilidade) de estudantes no estado i , no período t , que saíram do sistema no período $t + 1$.
- P_{FF} é a proporção (probabilidade) de estudantes no estado 3, no período t , que se formaram no período $t + 1$.

2.2.3 Hipóteses do modelo

Admite-se que independentemente do estado passado $t - 1$, o estado futuro $t + 1$ depende exclusivamente do estado presente t . Ou seja, admite-se vale a propriedade de Markov:

$$1- P[X_{t+1} = j | X_t = j_1, X_{t-1} = j_2, \dots, X_0 = j_0] = P[X_{t+1} = j | X_t = j_1]$$

Além disso, por hipótese, admite-se que:

2- Dado que um determinado aluno saiu do sistema em uma das séries, ele não retornará à mesma, durante o período deste estudo, isto é, $P_{EE} = 1$, isto é, este estado é absorvente (hipótese simplificadora);

3- Dado que um determinado aluno concluiu o ensino médio, é considerado que o mesmo concluiu o sistema, isto é, $P_{FF} = 1$, isto é, este estado é absorvente.

Dessa forma, tem-se uma cadeia de Markov com estados absorventes.

Sendo P a matriz de transição de probabilidades associada ao sistema em estudo, nas condições estabelecidas, tem-se:

$$P = \begin{bmatrix} P_{11} & P_{12} & 0.0 & 0.0 & P_{1E} \\ 0.0 & P_{22} & P_{23} & 0.0 & P_{2E} \\ 0.0 & 0.0 & P_{33} & P_{3F} & P_{3E} \\ 0.0 & 0.0 & 0.0 & P_{FF} & 0.0 \\ 0.0 & 0.0 & 0.0 & 0.0 & P_{EE} \end{bmatrix}$$

onde

$$P_{11} + P_{12} + P_{1E} = 1$$

$$P_{22} + P_{23} + P_{2E} = 1$$

$$P_{33} + P_{3F} + P_{3E} = 1$$

Assim, P é uma matriz estocástica, tal que:

$$P_{ii} = \frac{RT_{i,t+i}}{MI_{i,t+i-1}}, \text{ para } i = 1,2,3$$

$$P_{ij} = \frac{AN_{j,t+i}}{MI_{i,t+i-1}}, \text{ para } i = 1,2 \text{ e } j = i + 1$$

$$P_{3F} = \frac{AP_{3,t+2}}{MI_{3,t+2}}$$

onde

$RT_{i,t+i}$ é o número de repetentes da série i , no ano $t + i$

$MI_{i,t+i-1}$ é o número de matriculados na série i no ano $t + i - 1$

$AN_{j,t+i}$ é o número de alunos novos na série $j = i + 1$, no ano $t + i$

$AP_{3,t+2}$ é o número de alunos aprovados na 3ª série, no ano $t + 2$

P_{iE} é calculado por diferença com 1 (um), para $i = 1,2,3$

$P_{FF} = 1$ (estado absorvente): pois o aluno que entra no estado F (concluiu) e dele não sai mais.

$P_{EE} = 1$ (estado absorvente): por hipótese, o aluno que entra no estado EE (saiu do sistema), ou seja, não retorna ao sistema no período em estudo.

Por outro lado, pode-se particionar a matriz P em quatro submatrizes, a saber:

Figura 8 – Partição da matriz P de transições do sistema

$P =$	P_{11}	P_{12}	0	0	P_{1E}
	0	P_{22}	P_{23}	0	P_{2E}
	0	0	P_{33}	P_{3F}	P_{3E}
	0	0	0	1	0
	0	0	0	0	1

Fonte: Elaborado pelo autor

Dessa partição, obtém-se, a submatriz Q (3×3) cujos elementos são as probabilidades de transição entre os estados transientes, que aqui estão representadas pelos estados 1, 2 e 3 (1ª série, 2ª série e 3ª série, respectivamente) e a submatriz R (3×2) cujos elementos são as probabilidades de transição entre os estados transientes e os estados absorventes, que estão representados pelos estados F (conclusão do ensino médio) e E (evasão do ensino médio), onde

Q é a submatriz 3×3 , cujos elementos são as probabilidades de transição entre os estados transientes:

$$Q = \begin{bmatrix} P_{11} & P_{12} & 0 \\ 0 & P_{22} & P_{23} \\ 0 & 0 & P_{33} \end{bmatrix}$$

R é a submatriz 3×2 , cujos elementos são as probabilidades de transição entre os estados transientes e estados absorventes:

$$R = \begin{bmatrix} 0 & P_{1E} \\ 0 & P_{2E} \\ P_{3F} & P_{3E} \end{bmatrix}$$

$\mathbf{0}$ é a submatriz nula 2×3 ;

$$\mathbf{0} = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix}$$

$\mathbf{I}$ é a matriz identidade 2×2 .

$$\mathbf{I} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$$

Dessa forma, para se obter o número médio de períodos (anos) em que o aluno permanece na série j , dado que iniciou na série i , ou seja, o tempo de permanência nos estados transientes, basta observar os elementos da matriz

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

onde $\mathbf{I}$ é a matriz identidade (3×3), de acordo com item 1.3 dessa tese.

Além disso, considerando-se a matriz $\mathbf{T}$ (3×1), cujos elementos resultam das somas das linhas da matriz $\mathbf{N}$, obtém-se os tempos esperados para a absorção, isto é, o tempo esperado para o aluno concluir o ensino médio ou evadir-se, dado que ingressou no sistema no estado i , para $i = 1, 2, 3$.

2.4. ANÁLISE POR MEIO DE MATRIZES DE TRANSIÇÕES ESTOCÁSTICAS

Como exemplo de aplicação no Ensino Médio Integrado da metodologia apresentada nesse capítulo, a fonte de dados foi o INEP – Censo Escolar da Educação Básica, Inepdata, com recorte na série histórica do ensino médio do Centro-Oeste, na categoria Federal, com indicadores de matrícula, aprovação, reprovação e abandono, e escola que não está em localização diferenciada, no período 2021-2024 de acordo com a consulta realizada em 04/2026, com adaptações, visando apenas exemplificar os cálculos.

a) para o primeiro ano:

$$P_{11} = \frac{R_{1,2022}}{MI_{1,2021}} = \frac{6060}{24638} = 0,245962$$

$$P_{12} = \frac{AN_{1,2022}}{MI_{1,2021}} = \frac{17184}{24638} = 0,697459$$

$$P_{1E} = 1 - 0,245962 - 0,697459 = 0,056579$$

b) para o segundo ano:

$$P_{22} = \frac{R_{2,2023}}{MI_{2,2022}} = \frac{1550}{23244} = 0,066684$$

$$P_{23} = \frac{AN_{2,2023}}{MI_{2,2022}} = \frac{21548}{23244} = 0,927035$$

$$P_{2E} = 1 - 0,066684 - 0,927035 = 0,006281$$

c) para o terceiro ano:

$$P_{31} = \frac{R_{3,2024}}{MI_{3,2023}} = \frac{1550}{24212} = 0,064018$$

$$P_{32} = \frac{AN_{3,2024}}{MI_{3,2023}} = \frac{21913}{24212} = 0,905047$$

$$P_{3E} = 1 - 0,064018 - 0,905047 = 0,030935$$

Portanto a matriz de transições estocásticas é

$$P = \begin{bmatrix} 0,245962 & 0,697459 & 0,000000 & 0,000000 & 0,056579 \\ 0,000000 & 0,066684 & 0,927035 & 0,000000 & 0,006281 \\ 0,000000 & 0,000000 & 0,064018 & 0,905047 & 0,030935 \\ 0,000000 & 0,000000 & 0,000000 & 1,000000 & 0,000000 \\ 0,000000 & 0,000000 & 0,000000 & 0,000000 & 1,000000 \end{bmatrix}$$

Usando-se a partição, têm-se:

$$Q = \begin{bmatrix} 0,245962 & 0,697459 & 0,000000 \\ 0,000000 & 0,066684 & 0,927035 \\ 0,000000 & 0,000000 & 0,064018 \end{bmatrix};$$

$$R = \begin{bmatrix} 0,000000 & 0,056579 \\ 0,000000 & 0,006281 \\ 0,905047 & 0,030935 \end{bmatrix};$$

$$\mathbf{0} = \begin{bmatrix} 0,000000 & 0,000000 & 0,000000 \\ 0,000000 & 0,000000 & 0,000000 \end{bmatrix} \mathbf{e}$$

$$I_2 = \begin{bmatrix} 1,000000 & 0,000000 \\ 0,000000 & 1,000000 \end{bmatrix}$$

A partir da partição, segue:

$$I_3 - Q = \begin{bmatrix} 0,754038 & -0,697459 & 0,000000 \\ 0,000000 & 0,933316 & -0,927035 \\ 0,000000 & 0,000000 & 0,935982 \end{bmatrix} \Rightarrow$$

$$N = (I_3 - Q)^{-1} = \begin{bmatrix} 1,326193 & 0,991053 & 0,981579 \\ 0,000000 & 1,071448 & 1,061207 \\ 0,000000 & 0,000000 & 1,068397 \end{bmatrix}$$

Como os elementos da matriz $N = (I - Q)^{-1}$ especificam os tempos médios de permanência nos estados transientes, isto é, a matriz N fornece o número médio de períodos (anos) em que o aluno permanece na série j, tendo iniciado na série i, a partir da observação dos elementos $N_{11} = 1,326193$; $N_{22} = 1,071448$ e $N_{33} = 1,068397$, pode-se concluir que:

- i) dado que o aluno entra na 1ª série é de se esperar que ele permaneça, em média, **1,326193** períodos (anos) nessa série, aproximadamente, 1 ano e 4 meses;
- ii) dado que o aluno que entra na 2ª série é de se esperar que ele permaneça, em média, **1,071448** períodos (anos) na 2ª série, aproximadamente, 1 ano e 1 mês; e,
- iii) dado que o aluno entra na 3ª série, é de se esperar que ele permaneça nessa série, em média, **1,068397** períodos (anos) na 3ª série, aproximadamente, 1 ano e 1 mês.

Além disso, a matriz T , dos tempos esperados para a conclusão ou evasão do ensino médio, dado o ingresso no estado i , para $i = 1, 2, 3$, é

$$T = \begin{bmatrix} 3,299302 \\ 2,132655 \\ 1,068397 \end{bmatrix} \cong \begin{bmatrix} \mathbf{3,29} \\ \mathbf{2,13} \\ \mathbf{1,07} \end{bmatrix}$$

Portanto, dado que um aluno entra na 1ª série, ele levará, em média, **3,29** períodos (anos), para concluir o ensino médio ou sair do sistema por evasão. O aluno que ingressa no sistema na 2ª série levará, em média, **2,13** períodos (anos) para concluir o ensino médio ou evadir-se. O que entra na 3ª série, levará, em média, **1,07** períodos (anos), para concluir o ensino médio ou evadir-se. Além disso, tem-se

$$N \times R = \begin{bmatrix} 0,888375 & 0,111625 \\ 0,960442 & 0,039558 \\ 0,966949 & 0,033051 \end{bmatrix}$$

onde a matriz $N \times R = (I - Q)^{-1} \times R$ fornece as probabilidades do estado estacionário interpretadas aqui, como probabilidades de conclusão ou evasão, dado o estado inicial de ingresso no sistema.

Portanto um aluno que entra na 1ª série tem, aproximadamente, **88,8%** de probabilidade de concluir o ensino médio, enquanto o que entra na 2ª série tem, aproximadamente, **96,0%** de probabilidade de concluir essa etapa. Finalmente, o aluno que entrou na 3ª série tem, aproximadamente, **96,5%** de probabilidade de concluir o ensino médio.

CAPÍTULO III – SIMULAÇÃO ESTOCÁSTICA

3.1 INTRODUÇÃO

Simulação [Do lat. *Simulatione*] 1. Ato ou efeito de simular. 2. Disfarce, fingimento, simulacro. 3. Hipocrisia, fingimento, impostura. 4. Reprodução ou representação de um processo, fenômeno ou sistema relativamente complexo, por meio de outro, ger. para fins científicos de observação, análise e predição ou para treinamento ou diversão etc. 5. Experiência ou ensaio realizado com modelos, esp. de modelos computacionais, relativos a processos ou objetos concretos que não podem ser submetidos a experimentação direta. 6. *Jur.* Declaração enganosa de vontade, com o objetivo de produzir efeito diferente que nele se indica. 7. *Piscol.* Imitação de uma perturbação somática ou psíquica, com fins utilitários. Ferreira, Aurélio Buarque de Holanda, 2004, p. 1849).

Esta introdução inicia-se por meio de uma consulta ao Novo Dicionário AURÉLIO da Língua Portuguesa de 2004, 3ª. edição, revista e ampliada. O sentido da palavra simulação neste trabalho é o de experiência ou ensaio realizado com modelos, especialmente de modelos computacionais, relativos a processos ou objetos concretos que não podem ser submetidos a experimentação direta.

Além disso, por simulação estocástica vamos denotar a arte de gerar amostras aleatórias em um ambiente computacional, cujo objetivo é usar essas amostras para a obtenção de certo resultado.

Segundo Bustos e Orgambide (1992), a palavra arte é aqui apropriada, porque precisamos lançar mão de uma classe enorme de resultados, que incluem, basicamente, teoria dos números, probabilidades, processos estocásticos, estatística, ciência da computação, análise numérica, além das particularidades do problema sob consideração, para atingir o objetivo com o máximo de qualidade e o mínimo de custo.

Para o uso das simulações realizadas nesse capítulo pode-se observar dois resultados a seguir, da teoria que as justificam e que serão apresentados sem as demonstrações, mas que naturalmente poderão ser obtidos nas referências que serão indicadas.

3.1.1 Teorema - Lei Fraca dos Grandes Números

Seja $X = (X_1, X_2, \dots, X_n, \dots)$ uma sequência de variáveis aleatórias independentes e identicamente distribuídas (iid) com média e variâncias finitas. Sejam $E[X] = m$ e $Var[X] = \sigma^2$ e $S_n = X_1 + X_2 + \dots + X_n$ então, para todo $\epsilon > 0$, tem-se

$$\lim_{n \rightarrow \infty} P \left[\left| \frac{S_n}{n} - m \right| > \epsilon \right] = 0.$$

Esse teorema afirma que uma variável aleatória X tal que $E[X] = m$, quando n cresce, a média de n realizações dessa variável (S_n/n) é uma *boa aproximação* de m , no sentido de que a probabilidade, que S_n/n e m tenham como diferença um valor maior do que $\epsilon > 0$, tende para zero em probabilidade, quando n tende a infinito. Para a demonstração, consulte Gnedenko, 1976, p.216. Na referência Leon-Garcia, 1994, p.277, há excelente discussão sobre a lei fraca dos grandes números.

3.1.2 Teorema do Limite Central

Seja $\beta > 0$, então

$$\lim_{n \rightarrow \infty} P \left[\left| \frac{S_n}{n} - m \right| < \beta \frac{\sigma}{\sqrt{n}} \right] = \frac{1}{\sqrt{2\pi}} \int_{-\beta}^{\beta} e^{-\frac{x^2}{2}} dx,$$

onde o RHS dessa igualdade é a área sob a curva normal padrão entre $-\beta$ e β . Por exemplo, se $\beta = 3$, a área da normal padrão é, aproximadamente, 0,997. Para $n = 100$, essa expressão no LHS afirma que ao estimar a média m por $S_{100}/100$ o erro cometido é menor do que $3 \frac{\sigma}{10}$ em 99,7% dos casos. A demonstração desse teorema, pode ser encontrada na referência Leon-Garcia, 1994, p.280.

Como exemplo de aplicação, vamos supor que em certa localidade a probabilidade de nascer um menino seja a mesma de nascer uma menina.

Seja $S = X_1 + X_2 + X_3$, onde

$$X_i = \begin{cases} 1, & \text{se ocorrer menino, com probabilidade } 0.5 \\ 0, & \text{se ocorre menina, com probabilidade } 0.5 \end{cases}, \text{ para } i = 1, 2, 3$$

Vamos simular a distribuição da variável aleatória $S = X_1 + X_2 + X_3$, em 100 nascimentos.

3.1.3 - Programa 1 - Simulação do número de meninos em 100 nascimentos

Preâmbulo

- set.seed(123) = semente
- var n_alunos, n_simulações, n_filhos, prob_menino
- simulação: rbinom(n_simulações, size=n_filhos,prob=prob_menino)
- resultados: table(factor(simulacao_S,levels:0:3))
- porcentagem: prop.table(resutados/100)

software R

```
# Definir a semente para reprodutibilidade
set.seed(123)

# Parâmetros
n_simulacoes <- 100
n_filhos <- 3
prob_menino <- 0.5

# Simulação: gera 100 grupos de 3 crianças (0 ou 1) e soma os meninos
# rbinom(n, size, prob)
simulacao_S <- rbinom(n_simulacoes, size = n_filhos, prob = prob_menino)

# Contar a frequência de cada resultado (0, 1, 2, 3 meninos)
resultados <- table(factor(simulacao_S, levels = 0:3))

print(resultados)

# Visualizar em porcentagem
prop.table(resultados/100)
```

Fonte: Elaboração do Autor.

RESULTADO

Tabela 7 – Simulação do número de meninos em três nascimentos

Número de Meninos (S)	Frequência Aproximada (em 100)	Probabilidade Teórica
0	11	0,125
1	42	0,375
2	33	0,375
3	14	0,125

Fonte: Elaboração do autor

Nota: Os resultados teóricos esperados para 100 simulações seriam aproximadamente 12.5, 37.5, 37.5, 12.5, com base nas probabilidades de 1/8, 3/8, 3/8, 1/8.

Com $n = 100$, obtemos:

$$P(S = 0) = 11/100 ; P(S = 1) = 42/100$$

$$P(S = 2) = 33/100 ; P(S = 3) = 14/100$$

Pela Lei dos Grandes Números e o Teorema do Limite Central, aumentando-se o número (n) de simulações os valores se aproximarão cada vez mais dos valores da distribuição teórica.

A seguir, serão apresentados os algoritmos que serão utilizados nas simulações das cadeias de Markov estudadas. O algoritmo Metropolis-Hastings funciona gerando uma sequência de valores de tal maneira que, à medida que mais e mais valores de amostra são produzidos, a distribuição de valores se aproxima mais da distribuição desejada $P(x)$.

Esses valores de amostra são produzidos iterativamente, com a distribuição da próxima amostra dependendo apenas do valor atual da amostra (transformando a sequência de amostras em uma cadeia de Markov). Especificamente, a cada iteração, o algoritmo escolhe um candidato para o próximo valor de amostra com base no valor atual da amostra. Então, com alguma probabilidade, o candidato é aceito ou rejeitado. Caso seja rejeitado, o valor do candidato é descartado e o valor atual é reutilizado na próxima iteração. A probabilidade de aceitação é determinada comparando os valores da função $f(x)$ dos valores de amostra atuais e candidatos em relação à distribuição desejada $P(x)$.

3.2 Simulação estocástica de 10.000 alunos do Ensino Médio Integrado

O programa a seguir simula 10.000 alunos do ensino médio integrado, distribuídos na cadeia de Markov nos estados “1º ano”, “2º ano”, “3º ano”, “Evadido¹¹” e “Formado”, de maneira aleatória, com sorteio de probabilidades e atribuição de cortes em 0,15 e 0,95.

¹¹ Neste contexto, “Evadido” significa aluno que saiu desse sistema, seja por transferência se for menor de idade, ou seja por transferência ou abandono se for maior de idade.

3.2.1 - Programa 2 - Simulação estocástica de 10.000 alunos

Preâmbulo

- a. `set.seed(123)` = semente
- b. `var n_alunos, n_simulações, sorteio, estado_atual`
- c. simulação: **`runif(1)`**: sorteia um número entre 0 e 1 ##distribuição uniforme
- d. estados: 1_Ano, 2_Ano, 3_Ano, Evadido, Formado
- e. resultado: `prop.table(table(resultados))`

Software R

```
set.seed(123) # Para resultados reproduzíveis
n_alunos <- 10000
resultados <- character(n_alunos)
for (i in 1:n_alunos) {
  estado_atual <- "1_Ano"
  while (estado_atual != "Formado" && estado_atual != "Evadido") {
    sorteio <- runif(1) # Sorteia um número entre 0 e 1
    if (estado_atual == "1_Ano") {
      if (sorteio < 0.15) estado_atual <- "1_Ano"
      else if (sorteio < 0.95) estado_atual <- "2_Ano"
      else estado_atual <- "Evadido"
    } else if (estado_atual == "2_Ano") {
      if (sorteio < 0.15) estado_atual <- "2_Ano"
      else if (sorteio < 0.95) estado_atual <- "3_Ano"
      else estado_atual <- "Evadido"
    } else if (estado_atual == "3_Ano") {
      if (sorteio < 0.15) estado_atual <- "3_Ano"
      else if (sorteio < 0.95) estado_atual <- "Formado"
      else estado_atual <- "Evadido"
    }
  }
  resultados[i] <- estado_atual
}

# Cálculo da probabilidade final
prop.table(table(resultados))
```

Resultados

Evadido	Formado
0.1637	0.8363

Fonte: Elaboração do autor.

Como resultado desse algoritmo, foram obtidos 16,37% de alunos evadidos e 83,63% de alunos que se formaram, que são estados absorventes desse sistema.

3.2.2 Considerações sobre esse algoritmo

Esse tipo de algoritmo é muito útil em aplicações Bayesianas, onde a distribuição posterior é proporcional à distribuição a priori multiplicada pela verossimilhança, mas com constante de proporcionalidade frequentemente desconhecida. Portanto, é particularmente útil para amostrar distribuições posteriores e realizar cálculos Bayesianos analiticamente intratáveis e, além disso, é um algoritmo muito simples para produzir amostras de distribuições que, de outra forma, seriam difíceis de amostrar. A seguir, seguem simulações com probabilidades hipotéticas para ilustrar o processo.

3.2.3 - Programa 3 – Simulação de 10.000 alunos com código rápido e com cursos seriados genéricos

Preâmbulo

- a. `set.seed(123)` = semente
- b. `var X, n_alunos, n_simulações, sorteio, estado_atual`
- c. simulação: `runif(1)`: sorteia um número entre 0 e 1 ##distribuição uniforme
- d. estados: 1_Ano, 2_Ano, 3_Ano, Evadido, Formado
- e. resultado: `prop.table(table(resultados))`

Software R

```
X <- 3 # Altere aqui para qualquer número inteiro de 2 e 10
if (X < 2 || X > 10) {
  stop("O número de semestres (X) deve ser maior que 1 e menor que 11.")
}
n_alunos <- 10000 # pode ser alterado pelo usuário
resultados <- character(n_alunos)
for (i in 1:n_alunos) {
  semestre_atual <- 1 # Começa no 1º semestre
  evadido <- FALSE # Flag de evasão
  # Enquanto não ultrapassou o último semestre e não foi evadido
  while (semestre_atual <= X && !evadido) {
    sorteio <- runif(1) # Sorteia um número entre 0 e 1
    if (sorteio < 0.15) {
      # 15% de chance: REPETE o semestre (permanece no mesmo)
      semestre_atual <- semestre_atual
    } else if (sorteio < 0.95) {
```

```

# 80% de chance: AVANÇA para o próximo semestre
semestre_atual <- semestre_atual + 1
  } else {
# 5% de chance: EVADE
evadido <- TRUE
  }
}
# Define o resultado final
if (evadido) {
  resultados[i] <- "Evadido"
} else {
  resultados[i] <- "Formado"
}
}
# Exibe as proporções finais
prop.table(table(resultados))
Resultados
Evadido Formado
0.1739  0.8261

```

Fonte: Elaboração do autor

3.2.4 - Programa 4 - Simulação de 100.000 alunos com código mais rápido ainda e com cursos seriados genéricos

Preâmbulo

- set.seed(42) = semente
- var n_alunos, n_simulações, sorteio_destino, estado_atual
- simulação:
matrix(runif(n_alunos * X), nrow = n_alunos, ncol = X) ## Faz todos os sorteios de uma vez
- estados: 1_Ano, 2_Ano, 3_Ano, Evadido, Formado
- resultado: prop.table(table(resultados))

Software R – Código rápido mais simples e mais geral (vetores e matrizes)

```

X <- 3 # Altere aqui para qualquer número inteiro de 2 e 10

if (X < 2 || X > 10) {
  stop("O número de semestres (X) deve ser maior que 1 e menor que 11.")
}

set.seed(42)
n_alunos <- 100000

```

```

# 1. Probabilidades condicionadas ao fato de que o aluno mudou de estado
# Se ele não repete (85% das vezes), qual a chance de evadir ou avançar?
p_ao_repetir <- 0.85
p_avancar_cond <- 0.80 / p_ao_repetir # ~94.1%
p_evadir_cond <- 0.05 / p_ao_repetir # ~5.9%

# 2. Simular as escolhas para todos os alunos em todos os X semestres
#### possíveis
# Sorteamos matrizes completas de uma vez só
sorteios_destino <- matrix(runif(n_alunos * X), nrow = n_alunos, ncol = X)

# Se o sorteio for menor que p_evadir_cond, o aluno evadiria naquele ##
semestre
evadiu_no_semestre <- sorteios_destino < p_evadir_cond

# 3. Determinar a dinâmica de sobrevivência do aluno
# Encontra o primeiro semestre onde ocorreu uma evasão
primeira_evasao <- max.col(evadiu_no_semestre, ties.method = "first")
houve_evasao <- rowSums(evadiu_no_semestre) > 0

# Se o aluno evadiu antes ou no semestre X, ele é considerado "Evadido"
# Caso contrário, ele conseguiu passar por todos os X semestres e está
"Formado"
resultados <- ifelse(houve_evasao, "Evadido", "Formado")

# Exibe as proporções finais instantaneamente
prop.table(table(resultados))

Resultados
Evadido Formado
0.16748 0.83252

```

Fonte: Elaboração do autor

3.2.5 Considerações sobre esse último algoritmo

Esse algoritmo generaliza o anterior e é muito mais rápido por usar vetores e matrizes para as simulações.

3.2.6 Definição do modelo – Matriz de transição

Vamos assumir as seguintes probabilidades anuais:

- Aprovação: 80% (passa para o próximo ano ou formam-se).
- Reprovação: 15% (permanecem no mesmo ano).
- Evasão: 5% (saem desse sistema).

Tabela 8 – Simulação de resultados da matriz de transição

De /Para	1º ano	2º ano	3º ano	Formado	Evadido
1º ano	0,15	0,80	0,00	0,00	0,05
2º ano	0,00	0,15	0,80	0,00	0,05
3º ano	0,00	0,00	0,15	0,80	0,05
Formado	0,00	0,00	0,00	1,00	0,00
Evadido	0,00	0,00	0,00	0,00	1,00

Fonte: Elaboração do autor

3.2.7 - Programa 5 - Simulação estocástica de 10.000 alunos

Preâmbulo

- set.seed(123) = semente
- var n_alunos, n_simulações, sorteio, estado_atual
- simulação: **runif(1)**: sorteia um número entre 0 e 1 ##distribuição uniforme
- estados: 1_Ano, 2_Ano, 3_Ano, Evadido, Formado
- resultado: prop.table(table(resultados))

Software R

```
set.seed(123) # Para resultados reproduzíveis
n_alunos <- 100000
resultados <- character(n_alunos)
for (i in 1:n_alunos) {
  estado_atual <- "1_Ano"
  while (estado_atual != "Formado" && estado_atual != "Evadido") {
    sorteio <- runif(1) # Sorteia um número entre 0 e 1
    if (estado_atual == "1_Ano") {
      if (sorteio < 0.15) estado_atual <- "1_Ano"
      else if (sorteio < 0.80) estado_atual <- "2_Ano"
      else estado_atual <- "Evadido"
    } else if (estado_atual == "2_Ano") {
      if (sorteio < 0.15) estado_atual <- "2_Ano"
      else if (sorteio < 0.80) estado_atual <- "3_Ano"
      else estado_atual <- "Evadido"
    } else if (estado_atual == "3_Ano") {
      if (sorteio < 0.15) estado_atual <- "3_Ano"
      else if (sorteio < 0.80) estado_atual <- "Formado"
      else estado_atual <- "Evadido"
    }
  }
  resultados[i] <- estado_atual
}
```

```
# Cálculo da probabilidade final
prop.table(table(resultados))
```

Resultado**Evadido Formado****0.5539 0.4461****Fonte:** Elaboração do Autor.

Como resultado desse algoritmo, foram obtidos 55,39% de alunos evadidos e 44,61% de alunos que se formaram, que são estados absorventes desse sistema.

3.2.8 Discussão dos resultados desse último algoritmo

Embora este problema específico possa ser resolvido com álgebra linear (usando a matriz fundamental (Leontief) de uma Cadeia de Markov absorvente), a simulação permite:

1. **Flexibilidade:** Adicionar variáveis complexas, como um aluno que só pode repetir 2 vezes antes de ser jubilado obrigatoriamente.
2. **Incerteza:** Integrar distribuições de probabilidade para as taxas de aprovação, em vez de valores fixos (Incerteza Bayesiana).

3.3 SIMULAÇÃO COM A ÁLGEBRA DAS MATRIZES

3.3.1 Tópicos de álgebra das matrizes

No Capítulo II, verificou-se que uma matriz de transição com estados absorventes pode ser escrita na forma canônica:

$$P = \begin{bmatrix} Q & R \\ 0 & I \end{bmatrix}$$

onde

- **Q:** é a matriz constituída pelas probabilidades de transição entre estados transitórios (1º, 2º e 3º ano).

- **R**: é a matriz das probabilidades de transição dos estados transitórios para os estados absorventes (Formado/Evadido).
 - **I**: é a matriz identidade (estados absorventes permanecem neles mesmos).
- A probabilidade exata de absorção é dada por:

$$B = N \times R = (I - Q)^{-1} \times R$$

onde $N = (I - Q)^{-1}$ é a matriz de Leontief ou matriz fundamental.

O valor na primeira linha e primeira coluna de B é a probabilidade exata de um aluno que entra no 1º ano sair como "Formado". O valor da segunda linha e primeira coluna de B é a probabilidade exata de um aluno que entra no 2º ano sair como "Formado" e, analogamente, o valor da terceira linha primeira coluna é a probabilidade exata de um aluno do 3º ano sair "Formado". O valor da primeira linha e segunda coluna de B é a probabilidade exata de um aluno que entra no 1º ano sair "Evadido". O valor da segunda linha e segunda coluna de B é a probabilidade exata de um aluno que entra no 2º ano sair "Evadido" e, analogamente, o valor da terceira linha e segunda coluna de B é a probabilidade exata de um aluno que entra no 3º ano sair "Evadido".

3.3.2 Implementação no R (Método Exato sem Jubilamento)

Utilizando as mesmas probabilidades do exemplo anterior: 80% de aprovação, 15% de retenção e 5% de evasão, têm-se:

3.3.3 – Programa 6 - Matrizes de Transição de Probabilidades

```
# 1. Definir a submatriz Q: Transitórios: 1ano, 2ano, 3ano
Q <- matrix(c(0.15, 0.80, 0.00, # De 1º para 1º, 2º, 3º
              0.00, 0.15, 0.80, # De 2º para 1º, 2º, 3º
              0.00, 0.00, 0.15), # De 3º para 1º, 2º, 3º
            nrow = 3, byrow = TRUE)

Q

# 2. Definir a submatriz R: Transitórios para Absorventes: Formado, Evadido
R <- matrix(c(0.00, 0.05, # 1º ano pode evadir
              0.00, 0.05, # 2º ano pode evadir
              0.80, 0.05), # 3º ano pode formar ou evadir
            nrow = 3, byrow = TRUE)

R

# 3. Calcular a matriz de Leontief:  $N = (I - Q)^{-1}$ 
#onde
I <- diag(3)

I
```

```

N <- solve(I - Q) ## calcula a inversa
N
# 4. Calcular a Matriz de Absorção B
B <- N%*%R
B
Resultado
      Formado   Evadido
Entra no 1º 0.8337065 0.16629351
Entra no 2º 0.8858131 0.11418685
Entra no 3º 0.9411765 0.05882353

```

Fonte: Elaboração do autor.

Tabela 9 – Exemplo com dados simulados – Método exato

Ponto de Entrada	Probabilidade de formar	Probabilidade de evadir
1º ano	83,37%	16,63%
2º ano	88,58%	11,42%
3º ano	94,12%	5,88%

Fonte: Elaboração do autor

Discussão:

A simulação estocástica é uma **aproximação estatística**. Quanto mais simulacros de alunos rodar (ex: 1 milhão em vez de 10 mil), mais o resultado do se aproximará dos **83,3738...%** calculados pela álgebra de matrizes.

Conclusões:

- A Álgebra de Matrizes entrega o valor real instantaneamente para modelos simples.
- A simulação estocástica é preferível quando as regras são complexas (ex: "se o aluno reprovar 2 vezes seguidas, a chance de evasão triplica"), situações onde a matriz deixaria de ser constante.

3.4 INTRODUÇÃO DA REGRA DE JUBILAMENTO

Para incluir uma regra de jubramento (ex: o aluno pode repetir apenas uma vez cada ano; na segunda reprovação, ele é desligado), a lógica de matrizes precisa ser expandida. O estado "1º Ano" deve ser dividido em "1º Ano (1ª vez)" e "1º Ano (Repetente)". Se o aluno que repetir a primeira vez voltar a repetir a mesma série novamente, ele cai no estado absorvente de Jubilado/Evadido.

3.4.1 Modelagem dos estados

Para um limite de **uma reprovação por série**, teremos os seguintes estados transitórios:

- : 1º ano (primeira vez e repetente).
- : 2º ano (primeira vez e repetente).
- : 3º ano (primeira vez e repetente).

3.4.2 Implementação no R (Método Exato com Jubilamento)

Serão usadas as mesmas taxas: 80% aprovação, 15% reprovação e 5% evasão.

3.4.3– Programa 7 - Matrizes de Transição de Probabilidades - Jubilamento

```
# Estados: E1_1, E1_2, E2_1, E2_2, E3_1, E3_2
Q_jub <- matrix(0, nrow = 6, ncol = 6)
# Transições do 1º Ano
Q_jub[1, 2] <- 0.15 # E1_1 -> E1_2 (Reprova 1ª vez)
Q_jub[1, 3] <- 0.80 # E1_1 -> E2_1 (Passa)
Q_jub[2, 3] <- 0.80 # E1_2 -> E2_1 (Repetente passa)
# Transições do 2º Ano
Q_jub[3, 4] <- 0.15 # E2_1 -> E2_2 (Reprova 1ª vez)
Q_jub[3, 5] <- 0.80 # E2_1 -> E3_1 (Passa)
Q_jub[4, 5] <- 0.80 # E2_2 -> E3_1 (Repetente passa)
# Transições do 3º Ano
Q_jub[5, 6] <- 0.15 # E3_1 -> E3_2 (Reprova 1ª vez)
# Matriz R (Saída para estados absorventes: Formado e Jubilado/Evadido)
R_jub <- matrix(0, nrow = 6, ncol = 2)
# Evasão padrão (5%) + Jubilamento (quem é repetente e não passa nem repete
# mais)
R_jub[1, 2] <- 0.05 # E1_1 evade
R_jub[2, 2] <- 0.05+0.15 # E1_2 evade ou é jubilado (falhou na 2ª chance)
R_jub[3, 2] <- 0.05 # E2_1 evade
R_jub[4, 2] <- 0.05+0.15 # E2_2 evade ou é jubilado
R_jub[5, 2] <- 0.05 # E3_1 evade
R_jub[5, 1] <- 0.80 # E3_1 forma
R_jub[6, 2] <- 0.05+0.15 # E3_2 evade ou é jubilado
R_jub[6, 1] <- 0.80 # E3_2 forma
# Cálculo da Matriz de Absorção
N_jub <- solve(diag(6) - Q_jub)
B_jub <- N_jub %*% R_jub
rownames(B_jub) <- c("1-ano-1v", "1-ano-2v", "2-ano-1v", "2-ano-2v", "3-ano-1v", "3-ano-2v")
colnames(B_jub) <- c("Formado", "Evadido/Jubilado")
print(round(B_jub, 4))
```

Fonte: Elaboração do autor

3.4.4 Comparação dos resultados

Ao introduzir o limite de reprovação, a probabilidade de sucesso cai:

- **Sem Jubilamento:** 83,4% de chance de formar (o aluno pode repetir infinitamente).
- **Com Jubilamento (1 repetição):** A probabilidade cai para cerca de **74,2%**.

A diferença no percentual de formados é devida a:

- **Risco Acumulado:** O aluno agora tem "vidas" limitadas.
- **Estados de Saída:** O estado "Evadido" agora engloba tanto a desistência voluntária quanto o desligamento forçado pela escola.

A seguir, apresenta-se o código para gerar essa comparação visual:

3.4.5 Preparação dos dados

Primeiro reunimos os resultados que calculamos anteriormente (aproximadamente 83,3% para o modelo simples e 74,2% para o modelo com jubilamento).

#Criando o conjunto de dados para o gráfico

```
dados_comparacao <- data.frame(
  Cenario = c("Sem Limite de Reprovação", "Com Jubilamento (1 Rep.)"),
  Probabilidade = c(0.8337, 0.7424) # Valores calculados via Matrizes
)
```

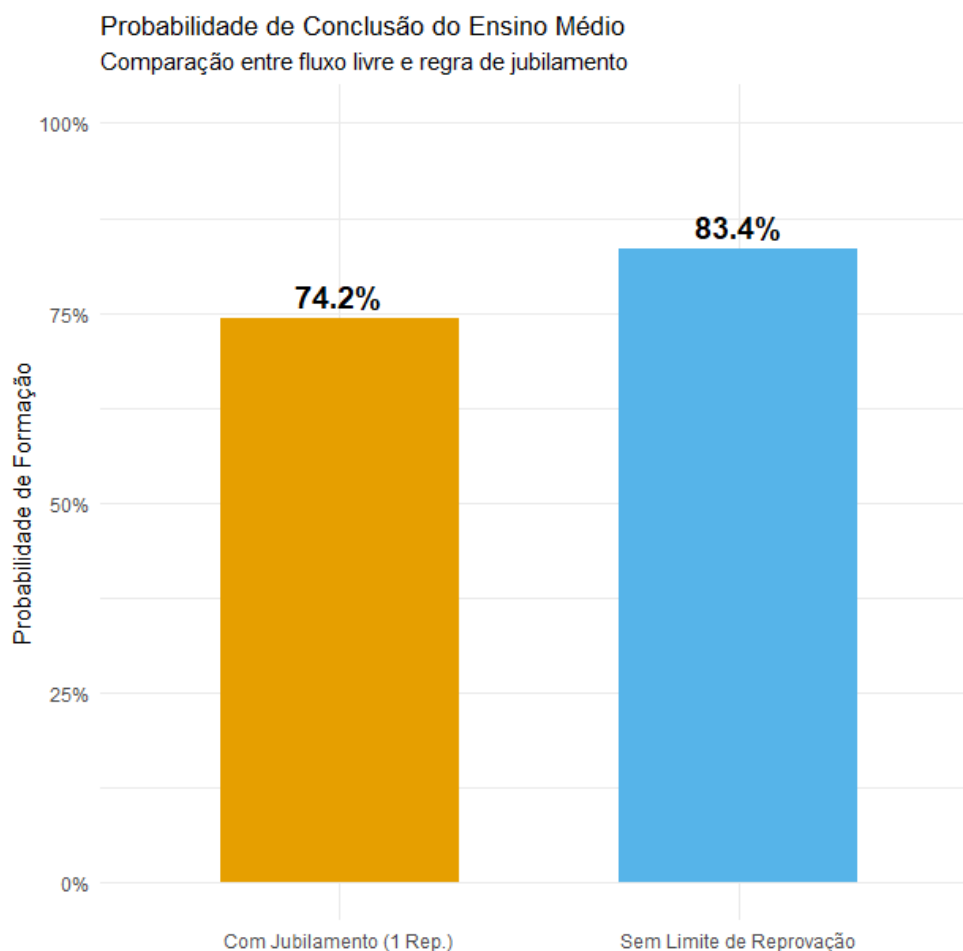
RESULTADO DO PROGRAMA 7

	Formado	Evadido/Jubilado
1-ano-1v	0.7787	0.2213
1-ano-2v	0.6771	0.3229
2-ano-1v	0.8464	0.1536
2-ano-2v	0.7360	0.2640
3-ano-1v	0.9200	0.0800
3-ano-2v	0.8000	0.2000

3.4.6 - Programa 8 - Comparação dos resultados com ou sem jubramento

```
# Carregando a biblioteca de visualização
dados_comparacao <- data.frame(
  Cenario = c("Sem Limite de Reprovação", "Com Jubramento (1 Rep.)"),
  Probabilidade = c(0.8337, 0.7424) # Valores calculados via Matrizes
)
library(ggplot2)
ggplot(dados_comparacao, aes(x = Cenario, y = Probabilidade, fill = Cenario)) + geom_bar(stat =
"identity", width = 0.6) +
  geom_text(aes(label = paste0(round(Probabilidade * 100, 1), "%")),
    vjust = -0.5, size = 5, fontface = "bold") +
  scale_y_continuous(labels = scales::percent, limits = c(0, 1)) +
  scale_fill_manual(values = c("#E69F00", "#56B4E9")) +
  labs(title = "Probabilidade de Conclusão do Ensino Médio",
    subtitle = "Comparação entre fluxo livre e regra de jubramento",
    x = "", y = "Probabilidade de Formação") +
  theme_minimal() +
  theme(legend.position = "none")
ggplot(dados_comparacao, aes(x = Cenario, y = Probabilidade, fill = Cenario)) + geom_bar(stat =
"identity", width = 0.6) +
  geom_text(aes(label = paste0(round(Probabilidade * 100, 1), "%")),
    vjust = -0.5, size = 5, fontface = "bold") +
  scale_y_continuous(labels = scales::percent, limits = c(0, 1)) +
  scale_fill_manual(values = c("#E69F00", "#56B4E9")) +
  labs(title = "Probabilidade de Conclusão do Ensino Médio",
    subtitle = "Comparação entre fluxo livre e regra de jubramento",
    x = "", y = "Probabilidade de Formação") +
  theme_minimal() +
  theme(legend.position = "none")
# Carregando a biblioteca de visualização
library(ggplot2)
ggplot(dados_comparacao, aes(x = Cenario, y = Probabilidade, fill = Cenario)) + geom_bar(stat =
"identity", width = 0.6) +
  geom_text(aes(label = paste0(round(Probabilidade * 100, 1), "%")),
    vjust = -0.5, size = 5, fontface = "bold") +
  scale_y_continuous(labels = scales::percent, limits = c(0, 1)) +
  scale_fill_manual(values = c("#E69F00", "#56B4E9")) +
  labs(title = "Probabilidade de Conclusão do Ensino Médio",
    subtitle = "Comparação entre fluxo livre e regra de jubramento",
    x = "", y = "Probabilidade de Formação") +
  theme_minimal() +
  theme(legend.position = "none")
ggplot(dados_comparacao, aes(x = Cenario, y = Probabilidade, fill = Cenario)) + geom_bar(stat =
"identity", width = 0.6) +
  geom_text(aes(label = paste0(round(Probabilidade * 100, 1), "%")),
    vjust = -0.5, size = 5, fontface = "bold") +
  scale_y_continuous(labels = scales::percent, limits = c(0, 1)) +
  scale_fill_manual(values = c("#E69F00", "#56B4E9")) +
  labs(title = "Probabilidade de Conclusão do Ensino Médio",
    subtitle = "Comparação entre fluxo livre e regra de jubramento",
    x = "", y = "Probabilidade de Formação") +
  theme_minimal() +
  theme(legend.position = "none")
```

Fonte: Elaboração do Autor.

Figura 9 – Número de formados com e sem jubilamento

Fonte: Elaboração do Autor

3.4.7 Análise do gráfico

- **Impacto das Regras:** Neste cenário hipotético, o gráfico evidencia que a introdução de uma regra restritiva (jubilamento) reduz a taxa de sucesso escolar em cerca de 9,2 pontos percentuais;
- **Decisão Pedagógica:** Modelo usado por gestores para prever o impacto de mudanças no regimento escolar sobre os índices de evasão.

A comparação visual demonstra que a probabilidade de um aluno concluir o curso cai de 83,4% (modelo sem restrições) para 74,2% quando se aplica o limite de uma reprovação por ano.

3.5 ANÁLISE DE SENSIBILIDADE

Este código calcula a solução exata (via álgebra de matrizes) para 31 cenários diferentes de evasão.

A seguir, apresenta-se o código para testar taxas de evasão de **0% a 30%** e ver o impacto direto no sucesso dos alunos.

3.5.1 – Programa 9 - Análise de sensibilidade

```
# Pacote necessário no R para o gráfico
library(ggplot2)

# Parâmetros fixos
taxa_reprovacao <- 0.15
ev <- seq(0, 0.30, by = 0.01) # Taxas de evasão de 0% a 30%

# Função para calcular prob. de formação (Modelo com 1 jubilamento)
calc_formacao <- function(ev) {
  ap <- 1 - taxa_reprovacao - ev # taxa de aprovação (corrigido: removida a
  chave extra)

  # Matriz Q (Transitórios)
  Q <- matrix(0, 6, 6)
  Q[1,2] <- taxa_reprovacao
  Q[1,3] <- ap
  Q[2,3] <- ap
  Q[3,4] <- taxa_reprovacao
  Q[3,5] <- ap
  Q[4,5] <- ap
  Q[5,6] <- taxa_reprovacao

  # Matriz R (Absorventes: Formado [col 1], Evadido [col 2])
  R <- matrix(0, 6, 2)
  R[1,2] <- ev
  R[2,2] <- ev + taxa_reprovacao # Jubilamento
  R[3,2] <- ev
  R[4,2] <- ev + taxa_reprovacao
  R[5,1] <- ap
  R[5,2] <- ev
  R[6,1] <- ap
  R[6,2] <- ev + taxa_reprovacao

  B <- solve(diag(6) - Q) %*% R
  return(B[1,1]) # Retorna prob. de formar saindo do 1º ano
}

# Gerar resultados (usando o vetor 'ev' como taxas de evasão)
```

```

resultados_sens <- data.frame(
  Evasao = ev,
  Sucesso = sapply(ev, calc_formacao)
)

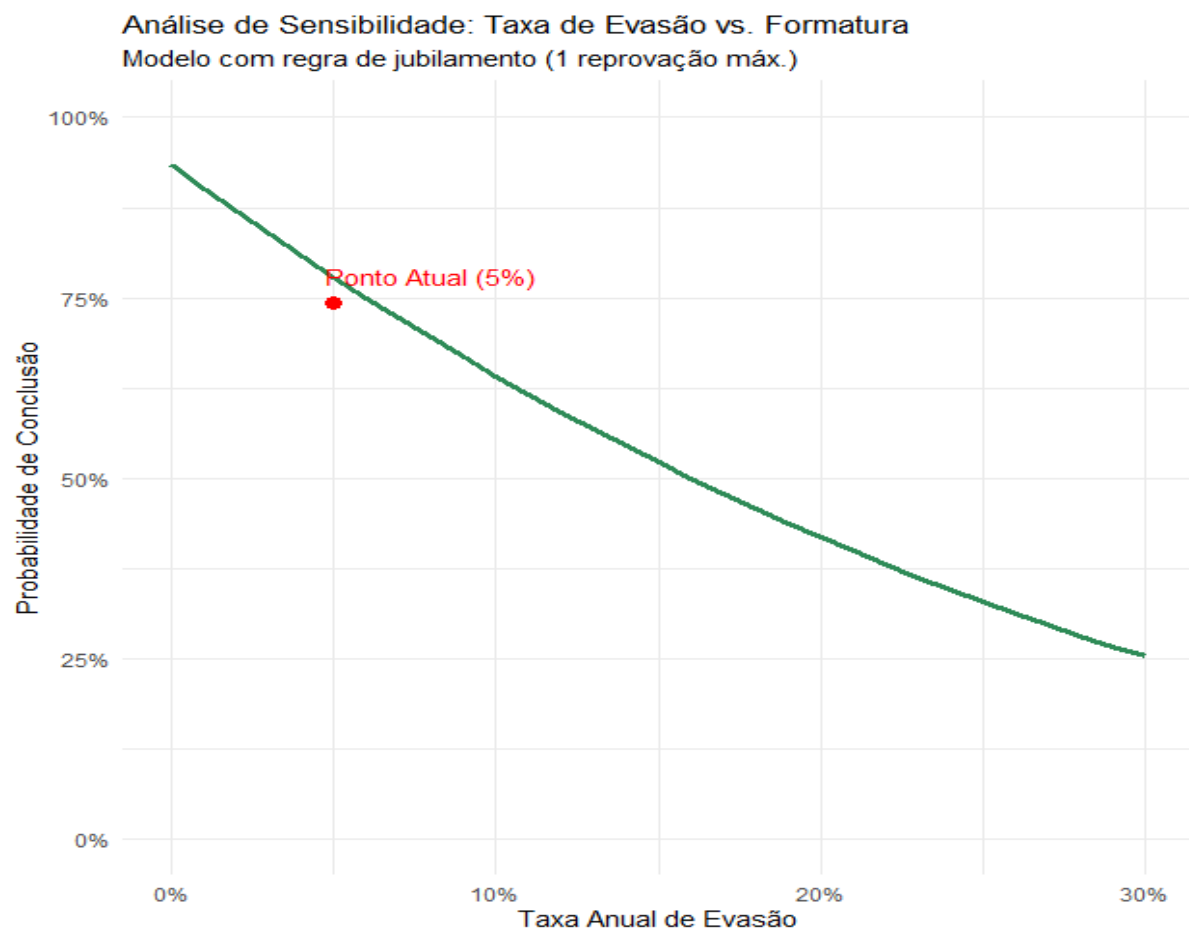
# Gráfico
ggplot(resultados_sens, aes(x = Evasao, y = Sucesso)) +
  geom_line(color = "#2E8B57", size = 1.2) +
  geom_point(aes(x = 0.05, y = 0.7424), color = "red", size = 3) + # Ponto
  atual (5% evasão)
  annotate("text", x = 0.08, y = 0.78, label = "Ponto Atual (5%)", color = "red")
  +
  scale_x_continuous(labels = scales::percent) +
  scale_y_continuous(labels = scales::percent, limits = c(0, 1)) +
  labs(title = "Análise de Sensibilidade: Taxa de Evasão vs. Formatura",
        subtitle = "Modelo com regra de jubramento (1 reprovação máx.)",
        x = "Taxa Anual de Evasão", y = "Probabilidade de Conclusão") +
  theme_minimal()

```

Fonte: Elaboração do Autor.

3.5.2 Visualização do Impacto

Figura 10 – Análise de sensibilidade para testar taxas de evasão



Fonte: Elaboração do Autor

Observações:

Elasticidade:

A queda não é linear. Pequenos aumentos na evasão anual geram perdas acumuladas grandes, pois o aluno "arrisca" a evasão em cada um dos 3 anos (e nas eventuais reprovações).

Ponto de Ruptura:

Se a evasão anual subir para **20%**, a probabilidade de um aluno que entra no 1º ano se formar cai para aproximadamente **45%**. (menos da metade).

Evasão Zero:

Mesmo com **0% de evasão**, a taxa de formação não seria 100% neste modelo, devido ao **jubilamento** (nos cursos em que há essa regra)

3.5.3 Resultados da análise

Para que o sistema garanta a formação de pelo menos **metade (50%)** dos alunos que ingressam no 1º ano:

- **Taxa de Evasão Máxima Aceitável: 13,11%** (ao ano).
- **Contexto:** Se a evasão anual das turmas ultrapassar este valor, a eficiência do sistema cai para menos de 50%, tornando a "formatura" um evento estatisticamente improvável para o ingressante médio.

Observação:

Este valor é relativamente baixo porque o risco é acumulativo:

- O aluno enfrenta o risco de evasão no **1º ano**.
- Se reprovar, enfrenta o risco novamente como **repetente**.
- Se passar, enfrenta o risco no **2º ano**, e assim por diante.
- O efeito cascata das probabilidades ($P_{Total} = P_1 \times P_2 \times P_3 \dots$) faz com que pequenas taxas anuais resultem em uma grande perda total.

3.5.4 Resumo comparativo

- **Evasão Atual (5%):** Formatura de **74,2%**.
- **Evasão Limite (13,1%):** Formatura de **50,0%**.
- **Evasão Crítica (>15%):** A maioria dos alunos **não** receberá o diploma.

A tabela a seguir consolida os indicadores de desempenho calculados via álgebra de matrizes (método exato) para o modelo de Ensino Médio com regra de jubramento, sendo no máximo uma reprovação por série e com taxa fixa de 15%. Os indicadores representam o destino final de um aluno que ingressa no 1º Ano.

Procedendo-se dessa forma, mas aumentando a taxa de evasão em repetições sucessivas do algoritmo, foi obtida a seguinte tabela.

Tabela 10 – Comparativa do destino final do Ingressante

Cenário de Evasão Anual	Formatura (%)	Evasão Voluntária (%)	Jubilamento (%)	Eficiência do Sistema
Otimista (2%)	88,2%	6,3%	5,5%	Altíssima
Atual (5%)	74,2%	14,4%	11,4%	Moderada
Crítico (10%)	57,1%	24,8%	18,1%	Baixa
Limite (13,1%)	50,0%	29,6%	20,4%	Ponto Equilíbrio
Severo (20%)	35,4%	38,5%	26,1%	Falência do Fluxo

Fonte: Elaboração do autor.

3.5.5 Análise dos Indicadores

- Formatura:** Representa a probabilidade de o aluno completar os 3 anos com sucesso, considerando que ele pode repetir cada ano no máximo uma vez.
- Evasão Voluntária:** Soma das probabilidades de o aluno desistir em qualquer um dos anos (seja na primeira tentativa ou na segunda).
- Jubilamento:** Probabilidade de o aluno ser desligado por não conseguir aprovação após a segunda tentativa em uma mesma série. Note que, quanto maior a evasão, maior o jubramento "teórico" acumulado, pois o sistema perde o aluno antes.
- Ponto de Equilíbrio:** Com uma evasão de 13,11%, o sistema torna-se um "jogo de uma moeda": um aluno que entra tem exatamente as mesmas chances de se formar ou de sair sem o diploma (50%).

CAPÍTULO IV - TESTES DIAGNÓSTICOS

4.1 VALIDAÇÃO E DIAGNÓSTICOS

Além de servirem como entradas, as cadeias de Markov possuem seus próprios "testes de diagnóstico" estatísticos para garantir que o modelo é confiável:

- **Teste de Estacionalidade:** verifica se as probabilidades de transição permanecem constantes ao longo do tempo (usando, por exemplo, o teste de quiquadrado).
- **Diagnósticos de Convergência:** no Método de Monte Carlo via ferramentas de diagnóstico são essenciais para decidir se a cadeia "convergiu" para a distribuição desejada antes de coletar os resultados.

4.1.1 Definição dos estados

Com a regra de jubilamento, teremos 8 estados no total:

- **Estados Transientes:** (1A) (1º ano), (1R) (Repetindo 1º), (2A) (2º ano), (2R) (Repetindo 2º), (3A) (3º ano), (3R) (Repetindo 3º).
- **Estados Absorventes:** (F) (Formado), (E) (Evadido/Jubilado).

Tabela 11 - Matriz de transição (P)

DE/PARA	1A	1R	2A	2R	3A	3R	F	E
1A	0.0	0.2	0.7	0.0	0.0	0.0	0.0	0.1
1R	0.0	0.0	0.7	0.0	0.0	0.0	0.0	0.3
2A	0.0	0.0	0.0	0.2	0.7	0.0	0.0	0.1
2R	0.0	0.0	0.0	0.0	0.0	0.7	0.0	0.3
3A	0.0	0.0	0.0	0.0	0.0	0.2	0.7	0.1
3R	0.0	0.0	0.0	0.0	0.0	0.0	0.7	0.3
F	0.0	0.0	0.0	0.0	0.0	0.0	1.0	0.0
E	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0

Fonte: Elaboração do autor.

Na matriz (P) acima, a título de exemplo, foram adotadas probabilidades hipotéticas de: 70% de aprovação, 20% de reprovação, 10% de evasão. Além disso, no estado de repetição (R), se o aluno não passa, ele é jubilado (vai para o estado E), por isso a probabilidade de evasão/saída aumenta.

4.2 DIAGNÓSTICO – CURVA ROC – ÍNDICE DE GINI

Em Cadeias de Markov, o "teste de diagnóstico" avalia a capacidade do modelo de prever quem se formará (Classe Positiva) versus quem evadirá (Classe Negativa).

As probabilidades de absorção são conceitos fundamentais em Cadeias de Markov absorventes, que determinam a chance de um sistema, iniciando em um determinado estado transiente, ser "absorvido" (ficar preso) em um estado absorvente específico ao longo do tempo.

Comportamento Limite: Em uma cadeia de Markov absorvente, a probabilidade de que o sistema seja eventualmente absorvido (seja qual for o estado absorvente) é 1, o que significa que o processo se estabiliza à medida que o tempo tende ao infinito. Assim, como o sistema é dinâmico, usamos a probabilidade de absorção que é a chance de um aluno que entra no 1º ano chegar ao estado "Formado".

4.2.1 A curva ROC (*Receiver Operating Characteristic*)

A respeito das probabilidades, do ponto de vista dos testes de hipóteses, com relação a nossas decisões, podemos ter

Tabela 12 – Resultados possíveis de um teste de hipóteses

		REALIDADE	
		Verdadeira	Falsa
DECISÃO	Verdadeira	$1 - \alpha$	β
	Falsa	α	$1 - \beta$

Fonte: Elaboração do autor

Notas: $(1 - \alpha)$ é o *nível de confiança*, (α) é o *nível de significância* (ou falso negativo), (β) (falso positivo), $(1 - \beta)$ é o *poder do teste*.

Vamos verificar se pelo menos uma variável explicativa é estatisticamente significativa para explicar a probabilidade de ocorrência do evento e se temos um modelo estatisticamente significativo para fins de previsão, ao nível de significância de α (alpha), será $\alpha = 5\%$ (default).

Chama-se *cutoff* ao *ponto de corte* (θ) para o qual as observações sejam classificadas como evento e as demais como não-evento, em função das probabilidades calculadas. Esse ponto é definido pelo pesquisador.

CRITÉRIO:

Se $p_i > \theta$, a observação i será classificada como **evento**.

Se $p_i < \theta$, a observação i será classificada como **não evento**.

Por exemplo, podemos definir para a probabilidade de absorção um *ponto de corte* (*cutoff*) de 0,5. Somente a título de ilustração, suponhamos que o resultado seja a seguinte matriz de confusão (*confusion matrix*)

Tabela 13 – Possibilidades de classificação de eventos, segundo o *cutoff*

	Incidência do evento	Incidência do não evento
Classificado como evento	VP = 46	FP = 16
Classificado como não evento	FN = 13	VN = 25

Fonte: Elaboração do autor.

A seguir, segue a métrica de uma matriz de confusão.

a) Acurácia (accuracy):

$$\frac{VP + VN}{VP + FP + VN + FN} = \frac{(46 + 25)}{100} = 0,71 = 71\%$$

b) Precisão (precision):

$$\frac{VP}{VP + FP} = \frac{46}{62} = 0,7419 = 74,19\%$$

c) Sensibilidade (recall): alta sensibilidade, tem-se menos falsos negativos

$$\frac{VP}{VP + FN} = \frac{46}{59} = 0,78 = 78\%$$

d) Especificidade: alta especificidade, tem-se menos falsos positivos, teste de maior poder.

$$\frac{VN}{VN + FP} = \frac{25}{41} = 0,61 = 61\%$$

e) F1_score

$$\frac{2 * \text{Precisão} * \text{Sensibilidade}}{\text{Precisão} + \text{Sensibilidade}} = \frac{2 * 0,7419 * 0,78}{0,7419 + 0,78} = 0,76 = 76\%$$

Portanto, nesse exemplo para um *ponto de corte* de 50%, temos:

Acurácia: 71%

Precisão: 74,19%

Sensibilidade: 78%

Especificidade: 61%

F1_score: 76%

Para gerar a curva ROC, variamos o *ponto de corte* (θ). Se a probabilidade calculada pelo modelo de um aluno se formar for maior ou igual a (θ), o diagnóstico será evento¹², caso contrário será não evento.

1. **Sensibilidade:** Proporção de alunos que realmente se formaram e o modelo previu corretamente.
2. **Especificidade:** Proporção de alunos que evadiram e o modelo previu que evadiriam.

Observação: Para a construção da curva ROC, área sob a curva e o índice de Gini por meio do software R, veja o Anexo IV, onde se encontra um banco de dados com cem alunos.

4.2.2 Resultado esperado

Para o modelo de Markov, a curva ROC costuma ser muito eficiente se os dados históricos forem bons.

- **Curva ROC:** Terá um formato arqueado para o canto superior esquerdo. Quanto mais "quadrada", melhor o modelo.
- **Índice de Gini:** É calculado a partir da Área Sob a Curva (**AUC**).

$$Gini = 2 \times AUC - 1$$

¹² No caso, “evento” significa “irá se formar”.

- Se **AUC = 0,85**, tem-se **Gini = 0,70**. (um valor muito bom indicando sistemas educacionais bem ajustados).

4.2.3 Resumo do diagnóstico

AUC (Área sob a curva): Indica a probabilidade de o modelo distinguir corretamente um aluno que terá sucesso de um que evadirá. É a métrica-resumo que quantifica o desempenho geral. Varia de 0 a 1. Para **AUC = 0,5**, o modelo não possui capacidade discriminativa. Se **AUC > 0,7**, tem-se um desempenho adequado e **AUC > 0,9**, tem-se excelente capacidade de separação.

Gini: Mede a desigualdade entre as previsões, ou seja, um **Gini** alto indica que o modelo separa muito bem os grupos de risco dos grupos de sucesso.

Eixos do Gráfico:

- **Eixo Y** (Vertical): Taxa de Verdadeiros Positivos (Sensibilidade)
- **Eixo X** (Horizontal): Taxa de Falsos Positivos (1 - Especificidade).

Considerações Importantes

- **Objetivo da Análise**: A Curva ROC e o GINI são métricas de avaliação de modelo preditivo, e não de medição de desigualdade (como o coeficiente de GINI usado em economia). Eles respondem à pergunta: "Quão bom é o meu modelo em prever corretamente os resultados?".
- **Cenário de Simulação**: O valor e a interpretação dessas métricas serão mais ricos se a simulação for usada para testar diferentes políticas, intervenções ou modelos preditivos. Por exemplo, comparar o Índice de GINI de um modelo simples com o de um modelo mais complexo para ver qual oferece o melhor poder de previsão.
- **Em resumo**: a Curva ROC e o Índice de GINI são ferramentas valiosas para sintetizar e comunicar a eficácia de modelos preditivos em simulações com resultados binários, sendo perfeitamente aplicáveis ao estudo de um curso de 3 anos.

4.3 SIMULAÇÕES DE CADEIAS DE MARKOV

4.3.1 Simulação de 1.000 alunos com taxas pré-fixadas

A simulação foi baseada em uma amostra de 1.000 alunos, seguindo as regras de fluxo (aprovação de 70%, reprovação de 20% e saída do sistema de

10%). A regra de jubramento foi aplicada transformando a segunda reprovação em uma transição direta para o estado de “Evadido”.

4.3.2 Simulação dos Resultados

Considerando uma Cadeia de Markov onde o estado final de cada aluno é determinado pelas probabilidades de transição:

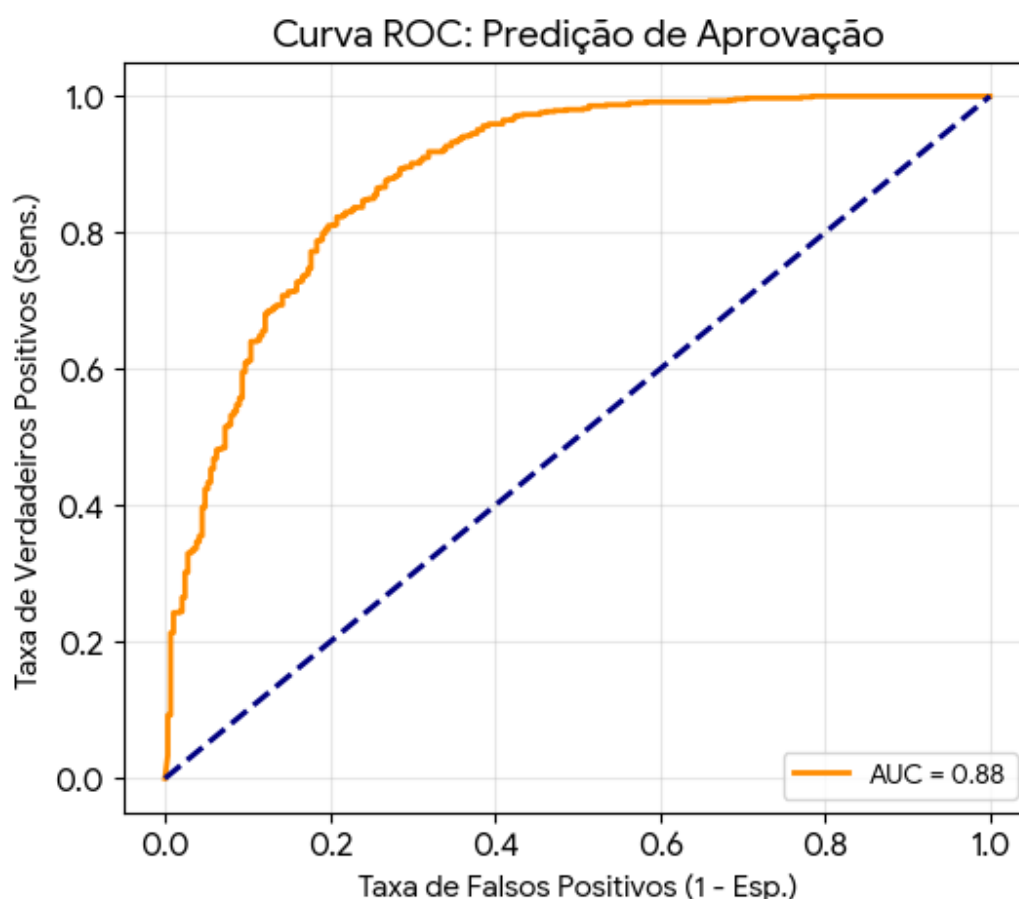
- **Aprovação (70%):** ~700 alunos.
- **Reprovação (20%):** ~200 alunos.
- **Evasão (10%):** ~100 alunos.

Observação: Nesta simulação específica de 1.000 indivíduos, foram obtidas 695 aprovações.

4.3.3 Visualização da Curva ROC

Para a curva ROC, definimos a **Aprovação** como a classe positiva (1,0) e a soma de Reprovação/Evasão como a classe negativa (0,0). O gráfico abaixo demonstra a capacidade de um modelo hipotético em distinguir esses alunos:

Figura 11 – Curva ROC e a área sob essa curva



Fonte: Elaboração do Autor.

4.3.4 Análise dos resultados

- **AUC (0.88):** O valor da Área Sob a Curva indica que o modelo tem uma excelente capacidade discriminatória. Isso significa que, ao escolher um aluno aprovado e um não aprovado ao acaso, o modelo tem cerca de 88% de chance de atribuir um score maior ao aluno aprovado.
- **Distribuição:** A curva se inclina rapidamente para o canto superior esquerdo, o que reflete que conseguimos uma alta taxa de acerto de aprovados (Sensibilidade) mantendo poucos erros de classificação de reprovados (Falsos Positivos).
- **Índice de Gini:** Calculado como $(2 \times AUC - 1)$. O valor de GINI igual a 0,76 reflete a taxa de erro na classificação de risco.
- **Significado Educacional:** Um Gini alto sugere que a regra de jubramento é o principal preditor de evasão no sistema, tornando o diagnóstico da situação do aluno o fator crítico para intervenções pedagógicas se for sua primeira vez na série ou uma repetição.

Tabela 14 – Testes diagnósticos segundo o índice de Gini

CENÁRIO	GINI ESPERADO	DIAGNÓSTICO PRINCIPAL
Com jubramento rígido	0,99	O status de repetente prediz quase perfeitamente a evasão
Com alta evasão voluntária	0,85	O "ruído" externo dificulta prever quem sairá por vontade própria
Com política de recuperação	0,90	O sucesso é democratizado, o risco é diluído e mais difícil de isolar

Fonte: Elaboração do autor

Pode-se observar um efeito multiplicador, de modo que, se diminuir o total de repetentes, o sucesso do sistema é democratizado e o risco diluído. Dessa forma, para aumentar o índice de eficiência dos cursos, esta tese conclui com as seguintes sugestões para a gestão escolar otimizar o número de formandos e minimizar o número de alunos que evadem:

Em resumo, a Curva ROC e o Índice de GINI são ferramentas valiosas para sintetizar e comunicar a eficácia de modelos preditivos em simulações com resultados binários, sendo perfeitamente aplicáveis ao estudo de um curso de 3 anos.

Tabela 15 – Resumo sobre a curva Roc e o índice de Gini

Ferramenta	Função na Análise
Curva ROC	Avalia o desempenho do modelo em todas as faixas de probabilidade, traçando a Taxa de Verdadeiros Positivos (Sensibilidade) contra a Taxa de Falsos Positivos (1 – Especificidade)
	No contexto de um curso seriado de 3 anos:
	Permite visualizar e escolher o melhor ponto de corte para classificar os alunos. Por exemplo, qual é a probabilidade mínima caracteriza um aluno em risco de evasão
Índice de GINI	Resume o poder de discriminação do modelo em um único número. É calculado a partir da Curva ROC, sendo $GINI = 2 * AUC - 1$

Tabela 16 – Sugestões para a Gestão Escolar

Atitude da Gestão Escolar	Consequência
Não adotar política de jubramento	Diminuir a saída dos alunos do sistema (evasão)
Implementar uma política de recuperação	Evitar repetência e saída dos alunos do sistema (evasão)
Controlar as taxas de evasão e de reprovação	Evasão de 14% e reprovação de 20% implica taxa de aprovação de, aproximadamente, 50% (no sistema simulado).
Aumentar as aprovações do 1º semestre	Obter o efeito multiplicador do processo
Diminuir os falsos negativos (β) Por meio da <i>confusion matrix</i>	Aumentar o poder do teste ($1 - \beta$)
Realizar simulações periodicamente	Interpretar os resultados
Realizar testes diagnósticos - construir a curva ROC	Controlar o impacto do índice de <i>Gini</i> e a área sob a curva (<i>AUC</i>)

Fonte: Elaboração do Autor.

Dessa forma, poderá haver aumento significativo do número de alunos formados, em relação ao número de alunos que entram no sistema e, além disso, uma diminuição do número de alunos que saem do sistema antes de se formar, aumentando a eficiência do sistema, com os atuais recursos

5. CONCLUSÕES E CONSIDERAÇÕES FINAIS

Inicialmente, na introdução desta tese, após revisitar o sistema educacional brasileiro e a norma ABNT NBR ISO 21001 (ISO 21001:2018), estabelecendo a tomada de decisão baseada em evidências como um dos pilares da gestão acadêmica, foi estabelecida essa análise com base em dados de avaliações, evasão e desempenho para as quais foram definidos os Indicadores Chave de Performance (KPIs) nos indicadores Avaliações (Sistema e Clima), Evasão (Escolar/Acadêmica), Desempenho (Acadêmico/Pedagógico). A seguir, no Capítulo I foi realizada uma breve revisão a respeito da modelagem de Markov e cadeias de Markov, por meio de conceitos, descrições e exemplos. A seguir, foram abordados os conceitos de estados estacionários e a evolução do sistema ao longo do tempo, seja por meio de diagrama de árvore ou por meio da álgebra matricial, com a introdução do conceito de matriz de transição estocástica. A seguir foi abordado o conceito de estado absorvente, que pode estar associado a falha ou término do sistema. A partir desses conceitos foi introduzida a matriz de Leontief ou matriz fundamental que obtém o número de intervalos de tempo necessário para o sistema entrar no estado absorvente e a matriz absorvente que contém as probabilidades dos estados absorventes do sistema (formado ou evadido). No Capítulo II, por meio de cadeias de Markov, foi apresentada uma metodologia para analisar a eficiência de um curso de ensino médio integrado de três anos, já utilizada na literatura e, a seguir, a título de exemplo de aplicação dessa metodologia, foi analisada a eficiência de um curso médio, com dados do INEP adaptados, no período de 2022 a 2025. Esta tese mostrou que esta metodologia poderia ter sido aplicada, com as devidas adaptações, a qualquer curso seriado e de qualquer nível do sistema de educação brasileiro. A seguir, no Capítulo III, foram apresentados a Lei Fraca dos Grandes Números e o Teorema do Limite Central, que justificam e balizam as simulações que a seguir foram realizadas. Nesta tese, a simulação estocástica foi adotada para a análise dos cursos, pois abrange uma grande gama de métodos de simulações para a realização de simulações das cadeias de Markov, que permitiram obter várias percepções sobre as melhores estratégias para que os cursos de ensino médio tenham melhor eficiência. A principal contribuição desta tese para a gestão escolar foi estabelecer essa metodologia para obter a probabilidade de evasão

dos estudantes e, conseqüentemente, aumentar a eficiência dos cursos em oferta, diminuindo a latência entre os dados oficiais obtidos e a tomada de decisão. Verificou-se que no Ensino Médio Integrado, se introduzimos o jubramento, o número de formados diminuirá drasticamente, por outro lado, se não introduzirmos o jubramento e houver um programa de recuperação para os alunos repetentes e/ou com dificuldades, o número de formados aumenta significativamente. Além disso, se diminuir o número de alunos repetentes e/ou evadidos no primeiro ano, o número de alunos formados aumenta consideravelmente, devido ao efeito multiplicador da cadeia de Markov correspondente. Uma outra contribuição desta tese foi introduzir os algoritmos e programas, para realização das simulações em diversas situações, que poderão ser úteis para as decisões da gestão escolar. Nesta tese além dos diversos experimentos realizados por meio de simulações, para testar o comportamento das taxas de evasão sem/com jubramento para medir a eficiência dos cursos, foi introduzida uma análise de sensibilidade e a curva ROC (*Receiver Operating Characteristic*) como ferramenta gráfica para avaliar o desempenho de modelos de classificação binária. Além disso, foram realizados os testes diagnósticos, em cenários diversos, obtendo-se a área sob a curva ROC e o índice de Gini, para interpretação dos resultados e analisar seu impacto. Todas essas informações estão disponíveis para que a gestão escolar possa tomar as melhores decisões a respeito dos cursos em oferta. Para trabalhos futuros, sugere-se à gestão escolar, que por meio da metodologia apresentada, realize estudos de simulação estocástica, como um dos instrumentos que visam avaliar e otimizar a eficiência de seus cursos.

REFERÊNCIAS

AARTS, E.; KORST, J. Simulated annealing and Boltzman Machines: A Stochastic Approach to Combinatorial Optimization and Neural Computing. New York: John Wiley & Sons, 1991, 284 p.

ABNT NBR ISO 21001 (ISO 21001:2018): Sistemas de gestão da qualidade para Organizações Educacionais - requisitos. Rio de Janeiro, 2018. ASSOCIAÇÃO BRASILEIRA DE NORMAS TÉCNICAS.

APOSTOL, T. M. Análisis Matemático. Massachusetts: Addison-Wesley Publishing Company, Inc., Version Española por el Dr. D. Francisco Vélaz Cantarell, Barcelona: Editorial Reverté, S.A., 1960, 534 p.

BARROS, Anna Carolina ... [et al]. Pedro Guilherme Costa Ferreira (Org.) Análise de Séries Temporais em R: curso introdutório. São Paulo: Editora Atlas, 2021, 264 p.

BILLINGSLEY, Patric. Probability and Measure. 3.ed. New York: John Wiley & Sons, 1995, 608 p.

BLANCHARD, Oliver J. Methods of Solution for Dynamic Rational Expectations Models: A Survey. Mathematical Programming Study, v.23, p.210-225, 1985b.

BOLDRINI, et al. Algebra Linear. 3. ed. São Paulo: Harper & Row do Brasil, 1980, 411 p.

BRASIL. **Lei nº 9.394, de 20 de dezembro de 1996.** Estabelece as diretrizes e bases da educação nacional. *Diário Oficial da União*, Brasília, DF, 23 dez. 1996. Disponível em: http://www.planalto.gov.br/ccivil_03/leis/l9394.htm. Acesso em: 03 mar. 2026.

_____. **Lei nº 14.945/2024, de 31 de julho de 2024.** Brasília/DF, 2024. Disponível em: <https://www.planalto.gov.br/ccivil_03/ato2023-2026/2024/lei/l14945.htm>. Acesso em: 03 de mar. 2026.

BRONSON, Richard. Pesquisa Operacional. São Paulo: McGraw-Hill do Brasil, 1985, 318 p.

BUSTOS, Oscar H.; ORGAMBIDE, A. C. Frery. Simulação Estocástica: Teoria e Algoritmos. Livro-Texto do 10º. Simpósio Nacional de Probabilidade e Estatística. UFRJ/ABE. Rio de Janeiro, 1992, 152 p.

CHUNG, Kai Lay. A Course in Probability Theory. New York: Academic Press. 2 ed., 2000, 419 p.

COLLARD, F. & JUILLARD, M. Stochastic Simulations with DYNARE: A practical guide. 2003. Disponível em: <https://www.dynare.org/assets/tutorial/guide.pdf>. Acesso em: 10/03/2026.

DIRAC, Paul Adrien Maurice. Directions in Physics: Lectures Delivered During a Visit to Australia and New Zealand. New York: John Wiley & Sons, 1978.

FÁVERO, Luiz Paulo; BELFIORI, Patrícia. ANÁLISE DE DADOS: Estatística e Modelagem com Excel, SPSS e STATA. Rio de Janeiro: LTC, 2020, 1.187 p.

FERRARI, Pablo. Passeios Aleatórios e Redes Elétricas. 16º Colóquio Brasileiro de Matemática. Rio de Janeiro: Instituto de Matemática Pura e Aplicada – IMPA – CNPq, 1987, 87 p.

FERREIRA, Aurélio Buarque de Holanda. Dicionário Aurélio. 3 ed. Curitiba: Editora Positivo, 2004, 2.120 p.

GNEDENKO, A. YA.; KHINCHIN, B.V. An Elementary Introduction to the Theory of Probability. New York: Dover Books, 1976.

GUJARATI, Damodar. Econometria Básica. 4. ed. Rio de Janeiro: Elsevier, 2006, 812 p.

HAMILTON, James D. Time Series Analysis. New Jersey: Princeton University Press, 1994, 799 p.

Hill, Carter, GRIFFITHS; William E.; JUDGE, George G. Econometria, 2.ed. São Paulo: Editora Saraiva, 471 p.

LACHTERMACHER, Gerson. Pesquisa Operacional na tomada de decisões. 5 ed. Rio de Janeiro: LTC – Livros Técnicos e Científicos Editora Ltda. Grupo GEN, 2016, 534 p.

LIPSCHUTZ, Seymour. Algebra Linear. 3.ed. São Paulo: Makron Books, 1994, 647 p.

LEON-GARCIA, Alberto. Probability and random processes for electrical engineering. 2.ed. Boston: Addison-Wesley Publishing Company, Inc, 1994, 596 p.

LOESCH, Cláudio; HOELTGEBAUM, Mariane. Métodos Estatísticos Multivariados. São Paulo: Editora Saraiva, 2012, 288 p.

MACDONALD, Iain; ZUCCHINI, Walter. Hidden Markov and Other Models for Discrete-valued Time Series. London: Chapman & Hall/CRC, 2000, 236 p.

MACHADO, Antônio dos Santos. Álgebra linear e Geometria Analítica. São Paulo: Atual Editora Ltda., 1982, 210 p.

MANÉ, Ricardo. Teoria Ergódica. Projeto Euclides. IMPA. Rio de Janeiro: Livros Técnicos e Científicos Editora S.A., 1983, 388 p.

MEYN, Sean P.; TWEEDIE, Richard L. Markov Chain and Stochastic Stability. Berlin: Springer-Verlag, 1996, 550 p.

MORETIN, Pedro A.; BUSSAB, Wilton de O. Análise de Séries Temporais. São Paulo: Edgard Blücher, 2006, 535 p.

MORETIN, Pedro A.; TOLOI, Clélia M. C. Estatística Básica. São Paulo: Editora Saraiva, 2011, 540 p.

NORRIS, J. R. Markov chains. Cambridge: Cambridge University Press, 1998. 254 p.

PEREIRA, José Dias; SILVA, Eduardo Freitas da. Aplicação de um método simplificado do progresso estudantil do Ensino de 2º grau, p. 51. In: Estudos e Informes Estatísticos. Coletânea - Brasília: Serviço de Estatística de Educação e Cultura, MEC/SG, 1986, 143 p.

RODRIGUES, Adriano et al.; Análise Multivariada. São Paulo: Editora Atlas, 2007, 541, p.

ROSS, Sheldon. Stochastic Processes. 2nd Ed. New York: John Wiley & Sons, 1996, 510 p.

VIANNA, Marcelo. Stochastic Dynamics of Deterministic Systems. 21º Colóquio Brasileiro de Matemática. Rio de Janeiro: IMPA, 21-27 de julho, 1997, 197 p.

WICKHAM, Hadley; GROLEMUND, GARRET. R para Data Science: Importe, Arrume, Transforme, Visualize e Modele Dados. Rio de Janeiro: Alta Books, 2019, 497 p.

APÊNDICE 1

FUNDAMENTOS DE PROBABILIDADE

O propósito desse apêndice é apresentar as definições e teoremas básicos da Teoria da Medida. Naturalmente, não demonstraremos esses teoremas, mas indicaremos onde as demonstrações poderão ser encontradas.

1.1 MEDIDA

Seja X um conjunto não vazio. Uma família $\mathfrak{F}$ de subconjuntos de X é uma álgebra se:

- i) $X \in \mathfrak{F}$
- ii) $A \in \mathfrak{F} \Rightarrow A^c \in \mathfrak{F}$
- iii) $A \in \mathfrak{F} \text{ e } B \in \mathfrak{F} \Rightarrow A \cup B \in \mathfrak{F}$

Por exemplo, seja $\mathfrak{F}$ o conjunto das partes do conjunto $X = \{1, 2, 3\}$. Para essa família, tem-se:

$$\mathfrak{F} = [\emptyset, \{1\}, \{2\}, \{3\}, \{1, 2\}, \{1, 3\}, \{2, 3\}, \{1, 2, 3\}]$$

Logo, $\mathfrak{F}$ é uma álgebra de subconjuntos de X .

Dessa definição segue que:

$$\text{iv) } A \in \mathfrak{F} \text{ e } B \in \mathfrak{F} \Rightarrow A \cap B \in \mathfrak{F}, \text{ pois } A \cap B = (A^c \cup B^c)^c$$

$$\text{v) } A \in \mathfrak{F} \text{ e } B \in \mathfrak{F} \Rightarrow A - B \in \mathfrak{F}, \text{ pois } A - B = A \cap B^c$$

Além disso, $\mathfrak{F}$ é uma σ -álgebra, se

$$A_i \in \mathfrak{F}, i = 1, 2, 3, \dots \Rightarrow \bigcup_{i=1}^{\infty} A_i \in \mathfrak{F}$$

Se $\mathfrak{F}$ é uma σ -álgebra de subconjuntos de X , dizemos que

$\mu: \mathfrak{F} \rightarrow [0, +\infty]$ é uma medida se, para toda família $A_i \in \mathfrak{F}$, $i = 1, 2, 3, \dots$

de conjuntos disjuntos tais que $\bigcup_i A_i \in \mathfrak{F}$, vale:

$$\mu\left(\bigcup_{i \geq 1} A_i\right) = \sum_{i \geq 1} \mu(A_i)$$

Uma σ –álgebra gerada pela família de conjuntos abertos do conjunto dos números reais é chamada de σ –álgebra de Borel. Nesse caso, diz-se que os conjuntos de $\mathcal{B}$ são denominados de borelianos de X .

Um *espaço de medida* é uma terna $(X, \mathfrak{I}, \mu)$ onde $\mathfrak{I}$ é uma σ –álgebra de subconjuntos de X e $\mu: \mathfrak{I} \rightarrow [0, +\infty]$ é uma medida.

1.2 PROBABILIDADE

Um *espaço de probabilidade* é um espaço de medida $(X, \mathfrak{I}, \mu)$ tal que a medida μ tem peso 1, isto é, $\mu(X) = 1$. Nesse caso, dizemos que $\mu: \mathfrak{I} \rightarrow [0, +\infty]$ é uma *medida de probabilidade*.

Sejam $X \neq \emptyset$ um conjunto e $\mathfrak{I}$ uma σ – álgebra de subconjuntos de X e $Y \neq \emptyset$ outro conjunto e $\mathcal{B}$ uma σ –álgebra de subconjuntos de Y . Uma função $T: X \rightarrow Y$ é *mensurável*, com relação às σ –álgebras $\mathfrak{I}$ e $\mathcal{B}$, respectivamente, se $T^{-1}(B) \in \mathfrak{I}$, para todo $B \in \mathcal{B}$.

APÊNDICE 2

Balanceamento de dados e códigos do RMarkdown para auxiliar na tomada de decisão baseada em evidências

1. Balanceamento de dados

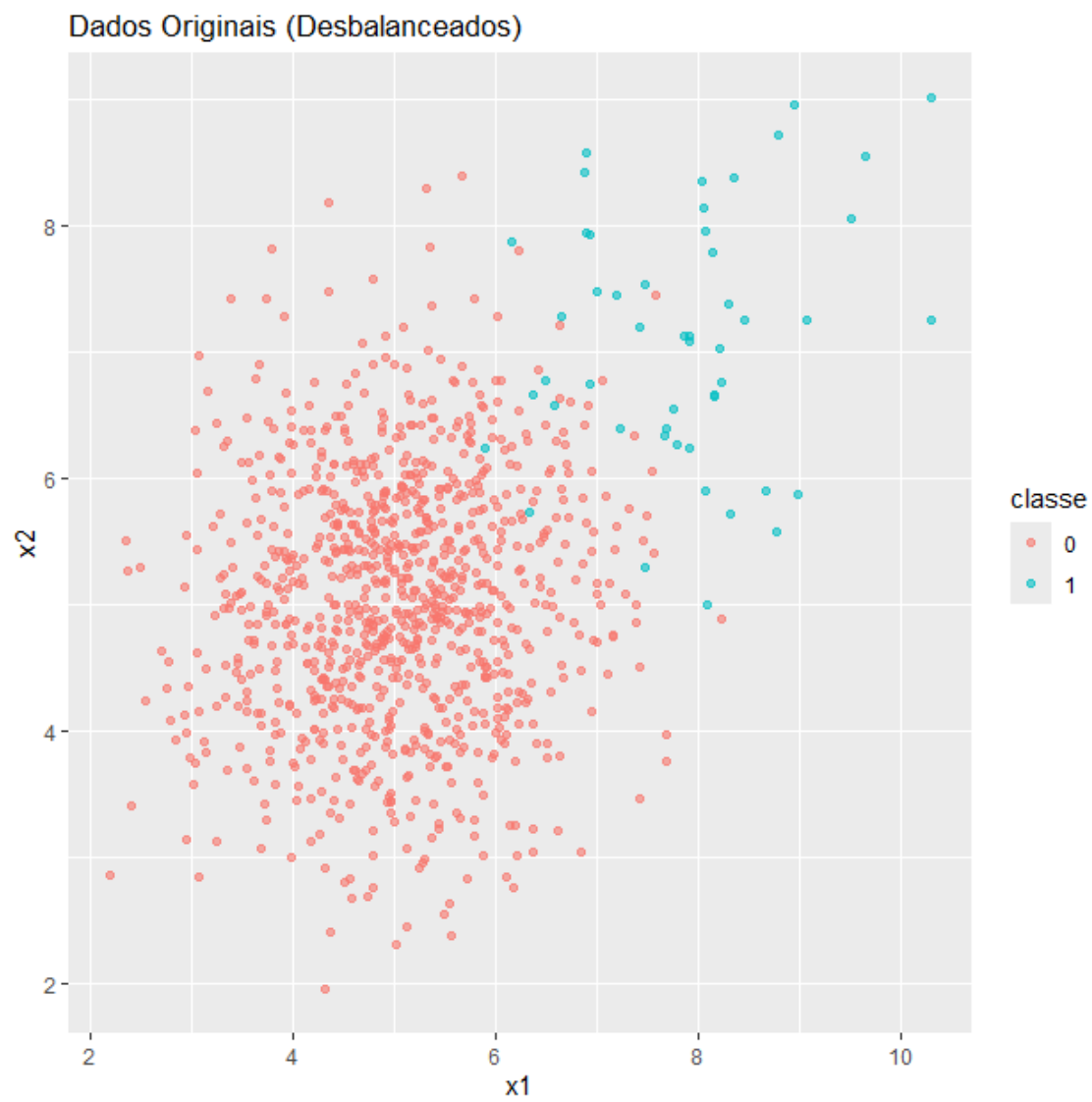
Visando auxiliar a gestão escolar na tomada de decisão baseada em evidências, seguem os códigos no aplicativo RMarkdown e para auxiliar, caso seja necessário, balancear dados antes da aplicação dos códigos, seguem exemplos de como isto pode ser realizado.

1.1 Programa 10 - Dados desbalanceados –

1.2 SOFTWARE R

```
# Carregar bibliotecas necessárias
library(ggplot2)
library(caret)
# Primeiro, instale o devtools se ainda não o tiver
install.packages("devtools")
# Instale o DMwR2 diretamente do GitHub
devtools::install_github("ltorgo/DMwR2")
library(DMwR2) # para SMOTE
library(ROSE) # para ROSE (oversampling avançado)
library(dplyr)
# Set seed para reprodutibilidade
set.seed(123)
# Criar dataset desbalanceado (95% classe 0, 5% classe 1)
n_majoritaria <- 950
n_minoritaria <- 50
dados <- data.frame(
  x1 = c(rnorm(n_majoritaria, mean = 5, sd = 1),
        rnorm(n_minoritaria, mean = 8, sd = 1)),
  x2 = c(rnorm(n_majoritaria, mean = 5, sd = 1),
        rnorm(n_minoritaria, mean = 7, sd = 1)),
  classe = factor(c(rep("0", n_majoritaria), rep("1", n_minoritaria)))
)
# Verificar desbalanceamento
table(dados$classe)
# 0 1
# 950 50
# Esta base está desbalanceada
# Visualizar
ggplot(dados, aes(x = x1, y = x2, color = classe)) +
  geom_point(alpha = 0.6) +
  ggtitle("Dados Originais (Desbalanceados)")
```

Fonte: Elaboração do Autor.

Figura 12 – Dados desbalanceados

Fonte: Elaboração do Autor.

2. POSSÍVEIS SOLUÇÕES

2.1. Programa 11 - Undersampling aleatório (Subamostragem)

```
# Método 1: Undersampling Aleatório Manual
library(ggplot2)
library(caret)
set.seed(456)
# Separar as classes
classe_0 <- dados[dados$classe == "0", ]
classe_1 <- dados[dados$classe == "1", ]

# Amostrar da classe majoritária o mesmo número da minoritária
classe_0_undersampled <- classe_0[sample(1:nrow(classe_0),
                                          size = nrow(classe_1),
                                          replace = FALSE), ]

# Combinar os dados balanceados
dados_undersampled <- rbind(classe_0_undersampled, classe_1)

# Verificar resultado: Base balanceada
table(dados_undersampled$classe)
# 0 1
# 50 50

# Visualizar
ggplot(dados_undersampled, aes(x = x1, y = x2, color = classe)) +
  geom_point(alpha = 0.6) +
  ggtitle("Após Undersampling (50 amostras por classe)")

# === SOLUÇÃO PRÁTICA: Usando o pacote 'caret' ===

# Undersampling com downSample()
dados_undersampled_caret <- downSample(x = dados[, c("x1", "x2")],
                                       y = dados$classe,
                                       yname = "classe")

# Dados balanceados
table(dados_undersampled_caret$classe)
# 0 1
# 50 50

# Visualizar
ggplot(dados_undersampled, aes(x = x1, y = x2, color = classe)) +
  geom_point(alpha = 0.6) +
  ggtitle("Após Undersampling (50 amostras por classe)")

#Vantagens do Undersampling: Rápido, reduz tamanho do dataset.
#Desvantagens: Pode perder informação importante da classe majoritária.
```

Fonte: Elaboração do Autor.

Figura 13 – Dados balanceados undersampling

Fonte: Elaboração do Autor.

2.2. OVERSAMPLING (Sobreamostragem)

#Aqui aumentamos a classe minoritária criando cópias ou amostras sintéticas.

2.1 - Programa 12 - Oversampling aleatório (Duplicação)

```
# Método 1: Oversampling Aleatório Manual
library(ggplot2)
library(caret)
set.seed(789)
# Separar as classes
classe_0 <- dados[dados$classe == "0", ]
classe_1 <- dados[dados$classe == "1", ]
# Duplicar a classe minoritária até igualar à majoritária
n_replicas <- nrow(classe_0) / nrow(classe_1) # 950/50 = 19
classe_1_oversampled <- classe_1[sample(1:nrow(classe_1),
size = nrow(classe_0),
```

```

replace = TRUE), ]
# Combinar
dados_oversampled <- rbind(classe_0, classe_1_oversampled)
#Dados balanceados
table(dados_oversampled$classe)
# 0 1
# 950 950
# Visualizar (note os pontos sobrepostos)
ggplot(dados_oversampled, aes(x = x1, y = x2, color = classe)) +
  geom_point(alpha = 0.4) +
  ggtitle("Após Oversampling Aleatório (950 amostras por classe)")

```

Fonte: Elaboração do Autor.

Figura 14 – Dados balanceados oversampling



Fonte: Elaboração do Autor.

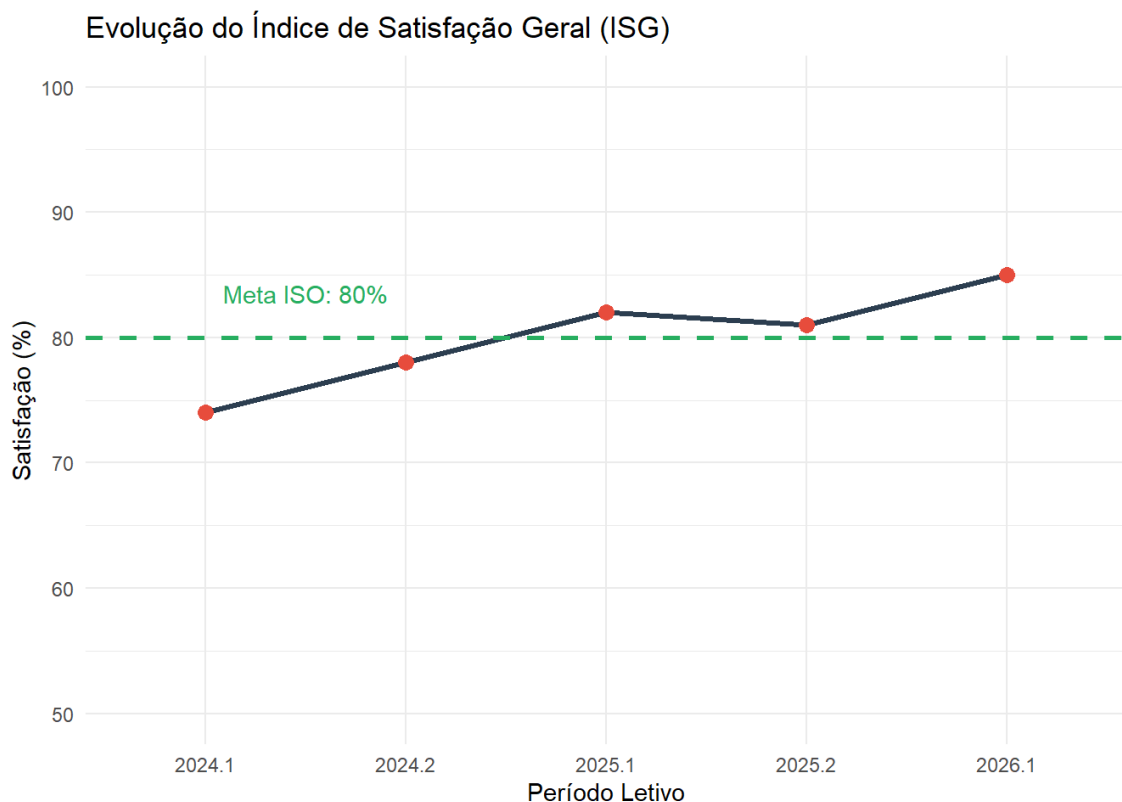
Códigos do RMarkdown para auxiliar na tomada de decisão baseada em evidências, de acordo com a norma ABNT NBR ISO 21001 (ISO 21001:2018)

2.3 - Programa 13 - Dados satisfação - Fonte: Elaboração do Autor.

```
dados_satisfacao <- data.frame(
  Semestre = c("2024.1", "2024.2", "2025.1", "2025.2", "2026.1"),
  ISG = c(74, 78, 82, 81, 85)
)

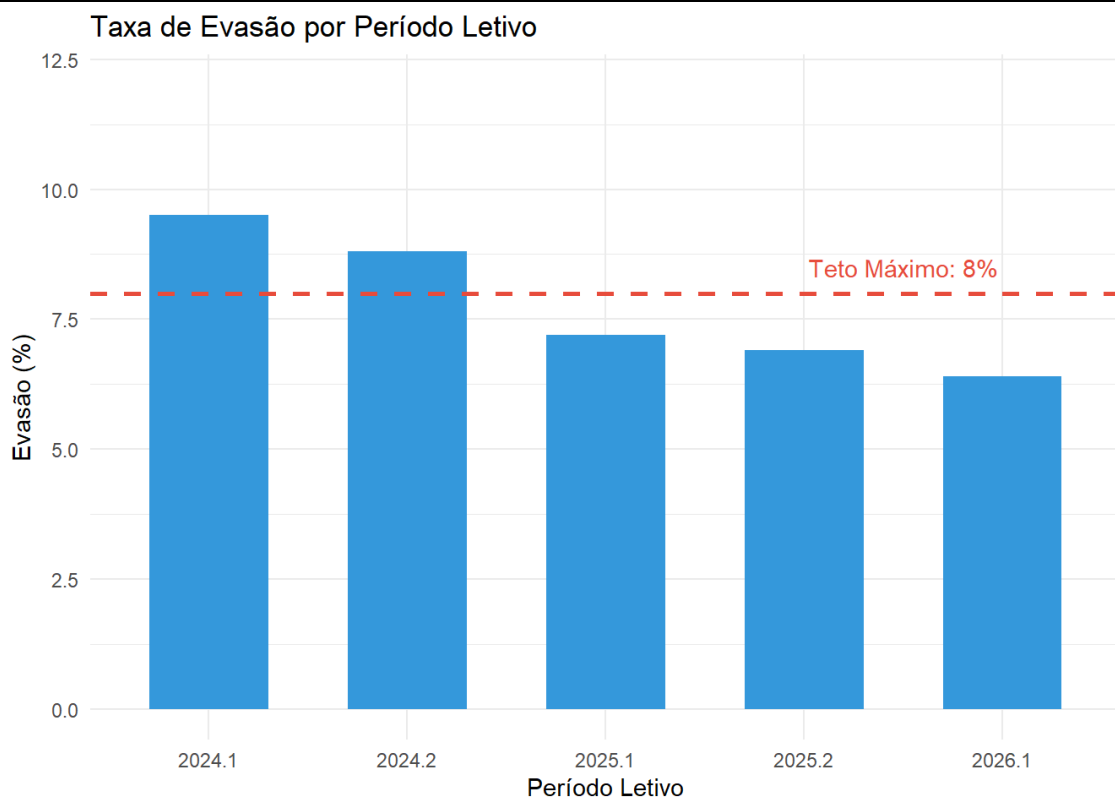
ggplot(dados_satisfacao, aes(x = Semestre, y = ISG, group = 1)) +
  geom_line(color = "#2c3e50", size = 1.2) +
  geom_point(color = "#e74c3c", size = 3) +
  geom_hline(yintercept = 80, linetype = "dashed", color = "#27ae60", size = 1) +
  annotate("text", x = 1.5, y = 82, label = "Meta ISO: 80%", color = "#27ae60", vjust = -0.5) +
  labs(title = "Evolução do Índice de Satisfação Geral (ISG)",
       x = "Período Letivo", y = "Satisfação (%)") +
  ylim(50, 100) +
  theme_minimal()
```

```
## Warning: Using `size` aesthetic for lines was deprecated in ggplot2 3.4.0.
## i Please use `linewidth` instead.
## This warning is displayed once per session.
## Call `lifecycle::last_lifecycle_warnings()` to see where this warning was
## generated.
```



2.4 - Programa 14 - Evasão Escolar

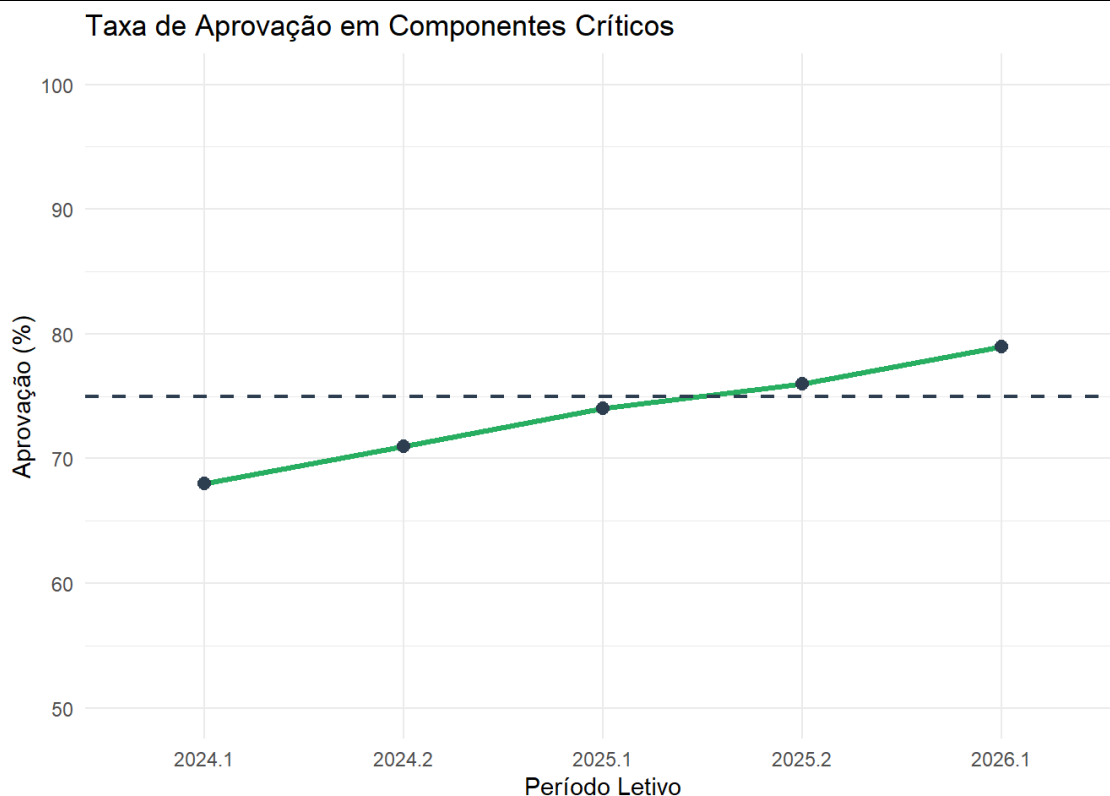
```
dados_evasao <- data.frame(  
  Semestre = c("2024.1", "2024.2", "2025.1", "2025.2", "2026.1"),  
  Taxa = c(9.5, 8.8, 7.2, 6.9, 6.4)  
)  
  
ggplot(dados_evasao, aes(x = Semestre, y = Taxa)) +  
  geom_bar(stat = "identity", fill = "#3498db", width = 0.6) +  
  geom_hline(yintercept = 8.0, linetype = "dashed", color = "#e74c3c", size = 1) +  
  annotate("text", x = 4.5, y = 8.5, label = "Teto Máximo: 8%", color = "#e74c3c") +  
  labs(title = "Taxa de Evasão por Período Letivo",  
        x = "Período Letivo", y = "Evasão (%)") +  
  ylim(0, 12) +  
  theme_minimal()
```



Fonte: Elaboração do Autor.

2.5 - Programa 15 - Desempenho (Acadêmico - Pedagógico)

```
dados_desempenho <- data.frame(
  Semestre = c("2024.1", "2024.2", "2025.1", "2025.2", "2026.1"),
  Aprovacao = c(68, 71, 74, 76, 79)
)
ggplot(dados_desempenho, aes(x = Semestre, y = Aprovacao, group = 1)) +
  geom_area(fill = "#2ecc71", alpha = 0.3) +
  geom_line(color = "#27ae60", size = 1.2) +
  geom_point(color = "#2c3e50", size = 2.5) +
  geom_hline(yintercept = 75, linetype = "dashed", color = "#2c3e50", size = 0.8) +
  labs(title = "Taxa de Aprovação em Componentes Críticos",
       x = "Período Letivo", y = "Aprovação (%)") +
  ylim(50, 100) +
  theme_minimal()
## Warning in max(ids, na.rm = TRUE): nenhum argumento não faltante para max;
## retornando -Inf
```



Fonte: Elaboração do Autor.

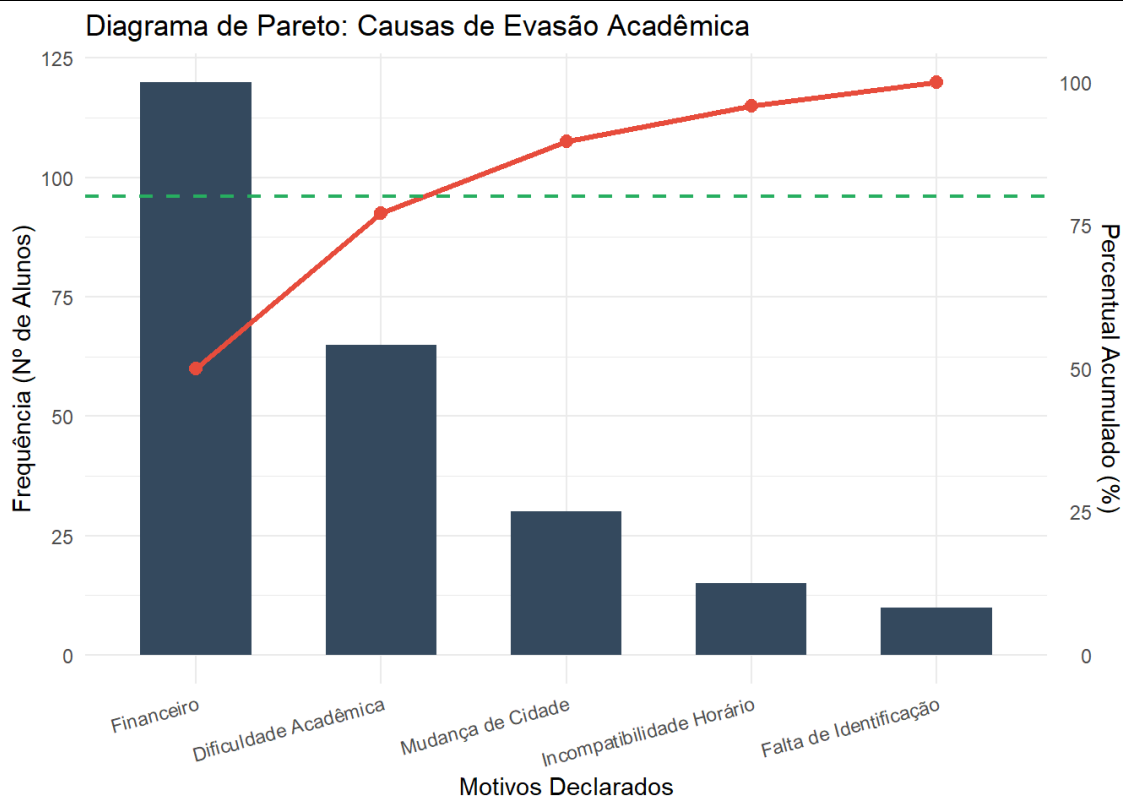
2.6 - Programa 16 - Análise de Causa Raiz: Diagrama de Pareto (Evasão)

```
# 1. Estruturação dos dados brutos de motivos de evasão
dados_causas <- data.frame(
  Motivo = c("Financeiro", "Dificuldade Acadêmica", "Mudança de Cidade", "Incompatibilidade Horário",
    "Falta de Identificação"),
  Frequencia = c(120, 65, 30, 15, 10)
)

# 2. Processamento para o cálculo do Pareto
dados_causas <- dados_causas %>%
  arrange(desc(Frequencia)) %>%
  mutate(
    Motivo = factor(Motivo, levels = Motivo),
    Freq_Acumulada = cumsum(Frequencia),
    Percent_Acumulado = (Freq_Acumulada / sum(Frequencia)) * 100
  )

# Fator de escala para alinhar o segundo eixo Y (Percentual) com o primeiro (Frequência)
escala_eixo <- max(dados_causas$Frequencia) / 100

# 3. Construção do Gráfico de Pareto
ggplot(dados_causas, aes(x = Motivo)) +
  # Barras de frequência absoluta
  geom_bar(aes(y = Frequencia), stat = "identity", fill = "#34495e", width = 0.6) +
  # Linha de percentual acumulado
  geom_line(aes(y = Percent_Acumulado * escala_eixo, group = 1), color = "#e74c3c", size = 1.2) +
  geom_point(aes(y = Percent_Acumulado * escala_eixo), color = "#e74c3c", size = 2.5) +
  # Linha guia dos 80%
  geom_hline(yintercept = 80 * escala_eixo, linetype = "dashed", color = "#27ae60", size = 0.8) +
  # Configuração dos dois eixos Y
  scale_y_continuous(
    name = "Frequência (Nº de Alunos)",
    sec.axis = sec_axis(~ . / escala_eixo, name = "Percentual Acumulado (%)")
  ) +
  labs(title = "Diagrama de Pareto: Causas de Evasão Acadêmica", x = "Motivos Declarados") +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 15, hjust = 1))
```



Fonte: Elaboração do Autor

APÊNDICE 3

SIMULAÇÕES DE VARIÁVEIS ALEATÓRIAS

Este apêndice é uma introdução a *Management Sciences*, com a qual o gestor escolar poderá iniciar sua experiência no processo de tomada de decisão e de modelagem matemática, com o auxílio de softwares adequados.

Quando os gestores se veem diante de uma situação na qual uma decisão deve ser tomada entre uma série de alternativas conflitantes e concorrentes, duas opções básicas se apresentam: (1) usar apenas a intuição gerencial; e (2) realizar um processo de modelagem da situação e exaustivas simulações dos mais diversos cenários de maneira a estudar mais profundamente o problema.

Denomina-se *Management Sciences* (MS) a área de estudos que utiliza computadores, estatística e matemática para resolver problemas de negócios. Pode ser considerada uma subárea da pesquisa operacional (PO). A definição de MS nos leva a três objetivos inter-relacionados:

- a) Converter dados em informações significativas: transformar dados brutos em dados de forma organizada;
- b) Apoiar o processo de tomada de decisão de formas transferíveis e independentes: por meio dos sistemas de apoio à decisão;
- c) Criar sistemas computacionais úteis para os usuários não técnicos: facilitar, por meio de sistemas de fácil utilização e/ou simulações, os processos de tomada de decisão operacional, gerencial e estratégico.

Em algumas questões de determinados estudos, mesmo com a disponibilidade de um modelo probabilístico, não se pode obter uma solução analítica.

Nesse caso, para obter aproximações de quantidades de interesse, pode-se recorrer a simulações, que tentam reproduzi-las, em ambiente controlado.

No caso desta tese, foram utilizadas simulações de variáveis aleatórias, para resolver problemas reais da gestão escolar.

Historicamente, este tipo de questão deu origem aos Métodos Monte Carlo (MMC), que pressupõem um gerador de números aleatórios equiprováveis.

No nosso estudo, uma simulação pode ser entendida como a realização de um modelo (binomial, uniforme, normal, etc.). Nesse sentido, os valores simulados podem ser considerados como uma amostra. Esse entendimento será útil para estudar as distribuições de estimadores e suas propriedades.

Originalmente, números aleatórios eram gerados manualmente ou mecanicamente, usando, por exemplo, roletas. Na atualidade, utilizam-se computadores para gerar números aleatórios, que na realidade são números pseudoaleatórios, pois dependem de relações matemáticas determinísticas, de modo que cada iteração depende do número gerado na iteração que a antecede. Para a simulação de variáveis aleatórias, pode-se usar o Teorema da Transformação Integral, apresentado a seguir.

Teorema da Transformação Integral. Se X for uma variável aleatória com função densidade acumulada $\mathfrak{F}$, então a variável aleatória $U = \mathfrak{F}(x)$ tem distribuição uniforme no intervalo $[0,1]$.¹³

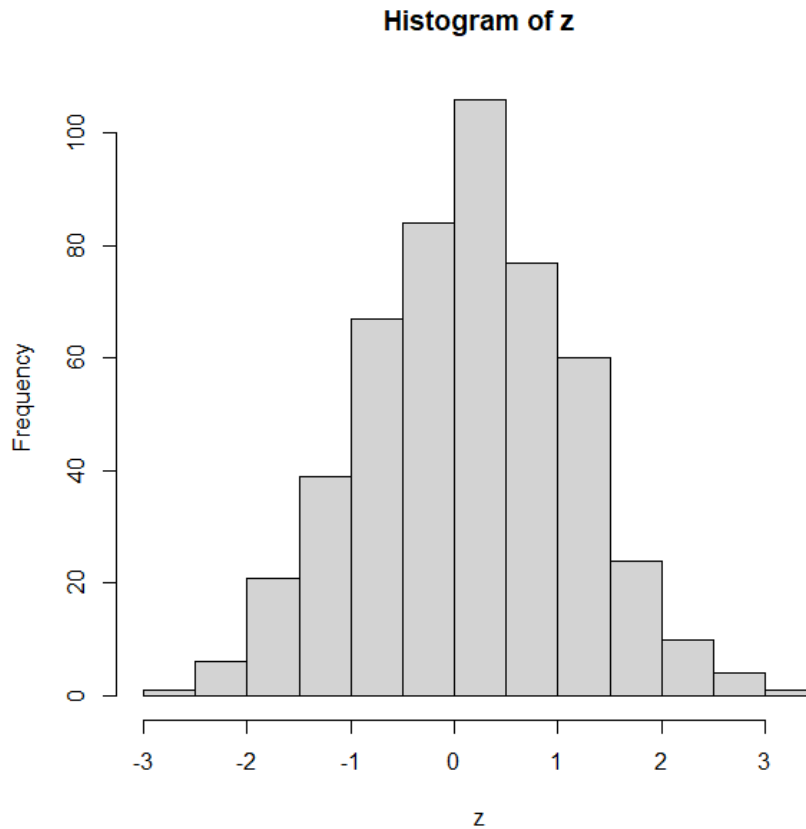
¹³ Pode-se obter a demonstração desse teorema na referência MORETTIN; BUSSAB, 2011, p.241.

EXEMPLOS DE SIMULAÇÃO DE ALGUNS MODELOS

Usando o software R, teremos:

Figura 15 - 500 valores da **distribuição normal** $Z \sim N(0,1)$:

$z < -rnorm(500,0,1), hist(z)$

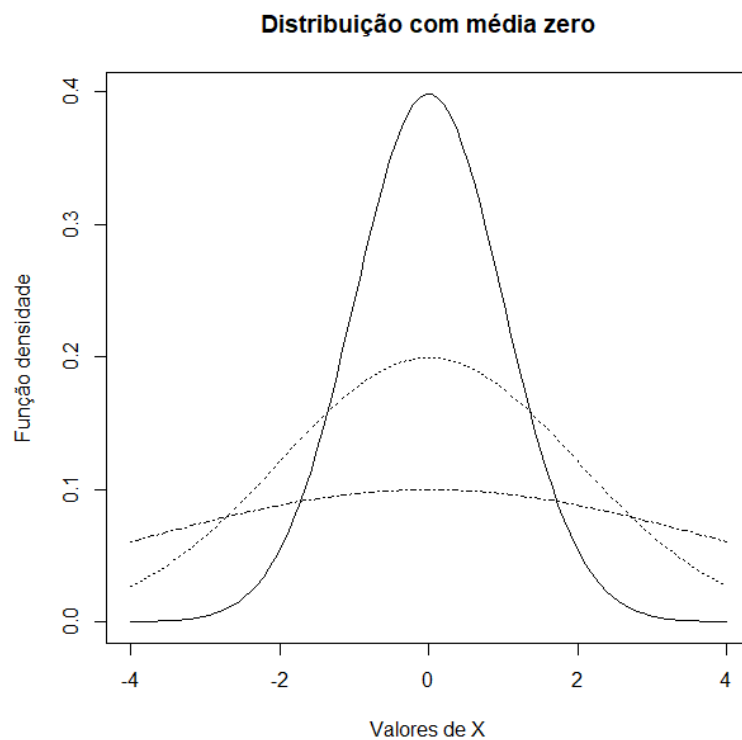


Fonte: Elaboração do Autor.

Na linguagem R, com relação à distribuição normal, usa-se a função $dnorm(x, \mu, \sigma)$ para avaliar função densidade de probabilidade no ponto x . Para avaliar a probabilidade acumulada no ponto x , isto é, $P(X \leq x)$, usa-se a função $pnorm(x, \mu, \sigma)$. Para obter o percentil (quantil) de ordem k , usa-se $qnorm(k, \mu, \sigma)$ e para gerar k números aleatórios, usa-se $rnorm(k, \mu, \sigma)$.

c) Gráfico das densidades de **distribuições normais com média $\mu = 0$** e três valores para o desvio-padrão $\sigma_1 = 1,5$, $\sigma_2 = 2$ e $\sigma_3 = 4$.

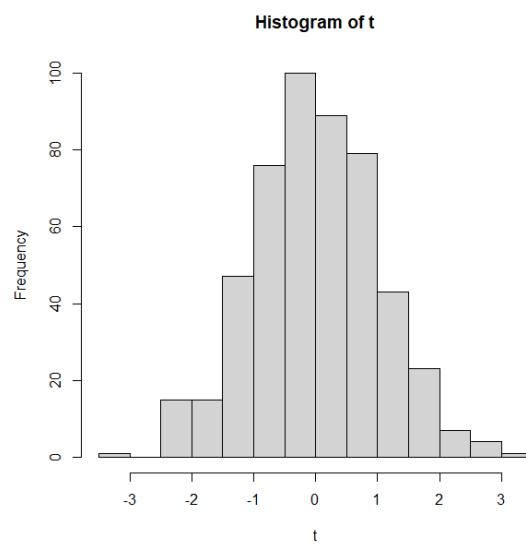
Figura 16 – Distribuições normais com mesma média



Fonte: Elaboração do Autor.

Figura 17 - 500 valores da **distribuição t de Student**

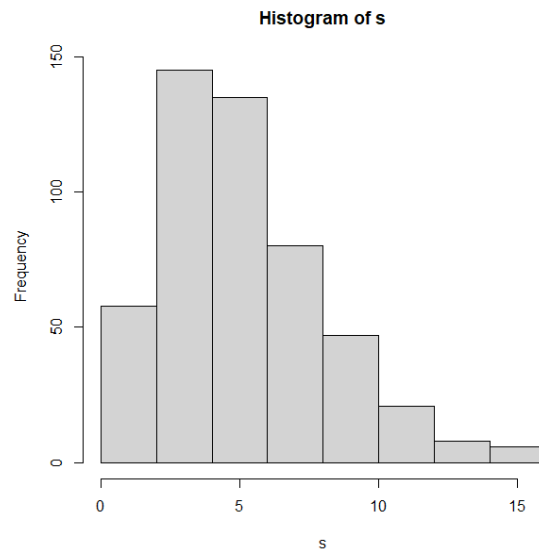
$t < -rt(500,70)$, $hist(t)$



Fonte: Elaboração do Autor.

Figura 18 - 500 valores da **distribuição quiquadrado** $\chi^2(5)$.

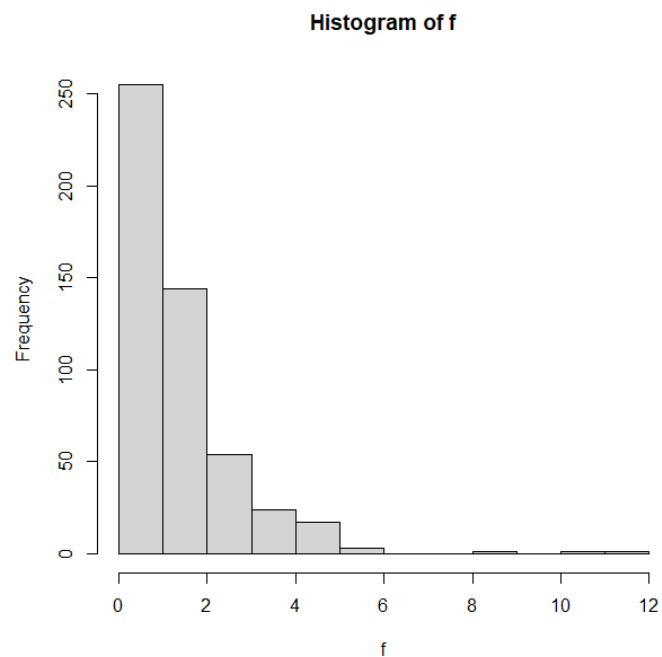
$$s < -rchisq(500,5), \text{ hist}(s)$$



Fonte: Elaboração do Autor.

Figura 19 - 500 valores da **distribuição F**:

$$f < -rf(500,9,7), \text{ hist}(f)$$



Fonte: Elaboração do Autor.

Tabela 17 - Funções correspondentes às principais distribuições de probabilidade

Distribuição de probabilidade	Nome da função no R
Beta	beta
Binominal	binom
Binomial negativa	nbinom
Cauchy	cauchy
Exponencial	exp
F Snedecor	f
Gama	gamma
Geométrica	geom
Hipergeométrica	hyper
Lognormal	lnorm
Normal	norm
Poisson	pois
Qui quadrado	chisq
T de Student	t
Uniforme	unif
Weibull	weibull

APÊNDICE 4

CONSTRUÇÃO DA CURVA ROC, AUC E GINI NO R

Neste apêndice, vamos introduzir o modelo de regressão logística binária, apresentar as estimações por meio da máxima verossimilhança e interpretar os resultados obtidos e fazer previsões.

A Regressão Logística é uma técnica supervisionada de *machine learning* utilizada para explicar ou prever a probabilidade de ocorrência de determinado evento em função de uma ou mais variáveis explicativas. Nesta, a variável dependente é qualitativa binária e as variáveis explicativas podem ser quantitativas ou qualitativas,

Os objetivos da regressão logística binária são estimar a probabilidade de ocorrência de determinado evento ou classificar indivíduos em categorias específicas. Mais especificamente, se a resposta a um estímulo (sim ou não) - pode ser a preferência por um produto, adimplência, aprovação em um curso, votar em certo candidato em uma eleição, etc. a probabilidade de ocorrência será indicada por p e a probabilidade de não ocorrência será indicada por $(1 - p)$.

O objetivo do modelo é estudar a probabilidade de ocorrência de um evento definido por uma variável Y indicadora, de modo que $Y = 1$ indica a ocorrência do evento de interesse e $Y = 0$ indica a ocorrência do não evento, com base nos valores das variáveis explicativas. Como Y é variável qualitativa, não se calcula sua média e variância, portanto não pode ser usado o método dos mínimos quadrados (OLS). O procedimento é definir um vetor Z de variáveis explicativas, com respectivos parâmetros estimados e logaritmo natural desse vetor:

$$Z_i = \alpha + \beta_{1i}X_{1i} + \beta_{2i}X_{2i} + \dots + \beta_{ki}X_{ki}$$

$$\ln(\text{chance}_{Y=1}) = Z_i$$

onde a chance de $Y = 1$ é definida por

$$\begin{aligned} chance_{Y=1} &= \frac{p}{1-p} \Rightarrow \\ \ln\left(\frac{p_i}{1-p_i}\right) &= Z_i \Rightarrow \\ \frac{p_i}{1-p_i} &= e^{Z_i} \Rightarrow \\ p_i &= \frac{e^{Z_i}}{1+e^{Z_i}} = \frac{1}{1+e^{-Z_i}} \Rightarrow \\ 1-p_i &= 1 - \frac{e^{Z_i}}{1+e^{Z_i}} = \frac{1}{1+e^{Z_i}} \end{aligned}$$

Como não se pode usar OLS, a regressão logística binária é estimada por máxima verossimilhança, pois $Y_i = 1$ ou $Y_i = 0$.

Nesse caso, temos:

$$p(Y_i) = p_i^{Y_i} \cdot (1-p_i)^{1-Y_i} \Rightarrow L = \prod_{i=1}^n [p_i^{Y_i} \cdot (1-p_i)^{1-Y_i}] \text{ (função V.S.)}$$

O objetivo obter o valor de p que maximiza o logaritmo da função V.S.:

$$\max LL = \sum_{i=1}^n \left\{ \left[Y_i \cdot \ln\left(\frac{1}{1+e^{-Z_i}}\right) \right] + \left[(1-Y_i) \cdot \ln\left(\frac{1}{1+e^{Z_i}}\right) \right] \right\}$$

Exemplo de aplicação da regressão logística binária

Sem dúvidas, duas das principais variáveis que afetam diretamente o tempo de chegada dos estudantes à escola são:

- a distância entre a residência do aluno e escola que frequenta;
- a velocidade do transporte.

Neste exemplo, vamos supor que:

1. Os estudantes pesquisados deslocam-se para a escola por meio de ônibus (transporte público);
2. Para cada estudante, foram obtidos a distância casa/escola em Km (X_1) e o número de paradas que o ônibus realizou nesse trajeto (X_2).

3. A variável de interesse será uma variável indicadora binária Y , tal que $Y = 1$, se chegou atrasado e $Y = 0$, caso contrário.

O que pretende é obter a probabilidade de um estudante, dentre os pesquisados, chegar atrasado para a aula. Para tanto, será utilizado o banco de dados da Tabela 17, na qual para cada aluno temos a distância da residência até a escola em Km e o número de paradas de ônibus casa/escola. Naturalmente, esse é um problema de regressão logística binária. O programa 15 a seguir, usando o software R, com os dados da Tabela 17, resolve essa questão, por meio de regressão logística binária.

Programa 17 - R: Construção da curva ROC, obtenção da AUC e do Gini

```
#####
#          INSTALAÇÃO E CARREGAMENTO DE PACOTES NECESSÁRIOS          #
#####

#Pacotes utilizados
pacotes <- c("ggplot2","plotly","tidyverse","knitr","kableExtra","fastDummies","rgl","car",
            "reshape2","jtools","lmtest","caret","pROC","ROCR","nnet")
if(sum(as.numeric(!pacotes %in% installed.packages())) != 0){
  instalador <- pacotes[!pacotes %in% installed.packages()]
  for(i in 1:length(instalador)) {
    install.packages(instalador, dependencies = T)
    break()}
  supply(pacotes, require, character = T)
} else {
  supply(pacotes, require, character = T)
}

#####
#          REGRESSÃO LOGÍSTICA BINÁRIA - PARTE CONCEITUAL          #
#####

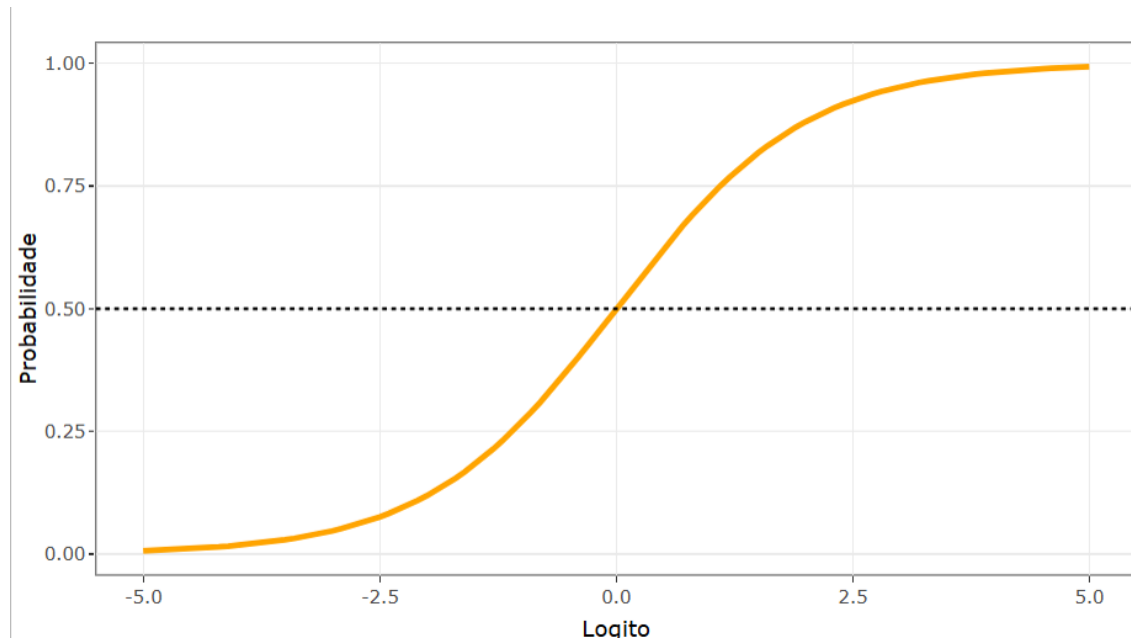
####Estabelecendo uma função para a probabilidade de ocorrência de um evento
library(ggplot2)
library(plotly)
# Definição da função (exemplo: função logística)
curva_evento <- function(x) {
  1 / (1 + exp(-x)) # ou plogis(x)
}
p <- ggplot(tibble::tibble()) +
  stat_function(
    fun = curva_evento,
    xlim = c(-5,5),
```

```

size = 1,
color = "orange"
) +
geom_hline(yintercept = 0.5, linetype = "dotted") +
labs(x = "Logito", y = "Probabilidade") +
theme_bw()
ggplotly(p)

```

Figura 20 – Curva logística no intervalo [-5,5]



Fonte: Elaboração do Autor

```

#####
REGRESSÃO LOGÍSTICA BINÁRIA - CARREGAMENTO DA BASE DE DADOS
#####

```

```

library(readxl)
getwd()
read_excel("C:/Users/USER/Desktop/maximavero.xlsx")
dados<-read_excel("C:/Users/USER/Desktop/maximavero.xlsx")
head(dados)
save(dados,file="dados.RData")
dados
head(dados)
load(file = "dados.RData")

```

```
#####
##      OBSERVAÇÃO DA BASE DE DADOS  ## dados ##
#####
```

```
###Visualizando a base de dados
dados %>%
  kable() %>%
  kable_styling(bootstrap_options="striped",
    full_width = F,
    font_size = 12)
#Estatísticas descritivas univariadas da base de dados
summary(dados)
#Tabela de frequências absolutas da variável 'Y'
table(dados$Y)
```

```
#####
#      ESTIMAÇÃO DE UM MODELO LOGÍSTICO BINÁRIO      #
#####
```

```
modelo_dados <- glm(formula = Y ~ X1 + X2,
  data = dados,
  family = "binomial")
#Parâmetros do modelo_dados
summary(modelo_dados)
#Outro modo de apresentar os outputs do modelo
summ(modelo_dados, confint = T, digits = 3, ci.width = .95) #função summ do pacote jtools
export_summs(modelo_dados, scale = F, digits = 6) #função export_summs do pacote jtools
#Extração do valor do LL
logLik(modelo_dados)
#Fazendo predições para o modelo_dados. Exemplo: qual a probabilidade média
#de se chegar atrasado quando o trajeto tem 7 km e passa-se por 10 paradas de ônibus no
percurso?
predict(object = modelo_dados,
  data.frame(X1 = 7, X2 = 10),
  type = "response")
```

```
#####
#      CONSTRUÇÃO DE UMA MATRIZ DE CONFUSÃO      #
#####
```

```
# Adicionando os valores previstos de probabilidade da base de dados
dados$probit <- modelo_dados$fitted.values
#Visualizando a base de dados com a variável probit
dados %>%
  kable() %>%
  kable_styling(bootstrap_options = "striped",
    full_width = F,
    font_size = 12)
#Matriz de confusão (para cada cutoff)
confusionMatrix(table(predict(modelo_dados, type = "response") >= 0.5,
  dados$Y == 1)[2:1, 2:1])
#função confusionMatrix do pacote caret#
```

```
#####
```

```
#          CONSTRUÇÃO DA CURVA ROC          #
#####
ROC <- roc(response = dados$Y,
            predictor = modelo_dados$fitted.values)
ggplotly(
  ggroc(ROC, color = "darkorchid", size = 1) +
    geom_segment(aes(x = 1, xend = 0, y = 0, yend = 1),
                 color="orange",
                 size = 0.2)+
  labs(x = "1 - Especificidade",
       y = "Sensitividade",
       title = paste("Area abaixo da curva:",
                     round(ROC$auc, 3),
                     "|",
                     "Coeficiente de Gini",
                     round((ROC$auc[1] - 0.5) / 0.5, 3))) +
  theme_bw()
)
#####
```

Fonte: Elaboração do Autor.

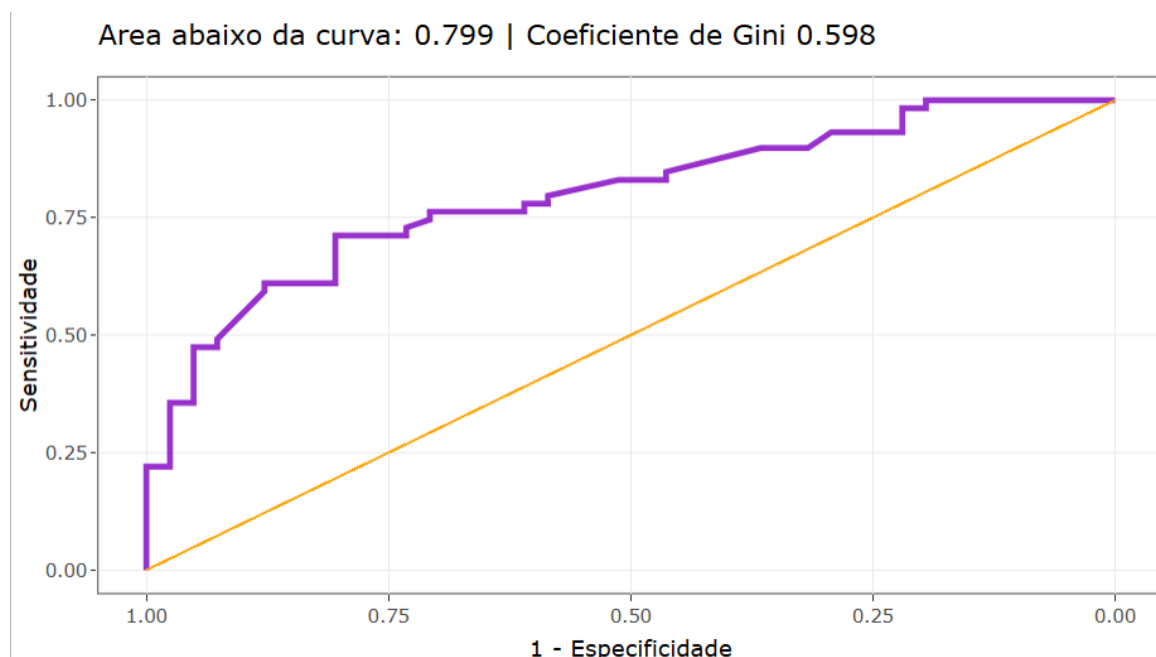
Tabela 18 - Matriz de confusão para cutoff igual a 0,5

	TRUE	FALSE
TRUE	46	16
FALSE	11	25

Métrica:

Acurácia: **0.71**; Sensibilidade: **0.7797**; Especificidade: **0.6098**

Figura 21 - Curva ROC, área sob a curva (AUC) e índice de GINI



Fonte: Elaboração do Autor.

Apresenta-se a seguir o banco de dados utilizado¹⁴

Tabela 19 - Banco de dados – maximavero.xlsx

Estudante	Atrasado (Y)	Distância (X1)	Nº. Paradas (X2)
Abel	1	10	10
Aderbal	1	12	10
Alberto	1	22	11
Alex Sandro	1	28	10
Alexandre	0	14	10
Alexandrina	0	11,6	10
Alison	1	13	11
Amélia	0	13	10
Ana Célia	1	10,5	10
Ana Paula	0	23,5	7
Ananda	0	12	7
Angélica	1	9,4	10
Ângelo	0	9,5	10
Anita	0	13	10
Antônio	0	11,5	10
Ariel	0	12	6
Bento	1	17	10
Bremer	1	1	12
Breno	1	14,8	10
Bruno	0	9,5	8
Carlos	0	12,3	10
Carmela	0	12,9	10
Carmem	1	12,8	11
Carolina	1	10,5	10
Casemiro	1	17	10
Cátia	1	17	10
Cícero	1	15	10
Clarice	1	12	10
Claudete	0	13,3	10
Claudio	0	11	10
Colbert	1	11	18
Cristina	1	11,5	10
Danilo	1	17	10
Deliane	0	12	10
Diego	1	15	10
Douglas	1	12	11
Ederson	1	13	10
Everson	1	11	17
Endrik	1	14	10

¹⁴ Fávero; Belfiori, 2020, p. 615, com adaptações.

Erika	0	13,4	8
Estevan	1	13,8	10
Fabio	0	12,5	7
Fatima	1	15,5	11
Felipe	1	19	10
Flávio	1	10,1	15
Gabriel	0	14,7	10
Giulia	0	15,4	10
Gustavo	1	22	11
Hamilton	1	20	10
Ida	0	13	5
Igor	1	18,7	10
Ivone	1	14,2	11
Jessica	1	17	10
Joao	0	11	11
João Carlos	1	15,4	11
Jorge	1	18,7	10
José Carlos	1	12	10
Leonardo	1	18	10
Leonídia	1	24	10
Leonor	0	14,4	10
Lourdes	1	17	10
Lucas	1	10,2	11
Lucia	0	14	11
Lucio	1	14,2	10
Luiza	0	15,8	10
Luiz			
Henrique	1	12,2	10
Marcela	1	11,5	10
Marcos	1	12,1	11
Maria	0	13,5	10
Marina	1	1	13
Marinalva	0	16	10
Nádia	0	13	10
Nasser	1	29	10
Natalia	0	14,7	10
Neymar	1	15	10
Nikoleta	0	13,7	10
Norai	1	15	10
Odete	0	11,7	10
Orlando	0	12	10
Oscar	0	12	10
Paulo	0	12	10
Pedro	0	17	10
Rafael	0	13,5	10

Regiane	0	18,4	10
Regina	1	19	10
Ricardina	0	11	10
Ricardo	0	13,5	10
Roberto	1	19	10
Ronaldo	1	12	11
Rosa	1	13,2	10
Sofia	1	16,5	10
Susana	0	11,2	10
Tereza	1	15	10
Thayro	1	13	10
Vicente	0	17	10
Victor	1	17	12
Vitoria	1	19	12
Wesley	1	1	11
Weverton	1	13	11
Zelia	0	12	9