UNIVERSIDADE FEDERAL DE SERGIPE

CENTRO DE CIÊNCIAS EXATAS E TECNOLOGIA

DEPARTAMENTO DE COMPUTAÇÃO

Uso de Redes Bayesianas Multinível para prever a evasão estudantil no Departamento de Computação da Universidade Federal de Sergipe

Trabalho de Conclusão de Curso

Eduardo Fillipe da Silva Reis

São Cristóvão – Sergipe

2023

UNIVERSIDADE FEDERAL DE SERGIPE

CENTRO DE CIÊNCIAS EXATAS E TECNOLOGIA

DEPARTAMENTO DE COMPUTAÇÃO

Eduardo Fillipe da Silva Reis

Uso de Redes Bayesianas Multinível para prever a evasão estudantil no Departamento de Computação da Universidade Federal de Sergipe

Trabalho de Conclusão de Curso submetido ao Departamento de Computação da Universidade Federal de Sergipe como requisito parcial para a obtenção do título de Bacharel em Ciência da Computação.

Orientador(a): Prof. Dr. Carlos Alberto Estombelo Montesco

São Cristóvão – Sergipe

2023

Dedico este trabalho com profundo agradecimento e gratidão aos meus familiares, professores e amigos, cujo inabalável apoio e encorajamento foram fundamentais para que eu pudesse concluir esta jornada com sucesso.

Agradecimentos

Primeiramente, agradeço à minha família por todo o apoio dado para concretizar este sonho. Principalmente à minha mãe, Vanda, que trabalhou incansavelmente e sempre investiu na minha educação desde os primeiros momentos da minha vida, abrindo para min os caminhos aos quais ela não teve acesso. Por isso, sempre estarei em dívida.

Agradeço também, à todos os professores que me apoiaram nesta jornada. Ao professor Carlos Estombelo, por se mostrar ser mais que um orientador, mas uma pessoas com quem podia contar e que sempre esteve disponível para ouvir as minhas dúvidas mais improváveis, mesmo nos horários mais inoportunos. À professora Leila Silva, por me ensinar não apenas os rigores da teoria da computação, mas também por iluminar os meus pensamentos com suas experiências de vida e da academia, fazendo um graduando sonhar em, um dia, ser uma fração da professora que ela é. Aos professores Ricardo e Edilaine Salgueiro, por sempre manterem as portas do laboratório de redes abertas para min e me darem a oportunidade de dar os primeiros passos na pesquisa acadêmica. Por fim, agradeço ao professor Márcio Borges, que mesmo sendo de outra instituição, nos apoiou na validação do *dataset* sintético utilizado neste trabalho, nos ajudando a seguir com mais segurança no restante dos estudos.

Não poderia deixar de agradecer também à *LabSit*, especialmente à Dimas Augusto e Bryanne Araújo, que sempre acreditaram no meu potencial e me apoiaram direta e indiretamente para dar mais esse passo, colocando minha educação como prioridade e abrindo portas para me tornar o profissional, ou melhor, *o humano*, que sou hoje.

À todas as pessoas que trabalham para fazer o DCOMP um lugar melhor. Adnésia, Elaine e Elder, com quem tive mais contato e que sempre me trataram da melhor maneira e estiveram dispostos a ajudar com o que fosse necessário, fazendo da minha curta estadia naqueles corredores mais simples e amena.

Por último, mas não menos importante, gostaria de agradecer à todos os meus colegas e amigos que me acompanharam nessa jornada. Thiago, Rodolfo, Ednilson, Filipe e Adisiel. Nunca esquecerei das conversas nos laboratórios, das noites em claro estudando, dos momentos de tensão antes das provas, das discussões nos trabalho em grupo ou das conversas durante o almoço. Vocês fizeram esses anos mais leves.

Muito obrigado, à todos vocês, por fazerem parte desta caminhada, irei sempre guardá-los na memória e no coração.

Ciência da computação não é a ciência dos computadores,
assim como a astronomia não é a
ciência dos telescópios.
(Edsger W. Dijkstra)

Resumo

O problema da evasão estudantil no ensino superior é um desafio que acarreta em prejuízos financeiros e emocionais aos envolvidos. Assim, este trabalho tem como objetivo estudar o tema no Departamento de Computação da UFS e desenvolver um modelo baseado em Redes Bayesianas Multinível (RBM) para prever quais estudantes estão mais propensos a evadir seus cursos. O modelo de RBM resultante foi capaz de identificar estudantes em risco de evasão com acurácia de 88% e AUC de 0.86. Adicionalmente, o trabalho disponibilizou um *dataset* com mais de 135 mil observações de componentes curriculares de 3588 alunos do DCOMP e ferramentas para atualização contínua desses dados.

Palavras-chave: evasão estudantil; ensino superior; redes bayesianas multinível; análise multinível;

Abstract

The problem of student dropout in higher education is a challenge that results in financial and emotional losses to those involved. Thus, this work aims to study the topic in the Department of Computing at UFS and develop a model based on Multilevel Bayesian Networks (MBN) to predict which students are more likely to drop out of their courses. The resulting MBN model was able to identify students at risk of dropping out with an accuracy of 88% and an AUC of 0.86. Additionally, the work made available a dataset with over 135,000 observations of curricular components from 3,588 DCOMP students, as well as tools for continuous updating of this data.

Keywords: student dropout; higher education; multilevel bayesian networks; multilevel analysis;

Lista de ilustrações

Lista de quadros

Lista de tabelas

Lista de algoritmos

Lista de abreviaturas e siglas

AES	*Advanced Encryption Standard*
AUC	*Area Under the ROC Curve*
CAPES	Coordenação de Aperfeiçoamento de Pessoal de Nível Superior
CBC	*Cipher Block Chaining*
CRISP-DM	*Cross Industry Standard Process for Data Mining*
DAG	*Directed acyclic graph*
DCOMP	Departamento de Computação
FMPC	Função de Massa de Probabilidade Conjunta
HTML	*HyperText Markup Language*
IA	Inteligência Artificial
KNN	*K Nearest Neighbours*
MOOC	*Massive Online Courses*
PDF	*Portable Document Format*
RBM	Rede Bayesiana Multinível
ROC	*Receiver Operating Characteristic Curve*
SIGAA	Sistema Integrado de Gestão de Atividades Acadêmicas
SQL	*Structured Query Language*
SVM	*Support Vector Machines*
UFS	Universidade Federal de Sergipe

Lista de símbolos

Ω/ω	Letra grega Ômega
$\subseteq$	Está contido ou é
$\cap$	Intersecção entre dois conjuntos
$\cup$	União entre conjuntos
$\in$	Pertence à um conjunto
$\sum$	Somatório
$\prod$	Produtório
$\neg$	Negação
$\leq$	Menor ou igual à
$\neq$	Diferente
$\geq$	Maior ou igual que
$\Rightarrow$	Implica
$\perp\!\!\!\perp$	Independência estatística

Sumário

1 Introdução

1.1 Motivação

A evasão no ensino superior é tema de debate em todo o mundo e que acontece tanto em países desenvolvidos quanto em desenvolvimento, segundo dados do NCES (2020). Além disso, é um problema que atinge instituições de ensino públicas e privadas, mostrando-se um obstáculo a ser superado por toda a sociedade.

Trata-se de uma questão grave e complexa, capaz de afetar não somente os estudantes, mas também todo o país. Tal falha no sistema educacional resulta em desperdício de recursos públicos, além de impedir a formação de profissionais capacitados e a produção científica e tecnológica. Assim, é fundamental que sejam adotadas medidas eficazes para solucionar esse problema e garantir um ensino superior de qualidade para todos.

Segundo Bovo (2022), no Brasil, para cada aluno de instituições federais do ensino superior são investidos, em média, R$27.850,00 por ano. Isso, aliado ao fato de que em média, no Brasil, 32,34% dos alunos matriculados em cursos do ensino superior abandonaram os estudos (LüDER, 2022), torna evidente que o problema da evasão é um tópico que merece atenção imediata quando analisado através do aspecto econômico.

Além disso, há também o aspecto psicológico. Este, afeta diretamente os estudantes que, pelos mais diversos motivos, optam ou são obrigados a abandonar os estudos. Santos, Pilatti e Bondarik (2022) argumentam que o ingresso na universidade é um momento complexo, marcado de incertezas e mudanças na vida do estudante. Dessa forma, além dos prejuízos institucionais, a evasão ainda produz prejuízos materiais e psicológicos nos evadidos.

Portanto, é necessário que o governo e as instituições de ensino atuem de forma conjunta afim de reduzir os níveis de evasão no sistema educacional brasileiro. Para isso, saber quais são as instituições mais afetadas pelo problema e as suas causas, além redirecionar recursos e desenvolver

estratégias à nível federal são ações essenciais para contornar o problema. Aliado a isso, também é possível investir em abordagens capazes de antecipar o problema, identificando, individualmente, quais estudantes possuem uma maior tendência de evadir o curso ou instituição, antes que venha acontecer. Assim, este trabalho sugere uma abordagem automatizada de identificação de estudantes em risco de evasão estudantil, que permita que as instituições de ensino possam prestar, de forma mais individualizada, atendimento e apoio à esses estudantes.

1.2 Objetivos

O objetivo principal deste trabalho é desenvolver um modelo computacional automatizado para prever a evasão escolar de estudantes no Departamento de Computação (DCOMP) da Universidade Federal de Sergipe, além de analisar as relações entre os dados estudados que mais contribuem para o aumento dos índices de evasão.

O objetivo secundário é identificar as possíveis causas da evasão escolar por meio de Redes Bayesianas Multinível, que permitem modelar a relação entre os fatores que influenciam o problema e identificar aqueles com maior impacto. Dessa forma, será possível identificar quais estudantes precisam de mais atenção durante o programa de graduação em computação, qual tipo de suporte cada um deles necessita e quais são as principais dificuldades enfrentadas por cada curso.

Por fim, este trabalho também visa a criação de um *dataset* sobre a performance estudantil no DCOMP, que poderá ser utilizado por terceiros em pesquisas científicas futuras, respeitando a privacidade dos envolvidos.

1.3 Metodologia

Gil (2002), descreve a metodologia de pesquisa como o processos de definição dos procedimentos e métodos que serão utilizados no decorrer do trabalho. Ele ainda afirma as pesquisas científicas podem ser de três naturezas, não mutuamente exclusivas: exploratória, descritiva ou explicativa. As pesquisas exploratórias tem por objetivo proporcionar maior familiaridade com o problema, proporcionando uma maior flexibilidade de planejamento. As pesquisas de natureza descritiva tem o objetivo de descrever as características de populações, grupos, fenômenos ou estabelecer relações entre variáveis. Já as pesquisas explicativas são as que mais aprofundam o conhecimento da realidade, sendo mais complexas, mais suscetíveis à erros e pretendem explicar as razões e porquês dos fenômenos estudados.

Assim, em consonância com os objetivos deste trabalho, optou-se por utilizar uma metodologia de pesquisa exploratória. Essa metodologia foi escolhida pois, neste trabalho, o problema da evasão estudantil não será apenas avaliado de forma descritiva, mas como um estudo de caso para a aplicação das Redes Bayesianas Multinível.

Para isso, o estudo aqui exposto utilizará dados dos alunos do Departamento de Computação (DCOMP) da Universidade Federal de Sergipe. O DCOMP, existe, com este nome, desde o ano de 2009 e já recebeu mais de 3500 alunos, possuindo 1088 alunos ativos até a data de 10/11/2022. Atualmente, o DCOMP é formado por três cursos de graduação: Ciência da Computação, Engenharia da Computação e Sistemas da Informação. Neste trabalho serão utilizados dados desses três cursos e de alunos ativos e inativos do departamento.

Os dados dessa amostra populacional da UFS contam com as informações da performance universitária de alunos ativos e inativos do departamento de computação da universidade. A coleta desses dados será feita de forma automatizada e anonimizada, através de softwares desenvolvidos para esse fim durante o estudo. As informações de performance de cada aluno serão extraídas do histórico do ensino superior de cada um dos envolvidos neste estudo. Para a coleta dos históricos, em parceria com a secretaria do DCOMP, serão utilizadas técnicas de WebScraping. Através desse método, será possível fazer o download automático de cada histórico, a partir do Sistema Integrado de Gestão de Atividades Acadêmicas (SIGAA), que é o sistema acadêmico adotado pela universidade. Isso irá permitir a obtenção dos dados de maneira mais eficiente e precisa.

Assim, será realizada a extração e análise dos dados dos históricos dos alunos. Para isso, se fará uso do Processo Padrão Transversal da Indústria para Mineração de Dados (CRISP-DM) (SHEARER, 2000), ilustrada na Figura 1. A metodologia CRISP-DM é um modelo de processo cíclico e de melhoria contínua utilizada na análise e ciência de dados, sendo composta por seis etapas: entendimento do negócio, entendimento dos dados, preparação dos dados, modelagem, avaliação e implantação. Essas etapas se retroalimentam, até que o resultado esteja pronto para ser disponibilizado ao usuário, na etapa de implantação.

1.4 Organização do documento

Dessa forma, os próximos capítulos descrevem como se deu a execução do trabalho, com base nos objetivos apresentados.

Assim, no capítulo 2 será exposto o Mapeamento sistemático (PETERSEN et al., 2008) em bases científicas com a finalidade de mapear o estado da arte da área de pesquisa, mostrando como outros pesquisadores trataram o problema, trazendo de forma visual um resumo dos estudos dentro do tema de pesquisa nos últimos anos.

No capítulo 3 será apresentado o embasamento teórico necessário para desenvolver o trabalho, ilustrando teoremas e modelos utilizados durante o estudo e necessários para a compreensão do restante do documento.

O capítulo 4 mostrará como se deu o experimento realizado a partir de um conjunto de dados sintéticos, com o objetivo didático de aprendizado da teoria estudada.

Nos capítulos 5, 6, 7 e 8, seguindo a metodologia CRISP-DM, serão abordados os temas

Figura 1 – Etapas da metodologia CRISP-DM

Fonte: Shearer (2000)

referentes ao desenvolvimento dos objetivos centrais deste trabalho, desde o Entendimento do negócio e dos dados, até a validação do modelo treinado.

O capítulo 9 trará os resultados da execução do modelo treinado sobre os dados de alunos ativos do DCOMP, realizando uma discussão sobre os resultados.

Por fim, o capítulo 9 trará as conclusões deste trabalho, evidenciando os resultados, discussões e futuros trabalhos a serem desenvolvidos.

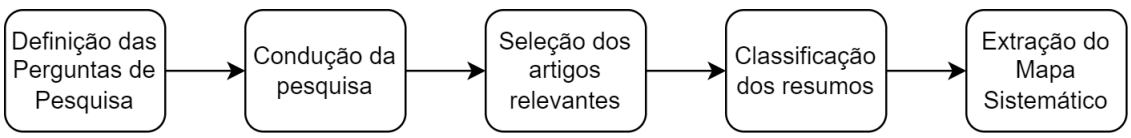
2

Mapeamento Sistemático

Dado o objeto de estudo desse trabalho, a evasão estudantil em programas de graduação da Universidade Federal de Sergipe, foi realizado um levantamento de como problemas similares foram solucionados por outros pesquisadores, a fim de descobrir quais técnicas e conjuntos de dados foram utilizados ao longo do tempo, dentro dessa área de pesquisa e, assim, aproveitar tais resultados de forma comparativa e norteadora para a realização desse trabalho. Para isso, foi utilizado como referência o trabalho de Petersen et al. (2008), que apresenta uma metodologia para mapeamentos sistemáticos no campo da Ciência da Computação, que, empregando um método sistematizado, tem como resultado final uma visão geral da área de pesquisa alvo na forma de um sumário visual.

Esta metodologia é composta por um processo de cinco etapas. Cada uma dessas etapas produz um resultado, uma parcela do mapeamento sistemático, que por si só possui valor, e, quando combinados, formam o mapeamento sistemático como um todo. O resultado de cada um desses estágios é fornecido como entrada para a seguinte, formando uma cadeia, como ilustra a Figura 2. A próxima listagem, resume o objetivo de cada uma dessas etapas e quais seus respectivos resultados.

Figura 2 – Sequência de etapas para a confecção de um mapeamento sistemático.



Fonte: Petersen et al. (2008)

1. **Definição das perguntas de pesquisa:** provê uma visão geral do objetivo do mapeamento sistemático e define quais questões o trabalho pretende responder.

2. **Condução da pesquisa:** onde o pesquisador realiza a busca de artigos a partir de palavras-chave extraídas das perguntas de pesquisa.
3. **Seleção de artigos relevantes:** onde o pesquisador filtra ou inclui os artigos encontrados na etapa anterior com base em condições pre-definidas.
4. **Classificação dos resumos:** onde são extraídas palavras-chave que representam os principais aspectos de cada artigo. Em seguida, os artigos são classificados com base nessas palavras-chave.
5. **Extração dos dados e mapeamento dos estudos:** nessa etapa, por fim, são extraídos gráficos e *insights* sobre os artigos classificados na etapa anterior, obtendo-se assim uma ampla visão do campo de pesquisa.

Dessa forma, nesse capítulo serão demonstrados os resultados da aplicação dessas etapas, culminando, por fim, no Mapeamento sistemático, interpretação e discussão dos resultados encontrados.

2.1 Definição das perguntas de pesquisa

A Tabela 1 exibe as três perguntas de pesquisa que foram utilizadas para nortear este mapeamento sistemático. As perguntas foram derivadas a partir do tema deste trabalho: *A evasão estudantil no âmbito do ensino superior*. A primeira pergunta, que tem o objetivo de esclarecer a quantidade e frequência dos estudos com esse tema de pesquisa, fornecendo uma visão quantitativa e atualizada da área. As duas últimas, buscam compreender, de forma qualitativa, como o problema da evasão estudantil no ensino superior vem sendo tratado por pesquisadores e instituições, obtendo assim uma ampla visão das técnicas utilizadas na abordagem deste tema.

Tabela 1 – Perguntas de pesquisa do mapeamento sistemático

Número	Pergunta de pesquisa
1	Quais e quantos são os artigos que tratam sobre a evasão estudantil em cursos do ensino superior?
2	Quais foram os métodos e técnicas usados pelos pesquisadores para realizar esses estudos?
3	Quais são os principais conjuntos de dados utilizados e analisados nesses artigos?

2.2 Condução da pesquisa

As buscas foram conduzidas nas 4 bases de artigos científicos mais utilizadas na Ciência da computação, além disso também do motor de busca da CAPES para enriquecer os resultados. A lista dessas bases e seus respectivos endereços eletrônicos é exibida na Tabela 2.

Tabela 2 – Bases utilizadas na busca do Mapeamento sistemático

Base	Endereço da *WEB*
ACM	<https://dl.acm.org/>
IEEE Explore	<https://ieeexplore.ieee.org/Xplore/>
Springer Link	<https://link.springer.com/>
Science Direct	<https://www.sciencedirect.com/>
Periódicos Capes	<https://www.periodicos.capes.gov.br/>

A pesquisa foi conduzida em cada uma das bases utilizando a *string* de busca do Quadro 1, construída a partir das perguntas de pesquisa da Tabela 1.

Quadro 1 – *String* de busca utilizada na condução da pesquisa

(("student"OR "school") AND ("dropout"OR "evasion"OR "drop out")) AND ("prediction"OR "estimation"OR "estimating"OR "predict"OR "study case") AND ("higher education"OR "high school")

Após a busca, foram localizados um total de 1156 artigos, conforme a Tabela 3 que explicita a quantidade de artigos localizada em cada base da Tabela 2.

Tabela 3 – Quantidade de artigos primários localizados em cada base de dados

Base	Quantidade de artigos primários
ACM	33
IEEE Explore	42
Springer Link	495
Science Direct	432
Periódicos Capes	156

2.3 Seleção de artigos relevantes

Para a filtragem dos artigos primários citados na Tabela 3, foram utilizados 4 critérios que aparecem listados na Tabela 4.

Após o processo de seleção, a quantidade de artigos foi reduzido à um total de 34 artigos relevantes para o processo de mapeamento sistemático, sumarizados na Tabela 5.

Tabela 4 – Critérios de seleção dos artigos primários da busca

Inclui-se documentos que	Excluem-se documentos que
Sejam livros, artigos ou relatórios técnicos que busquem prever as causas ou os índices de evasão escolar no ensino superior	Que não tratem da evasão estudantil no âmbito do ensino superior
Que exponham as técnicas e dados utilizados durante o experimento	Estão fora da área de estudos da computação

Tabela 5 – Resultado do processo de seleção

Título	Referência
Predicting academic performance using tree-based machine learning models	Zhang, Wang e Wang (2022)
Application of logistic regression to predict the failure of students in subjects of a mathematics undergraduate course	Costa e Diniz (2022)
FairEd: A Systematic Fairness Analysis Approach Applied in a Higher Educational Context	Verdugo et al. (2022)
Implementation of a Predictive Information System for University Dropout Prevention	Guzmán-Castillo et al. (2022)
Predicting students at risk of academic failure using ensemble model during pandemic in a distance learning system	Karalar, Kapucu e Gürüler (2021)
System framework of intelligent consulting systems with intellectual technology	Suaprae, Nilsook e Wannapiroon (2021)
A review: Predicting student success at various levels of their learning journey in a science programme	Mabunda, Jadhav e Ajoodha (2021)
A data-driven approach to predict first-year students' academic success in higher education institutions	Gil et al. (2021)
University dropout prevention through the application of big data	Shiau (2020)
Should College Dropout Prediction Models Include Protected Attributes?	Yu, Lee e Kizilcec (2020)
Student Dropout Prediction: A KMUTT Case Study	Tenpipat e Akkarajitsakul (2020)
n-Gage: Predicting in-class Emotional, Behavioural and Cognitive Engagement in the Wild	Gao et al. (2020)
EvolveDTree: Analyzing Student Dropout in Universities	Santos et al. (2020)

Continua na próxima página

Tabela 5 – *Continuação*

Título	Referência
Predictive model to reduce the dropout rate of university students in Perú: Bayesian Networks vs. Decision Trees	Medina et al. (2020)
Predicting student dropout in subscription-based online learning environments: The beneficial impact of the logit leaf model	Coussement et al. (2020)
Predicting dropout using high school and first-semester academic achievement measures	Kiss et al. (2019)
Bayesian Classifier Applied to Higher Education Dropout	Viloria, Lezama e Varela (2019)
Transfer Learning using Representation Learning in Massive Open Online Courses	Ding et al. (2019)
An analysis of student representation, representative features and classification algorithms to predict degree dropout	Manrique et al. (2019)
Ensemble regression models applied to dropout in higher education	Silva et al. (2019)
Mixture Structural Equation Models for Classifying University Student Dropout in Latin America	Viloria e Lezama (2019)
Predicting Dropout in Higher Education Based on Secondary School Performance	Nagy e Molontay (2018)
Comparative Analysis of Prediction Techniques to Determine Student Dropout: Logistic Regression vs Decision Trees	Pérez et al. (2018)
MOOC dropout prediction using machine learning techniques: Review and research challenges	Dalipi, Imran e Kastrati (2018)
Analyze and Predict Student Dropout from Online Programs	Kang e Wang (2018)
Running out of STEM: A Comparative Study across STEM Majors of College Students At-Risk of Dropping Out Early	Chen, Johri e Rangwala (2018)
Predicting high-risk students using Internet access logs	Zhou et al. (2018)
Higher education student dropout prediction and analysis through educational data mining	Hegde e Prageeth (2018)
Students' performance tracking in distributed open education using big data analytics	Hussein e Khan (2017)

Continua na próxima página

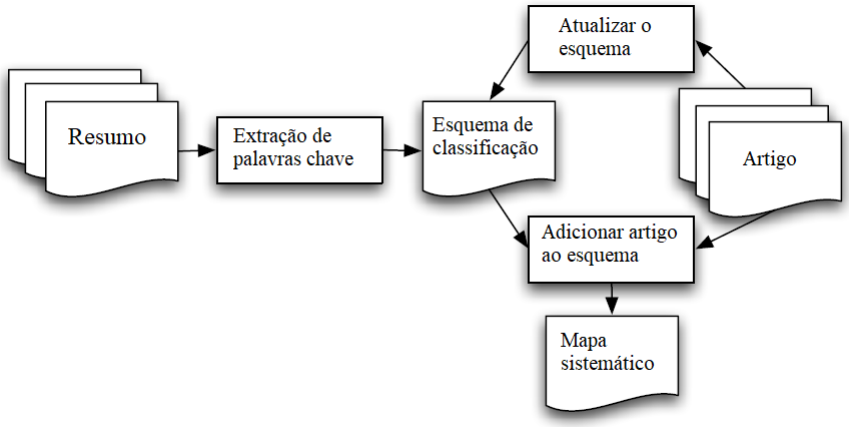
Tabela 5 – *Continuação*

Título	Referência
Survival analysis based framework for early prediction of student dropouts	Ameri et al. (2016)
Estimating student dropout in distance higher education using semi-supervised techniques	Kostopoulos, Kotsiantis e Pintelas (2015)
Predicting student failure at school using genetic programming and different data mining approaches with high dimensional and imbalanced data	Márquez-Vera et al. (2013)
An intelligent system for prediction of school dropout risk group in higher education classroom based on artificial neural networks	Martinho, Nunes e Minussi (2013)

2.4 Classificação dos resumos

Para a classificação do artigos listados na Tabela 3, foi utilizado o processo sistemático ilustrado na figura Figura 3.

Figura 3 – Processo sistemático para classificação dos artigos



Fonte: Petersen et al. (2008)

Este processo, apresentado por Petersen et al. (2008), consiste em ler os resumos e/ou introduções dos artigos, realizando a extração de palavras chave e, então, classificá-las com base num esquema, construído a partir das palavras chave, que ajude a responder as Perguntas de pesquisa, formuladas anteriormente.

Para esse mapeamento sistemático, foram definidas quatro facetas que juntas representam as perguntas de pesquisa e que, combinadas, constroem uma representação de alto nível da área

de pesquisa. O Quadro 2 mostra o esquema de classificação criado a partir da extração das palavras chave dos artigos selecionados.

Quadro 2 – Esquema de classificação dos artigos selecionados

Faceta	**Propósito**	**Possíveis Classificações**	**Número de Artigos**
Dados utilizados	Classificar os artigos quanto aos dados coletados e analisados	Performance do Ensino Médio	6
		Performance Universitária	25
		Dados Socioeconômicos	23
Técnicas empregadas	Classificar os artigos quanto às técnicas e tecnologias utilizadas para tratar o problema	Algoritmos Genéticos	2
		Análise de Sobrevivência	2
		Alg. baseados em Árvores	10
		Computação Cognitiva	2
		Ensemble	2
		KNN	2
		Mixture Structural Equations	1
		Modelo de folha Logit	1
		Naive Bayes	5
		Redes Bayesianas	2
		Redes Neurais	3
		Regressão Logística	6
		Alg. Semi Supervisionados	1
		SVM	1
Tipo de trabalho	Classificar os artigos quanto ao tipo de trabalho, como sugere Petersen et al. (2008)	Avaliação	17
		Validação	0
		Proposta de solução	16
		Trabalho Filosófico	0
Contribuição do trabalho	Classificar em relação à contribuição final que o artigo produz	Processo	6
		Modelo	24
		Ferramenta/Sistema	1

Uma característica importante encontrada nos trabalhos selecionados é a tendência a utilizar mais de um tipo de dado durante as análises, aumentando, assim, a precisão dos algoritmos e melhorando os resultados. O Quadro 3 mostra como os dados da faceta referente aos dados utilizados combinados nos artigos selecionados e quantos artigos se apresentam cada combinação.

2.5 Extração dos dados e mapeamento dos estudos

Conforme sugere Petersen et al. (2008), esta etapa de construção do mapeamento sistemático deve, de forma visual, sumarizar a classificação realizada sobre os artigos selecionados. Tal sumarização é ilustrada nas imagens 4, 5, 6, 7, 8.

As imagens 4 e 5, respectivamente, ilustram como as facetas "Dados utilizados"e "Técnicas empregadas"se relacionam com as facetas "Tipo de trabalho"e "Contribuição do trabalho", onde

Quadro 3 – Esquema de classificação dos artigos selecionados quanto à combinação utilizada dos dados

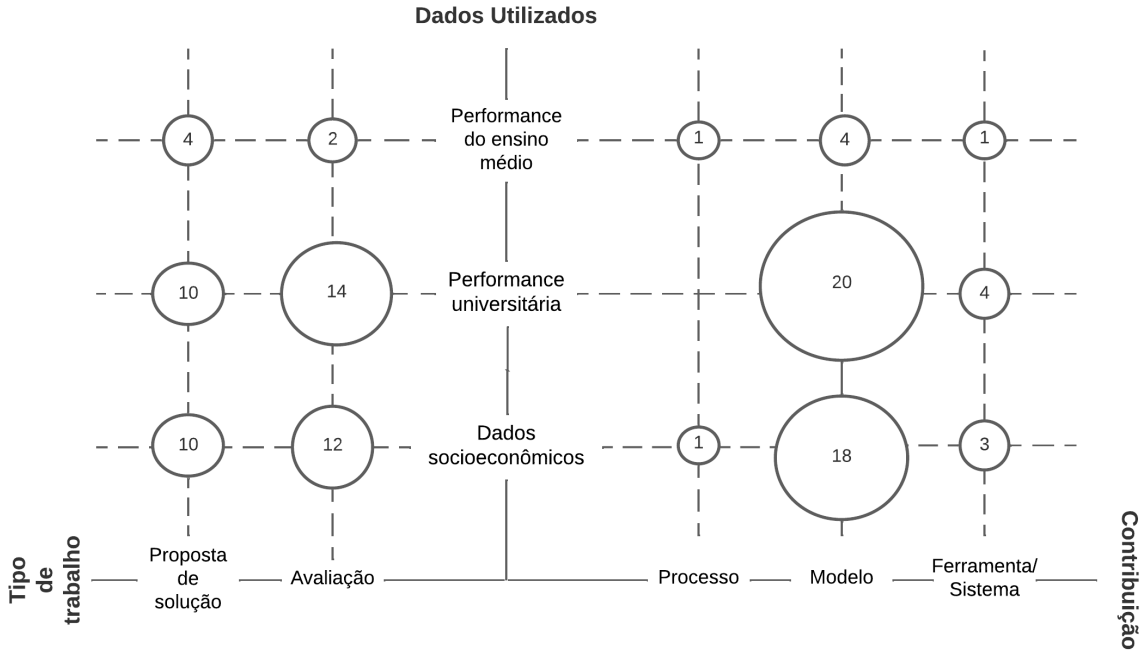
Faceta	**Propósito**	**Possíveis Classificações**	**Número de Artigos**
Dados Utilizados	Classificar os artigos quanto à combinação dos dados utilizados no trabalho	Dados da performance universitária	7
		Dados socioeconômicos	2
		Dados socioeconômicos e da performance universitária	17
		Dados socioeconômicos, da performance universitária e da performance do Ensino Médio	3
		Dados socioeconômicos e da performance do Ensino Médio	2
		Dados da Performance do Ensino Médio	1

cada interseção entre os eixos *x* e *y* indica quantos artigos pertencem a cada valor de faceta simultaneamente.

Por sua vez, as imagens 6 e 7 mostram como o interesse sobre a área de pesquisa variou entre os anos de 2012 e 2022, através da faceta "Dados Utilizados".

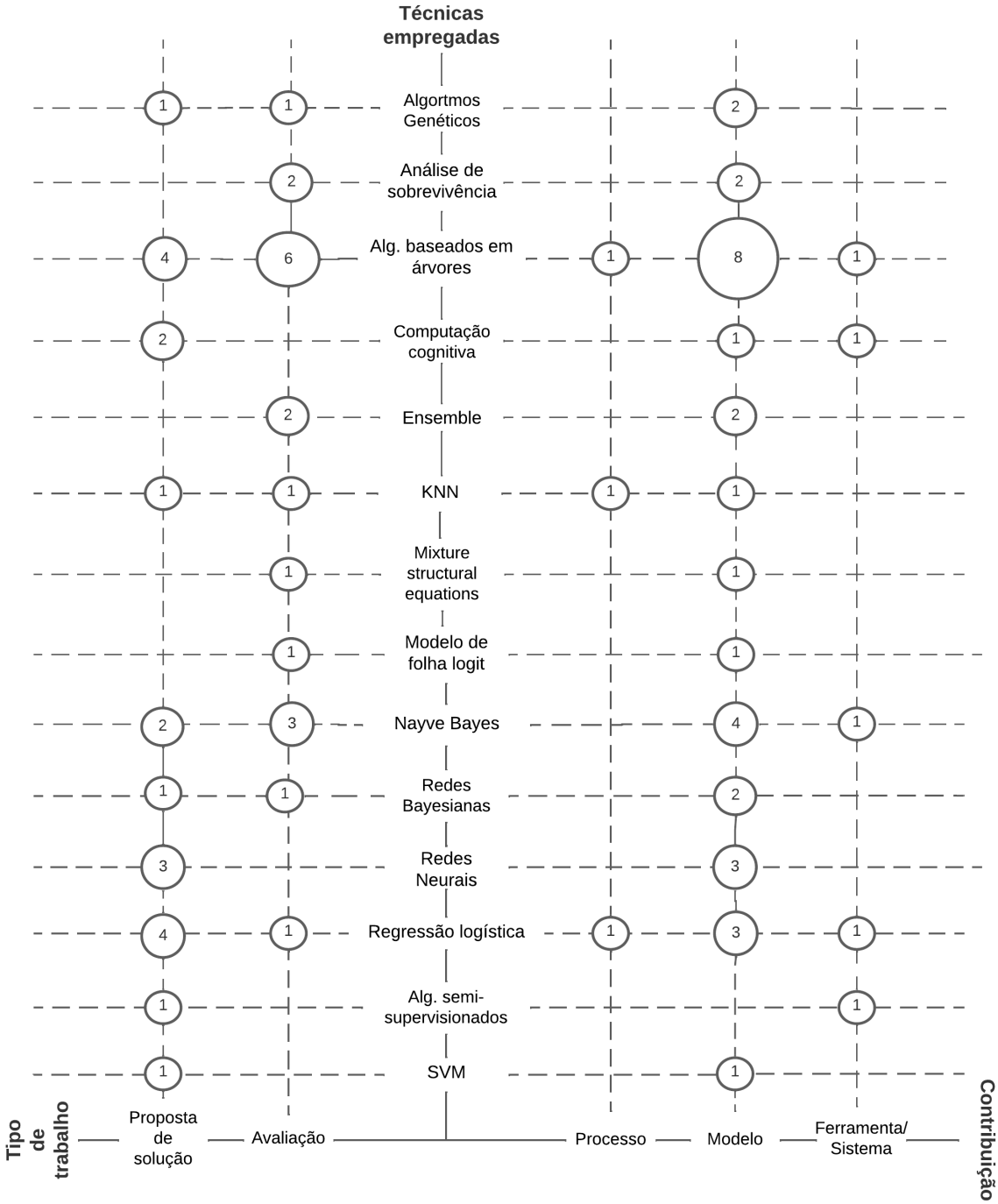
Por fim, a imagem 8 sumariza de forma quantitativa, mostrando quantas vezes técnicas ou algoritmos, foram utilizadas nos trabalhos selecionados.

Figura 4 – Processo sistemático para classificação dos artigos



Fonte: Autor

Figura 5 – Processo sistemático para classificação dos artigos

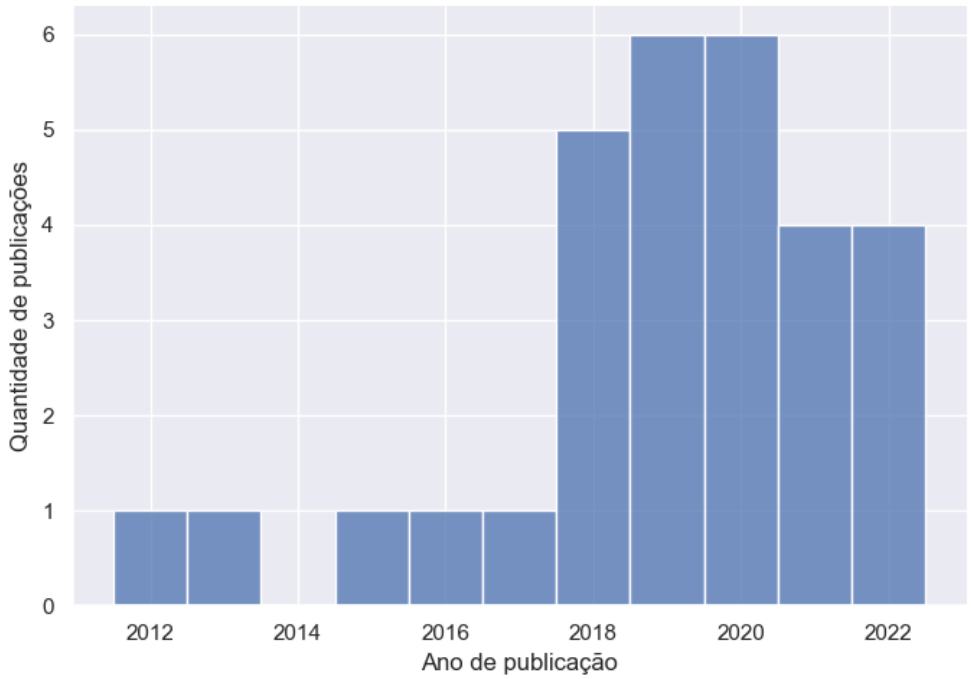


Fonte: Autor

2.6 Interpretação dos resultados

O sumário dos dados, ilustrado nas imagens da seção seção 2.5 deste capítulo, permite uma visão geral de como o problema da evasão estudantil no ensino superior vem sendo tratada dentro do âmbito da computação. Além disso, é possível uma leitura de como as metodologias

Figura 6 – Quantidade de publicações por ano



Fonte: Autor

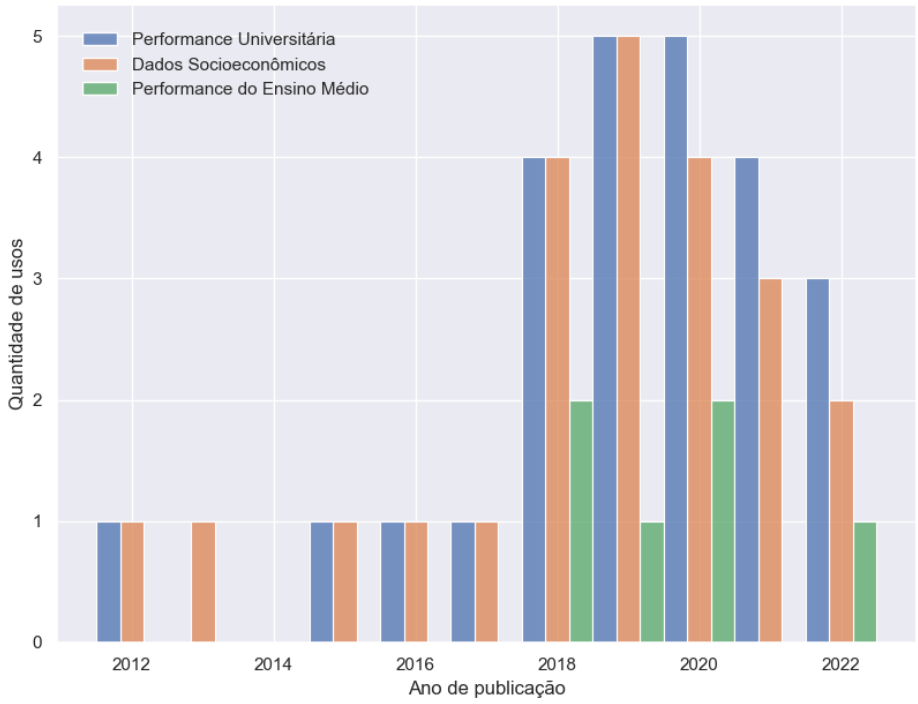
de solução do problema foram alteradas com o passar do tempo. Dessa forma, esta seção será voltada para discussões e conclusões obtidas a partir deste mapeamento sistemático.

Logo na Figura 4, pode-se observar que, em sua maioria, independentemente da família de dados utilizados, a grande maioria dos trabalhos selecionados produz um modelo como contribuição. Uma explicação para esse fenômeno está na intrínseca relação da área de pesquisa com um ramo da Inteligência artificial, o Aprendizado de Máquina. Como listado no Quadro 2, diversos algoritmos, que se encontram dentro do contexto da aprendizagem de máquina foram utilizados, desde os mais clássicos até abordagens mais recentes como a Computação Cognitiva e técnicas de *Ensemble*. Dessa forma, é natural a produção de trabalhos que produzam modelos que variam na parametrização e conjunto de dados utilizados.

Também, a Figura 4 mostra que não foram localizados, nos artigos selecionados, trabalhos do tipo "Validação"ou "Trabalho filosófico", indicando que a área de pesquisa possui um espectro voltado para estratégias de solução do problema e de avaliação de quais técnicas são mais adequadas para diferentes cenários do problema. Essas variações ocorrem principalmente devido às diferentes características dos dados utilizados pelos autores e também à variância introduzida nesses dados, pois cada trabalho tende a estudar diferentes instituições de ensino, dentro de sociedades distintas, tanto no âmbito social quanto político e econômico.

Outro aspecto a ser ressaltado é o aumento do interesse pelo tema de pesquisa na última década. A Figura 6 mostra que de 2016 à 2018 houve um aumento considerável no número de

Figura 7 – Histograma dos dados utilizados por ano



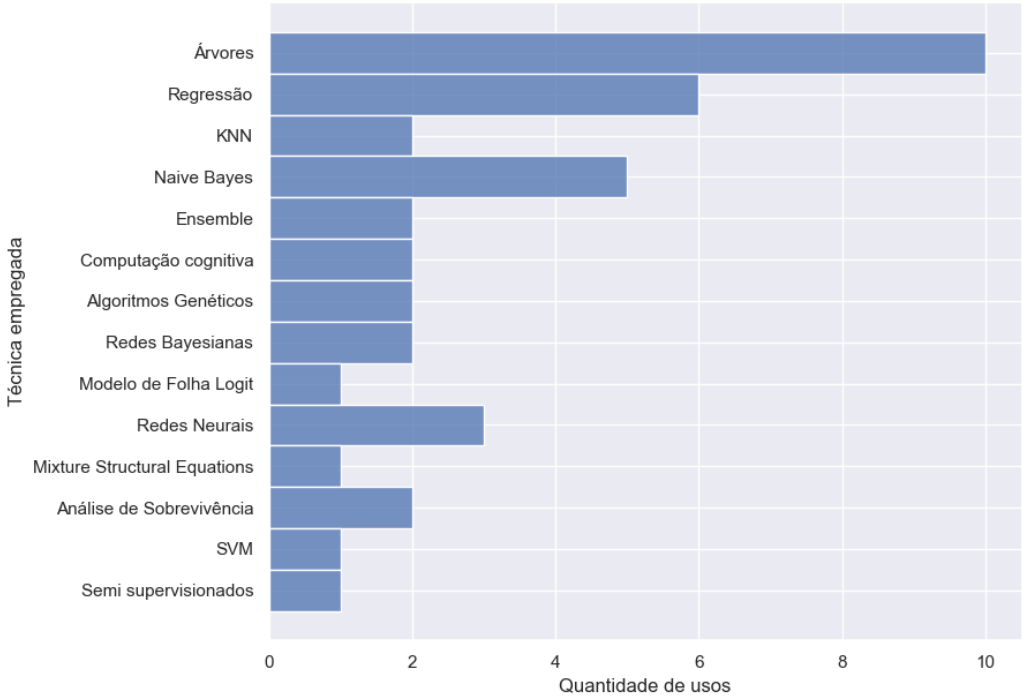
Fonte: Autor

artigos produzidos, quando se comparado aos anos anteriores. Essa tendência, que se preservou nos anos de 2019 e 2021, se explica, principalmente, pelo surgimento de novos desafios dentro da área de pesquisa. Esses desafios surgem principalmente no âmbito do ensino à distância, realizados, cada vez mais, de forma *online* e massiva, como explica Dalipi, Imran e Kastrati (2018).

Já na Figura 7 é possível se observar homogeneidade na forma com que os conjuntos de dados são utilizados ao passar do tempo. Isso mostra a preferência dos autores por utilizar dados socioeconômicos e de performance universitária como variáveis preditoras, e, na maioria das vezes, em conjunto, como mostra o Quadro 3. Entretanto, também existem autores que preferiram seguir uma abordagem menos ortodoxa, quando comparada ao demais. Esse é o caso de Zhou et al. (2018), que utilizou dados de logs de acesso à Internet para detectar alunos em risco de evasão.

Ainda, na Figura 8 e Figura 5 destaca-se o maior uso de técnicas baseadas em árvores, *Random Forests* e Árvores de decisão. Tal preferência é devida, em partes, ao conhecimento que a topologia desses tipos de modelo consegue entregar e que torna seu uso vantajoso. Como explica Zhang, Wang e Wang (2022), é possível usar a topologia de uma árvore de decisões não apenas para prever quais estudantes possuem tendência a evadir o ensino superior, como no caso de algoritmos baseados em regressão, mas também para saber *os porquês* dessa situação. Essa condição está intrinsecamente relacionada com este trabalho, uma vez que a topologia

Figura 8 – Histograma das técnicas empregadas



Fonte: Autor

das Redes Bayesianas Multinível, podem ser utilizadas com o mesmo propósito, como afirmam Lappenschaar et al. (2013a) e Russell e Norvig (2010).

Portanto, com base nos dados e sumários mostrados neste mapeamento sistemático, é possível afirmar que: 1) O problema do tema de pesquisa em questão vem sendo cada vez mais estudado e novas soluções vem sendo incorporadas para seus novos desafios, como é o caso dos cursos *online*. 2) O mapeamento sistemático demonstra que as Redes Bayesianas Multinível, ainda não foram objeto de estudo dentro deste tema, ressaltando a importância desse trabalho como ferramenta de avaliação para esse tipo de modelo.

3

Referencial Teórico

Este capítulo objetiva elucidar os conceitos necessários para a leitura do restante do trabalho. Primeiro serão apresentados alguns conceitos base da Probabilidade, então, a definição, propriedades e algoritmos utilizados nos modelos de Redes Bayesianas. Também será apresentado o formalismo das Redes Bayesianas Multinível, objeto de estudo deste trabalho, apresentado e empregado com sucesso por Lappenschaar et al. (2013a). Por fim, será definida a metodologia utilizada no processo de modelagem e tratamento dos dados deste trabalho, a CRISP-DM.

3.1 Inteligência Artificial

A inteligência humana é uma habilidade única que permite aos indivíduos perceber, interpretar e tomar decisões com base em informações sensoriais, cognitivas e emocionais, com o objetivo de interagir com o mundo à sua volta de maneira consciente e adaptativa (RUSSELL; NORVIG, 2010). Embora seja uma habilidade natural e inerente a cada indivíduo, o processo de pensamento ainda representa um desafio para a ciência até os dias atuais.

Nesse contexto, surge a Inteligência Artificial, IA. Sendo campo de estudo relativamente jovem (RUSSELL; NORVIG, 2010), a IA objetiva não apenas estudar e definir como funciona o pensamento humano, mas também criar modelos de inteligência, capazes de tomar decisões em ambientes complexos a fim de executar tarefas que auxiliem os seres humanos no seu desenvolvimento.

Dessa forma, a Inteligência Artificial é um campo de estudos amplo e diversificado, possuindo ramificações teóricas e práticas, que vão desde a compreensão do pensamento, passando por estudos éticos da a aplicação da IA, aprendizado de máquina, computação cognitiva, entre outros.

Uma dessas áreas é a do pensamento incerto, que estuda como tomar decisões quando o contexto externo é incerto ou nem todas as informações do problema estão disponíveis

no momento da decisão do indivíduo (PEARL, 1988). Nesse tipo de problema é necessário quantificar o grau convicção de que um determinado evento vá ou não ocorrer. Para isso, a teoria da probabilidade é largamente utilizada para estabelecer os meios que suportam o pensamento racional sobre incertezas.

3.2 Probabilidade

O conceito de probabilidade possui uma história rica, diversa e que inclui muitas abordagens filosóficas diferentes (NEAPOLITAN, 2003). A teoria da probabilidade está relacionada à experimentos, ou eventos, que podem possuir diversos resultados. O conjunto desses possíveis resultados é chamado de *espaço amostral* e é definida pela letra grega Ω. Cada e qualquer subconjunto $\{e_j, ..., e_k\} \subseteq \Omega$ é chamado de *evento*, de forma que:

$$\Omega = \{e_1, e_2, ..., e_i\}$$

Kolmogorov (1950) introduziu os três axiomas da probabilidade, que juntos definem uma função de probabilidade:

Definição 3.2.1 (Função probabilística)**.** *Seja Ω um espaço amostral de **n** elementos e **P** uma função que relaciona um número real, **P(E)**, à cada evento **E** $\subseteq \Omega$. P é dita uma função de probabilidade se:*

1. $0 \leq P(e_i) \leq 1, para 1 \leq i \leq n$
2. $P(\{e_1\}) + P(\{e_2\}) + ... + P(\{e_n\}) = 1$
3. *Para cada evento $E = \{e_{i1}, e_{i2}, ...e_{ik}\}$, tal que E não é um conjunto unitário:*

$$P(E) = P(e_{i1}) + P(e_{i2}) + ... + P(e_{ik})$$

*O par (Ω, P) é chamado de **espaço probabilístico**.*

Definição 3.2.2 (Variável Aleatória)**.** *Dado um espaço probabilístico (Ω, P), uma **variável aleatória** X é uma função em Ω, que atribui um valor único à cada evento e_i em Ω.*

*O conjunto de valores que X pode assumir é chamado de **espaço de X**.*

3.2.1 Probabilidade conjunta

Num espaço amostral Ω, a probabilidade de que os eventos E e F ocorram simultaneamente é chamada de **probabilidade conjunta**, denotada por $P(E \cap F)$, ou, de forma simplificada,

$P(E, F)$. Dessa forma, caso E e F sejam **mutualmente exclusivos**, a probabilidade conjunta desses eventos é:

$$P(E \cap F) = 0$$

Para eventos ou variáveis aleatórias discretas, o conjunto de todas as probabilidades conjuntas entre as variáveis é denominada de **Função de Massa de Probabilidade Conjunta**, FMPC. Uma representação gráfica dessa distribuição é a chamada **Tabela de distribuição conjunta completa**, a partir da qual é possível realizar inferências e consultas. O Quadro 4 exemplifica uma FMPC para duas variáveis, E e F, binárias no formato de tabela.

Quadro 4 – Exemplo de Tabela de distribuição conjunta completa para duas variáveis binárias

	e	$\neg e$
f	0,108	0,012
$\neg f$	0,016	0,064

A partir dessa distribuição é possível realizar procedimentos de inferências complexos, a fim de responder perguntas sobre o domínio dos dados. A distribuição conjunta é capaz de, através da **Marginalização**, calcular probabilidades de quando eventos acontecem sem levar em consideração os demais. Por exemplo, a fim de calcular o valor $P(e)$, é necessário realizar a soma de todos os eventos onde a condição é verdadeira:

$$P(e) = \sum_{z \in \{f, \neg f, \neg e\}} P(e, z)$$

$$P(e) = P(e, f) + P(e, \neg f) + P(e, \neg e)$$

$$P(e) = 0,108 + 0,016 + 0 = 0,124$$

Apesar de ser um poderoso mecanismo para tal, tabelas de distribuição conjuntas crescem exponencialmente conforme o número de variáveis aumenta. Assim, uma tabela de distribuição de probabilidade conjunta de n variáveis binárias terá um tamanho de 2^n. Essa condição impossibilita que algoritmos de inferência sejam aplicados diretamente à problemas com um grande número de variáveis. Isso pois, nem sempre é viável encontrar todas as probabilidades que a tabela necessita para estar completa, assim como o tempo computacional para execução do processo também cresce de maneira exponencial. As Redes Bayesianas, como será explorado na seção 3.3, surgem como uma estrutura capaz de contornar esse problema, possibilitando o processo de inferência em grandes domínios.

3.2.2 Probabilidade Condicional

A probabilidade condicional, ou à posteriori, é probabilidade de que um evento E ocorra dado que um evento F ocorre.

Definição 3.2.3 (Probabilidade Condicional)**.** *Sejam E e F eventos tal que $P(F) \neq 0$. Então, a probabilidade condicional de E dado F, denotada por $P(E|F)$, é:*

$$P(E|F) = \frac{P(E \cap F)}{P(F)} \tag{3.1}$$

Dessa definição, é direta a derivação da **Regra do produto**. Multiplicando ambos os termos da equação por $P(F)$, tem-se:

$$P(E \cap F) = P(E|F)P(F) \tag{3.2}$$

O denominador da Equação 3.1, pode ser visto como um fator constante no cálculo de probabilidades, dessa forma se tornando um fator de **normalização**. Por exemplo, no Quadro 4, no cálculo das probabilidades condicionais $P(E|F)$ e $P(\neg E|F)$, temos:

$$P(e|f) = \frac{P(e,f)}{P(f)} = \frac{0,108}{0,108+0,012} = 0.9$$

$$P(\neg e|f) = \frac{P(\neg e,f)}{P(f)} = \frac{0,012}{0,108+0,012} = 0.1$$

Percebe-se que o denominador $P(F)$ permanece constante. Dessa forma, podemos escrever uma única equação para o cálculo da probabilidade sobre um conjunto de variáveis: $\boldsymbol{P}(\boldsymbol{E}|f)$ em relação ao termo normalizador α:

$$\boldsymbol{P}(\boldsymbol{E}|f) = \alpha \boldsymbol{P}(\boldsymbol{E}, f) = \begin{bmatrix} \alpha P(e,f) \\ \alpha P(\neg e,f) \end{bmatrix} = \alpha \begin{bmatrix} 0.108 \\ 0.012 \end{bmatrix} = \begin{bmatrix} \frac{0.108}{0.108+0.012} \\ \frac{0.012}{0.108+0.012} \end{bmatrix} = \begin{bmatrix} 0.9 \\ 0.1 \end{bmatrix}$$

Por fim, outra importante regra é a chamada **Regra da Cadeia Geral**, largamente utilizada em redes bayesianas, uma vez que permite que distribuições conjuntas sejam representadas como probabilidades condicionais.

Definição 3.2.4 (Regra da Cadeia Geral)**.** *Seja $\boldsymbol{X}$ um conjunto de eventos ou variáveis aleatórias de n elementos, a distribuição de probabilidade conjunta de X, $P(\bigcap_{i=1}^{n} x_i)$, pode ser representada como (NEAPOLITAN, 2003):*

$$P(\bigcap_{i=1}^{n} x_i) = \prod_{k=1}^{n} P(x_k | \bigcap_{j=1}^{k-1} x_j) \tag{3.3}$$

Demonstração. Anexo A.1 □

3.2.3 Independência e Independência Condicional

A independência entre variáveis é uma característica importante, que ajuda a reduzir a complexidade durante os processos de inferência probabilística e que é vastamente utilizada nas análises Bayesianas.

Duas variáveis aleatórias, E e F, são ditas independentes quando uma não exerce influência sobre a outra dentro do espaço amostral.

Definição 3.2.5 (Independência)**.** *Dois eventos E e F são ditos independentes quando uma das seguintes condições ocorre (NEAPOLITAN, 2003):*

1. $P(E|F) = P(E)$ *ou* $P(E|F) = P(F)$ *e* $P(F) \neq 0, P(E) \neq 0$
2. $P(E) = 0$ *ou* $P(F) = 0$

A independência entre E e F implica diretamente que: $P(E \cap F) = P(E)P(F)$

Outra relação deste tipo é a chamada **independência condicional**. Essa relação estabelece que dois eventos E e F são independentes apenas sob à luz de um terceiro evento, G. Em outras palavras G explica totalmente os eventos E e F.

Definição 3.2.6 (Independência Condicional)**.** *Sejam os eventos E, F e G. E é dito* ***condicionalmente independente de*** *F* ***dado*** *G se uma das afirmações é verdadeira (NEAPOLITAN, 2003):*

1. $P(E|F \cap G) = P(E|G)$ *e* $P(E|G) \neq 0, P(F|G) \neq 0$
2. $P(E|G) = 0$ *ou* $P(F|G) = 0$

A notação mais comumente utilizada na literatura para representar essa propriedade é: $E \perp\!\!\!\perp F|G$. Caso E e F não sejam independentes dado G, então: $E \not\perp\!\!\!\perp F|G$.

Outra maneira útil de escrever esta relação é:

$$E \perp\!\!\!\perp F|G \Rightarrow P(E \cap F|G) = P(E|G)P(F|G) \tag{3.4}$$

Demonstração. Anexo A.2 □

3.2.4 Teorema de Bayes

Originalmente desenvolvido por Thomas Bayes em 1763, o **Teorema de Bayes** é, até os dias de hoje, largamente utilizado em sistemas computacionais inteligentes, e é definido da seguinte maneira:

Teorema 1 (Bayes)**.** *Sejam a e b possíveis valores que duas variáveis aleatórias A e B possam assumir, assim:*

$$P(b|a) = \frac{P(a|b)P(b)}{P(a)} \tag{3.5}$$

Demonstração. Anexo A.3 □

O teorema de Bayes permite calcular $P(b|a)$ quando se sabem os valores $P(a|b)$, $P(b)$ e $P(a)$. Essa forma de realizar o cômputo é útil durante o cálculo de probabilidades no mundo real. As probabilidades $P(b|a)$ e $P(a|b)$ possuem uma relação de **causa** e **diagnóstico**, respectivamente. A Equação 3.5 pode ser reescrita da seguinte maneira (RUSSELL; NORVIG, 2010):

$$P(causa|efeito) = \frac{P(efeito|causa)P(causa)}{P(efeito)} \tag{3.6}$$

Em problemas do mundo real, o valor de $P(causa|efeito)$ muitas vezes é difícil ou até impossível de se obter empiricamente (RUSSELL; NORVIG, 2010). Entretanto, as demais probabilidades podem ser mais facilmente calculadas, utilizando-se, por exemplo do histórico de uma população como base.

Lappenschaar et al. (2013a) evidencia esta relação causal no contexto da medicina, onde se objetiva identificar a doença que causa determinados sintomas. O cômputo de $P(doença|sintomas)$, em geral, é mais difícil que $P(sitomas|doença)$. Isso ocorre porque o doenças diferentes podem possuir sintomas muito semelhantes, ao mesmo tempo em que o conhecimento da medicina possuem bases sólidas pautadas em relações de causa, ou seja, os dados disponíveis hoje trazem muito mais informações sobre os sintomas que determinadas doenças causam.

3.3 Redes Bayesianas

Como mostrou a seção anterior, a inferência direta sobre distribuições conjuntas não é computacionalmente viável para problemas que envolvem um grande número de variáveis. As Redes Bayesianas tratam esse problema representando as tabelas de distribuição conjunta de maneira eficiente, na forma de um dígrafo acíclico, DAG, explorando a **Condição de Markov**.

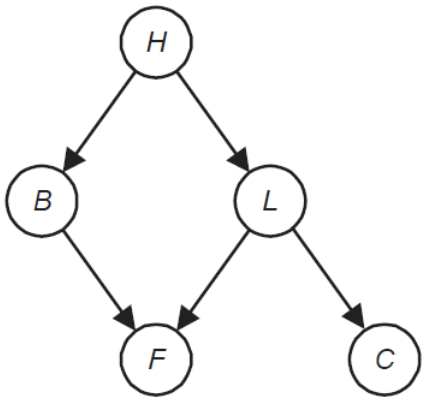
Definição 3.3.1 (Condição de Markov)**.** *Seja P uma distribuição conjunta sobre um conjunto de variáveis V, um DAG $G = (V, E)$. Diz-se que G satisfaz a* ***Condição de Markov*** *se para cada variável $X \in V$, X* ***é condicionalmente independente à todos os seus não descendentes dado o conjunto de seus pais****. Denotando os conjunto de pais e não descendentes de X por, respectivamente, PA_X e ND_X, então (NEAPOLITAN, 2003):*

$$X \perp\!\!\!\perp ND_X | PA_X \tag{3.7}$$

Se (G, P) satisfazem à condição de Markov, diz-se que G e P satisfazem a condição de Markov um com o outro.

A Figura 9 ilustra um DAG que satisfaz a condição de Markov. Assim, pode-se inferir as seguintes relação de independência entre as variáveis H, B, L, F e C, expostas no Quadro 5

Figura 9 – Exemplo de dígrafo fiel à condição de Markov



Fonte: Neapolitan (2003)

Quadro 5 – Relações de independência contidas no DAG da Figura 9

Vértice	Pais	Independência Condicional
H	$\{\}$	-
B	$\{H\}$	$B \perp\!\!\!\perp \{L, C\}\|\{H\}$
L	$\{H\}$	$L \perp\!\!\!\perp \{B\}\|\{H\}$
F	$\{B, L\}$	$F \perp\!\!\!\perp \{H, C\}\|\{B, L\}$
C	$\{L\}$	$C \perp\!\!\!\perp \{F, B, H\}\|L$

Teorema 2 (Distribuição conjunta num DAG fiel à condição de Markov)**.** *Seja $G = (V, E)$ um DAG com n vértices, P uma distribuição de probabilidade conjunta em V e (G, P) satisfaz a condição de Markov, então* ***P é igual ao produto das distribuições condicionais de todos os nós dados seus pais, sempre que essa probabilidade existir****. Assim (PEARL, 1988):*

$$P(\bigcap_{v \in V} v) = \prod_{v \in V} P(v|PA_v) \tag{3.8}$$

Demonstração. Anexo A.4

□

A condição de Markov e o Teorema 2 são essenciais para a redução da quantidade de probabilidades necessárias para representar uma distribuição conjunta num DAG. Como

demonstrado nas seções anteriores, uma tabela de distribuição conjunta para variáveis binárias tem um tamanho exponencial na quantidade de variáveis, n. Fazendo uso do Teorema 2, se cada vértice de G tiver no máximo 2 filhos, a complexidade da operação se torna linear. Em geral, fazendo uso do recursos apresentados acima, a quantidade de probabilidades necessárias para representar uma distribuição conjunta é de $2^k n$, onde k é a quantidade máxima de filhos de um vértice em G (NEAPOLITAN, 2003). Então, se k não é muito grande, o cômputo das probabilidades se torna um problema tratável.

Definição 3.3.2 (Rede Bayesiana)**.** *Seja P uma distribuição de probabilidade conjunta de um conjunto de variáveis aleatórias V e $G = (V, E)$ um DAG. (G, P) é uma **Rede Bayesiana** se (G, P) satisfaz a condição de Markov (PEARL, 1988). A especificação completa é como se segue:*

1. *Cada vértice pode representar uma variável aleatória discreta ou contínua*
2. *Cada aresta representa uma relação direta entre os dois vértices conectados*
3. *Cada vértice X_i na rede possui uma distribuição de probabilidade condicional que quantifica a influência dos pais sobre o nó, dada por: $\boldsymbol{P(X_i|PA_{X_i})}$*

Dessa forma, uma rede Bayesiana se trata de uma estrutura de dados capaz de representar uma distribuição de probabilidade conjunta. Intuitivamente, cada aresta numa Rede Bayesiana representa uma relação direta entre cada vértice dessa aresta.

Para um expert do domínio de um problema é relativamente fácil estipular relações entre as variáveis do problema, uma vez que cada aresta estabelece uma relação de causa e efeito sobre os nós envolvidos (RUSSELL; NORVIG, 2010), como ilustra a Figura 10. Nessa relação, o nó pai exerce uma influência causal direta sobre seu filho, que, por sua vez, se comporta como um efeito direto do pai.

Apesar das relações de Causa-Efeito não serem obrigatórias durante a definição de uma Rede Bayesiana, o emprego destas relações simplifica o processo, pois o levantamento dessas probabilidades é mais simples, ao mesmo tempo que menos probabilidades são necessárias para representar o conhecimento na Rede (RUSSELL; NORVIG, 2010).

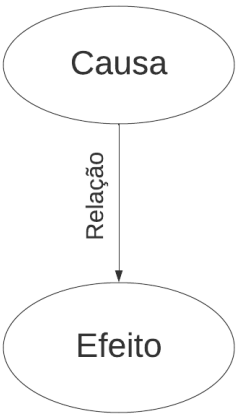
3.3.1 Relações de dependência e independência

As relações de independência entre variáveis em redes bayesianas são responsáveis, em grande parte, pela redução da complexidade do cômputo de probabilidades.

A condição de Markov define a principal relação de independência condicional em redes bayesianas, estabelecendo que todo nó X_i numa rede bayesiana é independente à qualquer não descendente dados seus pais:

$$X \perp\!\!\!\perp ND_X | PA_X$$

Figura 10 – Relação de causa e efeito em redes Bayesianas



Fonte: Autor

A partir desta desta relação é possível derivar outras duas importantes condições de independência. Uma delas está relacionada ao Envoltório de Markov de uma variável e a outra é a chamada D-separação, definidas nas próximas sub-seções.

3.3.1.1 D-Separação

Para definir a D-Separação, antes é necessário estabelecer o conceito de cadeias e caminhos bloqueantes.

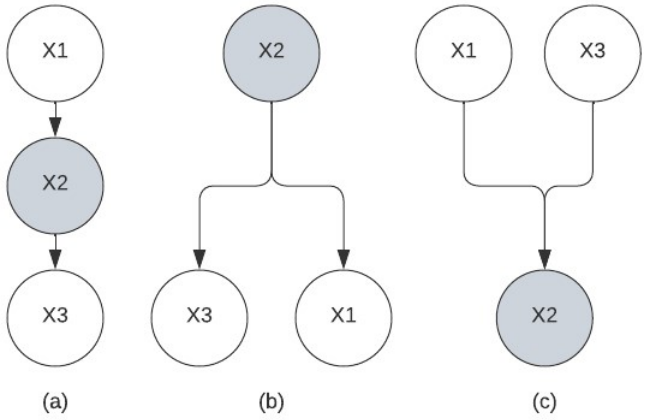
Definição 3.3.3 (Cadeia)**.** *Numa rede bayesiana $BN = (G, P)$ uma **cadeia** é definida como uma ligação, não necessariamente orientada, existente em G, que conecta dois vértices X_1, X_3, tal que existe um conjunto, B, não vazio, de vértices que formam o caminho entre X_1 e X_3 (NEAPOLITAN, 2003).*

Assim, uma cadeia numa rede bayesiana pode ter um *encontro* de três tipos, chamadas de *cadeias fundamentais*: ***cabeça-cauda***, ***cauda-cauda***, ***cabeça-cabeça*** (NEAPOLITAN, 2003). A Figura 11 ilustra os três tipos de cadeia, na qual o *encontro* ocorre no vértice $X2$, ou seja, $B = \{X2\}$.

Definição 3.3.4 (Cadeia bloqueada)**.** *Seja $BN = (G = (V, E), P)$ uma rede Bayesiana, B um conjunto de vértices, tal que $B \subset V$, $X1$ e $X3$ dois vértices distintos em V, tal que $X1$ e $X3 \notin B$ e, por fim, p uma cadeia entre $X1$ e $X3$, tal que B conecta $X1$ e $X3$. Então, diz-se que p é **bloqueado** por B se uma das seguintes condições é verdadeira:*

1. *Existe um vértice $X2 \in B$ em p tal que as arestas que incidem em $X2$ na cadeia p possuem um encontro **cabeça-cauda** ou **cauda-cauda** em $X2$, como nos casos (a) e (b) da Figura 11.*

Figura 11 – Tipos de cadeias fundamentais numa rede bayesiana



Tipos de encontro numa cadeia: (a) ***cabeça-cauda***, (b) ***cauda-cauda***, (c) ***cabeça-cabeça***

Fonte: Autor

2. *Existe um vértice $X2$, tal que $X2$ e todos os descendentes de $X2$* ***não*** *estão em B, na cadeia p, e as arestas incidentes para $X2$ em p possuem um encontro do tipo* ***cabeça-cabeça*** *em $X2$, como no caso (c) da Figura 11.*

Definição 3.3.5 (D-Separação)**.** *Seja BN uma rede bayesiana. Dois conjuntos de nós, X e Y, são* ***D-Separados*** *por um terceiro conjunto $Z \subseteq BN$ se todas as cadeias entre quaisquer par de vértices em X e Y são bloqueados por Z.*

A condição de D-Separação implica que os vértices contidos em X são condicionalmente independentes à qualquer nó em Y dado Z (NEAPOLITAN, 2003):

$$X \perp\!\!\!\perp Y|Z$$

3.3.1.2 Envoltório de Markov

O Envoltório de Markov é uma propriedade, cunhada por Pearl (1988), relacionada à variáveis aleatórias. O Envoltório de Markov de uma variável aleatória, X, é o subconjunto de variáveis M_x que explicam, completamente, o comportamento de X. Assim, como X é totalmente explicada por esse conjunto de variáveis, sendo desnecessária qualquer processo de inferência sobre variáveis que não pertencem à M_x. Assim:

Definição 3.3.6 (Envoltório de Markov)**.** *Seja P uma distribuição de probabilidade conjunta sobre um conjunto de variáveis aleatórias V e $X \in V$. O Envoltório Markov M_x de X é qualquer conjunto de variáveis tal que X é condicionalmente independente à todas as outras variáveis em V dado M_x.*

Teorema 3. *Seja $BN = (G = (V, E), P)$ uma Rede Bayesiana. Então para cada variável $X \in BN$, o conjunto de*

- *todos os pais de X, PA_X*
- *filhos de X*
- *pais dos filhos de X*

*compõem o **Envoltório de Markov** de X, MB_X (NEAPOLITAN, 2003).*

3.3.2 Classes de equivalência

Numa rede Bayesiana, diferentes conjuntos de arestas do dígrafo que compõe a estrutura da rede podem levar à mesma distribuição de probabilidades sobre as variáveis dessa rede. Isso decorre da condição local de Markov, que implica que os pais de um nó não são totalmente independentes do filhos, e, assim, uma alteração nos filhos pode gerar uma mudança na informação dos pais. Essa propriedade pode ser demonstrada com uma aplicação do teorema de Bayes:

Demonstração. Seja X_1 o único pai de um outro nó, X_2, numa rede Bayesiana, temos:

$$P(X_2|X_1) = \frac{P(X_1|X_2)P(X_2)}{P(X_1)}$$

$$\frac{P(X_2|X_1)P(X_1)}{P(X_2)} = P(X_1|X_2)$$

□

Além disso, também é possível mostrar que diferentes configurações nas arestas de uma cadeia fundamental são capazes de gerar as mesmas distribuições de probabilidade:

Demonstração. Sejam X_i, X_j e X_k nós pertencentes à mesma rede Bayesiana e que estejam conectados através de uma cadeia do tipo cabeça-cauda (Figura 11 (a)): $X_i \rightarrow X_j \rightarrow X_k$. Assim a distribuição conjunta fatorizada desta cadeia é dada por:

$$\begin{aligned} P(X_i, X_j, X_k) &= P(X_i)P(X_j|X_i)P(X_k|X_j) = P(X_j, X_i)P(X_k|X_j) \\ &= P(X_i|X_j)P(X_j)P(X_k|X_j) \end{aligned} \tag{3.9}$$

□

A equação 3.9 é a representação fatorizada da probabilidade conjunta da cadeia $X_i \leftarrow X_j \rightarrow X_k$ (Figura 11 (b)). Dessa forma, conexões do tipo cabeça-cauda e cauda-cauda podem ser representadas pela mesma distribuição de probabilidades, sendo assim equivalentes.

Entretanto, o mesmo não é verdade sobre estruturas do tipo cabeça-cabeça (Figura 11 (c)), conforme a definição de cadeia bloqueante e de d-separação. Esse tipo de cadeia é chamado de $estrutura - v$.

Definição 3.3.7 (Estruturas-v)**.** *Seja $B = (G, P)$ uma rede Bayesiana, $C \in G$ uma cadeia do tipo cabeça-cabeça, $X_i \rightarrow X_j \leftarrow X_k$. C é chamada de* ***estrutura-v*** *se, e somente se, as arestas $X_i \rightarrow X_k$ e $X_k \rightarrow X_i$ não existem em G. X_k é, então, chamado de colisor, pois bloqueia a relação de independência entre X_i e X_j.*

Assim, pode-se afirmar que as arestas que formam as estruturas-v de uma rede Bayesiana são as únicas responsáveis por definir a distribuição global de probabilidades da rede (SCUTARI, 2015). Assim, estas arestas não podem ter suas direções alteradas sem modificar a distribuição global da rede.

Definição 3.3.8 (Classe de equivalência)**.** *Dois DAGs definidos sobre o mesmo conjunto de variáveis são equivalentes se, e somente se, possuem o mesmo* ***esqueleto*** *e as mesmas* ***estruturas-v****.*

Definição 3.3.9 (Esqueleto de uma rede Bayesiana)**.** *O esqueleto de uma rede Bayesiana, $B = (G, P)$, é um grafo não dirigido, S, construído a partir de G e que representa a estrutura topológica base de G. S é construído da seguinte maneira:*

Para cada aresta $X \rightarrow Y \in G$, então $X - Y \in S$.

Portanto, para representar a distribuição global de probabilidades numa rede Bayesiana, B, é necessário apenas o conjunto de estruturas-v no DAG de B e o conjunto de arestas não direcionadas restantes. Isso leva a conclusão de que DAGs com estruturas topológicas diferentes são capazes de representar a mesma distribuição de probabilidades, sendo assim *equivalentes*. De fato, a quantidade de DAGs equivalentes cresce de maneira super-exponencial em relação ao número de nós na rede(MAUA, 2020).

A definição de classes de equivalência é empregada nos algoritmos durante a etapa de aprendizado da estrutura das redes Bayesianas. As equivalências tornam as etapas finais de aprendizagem mais simples e eficientes, uma vez que os algoritmos não se restringem a procurar por uma única solução, mas sim pela *classe de equivalência* dessa solução, reduzindo, assim, o espaço da busca e tornando a computação mais rápida.

3.3.2.1 Representando classes de equivalência

Para representar as classes de equivalência são utilizados DAGs parciais especializados, definidos abaixo, com propósitos específicos e que passam por processos de transformação durante cada etapa de aprendizado da estrutura da rede.

Definição 3.3.10 (Grafo Parcialmente Dirigido e acíclico (PDAG))**.** *O esqueleto, S, de uma rede Bayesiana, B, que também contém as estruturas-v de B é chamado de* ***Grafo Parcialmente Dirigido e acíclico (PDAG)****.*

Definição 3.3.11 (Arco Compelido)**.** *Um arco que não pertence à uma estrutura-v de uma classe de equivalência, mas que possui direção fixa e não reversível na estrutura topológica da rede. Aplicar uma mudança na direção desse arco implica diretamente na alteração da classe de equivalência dessa rede, pois cria uma nova estrutura-v ou um ciclo.*

Definição 3.3.12 (Grafo Parcialmente Dirigido, acíclico e completo (CPDAG))**.** *Um PDAG de uma rede Bayesiana, B, que contém o conjunto de arcos compelidos da classe de equivalência de B é chamado de* ***Grafo Parcialmente Dirigido, acíclico e completo (CPDAG)****.*

A Figura 12 ilustra os conceitos definidos acima. As estruturas-v, são as arestas com tons mais fortes. Cabe salientar que a cadeia $X1 \rightarrow X4 \leftarrow X2$ não é uma estrutura-v pois a aresta $X1 \rightarrow X2$ existe no DAG.

O verdadeiro DAG (topo esquerdo da imagem) possui a mesma distribuição de probabilidade global e o mesmo esqueleto (topo direito) do seu equivalente (abaixo, na direita).

Por fim, no CPDAG (abaixo, na esquerda) é possível observar a presença do único arco compelido do DAG, $X8 \rightarrow X6$. Essa aresta deve, obrigatoriamente, pertencer à classe de equivalência da rede, pois a sua reversão criaria uma nova estrutura-v, $X3 \rightarrow X8 \leftarrow X6$, que não existe no DAG original, alterando a distribuição global de probabilidade.

3.3.3 Benefícios do uso de Redes Bayesianas

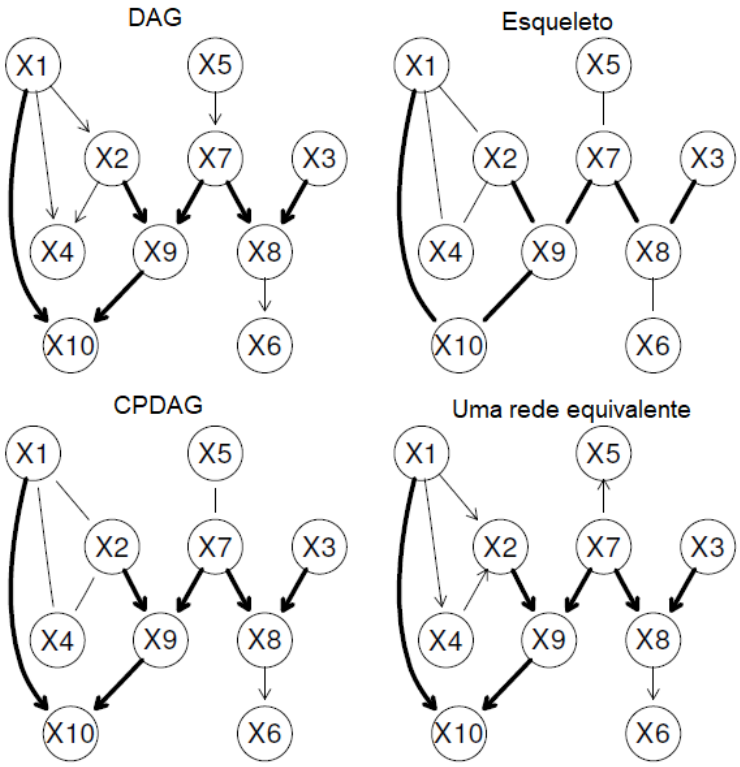
Numa Rede Bayesiana, cada variável é vista de forma incerta, diferentemente de outras técnicas, como as análises de Regressão (LAPPENSCHAAR et al., 2013a). Na análise de regressão, apenas a variável dependente é tratada como incerta, diferentemente das demais. Entretanto, se além do interesse na variável dependente, também se deseja saber quais são as relações entre as variáveis explicativas, a incerteza sobre as mesmas passa a ser uma forma mais viável de modelagem.

Dessa forma, a inferência em Redes Bayesianas pode ser considerada um processo qualitativo, já que, da mesma forma que em algoritmos baseados em árvores, é possível extrair conhecimento a partir da topologia da rede. Isso permite que sejam reveladas relações entre variáveis que podem passar despercebidas em outros métodos de observação.

3.4 Aprendizado em Redes Bayesianas

Assim como em outras áreas de estudo da Inteligência Artificial, o processo de estimar a estrutura e os parâmetros de uma rede Bayesiana é conhecido como aprendizado. Esse processo

Figura 12 – DAG, esqueleto, CPDAG e uma equivalência de uma rede Bayesiana $B = (G, P)$.



Fonte: Scutari (2015)

pode ocorrer de maneira supervisionada, não-supervisionado ou, ainda, de maneira combinada, utilizando ambas as técnicas (SCUTARI, 2015).

Segundo Russell e Norvig (2010), o aprendizado supervisionado é o processo de estimar parâmetros num modelo de aprendizado de máquina que necessita da intervenção humana para que o modelo possa aprender a partir de exemplos previamente classificados ou estimados por uma pessoa. O aprendizado supervisionado numa rede Bayesiana pode ser realizado empregando conhecimento prévio sobre o problema, obtido de especialistas ou do próprio domínio do problema. Esse conhecimento pode ser utilizado, por exemplo, para estabelecer uma relação inicial entre os nós da durante o aprendizado da estrutura.

O aprendizado não supervisionado, por sua vez, não necessita da intervenção humana, uma vez que o algoritmos utilizados possuem capacidade de extrair padrões utilizado o próprio conjunto de dados para isso. Por exemplo, algoritmos de agrupamento, como o k-Means, são capazes de agrupar dados estabelecendo pontos iniciais, chamados de *centroids*, calculando a distância de cada observação presente nos dados para os *centroids*, dessa forma, criando grupos de observações similares. Aprendizado não supervisionado pode ser utilizado para aprender distribuições de probabilidades em uma rede Bayesiana, sem necessidade de intervenção humana.

Por fim, a combinação das duas técnicas é comumente empregada durante o aprendizado

de redes Bayesiana (MAUA, 2020). Isso acontece pois nem sempre existem dados suficientes para descrever o fenômeno observado de forma completa. Para amenizar esse fator, conhecimento sobre o domínio do problema pode ser utilizado para estabelecer relações que não estejam contidas nesses dados. Por outro lado, há casos em que é impossível para um ser humano estabelecer todos os parâmetros do domínio do problema, mesmo quando a estrutura da rede é fornecida (SCUTARI, 2015). Por esses motivos o uso das duas técnicas, em geral, traz vantagens ao processo de aprendizado como um todo.

O aprendizado de redes Bayesiana geralmente é realizado em duas etapas:

1. Aprendizado estrutural: onde a estrutura topológica do DAG que representa a rede é aprendida;
2. Aprendizado de parâmetros: onde as distribuições probabilísticas locais de cada nó é estimada;

Essas etapas são inerentemente Bayesianas e todo o processo pode ser formalizado nos moldes da Equação 3.10 (SCUTARI, 2015). Considerando D o conjunto de dados sob análise e $B = (G, X)$ uma rede Bayesiana. Θ é o conjunto de parâmetros da distribuição X, e assim pode-se afirmar que Θ identifica X unicamente dentro da família dessa distribuição em D, então:

$$P(B|D) = P(G, \Theta|D) = P(G|D)P(\Theta|G, D) \tag{3.10}$$

A decomposição de $P(G, \Theta|D)$ em $P(G|D)$ e $P(\Theta|G, D)$ reflete as duas etapas listadas acima. O primeiro fator se refere ao aprendizado da estrutura e o último ao aprendizado das distribuição de probabilidades locais da rede.

Em geral, computar $P(\Theta|G, D)$ é uma tarefa mais simples que computar $P(G|D)$ (MAUA, 2020). Isso se deve ao fato de que, durante a aprendizagem dos parâmetros, G é dado como um parâmetro em conjunto com o $dataset$, D, conforme a representado na equação. Dessa maneira é possível estimar as distribuições de probabilidade de maneira local para cada nó v da rede: $P(v|PA_v)$. Além disso, como número de pais de qualquer vértice na rede tende a ser constante em relação à quantidade de variáveis (RUSSELL; NORVIG, 2010), o cômputo dessas distribuições também é eficiente.

Por outro lado, aprender a estrutura da rede a partir dos dados, $P(G|D)$, é um processo mais complexo, pois:

$$P(G|D) \overset{Bayes}{=} \frac{P(D|G)P(G)}{P(D)} \propto P(D|G)P(G) \tag{3.11}$$

A probabilidade a priori da estrutura da rede, $P(G)$, em geral é computada de maneira não informativa, atribuindo uma probabilidade uniforme para cada possível DAG. Essa estratégia é mais comumente adotada, pois como o espaço de possíveis DAG cresce de forma super-exponencial, inferir uma distribuição especializada seria um problema intratável mesmo para um

pequeno número de variáveis. Além disso, uma probabilidade uniformemente variada torna mais simples inserir conhecimento sobre o domínio do problema, sendo possível descartar os DAGs que não possuem um determinado conjunto de arestas, por exemplo.

Já a probabilidade $P(D|G)$ é impossível de ser calculada sem estimar também os parâmetros da rede: $P(D|G,\Theta)$ (SCUTARI, 2015). Entretanto, como visto, o aprendizado dos parâmetros da rede depende diretamente da estrutura do DAG, G. Para contornar esse problema existem duas estratégias que aproximam o valor de $P(D|G)$. A primeira estratégia possui raízes na Teoria da Informação e busca atribuir pontuações que medem quão bem o G representa os dados contidos em D. A segunda estratégia se baseia em testes de independência condicional sobre D, identificando as arestas entre variáveis na estrutura da rede. Ambas as estratégias possuem vantagens e desvantagens e o estado da arte desse ramo de pesquisa é o emprego de algoritmos híbridos, que combinam as duas estratégias afim de de estimar $P(G|D)$ (Equação 3.11) com mais precisão.

3.4.1 Aprendizado estrutural

Como citado no capítulo anterior, aprender a real estrutura de uma rede Bayesiana nem sempre é possível. Primeiro porque os dados podem não representar com fidelidade a real distribuição das variáveis da rede. E segundo porque as técnicas que existem para aproximar a estrutura da rede nem sempre são capazes captar todas as nuances da distribuição dos dados, mesmo com uma quantidade infinita de dados.

Em contraste, o aprendizado em redes Bayesianas permite uma fácil incorporação de conhecimento a priori sobre o domínio do problema (MAUA, 2020). Esse fator é aplicado como um mecanismo auxiliar em algoritmos de aprendizado, a fim de reduzir o espaço de busca, melhorar a aproximação e corrigir relações que possam ter sido inferidas de maneira incorreta por estes algoritmos.

Para realizar o aprendizado, existem, basicamente, dois tipos de algoritmos que utilizam diferentes métodos de inferência. Os Métodos Baseados em restrições se baseiam em testes de independência condicional entre as variáveis do problema para inferir as separações gráficas na estrutura da rede. Já os Métodos baseados em pontuação tiram proveito de técnicas já consolidadas na Inteligência Artificial, como *Hill-Clibing* e *Tabu-Search* (RUSSELL; NORVIG, 2010), para realizar buscas dentro do espaço de possíveis estruturas da rede e atribuindo uma pontuação a cada DAG encontrado, a fim de selecionar a estrutura que melhor representa os dados. Cada uma dessas técnicas será apresentada nas próximas seções, tendo como base os trabalhos de Maua (2020), Scutari (2015) e Russell e Norvig (2010).

3.4.1.1 Métodos baseados em restrições

Algoritmos baseados em restrições para aprendizado estrutural de redes Bayesianas possuem origens no algoritmo de Indução Causal, IC, desenvolvido por Verma e Pearl (1990). O IC, possui três passos para a identificação do CPDAG da rede Bayesiana e são descritos no Algoritmo 1.

Algoritmo 1: Indução Causal - IC

Entrada:

D - *dataset* de observações de n eventos sobre um conjunto de variáveis V de tamanho k;

$G = (V, E)$ - Um grafo, geralmente completo, contendo V como vértices;

Saída: Ao fim da execução, G é um CPDAG que comporta a classe de equivalência de um DAG que representa D.

1 Para cada aresta $A - B \in G$, procure por um conjunto $S_{AB} \subset V$ tal que $A \perp\!\!\!\perp B | S_{AB}$ e $A, B \notin S_{AB}$. Se S_{AB} existe, remova $A - B$ de G e guarde S_{AB} como um conjunto testemunha de $A - B$.

2 Para cada par de nós **não-adjacentes** $A, B \in G$, com um vizinho comum C, verifique se $C \in S_{AB}$. Se isso não for verdade, configure a cadeia $A - C - B$ como uma estrutura-v em que colide em C: $A \rightarrow C \leftarrow B$.

3 Direcione, de forma recursiva, as arestas $A - B \in G$ em $A \rightarrow B$ sempre que ocorrer uma das seguintes situações, até que não hajam mais arestas não direcionadas em G:

1. Existe um arco $C \rightarrow A$, tal que C e B não são adjacentes.
2. Existe uma cadeia $A \rightarrow C \rightarrow B$.
3. Existem duas cadeias: $A - C \rightarrow B$ e $A - D \rightarrow B$, tal que C e D não são adjacentes.
4. Existem duas cadeias: $A - C \rightarrow D$ e $C \rightarrow D \rightarrow B$, tal que C e B não são adjacentes.

O algoritmo parte de um grafo completo, G que possui as variáveis de D como vértices. No primeiro passo são realizado testes de independência sobre cada uma das arestas a fim de localizar os conjuntos de vértices que d-separam cada uma delas. Se dois vértices são independentes dado qualquer conjunto de outros vértices, inclusive o vazio, não deve haver uma aresta entre eles. Ao fim do primeiro passo G contém o esqueleto da sua classe de equivalência.

O segundo passo, é responsável pela identificação das estruturas-v, procurando por cadeias de vértices $A - C - B$ e verificando se $C \in S_{AB}$. Se $C \in S_{AB}$, então C bloqueia a relação de independência entre A e B e, segundo a definição de d-separação, C precisa ser um colisor, e essa cadeia é uma estrutura-v. Após o fim dessa etapa, G é o PDAG desta classe de equivalência, pois possui todas as estruturas-v da classe de equivalência.

O terceiro e último passo direciona todas as arestas não direcionadas restantes e que precisam ser compelidas para formar a classe de equivalência. O processo real de identificação

dessas arestas pode variar (VERMA; PEARL, 1990), mas nunca deve criar novas estruturas-v e nem adicionar ciclos ao grafo. Verma e Pearl (1990), demonstraram que repetidas aplicações das quatro regras dispostas no Algoritmo 1 são suficientes para identificar e compelir todos os arcos necessários, uma vez que todas as estruturas-v já foram identificadas no passo anterior do algoritmo. Ao fim dessa etapa, portanto, o CPDAG que representa D é retornado.

Para ser considerado correto, o IC supõe algumas condições para que, ao final do procedimento, seja retornado o verdadeiro CPDAG da classe de equivalência:

- Existe um Oráculo capaz de responder com exatidão perguntas sobre independências e independências condicionais entre variáveis.
- Os dados em D foram gerados por uma rede Bayesiana, B, hipotética, que pertence a mesma classe de equivalência do CPDAG retornado.
- B mapeia perfeitamente todas as relações de independência condicional da distribuição real da sua classe de equivalência.

Scutari (2015) argumenta que nenhuma dessas suposições pode ser verificada e alcançada na prática, mesmo com uma quantidade infinita de dados. Mesmo que bastante precisos, os testes de independência condicional hoje existentes não são capazes de realizar inferências com exatidão, o que descarta a existência de um oráculo. A segunda suposição não pode ser verificada, pois, em geral, a classe de equivalência é desconhecida. Por fim, nem toda distribuição $P(D)$ pode ser representada em sua totalidade como um DAG, mesmo supondo uma quantidade infinita de observações.

Lauritzen e Spiegelhalter (1988) mostraram que existem estruturas que não podem ser aprendidas pelos algoritmos baseados em restrições. Isso acontece pois os falsos positivos gerados pelos testes de independência influem nos próximos testes, gerando um efeito em cascata, um fenômeno conhecido como testagem de hipóteses múltiplas (MAUA, 2020). Eles ainda demonstraram que diferentes tipos de testes performam melhor em determinados conjuntos de dados.

Dessa forma, a eficácia do algoritmo depende diretamente dos testes de independência realizados durante o processo. Essa condição pode ser atenuada se os testes de independência forem considerados também como um parâmetro a ser escolhido afim de reduzir a quantidade de erros gerados pelos testes. Isso pode ser feito utilizando, técnicas de Validação Cruzada (BOUCKAERT, 2003) sobre os dados, com diferentes tipos testes de independência, durante o aprendizado estrutural e selecionando o modelo que melhor absorver as nuances do conjunto de dados.

3.4.1.1.1 Algoritmos de indução causal

O IC se trata de um algoritmo teórico, que não possui uso prático, devido ao tempo de execução exponencial (MAUA, 2020). Entretanto, é de extrema importância, pois estabelece as bases para qualquer outro algoritmo que utilize testes de independência para mapear os arcos no DAG de uma rede Bayesiana. Dessa forma, com o avanço da área de pesquisa, novos algoritmos para inferência estrutural surgiram, preservando os três passos fundamentais do IC, aumentando sua eficiência com a incorporação de novas heurísticas e propriedas ao processo.

Devido à quantidade exponencial de testes de independência realizados, a identificação do esqueleto (passo 1), é a mais custosa em tempo de execução, e, portanto concentra a maioria das variações e melhorias de performance entre os algoritmos derivados do IC.

A primeira aplicação prática do IC foi o algoritmo PC [1] (SPIRTES; MEEK, 1995), capaz de identificar o esqueleto da classe de equivalência fazendo um número polinomial de testes de independência, ao mesmo tempo que envolve, no máximo, um número constante de variáveis nos testes (MAUA, 2020). Isso é possível devido à incorporação do corolário 3.4.1 no primeiro passo do IC.

Corolário 3.4.1. *Seja $G = (V, E)$ o DAG de uma rede Bayesiana, se dois nós X e Y não são adjacentes em G, então uma das seguintes afirmações é verdadeira:*

- *PA_X d-separa X e Y*
- *PA_Y d-separa X e Y*

Demonstração. A prova segue diretamente da Condição Local de Markov de que todo nó é independente aos seus não-descendentes dados o conjunto dos seus pais: $A \perp\!\!\!\perp ND_A | PA_A$.

Assim, como X e Y não são adjacentes e não existem ciclos num DAG, obrigatoriamente $X \in ND_Y$ ou $Y \in ND_X$. Portanto, $X \perp\!\!\!\perp Y | PA_X$ ou $Y \perp\!\!\!\perp X | PA_Y$. □

Assim, no algoritmo PC, os candidatos à conjunto testemunha da d-separação entre dois vértices A e B estão contidos em $PA_A \cup PA_B$. Ainda, o número de pais de um vértice tende a ser constante em relação ao número de variáveis na rede (RUSSELL; NORVIG, 2010) (MAUA, 2020), e é denotado por d. Assim, no máximo, $2d$ vértices são candidatos a fazerem parte do conjunto testemunha de A e B. Como, inicialmente, o grafo é completo e não direcionado, esse número se expande para todos os vizinhos de A e B, $NB_A \cup NB_B$, então é possível afirmar que, no máximo, são observadas 2^{2d} variáveis para determinar a separação entre A e B.

Dessa maneira o PC, Algoritmo 2, consegue, até os dias atuais, ser competitivo em relação aos demais algoritmos baseados em restrições.

[1] O nome PC é derivado das iniciais dos nome dos autores, **P**eter Spirts e **C**lark Meek.

Algoritmo 2: PC

Entrada:

D - *dataset* de observações de n eventos sobre um conjunto de variáveis V de tamanho k;
$G = (V, E)$ - Um grafo, geralmente completo, contendo V como vértices;
$0 > d <= |V|$ - O número máximo de pais que um vértice poderá possuir no CPDAG;

Saída: Ao fim da execução, G é um CPDAG que comporta a classe de equivalência de um DAG que representa D.

1 Para cada aresta $A - B \in G$, procure por um conjunto $S_{AB} \subset (NB_A \cup NB_B)$, de tamanho máximo d, tal que $A \perp\!\!\!\perp B | S_{AB}$ e $A, B \notin S_{AB}$. Se S_{AB} existe, remova $A - B$ de G e guarde S_{AB} como um conjunto testemunha de $A - B$.

2 Para cada par de nós **não-adjacentes** $A, B \in G$, com um vizinho comum C, verifique se $C \in S_{AB}$. Se isso não for verdade, configure a cadeia $A - C - B$ como uma estrutura-v em que colide em C: $A \rightarrow C \leftarrow B$.

3 Direcione, de forma recursiva, as arestas $A - B \in G$ em $A \rightarrow B$ sempre que ocorrer uma das seguintes situações, até que não hajam mais arestas não direcionadas em G:

1. Existe um arco $C \rightarrow A$, tal que C e B não são adjacentes.
2. Existe uma cadeia $A \rightarrow C \rightarrow B$.
3. Existem duas cadeias: $A - C \rightarrow B$ e $A - D \rightarrow B$, tal que C e D não são adjacentes.
4. Existem duas cadeias: $A - C \rightarrow D$ e $C \rightarrow D \rightarrow B$, tal que C e B não são adjacentes.

Além do PC, outros algoritmos de indução causal também foram desenvolvidos. Entretanto, diferentemente do PC, esses algoritmos possuem uma etapa preliminar que visa identificar o Envoltório de Markov de cada variável a fim de reduzir o número de testes condicionais aplicados. Scutari (2015) fornece um resumo do funcionamento desses algoritmos:

- *Grow-Shrink* (GS): baseado no algoritmo Grow-Shrink para detecção do Envoltório de Markov (Algoritmo 3). É o algoritmo mais simples desse tipo. Funciona aplicando uma estratégia que primeiro encontra todos os candidatos a pertencer ao Envoltório de Markov *(Grow)* e depois realiza um filtragem nesse conjunto *(Shrink)*, procurando por independências dentro do conjunto encontrado até então.
- Incremental Association (IAMB): utiliza o algoritmo IAMB para detectar o Envoltório de Markov.
- Fast Incremental Association: uma variante do IAMB que utiliza um conjunto de heurísticas para reduzir a quantidade de testes de independência condicionais.
- Interleaved Incremental Association: outra variante do IAMB que utiliza *forward stepwise*

selection para reduzir o número de falsos positivos durante a descoberta do Envoltório de Markov.

Algoritmo 3: Grow Shrink para deteção do Envoltório de Markov

Entrada:
V - Conjunto de variáveis, tal que $MB(v) \in V$
v - Uma variável que pertence à V
Saída: O Envoltório de Markov de v

```
1 MB_v = {}
2 Grow: enquanto ∃w ∈ {V − v} tal que w ⫫̸ v|MB_v faça
3 |   MB_v = MB_v ∪ {w}
4 fim
5 Shrink: enquanto ∃w ∈ MB_v tal que w ⫫ v|MB_v − {w} faça
6 |   MB_v = MB_v − {w}
7 fim
8 Retorne MB_v
```

3.4.1.1.2 Testes de independência

Os algoritmos descritos na seção anterior dependem de testes capazes de inferir a independência condicional entre duas variáveis, X e Y dado um conjunto de variáveis Z, que pertencem à um conjunto de dados D. Tais testes podem ser formalizados como testes estatísticos com as seguintes hipóteses:

$$H_0 : X \perp\!\!\!\perp Y|Z$$

$$H_1 : X \not\perp\!\!\!\perp Y|Z$$

Como a real distribuição das variáveis em D não é conhecida, os testes ocorrem de maneira aproximada e, portanto, existe uma probabilidade de erro associada a cada teste. Assim, dois tipos de erros podem ocorrer:

- Erro tipo I (falso positivo): O teste rejeita H_0, mas H_0 é verdadeira e deve ser aceita.
- Erro tipo II (falso negativo): O teste aceita H_0, mas H_0 é falsa e deve ser rejeitada.

A probabilidade de que um falso positivo ocorra, α, é chamado de *nível de significância* do teste. A escolha de um α pequeno tende a reduzir a ocorrência de falsos positivos. Isso, por consequência, leva os algoritmos baseados em testes de independência a inferir estruturas mais esparsas, uma vez que menos relações de independência são rejeitadas.

No caso discreto, os testes de independência condicional são funções das frequências observadas em D, para as variáveis X, Y e todas as configurações possíveis de Z. Assim, os testes χ^2 de Pearson e G^2 são comumente utilizados para validar H_0 (SCUTARI, 2015).

Definição 3.4.1 (Teste χ^2 de Pearson)**.**

$$\chi^2(X,Y|Z) = \sum_{x\in X}\sum_{y\in Y}\sum_{z\in\{A\times B,\forall_{(A,B)}\in Z\}} \frac{(n_{xyz}-m_{xyz})^2}{m_{xyz}}, \text{ onde: } m_{xyz} = \frac{(n_{x+z}n_{+yz})}{n_{++z}} \tag{3.12}$$

Definição 3.4.2 (Teste G^2)**.**

$$G^2(X,Y|Z) = \sum_{x\in X}\sum_{y\in Y}\sum_{z\in\{A\times B,\forall_{(A,B)}\in Z\}} \frac{n_{xyz}}{n} log\frac{n_{xyz}n_{++z}}{n_{t+k}n_{+ek}} \tag{3.13}$$

Em ambas as definições, a notação $c \in V$ simboliza as categorias de uma variável V, enquanto que n_{xyz} denota a frequência que uma configuração de categorias x, y, z ocorrem em D. Por fim, o índice "+"denota a marginalização de uma categoria, indicando a soma sobre as demais.

A Equação 3.12 se trata de uma adaptação para independências condicionais do teste χ^2 para tabelas de contingência (MAUA, 2020). Enquanto que a Equação 3.13 é uma variante do teste *log-likelihood ratio*, sendo equivalente ao teste de Informação Mútua (SCUTARI, 2015).

Ambos os testes possuem uma distribuição χ^2_d sob H_0, onde a quantidade de graus de liberdade, d, da distribuição é dada por:

$$d = (|X|-1)(|Y|-1)(\prod_{z\ in\ Z}|z|) \tag{3.14}$$

Dessa forma, o resultado dos testes depende da aceitação ou rejeição de H_0.

Definição 3.4.3 (Condição de aceitação de H_0)**.** *Aceita-se H_0 quando o resultado do teste estatístico t, χ^2 ou G^2, possui p-valor estatisticamente significante, ou seja:*

$$p\text{-}valor(t) < \alpha \tag{3.15}$$

Durante a execução de um algoritmo baseado em restrições, como o IC, a aceitação de H_0 significa que existem evidências estatisticamente significantes em D de que $X \perp\!\!\!\perp Y|Z$. Dessa maneira, o algoritmo passa a considerar hipótese de independência real e fiel à sua classe de equivalência, tomando as ações necessárias, como eliminar a aresta $X - Y$.

Por outro lado, se H_0 for rejeitada, não existem evidências estatisticamente significantes em D que apontem a independência entre X e Y. Nesse caso, o IC passa a considerar a relação de independência como *falsa* e não realiza quaisquer alterações no grafo.

3.4.1.2 Métodos baseados em pontuação

Algoritmos baseados em pontuação codificam o processo de aprendizado da estrutura de uma rede Bayesiana como um problema de otimização do conjunto de parâmetros que define a distribuição de probabilidade condicional da rede (MAUA, 2020). A busca, em geral, ocorre de

maneira gananciosa no espaço dos possíveis DAG que representam a distribuição. A escolha gananciosa, por sua vez, é realizada com base em uma ação que altere o DAG, de forma que o novo DAG passe a representar *melhor* a distribuição dos dados.

Esse processo resume o funcionamento dos chamados Algoritmos de busca local (RUSSELL; NORVIG, 2010). Os algoritmos desse tipo, no âmbito do aprendizado estrutural de redes Bayesianas, maximizam uma função $S(G, D)$, também chamada de **função objetivo**, que mede quão bem a estrutura de um DAG, G, representa a distribuição em um conjunto de dados D. Formalmente,

Definição 3.4.4 (Função objetivo)**.** *Seja G o DAG de uma rede Bayesiana e D a distribuição de probabilidades que representa as variáveis da rede. Então, $S(G, D)$ é uma função mapeia a qualidade do ajuste de G em D como um número real (MAUA, 2020).*

$$S : G, D \rightarrow \mathbb{R} \tag{3.16}$$

Além disso, os algoritmos baseados em pontuação assumem quatro propriedades sobre as funções objetivo:

Definição 3.4.5 (Minimalidade)**.** *S deve retornar pontuações melhores para configurações mais sucintas da rede. Assim, se G e G' possuem qualidade de ajuste, mas G' possui menos parâmetros então:*

$$S(G', D) < S(G, D)$$

Definição 3.4.6 (Consistência)**.** *Seja $B = (G, X)$ uma rede Bayesiana, D um conjunto de dados que possui N possíveis configurações de estruturas para G, GN o conjunto dessas estruturas e SN um conjunto de N elementos onde cada $SN_i \in SN$ é o valor da função objetivo de $GN_i \in GN$, então:*

$$\lim_{|D| \rightarrow \infty} P(|S(G, D) - max(SN)| > 0) = 0$$

Definição 3.4.7 (Eficiência)**.** *Uma função objetivo é dita eficiente se pode ser computada em tempo polinomial em relação ao tamanho de G.*

Definição 3.4.8 (Fatoração)**.** *Uma função objetivo é dita fatorizável se pode ser decomposta da seguinte forma:*

$$S(G, D) = \sum_{X \in V} f(X, PA_X, D)$$

onde V é o conjunto de vértices da rede.

As duas primeiras definições trazem uma visão qualitativa sobre o resultado do mapeamento, enquanto as últimas se preocupam em estabelecer parâmetros para o cômputo eficiente destas funções.

A definição de Minimalidade implica que as funções objetivo devem priorizar as soluções mais simples, sempre que possível. A segunda definição, Consistência, descreve que, quando existe um montante infinito de dados, as funções objetivo devem ser capazes de identificar o DAG real que descreve a distribuição sobre os dados.

Por sua vez, a definição de eficiência impõe que, idealmente, as funções objetivo precisam possuir um tempo de cômputo polinomial. E por fim, a Fatoração implica que as funções devem poder ser reescritas dentro do formalismo das redes Bayesianas, a fim de facilitar e agilizar seu cômputo, fazendo-o de forma individual para cada vértice da rede.

3.4.1.2.1 Algoritmos baseados em pontuação

Os algoritmos baseados em pontuação necessitam de uma função objetivo mínima e consistente, para serem considerados corretos. Além disso, sua aplicabilidade em problemas reais depende da eficiência e da possibilidade de decomposição da função objetivo em fatores da rede. A partir dessas suposições é possível empregar uma ampla gama de algoritmos de otimização para realizar a tarefa de aprendizado.

O *Hill Climbing*, Algoritmo 4, é um dos algoritmos de busca local mais simples , mas que ilustra o processo de otimização, de forma transparente.

Algoritmo 4: Hill Climbing

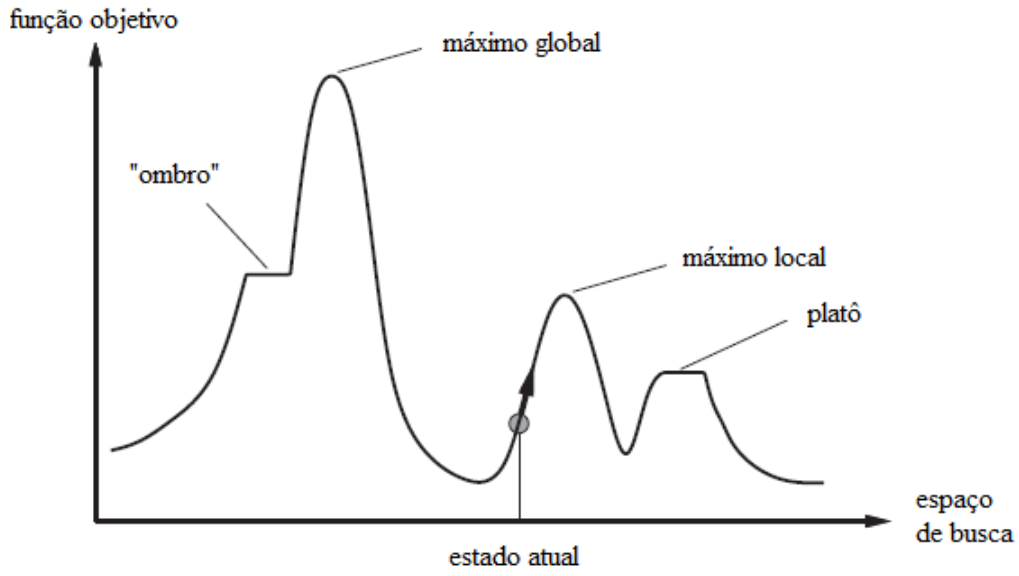
Entrada:
D - Conjunto de dados
G - um DAG, em geral vazio, mas não necessariamente S - Uma função objetivo a ser minimizada

Saída: Um DAG, G' que pertence à mesma classe de equivalência que representa a distribuição em D

```
1  G_r = G
2  S_{G_r} = S(G_r, D)
3  pontuacaoMaxima = −∞
4  enquanto S_{G_r} > pontuacaoMaxima faça
5      pontuacaoMaxima = S_{G_r}
6      para cada adição, remoção ou reversão de aresta que não crie um ciclo em G_r
         faça
7          Seja G* o grafo modificado;
8          S_{G*} = S(G*, D)
9          se S_{G*} > S_G então
10             G_r = G*
11             S_{G_r} = S_{G*}
12         fim
13     fim
14 fim
15 Retorne G_r
```

Para que esteja correto, o Algoritmo 4 assume que a função, S, possui apenas um ponto máximo. Essa suposição, entretanto, é utópica em problemas reais, como explica Russell e Norvig (2010). Problemas do mundo real possuem funções complexas, com diversos máximos e mínimos locais, além de de regiões onde não há variações, chamados de platôs e ombros, conforme a Figura 13. Dessa forma, algoritmos como o Hill Climbing possuem dificuldade em encontrar o máximo global das funções objetivo, pois a sua execução termina assim que o valor da função passa a decair, logo após um máximo local.

Figura 13 – Variações que uma função objetivo pode possuir.



Fonte: Russell e Norvig (2010)

Uma abordagem para resolver esse problema é incorporar processos estocásticos ao algoritmo. Isso permite que o algoritmo saia de um máximo local através da aplicação de perturbações ao estado atual do DAG. O Algoritmo 5, *Hill Climbing com reinícios randômicos*, realiza execuções sucessivas do *Hill Climbing*, aplicando um número fixo de operações aleatórias no DAG a cada execução. Essas operações aleatórias movem o DAG para longe do máximo local. Se a pontuação do DAG permanecer inalterada após R reinícios, o algoritmo encerra a busca e retorna o resultado encontrado.

Além do Hill Climbing, também é possível utilizar outros métodos de busca local, entre eles:

- *Tabu Search*: Uma variação do Hill Climbing que mantém uma lista (Tabu) dos últimos K estados do DAG e finaliza a execução se o estado atual não for melhor que nenhum dos últimos estados, impedindo que o algoritmo volte para estados já localizados (SCUTARI, 2015).

Algoritmo 5: Hill Climbing com reinícios randômicos

Entrada:
D - Conjunto de dados
G - um DAG, em geral vazio, mas não necessariamente
S - Uma função objetivo a ser minimizada
R - Um inteiro que representa o número de reinícios
P - O número de perturbações aleatórias antes de cada reinício

Saída: Um DAG, G' que pertence à mesma classe de equivalência que representa a distribuição em D

```
1  G* = G
2  melhorPontuacao = S(G*)
3  reinicios = 1
4  enquanto reinicios ≤ R faça
5      G** = hillClimbing(D, G*)
6      se S(G**) > melhorPontuacao então
7          melhorPontuacao = S(G**)
8          G* = G**
9          reinicios = 1
10     fim
11     senão
12         para i de 1 até P faça
13             Perturbe G*, adicionando, removendo ou revertendo uma aresta, de modo
                que ciclos não sejam criados.
14         fim
15         reinicios = reinicios + 1
16     fim
17 fim
18 Retorne G*
```

- Algoritmos genéticos: Um tipo de algoritmo que se baseia na teoria da evolução de Darwing. Em contraste ao Hill Climbing, começa com uma quantidade k de DAGs randômicos. A partir daí, utiliza processos de seleção, cruzamento e mutação entre os DAGS (RUSSELL; NORVIG, 2010).

3.4.1.2.2 Funções de pontuação

Diferentemente dos testes de independência em algoritmos baseados em restrições, as funções de pontuação buscam descrever o DAG como um todo e quão bem ele representa as relações entre as variáveis do problema.

As funções de pontuação podem então ser divididas em dois grupos: baseadas na teoria da informação e baseadas no formalismo Bayesiano. Entretanto, apesar de possuírem diferentes origens, as funções possuem o mesmo objetivo, com, até mesmo, algumas equivalências entre os diferentes grupos.

A função de pontuação *Log Likelihood*, LL, deriva diretamente da equação de Informação Mútua (MAUA, 2020) que mede a quantidade de informação que uma variável X revela sobre uma variável Y. No caso das redes Bayesianas, a Log Likelihood mede o quanto um DAG G revela sobre uma distribuição de probabilidade P, utilizando probabilidades empíricas colhidas de um conjunto de dados distribuído por P.

Definição 3.4.9 (Log Likelihood)**.**

$$LL^{MLE} = N \sum_{X \in \mathcal{V}} \hat{I}(X, PA_X) - N \sum_{X \in \mathcal{V}} \hat{H}(X)$$

onde I e H são definidas como:

$$I(\mathcal{X}, \mathcal{Y}) = \sum_{\mathcal{X},\mathcal{Y}} P(\mathcal{X} \cup \mathcal{Y}) \ln \frac{P(\mathcal{X} \cup \mathcal{Y})}{P(\mathcal{X})P(\mathcal{Y})} \textit{ (Informação Mútua)}$$

$$H(\mathcal{X}) = -\sum_{\mathcal{X}} P(\mathcal{X}) \ln P(\mathcal{X}) \textit{ (Entropia de Shanonn)}$$

Por outro lado, uma das alternativas à $LogLikelihood$, mais utilizadas pela literatura(SCUTARI, 2015) é a *Bayesian Information Criterion*, BIC.

Definição 3.4.10 (Bayesian Information Criterion)**.**

$$BIC(G) = \sum_{X \in G} logP(X|PA_X) - \frac{d_X}{2} logn$$

onde d_X é a quantidade de parâmetros em cada nó e n o tamanho da amostra.

Outra função, similar com à BIC é a *Akaike information criterion*, AIC, definida como:

Definição 3.4.11 (Akaike information criterion)**.**

$$AIC(G) = \sum_{X \in G} logP(X|PA_X) - \frac{d_X}{2}$$

onde d_X é a quantidade de parâmetros em cada nó.

Segundo (LAPPENSCHAAR et al., 2013a), a função LL^{MLE} ainda pode ser reescrita com base no formalismo Bayesiano:

$$LL^{MLE} = \sum_{x \in V} \sum_{i=1}^{n} logP(v_i|PA_{v_i})$$

Assim, as funções BIC e AIC se tornam variantes da LL^{MLE}, diferenciando por um termo de *penalidade*. A presença de uma função de penalidade e necessária para evitar *overfitting* durante o aprendizado da estrutura do DAG, pois a função LL^{MLE} tende a favorecer redes mais densas quando não há uma quantidade suficiente de dados. Assim, assim o fator de penalização atua sobre a quantidade de parâmetros na rede, d_X, que acaba por punir DAGs com uma grande concentração de arestas.

3.4.1.3 Métodos Híbridos

Métodos de aprendizado híbrido são atualmente considerados o estado da arte no campo de pesquisa (SCUTARI, 2015). Eles combinam estratégias de algoritmos baseados em restrições e pontuação para criar abordagens confiáveis e adaptáveis a diferentes cenários. Para isso, os algoritmos híbridos unem as melhores características dos demais métodos em duas etapas básicas: restringir e maximizar, conforme o Algoritmo 6 (FRIEDMAN et al., 1999).

Algoritmo 6: Candidatos esparsos

Entrada:
D - Conjunto de dados
G - um DAG, em geral vazio, mas não necessariamente
S - uma função objetivo

Saída: Um DAG, G' que pertence à mesma classe de equivalência que representa a distribuição em D

1 **enquanto** *não houver convergência* **faça**
2 **restringir:** Selecione um conjunto C_i de candidatos a pais de cada variável X_i. C_i deve conter os pais de X_i.
3 **maximizar:** Encontre um DAG, G^* que maximize $S(K)$, tal que o conjunto de arestas de K contém todas as arestas em G e as arestas que ligam cada X_i a seu conjunto de pais, C_i.
4 $G = G^*$
5 **fim**
6 Retorne G

A fase "restringir", no algoritmo seleciona os conjuntos de nós que possam ser pais dos vértices em G. O conjunto de candidatos de cada vértice não precisa ser exato, entretanto os verdadeiros pais devem estar entre os candidatos. Essa tarefa é realizada através de testes de independência para reduzir o Envoltório de Markov de cada nó. Também existem heurísticas específicas que realizam essa tarefa de maneira mais eficiente, como o algoritmo *Max-min parents and children*, MMPC, descrito por Tsamardinos, Brown e Aliferis (2006).

A fase "otimizar"por sua vez, utiliza um algoritmo de busca local para maximizar a função objetivo S, a partir das restrições obtidas na fase anterior. As restrições diminuem, de maneira considerável, o espaço de busca. Ao mesmo tempo, a partir da segunda iteração, a etapa de "restrição"funciona como um tipo de pertubação ao estado do DAG, capaz de mover o algoritmo de otimização de um máximo local.

Uma versão especializada do Algoritmo 6 é o *Max-Min Hill Climbing*, MMHC. O algoritmo emprega uma única iteração de restrição-otimização, usando a heurística MMPC durante a fase de restrição e o *Hill Climbing com reinícios randômicos* para otimização.

Apesar de existirem algoritmos especialistas, como o MMHC, Scutari (2015) mostra que é possível utilizar qualquer combinação de algoritmos de restrição e pontuação para cumprir as etapas de restrição e otimização dos métodos híbridos, podendo-se utilizar a combinação que

mais se adapte ao cenário do problema.

3.4.2 Aprendizado de parâmetros

O aprendizado de parâmetros numa rede Bayesiana consiste em estimar a distribuição de probabilidade conjunta que representa os dados. Uma vez que se conhece a estrutura da rede, pela Equação 3.8, essa tarefa se resume em estimar, para cada vértice da estrutura, a sua tabela de probabilidade condicional:

$$P(X|PA_x)$$

Assim, sejam X e Y dois vértices na estrutura de uma rede Bayesiana, tal que $PA_Y = X$, então as distribuições de probabilidade condicional em X e Y são dadas por:

$$P(X,Y) = \prod_{v \in X,Y} P(v|PA_v)$$

$$P(X,Y) = P(Y|X)P(X)$$

Esse é um processo conhecido como Inferência Estatística e que pode ter uma abordagem Frequentista ou Bayesiana (SCUTARI, 2015).

A abordagem frequentista se baseia na frequência com que eventos ocorrem no conjunto de dados observados para realizar o cômputo das probabilidades. Assim:

$$\hat{P}(X) = \frac{N[X=x]}{N} \text{ e } \hat{P}(Y,X) = \frac{N[Y=y, X=x]}{N}$$

onde N denota o total de eventos observados no conjunto de dados e $N[X=x]$ a quantidade de vezes que o evento $X=x$ ocorre nos dados. Dessa forma:

$$\hat{P}(X|Y) = \frac{\hat{P}(Y,X)}{\hat{P}(X)} = \frac{N[Y=y, X=x]}{N[X=x]} \tag{3.17}$$

Essa abordagem pode ser formalmente definida como o estimador *Maximum Likelihood*, *MLE*.

Definição 3.4.12 (Estimador Maximum Likelihood)**.**

$$MLE(O, PA_O) = \hat{P}(O|PA_O) = \frac{\hat{P}(\bigcap_{V \in \{O \cup PA_X\}} V)}{\hat{P}(\bigcap_{K \in PA_O} K)} = \frac{N[V_1 = v_{1_v}, ..., V_i = v_{i_v}]}{N[K_1 = k_{1_v}, ..., K_j = k_{j_v}]}$$

para todo $V_j \in \{O \cup PA_X\}$ *e todo* $K_j \in PA_x$*, onde* v_{i_v} *e* k_{j_v} *são os valores dos eventos das variáveis* V_i *e* K_j *respectivamente.*

Por outro lado, a abordagem Bayesiana difere da abordagem frequencista ao incorporar a probabilidade a priori das variáveis além de inferir os valores de probabilidade a partir da frequência dos dados. A probabilidade a priori é distribuída uniformemente entre os possíveis valores da variável. Essa inclusão permite equilibrar a probabilidade empírica e funciona como um fator de regularização, pois a probabilidade empírica depende de quão bem os dados representam a distribuição real. Esse fator de regularização é necessário para lidar com o viés da inferência baseada em dados limitados.

Para controlar o balanceamento entre a distribuição a priori e a posteriori na inferência Bayesiana, o estimador é equipado com um parâmetro chamado *Imaginary sample size*, iss. Esse parâmetro informa o quão influente a distribuição a priori será sobre a distribuição a posteriori. Tipicamente, o valor do parâmetro iss é definido como um número pequeno, geralmente entre 1 e 15 (SCUTARI, 2015), para garantir que a probabilidade empírica tenha maior peso na distribuição a posteriori. Essa escolha é feita para evitar que a distribuição a priori influencie excessivamente os resultados da inferência estatística.

Assim, denotando como π a probabilidade uniforme a priori das variáveis envolvidas na distribuição local de cada vértice da rede, $P(X|PA_X)$ pode ser calculada como:

$$\hat{P}(O|PA_O) = \frac{\hat{P}(\bigcap_{V \in \{O \cup PA_X\}} V)}{\hat{P}(\bigcap_{K \in PA_O} K)}$$

Onde $\hat{P}(\bigcap_{E \in V} E)$, para qualquer conjunto de vértices V é dado por:

$$\hat{P}(\bigcap_{E \in V} E) = (\frac{iss}{n + iss})\pi_V + (\frac{n}{n + iss})\hat{p}_V \tag{3.18}$$

onde, π_v e $\hat{p}_v$ denotam, respectivamente, a probabilidade a priori e a frequência normalizada dos eventos em V normalizada pela quantidade de observações n.

3.5 Redes Bayesianas Multinível

A Rede Bayesiana Multinível, RBM, é um formalismo introduzido pela primeira vez por Lappenschaar et al. (2013a). As RBM, foram sugeridas como uma alternativa às técnicas Regressão Linear Multinível, pois conferem ao processo um visão das relações entre as variáveis e das causas dos eventos no problema, diferentemente das técnicas de regressão, que, por sua vez, fornecem uma visão apenas quantitativa do problema.

Por exemplo, na medicina, existe uma necessidade inerente às análises multinível, como argumenta Lappenschaar et al. (2013a). Isso pois, os dados históricos de pacientes são naturalmente hierárquicos, uma vez que pacientes e doenças são sensíveis à variação do ambiente, relações familiares e costumes.

3.5.1 Análises multinível

As técnicas de análise estatísticas padrão são eficazes na análise de dados não hierarquizados. Entretanto, esses tipos de dados nem sempre compreendem a realidade e podem levar a inconsistências estatísticas (HOX, 2017). Isso ocorre pois a análise de dados não hierárquicos não é capaz de detectar a variância introduzida pelas diferenças entre grupos distintos de indivíduos, além de presumir independência entre esses grupos.

Como citado anteriormente, as análises multinível são empregadas na análise de dados estruturados de maneira hierárquica. Por exemplo, no contexto educacional uma tarefa comum é a necessidade de prever a performance estudantil numa determinada disciplina, dadas algumas informações *explicativas* sobre os alunos e o ambiente escolar:

1. **A experiência do professor** - Em anos
2. **A participação do aluno durante as aulas** - Um valor de 0 à 10, sendo zero nada participativo e 10 extremamente participativo
3. **Gênero** - Por simplicidade, feminino ou masculino

Assim, a coleta dos dados segue-se numa escola, turma a turma. Cada turma é vista como um **grupo**, composta por n alunos, onde cada aluno possui como atributos o gênero e a participação. Por sua vez, a experiência de ensino do professor não é uma variável inerente à um aluno específico, mas à todo um grupo de alunos, afetando cada grupo de forma diferente. Na análise multinível, esse tipo de variável é conhecida como **coeficiente randômico** (HOX, 2017). A Figura 14 ilustra a relação entre essas variáveis.

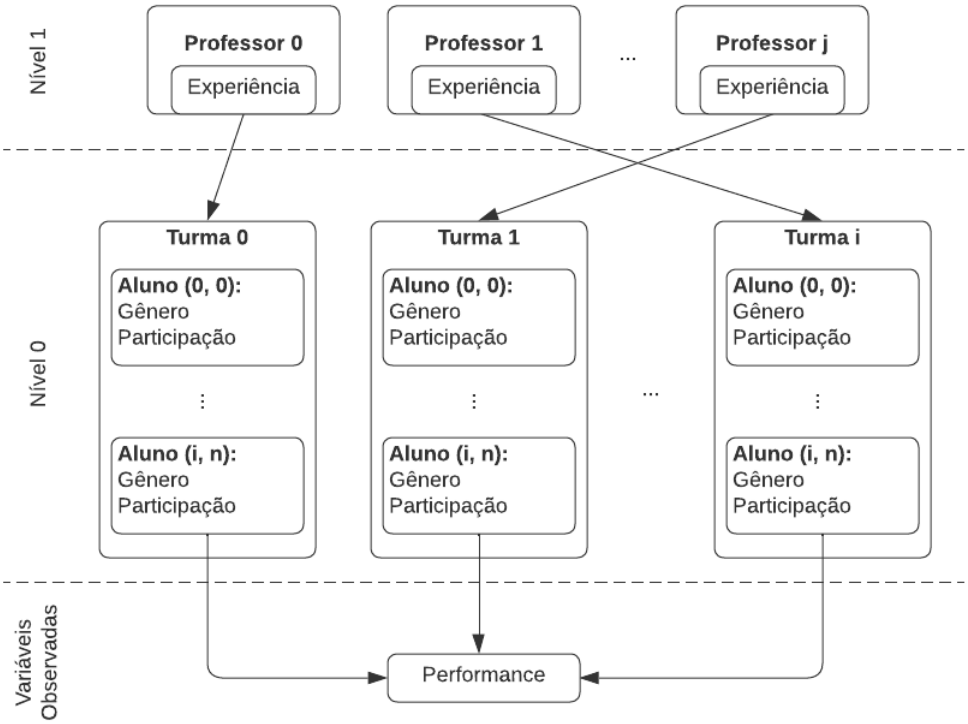
Dessa forma, os dados são estruturados em níveis. No nível 0 estão as variáveis individuais, os alunos no exemplo. Já no nível 1, estão presentes os coeficientes randômicos que exercem influência direta sobre os grupos no nível 0. Na análise multinível, essas variáveis possuem denominações específicas, que representam sua função no modelo: *globais*, *estruturais* e *contextuais* (HOX, 2017).

As variáveis ***globais*** se referem somente ao contexto que estão inseridos e fora dele perdem significado. Por exemplo, na Figura 14 as variáveis *Gênero* e *Participação* são variáveis pertencentes à um indivíduo do tipo *Aluno*. Já a variável *Experiência* pertence à um professor específico, sendo uma uma variável global dentro do contexto do professor.

As variáveis ***estruturais*** pertencem à sub-unidades do domínio, referindo-se sempre à um grupo de indivíduos, *agregando* suas características e para utilizá-las de forma explanatória. No exemplo acima, no nível 1 poderia ser adicionado uma variável estrutural que agrega a renda *per capita* de cada turma. Esse tipo procedimento é chamado de ***agregação*** na análise multinível.

Por fim, as variáveis **contextuais** são resultado de **desagregações** no modelo multinível. As desagregações ocorrem sempre de um nível mais alto para um nível mais baixo, trazendo

Figura 14 – Exemplo de organização de dados hierárquicos com dois níveis



Fonte: Autor

contexto explanatório para variáveis individuais. O ato de desagregar informações consiste em mover variáveis de níveis mais altos para contextos de baixo nível. Por exemplo, na Figura 14, seria possível *desagregar* a variável *Experiência* do nível 1 para se tornar uma variável inerente à cada aluno, de forma individual. As desagregações, em geral, não devem fazer parte da Análise Multinível tradicional, entretanto podem ser usadas em alguns casos para estabelecer simplificar modelos que possuem muitos níveis, tornando mais simples a interpretação dos resultados (HOX, 2017).

A análise multinível precisa ser empregada em conjunto à outras técnicas de análise estatística para levar aos resultados esperados. O poder de síntese em problemas complexos e de grande variância desperta cada vez mais interesse na criação de variações multiníveis para técnicas já existentes (HOX, 2017), como as Regressões Multivariadas Multinível e Análises de Sobrevivência Multinível.

3.5.2 Formalismo das Redes Bayesianas Multinível

Esta subseção objetiva definir o formalismo apresentado por Lappenschaar et al. (2013a), as **Redes Bayesianas Multinível**, que combina a metodologia da Análise multinível com o formalismo das Redes Bayesianas para analisar as interações e dependências probabilísticas entre

dados obtidos a partir de dados hierárquicos. A fundamentação teórica a seguir será apresentada de maneira construtiva, e sumarizada ao final dessa subseção, com o conjunto de definições do formalismo.

As Redes Bayesianas Multinível podem ser definidas em termos de uma variável aleatória dependente O, um conjunto de variáveis explicativas, E, um conjunto de variáveis de nível, L, e um conjunto de variáveis indicativas, I, que modelam a probabilidade de que O ocorra dado o conjunto de suas variáveis explicativas e de nível:

$$P(O|\{E_1, ..., E_n\} \cup (\bigcup_{j=1}^{m} L^j))$$

Onde:

- $\{E_1, ..., E_n\}$ é o conjunto de variáveis aleatórias explicativas no nível 0 da hierarquia. As variáveis contidas nesse nível são grupo-independentes dentro da definição das RBM.
- $L^j = L^j_1, ..., L^j_{m_j}$, com j = 1, ..., m, sendo $m + 1$ a quantidade de níveis na rede.
- I_j uma variável *indicativa* que relaciona um certo grupo de indivíduos num nível j.

A variável indicadora I faz parte da fundação das RBM. Ela possui a função de dividir o domínio do problema em categorias e é definida pela hierarquia dos dados de forma **determinística**.

Outra importante característica numa RBM é a inexistência de filiações que partem de um nó mais baixo para um nó mais alto. Assim, a rede Bayesiana descrita até o momento tem a forma apresentada na Figura 15.

Uma importante característica da variável indicativa I_j é a sua relação determinística com um grupo de variáveis no nível j. Essa relação pode ser utilizada para simplificar a RBM definida até então.

Teorema 4. *Sejam X e Y duas variáveis aleatórias tal que Y é **deterministicamente dependente** de X, isto é, existe uma função tal que $Y = f(X)$. Então, para todos os conjuntos de variáveis aleatórias Z disjuntas de X e Y:*

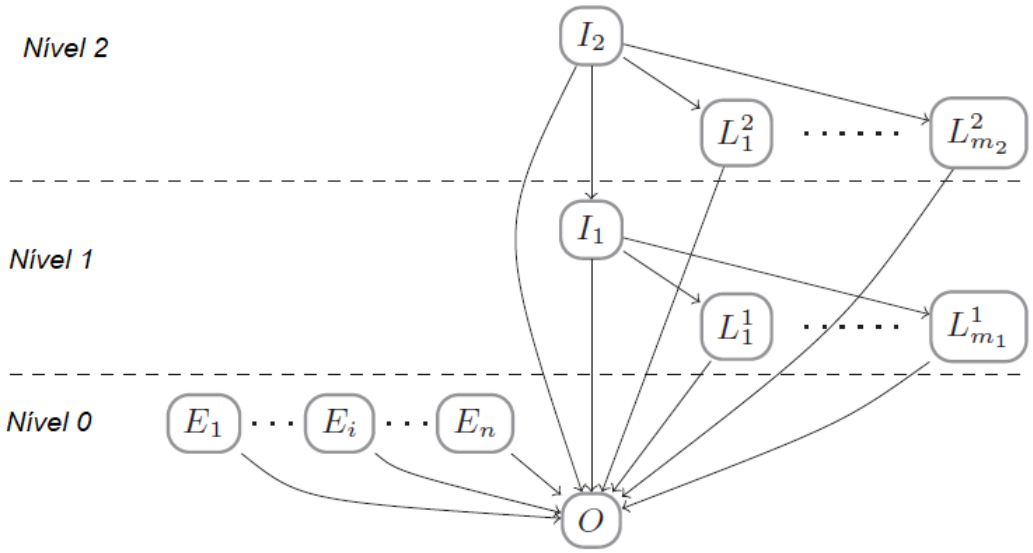
$$Z \perp\!\!\!\perp Y|X$$

Demonstração. Seja P uma distribuição de probabilidade sobre um conjunto de variáveis $\mathbf{Z}$. Então a seguinte igualdade é verdadeira:

$$P(Z|X) = \frac{P(Z,X)}{P(X)} = \frac{\sum_{y\in Y} P(Z,X,y)}{P(X)} = \sum_{y\in Y} P(Z,y|X)$$

$$= \sum_{y\in Y} \frac{P(Z,X,y)}{P(X)} \cdot \frac{P(y,X)}{P(y,X)} = \sum_{y\in Y} \frac{P(Z,X,y)}{P(y,X)} \cdot \frac{P(y,X)}{P(X)}$$

Figura 15 – Rede Bayesiana Multinível Complexa de três níveis



Fonte: Lappenschaar et al. (2013a)

$$= \sum_{y \in Y} \boldsymbol{P(Z|X,y)P(y|X)} \tag{3.19}$$

Como a relação entre X e Y é determinística, também é verdade que:

$$P(Y|X) = \begin{cases} 1 & \text{se } Y = f(X) \\ 0 & \text{se } Y \neq f(X) \end{cases} \tag{3.20}$$

Daí, segue-se que:

$$P(Z|X) = \sum_{y \in Y} P(Z|X,y)P(y|X) = P(Z|X, f(X)) = P(Z|X,Y)$$

Logo,

$$\boldsymbol{Z \perp\!\!\!\perp Y|X}$$

□

A partir do Teorema 4, é possível inferir relações de independência condicional entre as variáveis indicadoras e o restante da Rede Bayesiana da Figura 15, reduzindo o número de arestas na estrutura da rede:

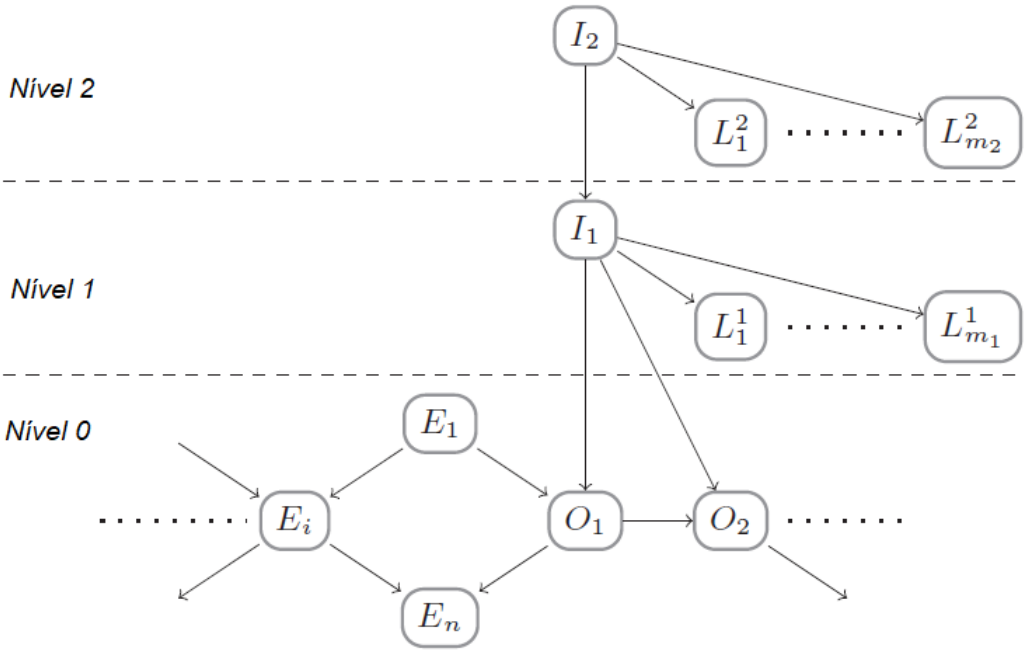
1. Como $P(L_i^j|I_j)$ é determinística, então: $O \perp\!\!\!\perp L_i^j|I_j$. Essa independência implica que **numa RMB não existem arestas que partem das variáveis de grupo, L, para a variável observada O.**

2. Como $P(I_{j-1}|I_j)$ é determinística, então: $O \perp\!\!\!\perp I_j|I_1$ para todo I_j. Essa relação implica que a **variável observada, O, possui como apenas a variável indicadora do nível 1 como pai**. Além disso, **qualquer vértice I_j só tem como pai o vértice I_{j+1}**.

Pelo Teorema 4 as variáveis de saída, O, são independentes das variáveis de nível, L, dado o valor da variáveis indicadoras I. Entretanto, isso não significa que as variáveis de nível não tenham significado. Pelo contrário, as variáveis de nível são uma importante fonte de informações sobre como essas variáveis introduzem variância entre os níveis e grupos de cada nível, sendo este um dos objetivos da metodologia multinível.

Nessas condições, pode-se reescrever a RBM como exposto na Figura 16.

Figura 16 – Uma Rede Bayesiana Multinível genérica



Fonte: Lappenschaar et al. (2013a)

Definição 3.5.1 (Rede Bayesiana Multinível)**.** *Uma Rede Bayesiana, $BN = (G = (V, A), P)$ é uma **Rede Bayesiana Multinível**, **RMB**, se seu conjunto de vértices V é descrito por pela tupla (m, O, E, L, I) tal que, par a par, os conjuntos O, E, L, I são disjuntos e as seguintes afirmações são verdadeiras (LAPPENSCHAAR et al., 2013a):*

- *$m \in \mathbb{N}$ denota o número de níveis na RMB, onde o nível 0 é chamado de **nível base**;*
- *O, o conjunto de **variáveis dependentes** ou de **saída**, está no nível base tal que se $(V \rightarrow O_i) \in A$, então $V \in E \cup (O - O_i) \cup I$;*

- *E, o conjunto de **variáveis explicativas**, está no nível base, tal que se $(V \rightarrow E_i) \in A$, então $V \in (E - E_i) \cup O \cup I$;*
- *$L = \{L^1, ..., L^m\}$, onde cada L^j é um conjunto de **variáveis de grupo** no nível $j \geq 1$. Para a variável de grupo L_i^j, é verdade que:*
  1. *$(V \rightarrow L_i^j) \in A \Rightarrow V = I_j$;*
  2. *$P(L_i^j | I_j)$ é determinística;*
- *$I = I_1, ..., I_m$ são **variáveis indicadoras**, tal que I_j é o único pai de I_{j-1} em G, para todo $1 \leq j \leq m$ e $P(I_{j-1}|I_j)$ é determinística;*
- *O conjunto de variáveis aleatórias X distribuídas por P é dado por $X = \{I \cup E \cup O \cup L\}$;*

3.6 Metodologia CRISP-DM

CRISP-DM é a sigla inglesa para Processo Padrão Transversal da Indústria para Mineração de Dados. CRISP-DM consiste em uma sequência de etapas base que servem para organizar processos relacionados à ciência de dados. A metodologia foi desenvolvida, inicialmente por Chapman (2000), sofrendo algumas atualizações e melhorias ao longo do tempo. Hoje essa metodologia é uma das mais utilizadas no mercado (SUAPRAE; NILSOOK; WANNAPIROON, 2021), sendo adotada como um padrão industrial.

O padrão segue um formato cíclico e de melhoria continua, que inicia no entendimento do negócio e segue até a visualização dos resultados da análise e processamento dos dados envolvidos no problema. Ao todo, seis etapas fazem parte da metodologia. Cada etapa possui responsabilidades e podem ser revisitadas sempre que haja a necessidade, afim de melhorar o processo como um todo. Dessa forma, se trata de uma metodologia flexível, para que erros ou enganos sejam detectados e corrigidos de maneira ágil, sempre visando a solução do próblema sob a ótica do negócio.

Assim, Shearer (2000) define os as seis etapas da metodologia CRISP como: Entendimento do negócio, entendimento dos dados, preparação dos dados, modelagem, avaliação e implantação, ilustrados na Figura 1.

1. **Entendimento do negócio:** Nessa etapa ocorre o levantamento do contexto do problema a ser resolvido, seus critérios para aceite de solução, além dos dados necessários para realização do processo. Apesar de estar principalmente presente na primeira fase, o domínio do negócio deve sempre possuir um papel central durante todo o processo, servindo como guia durante todas as etapas seguintes.
2. **Entendimento dos dados:** Os dados trazem consigo a informação agregada sobre o problema sendo tratado. Por isso, é necessário saber como os dados estão organizados

e distribuídos, afim de realizar as próximas etapas. Assim, primeiro é necessária uma coleta inicial dos dados, para que então seja possível proceder com um processo de análise exploratória, verificando a qualidade e a integridade desse conjunto de dados.

3. **Preparação dos dados:** Os dados levantados na etapa anterior podem apresentar uma série imperfeições ou estar até mesmo incompletos. Dessa forma, a etapa de preparação compreende um conjunto de processos aplicados sobre os dados afim de torná-los mais propícios para a extração de conhecimento. Assim, nessa etapa, primeiro é realizado um processo de *seleção*, onde apenas os dados úteis para a solução do problema são mantidos. Depois, os dados selecionados são tratados num processo de limpeza, onde são removidos os dados ambíguos e estimados os dados faltantes, com uso de técnicas específicas para tal finalidade. Por fim, os dados são *integrados* com fontes de informação complementares ao problema e *formatados*, afim de trazer mais consistência ao processo de modelagem.

4. **Modelagem:** Essa etapa consiste na seleção, construção e teste de um modelo computacional capaz de resolver o problema definido na etapa de entendimento do negócio. É nessa etapa em que a técnica a ser utilizada para o processamento dos dados deve ser escolhida, assim como deve ser feito o ajuste dos parâmetros dessa técnica, para que o modelo se ajuste ao cenário do problema.

5. **Avaliação:** Esta é a etapa subsequente à implantação dos dados, responsável por realizar testes de performance e aceite aos critérios definidos na etapa de entendimento do negócio, antes da disponibilização final para o usuário.

6. **Implantação:** Nessa etapa, o modelo deve ser disponibilizado para uso do conjunto de interessados. Deve ser definido um plano de manutenção e aperfeiçoamento do sistema entregue. Além disso, também deve ser entregue ao usuário final as instruções necessárias para plena utilização da plataforma construída.

4

Experimento com dados sintéticos

Com objetivos didáticos, foi realizado um experimento que reproduz o estudo conduzido por Lappenschaar et al. (2013a). O experimento consistiu em três etapas: a geração do conjunto de dados, o treinamento da Rede Bayesiana Multinível e, por fim, a validação do modelo.

4.1 Geração do *dataset*

Os dados sintéticos utilizados no experimento possuem as variáveis listadas na Tabela 6. Essas variáveis simulam os dados de pacientes portadores de múltiplas comorbidades, que realizam seu tratamento em clínicas familiares.

As funções determinísticas e probabilísticas associadas à cada variável também é definida por Lappenschaar et al. (2013a). Assim, utilizando a linguagem R, foram amostrados 10000 pacientes de maneira uniforme sobre as 50 clínicas, seguindo as regras e probabilidades definidas abaixo, onde o valor $\mathcal{N}(\mu_s, \sigma_s)$ é amostrado da distribuição $\mathcal{N}(\mu_s, 0.01)$, com $\mu_s = 0.25, 0.30, 0.35, 0.40, 0.45$ para $s = 1, 2, 3, 4, 5$ correspondendo ao valor de I2.

- $G = Binomial(p = 0.5)$
- $D1 = Binomial(p = 0.5 + 0.1G + \mathcal{N}(0, 0.1))$
- $D2 = Binomial(p = 0.2 + \mathcal{N}(\mu_s, \sigma_s))$
- $D3 = Binomial(p = 0.2 + 0.1D1 + 0.2D2 + 0.2D1D2 + \mathcal{N}(0, 0.1))$

- $I2 = \begin{cases} 1 & \text{se } I1 \in \{1, ..., 10\} \\ 2 & \text{se } I1 \in \{11, ..., 20\} \\ 3 & \text{se } I1 \in \{21, ..., 30\} \\ 4 & \text{se } I1 \in \{31, ..., 40\} \\ 5 & \text{se } I1 \in \{41, ..., 50\} \end{cases}$
- $L1 = \begin{cases} 1 & \text{se } I1 \mod 10 \in \{0, 1\} \\ 2 & \text{se } I1 \mod 10 \in \{2, 3\} \\ 3 & \text{se } I1 \mod 10 \in \{4, 5\} \\ 4 & \text{se } I1 \mod 10 \in \{6, 7\} \\ 5 & \text{se } I1 \mod 10 \in \{8, 9\} \end{cases}$
- $L2 = \begin{cases} 0 & \text{se } I2 \in \{1, 2\} \\ 1 & \text{se } I2 \in \{3, 4, 5\} \end{cases}$

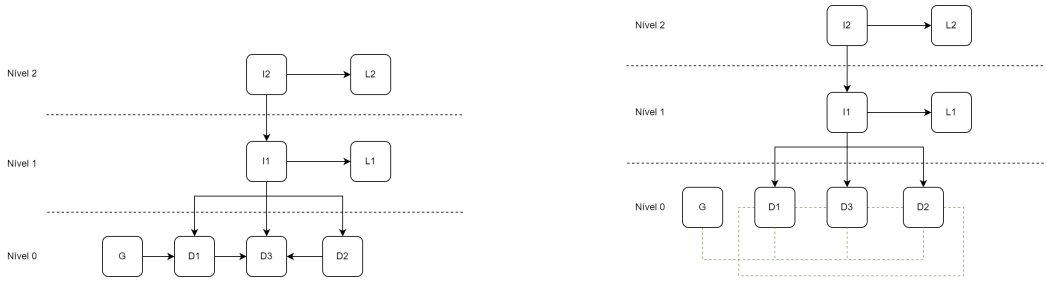
Tabela 6 – Variáveis envolvidas no experimento

Nome da variável	Tipo	Domínio	Descrição
I2	Discreta	1, 2, 3, 4, 5	Variável indicadora do nível 2. Sinaliza a região em que uma clínica está situada.
I1	Discreta	1, 2, ..., 49, 50	Variável indicadora de nível 1. Indica à qual clínica um determinado paciente pertence.
L1	Discreta	1, 2, 3, 4, 5	Variável de grupo do nível 1. Indica o tipo de uma determinada clínica familiar.
L2	Discreta	0, 1	Variável de grupo do nível 2. Determina o tipo de uma clínica.
D1	Discreta	0, 1	Variável observada 1. Uma comorbidade associada à um paciente, como a diabetes.
D2	Discreta	0, 1	Variável observada 2. Uma comorbidade associada à um paciente, como a retinopatia.
D3	Discreta	0, 1	Variável observada 3. Uma comorbidade associada à um paciente, como a hipertensão.
G	Discreta	0, 1	Variável explicativa. Modela o sexo do paciente.

4.2 Especificação do modelo

Dadas as definições acima, a MBN que define o contexto apresentado é representada na Figura 17a. Então, de acordo com o formalismo das MBN, as relações topológicas a serem

aprendidas na rede são as arestas entre as variáveis observadas e as variáveis explicativas, representadas na Figura 17b pelas linhas tracejadas em verde.



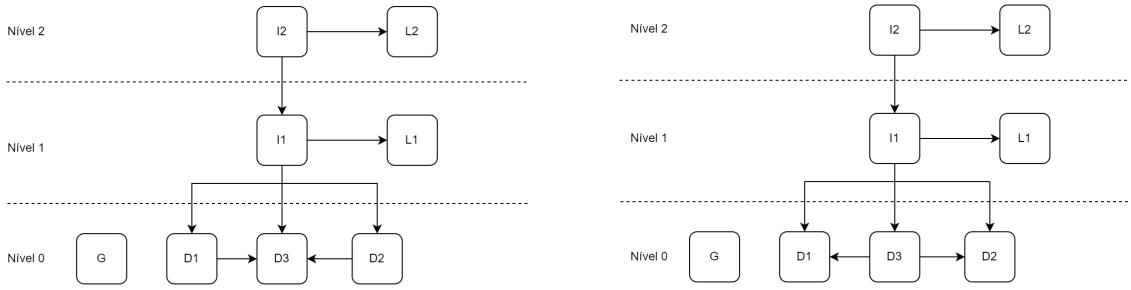
(a) DAG real (b) Arestas a disponíveis para o treinamento

Figura 17 – Estrutura real da MBN e a rede a ser fornecida como entrada o treinamento

Para codificação das restrições da MBN da Figura 17b foi utilizado um sistema de *blacklists* e *whitelists* de arestas. Assim, as arestas contidas na *blacklist* não podem existir na rede, enquanto que as arestas contidas na *whitelist* devem existir na rede. Para gerar estas listas foi desenvolvido um utilitário, chamado de *mbnlearn* (REIS, 2023), que recebe a definição das variáveis da MBN e retorna as devidas restrições. A *whitelist* e a *blacklist* que definem a MBN da Figura 17b seguem listadas nos anexos B.1 e B.2.

4.3 Treinamento

O treinamento da estrutura da rede foi realizada utilizando três configurações de algoritmos: PC, *Hill Climbing*(HC) com o *score* BIC e *Hill Climbing*(HC) com o *score* AIC. O DAG resultante da aplicação de cada método segue na Figura 18.



(a) PC e HC-AIC (b) HC-BIC

Figura 18 – DAGs resultantes da aplicação dos métodos de treinamento

Na Figura 18a é possível observar que ambos os métodos, PC e HC-AIC, chegaram ao mesmo DAG, enquanto que o HC-BIC(Figura 18b) possui uma inversão nas arestas $D1 - D3 - D2$. Além disso, também é possível observar que nenhum dos métodos foi capaz de identificar o arco $G \rightarrow D1$, existente no DAG real(Figura 17a). Analisando os *scores* de cada DAG é possível observar que o DAG real, em geral possui um score menor que os demais, indicando que os

métodos de treinamento não são capazes de identificar a variação de 0.01 na distribuição de probabilidade de $D1$, quando G é positivo.

Quadro 6 – Comparativo dos *Scores* dos DAGs resultantes do treinamento da estrutura da rede

Método	**BIC**	**AIC**	**BDE**
DAG Real	-69394.52	-66492.36	-67941.83
PC	-69185.70	-66463.80	-67692.85
HC-AIC	-69185.70	-66463.80	-67692.85
HC-BIC	-69027.32	-66485.68	-67472.97

Após o treinamento da estrutura da rede, foi realizado o aprendizado das tabelas de probabilidade condicional de cada vértice da rede. Para isso foi utilizado o método Bayesiano com um tamanho de amostra imaginária (iss) 10, para cada um dos três DAG.

4.4 Validação

Para a validação do modelo, foi realizada a validação cruzada dos resultados do treinamento, em conjunto com a avaliação da Área Sob a Curva ROC (AUC) dos modelos. Conforme Lappenschaar et al. (2013a), a validação do modelo foi feito realizando a previsão do valor da variável $D3$ em relação às variáveis $D1$, $D2$ e $L1$: $P(D3|D1, D2, L1)$. O Quadro 7 mostra a acurácia e a AUC de cada modelo treinado.

Quadro 7 – Comparativo da acurácia dos modelos treinados

Método	**Acurácia**	**AUC**
DAG Real	0.7015	0.6582
PC	0.7012	0.6580
HC-AIC	0.7012	0.6580
HC-BIC	0.7	0.6577

Assim, é possível observar que os modelos, apesar de suas particularidades, obtiveram métricas de performance similares. As performances similares entre os modelos, demonstram que a inversão da relação causal das arestas D1, D2 e D3 entre os modelos(Figura 18), não possui grande impacto no poder de previsão do modelo. Isso pode ser explicado pela influência causada pelos vértices filhos nos vértices pais, conforme abordado na Seção 3.3.2.

5

Entendimento do negócio e dos dados

Este capítulo visa definir os critérios para o aceite do modelo apresentado no Capítulo 7, discorrendo também sobre os dados necessários para alcançar tal propósito. Além disso, será feita também uma revisão sobre os dados utilizados, discorrendo sobre as características dos históricos escolares que serão a base desse trabalho.

5.1 Entendimento do negócio

Como disposto nos capítulos 1 e 2, a evasão escolar no ensino superior é um problema multifacetado e que aflige diversos países ao redor do mundo, fazendo estudantes e instituições mais vulneráveis.

Assim, este estudo visa criar uma solução computacional robusta e capaz de identificar estudantes em risco de evasão escolar, para que, então, gestores e responsáveis sejam capazes de tomar decisões a fim de mitigar o problema.

Para isso, deverá ser desenvolvido um modelo capaz de analisar dados da performance universitária dos estudantes envolvidos e decidir quando um indivíduo está mais propenso ou não à evadir a instituição de ensino. O modelo também possui alguns requisitos que devem ser alcançados afim de atingir os demais objetivos deste trabalho. Assim, o Quadro 8 lista os requisitos necessários para o aceite do modelo apresentado neste trabalho.

Quadro 8 – Requisitos de aceite para o modelo desenvolvido neste trabalho

O modelo deve ser capaz de identificar quando um estudante estiver em risco de evasão.
O modelo deve possuir uma acurácia de ao menos 70%.
O modelo deve ser capaz de trabalhar com dados hierárquicos.
O modelo deve possuir bases no formalismo das Redes Bayesianas Multinível.

Dessa forma, para atingir as necessidades acima, serão disponibilizados os dados da performance universitária dos estudantes envolvidos no trabalho.

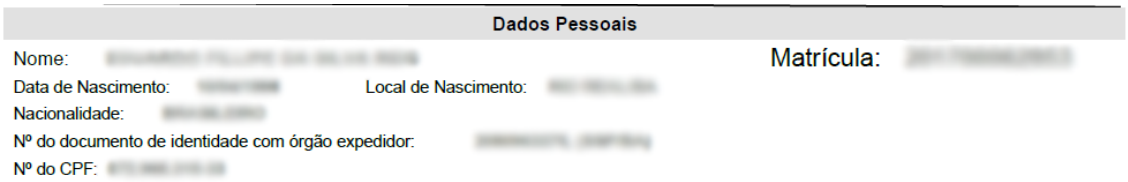
5.2 Entendimento dos dados

Como citado anteriormente, este trabalho utilizará os dados da performance universitária dos estudantes do Departamento de Computação da Universidade Federal de Sergipe. Esses dados serão obtidos a partir do histórico universitário, gerado pelo sistema acadêmico da instituição, dos discentes do departamento.

Os históricos escolares utilizados neste trabalho representam um resumo da vida estudantil dos alunos na universidade, contando também com algumas informações pessoais que identificam cada indivíduo. Cada históricos escolar possui, em geral, 6 seções principais: A de identificação pessoal, de identificação do curso, de índices acadêmicos, componentes curriculares cursados, componentes curriculares pendentes e resumo de progresso no curso.

A seção de identificação pessoal reúne os dados sobre o estudante, contendo sua matrícula, nome, data de nascimento, número de documento de identificação e outras informações, conforme a Figura 19. Estes dados são sensíveis em sua maioria, dada sua capacidade de identificar o portador de cada histórico.

Figura 19 – Seção de dados pessoais do históricos escolar



Já a seção de identificação do curso traz dados que identificam o curso do indivíduo, como o nome e certificação junto ao Ministério da Educação, MEC, ilustrada na Figura 20.

As seções de índices acadêmicos e resumo de progresso mostram de maneira agregada o desempenho do aluno no decorrer do programa de graduação. Os índices acadêmicos trazem quatro medidas, listadas na Tabela 7 e exibidas na Figura 20, que descrevem o comportamento do estudante até aquele momento no tempo. Por sua vez, o resumo do progresso no curso, exemplificado na Figura 21, mostra qual porcentagem das componentes curriculares exigidas para formação do estudantes já foram finalizadas, dividindo a informação em Total Integralizado e Pendente, para componentes curriculares obrigatórios, componentes curriculares optativos e atividades curriculares obrigatórias.

Por fim, no currículo também podem ser encontradas dados relacionados à cada componente curricular cursada ou pendente de integralização pelo discente. Essas informações

Figura 20 – Seção de dados do curso do histórico escolar

Dados do Curso

Curso: **CIÊNCIA DA COMPUTAÇÃO - SÃO CRISTÓVÃO - PRESENCIAL - DCOMP - BACHARELADO - V**

Ênfase:
Currículo: **08** Status: **ATIVO**
Ano/Período Letivo Inicial: **2017.1**
Forma de Ingresso: **Vestibular**
Renovação de Reconhecimento do Curso: **Portaria 921/2018/MEC de 27/12/2018, publicado(a) em 28/12/2018**

Índices Acadêmicos
IECH: **0.865** IEPL: **0.757**
MC: **8.29** IEA: **5.4283**

Período Letivo Atual: **11** Prazo para Conclusão:**2025.2**
Trancamentos: **Nenhum** Perfil Inicial:**0**
Dispensas: **Nenhuma**
Prorrogações: **4 períodos letivos**
Prorrogações Não Contabilizadas: **2021.2, 2021.1, 2020.2, 2020.1**
Ano/Período Letivo de Saída: Data da Colação de Grau:
Tipo Saída:
Data de Saída:
Trabalho de Conclusão de Curso:
Data de Expedição do Diploma:

Figura 21 – Seção de resumo do progresso no curso do histórico escolar

	Obrigatórias				Optativos	Total	
	Comp. Curricular		Atividade	CH Total	Comp.Curricular/Atividade		
	CR	CH	CH		CH	CR	CH
Exigido	152	2280	510	2790	420	152	3210
Integralizado	152	2280	210	2490	270	152	2760
Pendente	0	0	300	300	150	0	450

Tabela 7 – Índices acadêmicos e suas descrições

Sigla	Índice	Descrição
IECH	Índice de Eficiência em Carga Horária	Divisão da carga horária com aprovação pela carga horária utilizada
IEPL	Índice de Eficiência em Períodos Letivos	Divisão da carga horária acumulada pela carga horária esperada
MC	Média de Conclusão	Média do rendimento acadêmico final obtido pelo estudante nos componentes curriculares, ponderadas pela carga horária discente dos componentes
IEA	Índice de Eficiência Acadêmica	Produto da MC pelo IECH e pelo IEPL

Fonte: SIGAA

estão presentes, respectivamente, nas seções Componentes Curriculares Cursados (Figura 22) e Componentes Curriculares Pendentes (Figura 23). As informações são dispostas no formato de uma tabela, contendo o nome e código das disciplinas, nome do professor, turma, período em que a componente foi, ou será, cursada, nota final e situação.

Dessa forma, é possível afirmar que o histórico escolar estudantil é uma importante fonte

Figura 22 – Seção de componentes curriculares cursados do históricos escolar

Componentes Curriculares Cursados/Cursando

Ano/Período Letivo	Componente Curricular			CH	Turma	Freq %	Nota	Situação
2017.1	#	COMP0196	FUNDAMENTOS DA COMPUTAÇÃO *Prof. Dr. THIAGO XAVIER ROCHA DE SOUZA (30h)*	30	02	100.0	10.0	APROVADO
2017.1	e	COMP0197	PROGRAMAÇÃO IMPERATIVA *Prof. Dr. GIOVANNY FERNANDO LUCERO PALMA (90h)*	90	08	100.0	9.8	APROVADO
2017.1		COMP0337	MÉTODOS E TÉCNICAS DE PESQUISA *Profa. Dra. MARIA AUGUSTA SILVEIRA NETTO NUNES (60h)*	60	02	--	--	TRANCADO
2017.1	e	MAT0064	CÁLCULO I *Prof. Dr. RICARDO PINHEIRO DA COSTA (90h)*	90	21	80.0	6.5	APROVADO
2017.1	e	MAT0067	VETORES E GEOMETRIA ANALÍTICA *Profa. DANIELE COSTA FONSECA MENEZES (60h)*	60	10	100.0	6.8	APROVADO
2017.1	e	MAT0104	FUNDAMENTOS DE MATEMATICA PARA COMPUTACAO *Prof. Dr. LUIS JONATHA RODRIGUES DE OLIVEIRA (90h)*	90	02	100.0	7.2	APROVADO
2017.2	e	COMP0198	PROGRAMAÇÃO ORIENTADA A OBJETOS *Prof. Dr. ANDRÉ BRITTO DE CARVALHO (60h)*	60	03	96.66	9.6	APROVADO
2017.2	e	COMP0212	ESTRUTURA DE DADOS I *Prof. Dr. CARLOS ALBERTO ESTOMBELO MONTESCO (60h)*	60	02	96.66	8.4	APROVADO
2017.2	e	COMP0219	CIRCUITOS DIGITAIS I *Prof. Dr. LUIZ BRUNELLI (60h)*	60	01	83.33	8.3	APROVADO
2017.2	e	COMP0220	LABORATÓRIO DE CIRCUITOS DIGITAIS I *Prof. Dr. LUIZ BRUNELLI (30h)*	30	04	86.66	8.9	APROVADO

Figura 23 – Seção de componentes curriculares pendentes do históricos escolar

Componentes Curriculares Obrigatórios Pendentes: 5

Código	Nível	Componente Curricular		CH
COMP0485	8	TRABALHO DE CONCLUSÃO DE CURSO I	Matriculado	60 h
COMP0308	9	ATIVIDADES COMPLEMENTARES		120 h
COMP0486	9	TRABALHO DE CONCLUSÃO DE CURSO II		120 h
ENADE	-	ENADE INGRESSANTE PENDENTE		0 h
ENADE	-	ENADE CONCLUINTE PENDENTE		0 h

de informações sobre a performance de um discente durante o processo de graduação. Isso pois é possível analisar informações sobre cada componente curricular já cursada pelos estudantes, comparar seu desempenho com relação ao progresso do curso, além de fornecer uma síntese da performance estudantil a partir dos índices acadêmicos. Assim, essas informações podem ser utilizadas não apenas neste trabalho, mas como também na realização de diversos estudos, análises e processos úteis para a universidade.

6

Preparação dos Dados

Este capítulo visa esclarecer o processo de extração e tratamento dos dados envolvidos neste trabalho, elencando os métodos e etapas que compreendem o download dos históricos escolares, extração, anonimização e limpeza dos dados.

6.1 Download dos históricos escolares

Como informado no Capítulo 5 a principal fonte de dados utilizada neste trabalho será o histórico escolar da graduação dos estudantes envolvidos. Dessa forma, a primeira etapa do processo de preparação dos dados é a obtenção dos históricos escolares.

Assim, com a colaboração da secretaria e da coordenação do Departamento de Computação, foi possível realizar o *download* de todos os históricos, de alunos ativos e inativos do departamento, a partir do Portal do Coordenador disponível no SIGAA.

Entretanto, o sistema acadêmico não disponibiliza uma função de *download* em lotes dos históricos, existe apenas a possibilidade de realizar o *download* de um histórico por vez. Ainda, devido ao grande número de informações que o sistema agrega sobre os discentes no histórico escolar, cada *download* costuma levar, em média, 30 segundos. Essa limitação torna o processo manual de *download* dos históricos, dos mais de 3000 discentes, inviável. Além disso, os históricos também possuem informações sensíveis e sigilosas sobre os estudantes envolvidos. Dessa forma, foi necessário o desenvolvimento de um sistema utilitário que fizesse de forma automática o *download* desses históricos.

Assim, esse utilitário foi implementado utilizando técnicas de *WebScraping*, sendo capaz de interagir com o sistema acadêmico de através da sua interface *Web*, o SIGAA. O programa foi desenvolvido em linguagem Python e fazendo uso, principalmente, do projeto Selenium.

Selenium é um framework multiplataforma com capacidade de interagir diretamente

com *WebBrosers*, enviando comandos e consultando informações sobre as páginas *web* neles processados, através da estrutura *HTML* da página. Assim, uma vez que se conheça a estrutura da página alvo da análise, é possível realizar ações na página como selecionar conteúdos, acionar botões e enviar comandos ao próprio *browser*.

Dessa maneira, inicialmente foi necessário realizar o mapeamento das ações e interações necessárias para realizar o download de cada histórico no Portal do SIGAA. Após, foi realizada a implementação dessas ações, via Selenium, com cada elemento HTML responsável por disparar as ações mapeadas na página. Também, foi necessário implementar alguns comportamentos para lidar com eventos que possam ocorrer durante o processo, como recuperação de erros no SIGAA. A implementação de tais comportamentos permitiram o desenvolvimento de uma aplicação resiliente e rastreável, capaz de finalizar o processo mesmo durante a ocorrência de erros fatais no sistema acadêmico. Assim, a aplicação possui uma interface de comunicação com o usuário através de linha de comando, onde os argumentos de execução são listados na Tabela 8.

Tabela 8 – Argumentos de execução do utilitário desenvolvido para realizar o download dos currículos

Argumento	Descrição
-h, –help	Mostra uma tela de ajuda, listando todos os argumentos e como utilizá-los.
-m, –matriculas	O caminho do arquivo onde se encontram as matrículas a serem processadas. O formato do arquivo é CSV com uma única coluna 'matricula'.
-db, –database	Arquivo onde o banco de dados local do progresso do download dos históricos ficará armazenado.
-al, –auto_login	Ativa ou desativa o auto login. Se ativada, os parâmetros –user e –password são obrigatórios. Se falso, o utilitário irá aguardar por um tempo determinado (argumento –login_timeout) para que um ser humano efetue a autenticação manual na tela de login do SIGAA.
-u, –user	Usuário para auto login no SIGAA.
-p, –password	Senha para auto login no SIGAA.
-lt, –login_timeout	Tempo em segundos que a aplicação espera para que um humano efetue login no SIGAA.
-crs, –course	Iniciais do curso que será utilizado, escolhido logo após o login: CC, SI ou EC.
-ne, –notify_email	Ativa ou desativa a notificação do progresso do download via e-mail.
-nt, –notification_targets	Lista de emails, separados por vírgula, que irão receber as notificações de progresso.
-t, –notification_threshold	Limiar de notificação. Esse valor é utilizado para enviar noticações via e-mail. Ex: Se o limiar é 15, a cada 15 matrículas processadas uma notificação será enviada.

Assim, através dessa aplicação, foi realizado o download do histórico dos TODO: NUMERO DE HISTÓRICOS discentes envolvidos neste trabalho.

6.2 Extração dos dados dos históricos

A versão em PDF dos históricos escolares apresenta os dados de maneira não estruturada, dada a própria natureza do formato dos arquivos. Por esse motivo, foi necessário realizar a extração dos dados não estruturados para um formato no qual pudessem ser mais facilmente utilizados.

Para isso, foi desenvolvido um segundo *software* auxiliar, dessa vez com o objetivo de transformar os dados não estruturados contidos na versão em PDF para um modelo relacional, ilustrado na Figura 24, capaz de representar as informações e relações presentes nos históricos.

Nem todas as informações mostradas na Figura 24 serão utilizados neste trabalho, entretanto um modelo de dados robusto é essencial para que seja possível a criação de um *dataset* que possa ser reutilizado, posteriormente, na realização de outros tipos de análises. Por esse motivo, todos os dados contidos nos históricos foram extraídos e estruturados nesse formato.

Para fazer a extração, novamente foi implementada uma aplicação utilizando a linguagem Python, dessa vez com o intuito de transformar os dados não estruturados, contidos no PDF, no modelo normalizado da Figura 24.

Para extrair os dados mostrados de forma tabular no PDF, foi utilizada o *Tabula*, um *framework* implementado em Java e com interfaces Python, capaz de identificar tabelas contidas em documentos PDF e convertê-las para objetos de baixo nível como *dataframes* e matrizes. Essa estratégia teve ser adotada devido à complexidade da disposição desses dados como texto na estrutura do PDF, que tornava inviável a implementação de uma estratégia de extração baseada puramente em texto, dado o curto prazo e demais objetivos deste trabalho.

Os demais dados puderam ser extraídos com o uso de expressões regulares capazes de segregar as informações requeridas com base na distribuição espacial do texto dentro de cada PDF. Para garantir que essa distribuição fosse sempre a mesma em todos os arquivos, o texto foi extraído com base no *layout* dos PDFs, fazendo-se uso da biblioteca *PDFMiner*.

Além disso, dada a quantidade de históricos a serem manipulados, foi aplicada uma estratégia de processamento paralelo, utilizando a abordagem de múltiplos processos a nível do sistema operacional. Para isso, foi utilizada uma fila sincronizada compartilhada entre os processos. A fila contém o caminho de cada arquivo PDF a ser processado e cada processo possui uma conexão SQL com a base de dados onde os dados serão salvos. A criação dessa fila e dos processos, além da coleta de *logs* é responsabilidade do processo principal, chamado de executor. O fluxo resumido de interação do programa é mostrado na Figura 25.

6.3 Anonimização dos dados

Os dados contidos no histórico escolar não se restringem apenas à informações sobre a performance dos discentes, mas também informações pessoais que os identificam. Dessa forma, foi necessário realizar a anonimização das informações sensíveis contidas nos históricos, sumarizadas na Tabela 9.

Tabela 9 – Dados sensíveis encontrados nos históricos escolares

Nome do campo	Tipo de dado	Descrição
Nome	Alfanumérico	Nome do discente
Matrícula	Alfanumérico	Matrícula do discente
Nº do documento de identidade com órgão identificador	Alfanumérico	RG do discente
Nº do CPF	Alfanumérico	CPF do discente

Assim, foram gerados dois *datasets*, um público e outro restrito. Ambos os *datasets* foram submetidos a diferentes processos de anonimização, afim de garantir a proteção dos dados pessoais dos indivíduos envolvidos.

O *dataset* público passou por um processo de anonimização, com a remoção completa dos dados sensíveis (Tabela 9), garantindo que não haja qualquer possibilidade de identificação dos envolvidos no trabalho. No entanto, é importante destacar que os dados ainda possuem campos identificadores incrementais que relacionam um indivíduo a seus determinados históricos. Embora não permitam a identificação direta dos participantes, é possível realizar a individualização dos dados de cada histórico. Dessa forma, o *dataset* pode ser utilizado por outros pesquisadores interessados em realizar estudos sobre a performance estudantil de alunos de graduação, especialização ou pós-graduação, desde que sigam as normas éticas e legais para a utilização de dados anonimizados.

Por sua vez, o *dataset* restrito foi submetido a um processo de encriptação utilizando o algoritmo *AES* (*Advanced Encryption Standard*) (NIST, 2001) em modo *CBC* (*Cipher Block Chaining*) e uma chave de criptografia de 256 *bits* para proteger as informações sensíveis. O acesso a esse conjunto de dados é restrito ao orientador do trabalho, que é o único detentor da chave de criptografia. É importante salientar que esse *dataset* contém informações que, após a desencriptação, são capazes de identificar os estudantes envolvidos, por isso seu acesso é restrito. A finalidade é realizar atualizações periódicas nos dados, adicionando novos históricos e atualizando os já existentes, a fim de manter a base de dados atualizada e completa ao longo do tempo. A utilização do *AES* com chave de criptografia de 256 *bits* em modo *CBC* é amplamente reconhecida como uma das formas mais seguras de encriptação (STALLINGS, 2017), garantindo que os dados permaneçam protegidos contra acesso não autorizado e ameaças potenciais à segurança da informação. Isso possibilita a realização de análises mais precisas, detalhadas e

atualizadas sobre a performance acadêmica dos alunos de graduação.

6.4 Validação dos dados extraídos

Dado o caráter automatizado do processo de extração dos dados, fez-se necessária a validação das informações obtidas. Essa validação foi feita em duas etapas.

A primeira etapa foi realizada durante o processo de desenvolvimento do utilitário de extração. Utilizando-se técnicas de desenvolvimento orientado à testes, TDD, foi possível realizar a validação do software de extração desde a sua concepção, usando o histórico escolar de alunos voluntários de diferentes cursos. Ao todo foram implementados e executados 235 casos de teste, que validaram se o comportamento do utilitário era o esperado.

A segunda etapa se deu após o processo de extração e que cruzou os dados extraídos dos históricos com os à disposição no portal Power BI disponibilizado pela Superintendência de Indicadores de Desempenho Institucional, SIDI. Este portal, ilustrado na Figura 26, dispõe dados sobre o desempenho escolar de toda a UFS, sendo possível realizar agregações por turma, disciplina, centro acadêmico, departamento e período letivo. Apesar de não conter todas as informações extraídas dos históricos escolares, o portal apresenta dados quantitativos sobre o total de alunos matriculados, trancados, aprovados e reprovados em componentes curriculares, além exibir a média dos alunos nestas componentes.

Após a comparação, atestou-se uma diferença entre os valores de alunos matriculados e médias de notas. A diferença foi, em média, de 3 alunos, para números de matrículas, e de 0,15, para as médias de notas, por componente curricular. Assim, constatou-se que a causa da diferença se deve a desatualização do portal com relação aos históricos escolares.

Segundo o setor responsável pelo portal na UFS, os dados extraídos dos históricos refletem a situação do DCOMP após a integralização do período 2021.2, enquanto que o portal, durante a validação, não possuía a situação atualizada das notas de algumas disciplinas.

Dessa maneira, a validação foi capaz de comprovar a confiabilidade dos dados extraídos, assegurando que o *dataset* retrata a situação, de forma detalhada, da performance estudantil dos discentes ativos do Departamento de Computação da UFS no período letivo de 2021.2.

Figura 24 – Diagrama representativo do modelo entidade relacionamento utilizado para representar os dados do históricos escolar

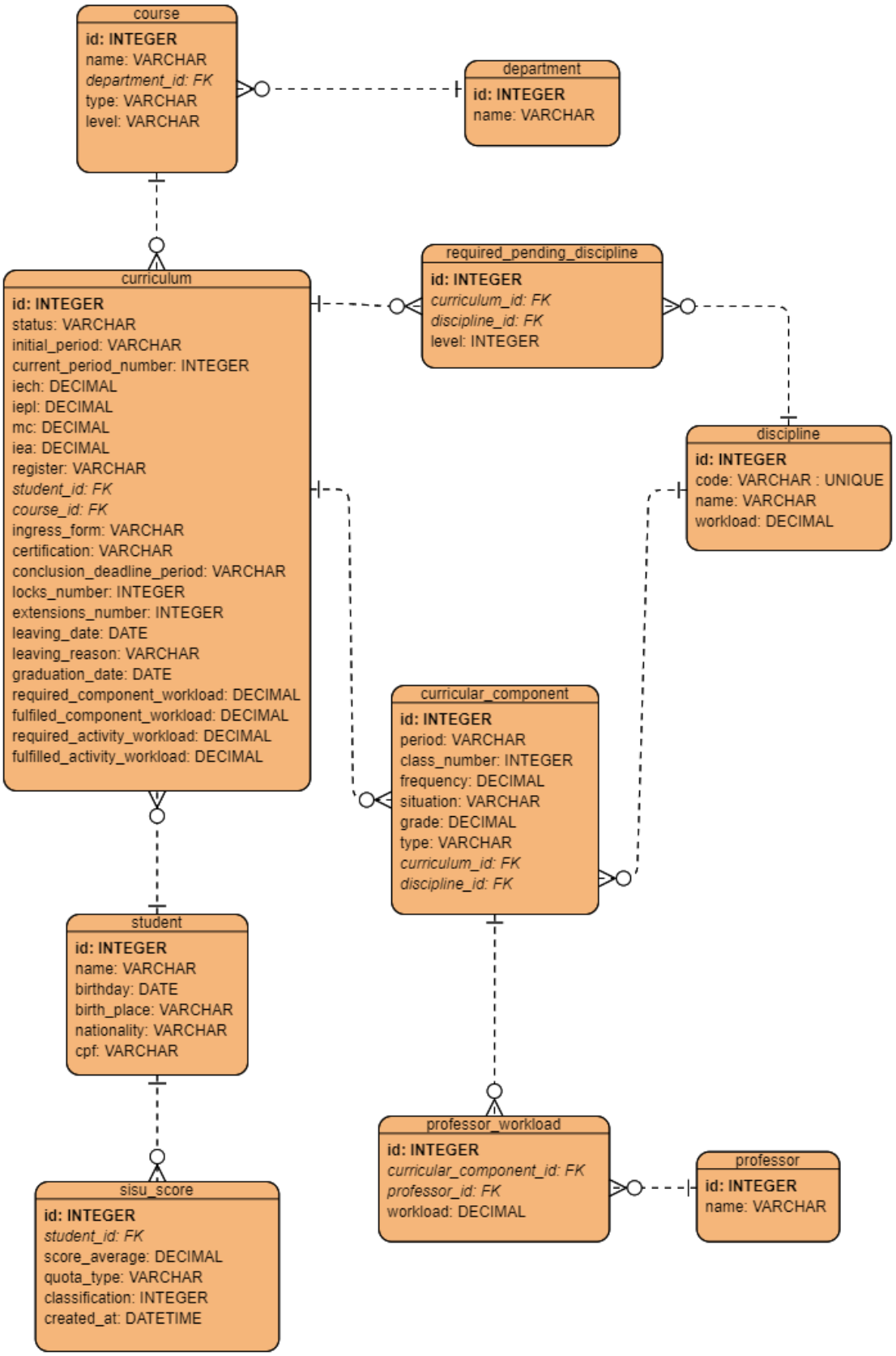


Figura 25 – Fluxo resumido do funcionamento do utilitário de extração

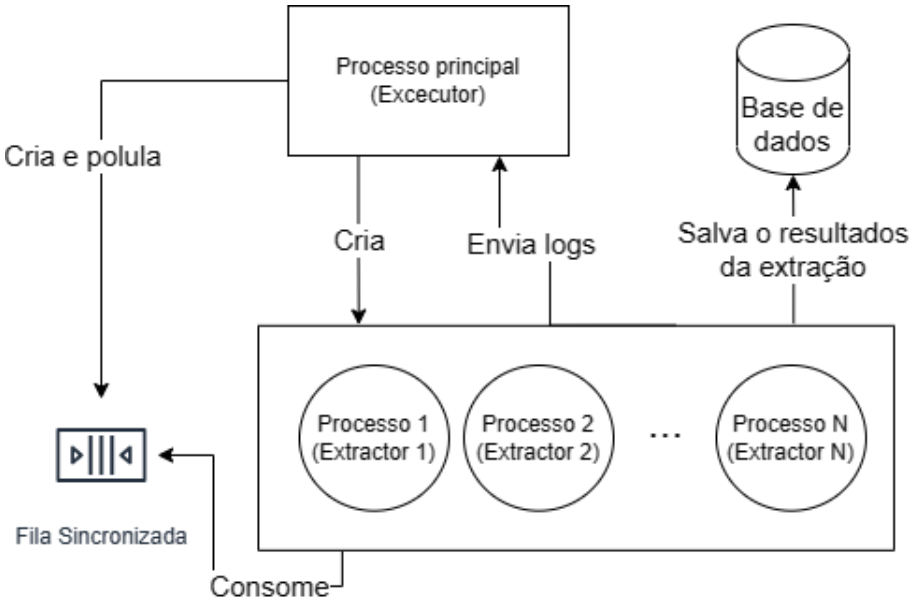


Figura 26 – Página principal do Portal PowerBI disponibilizado pela SIDI.

7

Modelagem e Treinamento

Esta seção descreve o processo de criação e treinamento do modelo de Rede Bayesiana Multinível utilizado para analisar a relação entre a evasão estudantil e o desempenho de alunos de graduação. Inicialmente, são descritos os dados utilizados para treinar o modelo, incluindo suas características e limitações. O processo de treinamento da RBM é, então, descrito em detalhes, incluindo os métodos de inferência utilizados e os parâmetros adotados. Por fim, são apresentados os resultados do treinamento e algumas discussões sobre o processo.

7.1 Modelagem

O modelo escolhido como base da RBM é fundamentado, principalmente, em um subconjunto de dados agregados que se referem aos cursos e componentes curriculares cursados pelos estudantes. Essa escolha foi motivada por dois fatores principais: o primeiro é a disponibilidade dos dados, uma vez que as médias e índices estão disponíveis no mesmo formato durante toda a vida acadêmica do aluno, o que facilita a sua coleta e utilização. Em contrapartida, a disponibilidade do desempenho de componentes individuais varia de estudante para estudante, o que dificulta a sua utilização em um modelo.

O segundo motivo se refere à aplicabilidade do modelo em cursos heterogêneos no tocante às disciplinas cursadas, uma vez que se torna inviável a criação de um modelo genérico que englobe todas as componentes curriculares das diferentes grades curriculares. Portanto, o uso de dados agregados sobre cursos e componentes curriculares permite que o modelo seja aplicado de forma mais ampla e consistente, possibilitando a análise de fatores que influenciam o desempenho escolar em diferentes contextos e cursos.

Assim, a Tabela 10 lista o conjunto inicial de dados utilizados para o treinamento do modelo. Inicialmente, foi feita a extração dos dados brutos a partir do banco de dados mostrado na Figura 24, selecionando-se os dados dos alunos que possuem os índices acadêmicos mais

recentes presentes em seus históricos. Essa consulta resultou em 2603 resultados.

Tabela 10 – Dados utilizados durante o treinamento da RBM

Nome do campo	Descrição	Tipo	Domínio
Curso	Curso ao qual o currículo pertence	Categórica Multinominal	CC, EC, SI
IECH	Índice de Eficiência em carga horária	Categórica Multinominal	[0.3...0.4), [0.4...0.5), [0.5...0.6), ..., [0.9...1]
MC	Média de Conclusão	Categórica Multinominal	[0...1), [1...2), [2...3), ..., [9...10]
IEPL	Índice de Eficiência em Períodos Letivos	Categórica Multinominal	[0.3...0.4), [0.4...0.5), [0.5...0.6), ..., [1...1.1]
IEA	Índice de Eficiência em Períodos Letivos	Categórica Multinominal	[0...1), [1...2), [2...3), ..., [9...10]
Faixa etária	Idade do estudante no período letivo inicial	Categórica Multinominal	[0...17), [17...20), [20...25), ..., [55...∞]
CH Obrigatória Cumprida	Percentual da carga horária obrigatória cumprida pelo discente	Categórica Multinominal	[0...20), [20...40), [40...50), ..., [90...100]
CH Optativa Cumprida	Percentual da carga horária optativa cumprida pelo discente	Categórica Multinominal	[0...20), [20...40), [40...50), ..., [90...100]
Naturalidade	Estado de nascimento do discente	Categórica Multinominal	ESTADOS BRASILEIROS
Forma de Ingresso	Forma de ingresso ao curso	Categórica Multinominal	Vestibular, Transferência, Portador de Diploma
Nº Trancamentos	Número de trancamento que o discente realizou no curso	Categórica Multinominal	[0...1), [1...2), [2...4), [4...∞]
Razão de Saída	Razão de saída do curso de graduação	Categórica Binária	Abandono, Concluído

Dessa forma, os dados foram tratados e limpos, sendo normalizados, discretizados e as conclusões e evasões da universidade foram agregadas em uma única variável (razão de saída). A discretização foi realizada seguindo a definição dos dados na Tabela 10, com a divisão do domínio contínuo das variáveis em intervalos predefinidos. Por sua vez, a agregação da variável razão de saída foi feita de acordo com o mapeamento apresentado na Tabela 11.

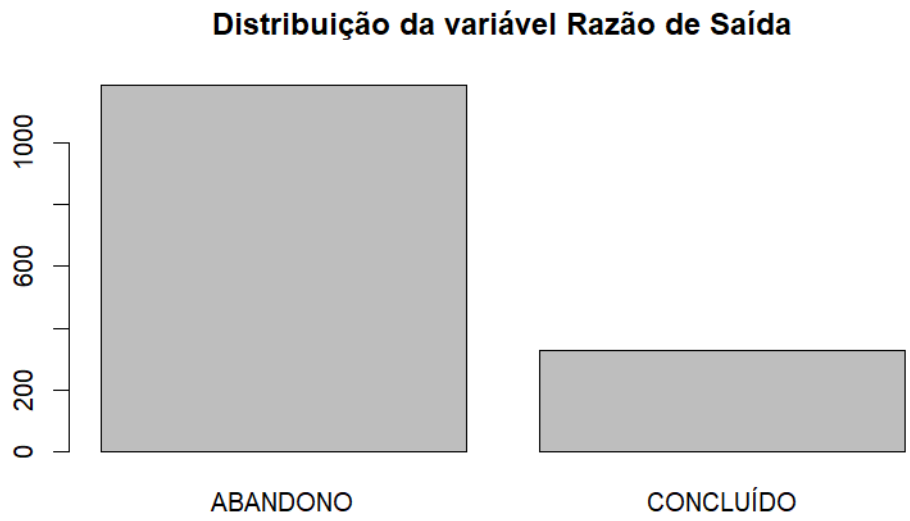
Além disso, também foram removidos os casos de históricos escolares com status *ATIVO*, uma vez que o valor da variável "Razão de saída"ainda não existe, pois o aluno ainda segue cursando a graduação. Após a limpeza, o conjunto de dados foi reduzido a um total de 1563 registros. Entretanto, devido aos altos índices de evasão no departamento de computação, existe uma disparidade entre o número de observações com razão de saída "ABANDONO"e "CONCLUÍDO", como se pode observar na Figura 27.

Para criar as restrições do modelo de RBM, utilizou-se o utilitário *mbnlearn* (REIS, 2023). Esse utilitário recebe a especificação dos tipos de variáveis e retorna as restrições da rede no formato de uma *blacklist* e uma *whitelist*, conforme sugerido por Lappenschaar et al.

Tabela 11 – Mapeamento do campo Razão de saída

Classe original	Classe mapeada
ABANDONO	ABANDONO
CADASTRO CANCELADO	ABANDONO
CANC.. NOVO VESTIBULAR	ABANDONO
CANCELAMENTO ESPONTÂNEO	ABANDONO
DECURSO DE PRAZO MÁXIMO P/ CONCLUSÃO DE CURSO	ABANDONO
NÃO CONFIRMAÇÃO VÍNCULO	ABANDONO
Não atendeu à convocação para o Cadastro Específico conforme Edital 21/2017/PROGRAD	ABANDONO
TRANSF.P/OUTRA IES	TRANSFERÊNCIA
Transferência Interna	TRANSFERÊNCIA
CONCLUÍDO	CONCLUÍDO

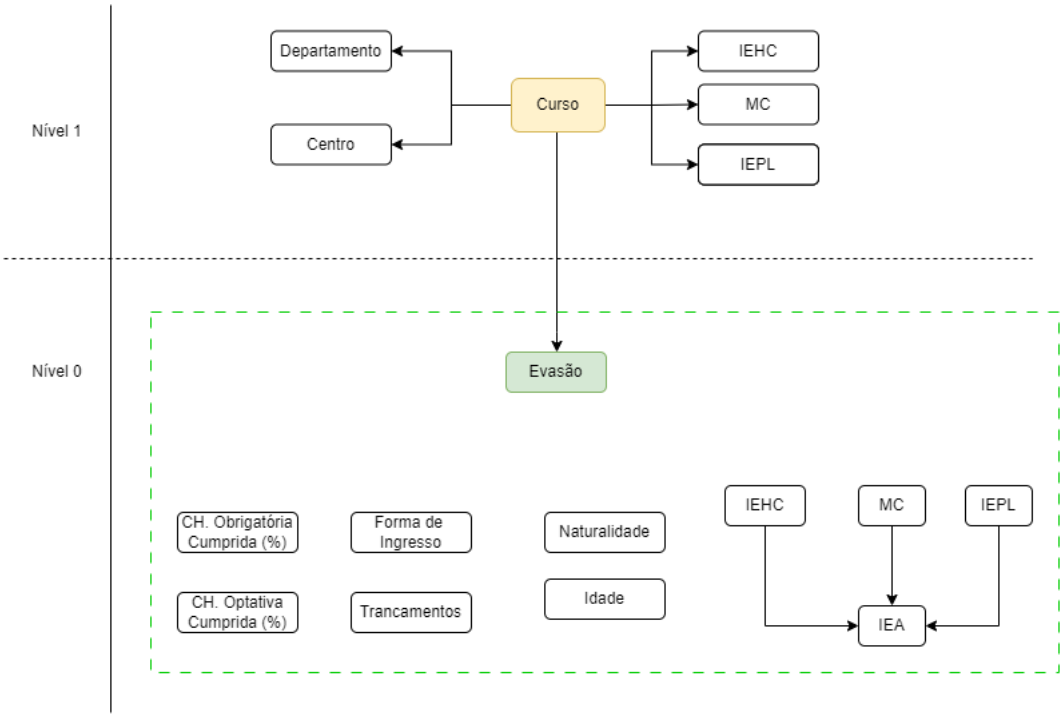
Figura 27 – Distribuição da variável Razão de saída



(2013a). No total, o modelo foi representado por um conjunto de 96 arcos na *blacklist* e 7 arcos na *whitelist*, utilizando a mesma estrutura empregada na Seção 4.2.

Assim, a RMB modelada para o processo de treinamento é a ilustrada na Figura 28. É possível observar que as variáveis relacionadas diretamente ao estudante estão no nível 0, enquanto que as variáveis de nível 1 agregam informações sobre o curso do aluno, que por sua vez divide o domínio da rede como a variável indicadora do nível 1. Na imagem, também é possível observar que as relações a serem aprendidas durante o treinamento da estrutura da rede estão contidas na área tracejada em verde.

Figura 28 – Modelo base da RBM para o treinamento.



7.2 Treinamento

O treinamento foi realizado utilizando o método de Validação cruzada com 10 dobras e utilizou um subconjunto de 80% dos 1563 exemplos obtidos após a limpeza dos dados, discriminada na seção anterior. O subconjunto foi selecionado de maneira randômica, com distribuição uniforme e independente, a fim de evitar enviesamento durante o treinamento. Os demais 20% do *dataset* foram reservados para utilização durante a etapa de validação do modelo.

A etapa de avaliação da Validação Cruzada foi homogênea entre todos os modelos treinados, utilizando inferência aproximada para prever o valor da variável Razão de Saída a partir da Cobertura de Markov para esse nó na estrutura e 1000 amostras.

No processo de treinamento da estrutura da rede, foram explorados diferentes tipos de algoritmos para a seleção da melhor estrutura. Para isso, utilizamos algoritmos dos três tipos existentes: baseados em restrições, pontuação e algoritmos híbridos, abordados na Seção 3.4. Essa abordagem foi adotada para garantir uma análise mais ampla e completa das possibilidades de estruturas, além de permitir uma comparação mais robusta entre os resultados obtidos. A performance de cada algoritmo durante a validação cruzada é sumarizada na Tabela 12.

Os algoritmos baseados em pontuação demonstraram uma habilidade superior em capturar as nuances do dataset, o que resultou em uma acurácia de mais de 95% para o *score* AIC. Com base nesse resultado, o modelo *Hill Climbing - AIC* foi selecionado para avançar no processo

Tabela 12 – Resultado do treinamento dos modelos em validação cruzada

Algoritmo	Tipo	Tipo de Inferência para CPTs	Acurácia	Parâmetros
PC Stable	Restrição	Bayesiana	79.05%	Teste: *Pearson* X^2 ISS: 5
PC Stable	Restrição	*Maximum Likelihood*	77.3%	Teste: *Pearson* X^2
Hill Climbing	Pontuação	Bayesiana	94.89%	*Score* : *BIC* ISS: 5 Reinícios: 2000 Perturbações: 2
Hill Climbing	Pontuação	Bayesiana	95.98%	*Score* : *AIC* Reinícios: 2000 Perturbações: 2
Max Min Hill Climbing	Híbrido	Bayesiana	77.1%	*Score* : *BIC* ISS: 5 Reinícios: 2000 Perturbações: 2 Teste: *Pearson* X^2

de avaliação. No entanto, para prosseguir, foi necessário escolher uma única estrutura gerada a partir de cada uma das 10 dobras do processo de validação cruzada.

A fim de realizar essa tarefa, foi necessário avaliar o nível de confiança de cada uma das arestas geradas durante o treinamento. Esse processo foi conduzido por meio do método desenvolvido por Friedman, Goldszmidt e Wyner (1999), o qual se baseia no cálculo da frequência com que as arestas ocorrem em múltiplos modelos. Para cada aresta, um valor entre 0 e 1 foi atribuído, onde 0 significa que a aresta não foi gerada em nenhum modelo e 1 significa que a aresta foi gerada por todos os modelos. O resultado desse procedimento está resumido na Tabela 13, que apresenta as arestas que não foram geradas em todas as dobras do treinamento, ou seja, que possuem um nível de confiança menor que 1.

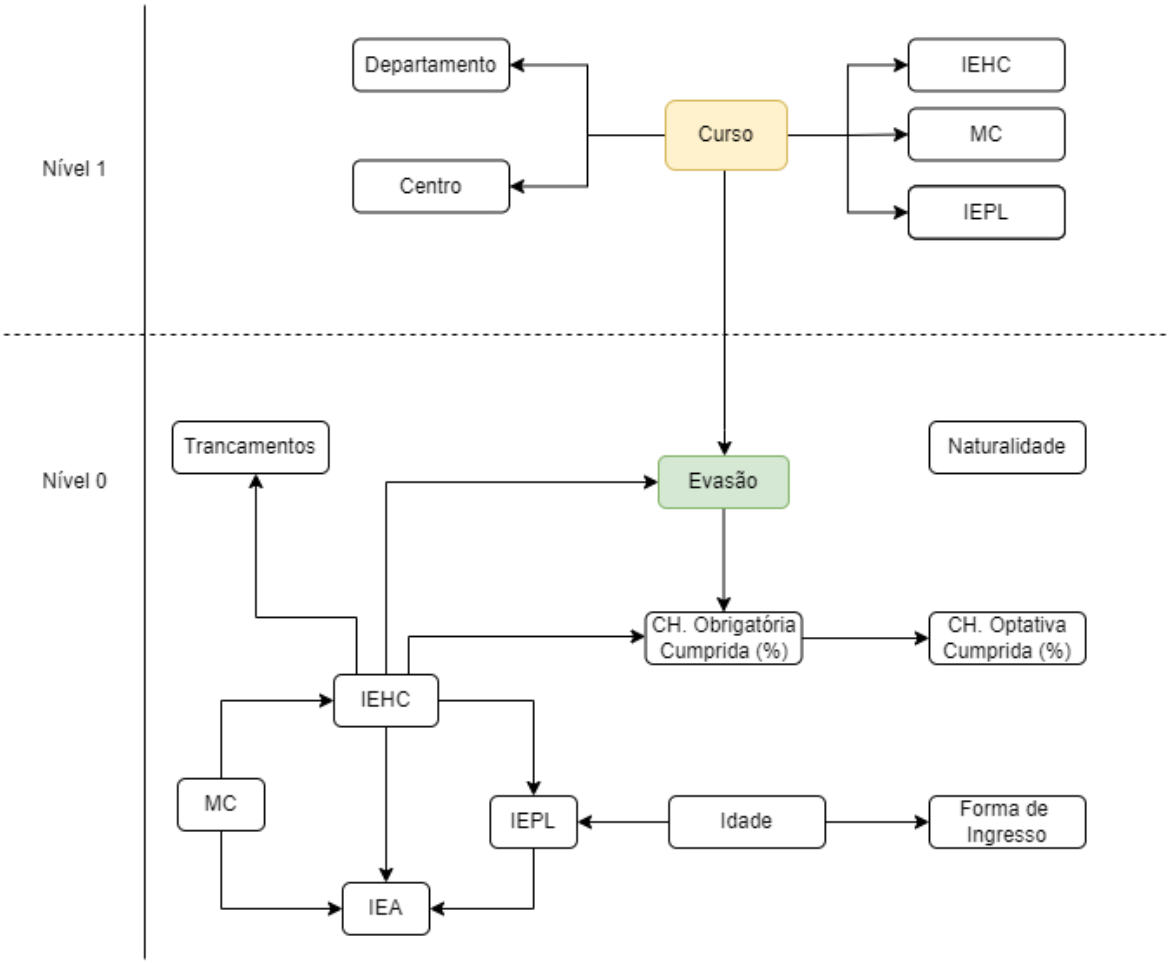
Tabela 13 – Avaliação da confiança das arestas do modelo HC-AIC

Vérice Origem	Vértice Destino	Confiança
MC	CH Obrigatória Cumprida	0.1
IECH	CH Obrigatória Cumprida	0.9

Com base na tabela, é possível verificar que o processo de treinamento como um todo foi consistente através das diferentes dobras da Validação Cruzada. Apenas duas arestas possuem um nível de confiança menor que 1. A aresta que possui o vértice *MC* foi removida devido ao seu baixo nível de confiança, enquanto que a aresta que tem o *IECH* como origem foi mantida, pois foi gerada por 9 dos 10 modelos. Assim, o modelo resultante do treinamento é o mostrado na

imagem Figura 29.

Figura 29 – Estrutura da rede resultante do treinamento



A partir da estrutura resultante do treinamento, é possível observar que a variável "Naturalidade"é independente do restante da rede e, portanto, não é relevante para prever a probabilidade de evasão de um discente.

Ademais, a lista de variáveis que influenciam diretamente na variável "Razão de saída", sua Cobertura de Markov (Definição 3.3.6), é composta pelos nós "Curso", "CH Obrigatória Cumprida"e "IECH". Dessa forma, a probabilidade de evasão de um determinado aluno neste modelo pode ser calculada da seguinte maneira:

$$P(Evasão|CH\ Obrigatória\ Cumprida, IECH, Curso) \tag{7.1}$$

Com base na Equação 7.1, é possível verificar que além do desempenho acadêmico médio em componentes curriculares(MC), as medidas de regularidade no curso, como as variáveis $IECH$ e $CH.\ Obrigatória\ Cumprida$, têm maior influência nos índices de evasão. Isso indica que alunos

com menor frequência ou cumprimento de disciplinas obrigatórias têm maior probabilidade de evadir. O modelo sugere que essas variáveis são fatores importantes na detecção da evasão e devem ser consideradas pelos gestores educacionais ao desenvolver estratégias de prevenção e intervenção.

Por outro lado, a variável *MC* possui influência direta sobre a distribuição de *IECH*, que por sua vez influencia a probabilidade de evasão. Essa relação em cadeia é representada pela seguinte equação de probabilidade, através da aplicação do Teorema 2:

$$P(Evasão, IECH, MC) = P(Evasão|IECH)P(IECH|MC)$$

Essa discussão pode ser ampliada para o âmbito da causalidade. É possível observar relações de causa e efeito que fluem entre as variáveis de índices acadêmicos. Por exemplo, um baixo desempenho em *MC* tende a comprometer o *IECH* de um estudante, uma vez que há um aumento nas reprovações. Da mesma forma, um baixo *IECH* implica em um atraso na carga horária do curso, resultando em um menor *IEPL*. Além disso, a variável *Idade* não é influenciada diretamente por outros vértices da rede, mas afeta a *Forma de Ingresso* e o *IEPL*, indicando que a rede capturou a variação de oito anos na média de idade dos estudantes, quando agrupados pela forma de ingresso (Tabela 14). Podemos fazer uma análise semelhante para as variáveis *CH. Obrigatória Cumprida* e *CH. Optativa Cumprida*, que mostram que a rede conseguiu inferir que à medida que as disciplinas obrigatórias são cumpridas, as disciplinas optativas começam a ser cursadas pelos discentes. As relações causais identificadas pela rede podem fornecer insights valiosos para os gestores educacionais na concepção de políticas e intervenções voltadas para a melhoria do desempenho acadêmico e a prevenção da evasão escolar.

Tabela 14 – Idade média dos estudantes agrupados pela forma de ingresso

Forma de ingresso	Média de Idade
Portador de Diploma	28.63
Transferência Compulsória	24.13
Transferência Interna	24.03
Transferência Voluntária	23.21
Sob Judice	22
Vestibular	20.83

A topologia do modelo selecionado e os resultados obtidos por Lappenschaar et al. (2013a) confirmam que as Redes Bayesianas Multinível são uma ferramenta eficaz para mapear as relações causais entre as variáveis usando métodos clássicos de treinamento. No entanto, é importante destacar que a incorporação do conhecimento prévio de especialistas pode melhorar

ainda mais a estrutura da rede, especialmente quando se trata de relações ainda não observadas no modelo. Isso pode tornar o processo de previsão de probabilidades ainda mais consistente e preciso.

8 Avaliação

Neste capítulo, será realizada a avaliação do modelo treinado e selecionado no capítulo anterior, abordando sua performance de classificação sobre o conjunto de dados de teste. Será apresentada uma análise detalhada, incluindo a curva ROC e a área sob a curva (AUC) do modelo.

8.1 Performance no *dataset* de testes

No capítulo anterior, foi mencionado que o conjunto de dados utilizado para o treinamento dos modelos corresponde a 80% do total de dados disponíveis. Na etapa de avaliação, utilizou-se os 20% restantes para testar o desempenho do modelo, permitindo a análise de dados ainda não processados pelo mesmo.

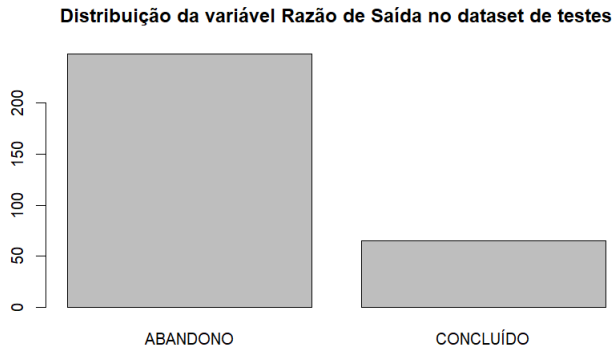
Na etapa de avaliação, utilizou-se a Cobertura de Markov para prever a variável "Razão de saída"a partir das observações presentes no conjunto de dados de teste, conforme apresentado na Equação 7.1. Assim como o *dataset* de treinamento, o *dataset* de testes possui um desbalanceamento entre as classes observadas, conforme a Figura 30.

Assim, o modelo selecionado (Figura 29) obteve uma acurácia de 88.17%, quando testado contra as 313 observações contidas no *dataset* de testes.

8.2 Análise da curva ROC

Embora a acurácia seja uma métrica importante, sua utilização isolada pode ser insuficiente para sumarizar o desempenho do modelo, especialmente quando há um desbalanceamento nos dados observados. Portanto, além da acurácia, também foi realizada a análise da curva ROC, que permite avaliar a quantidade de falsos positivos, falsos negativos, verdadeiros positivos e verdadeiros negativos detectados pelo modelo durante a avaliação. Assim, os tipos de detecções feitas pelo modelo é sumarizada na Quadro 9.

Figura 30 – Distribuição das classes observadas no *dataset* de treino



Quadro 9 – Matriz de confusão do modelo selecionado (Figura 29)

		Referência	
		ABANDONDO	CONCLUÍDO
Previsão	ABANDONO	222 (Verdadeiros Positivos)	11 (Falsos Positivos)
	CONCLUÍDO	26 (Falsos Negativos)	54 (Verdadeiros Negativos)

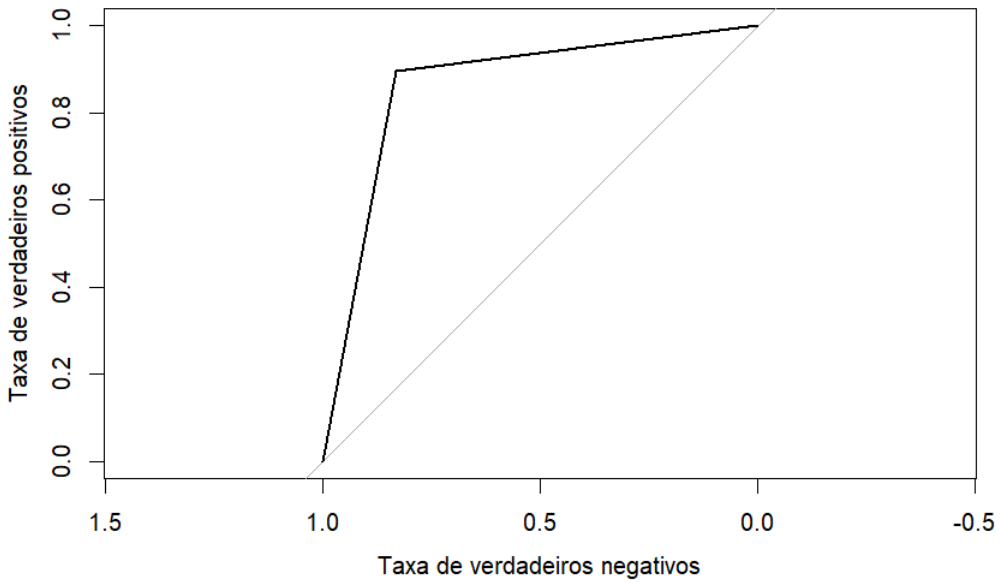
A partir da matriz de confusão do modelo é possível observar que a taxa de detecção de verdadeiros negativos é 6.44% menor que a taxa de detecção de verdadeiros positivos. Esse é um indicativo de possível *overfit* do modelo, devido à grande quantidade de observações de alunos que abandonaram o curso no *dataset* de treinamento.

Em relação à curva ROC, ilustrada na Figura 31, observou-se que o modelo apresentou um desempenho superior ao de um classificador aleatório para qualquer valor de *threshold*, indicando que há margem para ajustar o limiar de classificação do modelo e personalizá-lo para diferentes aplicações. Além disso, a Área Sob a Curva (AUC) obtida foi de 0.86, o que indica um bom desempenho do modelo na tarefa de classificação.

A curva ROC foi gerada a partir do cálculo das taxas de verdadeiros positivos (TVP) e de verdadeiros negativos (TVN). A TVP mede a proporção de vezes em que o modelo classificou corretamente um abandono em relação ao total de classificações de abandono, enquanto a TVN mede a proporção de vezes em que o modelo classificou corretamente uma conclusão em relação ao total de classificações de conclusão. As definições são matematicamente expressas nas Equações 8.1 e 8.2.

$$TVP = \frac{Nº\ de\ verdadeiros\ positivos}{Nº\ de\ falsos\ positivos + Nº\ de\ verdadeiros\ positivos} = 0.8952 \tag{8.1}$$

Figura 31 – Curva ROC do modelo selecionado (Figura 29)



$$TVN = \frac{N^o \textit{ de verdadeiros negativos}}{N^o \textit{ de falsos negativos} + N^o \textit{ de verdadeiros negativos}} = 0.8308 \tag{8.2}$$

Além disso, apesar da quantidade limitada de dados e do desbalanceamento do *dataset*, o modelo atendeu a todos os requisitos necessários para ser aprovado, conforme o Quadro 8, inclusive superando a pontuação mínima exigida. Além disso, o modelo é capaz de operar em diferentes situações, pois apresenta margem para variação de limiar de classificação.

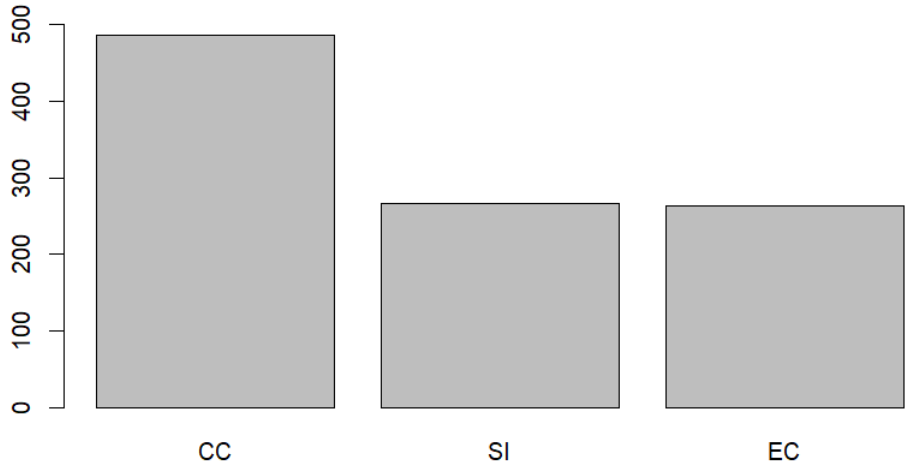
Por fim, foi constatado que o modelo apresentou uma acurácia superior à média dos modelos que utilizaram somente dados de performance universitária, evidenciados pelo mapeamento sistemático realizado no Capítulo 2, que mostrou uma média de 83%. No entanto, é necessário conduzir estudos comparativos com outros modelos, específicos para a Universidade Federal de Sergipe, a fim de identificar as melhores soluções para a problemática em questão.

8.3 Detectando a probabilidade de evasão de alunos ativos do DCOMP

Após o treinamento e avaliação do modelo selecionado, HC-AIC, os dados dos alunos ativos foram adquiridos a partir dos dados brutos e submetidos aos mesmos processos dos dados utilizados para o treinamento, descritos no Capítulo 7.

Assim, 1016 currículos de alunos ativos foram selecionados, com a distribuição ilustrada na Figura 32.

Figura 32 – Distribuição dos currículos ativos por curso

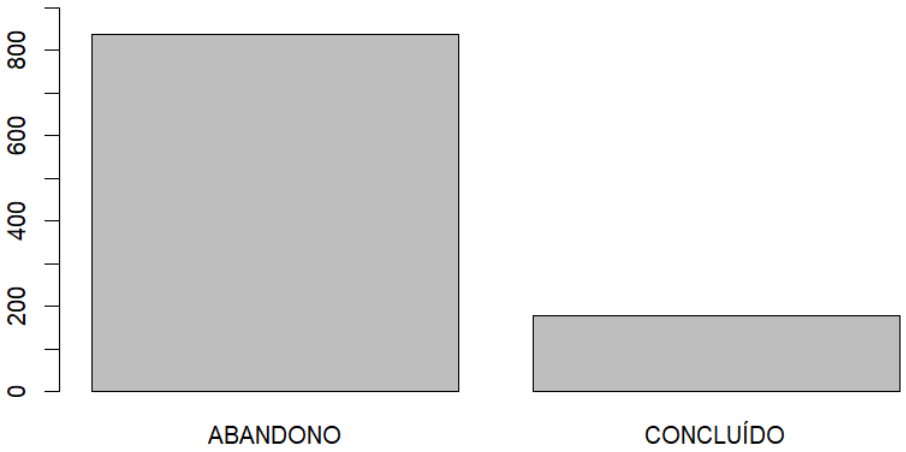


Para cada um dos currículos foi realizada a predição da probabilidade de abandono utilizando o modelo HC-AIC dadas as variáveis que compõem a Cobertura de Markov da variável *Razão de saída*, conforme a Equação 7.1.

Ao total, 835 currículos foram classificados como em risco de evasão, ou seja com valor da variável *Razão de Saída* como *ABANDONO*. Os demais 181 currículos foram classificados como "CONCLUÍDOS", indicando baixo risco de evasão. O limiar de classificação utilizado para essa tarefa foi de 0.5.

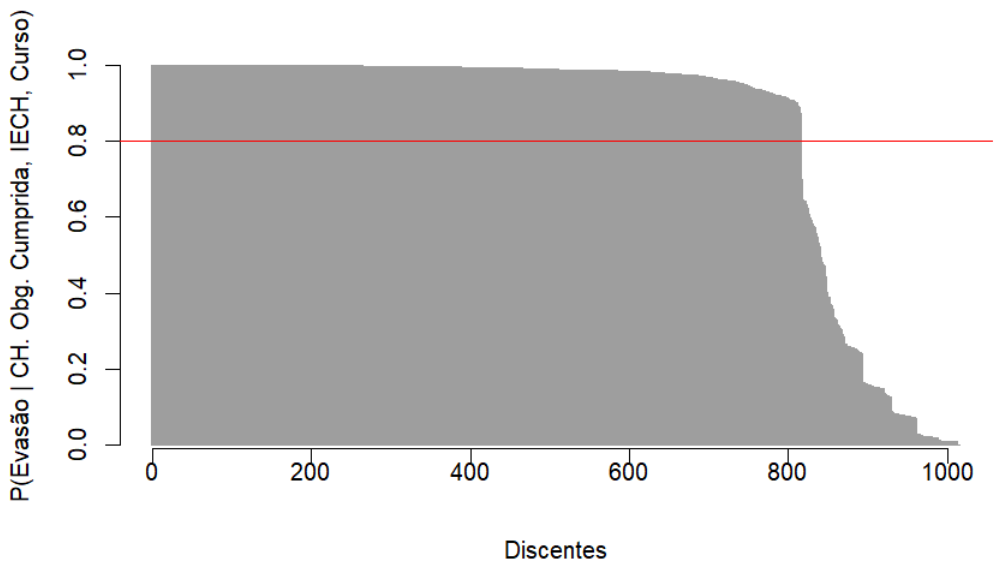
Dessa forma, a Figura 33 mostra a quantidade de currículos classificado em cada classe. Já a Figura 34 mostra a probabilidade de abandono de cada um dos currículos analisados.

Figura 33 – Distribuição dos valores previstos para a variável *Razão de saída*



É possível observar na Figura 33 que 82% dos estudantes foram classificados como em risco de evasão. Apesar de alto, esse número é próximo à ao percentual de evasão observados

Figura 34 – Probabilidade de evasão para cada currículo processado pelo modelo



nos dados atuais do departamento (Figura 27).

Entretanto, é preciso levar em consideração que o modelo não considera em suas previsões os alunos passíveis de realizar algum tipo de transferência entre cursos. Dessa forma, as previsões de evasão e conclusão também agregam os alunos que podem vir a realizar algum tipo de transferência.

Por sua vez, a Figura 34 mostra que as classificações realizadas de evasão realizadas ocorrem com um elevado grau de certeza pelo modelo. É possível observar que 90% dos currículos classificados como *ABANDONO* possuem uma probabilidade associada de mais de 0.9, sendo esse outro indicativo de superespecialização do modelo, devido aos altos níveis de desbalanceamento dos dados utilizados.

9
Conclusão

Neste trabalho, foi proposto o desenvolvimento de um modelo computacional automatizado para prever a evasão escolar de estudantes no Departamento de Computação (DCOMP) da Universidade Federal de Sergipe e analisar as relações entre os fatores estudados. A partir da análise desses dados, foi possível desenvolver um modelo Baseado em Redes Bayesianas Multinível, que alcançou uma acurácia de 88% na identificação de alunos em risco de evasão.

Seguindo a metodologia CRISP, adotada para o desenvolvimento do trabalho, pode-se afirmar que o modelo proposto foi aceito, uma vez que atendeu a todos os requisitos previamente estipulados no Quadro 8. Assim, é possível concluir que o modelo proposto apresenta uma solução viável e eficiente para o problema da evasão escolar no DCOMP, permitindo a identificação precoce de estudantes em situação de risco e a adoção de medidas preventivas e de suporte adequadas.

Ademais, outro objetivo deste trabalho foi realizar a análise das relações fornecidas entre pela topologia da rede Bayesiana resultante do treinamento. Assim, a análise permitiu identificar que a variável *Naturalidade* possui pouco impacto sobre a probabilidade de evasão estudantil e, portanto, pôde ser removida do modelo. Em contrapartida, as variáveis relacionadas à regularidade do aluno no curso, como *Nº de Trancamentos*, *CH. Obrigatória Cumprida* e *IECH*, demonstraram ser mais influentes no índice de evasão escolar do que aquelas que medem diretamente o desempenho acadêmico em componentes curriculares, como o MC.

Adicionalmente, o modelo treinado revelou que a idade do estudante é um fator importante na previsão da probabilidade de evasão, evidenciando a necessidade de ações específicas por parte da administração universitária para atender às diferentes faixas etárias e reduzir os índices de evasão.

Assim, os resultados obtidos através deste estudo apresentam importantes implicações para a tomada de decisões e o planejamento estratégico da universidade, fornecendo *insights* valiosos para a adoção de medidas preventivas e de suporte aos estudantes em risco de evasão.

Entretanto, apesar do desempenho satisfatório do modelo desenvolvido, é importante salientar a possibilidade de ocorrência de *overfitting*, devido ao desbalanceamento do *dataset* utilizado no treinamento. Para minimizar essa possibilidade e garantir a robustez do modelo, é recomendável que a rede Bayesiana seja treinada com um *dataset* maior e mais diversificado, contendo informações de alunos de outros departamentos, além de considerar uma abordagem mais rigorosa para lidar com o desbalanceamento de classes. Ademais, é válido mencionar que a implantação do modelo em um ambiente real também requer a avaliação contínua de seu desempenho e a realização de ajustes necessários, para garantir sua eficácia na previsão da evasão estudantil.

Além do modelo computacional automatizado para prever a evasão escolar de estudantes no DCOMP, este trabalho também objetivou a criação de um *dataset* contendo informações sobre a performance a de alunos da graduação. Assim, o trabalho resultou no desenvolvimento de um *dataset* com mais de 135 mil observações de componentes curriculares e 3925 históricos escolares de 3589 estudantes do departamento, juntamente com as ferramentas desenvolvidas para manter esses dados atualizados. O *dataset* possui potencial para ser utilizado em futuras pesquisas e estudos que visem usar a performance acadêmica de alunos de graduação para compreender e sugerir melhorias no sistema educacional. No entanto, é importante ressaltar a necessidade de se preservar a privacidade dos envolvidos e garantir a segurança dos dados. Portanto, é fundamental que medidas de proteção de dados sejam aplicadas durante o seu uso e que apenas pesquisadores devidamente autorizados possam ter acesso a essas informações.

9.1 Trabalhos futuros

Este trabalho apresenta uma análise estática da performance estudantil em relação ao tempo. Entretanto, uma possível melhoria seria a realização de uma análise com viés temporal, que consiga mensurar a evolução de um determinado estudante ao longo do tempo e levar isso em conta para realizar previsões mais precisas. O uso de Redes Bayesianas Dinâmicas pode auxiliar neste processo. O trabalho de Lappenschaar et al. (2013b), que define as Redes Bayesianas Multiníveis Temporais, também pode ser utilizado durante a exploração desse viés.

Outro fator que pode ser explorado em trabalhos futuros é a criação de um modelo ou método de aprendizado que utilize também os dados individuais das componentes curriculares cursadas, de maneira que sua aplicação continue sendo possível em diferentes departamentos e centros da universidade.

Há também espaço para melhorias no modelo em si. O trabalho em conjunto com entidades da administração estudantil da universidade para adicionar conhecimento especialista prévio ao modelo pode trazer melhorias significativas na acurácia do modelo e nos *insigths* que o mesmo venha fornecer.

Além disso, investigar a possível superespecialização do modelo e suas causas de forma

mais aprofundada é crucial para construir uma rede confiável e estável. A aplicação de técnicas de subamostragem e superamostragem pode ajudar a reduzir o desequilíbrio no conjunto de dados, o que possibilita um treinamento mais completo e preciso do modelo.

Além disso, o *dataset* construído neste trabalho pode ser utilizado para estudos em diversas áreas, por isso a democratização do acesso a esses dados é de extrema importância. Para isso, é necessário realizar estudos que construam as bases para tal, respeitando as normas da Lei Geral de Proteção de Dados, vigente no Brasil, a fim de garantir a proteção de estudantes e pesquisadores envolvidos em trabalhos posteriores.

Por fim, é importante ressaltar que o conjunto de dados produzido neste trabalho não inclui informações sobre as notas dos estudantes no Sistema de Seleção Unificada (SISU), utilizado como critério de ingresso nas universidades(e cujas relações já estão modeladas no conjunto de dados, conforme a Figura 24). No entanto, é possível utilizar o utilitário de extração de dados dos históricos escolares para adicionar essas informações ao conjunto de dados. Para isso, será necessário adquirir os dados do SISU e configurar o formato das informações para que o processo seja executado corretamente. A inclusão dos dados de desempenho no SISU pode ser valiosa para identificar casos de evasão estudantil no início da graduação, fornecendo informações sobre o desempenho acadêmico dos alunos antes mesmo do início do curso.

Referências

AMERI, S. et al. Survival analysis based framework for early prediction of student dropouts. *International Conference on Information and Knowledge Management, Proceedings*, Association for Computing Machinery, v. 24-28-October-2016, p. 903–912, 10 2016. Disponível em: <http://dx.doi.org/10.1145/2983323.2983351>. Citado na página 25.

BOUCKAERT, R. R. Choosing between two learning algorithms based on calibrated tests. In: *ICML*. [S.l.: s.n.], 2003. v. 3, n. 1. Citado na página 49.

BOVO, J. M. *Cobrança de Mensalidade Não É a solução para o financiamento da universidade pública*. 2022. Disponível em: <https://jornal.unesp.br/2022/06/08/cobranca-de-mensalidade-nao-e-a-solucao-para-o-financiamento-da-universidade-publica/>. Citado na página 16.

CHAPMAN, P. *CRISP-DM 1.0: Step-by-step data mining guide*. 1. ed. [S.l.]: SPSS, 2000. Citado na página 67.

CHEN, Y.; JOHRI, A.; RANGWALA, H. Running out of stem: A comparative study across stem majors of college students at-risk of dropping out early. *Proceedings of the 8th International Conference on Learning Analytics and Knowledge*, ACM, p. 10, 2018. Disponível em: <https://doi.org/10.1145/3170358.3170410>. Citado na página 24.

COSTA, S. F.; DINIZ, M. M. Application of logistic regression to predict the failure of students in subjects of a mathematics undergraduate course. *Education and Information Technologies*, 2022. ISSN 1573-7608. Disponível em: <https://doi.org/10.1007/s10639-022-11117-1>. Citado na página 23.

COUSSEMENT, K. et al. Predicting student dropout in subscription-based online learning environments: The beneficial impact of the logit leaf model. *Decision Support Systems*, North-Holland, v. 135, p. 113325, 8 2020. ISSN 0167-9236. Citado na página 24.

DALIPI, F.; IMRAN, A. S.; KASTRATI, Z. Mooc dropout prediction using machine learning techniques: Review and research challenges. *IEEE Global Engineering Education Conference, EDUCON*, IEEE Computer Society, v. 2018-April, p. 1007–1014, 5 2018. ISSN 21659567. Citado 2 vezes nas páginas 24 e 30.

DING, M. et al. Transfer learning using representation learning in massive open online courses. *Proceedings of the 9th International Conference on Learning Analytics & Knowledge*, ACM, 2019. Disponível em: <https://doi.org/10.1145/3303772.3303794>. Citado na página 24.

FRIEDMAN, N.; GOLDSZMIDT, M.; WYNER, A. Data Analysis with Bayesian Networks: A Bootstrap Approach. In: *UAI '99: Proceedings of the 15th Annual Conference on Uncertainty in Artificial Intelligence*. [S.l.]: Morgan Kaufmann, 1999. p. 196–205. Citado na página 88.

FRIEDMAN, N. et al. Using bayesian networks to analyze expression data. *Journal of computational biology*, Mary Ann Liebert, Inc., v. 6, n. 3, p. 301–22, 1999. Citado na página 59.

GAO, N. et al. n-gage. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, ACM PUB27 New York, NY, USA, v. 4, 9 2020. ISSN 24749567. Disponível em: <https://dl-acm-org.ez20.periodicos.capes.gov.br/doi/10.1145/3411813>. Citado na página 23.

GIL, A. C. *Como elaborar projetos de pesquisa*. 4. ed. [S.l.]: Atlas S.A, 2002. Citado na página 17.

GIL, P. D. et al. A data-driven approach to predict first-year students' academic success in higher education institutions. *Education and Information Technologies*, Springer, v. 26, p. 2165–2190, 3 2021. ISSN 15737608. Disponível em: <https://link-springer-com.ez20.periodicos.capes.gov.br/article/10.1007/s10639-020-10346-6>. Citado na página 23.

GUZMáN-CASTILLO, S. et al. Implementation of a predictive information system for university dropout prevention. *Procedia Computer Science*, v. 198, p. 566–571, 2022. ISSN 1877-0509. Disponível em: <https://www.sciencedirect.com/science/article/pii/S1877050921025266>. Citado na página 23.

HEGDE, V.; PRAGEETH, P. P. Higher education student dropout prediction and analysis through educational data mining. *Proceedings of the 2nd International Conference on Inventive Systems and Control, ICISC 2018*, Institute of Electrical and Electronics Engineers Inc., p. 694–699, 6 2018. Citado na página 24.

HOX, J. *Artificial Intelligence: A Modern Approach*. 3. ed. [S.l.]: Routledge, 2017. ISBN 9781315650982. Citado 2 vezes nas páginas 62 e 63.

HUSSEIN, A. S.; KHAN, H. A. Students' performance tracking in distributed open education using big data analytics. *ACM International Conference Proceeding Series*, Association for Computing Machinery, 3 2017. Disponível em: <http://dx.doi.org/10.1145/3018896.3018975>. Citado na página 24.

KANG, K.; WANG, S. Analyze and predict student dropout from online programs. *Proceedings of the 2nd International Conference on Compute and Data Analysis - ICCDA 2018*, ACM Press, 2018. Disponível em: <https://doi.org/10.1145/3193077.3193090>. Citado na página 24.

KARALAR, H.; KAPUCU, C.; GüRüLER, H. Predicting students at risk of academic failure using ensemble model during pandemic in a distance learning system. *International Journal of Educational Technology in Higher Education*, v. 18, p. 63, 2021. ISSN 2365-9440. Disponível em: <https://doi.org/10.1186/s41239-021-00300-y>. Citado na página 23.

KISS, B. et al. Predicting dropout using high school and first-semester academic achievement measures. *ICETA 2019 - 17th IEEE International Conference on Emerging eLearning Technologies and Applications, Proceedings*, Institute of Electrical and Electronics Engineers Inc., p. 383–389, 11 2019. Citado na página 24.

KOLMOGOROV, A. N. *Foundations of the theory of probability*. USA: New York: Chelsea Pub. Co., 1950. Citado na página 33.

KOSTOPOULOS, G.; KOTSIANTIS, S.; PINTELAS, P. Estimating student dropout in distance higher education using semi-supervised techniques. *ACM International Conference Proceeding Series*, Association for Computing Machinery, v. 01-03-October-2015, p. 38–43, 10 2015. Disponível em: <http://dx.doi.org/10.1145/2801948.2802013>. Citado na página 25.

LAPPENSCHAAR, M. et al. Multilevel bayesian networks for the analysis of hierarchical health care data. *Artificial Intelligence in Medicine*, v. 57, p. 171–183, 3 2013. ISSN 09333657. Citado 13 vezes nas páginas 31, 32, 37, 44, 58, 61, 63, 65, 66, 69, 72, 86 e 90.

LAPPENSCHAAR, M. et al. Multilevel temporal bayesian networks can model longitudinal change in multimorbidity. *Journal of Clinical Epidemiology*, v. 66, n. 12, p. 1405–1416, 2013. ISSN 0895-4356. Disponível em: <https://www.sciencedirect.com/science/article/pii/S0895435613002655>. Citado na página 98.

LAURITZEN, S. L.; SPIEGELHALTER, D. J. Local computations with probabilities on graphical structures and their application to expert systems. *Journal of the Royal Statistical Society. Series B (Methodological)*, [Royal Statistical Society, Wiley], v. 50, n. 2, p. 157–224, 1988. ISSN 00359246. Disponível em: <http://www.jstor.org/stable/2345762>. Citado na página 49.

LüDER, A. *Quase 3,5 Milhões de Alunos Evadiram de Universidades Privadas no Brasil em 2021*. 2022. Disponível em: <https://g1.globo.com/educacao/noticia/2022/01/02/quase-35-milhoes-de-alunos-evadiram-de-universidades-privadas-no-brasil-em-2021.ghtml>. Citado na página 16.

MABUNDA, J. G. K.; JADHAV, A.; AJOODHA, R. A review: Predicting student success at various levels of their learning journey in a science programme. *2021 IEEE International IOT, Electronics and Mechatronics Conference, IEMTRONICS 2021 - Proceedings*, Institute of Electrical and Electronics Engineers Inc., 4 2021. Citado na página 23.

MANRIQUE, R. et al. An analysis of student representation, representative features and classification algorithms to predict degree dropout. *ACM International Conference Proceeding Series*, Association for Computing Machinery, p. 401–410, 3 2019. Disponível em: <https://doi.org/10.1145/3303772.3303800>. Citado na página 24.

MARTINHO, V. R.; NUNES, C.; MINUSSI, C. R. An intelligent system for prediction of school dropout risk group in higher education classroom based on artificial neural networks. *Proceedings - International Conference on Tools with Artificial Intelligence, ICTAI*, p. 159–166, 2013. ISSN 10823409. Citado na página 25.

MAUA, D. D. *Probabilistic Graphical Models*. 2020. Disponível em: <https://www.ime.usp.br/~ddm/courses/mac6916>. Citado 8 vezes nas páginas 43, 46, 47, 49, 50, 53, 54 e 58.

MEDINA, E. C. et al. Predictive model to reduce the dropout rate of university students in perú: Bayesian networks vs. decision trees. *Iberian Conference on Information Systems and Technologies, CISTI*, IEEE Computer Society, v. 2020-June, 6 2020. ISSN 21660735. Citado na página 24.

MáRQUEZ-VERA, C. et al. Predicting student failure at school using genetic programming and different data mining approaches with high dimensional and imbalanced data. *Applied Intelligence*, Springer, v. 38, p. 315–330, 4 2013. ISSN 0924669X. Disponível em: <https://link-springer-com.ez20.periodicos.capes.gov.br/article/10.1007/s10489-012-0374-8>. Citado na página 25.

NAGY, M.; MOLONTAY, R. Predicting dropout in higher education based on secondary school performance. *INES 2018 - IEEE 22nd International Conference on Intelligent Engineering Systems, Proceedings*, Institute of Electrical and Electronics Engineers Inc., p. 000389–000394, 11 2018. Citado na página 24.

NCES. *What are the dropout rates of high school students?* 2020. Disponível em: <https://nces.ed.gov/fastfacts/display.asp?id=16>. Citado na página 16.

NEAPOLITAN, R. E. *Learning Bayesian Networks*. USA: Prentice-Hall, Inc., 2003. ISBN 0130125342. Citado 9 vezes nas páginas 33, 35, 36, 37, 38, 39, 40, 41 e 42.

NIST. *Advanced Encryption Standard (AES)*. 2001. Federal Information Processing Standards Publication 197. Accessado em 19 de março de 2023, de <https://csrc.nist.gov/publications/detail/fips/197/final>. Citado na página 80.

PEARL, J. *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 1988. ISBN 1558604790. Citado 4 vezes nas páginas 33, 38, 39 e 41.

PETERSEN, K. et al. Systematic mapping studies in software engineering. *Proceedings of the 12th International Conference on Evaluation and Assessment in Software Engineering*, v. 17, 06 2008. Citado 4 vezes nas páginas 18, 20, 25 e 26.

PéREZ, A. et al. Comparative analysis of prediction techniques to determine student dropout: Logistic regression vs decision trees. *Proceedings - International Conference of the Chilean Computer Science Society, SCCC*, IEEE Computer Society, v. 2018-November, 7 2018. ISSN 15224902. Citado na página 24.

REIS, E. *Create ready-to-learn Multilevel Bayesian Networks models*. 2023. Accessado em 19 de março de 2023, de <https://github.com/eduardo-fillipe/mbnlearn>. Citado 2 vezes nas páginas 71 e 85.

RUSSELL, S.; NORVIG, P. *Artificial Intelligence: A Modern Approach*. 3. ed. [S.l.]: Prentice Hall, 2010. Citado 11 vezes nas páginas 31, 32, 37, 39, 45, 46, 47, 50, 54, 56 e 57.

SANTOS, C. O. d.; PILATTI, L. A.; BONDARIK, R. Evasão no ensino superior brasileiro: conceito, mensuração, causas e consequências. *Debates em Educação*, v. 14, n. 35, p. 294–314, ago. 2022. Disponível em: <https://www.seer.ufal.br/index.php/debateseducacao/article/view/12555>. Citado na página 16.

SANTOS, G. A. et al. Evolvedtree: Analyzing student dropout in universities. *International Conference on Systems, Signals, and Image Processing*, IEEE Computer Society, v. 2020-July, p. 173–178, 7 2020. ISSN 21578702. Citado na página 23.

SCUTARI, M. *Bayesian Networks with examples in R*. 1. ed. [S.l.]: CRC Press, 2015. Citado 13 vezes nas páginas 43, 45, 46, 47, 49, 51, 52, 53, 56, 58, 59, 60 e 61.

SHEARER, C. The crisp-dm model:the new blueprint fordata mining. *Journal of Data Warehousing*, 2000. Citado 3 vezes nas páginas 18, 19 e 67.

SHIAU, Y. University dropout prevention through the application of big data. *ACM International Conference Proceeding Series*, Association for Computing Machinery, p. 1–7, 8 2020. Citado na página 23.

SILVA, P. M. D. et al. Ensemble regression models applied to dropout in higher education. *Proceedings - 2019 Brazilian Conference on Intelligent Systems, BRACIS 2019*, Institute of Electrical and Electronics Engineers Inc., p. 120–125, 10 2019. Citado na página 24.

SPIRTES, P.; MEEK, C. Learning bayesian networks with discrete variables from data. In: . [S.l.: s.n.], 1995. p. 294–299. Citado na página 50.

STALLINGS, W. *Cryptography and network security: principles and practice*. [S.l.]: Pearson Education, 2017. Citado na página 80.

SUAPRAE, P.; NILSOOK, P.; WANNAPIROON, P. System framework of intelligent consulting systems with intellectual technology. *ACM International Conference Proceeding Series*, Association for Computing Machinery, v. 21, p. 31–36, 2021. Disponível em: <https://doi.org/10.1145/3479162.3479167>. Citado 2 vezes nas páginas 23 e 67.

TENPIPAT, W.; AKKARAJITSAKUL, K. Student dropout prediction: A kmutt case study. *2020 1st International Conference on Big Data Analytics and Practices, IBDAP 2020*, Institute of Electrical and Electronics Engineers Inc., 9 2020. Citado na página 23.

TSAMARDINOS, I.; BROWN, L.; ALIFERIS, C. F. The max-min hill-climbing bayesian network structure learning algorithm. *Machine Learning*, v. 65, p. 31–78, 2006. Citado na página 59.

VERDUGO, J. V. et al. Faired: A systematic fairness analysis approach applied in a higher educational context. In: . Association for Computing Machinery, 2022. p. 271–281. ISBN 9781450395731. Disponível em: <https://doi-org.ez20.periodicos.capes.gov.br/10.1145/3506860.3506902>. Citado na página 23.

VERMA, T.; PEARL, J. Equivalence and synthesis of causal models. In: *Proceedings of the Sixth Annual Conference on Uncertainty in Artificial Intelligence*. USA: Elsevier Science Inc., 1990. (UAI '90), p. 255–270. ISBN 0444892648. Citado 2 vezes nas páginas 48 e 49.

VILORIA, A.; LEZAMA, O. B. P. Mixture structural equation models for classifying university student dropout in latin america. *Procedia Computer Science*, Elsevier, v. 160, p. 629–634, 1 2019. ISSN 1877-0509. Citado na página 24.

VILORIA, A.; LEZAMA, O. B. P.; VARELA, N. Bayesian classifier applied to higher education dropout. *Procedia Computer Science*, Elsevier, v. 160, p. 573–577, 1 2019. ISSN 1877-0509. Citado na página 24.

YU, R.; LEE, H.; KIZILCEC, R. F. Should college dropout prediction models include protected attributes? *Proceedings of the Eighth ACM Conference on Learning @ Scale*, ACM, 2020. Disponível em: <http://dx.doi.org/10.1145/3430895.3460139>. Citado na página 23.

ZHANG, W.; WANG, Y.; WANG, S. Predicting academic performance using tree-based machine learning models: A case study of bachelor students in an engineering department in china. *Education and Information Technologies*, 2022. ISSN 1573-7608. Disponível em: <https://doi.org/10.1007/s10639-022-11170-w>. Citado 2 vezes nas páginas 23 e 30.

ZHOU, Q. et al. Predicting high-risk students using internet access logs. *Knowledge and Information Systems*, Springer London, v. 55, p. 393–413, 5 2018. ISSN 02193116. Disponível em: <https://link-springer-com.ez20.periodicos.capes.gov.br/article/10.1007/s10115-017-1086-5>. Citado 2 vezes nas páginas 24 e 30.

Anexos

ANEXO A – Demonstrações

A.1 Regra da cadeia

Demonstração. Pela Equação 3.2, sabe-se que a equação é verdadeira para o caso em que X possui dois elementos:

$$P(x_1 \cap x_2) = P(x_1|x_2)P(x_2)$$

Dessa forma, suponha por indução que a Equação 3.3 é valida para n = k-1:

$$P(x_1 \cap x_2 \cap ... \cap x_{k-1}) = P(x_{k-1}|x_{k-2} \cap x_{k-3} \cap ... \cap x_1)P(x_{k-2} \cap ... \cap x_1)$$

Assim, prova-se que a regra vale para n = k. Aplicando-se a definição de probabilidade condicional:

$$P(x_1 \cap x_2 \cap ... \cap x_k) = P(x_k|x_{k-1} \cap x_{k-2} \cap ... \cap x_1)P(x_{k-1} \cap x_{k-2} \cap ... \cap x_1)$$

Aplicando-se a hipótese de indução no último termo:

$$P(x_1 \cap x_2 \cap ... \cap x_k) = P(x_k|x_{k-1} \cap x_{k-2} \cap ... \cap x_1)P(x_{k-1}|x_{k-2} \cap x_{k-3} \cap ... \cap x_1)P(x_{k-2} \cap ... \cap x_1)$$

$$P(x_1 \cap x_2 \cap ... \cap x_k) = \prod_{k=1}^{n} P(x_k|\bigcap_{j=1}^{k-1} x_j)$$

□

A.2 Relação de independência condicional

$$E \perp\!\!\!\perp F|G \Rightarrow P(E \cap F|G) = P(E|G)P(F|G)$$

Demonstração.

$$E \perp\!\!\!\perp F|G \Rightarrow P(E|F \cap G) = P(E|G)$$

$$\Rightarrow \frac{P(E \cap F \cap G)}{P(F \cap G)} = \frac{P(E \cap G)}{P(G)}$$

$$\Rightarrow P(E \cap F \cap G) = \frac{P(E \cap G)P(F \cap G)}{P(G)}$$

$$\Rightarrow \frac{P(E \cap F \cap G)}{P(G)} = \frac{\frac{P(E \cap G)P(F \cap G)}{P(G)}}{P(G)}$$

$$\Rightarrow \frac{P(E \cap F \cap G)}{P(G)} = \frac{P(E \cap G)}{P(G)} \frac{P(F \cap G)}{P(G)}$$

$$\Rightarrow P(E \cap F|G) = P(E|G)P(F|G)$$

□

A.3 Teorema de Bayes

$$P(b|a) = \frac{P(a|b)P(b)}{P(a)}$$

Demonstração. Da definição de probabilidade condicional, tem-se que:

$$\boldsymbol{P(a,b) = P(a|b)P(b)} \text{ e } \boldsymbol{P(b,a) = P(b|a)P(a)}$$

Como:

$$P(a,b) = P(b,a)$$

Então:

$$P(b|a)P(a) = P(a|b)P(b)$$
$$P(b|a) = \frac{P(a|b)P(b)}{P(a)}$$

□

A.4 Distribuição de probabilidades numa rede Bayesiana

$$P(\bigcap_{v \in V} v) = \prod_{v \in V} P(v|PA_v)$$

Demonstração. Ordene V de forma que os nós pais precedam seus descendentes. Assim, se $Y \in V$ é pai de $X \in V$ então Y precede X na ordem. Seja $[X_1, X_2, ..., X_n]$ o resultado da ordenação e para um dado conjunto de valores $\{x_1, x_2, ..., x_n\}$, pa_i será o conjunto dos valores contidos em PA_{X_i}. Assim, quer-se provar que:

$$P(x_n, x_{n-1}, ..., x_1) = P(x_n|pa_n)P(x_{n-1}|pa_{n-1})...P(x_1|pa_1)$$

A prova segue por indução no número de vértices de G.

Caso base: G possui 1 vértice, então pa_1 é vazio, assim:

$$P(x_1) = P(x_1|pa_1) = P(x_1)$$

Hipótese de Indução: Supõe-se que a Equação 3.8 é verdade para um sub-conjunto de tamanho i, então:

$$P(x_i, x_{i-1}, ..., x_1) = P(x_i|pa_n)P(x_{i-1}|pa_{i-1})...P(x_1|pa_1)$$

Caso indutivo: Prova-se que o teorema é valido para:

$$P(x_{i+1}, x_i, x_{i-1}, ..., x_1) = P(x_{i+1}|pa_{i+1})P(x_i|pa_i)P(x_{i-1}|pa_{i-1})...P(x_1|pa_1)$$

Existem dois casos:

Caso 1: $P(x_i, x_{i-1}, ..., x_1) = 0$

Nesse caso, existe um k $(1 \leq k \leq i)$ tal que $P(x_k|pa_k) = 0$, então o produto $P(x_{i+1}, x_i, x_{i-1}, ..., x_1) = 0$ é verdadeiro.

Caso 2: $P(x_i, x_{i-1}, ..., x_1) \neq 0$, assim aplicando a definição de probabilidade condicional:

$$P(x_{i+1}, x_i, x_{i-1}, ..., x_1) = P(x_{i+1}|x_i, ..., x_1)P(x_i, ..., x_1)$$

Agora, como (G, P) satisfazem a condição de Markov um com o outro, $PA_{X_i} \subseteq ND_{X_i}$ e, graças à ordenação inicial, $X_i, ..., X_1$ são todos os vértices não descendentes de X_{i+1}, então:

$$P(x_{i+1}, x_i, x_{i-1}, ..., x_1) = P(x_{i+1}|pa_{i+1})P(x_i, ..., x_1)$$

Daí, aplicando-se a hipótese de indução no último termo:

$$P(x_{i+1}, x_i, x_{i-1}, ..., x_1) = P(x_{i+1}|pa_{i+1})P(x_i|pa_i)P(x_{i-1}|pa_{i-1})...P(x_1|pa_1)$$

□

ANEXO B – Experimento com dados sintéticos

B.1 Blacklist utilizada no treinamento

Vértice de Origem	Vértice de Destino
I2	D1
I2	D2
I2	D3
L1	D1
L1	D2
L1	D3
L2	D1
L2	D2
L2	D3
D1	I1
D1	I2
D1	L1
D1	L2
D2	I1
D2	I2
D2	L1
D2	L2
D3	I1
D3	I2
D3	L1
D3	L2
I1	G
I2	G

Continua na próxima página

Tabela 15 – *Continuação*

Vértice de Origem	Vértice de Destino
L1	G
L2	G
G	I1
G	I2
G	L1
G	L2
I2	L1
L2	L1
L1	I1
L1	I2
L1	L2
I1	L2
L2	I1
L2	I2
I1	I2

B.2 White list utilizado no treinamento

Quadro 10 – Whitelist utilizada durante o experimento com dados sintéticos

Vértice de Origem	**Vértice de Destino**
I1	D1
I1	D2
I1	D3
I1	L1
I2	I1
I2	L2