UNIVERSIDADE FEDERAL FLUMINENSE

LUCAS DIRK GOMES FERREIRA

MODELAGEM DAS RELAÇÕES PROBABILÍSTICAS ENTRE VARIÁVEIS METEOROLÓGICAS E PRECIPITAÇÃO NO MUNICÍPIO DO RIO DE JANEIRO POR MEIO DE REDES BAYESIANAS

NITERÓI

2026

LUCAS DIRK GOMES FERREIRA

MODELAGEM DAS RELAÇÕES PROBABILÍSTICAS ENTRE VARIÁVEIS METEOROLÓGICAS E PRECIPITAÇÃO NO MUNICÍPIO DO RIO DE JANEIRO POR MEIO DE REDES BAYESIANAS

Dissertação de Mestrado apresentada ao Programa de Pós-Graduação em Computação da Universidade Federal Fluminense como requisito parcial para a obtenção do Grau de Mestre em Computação. Área de concentração: Ciência da Computação

Orientador:
MARIZA FERRO

NITERÓI

2026

Ficha catalográfica automática - SDC/BEE
Gerada com informações fornecidas pelo autor

F383m Ferreira, LUCAS DIRK GOMES
MODELAGEM DAS RELAÇÕES PROBABILÍSTICAS ENTRE VARIÁVEIS METEOROLÓGICAS E PRECIPITAÇÃO NO MUNICÍPIO DO RIO DE JANEIRO POR MEIO DE REDES BAYESIANAS / LUCAS DIRK GOMES Ferreira. - 2026.
136 f.: il.

Orientador: MARIZA FERRO.
Dissertação (mestrado)-Universidade Federal Fluminense, Instituto de Computação, Niterói, 2026.

1. REDES BAYESIANAS. 2. PRECIPITAÇÃO. 3. Modelagem Probabilística. 4. Produção intelectual. I. FERRO, MARIZA, orientador. II. Universidade Federal Fluminense. Instituto de Computação. III. Título.

CDD - XXX

Bibliotecário responsável: Debora do Nascimento - CRB7/6368

LUCAS DIRK GOMES FERREIRA

MODELAGEM DAS RELAÇÕES PROBABILÍSTICAS ENTRE VARIÁVEIS METEOROLÓGICAS E PRECIPITAÇÃO NO MUNICÍPIO DO RIO DE JANEIRO POR MEIO DE REDES BAYESIANAS

Dissertação de Mestrado apresentada ao Programa de Pós-Graduação em Computação da Universidade Federal Fluminense como requisito parcial para a obtenção do Grau de Mestre em Computação. Área de concentração: Ciência da Computação

Aprovada em Março de 2026.

BANCA EXAMINADORA

Profa. Mariza Ferro - Orientador, UFF

Profa. Aline Marins Paes Carvalho, UFF

Profa. Fernanda Cerqueira Vasconcellos, UFRJ

Prof. Thiago Malheiros Porcino, LNCC

Niterói

2026

Dedicatória(s): Ao meu pai, que desde cedo me ensinou o valor do estudo.

"The best thing about being a statistician is that you get to play in everyone's backyard"
— John Tukey

Agradecimentos

Dedico este trabalho às pessoas que, assim como eu, vêm de contextos periféricos e enfrentam grandes desafios para alcançar o ensino superior e a pesquisa. Reconheço também o papel dos contribuintes que sustentam a educação pública no país e tornam possível a existência da ciência e das universidades.

À Universidade Federal Fluminense, expresso meu apreço pelo ambiente acadêmico, pelos recursos disponibilizados e pelo suporte institucional que viabilizaram a realização deste mestrado. Minha orientadora, professora doutora Mariza Ferro, merece especial menção pela orientação atenta, pela dedicação e pela disponibilidade ao longo de todo o processo. À professora doutora Fernanda Cerqueira Vasconcellos, sou igualmente grato pelas contribuições e sugestões que enriqueceram a pesquisa.

À minha noiva, Marcelle Gomes, deixo meu reconhecimento pelo apoio constante, pela compreensão nas renúncias e pela parceria incondicional que me acompanhou até aqui. Estendo minha gratidão ao meu pai, cuja influência foi decisiva na minha relação com os estudos.

Durante o percurso, tive a felicidade de conviver com colegas que se tornaram amigos. Miguel Domingos Brito e Matheus Veloso da Silva contribuíram para tornar essa jornada mais leve e bem-humorada. Por fim, agradeço aos professores doutores Aline Marins Paes Carvalho, Fernanda Cerqueira Vasconcellos, Thiago Malheiros Porcino, Taiane Coelho Ramos e Fabrício Polifke da Silva, que gentilmente aceitaram integrar a banca avaliadora desta dissertação.

Resumo

Redes Bayesianas (RBs) fornecem uma estrutura probabilística interpretável para a modelagem da incerteza em sistemas ambientais complexos. Este estudo investiga se a modelagem por Rede Bayesiana (RB) é capaz de representar, de forma interpretável, as relações probabilísticas entre variáveis meteorológicas e a precipitação no Rio de Janeiro, utilizando dados horários de estações telemétricas (2002–2024). Os modelos candidatos foram avaliados quanto à interpretabilidade e ao desempenho preditivo. Restrições estruturais baseadas em conhecimento meteorológico foram incorporadas ao aprendizado da rede, de modo a favorecer coerência temporal e plausibilidade física. Entre os modelos avaliados o algoritmo *Hill-Climbing* (HC) com o critério K2 e AIC (*Akaike Information Criterion*) apresentaram os melhores resultados sob diferentes métricas preditivas. No entanto, para fins interpretativos, optou-se pelo algoritmo HC com o critério BIC (*Bayesian Information Criterion*), por apresentar maior parcimônia e desempenho global comparável. A estrutura aprendida indicou que a ocorrência de chuva depende principalmente de condições atmosféricas de curto prazo, especialmente alta umidade, baixa radiação solar e velocidades moderadas do vento com defasagem de uma hora. A validação confirmou resultados consistentes na validação cruzada e padrões fisicamente coerentes associados à convecção e ao transporte de umidade. Ao representar explicitamente dependências condicionais, a RB proposta melhora a interpretabilidade em relação a métodos do tipo "caixa-preta" e oferece suporte ao raciocínio probabilístico para avaliação de precipitação em curto prazo. Os resultados destacam o potencial das RBs como ferramentas transparentes e fisicamente consistentes para modelagem meteorológica e apoio à decisão ambiental. Além disso, no contexto experimental deste trabalho, os procedimentos de aprendizado estrutural das RBs apresentaram baixo consumo energético estimado e reduzida demanda computacional durante o treinamento, com tempos de execução entre 2354 e 3030 segundos por modelo e consumo aproximado de 0,05 kWh por execução.

Palavras-chave: Redes Bayesianas, Modelagem Probabilística, Previsão de Precipitação, Aprendizado de Estrutura, Dados Meteorológicos, Interpretabilidade.

Abstract

Bayesian Networks (BNs) provide an interpretable probabilistic framework for modeling uncertainty in complex environmental systems. This study investigates whether Bayesian Network (BN) modeling can interpretably represent the probabilistic relationships between meteorological variables and precipitation in Rio de Janeiro, Brazil, using hourly data from telemetric stations (2002–2024).Candidate models were evaluated in terms of interpretability and predictive performance. Structural constraints informed by meteorological knowledge were incorporated into the network learning process to promote temporal coherence and physical plausibility. Among the evaluated models, the *Hill-Climbing* (HC) algorithm with the K2 and AIC (*Akaike Information Criterion*) scoring criteria achieved the best results across different predictive metrics. However, for interpretative purposes, the HC algorithm with the Bayesian Information Criterion (BIC) was selected due to its more parsimonious structure, which aligns more closely with physical reasoning. The learned structure indicated that rainfall occurrence mainly depends on short-term atmospheric conditions, particularly high humidity, low solar radiation, and moderate wind speeds with a one-hour lag. Validation confirmed consistent cross-validation results and physically coherent patterns associated with convection and moisture transport. By explicitly representing conditional dependencies, the proposed BN improves interpretability over black-box methods and supports probabilistic reasoning for short-term precipitation assessment. The findings highlight the potential of BNs as transparent and physically consistent tools for weather modeling and environmental decision support. Furthermore, within the experimental setting adopted in this study, the BNs structure-learning procedures showed low estimated energy consumption and reduced computational demand during training, with execution times ranging from 2354 to 3030 seconds per model and an approximate consumption of 0.05 kWh per run.

Keywords: Bayesian Networks, Probabilistic Modeling, Rainfall Prediction, Structure Learning, Meteorological Data, Interpretability.

Lista de Figuras

Lista de Tabelas

Sumário

1 Introdução

A ocorrência de precipitação constitui um dos fenômenos meteorológicos mais relevantes para o planejamento urbano, a gestão de recursos hídricos e a prevenção de desastres naturais. No Sudeste do Brasil, em particular no município do Rio de Janeiro, a chuva exerce papel decisivo na dinâmica climática local. A variabilidade pluviométrica na região resulta da interação entre fatores atmosféricos e características geográficas específicas, incluindo a proximidade com o oceano e uma topografia heterogênea formada por maciços montanhosos e áreas costeiras. Essa configuração espacial produz elevada variabilidade espaço temporal e impõe desafios à representação adequada dos processos meteorológicos em modelos preditivos (SILVA; SILVA; SILVA, 2022).

A complexidade climática local é intensificada por fatores urbanos e sociais. O crescimento populacional, a ocupação desordenada do território e alterações no uso do solo ampliam a vulnerabilidade das cidades brasileiras a eventos meteorológicos extremos (SELUCHI; BEU; ANDRADE, 2016). A alta suscetibilidade das cidades brasileiras aos acidentes naturais está associada à incapacidade histórica de prover moradia adequada para as camadas populacionais mais baixas e promover um ordenamento territorial que imponha o interesse social sobre o interesse privado dos proprietários de terras (CARVALHO; GALVÃO, 2016).

No município do Rio de Janeiro, a ocupação de encostas e a remoção da cobertura vegetal contribuíram para a formação de áreas suscetíveis a deslizamentos de terra (COSTA; SILVA CONCEIÇÃO; OLIVEIRA AMANTE, 2018) fazendo com que eventos de precipitação frequentemente estejam associados a enchentes e movimentos de massa, configurando importantes riscos socioambientais. Além dos fatores urbanos, a dinâmica atmosférica local é influenciada por elementos geográficos e oceânicos que modulam a ocorrência de precipitação. A presença dos maciços da Tijuca, Pedra Branca e Mendanha, associada à proximidade com o Oceano Atlântico e às influências das baías de Guanabara e Sepetiba, contribuem para a formação de padrões atmosféricos complexos que favorecem episódios de chuva intensa (DERECZYNSKI; SILVA; MARENGO, 2013). Como consequência,

compreender a dinâmica das chuvas e os fatores atmosféricos associados à sua ocorrência torna-se fundamental para o monitoramento de eventos pluviométricos.

Mudanças observadas nos padrões climáticos globais também têm sido associadas a alterações na frequência e na intensidade de fenômenos meteorológicos em diversas regiões do planeta (HERRING et al., 2021). Em consonância, o Painel Intergovernamental sobre Mudanças Climáticas (*Intergovernmental Panel on Climate Change*, IPCC) também sintetiza esse quadro em seu relatório de avaliação (LEE et al., 2024). Esse cenário reforça a necessidade de aprofundar o entendimento sobre os mecanismos que condicionam a formação de eventos de precipitação em escala local. Embora exista ampla disponibilidade de dados meteorológicos observacionais, permanece o desafio de identificar quais variáveis atmosféricas são mais relevantes e de que forma suas interações contribuem para a ocorrência de chuva.

Nesse contexto, métodos de modelagem probabilística desempenham papel relevante na investigação de sistemas atmosféricos complexos. Avanços recentes em modelagem numérica e técnicas de aprendizado de máquina ampliaram a capacidade de previsão meteorológica (CHRISTENSEN; DELAUNAY, 2022). Entretanto, muitos modelos ainda apresentam limitações relacionadas à interpretabilidade e à adaptação a condições locais específicas. Abordagens baseadas exclusivamente em redes neurais frequentemente operam como "caixas-pretas", dificultando a compreensão dos mecanismos físicos subjacentes aos resultados obtidos (CHEN et al., 2025).

Embora existam diferentes abordagens de aprendizado de máquina consideradas explicáveis, essas técnicas apresentam limitações quando o objetivo central da análise é representar explicitamente relações probabilísticas condicionais entre variáveis. Neste trabalho, a escolha pelas RBs não se fundamenta apenas na capacidade preditiva, mas na possibilidade de representar dependências condicionais entre variáveis atmosféricas, incorporar conhecimento meteorológico por meio de restrições estruturais, quantificar incertezas e realizar inferências probabilísticas sob diferentes cenários de evidência. Assim, as RBs são particularmente adequadas ao objetivo desta dissertação, pois permitem investigar não apenas quais variáveis contribuem para a ocorrência de precipitação, mas também como essas variáveis se relacionam probabilisticamente em uma estrutura coerente com o fenômeno físico analisado.

As Redes Bayesianas (RBs) constituem uma alternativa apropriada para a modelagem probabilística de sistemas complexos caracterizados por incerteza e interdependência entre variáveis. Uma Rede Bayesiana (RB) representa um modelo gráfico probabilístico capaz de

descrever explicitamente relações de dependência condicional entre variáveis aleatórias por meio de grafos direcionados acíclicos. Essa representação permite integrar conhecimento de domínio e evidências observacionais, possibilitando a análise transparente das relações probabilísticas entre variáveis meteorológicas e a ocorrência de precipitação (KOLLER; FRIEDMAN, 2009).

A utilização de RBs em aplicações meteorológicas permite identificar condições atmosféricas associadas à ocorrência de chuva, quantificar a influência relativa de diferentes variáveis e realizar inferências probabilísticas (PUTRI; WIJAYANTO, 2021). Essa abordagem favorece a construção de modelos interpretáveis e fisicamente coerentes, contribuindo para uma compreensão mais aprofundada dos processos atmosféricos que atuam em escala local.

Historicamente, a aplicação de RBs em meteorologia tem demonstrado eficácia substancial na representação de incertezas em diferentes contextos. Estudos iniciais pioneiros exploraram o uso dessas redes na modelagem atmosférica e hidrometeorológica em regiões ibéricas e mediterrâneas, incluindo aplicações em previsão de precipitação e vento, fornecendo suporte probabilístico à tomada de decisão (COFINO et al., 2002; CANO; SORDO; GUTIÉRREZ, 2004; GARROTE; MOLINA; MEDIERO, 2007).

Pesquisas subsequentes investigaram a previsão de precipitação em diferentes regimes climáticos, como na Índia e na Indonésia, relatando desempenho preditivo promissor e destacando o papel central de variáveis relacionadas à umidade na ocorrência de chuva (SHARMA; GOYAL, 2016; DAS; CHANDA, 2020; PUTRI; WIJAYANTO, 2021). Além disso, estruturas baseadas em RBs aplicadas a séries climáticas espaço temporais têm demonstrado maior acurácia e redução de incerteza quando comparadas a abordagens estatísticas tradicionais baseadas em modelos não lineares e vetoriais (DAS; GHOSH, 2019).

No entanto, lacunas metodológicas significativas ainda persistem na literatura atual. Evidências comparativas indicam que a escolha da família de algoritmos de aprendizado de estrutura (*score-based*, *constraint-based* ou *hybrid-based*) afeta substancialmente tanto a topologia do grafo quanto o comportamento preditivo dos modelos (SCUTARI; GRAAFLAND; GUTIÉRREZ, 2019; KITSON et al., 2023). Embora estudos recentes enfatizem que técnicas de reamostragem e ajustes fora da amostra podem melhorar a robustez e reduzir a instabilidade nas estruturas aprendidas (CHOBTHAM; CONSTANTINOU, 2024; CARAVAGNA; RAMAZZOTTI, 2021; BARTH; SERRÃO; MACIEL, 2024), a maioria das aplicações meteorológicas avalia apenas um conjunto restrito de algoritmos

e carece de uma integração sistemática entre grandes bases de dados observacionais com baixa granularidade, restrições estruturais informadas por conhecimento de domínio e procedimentos de validação que considerem explicitamente a incerteza (COFINO et al., 2002; CANO; SORDO; GUTIÉRREZ, 2004; GARROTE; MOLINA; MEDIERO, 2007; SHARMA; GOYAL, 2016; DAS; GHOSH, 2019; DAS; CHANDA, 2020; PUTRI; WIJAYANTO, 2021).

Apesar dos avanços observados na literatura, ainda são escassos os estudos que integrem longas séries temporais observacionais com baixa granularidade, comparação sistemática entre diferentes famílias de algoritmos de aprendizado de estrutura e restrições estruturais baseadas em conhecimento físico em um único protocolo de modelagem aplicado ao contexto urbano tropical do Rio de Janeiro.

Diante desse contexto, o **objetivo geral** desta dissertação consiste em analisar se a modelagem por RBs é capaz de representar, de forma interpretável, as relações probabilísticas entre variáveis meteorológicas e a ocorrência de precipitação no município do Rio de Janeiro. Pretende-se, com isso, contribuir para a compreensão das interdependências atmosféricas associadas à formação da chuva e para a construção de uma estrutura probabilística que favoreça a interpretação do fenômeno. A escolha das RBs fundamenta-se em sua capacidade de representar distribuições de probabilidade conjunta de maneira compacta e interpretável, além de possibilitar inferências.

Dentre os **objetivos específicos** deste trabalho, destacam-se a investigação de diferentes algoritmos de aprendizado de estrutura de RB pertencentes às abordagens *Score-Based*, *Constraint-Based* e *Hybrid-Based*, a estimação das estruturas e dos parâmetros das RBs com a incorporação de restrições baseadas em conhecimento de domínio, a avaliação quantitativa do desempenho preditivo dos modelos e a realização de inferências probabilísticas para investigar condições atmosféricas associadas à ocorrência de precipitação.

As contribuições deste trabalho concentram-se em quatro aspectos principais. Primeiramente, foi construída uma estrutura gráfica baseada em RB capaz de representar explicitamente as dependências probabilísticas entre variáveis meteorológicas associadas à ocorrência de precipitação, permitindo visualizar e interpretar as relações condicionais entre os fatores atmosféricos. Em seguida, foram identificados padrões atmosféricos associados à ocorrência de chuva, destacando a combinação de elevada umidade relativa, baixos níveis de radiação solar e velocidades moderadas do vento em curtas defasagens temporais. Além disso, a inferência probabilística realizada na rede permitiu quantificar como diferentes combinações de estados das variáveis meteorológicas influenciam a proba-

bilidade de precipitação, contribuindo para a interpretação probabilística dos mecanismos associados à formação de chuva em ambientes urbanos tropicais. Por fim, o estudo fornece evidências empíricas sobre o desempenho de diferentes algoritmos de aprendizado estrutural de RBs aplicados a dados meteorológicos de longa duração do município do Rio de Janeiro, permitindo identificar abordagens mais adequadas para representar as relações probabilísticas associadas à precipitação.

Esta dissertação está organizada da seguinte forma: O Capítulo 1 apresenta o contexto do problema de pesquisa, a motivação do estudo e os objetivos desta dissertação. O Capítulo 2 apresenta a fundamentação teórica necessária para o desenvolvimento do trabalho, abordando os principais conceitos relacionados às RBs e aos métodos utilizados na modelagem. O Capítulo3 apresenta a revisão bibliográfica, discutindo aplicações de RBs em diferentes contextos e em problemas meteorológicos. O Capítulo 4 apresenta uma etapa de validação metodológica, na qual são reproduzidos e comparados modelos de RBs a partir de um estudo de referência. O Capítulo 5 descreve o design experimental e a metodologia adotada, incluindo os procedimentos de aquisição, pré-processamento dos dados e construção das redes. O Capítulo 6 apresenta a análise exploratória dos dados. O Capítulo 7 apresenta os resultados obtidos, a análise das inferências probabilísticas relacionadas à ocorrência de precipitação, consumo de energia e impacto ambiental. Finalmente, o Capítulo 8 apresenta as conclusões do trabalho e sugestões para pesquisas futuras.

2 Fundamentação Teórica

Este Capítulo apresenta a fundamentação teórica que embasa esta dissertação. Inicialmente, são introduzidos os conceitos centrais das RBs, abrangendo sua representação por grafos acíclicos direcionados, propriedades de independência condicional, d-separação, *Markov Blanket* e os principais métodos de aprendizado utilizados na construção dessas redes. Nesse contexto, são descritas as abordagens *Search and Score*, *Constraint-Based* e *Hybrid*, bem como os algoritmos e funções de pontuação avaliados neste trabalho. Posteriormente, são apresentadas as técnicas complementares empregadas na pesquisa, incluindo discretização, redução de dimensionalidade, reamostragem, balanceamento de classes e medidas de avaliação dos modelos. No contexto desta dissertação, esses conceitos são mobilizados para fundamentar a construção de RBs aplicadas à modelagem da precipitação a partir de variáveis meteorológicas observacionais, com ênfase na interpretabilidade estrutural, na coerência probabilística das dependências aprendidas e na avaliação comparativa dos algoritmos de aprendizado.

2.1 Redes Bayesianas

As Redes Bayesianas (*Bayesian Networks*), também chamadas de Redes Causais, Rede de Crença ou Gráficos de Dependência Probabilística (MACHADO et al., 2021) são um tipo de modelos gráficos probabilísticos que representam relações entre variáveis aleatórias e suas dependências condicionais por meio de um grafo acíclico direcionado (*Directed Acyclic Graph* - DAG) (NANDAR, 2009). Elas constituem uma forma natural para representação de informações condicionalmente independentes e possibilitam a representação compacta de uma tabela de probabilidades conjuntas (MARQUES; DUTRA, 2002). Além disso, essas redes ajudam a lidar com a grande quantidade de incertezas presentes no mundo (KOLLER; FRIEDMAN, 2009). Elas permitem avaliar o quanto a mudança de estado de uma variável altera a probabilidade de ocorrência de outra. Dessa forma, fornecem um sistema capaz de atribuir níveis de confiabilidade para todas as sentenças em sua base de

conhecimento e estabelecer relações entre elas (MARQUES; DUTRA, 2002).

Em matemática, um DAG é um grafo dirigido sem ciclo, ou seja, para qualquer vértice V, não há nenhuma ligação dirigida começando e acabando em V (BANG-JENSEN; GUTIN, 2008). Em outras palavras, não é possível seguir uma sequência de arestas que eventualmente retorne ao nó de origem, garantindo assim que a estrutura da rede permaneça acíclica. Como as RBs são grafos acíclicos direcionados, representam relações de dependência probabilística entre variáveis aleatórias sem a presença de ciclos na estrutura da rede. Além disso, os nós ou vértices representam as variáveis aleatórias, enquanto as relações de dependência condicional são representadas por meio de arestas ou arcos. Esses elementos são a representação da influência probabilística entre as variáveis.

A orientação das arestas indica a direção da influência probabilística, com o nó de origem influenciando o nó de destino, que é chamado de nó filho, ou seja, cada variável é um valor estocástico em função de seus pais. Além disso, as probabilidades condicionais associadas às arestas da rede são expressas por meio de tabelas de probabilidade condicional (*Conditional Probability Table* - CPT), no caso de variáveis discretas, ou por meio de distribuições de probabilidade condicional para variáveis contínuas (KOLLER; FRIEDMAN, 2009). Logo, a utilização de DAGs permite a representação gráfica de estruturas hierárquicas, o que facilita encontrar padrões e tendências durante a análise dessas estruturas (MARCHETE FILHO, 2013).

Uma estrutura do tipo DAG é denominada RB se, por definição, a distribuição conjunta de todas as variáveis puder ser fatorada como

$$P(X_1, X_2, \ldots, X_N) = \prod_{i=1}^{N} P(X_i \mid \text{Pais}(X_i)) \tag{2.1}$$

Na expressão acima, para qualquer variável X_i que não possua variáveis predecessoras (ou seja, não tenha pais no grafo), considera-se $P(X_i \mid \text{Pais}(X_i)) = P(X_i)$. Dessa forma, uma RB é especificada pelas distribuições de probabilidade condicionais associadas a cada variável em função de seus respectivos nós pais (CUNHA, 2009). Neste trabalho, as RBs serão utilizadas para representar as dependências probabilísticas entre variáveis meteorológicas observadas e a ocorrência de precipitação no município do Rio de Janeiro.

2.1.1 Redes Causais e Não Causais

Uma RB é denominada causal quando suas arestas são interpretadas como representações de relações de causa e efeito entre as variáveis. Por outro lado, RBs não causais não assumem relações de causalidade, limitando-se a descrever padrões de dependência probabilística (HECKERMAN, 2013).

Por exemplo, considere uma RB causal representada por *Chuva* $\rightarrow$ *Grama Molhada*, em que os parâmetros são conhecidos. Neste caso, a seta indica que a ocorrência de chuva é a causa da grama ficar molhada, e a probabilidade da grama estar molhada depende da chuva. Logo, ao observar que a grama está molhada, podemos inferir algo sobre a probabilidade de ter ocorrido chuva (CUNHA, 2009).

Por outro lado, podemos construir uma RB não causal, invertendo a direção da seta, Grama Molhada $\rightarrow$ Chuva. Neste cenário, a dependência entre as variáveis permanece, mas a interpretação muda. Ou seja, observar o estado da grama tem alguma informação sobre a probabilidade de ter ocorrido chuva, mas não estamos afirmando que a grama molhada causa chuva (CUNHA, 2009).

(PEARL et al., 2000) propôs uma formalização da causalidade em RBs, destacando que relações causais possuem caráter ontológico, enquanto dependências probabilísticas refletem apenas o conhecimento do observador. Para diferenciar observação de intervenção, o autor introduziu o operador $do(X = x)$, que fixa o valor de uma variável e remove as arestas de seus pais, resultando em um grafo mutilado. Uma RB é considerada causal se, após intervenções, o grafo resultante continuar válido, preservando a consistência probabilística e as independências condicionais. Assim, RBs causais permitem não apenas descrever dependências, mas também prever efeitos de intervenções e distinguir correlação de causalidade.

2.1.2 Variáveis Independentes e Condicionalmente Independentes

O entendimento sobre a independência entre variáveis desempenha um papel crucial na aplicação e construção de RBs, pois permite uma modelagem mais precisa das relações entre as variáveis e uma interpretação mais robusta dos resultados obtidos.

Diz-se que dois eventos são independentes quando a ocorrência de um deles não altera a probabilidade de ocorrência do outro (MORETTIN; BUSSAB, 2017). Ou seja, dados dois eventos independentes, A e B, a probabilidade de ocorrência de ambos é definida por

$$P(A \cap B) = P(A) \times P(B) \tag{2.2}$$

Embora a independência seja uma propriedade útil, nem sempre são observados dois eventos independentes. Uma situação mais comum é quando dois eventos são independentes, dado um evento adicional (KOLLER; FRIEDMAN, 2009). Ou seja, eventos condicionalmente independentes. Diz-se que o evento A é condicionalmente independente do evento B dado o evento C sob a distribuição P quando

$$P \models (A \perp B \mid C), \quad \text{se } P(A \mid B, C) = P(A \mid C), \tag{2.3}$$

No contexto desta dissertação, a noção de independência condicional é central para a aprendizagem estrutural das RBs, pois a presença ou ausência de dependência entre variáveis meteorológicas, dado um conjunto de condicionamento, influencia diretamente a identificação das conexões da rede. Esse conceito é particularmente relevante para compreender por que determinadas variáveis podem deixar de se conectar diretamente ao nó de precipitação após o condicionamento em outras variáveis atmosféricas.

2.1.3 Teorema de Bayes

Uma das relações mais importantes em probabilidade condicional é dada pelo teorema de Bayes. Este resultado descreve a probabilidade de um evento ocorrer, dado que um outro evento foi observado. Partindo da probabilidade inicial $P(A)$, a probabilidade a posteriori $P(A \mid B)$ é obtida ao considerar a informação de que B ocorreu ou a suposição de que B venha a ocorrer (MORETTIN; BUSSAB, 2017). Essa atualização probabilística é fundamental para analisar eventos dependentes, onde a ocorrência de um evento A influencia a probabilidade de outro evento B. Logo, o teorema de Bayes é um resultado matemático que propõe caracterizar a aprendizagem com a experiência, isto é, a modificação da atitude inicial em relação aos "Antecedentes", "Causas", "Hipóteses" ou "Estados" (LOPES, 2007). Esta probabilidade é determinada pela seguinte expressão:

$$P(A|B) = \frac{P(B|A) \cdot P(A)}{P(B)} \tag{2.4}$$

Onde $P(A)$ e $P(B)$ representam as probabilidades a priori dos eventos A e B; $P(A \mid B)$ é chamada de probabilidade a posteriori de A condicionada a B; $P(B \mid A)$ representa a probabilidade condicional de B dado A.

2.1.4 Regra da Cadeia para Redes Bayesianas

Na Figura 1, é apresentado um exemplo de RB, retirada do livro (KOLLER; FRIEDMAN, 2009), onde são observadas as relações probabilísticas entre variáveis aleatórias.

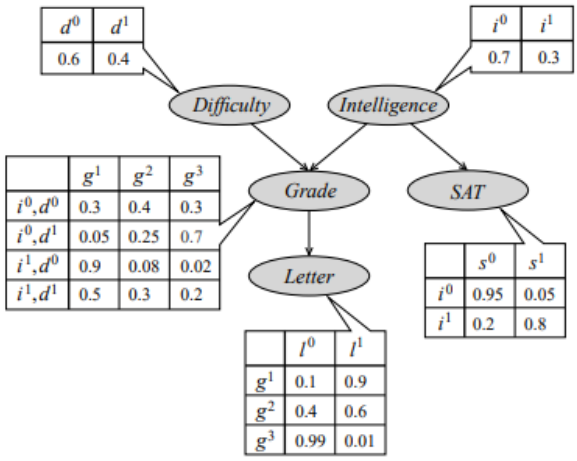

Figura 1: Exemplo de uma Rede Bayesiana (KOLLER; FRIEDMAN, 2009).

A variável *Grade* denota as notas de um aluno em uma determinada matéria, com possíveis estados correspondentes às notas A, B e C, representadas por g^1, g^2 e g^3, respectivamente. A variável *Difficulty* indica a dificuldade da matéria, sendo d^0 representada como fácil e d^1 como difícil.

A variável *Intelligence* representa o tempo de estudo dedicado pelo aluno, classificado como poucas horas de estudo para i^0 e muitas horas de estudo para i^1.. *SAT* indica a pontuação do aluno no exame, com pontuações baixas representadas por s^0 e altas por s^1. A variável *Reference Letter* representa se o aluno obteve ou não uma carta de referência após a conclusão da disciplina, sendo representada por l^1 em caso afirmativo e l^0 em caso negativo.

Com isso, é possível observar o fluxo probabilístico de influência que uma variável exerce sobre a outra. Por exemplo, considerando as variáveis *Difficulty* e *Intelligence* apontando para a variável *Grade*, observa-se que os estados das variáveis *Difficulty* e *Intelligence* afetam a probabilidade dos diversos estados possíveis da variável *Grade*.

Isso fica evidente na tabela de distribuição de probabilidade conjunta presente na RB, onde é observado que se tanto a *Intelligence* quanto a *Difficulty* forem 0, indicando baixo tempo de estudo dedicado pelo aluno e baixa dificuldade da matéria, então a probabilidade de observarmos um *Grade* correspondente a g^1 é de 0,3, g^2 é 0,4 e g^3 é 0,3. Esse exemplo de RB ilustra de que forma a observação de uma variável pode influenciar probabilisticamente

o estado de outra.

Um exemplo adicional de inferência extraído dessa rede é o seguinte: Se observarmos que a variável *Intelligence* é classificada como baixa, notamos que a probabilidade do aluno obter uma pontuação baixa no SAT é alta, equivalente a 0,95, enquanto a probabilidade de obter uma pontuação alta é baixa, equivalente a 0,05. Por outro lado, se observarmos que o tempo de estudo dedicado pelo aluno é alto, as probabilidades que ele tire uma pontuação baixa e alta no SAT são 0,2 e 0,8, respectivamente. Isso evidencia que a observação da variável *Intelligence* tem uma influência direta no estado da variável SAT.

Dado um grafo acíclico direcionado $\mathcal{G}$ que descreve uma RB, composta por um conjunto de variáveis aleatórias $X = (X_1, X_2, \ldots, X_n)$, uma distribuição P sobre esse espaço é dita fatorada de acordo com $\mathcal{G}$ se puder ser expressa como um produto das distribuições condicionais locais. Essa equação é chamada de regra da cadeia para RBs (KOLLER; FRIEDMAN, 2009).

$$P(X_1, X_2, \ldots, X_n) = \prod_{i=1}^{n} P\left(X_i \mid \text{Pais}(X_i)\right) \tag{2.5}$$

em que a expressão $P(X_1,X_2,...,X_n)$ denota a distribuição de probabilidade conjunta das variáveis aleatórias, enquanto $P_G(X_i)$ representa o conjunto de pais do nó X e $P(X_i|\text{P}_G(X_i))$ expressa as distribuições de probabilidade condicional de cada nó. Consequentemente, ao atribuirmos valores específicos a uma parte das variáveis, torna-se possível calcular a probabilidade de ocorrência dos valores associados às variáveis restantes, cujos valores são desconhecidos (MACHADO et al., 2021).

Logo, no contexto do nosso exemplo ilustrado na Figura 1, com as variáveis aleatórias *Difficulty* (D), *Intelligence* (I), *Grade* (G), *SAT* (S) e *Letter* (L), obtemos a seguinte equação utilizando a Regra da Cadeia para RBs.

$$P(D, I, G, S, L) = P(D).P(I).P(G|I,D).P(S|I).P(L|G) \tag{2.6}$$

A expressão $P(D, I, G, S, L)$ denota a distribuição de probabilidade conjunta das variáveis aleatórias representando a probabilidade de todas essas variáveis ocorrerem simultaneamente. Enquanto as variáveis $P(D)$ e $P(I)$ não dependem de nenhum conjunto de nós pais, sendo tratadas separadamente, as variáveis G, S e L dependem de seus respectivos conjuntos de nós pais dentro da rede. Com isso, $P(G|I, D)$ indica a probabilidade da variável aleatória *Grade* ocorrer, dado que *Intelligence* e *Difficulty* foram observados

e são seus pais na RB; a mesma interpretação pode ser feita para $P(S|I)$ e $P(L|G)$. Assim, a equação representa a probabilidade conjunta de todas essas variáveis, onde a probabilidade de cada variável é condicionada aos valores de suas variáveis pai na RB.

2.1.5 D-separação

Esta seção se propõe a explorar a noção de d-separação (*D-Separation*) em RBs, uma ferramenta fundamental para entender as relações de independência condicional entre as variáveis. Ou seja, quando podemos garantir que uma independência ($X \perp Y \mid Z$) ocorre em uma distribuição associada a uma estrutura de RB (KOLLER; FRIEDMAN, 2009). Logo, o conhecimento sobre *D-Separation* nos permite analisar quando é possível que X influencie Y dado Z.

Nesse contexto, é necessário considerar o conceito de *D-Separation* no raciocínio, pois somente por meio dessa abstração é possível formalizar e compreender melhor as relações de dependência complexas e as complicações associadas entre variáveis em uma RB (CONTALDI, 2016). Logo, concentrando nossa discussão em quando é possível que X influencie Y dado Z, de acordo com (KOLLER; FRIEDMAN, 2009), obtêm-se os seguintes casos.

A conexão direta é o caso em que X e Y estão diretamente conectados por uma aresta, indicando um fluxo probabilístico direto. Ou seja, para qualquer estrutura da rede que possua a aresta $X \rightarrow Y$, é possível construir uma distribuição onde X e Y estão correlacionados independentemente de qualquer evidência sobre as outras variáveis na rede.

A conexão indireta é o caso em que X e Y não estão diretamente conectados, mas há um caminho entre eles no grafo via Z. Existem quatro possibilidades em que X e Y estão conectados via Z, conforme ilustrado na Figura 2. Os dois primeiros correspondem a ideias de causalidade, o terceiro a uma causa comum e o quarto a um efeito comum.

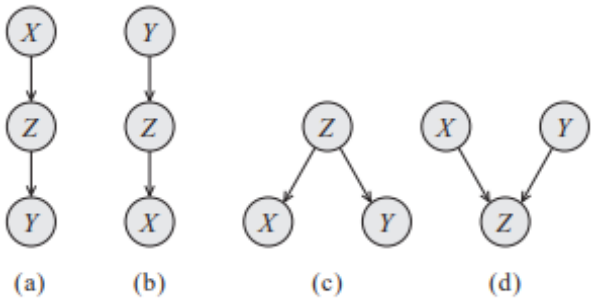

Figura 2: As quatro possíveis trilhas de duas arestas de X a Y via Z (KOLLER; FRIEDMAN, 2009).

No item (a) e (b), da Figura 2, observamos um efeito causal indireto nas variáveis X e Y mediado por Z. Logo, é evidente que as variáveis ainda mantêm uma influência probabilística entre si por meio de Z.

Um exemplo ilustrativo para o item (a) pode ser observado na Figura 1. Ao observar a variável *Intelligence* (X), ela influencia o estado da variável *Grade* (Z) que, por sua vez, afeta a probabilidade da variável *Letter* (Y). Essa dinâmica é análoga à observada no item (b). Logo, fica claro que essa estrutura implica uma relação probabilística entre as variáveis.

No item (c), observa-se que a variável Z tem dois efeitos: um sobre X e outro sobre Y. Um exemplo para esse item pode ser observado pela Figura 1, onde a variável *Intelligence* influencia tanto o estado da variável *SAT* quanto o da variável *Grade*. Em outras palavras, ao saber sobre a inteligência do aluno, nossa crença sobre sua nota na matéria e sua pontuação no SAT é alterada. Isso evidencia claramente a influência probabilística presente.

No item (d), nota-se que duas variáveis, X e Y, influenciam conjuntamente Z. Essa configuração é conhecida como estrutura em V, onde as variáveis X e Y influenciam Z, sem que exista uma conexão direta entre elas. Exemplo: Quando conhecemos o nível de dificuldade de uma matéria, representado pela variável *Difficulty* (X), isso afeta o estado da variável *Grade* (Z). No entanto, claramente, essa informação não nos fornece nenhuma indicação sobre a variável *Intelligence*. Assim, neste cenário, não identificamos um fluxo de influência probabilística, indicando a ausência de uma trilha ativa entre as variáveis. Logo, dizemos que X e Y são *D-separados* dado Z se não houver um caminho ativo entre X e Y dado Z. Denotamos Z como o conjunto de separação (*SepSet*) de X e Y.

2.1.6 Markov Blanket

Em RBs, o conceito de *Markov Blanket* (MB) é fundamental para compreender as relações de independência local entre as variáveis. Dado um grafo H, define-se o MB de uma variável X em H, denotada por $MB_H(X)$, como o conjunto de nós vizinhos imediatos de X no grafo (KOLLER; FRIEDMAN, 2009). Com base nessa definição, as independências locais associadas ao grafo H estabelecem que a variável X é condicionalmente independente de todos os demais nós do grafo, dado seu MB. Em outras palavras, todo o conhecimento necessário para inferir o comportamento de X está contido nos seus vizinhos diretos.

Complementando essa definição, (PEARL, 2014) descreve que o MB de um nó X é composto pelos pais de X, seus filhos, e os outros pais desses filhos. Essa estrutura permite que, ao condicionarmos X sobre seu MB, eliminemos qualquer dependência com o restante das variáveis da rede. Portanto, o MB representa o menor conjunto de variáveis necessário para isolar X de influências externas dentro da estrutura de uma RB.

Neste trabalho, o conceito de MB é útil para interpretar quais variáveis contêm a informação local mais relevante para a variável de interesse, isto é, a precipitação. Em termos aplicados, ele auxilia na compreensão de quais variáveis meteorológicas estão mais diretamente associadas ao comportamento probabilístico do nó de chuva.

2.1.7 Algoritmos de aprendizado das Redes Bayesianas

Do ponto de vista prático, a utilização das RBs em problemas do mundo real está condicionada à existência de métodos de aprendizado automático capazes de inferir tanto a estrutura gráfica quanto os parâmetros das probabilidades condicionais (CANO; SORDO; GUTIÉRREZ, 2004).

A aprendizagem da estrutura é o processo para obter uma representação estrutural mais otimizada do conjunto de dados, de forma que forme um DAG (MACHADO et al., 2021). Atualmente, a literatura apresenta três principais abordagens para o problema de aprendizagem estrutural, sendo elas *Search and Score* (S&S), *Constraint-Based (CB)* e os *métodos híbridos* (CONTALDI, 2016).

Essas três abordagens são particularmente relevantes para esta dissertação porque o objetivo metodológico do trabalho inclui comparar diferentes estratégias de aprendizado estrutural no contexto da modelagem probabilística da precipitação. Assim, a fundamentação dessas abordagens fornece a base conceitual necessária para interpretar as diferenças

estruturais e preditivas observadas entre os modelos avaliados.

2.1.7.1 Search and Score

O método *Search and Score* ou *Score-Based* aborda o problema do aprendizado da estrutura de uma RB como um problema de otimização. Este método tem como objetivo encontrar a estrutura de rede que maximize uma função de pontuação específica, a qual avalia o quão bem o modelo se ajusta aos dados observados (KOLLER; FRIEDMAN, 2009). Logo, o objetivo é encontrar a estrutura g^* no espaço G que maximize a função objetivo $f(g,D)$, onde D representa o conjunto de dados disponíveis para a estimação (DE CAMPOS; FRIEDMAN, 2006).

$$g^* = \arg\max_{g \in G} f(g, D) \tag{2.7}$$

É definido um espaço de hipóteses de modelos potenciais, representando as possíveis estruturas de rede consideradas, juntamente com uma função de pontuação que quantifica o ajuste do modelo aos dados observados. Uma característica do método *Score-Based* é sua capacidade de considerar a estrutura como um todo, tornando-o menos sensível a falhas individuais nos testes de independência como no método *Constraint-Based*. Porém, encontrar a estrutura de rede com a pontuação mais alta pode ser um problema NP-difícil, o que muitas vezes requer o uso de técnicas de busca heurística para encontrar uma solução aproximada (KOLLER; FRIEDMAN, 2009). Com isso, as métricas desempenham o papel de qualificar as RBs, pois não apenas buscam maximizar a adequação do modelo aos dados, mas também penalizam estruturas excessivamente complexas.

Conforme destacado por (LIU; MALONE; YUAN, 2012), a resolução exata do problema de aprendizado da estrutura torna-se inviável à medida que o número de variáveis aumenta significativamente. Diante dessa limitação, diversas abordagens iniciais passaram a adotar métodos aproximativos, entre os quais se destacam os algoritmos gananciosos como o *Hill Climbing* (HC) e o *Tabu Search* com reinicializações aleatórias. Tais métodos baseiam-se majoritariamente em estratégias de busca local para identificar estruturas de rede com boa pontuação, embora não assegurem a obtenção da solução ótima global.

O algoritmo HC (BUNTINE, W. L., 1994) é uma técnica de busca local aplicada ao aprendizado da estrutura de RBs. A cada iteração, o algoritmo avalia todas as operações locais disponíveis, como inserção, remoção ou reversão de arcos, e seleciona a operação que proporciona o maior ganho na função de pontuação.

O procedimento tem início com um conjunto de dados definido sobre o conjunto de variáveis V e um DAG inicial, frequentemente o vazio. Em cada passo, calcula-se o impacto de cada operação local no valor da função de escore, e aplica-se a alteração que resultar na maior melhoria positiva. Esse ciclo continua até que nenhuma modificação adicional leve a um aumento no escore total da estrutura.

O algoritmo *Tabu Search* (GLOVER, 1990) é uma técnica heurística voltada à resolução de problemas de otimização, distinguindo-se por utilizar mecanismos de memória adaptativa para superar as limitações das buscas locais e aproximar-se do ótimo global. Um dos principais componentes dessa abordagem é a lista tabu, que armazena temporariamente movimentos recentemente realizados, impedindo o retorno a soluções já exploradas e, assim, evitando ciclos. Essa lista é dinamicamente atualizada e contribui para diversificar a trajetória da busca. Adicionalmente, o método emprega critérios de aspiração, os quais permitem flexibilizar as restrições da lista tabu quando uma solução promissora, normalmente significativamente melhor, é identificada. O processo iterativo continua até o atendimento de um critério de parada, como o número máximo de iterações ou a estagnação da função de avaliação.

2.1.7.2 Funções de Pontuação

As funções de pontuação exercem um papel central na abordagem *Search and Score*, pois são responsáveis por quantificar a qualidade de uma estrutura de RB candidata com base em um conjunto de dados observado. Elas atuam como critério de avaliação durante o processo de busca, influenciando diretamente as decisões tomadas pelos algoritmos heurísticos (KOLLER; FRIEDMAN, 2009).

Diferentes funções de pontuação incorporam distintas suposições estatísticas e penalidades associadas à complexidade do modelo, buscando equilibrar a fidelidade do ajuste aos dados com a simplicidade estrutural (HECKERMAN; GEIGER; CHICKERING, 1995; KOLLER; FRIEDMAN, 2009). Entre as funções mais amplamente utilizadas na literatura destacam-se o *Bayesian Information Criterion* (BIC), *Akaike Information Criterion* (AIC), *K2 score* (K2) e *Bayesian Dirichlet score* (BDe).

As métricas BIC e AIC pertencem à classe de funções de pontuação fundamentadas na teoria da informação, cujo objetivo principal é evitar o sobreajuste. Para isso, buscam equilibrar a qualidade do ajuste do modelo aos dados com a complexidade estrutural da rede, penalizando modelos excessivamente parametrizados em relação ao tamanho do conjunto de dados disponível (KITSON et al., 2023).

O BIC, baseado em (SCHWARZ, 1978), é definido como a diferença entre a log-verossimilhança do modelo e um termo de penalização proporcional à complexidade da estrutura da rede. No contexto das RBs, a log-verossimilhança (LL) de uma estrutura R, dado um conjunto de dados D, é expressa por:

$$LL(R|D) = \sum_{i=1}^{n} \sum_{j=1}^{q_i} \sum_{k=1}^{r_i} N_{ijk} \log \left(\frac{N_{ijk}}{N_{ij}} \right) \tag{2.8}$$

Nessa equação, calcula-se a soma das contribuições de cada variável e de suas configurações de estados, considerando a frequência observada nos dados. A função BIC é então formalizada como:

$$BIC(R|D) = LL(R|D) - \frac{1}{2} \log(N) \sum_{i=1}^{n} (r_i - 1) q_i \tag{2.9}$$

em que

- n é o número total de nós na rede,
- q_i é o número de pais do i-ésimo nó,
- r_i é o número de estados do i-ésimo nó,
- N_{ijk} é o número total de observações na amostra em que o i-ésimo nó assume seu k-ésimo estado e o seus pais assumem o j-ésimo estado,
- N_{ij} é a quantidade de observações do i-ésimo nó quando seus pais assumem o j-ésimo estado.

A métrica AIC (AKAIKE, 2003) pode ser interpretada como uma derivação do BIC, diferenciando-se principalmente pela ausência do termo de penalização proporcional ao tamanho da amostra. A função AIC é definida como:

$$AIC(R|D) = LL(R|D) - \sum_{i=1}^{n} (r_i - 1) q_i \tag{2.10}$$

Esse critério penaliza a complexidade do modelo de forma mais branda em comparação ao BIC. No contexto das RBs, o AIC tende a selecionar estruturas mais complexas, priorizando o ajuste aos dados observados, ainda que isso implique em uma maior quantidade de parâmetros. Em contrapartida, o BIC impõe uma penalização mais severa à

complexidade estrutural. Essa penalização aumenta com o tamanho do conjunto de dados, favorecendo modelos mais simples à medida que mais informações estão disponíveis.

As métricas BDe e K2 integram o conjunto de funções de pontuação bayesianas, responsáveis por atribuir uma probabilidade posterior relativa a distintas estruturas de grafos, condicionada aos dados observacionais. Tais abordagens articulam a evidência empírica com crenças a priori acerca da estrutura gráfica ou dos parâmetros de dependência (KITSON et al., 2023).

A função de pontuação *K2* (COOPER; HERSKOVITS, 1992) pode ser entendida como uma simplificação da pontuação bayesiana geral derivada da distribuição *Dirichlet*. Nessa abordagem, assume-se uma distribuição a priori onde todos os valores possíveis dos parâmetros têm o mesmo peso ($N'_{ijk} = 1$), o que resulta na seguinte função de pontuação:

$$S_{K2}(R,D) = logP(R) + \sum_{i=1}^{n}\sum_{j=1}^{q_i}\left[\log\left(\frac{(r_i-1)!}{(N_{ij}+r_i-1)!}\prod_{k=1}^{r_i}N_{ijk}!\right)\right] \tag{2.11}$$

onde:

- $P(R)$ é a probabilidade a priori de uma determinada estrutura de grafo, que geralmente é assumida como igual para todos os grafos e, portanto, pode ser desconsiderada.
- i é o índice sobre as n variáveis,
- j é o índice sobre as q_i combinações de valores dos pais do nó X_i,
- r_i é número de estados possíveis que a variável X_i pode assumir,
- k é o índice sobre os r_i possíveis valores (estados) do nó X_i,
- N_{ijk} representa o número de instâncias nos dados D em que o nó X_i assume o k-ésimo valor , e seus pais assumem a j-ésima combinação de valores, sendo que $N_{ij} = \sum_{k=1}^{r_i} N_{ijk}$ representa o número total de instâncias nos dados D em que os pais do nó X_i assumem a j-ésima combinação de valores.

Entretanto, o *K2* não é equivalente em escore, ou seja, diferentes estruturas de rede que representam a mesma independência condicional podem receber pontuações distintas. Esse aspecto pode levar a preferências enviesadas durante a busca estrutural.

Para lidar com essa limitação, foi proposto o escore BDe (HECKERMAN; GEIGER; CHICKERING, 1995), o qual introduz a equivalência de escore, garantindo que redes equivalentes em termos de independência condicional recebam a mesma pontuação. A formulação do BDe baseia-se em uma distribuição *Dirichlet* e pode ser expressa por

$$S_{\mathrm{BDe}}(R, D) = \log P(R) + \sum_{i=1}^{n} \sum_{j=1}^{q_i} \left[\log \left(\frac{\Gamma\left(N' \sum_{k=1}^{r_i} \theta'_{ijk}\right)}{\Gamma\left(N_{ij} + N' \sum_{k=1}^{r_i} \theta'_{ijk}\right)} \right) + \sum_{k=1}^{r_i} \log \left(\frac{\Gamma\left(N_{ijk} + N'\theta'_{ijk}\right)}{\Gamma\left(N'\theta'_{ijk}\right)} \right) \right] \tag{2.12}$$

onde:

- Γ denota a função Gama.
- θ'_{ijk} é a probabilidade condicional *a priori* de o nó X_i assumir o k-ésimo valor, dado que seus pais apresentam a j-ésima combinação de valores na distribuição a priori.
- N' é o tamanho amostral equivalente e expressa o grau de confiança nos parâmetros a priori.

Em síntese, cada função de pontuação avalia a qualidade das estruturas candidatas segundo diferentes princípios estatísticos, podendo levar a modelos mais parcimoniosos ou mais ajustados. A escolha do critério é, portanto, um aspecto decisivo do desempenho da abordagem *Search and Score*.

2.1.7.3 Constraint-Based

A abordagem *Constraint-Based* foi um dos primeiros métodos para a aprendizagem da estrutura de uma RB. A ideia geral desse método é encontrar uma estrutura ótima entre todas as possíveis, restringindo progressivamente o espaço de busca (CONTALDI, 2016). Este conjunto de algoritmos fundamenta-se em testes de independência condicional para determinar a presença de arcos. Dessa forma, ao constatar-se a independência entre duas variáveis, o arco que as conecta é removido (OZELAME et al., 2021). No entanto, este método pode ser sensível a falhas nos testes de independências individuais (KOLLER; FRIEDMAN, 2009).

Os algoritmos que utilizam esse método de aprendizagem seguem três passos baseados no teste de independência condicional. O primeiro passo, que é opcional, define o MB para cada nó, eliminando candidatos na criação das estruturas. O segundo passo consiste na realização de testes de simetria para verificar e remover nós assimétricos, com o objetivo

de criar a estrutura da RB de forma não direcionada. O terceiro passo é a definição da direção dos arcos da estrutura criada anteriormente, visando estabelecer a orientação causal da RB (MACHADO et al., 2021).

O algoritmo *Causality Search* (PC) (SPIRTES; GLYMOUR; SCHEINES, 2000), é um método tradicional baseado em *Constraint-Based* para RBs. Esse algoritmo inicia com a construção de um grafo completo sobre o conjunto de variáveis V, no qual cada par de variáveis está conectado por uma aresta. Para cada par (X, Y), é criado um conjunto de separação SepSet(X, Y), inicialmente vazio.

Na etapa seguinte, o algoritmo entra em um processo iterativo caracterizado por um índice n, que começa em zero e é incrementado ao longo das iterações. Em cada ciclo, são examinados todos os pares de nós adjacentes (X, Y) cujos conjuntos de vizinhos tenham cardinalidade igual ou superior a n. Para esses pares, são realizados testes de independência condicional com todas as combinações possíveis de n vizinhos como conjunto condicionante. Quando for detectada independência condicional entre X e Y dado um subconjunto de seus vizinhos, a aresta entre eles é eliminada e o subconjunto responsável por essa separação é registrado em SepSet(X, Y). Este processo se repete até que não existam mais pares de nós com vizinhança suficiente para formar conjuntos condicionantes de tamanho n.

Finalizada a fase de remoção de arestas, o algoritmo passa para a etapa de orientação das conexões. Essa fase se baseia na identificação de *v-estruturas* e na aplicação de regras de inferência causal. Uma *v-estrutura* é identificada em um trio de variáveis (X, Y, Z) quando Y é adjacente a X e Z, mas não há conexão direta entre X e Z, e Y não pertence ao conjunto de separação entre X e Z. Nessa configuração, Y é considerado um *colisor*, e as arestas são orientadas como $X \rightarrow Y \leftarrow Z$.

Após a identificação das *v-estruturas*, aplicam-se regras de propagação para orientar as arestas restantes, conforme descrito por (PEARL et al., 2000). Essas regras são aplicadas iterativamente até que não restem arestas não direcionadas passíveis de orientação com base nas evidências disponíveis.

O algoritmo Grow-Shrink (GS) (MARGARITIS, 2003) adota uma abordagem centrada na determinação do MB de cada variável. Inicialmente, o algoritmo identifica os vizinhos de cada nó por meio de testes de independência condicional, realizados de forma iterativa, considerando um vértice por vez. Esse processo baseia-se na propriedade de condicionamento total, segundo a qual, em um grafo causal fiel, pais, filhos e cônjuges de um nó retêm informações exclusivas sobre ele

Após a identificação do MB, o grafo é convertido em sua versão moral, uma forma não direcionada em que cônjuges são conectados por arestas adicionais. Em seguida, são realizados testes de independência entre o nó em análise e cada vértice do seu MB, condicionados a subconjuntos do restante desse conjunto. Se for detectada independência, a aresta correspondente é removida.

Na fase final de orientação, o algoritmo examina pares (X, Y), com $Y \in \text{MB}(X)$, e analisa os vizinhos Z de X que não estão conectados a Y. Para cada Z, tenta-se orientar a conexão de Y em direção a X. A orientação é mantida apenas se Y e Z forem condicionalmente dependentes, dado qualquer subconjunto das respectivas coberturas de Markov.

O algoritmo *Incremental Association Markov Blanket* (IAMB) foi desenvolvido com o propósito de tornar mais eficiente a identificação do MB, permitindo sua aplicação em redes que contêm milhares de nós (TSAMARDINOS; ALIFERIS; STATNIKOV; STATNIKOV, 2003). De acordo com os autores, a etapa de expansão do MB no algoritmo GS apresenta limitações, pois tende a ser lenta na identificação de cônjuges em $\text{MB}(T)$, pois esses tipicamente exibem associação fraca com T. Essa limitação acaba resultando em um número elevado de testes de independência condicional nas fases de crescimento e de redução.

Como alternativa, os autores sugerem utilizar a informação mútua condicional $\text{MI}(X, T \mid \text{MB}(T))$ para definir a sequência em que os nós X são avaliados para possível inclusão no conjunto $\text{MB}(T)$ durante a etapa de expansão. A informação mútua condicional (MARGARITIS; THRUN, 1999) quantifica a dependência estatística entre duas variáveis aleatórias X e T, condicionada a uma terceira variável ou conjunto de variáveis.

O algoritmo *Max-Min Parents and Children* (MMPC) foi inicialmente introduzido por Tsamardinos, Aliferis e Statnikov (2003) e posteriormente descrito de forma mais completa por Tsamardinos, Brown e Aliferis (2006). O MMPC é uma técnica de aprendizado de estrutura local que visa identificar o conjunto de pais e filhos (*Parents and Children*, ou PC) de um nó-alvo T . Ele opera em duas fases principais: uma fase forward (expansiva) e uma fase backward (refinadora). Na fase forward, o algoritmo constrói um conjunto candidato de pais e filhos (CPC) utilizando uma heurística baseada em testes de associação condicional. A cada iteração, a variável com a maior associação mínima com T, mesmo após considerar todos os subconjuntos de CPC para tentar torná-la independente, é adicionada ao conjunto.

Na fase backward, o algoritmo aplica testes de independência condicional para re-

mover falsos positivos do conjunto CPC, verificando se alguma das variáveis pode ser d-separada de T ao se condicionar em subconjuntos de CPC. Se uma variável for considerada independente de T em algum subconjunto, ela é excluída (TSAMARDINOS; BROWN; ALIFERIS, 2006). Embora a complexidade teórica do MMPC possa ser alta no pior caso, na prática o número de candidatos analisados tende a ser pequeno, tornando-o viável mesmo em conjuntos de dados com muitas variáveis. O algoritmo utiliza testes estatísticos, como o teste G^2 (SPIRTES; GLYMOUR; SCHEINES, 2000), para avaliar independência, respeitando limitações amostrais para garantir confiabilidade estatística. Por sua abordagem sistemática e escalabilidade, o MMPC tem se mostrado eficaz na indução local de estruturas em RBs, sendo amplamente utilizado em contextos onde se busca interpretabilidade e precisão estrutural.

2.1.7.4 Hybrid

A abordagem *Hybrid* é a combinação dos métodos *Search and Score* e *Constraint-Based*, integrando os testes de independência com as buscas baseadas em pontuação. Inicialmente, os testes de independência são utilizados para criação da estrutura, em seguida, aplica-se a métrica de pontuação que irá direcionar os arcos. Embora ainda sejam menos comuns em comparação com as metodologias anteriores, a integração dessas abordagens pode ajudar a reduzir as limitações inerentes a cada algoritmo (OZELAME et al., 2021).

O algoritmo *Max–Min Hill Climbing (MMHC)* (TSAMARDINOS; BROWN; ALIFERIS, 2006) representa uma abordagem híbrida amplamente utilizada na aprendizagem de estruturas de RBs. Sua operação é dividida em duas etapas principais, na primeira, chamada fase de restrição, o MMHC utiliza o algoritmo baseado MMPC para construir o esqueleto do grafo da RB. Já na segunda etapa, denominada fase de maximização, o algoritmo aplica o método de busca *Tabu hill-climbing* com a limitação de considerar apenas as conexões identificadas anteriormente no esqueleto, restringindo o espaço de busca da DAG.

O algoritmo *Restricted Maximization* (rsmax2) (SCUTARI; SCUTARI; MMPC, 2019) é uma implementação geral dos algoritmos Sparse Candidate. Esses algoritmos (FRIEDMAN; NACHMAN; PE'ER, 2013) têm como objetivo de tornar mais eficiente o processo de aprendizagem de estruturas em RBs. O método propõe uma redução inicial no conjunto de possíveis pais para cada nó, utilizando critérios métricos específicos. Com base nesse conjunto reduzido, aplica-se um algoritmo tradicional de HC para construir uma estrutura inicial da rede. A partir dessa rede preliminar, os conjuntos de candidatos são atualiza-

dos e o processo é repetido de forma iterativa. A principal contribuição do algoritmo, em relação ao HC convencional, está na sua capacidade de reduzir o espaço de busca, proporcionando ganhos substanciais de desempenho computacional, particularmente em bases de dados com alta dimensionalidade (CONTALDI, 2016).

2.1.7.5 Aprendizagem de parâmetros

Para obter a distribuição de probabilidade conjunta *(Joint Probability Distribution - JPD)* de todas as variáveis, é utilizado o processo conhecido como aprendizado de parâmetros. As abordagens principais para este processo incluem a máxima verossimilhança e o algoritmo de maximização da expectativa (MACHADO et al., 2021).

A aprendizagem de parâmetros em RBs acausais será tratada a partir da abordagem apresentada por (HECKERMAN, 2013), a qual revisa métodos discutidos na literatura, como os descritos por (SPIEGELHALTER; LAURITZEN, 1990),(SPIEGELHALTER; DAWID et al., 1993) (COOPER; HERSKOVITS, 1992), (BUNTINE, W., 1991), (BUNTINE, W. L., 1994), (MADIGAN; RAFTERY, 1994), (HECKERMAN; GEIGER; CHICKERING, 1995).

Seja $U = \{X_1, X_2, \ldots, X_n\}$ um conjunto de variáveis aleatórias que definem o domínio de interesse da RB. Além disso, suponha a existência de um banco de dados composto por um conjunto de casos $C = \{C_1, C_2, \ldots, C_m\}$ onde cada instância de C_l corresponde a um vetor de observações parciais ou completas das variáveis em U.

Assumindo que C seja uma amostra aleatória de um conjunto de variáveis U, seguindo uma distribuição de probabilidade $p(U)$. Essa distribuição pode ser caracterizada por um conjunto finito de parâmetros Θ_U. Considerando-se que todas as variáveis em análise são discretas, Θ_U representa diretamente as probabilidades da distribuição.

A aprendizagem de parâmetros envolve avaliar a distribuição a priori e posteriori, modelando $p(U)$ por meio de uma estrutura acausal cuja parametrização é desconhecida. A atualização dos parâmetros será analisada para uma estrutura de rede acausal arbitrária, permitindo a atualização independente dos parâmetros, sob a suposição de independência e dados completos. O conjunto de dados é tratado como uma amostra aleatória dessa rede.

Seja r_i o número de estados da variável aleatória x_i. Definimos q_i como o número de combinações possíveis dos estados das variáveis que são pais de x_i na rede, onde $Pa(x_i)$ denota o conjunto de variáveis que são pais de x_i e r_l é o número de estados da variável

$x_l \in \mathrm{Pa}(x_i)$

$$q_i = \prod_{x_l \in \mathrm{Pa}(x_i)} r_l \tag{2.13}$$

Seja θ_{ijk} o parâmetro associado à probabilidade condicional, onde ξ representa o conhecimento ou informação prévia disponível sobre o domínio de interesse, $\theta_{ijk} > 0$ e $\sum_{k=1}^{r_i} \theta_{ijk} = 1$. Assim, θ_{ijk} representa a probabilidade condicional de $x_i = k$, dada a configuração j dos pais de x_i

$$\theta_{ijk} = P(x_i = k \mid \mathrm{Pa}(x_i) = j, \xi) \tag{2.14}$$

Além disso, define-se Θ_{ij} como o vetor de parâmetros associado à distribuição condicional de x_i, dado $\mathrm{Pa}(x_i) = j$.

$$\Theta_{ij} \equiv \bigcup_{k=1}^{r_i} \{\theta_{ijk}\} \tag{2.15}$$

Ademais, o conjunto completo de parâmetros da RB, denotado por Θ_{Bs}, onde n representa o número total de variáveis na rede e q_i é o número de configurações distintas dos pais de x_i. Esse conjunto representa todas as probabilidades condicionais possíveis, abrangendo o espaço completo de distribuições que podem ser atribuídas aos parâmetros da rede.

$$\Theta_{Bs} \equiv \bigcup_{i=1}^{n} \bigcup_{j=1}^{q_i} \Theta_{ij} \tag{2.16}$$

Supondo que Θ_{ij} segue uma distribuição de probabilidade Dirichlet a priori, onde B_h^s é a afirmação que o conjunto de dados é uma amostra aleatória da RB, C uma constante normalizadora e N'_{ijk} representa uma pseudo-contagem,

$$p(\Theta_{ij} \mid B_h^s, \xi) = c \cdot \prod_{k=1}^{r_i} \theta_{ijk}^{N'_{ijk}-1} \tag{2.17}$$

Assumindo a independência entre os parâmetros e a completude do banco de dados, se N_{ijk} denota o número de ocorrências no conjunto de dados C em que $x_i = k$ e $Pa(x_i) = j$, então a distribuição a posteriori de Θ_{ij} é denotada por

$$p(\Theta_{ij} \mid C, B_h^s, \xi) = c \cdot \prod_{k=1}^{r_i} \theta_{ijk}^{N'_{ijk}+N_{ijk}-1} \tag{2.18}$$

Ao tomar a esperança matemática de θ_{ijk} em relação à distribuição de Θ_{ij}, para cada i e j obtém-se a probabilidade de ocorrência de $x_i = k$ e $Pa(x_i) = j$ no próximo caso a ser observado no conjunto de dados, denotado por C_{m+1}, onde $N'_{ij} = \sum_{k=1}^{r_i} N'_{ijk}$ e $N_{ij} = \sum_{k=1}^{r_i} N_{ijk}$.

$$p(C_{m+1} \mid C, B_h^s, \xi) = \prod_{i=1}^{n} \prod_{j=1}^{q_i} \frac{N'_{ijk} + N_{ijk}}{N'_{ij} + N_{ij}} \tag{2.19}$$

Além do método apresentado, também é possível utilizar a abordagem de Estimação de Máxima Verossimilhança, que, ao contrário da abordagem Bayesiana, desconsidera as informações a priori e trata a estimação exclusivamente como um problema de contagem, maximizando a verossimilhança dos dados observados, onde $N(C \mid A)$ é a quantidade de casos em que ocorrem C e A.

$$P(C \mid A) = \frac{N(C \mid A)}{N(A)} \tag{2.20}$$

2.2 Técnicas Complementares

Além dos conceitos diretamente relacionados às RBs, esta dissertação recorre a técnicas complementares que apoiam a preparação dos dados, a redução da complexidade do problema e a avaliação dos modelos ajustados. Essas técnicas foram incorporadas ao fluxo metodológico com o objetivo de tornar a modelagem da precipitação mais estável, interpretável e adequada às características do conjunto de dados utilizado.

2.2.1 K-Means

As técnicas de agrupamento não hierárquicas consistem na definição de sementes iniciais que formarão os núcleos de k grupos de dados. Dentre essas técnicas, destaca-se o algoritmo amplamente utilizado, *k-means*. Esse método divide os dados em K grupos (ou clusters), alocando cada elemento ao grupo cujo centróide esteja mais próximo (ABDULNASSAR; NAIR, 2023).

Embora diferentes métricas de distância possam ser empregadas, a mais comum é a

distância euclidiana, dada por

$$d(p,q) = \sqrt{\sum_{i=1}^{n}(p_i - q_i)^2} \tag{2.21}$$

onde $p = (p_1, p_2, ..., p_n)$ e $q = (q_1, q_2, ..., q_n)$ são dois pontos em um espaço.

O procedimento do algoritmo segue as etapas descritas por (MACQUEEN, 1967):

- Especificar as k sementes iniciais que criarão os k grupos;
- Calcular a distância de cada elemento a cada uma das sementes e definir em qual grupo aquele elemento será alocado;
- Recalcular os centroides com base nos grupos formados;
- Recalcular as distâncias entre os elementos e os novos centroides;
- Realocar os elementos conforme as menores distâncias entre eles e os centroides;
- Repetir essas etapas até que não seja mais necessário realocar elementos.

Assim como outras técnicas de agrupamento não hierárquicas, o método k-means apresenta como vantagem a flexibilidade de realocação dos elementos entre os grupos. No entanto, apresenta como limitação a necessidade de definir previamente o número de grupos a ser utilizado. Ademais, o k-means é sensível à presença de valores discrepantes, característica que pode representar uma vantagem ou uma desvantagem, a depender das características do conjunto de dados e dos objetivos da análise.

No contexto desta dissertação, o algoritmo k-means foi utilizado como estratégia de discretização de variáveis contínuas, favorecendo a compatibilização dos dados com a modelagem por RBs discretas. Além de reduzir a granularidade dos valores observados, essa etapa contribui para a construção de tabelas de probabilidade condicional mais manejáveis e interpretáveis.

2.2.2 Análise de Componentes Principais

De acordo com (IZBICKI; SANTOS, 2020), os métodos de redução de dimensionalidade, como o de Análise de Componentes Principais (ACP), têm como objetivo identificar transformações das variáveis originais que consigam reter uma parte significativa das informa-

ções contidas nelas. Isso permite eliminar redundâncias entre as variáveis e reduzir sua quantidade, buscando assim um modelo mais parcimonioso.

Nesta técnica, a variabilidade de uma variável Z_i é medida através de sua variância, e a redundância entre duas variáveis Z_i e Z_j é avaliada pela correlação entre elas. A ACP busca encontrar novas variáveis que são combinações lineares das covariáveis originais (IZBICKI; SANTOS, 2020).

De modo genérico, o i-ésimo componente principal de X, para $i \geq 1$, é a variável Z_i que:

1. É uma combinação linear das variáveis originais, expressa como:

$$Z_i = \phi_{1i}X_1 + \cdots + \phi_{di}X_d \tag{2.22}$$

   com a restrição de normalização: $\sum_{k=1}^{d} \phi_{ki}^2 = 1$.

2. Possui a maior variância possível, capturando o máximo de informação.

3. Tem correlação zero com todos os componentes principais já calculados $(Z_1, \ldots, Z_{i-1})$.

Cada componente é definida para ter a maior variância possível, mas sob a condição de não conter informação redundante em relação aos componentes anteriores. No contexto desta dissertação, a ACP foi empregada como ferramenta auxiliar para analisar redundâncias e reduzir a dimensionalidade do conjunto de variáveis contínuas. Essa etapa foi motivada pela elevada correlação entre atributos meteorológicos derivados de diferentes defasagens temporais, o que poderia aumentar desnecessariamente a complexidade estrutural da RB e dificultar a identificação de dependências probabilísticas mais estáveis.

2.2.3 Teste Qui-Quadrado

O teste qui-quadrado χ^2 é um procedimento estatístico utilizado para verificar se a distribuição observada de uma variável categórica é compatível com uma distribuição esperada sob uma hipótese nula pré-definida. O teste avalia discrepâncias entre frequências observadas O_i e esperadas E_i, onde $E_i = Np_i$, sendo N o total de observações e p_i a proporção teórica associada à classe i. Essa abordagem permite avaliar se desvios entre observado e esperado podem ser atribuídos ao acaso ou indicam diferenças sistemáticas (BALAKRISHNAN; VOINOV; NIKULIN, 2013).

$$\chi^2 = \sum_{i=1}^{k} \frac{(O_i - E_i)^2}{E_i},$$

A qual, sob condições adequadas, segue aproximadamente uma distribuição qui-quadrado com $k - 1$ graus de liberdade (PEARSON, 1900). Essa abordagem permite avaliar se desvios entre observado e esperado podem ser atribuídos ao acaso ou indicam diferenças sistemáticas. Se o valor observado da estatística for maior que o valor crítico da distribuição qui-quadrado para o nível de significância adotado, rejeita-se a hipótese nula H_0 em favor da hipótese alternativa H_1.

De forma geral, a hipótese nula H_0 representa a suposição de que não há diferença significativa entre a distribuição observada e a distribuição teórica esperada , ou seja, quaisquer desvios são atribuídos ao acaso. Por outro lado, a hipótese alternativa H_1 indica que existe pelo menos uma discrepância significativa entre as frequências observadas e as esperadas, sugerindo que a variável categórica não segue a distribuição presumida (BALAKRISHNAN; VOINOV; NIKULIN, 2013).

2.2.4 Bootstrap

O método *bootstrap* é uma técnica estatística de reamostragem introduzida por (EFRON, 1992), amplamente utilizada para estimar a variabilidade de estatísticas em situações nas quais a distribuição amostral é desconhecida ou de difícil obtenção analítica. A abordagem consiste em gerar múltiplas amostras com reposição a partir da amostra original e, sobre cada uma dessas reamostragens, calcular a estatística de interesse, permitindo uma aproximação empírica de sua variabilidade. Trata-se, portanto, de um procedimento não paramétrico que fornece estimativas robustas mesmo em cenários com dados limitados.

No contexto das RBs, o *bootstrap* é utilizado como ferramenta para avaliar a estabilidade e a confiabilidade das estruturas inferidas. Como o processo de aprendizagem estrutural é sensível às variações nos dados, a reamostragem permite examinar a consistência das arestas, ou relações de dependência, obtidas a partir de diferentes subconjuntos da base de dados (FRIEDMAN; GOLDSZMIDT; WYNER, 2013). Dessa forma, torna-se possível identificar conexões estatisticamente mais robustas e distinguir relações espúrias.

Métodos de reamostragem, particularmente o *bootstrap*, têm sido aplicados para estimar a confiança nas conexões ao repetir o processo de aprendizado estrutural em múltiplos conjuntos de dados reamostrados. Em cada iteração, registra-se a presença ou ausência

de arcos e, ao final, calcula-se a frequência com que cada dependência aparece no conjunto das redes aprendidas. Estudos como o de (CARAVAGNA; RAMAZZOTTI, 2021) demonstram que esse procedimento pode melhorar a recuperação da estrutura verdadeira da rede, fornecendo informações relevantes sobre a robustez estrutural.

A necessidade de medidas de confiança associadas às características de uma RB torna-se especialmente evidente em situações em que a quantidade de dados disponível não é suficiente para induzir uma estrutura com alta pontuação. Entre as questões centrais destacam-se: a justificativa para a existência de uma aresta entre dois nós, a robustez do MB de um nó específico e a possibilidade de inferência sobre a ordenação das variáveis. O *bootstrap* oferece uma abordagem computacionalmente eficiente para responder a essas questões, conforme discutido em (FRIEDMAN; GOLDSZMIDT; WYNER, 2013). Além disso, as medidas de confiança resultantes podem ser utilizadas para induzir estruturas mais adequadas a partir dos dados e até mesmo para detectar a presença de variáveis latentes, contribuindo para análises mais robustas e interpretáveis.

No contexto desta dissertação, o *bootstrap* será utilizado para avaliar a estabilidade das estruturas aprendidas, permitindo identificar arcos recorrentes ao longo de múltiplas reamostragens. Essa estratégia é importante para reduzir a influência de flutuações específicas da amostra original e favorecer a seleção de dependências probabilísticas mais robustas.

2.2.5 Undersampling

A técnica de *undersampling* é amplamente utilizada para lidar com problemas de desbalanceamento de classes em conjuntos de dados, especialmente em tarefas de classificação supervisionada. O desbalanceamento ocorre quando a proporção entre as classes é significativamente desigual, o que pode levar modelos preditivos a apresentarem vieses em favor da classe majoritária, comprometendo o desempenho geral, principalmente em relação à classe minoritária (HE; GARCIA, 2009).

Essa técnica consiste na redução do número de instâncias da classe majoritária por meio da remoção aleatória ou seletiva de exemplos, de forma a equilibrar a distribuição entre as classes. Essa abordagem visa mitigar o impacto do desbalanceamento e possibilitar que o modelo aprenda padrões relevantes em ambas as classes de maneira mais eficaz. Contudo, o método apresenta limitações importantes. Ao descartar dados da classe majoritária, há o risco de perda de informações potencialmente relevantes (LIU; WU; ZHOU, 2008). Nesta dissertação, o *undersampling* foi adotado para mitigar o forte desbalan-

ceamento entre as classes associadas à ocorrência e à não ocorrência de precipitação no conjunto de treinamento.

2.2.6 Medidas de Avaliação dos Modelos

A avaliação de modelos baseados em RBs requer a aplicação de técnicas quantitativas e qualitativas capazes de mensurar tanto o desempenho preditivo quanto a qualidade do ajuste probabilístico (MARCOT, 2012). Dentre os métodos comumente empregados nesse contexto, destacam-se a validação cruzada, a análise da curva ROC (*Receiver Operating Characteristic*), as curvas de calibração, a função de perda logarítmica e a validação por especialistas. Esses instrumentos permitem avaliar diferentes aspectos do modelo, como acurácia, confiabilidade das previsões, capacidade discriminativa e aderência às expectativas do domínio aplicado.

2.2.6.1 Validação Cruzada K-Fold

A validação cruzada *k-fold* é uma técnica estatística amplamente utilizada para avaliar a capacidade de generalização de modelos preditivos. Nesse procedimento, o conjunto de dados é particionado em k subconjuntos aproximadamente do mesmo tamanho, denominados *folds*. Em cada iteração, um dos *folds* é utilizado como conjunto de teste, enquanto os demais servem para o treinamento do modelo. Esse processo é repetido k vezes, de forma que cada instância do conjunto de dados seja utilizada uma única vez para validação (HASTIE; TIBSHIRANI; FRIEDMAN, 2009).

No contexto desta dissertação, o modelo é avaliado conforme sua capacidade de classificar a ocorrência de chuva, isto é, identificar se haverá precipitação ($y = 1$) ou ausência de chuva ($y = 0$). Ao final de cada iteração da validação cruzada, os valores preditos são comparados aos valores observados no conjunto de teste, permitindo a construção da chamada matriz de confusão. Essa matriz organiza os resultados da classificação em quatro categorias: Verdadeiros Positivos (VP), quando o modelo prevê chuva e ela de fato ocorre; Verdadeiros Negativos (VN), quando o modelo prevê ausência de chuva e realmente não há; Falsos Positivos (FP), quando o modelo prevê chuva, mas ela não ocorre; e Falsos Negativos (FN), quando o modelo prevê ausência de chuva, embora ela tenha sido observada (MINING, 2006).

A estrutura formal da matriz de confusão pode ser representada conforme a Figura 3.

		Valor Predito	
		Sim	**Não**
Real	**Sim**	Verdadeiro Positivo (TP)	Falso Negativo (FN)
	Não	Falso Positivo (FP)	Verdadeiro Negativo (TN)

Figura 3: Matriz de confusão.

A partir das quantidades de VP, VN, FP e FN, são calculadas métricas agregadas ao longo das k iterações, como a acurácia, sensibilidade, especificidade e seus respectivos desvios padrão (MINING, 2006). A acurácia mede a proporção total de classificações corretas realizadas pelo modelo, sendo definida por:

$$Ac = \frac{VP + VN}{VP + VN + FP + FN}.$$

A sensibilidade, também denominada taxa de verdadeiros positivos, mede a proporção de observações classificadas como positivas que são, de fato, positivas, sendo dada por:

$$Sensibilidade = \frac{VP}{VP + FN}.$$

Já a especificidade refere-se à proporção de verdadeiros negativos corretamente classificados, indicando a eficácia do modelo em reconhecer casos negativos.

$$Especificidade = \frac{VN}{VN + FP}.$$

Estudos comparativos, como o de (KOHAVI et al., 1995), demonstram que a validação cruzada estratificada apresenta um equilíbrio adequado entre viés e variância, sendo particularmente eficaz na estimativa da acurácia e na seleção de modelos em tarefas de classificação supervisionada.

2.2.6.2 Função Perda de Log-Verossimilhança

A função de perda por log-verossimilhança, também conhecida como entropia negativa ou negentropia, corresponde à negação da log-verossimilhança esperada do conjunto de teste para uma RB ajustada com base no conjunto de treinamento (SCUTARI; SCUTARI;

MMPC, 2019). Essa função está fundamentada no estimador frequentista de máxima verossimilhança, segundo o qual os parâmetros w são estimados de modo a maximizar a função $p(D \mid w)$.

Na literatura de aprendizado de máquina, o logaritmo negativo da verossimilhança é comumente tratado como uma *função de erro*, cuja minimização é equivalente à maximização da verossimilhança (BISHOP; NASRABADI, 2006). Dessa forma, valores menores da função de perda indicam melhor desempenho do modelo, pois refletem maior concordância entre as probabilidades estimadas e os dados observados. Essa equivalência fornece uma base teórica robusta para o uso da perda logarítmica como medida de qualidade do ajuste probabilístico, sendo amplamente empregada em tarefas de aprendizagem supervisionada (MURPHY, 2012).

No contexto deste trabalho, a perda de log-verossimilhança complementa as métricas de classificação ao avaliar a qualidade probabilística das previsões produzidas pelas RBs. Sua inclusão é importante porque, em modelos probabilísticos, não basta classificar corretamente os eventos, sendo também necessário atribuir probabilidades coerentes aos desfechos observados.

2.2.6.3 Curvas ROC

As curvas ROC (*Receiver Operating Characteristic*) são amplamente utilizadas para avaliar o desempenho de modelos, especialmente em contextos probabilísticos e com desbalanceamento entre classes. Essas curvas representam a relação entre a taxa de verdadeiros positivos (sensibilidade) e a taxa de falsos positivos, para diferentes limiares de decisão do modelo (HASTIE; TIBSHIRANI; FRIEDMAN, 2009).

A métrica derivada mais comum associada à curva ROC é a Área Sob a Curva (AUC), que fornece uma medida agregada da capacidade do modelo em discriminar entre as classes. A AUC varia entre 0 e 1, sendo que valores mais próximos de 1 indicam maior capacidade discriminativa. Em termos interpretativos, a AUC representa a probabilidade de o modelo atribuir uma pontuação maior a uma instância positiva do que a uma instância negativa escolhida aleatoriamente (BRADLEY, 1997).

Segundo (BRADLEY, 1997), o uso da AUC é especialmente vantajoso por não depender de um ponto de corte específico e por ser robusto a distribuições desbalanceadas. Estudos mais recentes, como o de (ÇORBACIOĞLU; AKSEL, 2023), confirmam a utilidade da AUC como métrica padronizada para a comparação entre modelos probabilísticos

em diferentes domínios de aplicação.

2.2.6.4 Curvas de Calibração

As curvas de calibração são uma técnica que avaliam a confiabilidade das previsões probabilísticas geradas por modelos, como as RBs. Os gráficos gerados têm por objetivo comparar, para diferentes faixas de probabilidade prevista, a frequência observada de ocorrência do evento de interesse. Em outras palavras, avalia-se o quão bem as probabilidades atribuídas pelo modelo correspondem às proporções empíricas dos resultados reais (NICULESCU-MIZIL; CARUANA, 2005).

A calibração é uma propriedade distinta do desempenho preditivo, no sentido de que modelos com alta acurácia ou AUC ainda podem apresentar previsões mal calibradas. Segundo (VAN CALSTER et al., 2019), uma boa calibração implica que, entre todas as instâncias para as quais o modelo previu uma determinada probabilidade, a frequência observada do evento seja compatível com essa probabilidade. Assim, as curvas de calibração complementam métricas tradicionais como acurácia, sensibilidade e AUC, oferecendo uma análise mais abrangente da utilidade prática das probabilidades estimadas (HUANG et al., 2020).

Na prática, a curva ideal é representada por uma linha diagonal de 45 graus, indicando perfeita concordância entre as probabilidades previstas e as frequências observadas. Desvios em relação a essa linha sinalizam sobre ou subestimação dos riscos por parte do modelo, o que pode comprometer a interpretação probabilística das previsões (HUANG et al., 2020).

2.2.7 Teorema Central do Limite e Intervalos de Confiança

O Teorema Central do Limite é um dos resultados mais importantes da estatística, pois estabelece as condições sob as quais a distribuição da média amostral se aproxima de uma distribuição normal. Seja $X_1, X_2, \ldots, X_n$ uma sequência de variáveis aleatórias independentes e identicamente distribuídas (i.i.d.), com média μ e variância finita σ^2 (CASELLA; BERGER, 2002). Define-se a média amostral como:

$$\bar{X} = \frac{1}{n} \sum_{i=1}^{n} X_i.$$

O TCL afirma que, quando n é suficientemente grande, a distribuição da média amos-

tral $\bar{X}$ converge em distribuição para uma normal com média μ e variância σ^2/n, ou seja:

$$\bar{X} \xrightarrow{d} \mathcal{N}\left(\mu, \frac{\sigma^2}{n}\right), \quad \text{quando } n \to \infty.$$

Em outras palavras, mesmo que os dados não sigam uma distribuição normal, a distribuição da média tende à normalidade com o aumento do tamanho amostral (BUSSAB; MORETTIN, 2010). Esse comportamento permite o uso de ferramentas inferenciais baseadas na normalidade, como os intervalos de confiança e testes de hipóteses.

Um intervalo de confiança (IC) é um estimador intervalar construído a partir de uma estatística amostral, com o objetivo de conter o valor verdadeiro de um parâmetro populacional com um dado nível de confiança. Um IC de 95% para a média populacional indica que, em 95% das amostras aleatórias extraídas da população, o intervalo construído incluirá o valor real da média (BUSSAB; MORETTIN, 2010; CASELLA; BERGER, 2002; MONTGOMERY; RUNGER, 2010).

O IC para a média μ de uma população, quando a variância σ^2 é conhecida e os dados são aproximadamente normais, é dado por:

$$\bar{X} \pm z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}},$$

em que $z_{\alpha/2}$ representa o quantil da distribuição normal padrão associado ao nível de confiança desejado. Contudo, em situações nas quais a variância populacional é desconhecida, substitui-se σ pelo desvio padrão amostral, e o quantil da normal é substituído pelo quantil da distribuição t de Student com $n-1$ graus de liberdade (BUSSAB; MORETTIN, 2010; CASELLA; BERGER, 2002; MONTGOMERY; RUNGER, 2010).

$$\bar{X} \pm t_{(1-\alpha/2,\, n-1)} \cdot \frac{s}{\sqrt{n}}$$

3 Revisão Bibliográfica

A interdependência entre variáveis em uma RB é evidenciada por meio de distribuições de probabilidade conjunta, oferecendo transparência ao modelo e permitindo capturar as relações probabilísticas entre variáveis aleatórias (PEARL; GLYMOUR; JEWELL, 2016). Essa capacidade de modelar e interpretar essas relações tem impulsionado o desenvolvimento de diversos estudos que utilizam esses modelos para identificar relações entre variáveis por trás de diferentes tipos de eventos.

Os artigos discutidos nesta Seção têm como objetivo empregar RBs na modelagem de uma variedade de fenômenos, buscando não apenas compreendê-los, mas também prever suas ocorrências. Além disso, serão referenciados trabalhos que investigaram diferentes métodos para o aprendizado estrutural das RBs, contribuindo para aprimorar a precisão e a eficácia desses modelos em diversas aplicações.

3.1 Redes Bayesianas: aplicações gerais e fundamentos

Contaldi (2016) concentrou sua tese em fornecer um método para aprendizagem da estrutura das RBs a partir de conjuntos de dados limitados. Especificamente, o autor destacou o desenvolvimento de um algoritmo híbrido de aprendizado da estrutura da RB com o objetivo de reduzir o espaço de busca de forma confiável, aproveitando informações derivadas dos dados. A abordagem envolveu a utilização de uma fase inicial de busca condicional para restringir o escopo da busca, seguida pela aplicação de um algoritmo genético parametrizado. Concluiu-se que a proposta desse algoritmo híbrido representou um avanço significativo na área, destacando-se sua capacidade de reduzir o espaço de busca de maneira confiável e explorá-lo exaustivamente, aproveitando as informações disponíveis nos dados.

Kaikkonen et al. (2021) fizeram uma revisão bibliográfica sobre o uso das RBs na avaliação de riscos ambientais (ARA), onde essa avaliação é um processo de estimativa da probabilidade e das consequências dos efeitos adversos das atividades humanas e de

outros fatores de estresse no meio ambiente. Foi observado que as vantagens mais comuns relacionadas ao uso de RBs em ARA incluem o tratamento explícito da incerteza, a capacidade das RBs de integrar conhecimento de diferentes fontes e os meios para atualizar facilmente os modelos à medida que novos conhecimentos se tornam disponíveis.

Machado et al. (2021) adaptaram o Algoritmo Genético Multi-Agente (MAGA) para o aprendizado estrutural de RBs e o comparou com os algoritmos utilizados na literatura, como ABC, PSO-BNC, SAGA, GES e GA. Os resultados mostraram que o PSO-BNC teve desempenho insatisfatório devido ao tempo computacional elevado, especialmente em grandes conjuntos de dados. O GES apresentou resultados melhores em instâncias com menos dados e tempo computacional mais baixo. O SAGA, ABC e os MAGA foram os que se destacaram em grandes conjuntos de dados, tanto em termos de tempo computacional quanto de convergência.

Ozelame et al. (2021) propõem uma metodologia híbrida para a estimação de estruturas em RBs com foco na previsão de falhas em plantios de cana-de-açúcar. Essa metodologia, denominada *scoring and restrict*, combina dois tipos de algoritmos para encontrar as estruturas da rede. Neste algoritmo, a estrutura que maximiza a adequação aos dados é identificada, seguida pela inclusão restritiva da informação relacionada à variável de interesse, baseada em testes de independência condicional. O estudo utiliza dados de falhas em talhões de plantio de cana-de-açúcar, considerando variáveis climáticas, do solo, práticas agrícolas e a variável de interesse e o percentual de falha. Os resultados mostraram que a metodologia aplicada resultou em melhorias nas medidas de adequação e interpretação do modelo.

Oliveira Barth (2023) estudou uma metodologia Bayesiana orientada a dados para o aprendizado estrutural das RBs, aplicado na análise de sinais de eletromiografia e cinemática, utilizando o algoritmo baseado em Monte Carlo via Cadeias de Markov (MCMC). A conclusão destacou que o método proposto apresentou desempenho semelhante a algoritmos tradicionais, como o Hill Climb. Além disso, os resultados sugerem que a metodologia pode ser útil para aprender modelos de RBs em conjuntos de dados não homogêneos e com poucos dados disponíveis.

Bates e Dowdy (2024) analisaram a influência de modos de variabilidade climática, características meteorológicas regionais e persistência temporal sobre a vazão fluvial na costa leste da Austrália. Utilizando RBs, e comparando-as à regressão linear múltipla, os autores identificaram relações diretas e indiretas entre variáveis atmosféricas e o escoamento. As redes mostraram-se transparentes e interpretáveis, revelando dependências

hierárquicas consistentes com a física do sistema. Embora a regressão tenha apresentado ajuste ligeiramente melhor, as RBs forneceram maior capacidade explicativa e suporte à análise da variabilidade hidrológica.

Natalia et al. (2025) utilizaram das RBs para analisar incidentes de terrorismo na aviação. A análise evidenciou a utilidade das RBs na avaliação de riscos complexos e interdependentes, permitindo identificar padrões probabilísticos relevantes para políticas de segurança e resposta a emergências no setor da aviação.

Yu et al. (2025) desenvolveram e validaram um modelo, baseado em RBs, de triagem digital para transtornos neurocognitivos (NCDs). O modelo identificou oito fatores diretamente ligados aos NCDs, tais como doenças cerebrovasculares, traumatismo craniano e distúrbios afetivos. O algoritmo de busca Tabu, guiado pelo critério BIC, gerou uma estrutura interpretável e robusta. Comparado à regressão logística, o modelo apresentou melhor calibração e resistência a dados ausentes. Os autores concluíram que a abordagem em RB é eficaz, interpretável e escalável para triagem precoce de NCDs em grandes sistemas de saúde eletrônicos.

3.2 Aplicações de Redes Bayesianas à Modelagem da Precipitação

Cofino et al. (2002) introduzem as RBs para modelar as dependências espaciais e temporais entre diferentes estações meteorológicas entre 100 estações na bacia Norte da Península Ibérica durante o inverno de 1999. A conclusão destacou como tais modelos podem ser construídos e os tipos de inferências que podem ser realizadas sobre elas.

Cano, Sordo e Gutiérrez (2004) exploraram o uso de RBs, sob a perspectiva da mineração de dados, na análise de dados meteorológicos e na previsão do tempo. O banco de dados utilizado tinha informações sobre precipitação diária e velocidade máxima do vento em 100 estações na Península Ibérica e com os correspondentes padrões atmosféricos, abrangendo o período de 1959 a 2000. O autor elaborou o Grafo Acíclico Direcionado da RB utilizando o algoritmo K2 tradicional, bem como uma adaptação local do algoritmo K2 que considerava as dependências espaciais presentes no problema. A conclusão do trabalho destacou a eficácia das RBs na modelagem e previsão de eventos meteorológicos, demonstrando sua utilidade em lidar com a complexidade e as interdependências dos dados atmosféricos.

Garrote, Molina e Mediero (2007) utilizaram as RBs em conjunto com modelos de-

terminísticos para previsões probabilísticas de precipitação, visando aprimorar a precisão das previsões de enchentes. A metodologia empregada envolveu a calibração dos modelos Bayesianos com conjuntos de dados gerados por simulação Monte Carlo a partir de um modelo determinístico de precipitação-escoamento. Os resultados evidenciam a capacidade das RBs em produzir distribuições de probabilidade que reproduzem o comportamento do modelo determinístico. Concluiu-se que a abordagem proposta é eficaz para realizar previsões em tempo real na região do Mediterrâneo espanhol.

Sharma e Goyal (2016) tiveram como objetivo utilizar um modelo baseado em RBs para prever a precipitação mensal em 21 estações em Assam, Índia, empregando uma série histórica de dados de 1981 a 2000. Foram utilizadas duas técnicas de aprendizado de estrutura, K2 e algoritmo Markov Chain Monte Carlo (MCMC), para desenvolver o modelo de previsão. Os resultados mostraram que os modelos baseados em RB foram capazes de prever a precipitação mensal com precisão; a comparação entre as técnicas K2 e MCMC demonstrou que o algoritmo K2 superou o MCMC em todas as estações, indicando sua eficácia na previsão de chuva.

Das e Ghosh (2019) propõem um novo framework, chamado FB-STEP, para previsão espacial e temporal de séries temporais climatológicas, utilizando uma abordagem baseada em RBs e análise multifractal. O autor comparou o modelo proposto com métodos estatísticos tradicionais, modelos baseados em processos vetoriais e outras abordagens não lineares na predição das condições climáticas de cinco cidades da Índia, utilizando dados históricos de temperatura, umidade, taxa de precipitação e umidade do solo. Foi observado que o FB-STEP foi superior às outras técnicas no caso testado, produzindo o mínimo erro e menor incerteza na previsão de cada variável considerada. Essa abordagem foi capaz de lidar com a natureza caótica inerente dos dados climatológicos, evidenciando sua eficácia diante da complexidade dos fenômenos climáticos.

Das e Chanda (2020) aplicaram as RBs para modelar a estrutura de independência condicional entre a precipitação mensal e diversas variáveis locais, tais como temperatura do ar, umidade relativa, direção e intensidade do vento, umidade do solo, entre outras, ao longo de um período de seis meses em Gangestic West Bengal na Índia. Para o aprendizado estrutural das RBs o autor usou o algoritmo Hill-Climb e Max-Min Hill. O estudo concluiu que as RBs são eficazes para a seleção de variáveis e para a redução da complexidade do modelo, fornecendo um método eficiente para a previsão de chuvas. Os modelos propostos tiveram um desempenho comparável aos resultados de outros estudos de previsão de chuva na mesma região.

Putri e Wijayanto (2021) utilizaram as RBs para modelar os padrões de chuva e os relacionamentos entre as variáveis preditoras em Jacarta, Indonésia. Eles realizaram previsões sobre a ocorrência de chuva na região com base em dados de parâmetros atmosféricos obtidos do satélite NOAA. Os modelos com melhor desempenho foram os construídos utilizando os algoritmos Hill Climbing com pontuação BIC e BDE, algoritmos MMHC, RSMAX2 e H2PC para o aprendizado estrutural das RBs. Além disso, a estrutura do DAG revelou que a ocorrência de chuva em Jacarta é diretamente afetada pelas variáveis de umidade relativa e cobertura de nuvens. Verificou-se também que a variável temperatura afeta indiretamente a ocorrência desses fenômenos, enquanto a variável de pressão do nível médio do mar não exerce influência direta ou indireta sobre a ocorrência de chuva.

3.3 Oportunidade de pesquisa

A literatura revisada evidencia que as RBs têm sido empregadas na modelagem de incertezas e na representação de dependências probabilísticas em diferentes domínios, incluindo avaliação de riscos ambientais, diagnóstico médico, previsão de falhas e aplicações meteorológicas. No campo da precipitação, os estudos analisados mostram o potencial dessas redes para descrever relações entre variáveis atmosféricas e apoiar tarefas de previsão.

Entretanto, a análise conjunta desses trabalhos também revela limitações recorrentes. Em termos de dados, observa-se predominância de estudos conduzidos com séries temporais agregadas, especialmente em escalas diária ou mensal, ou ainda com recortes temporais curtos. Em termos metodológicos, parte relevante da literatura compara apenas um número reduzido de algoritmos de aprendizado estrutural. Além disso, muitos trabalhos privilegiam o desempenho preditivo, com menor aprofundamento da interpretação física das dependências aprendidas e de sua coerência com o comportamento atmosférico observado.

Conforme sintetizado na Tabela 1, embora os estudos revisados demonstrem a utilidade das RBs na modelagem meteorológica, permanece pouco explorada uma abordagem que combine, de forma simultânea, dados observacionais locais com resolução temporal horária, comparação sistemática entre diferentes famílias de algoritmos de aprendizado estrutural e incorporação explícita de restrições. Em particular, verifica-se que a literatura ainda carece de investigações voltadas à construção de redes que, além de preditivamente úteis, sejam também estruturalmente interpretáveis e coerentes com o conhecimento meteorológico do domínio analisado.

Tabela 1: Síntese comparativa dos estudos revisados

Estudo	Dados	Abordagem	Limitações e contribuições
Cofino et al. (2002)	Península Ibérica; 100 estações; precipitação diária; 1979–1993.	RB para dependências espaço-temporais; precipitação em três classes; K2.	Mostra a viabilidade de RBs para chuva. Contudo, tem caráter preliminar, usa apenas um algoritmo estrutural e trabalha em escala diária.
Cano, Sordo e Gutiérrez (2004)	Península Ibérica; dados diários de precipitação; 100 estações; 1959–2000; dados ECMWF/ERA-15.	RBs para mineração de dados meteorológicos e previsão; precipitação em quatro classes; K2 e LK2.	Evidencia a utilidade das RBs em problemas meteorológicos. Contudo, usa escala diária e utiliza apenas dois algoritmos de aprendizado estrutural
Garrote, Molina e Mediero (2007)	Bacias mediterrâneas da Espanha; dados sintéticos gerados por Monte Carlo.	RBs combinadas a modelos determinísticos para previsão probabilística de cheias; variáveis discretizadas.	RBs e modelos determinísticos para representar incerteza. Contudo, depende de simulação, não foca na precipitação observada e não compara algoritmos estruturais.
Sharma e Goyal (2016)	21 estações em Assam, Índia; série de 1981–2000; uso de variáveis atmosféricas locais.	RB para previsão mensal de chuva; comparação entre K2 e MCMC.	Indica melhor desempenho do K2. Contudo, restringe-se à escala mensal, compara apenas dois algoritmos e privilegia o desempenho preditivo em detrimento da interpretabilidade física.
Das e Ghosh (2019)	Kolkata, Raipur e Lucknow, Índia; dados diários de temperatura, umidade relativa, precipitação e umidade do solo; 2001–2016.	*Fuzzy Bayesian network* com análise multifractal; incorporação explícita de atributos espaço-temporais.	Incorpora dependências espaço-temporais e compara o desempenho com outros métodos. Contudo, não compara algoritmos estruturais, adota estrutura pré-definida e foca na predição contínua.
Das e Chanda (2020)	Gangetic West Bengal, Índia; dados mensais de precipitação e variáveis meteorológicas; 1980–2016; 72 preditores com defasagens.	RB com dois algoritmos estruturais; comparação com outros modelos.	Explora seleção de preditores via RB. Contudo, restringe-se à escala mensal, compara apenas dois algoritmos e privilegia o desempenho preditivo em detrimento da interpretação física.
Putri e Wijayanto (2021)	Jacarta, Indonésia; dados NOAA/CFSR a cada 6 horas; 2020–2021; variáveis meteorológicas e precipitação.	RBs para previsão de chuva; HC, MMHC, RSMAX2 e H2PC; comparação de critérios de pontuação; K-Means.	Compara múltiplos algoritmos e critérios. Contudo, restringe-se a uma série curta, a uma única área, à precipitação binária e não aprofunda a interpretação da estrutura aprendida.

Adicionalmente, observa-se a escassez de estudos aplicados a contextos urbanos brasileiros, especialmente em regiões tropicais marcadas por elevada variabilidade da precipitação, como o município do Rio de Janeiro. Essa lacuna é relevante, pois a complexidade atmosférica local, associada a fatores geográficos e urbanos, impõe desafios específicos que não são plenamente contemplados por estudos desenvolvidos em outros contextos climáticos ou com bases de dados mais agregadas.

Nesse sentido, a presente pesquisa busca contribuir para o avanço da literatura ao investigar se a modelagem por RBs, orientada por dados e guiada por restrições estruturais, é capaz de representar, de forma interpretável, as relações condicionais entre variáveis atmosféricas e a ocorrência de precipitação no município do Rio de Janeiro.

4 Validação Metodológica

A primeira etapa desta pesquisa de dissertação consistiu na reprodução dos resultados apresentados por (PUTRI; WIJAYANTO, 2021). Esse estudo foi selecionado como referência por empregar RBs na modelagem probabilística da precipitação a partir de variáveis atmosféricas, utilizando um conjunto de dados meteorológicos derivados de sensoriamento remoto e um procedimento de aprendizado estrutural comparável ao adotado nesta pesquisa. A reprodução teve como objetivo validar o conhecimento adquirido e aprofundar a compreensão sobre abordagens previamente aplicadas à modelagem de precipitação com RBs.

4.1 Aquisição e preparação dos dados

No estudo de referência, os autores buscaram modelar os padrões de chuva e os relacionamentos probabilísticos entre variáveis atmosféricas na cidade de Jacarta, Indonésia, realizando previsões sobre a ocorrência de precipitação com base em dados obtidos por sensoriamento remoto. Os dados obtidos foram extraídos do sistema NOAA CFSR, considerando o período de 1º de janeiro de 2020 a 21 de abril de 2021 para a construção do modelo, e de 22 de abril de 2021 a 13 de maio de 2021 para avaliação da ocorrência de chuvas.

Cabe destacar, contudo, que o contexto empírico deste estudo difere daquele adotado nesta dissertação. Em Jacarta, foram utilizados dados derivados de sensoriamento remoto, agregados espacialmente sobre uma área retangular representativa da cidade, com resolução temporal de 6 horas e série relativamente curta. Em contraste, nesta pesquisa, empregam-se dados observacionais horários provenientes de estações meteorológicas automáticas do INMET, abrangendo um período mais longo e associados ao contexto urbano do município do Rio de Janeiro. Além das diferenças de fonte, suporte espacial e extensão temporal, o caso do Rio de Janeiro envolve condicionantes geográficos e urbanos que contribuem para a elevada variabilidade da precipitação. Assim, a reprodução do estudo

de Jacarta deve ser compreendida como uma etapa de validação metodológica do pipeline de modelagem, e não como uma expectativa de equivalência direta entre as estruturas probabilísticas aprendidas em contextos meteorológicos distintos.

A ausência das coordenadas geográficas no estudo de referência exigiu a definição de um retângulo geográfico que representa a área da cidade. Para essa delimitação, foi utilizada a plataforma Google Earth. As coordenadas escolhidas estão apresentadas na Tabela 2.

Tabela 2: Coordenadas Geográficas - Jacarta.

Descrição	**Graus decimais**
Latitude Norte	-6.1546
Latitude Sul	-6.2146
Longitude Leste	106.9356
Longitude Oeste	106.8456

Os dados foram extraídos via *JavaScript*, por meio do ambiente Earth Engine Code Editor. As imagens do conjunto de dados NOAA CFSR possuem resolução temporal de 6 horas. Para fins de análise, foi realizada a média espacial das variáveis explicativas em cada imagem, enquanto a precipitação foi obtida pelo somatório espacial do acumulado sobre a área delimitada, para cada instante temporal. O código utilizado para a extração dos dados está disponível em repositório público e pode ser acessado por meio deste link. Os dados extraídos podem ser visualizados na Tabela 3.

As variáveis consideradas na extração foram: Temperatura, que representa a temperatura na superfície; Umidade Relativa, que representa a umidade relativa da atmosfera em camada única; Pressão ao Nível do Mar, que representa a pressão atmosférica reduzida ao nível médio do mar; Cobertura de Nuvens, que representa o conteúdo de água de nuvem em toda a atmosfera; e Precipitação, que representa a precipitação total acumulada na superfície.

Tabela 3: Dados extraídos do NOAA - Jacarta.

Variáveis	**Abreviações**	**Unidade de Medida**
Temperatura	temp	K
Umidade Relativa	rh	%
Pressão ao nível do mar	pmsl	Pa
Cobertura de nuvens	cc	kg/m^2
Precipitação	Precip	kg/m^2

Para a discretização das variáveis contínuas, o autor empregou o método K-Means utilizando a distância euclidiana como métrica [1]. Contudo, os valores dos centroides obtidos no agrupamento não foram apresentados no trabalho original. Diante disso, nesta dissertação, procedeu-se ao cálculo desses valores, os quais estão detalhados na Tabela 4. Quanto à variável precipitação, esta foi binarizada, atribuindo-se o valor 1 para a ocorrência de chuva e 0 para a sua ausência.

Tabela 4: Centroides da discretização das variáveis contínuas – NOAA (Jacarta).

Variáveis	**Categoria**	**Valor do Centroide**
Temperatura	Alta	30,29
	Baixa	28,56
Umidade Relativa	Alta	62,69
	Moderada	54,18
	Baixa	42,02
Pressão Atmosférica	Alta	101162,74
	Moderada	100969,98
	Baixa	100757,73
Cobertura de nuvens	Muito Alta	4,09
	Alta	1,21
	Baixa	0,37
	Muito Baixa	0,03

4.2 Reprodução e Comparação dos Modelos Estruturais

Para a construção das estruturas das RBs, foram empregadas diferentes abordagens de aprendizado estrutural, abrangendo métodos baseados em pontuação e algoritmos híbridos. No grupo dos métodos baseados em pontuação, utilizou-se o HC com as seguintes

[1] Mais detalhes sobre essa técnica podem ser encontrados na Seção 2.2.1.

funções de avaliação, BIC, K2, Log-Likelihood, BDe e AIC. Entre os métodos *hybrid-based*, foram utilizados os algoritmos MMHC, H2PC e RSMAX2.

Na Figura 4 é apresentada uma comparação entre as DAGs obtidas (à esquerda) e aquelas reportadas por (PUTRI; WIJAYANTO, 2021), a partir da aplicação dos algoritmos de aprendizado estrutural HC-BIC, HC-BDe, MMHC, RSMAX2 e H2PC. Essas DAGs representam as relações probabilísticas entre as variáveis atmosféricas utilizadas para a modelagem da precipitação, permitindo visualizar as dependências aprendidas entre os nós.

Observa-se que, embora as duas estruturas compartilhem conexões semelhantes, como a influência da temperatura sobre a umidade relativa, há uma diferença importante entre os grafos. No modelo desenvolvido por (PUTRI; WIJAYANTO, 2021), a variável de precipitação é diretamente influenciada pela umidade relativa e pela cobertura de nuvens. Já no modelo gerado nesta pesquisa, a variável precipitação é impactada diretamente somente pela umidade relativa. Contudo, é importante destacar que variáveis como temperatura e cobertura de nuvens mantêm influência indireta sobre a precipitação. A temperatura afeta a precipitação por meio de sua relação com a umidade relativa, estabelecendo uma cadeia de dependência. Da mesma forma, a cobertura de nuvens exerce influência indireta via umidade relativa.

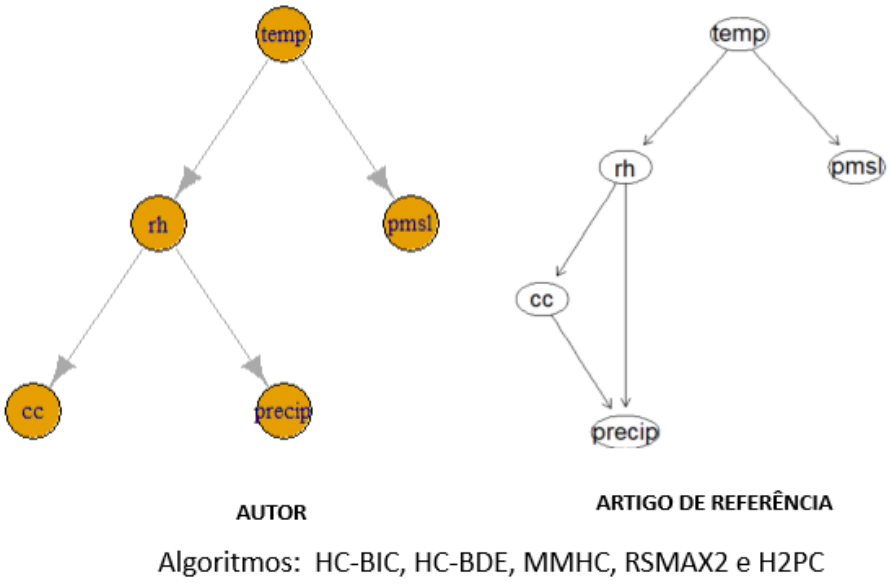

Figura 4: Comparação entre as DAGs obtidas nesta pesquisa (à esquerda) e aquelas reportadas por Putri e Wijayanto (2021) (à direita), considerando os algoritmos HC-BIC, HC-BDe, MMHC, RSMAX2 e H2PC.

A Figura 5 ilustra as diferenças entre as estruturas geradas pelo algoritmo HC-K2 na presente pesquisa e no estudo de referência. Verifica-se que, no modelo desenvolvido neste trabalho, a variável precipitação influencia diretamente a umidade relativa e a pressão ao nível do mar, enquanto no modelo de referência, a precipitação é impactada diretamente pela umidade relativa e pela cobertura de nuvens.

Essas diferenças mostram que a reprodução não resultou em uma estrutura idêntica à original, apesar do uso do mesmo algoritmo e abordagem metodológica. Esse desvio pode estar associado à ausência de informações metodológicas detalhadas no estudo de referência, como os valores exatos adotados no processo de discretização das variáveis, a não especificação das coordenadas geográficas utilizadas para a extração dos dados referentes a Jacarta, bem como à presença de componentes aleatórias inerentes a determinadas etapas do aprendizado estrutural.

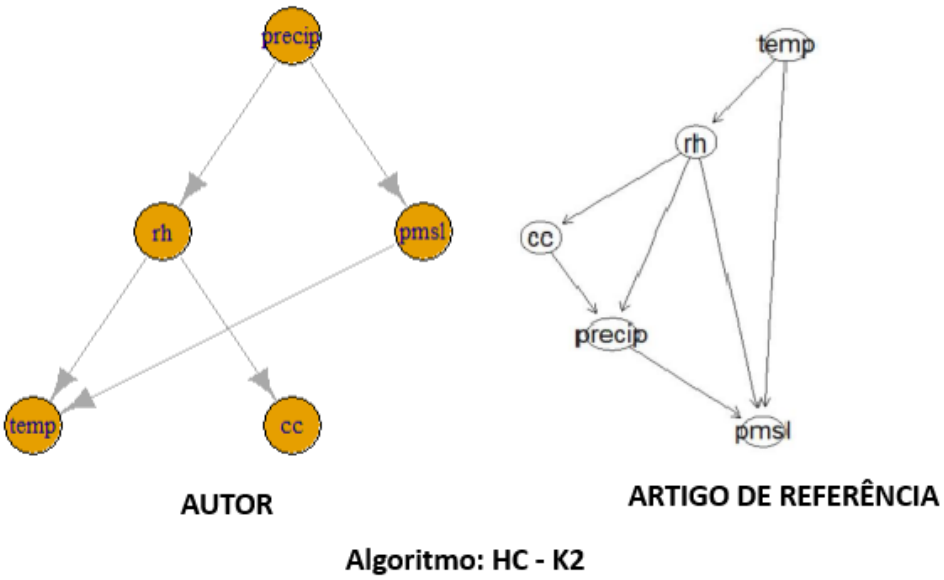

Figura 5: Comparação entre as DAGs obtidas nesta pesquisa (à esquerda) e aquelas reportadas por Putri e Wijayanto (2021) (à direita), considerando o algoritmo HC-K2.

O mesmo problema de reprodutibilidade ocorre ao se utilizar o algoritmo HC-LL, como mostrado pela Figura 6. No modelo gerado nesta pesquisa, observa-se uma estrutura concentrada, em que todas as variáveis atmosféricas são diretamente influenciadas pela temperatura. Já no modelo apresentado no artigo de referência, a rede de dependências é mais distribuída, evidenciando interações diretas e indiretas entre umidade relativa, cobertura de nuvens, pressão ao nível do mar e precipitação. Essas diferenças reforçam os desafios de reprodutibilidade enfrentados, os quais podem estar associados à falta de informações metodológicas completas no estudo de referência.

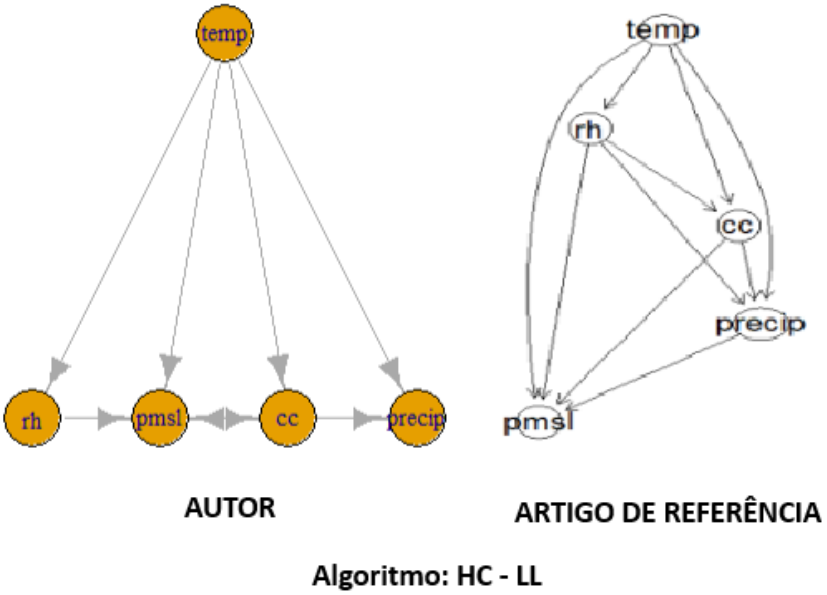

Figura 6: Comparação entre as DAGs obtidas nesta pesquisa (à esquerda) e aquelas reportadas por Putri e Wijayanto (2021) (à direita), considerando o algoritmo HC-LL.

A Figura 7 ilustra, mais uma vez, a inconsistência entre os resultados obtidos nesta pesquisa e aqueles apresentados por (PUTRI; WIJAYANTO, 2021), agora com a aplicação do algoritmo HC-AIC. No modelo reproduzido, a precipitação é influenciada diretamente pela umidade relativa e pressão ao nível do mar. Em contraste, no grafo do estudo de referência, a precipitação recebe influência direta da umidade relativa e cobertura de nuvens.

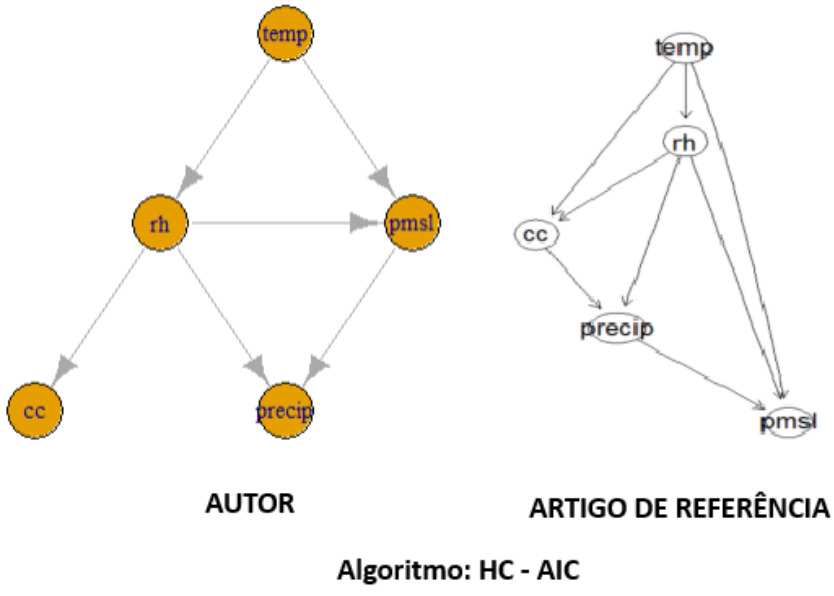

Figura 7: Comparação entre as DAGs obtidas nesta pesquisa (à esquerda) e aquelas reportadas por Putri e Wijayanto (2021) (à direita), considerando o algoritmo (HC-AIC).

De forma geral, as análises realizadas evidenciam limitações significativas no processo de reprodutibilidade dos modelos estruturais apresentados, mesmo diante da adoção dos mesmos algoritmos e diretrizes metodológicas indicadas no estudo original. Embora algumas relações esperadas tenham sido parcialmente preservadas, as diferenças estruturais entre os grafos obtidos sugerem a existência de lacunas na documentação metodológica do trabalho de referência, comprometendo a fidelidade da replicação.

Para obter modelos estruturalmente mais coerentes com a dinâmica física do fenômeno, algumas medidas podem ser adotadas. Em primeiro lugar, pode-se utilizar restrições estruturais baseadas em conhecimento de domínio, de modo a impedir direções de arcos incompatíveis com a ordem temporal das variáveis ou com interpretações meteorológicas pouco plausíveis. Esse tipo de estratégia contribui para reduzir a ocorrência de dependências estatisticamente ajustadas aos dados, mas inconsistentes do ponto de vista físico, favorecendo a aprendizagem de estruturas mais interpretáveis e compatíveis com os mecanismos atmosféricos envolvidos.

4.3 Comparação do Desempenho dos Modelos

Nesta etapa, optou-se por comparar apenas os modelos cujas DAGs apresentaram maior semelhança com aquelas reportadas. Foram considerados, portanto, os algoritmos HC-BIC, HC-BDE, MMHC, RSMAX2 e H2PC. A semelhança entre os resultados pode ser observada na Figura 4. Os demais algoritmos produziram grafos significativamente distintos, inviabilizando uma comparação justa de desempenho preditivo com base nas mesmas premissas estruturais.

A Tabela 5 apresenta a comparação das métricas de desempenho obtidas no estudo de referência e na reprodução. No estudo original, a acurácia reportada foi de 81,18%, enquanto na reprodução obteve-se 73,81%. Essa diferença de aproximadamente 7,4 pontos percentuais indica uma redução significativa na capacidade geral do modelo reproduzido em acertar as classificações.

O modelo apresentado obteve uma sensibilidade de 78,26%, enquanto o modelo reproduzido alcançou 73,47%. Embora essa diferença seja menos acentuada do que a observada na acurácia, ela indica uma redução na capacidade do modelo reproduzido em identificar corretamente os eventos de precipitação.

A especificidade, por sua vez, foi a métrica com maior discrepância observada entre os resultados, enquanto o estudo de referência alcançou uma especificidade de 84,62%,

o modelo reproduzido obteve apenas 74,29%. Essa diferença de mais de 10 pontos percentuais indica uma redução significativa na habilidade do modelo reproduzido em evitar falsos positivos, o que implica em uma maior tendência a prever chuva em situações onde ela não ocorre.

Tabela 5: Desempenho preditivo dos modelos com estruturas similares às do estudo de referência (HC-BIC, HC-BDE, MMHC, RSMAX2 e H2PC).

Modelo	**Acurácia(%)**	**Sensibilidade (%)**	**Especificidade (%)**
Referência	81,18	78,26	84,62
Reprodução	73,81	73,47	74,29

Os resultados obtidos evidenciam uma perda de desempenho na reprodução do modelo original, com impacto em todas as métricas avaliadas. A redução da acurácia indica que o modelo reproduzido apresenta menor capacidade geral de classificação correta, o que compromete sua efetividade como ferramenta preditiva. Ainda que a sensibilidade tenha apresentado uma diferença menos acentuada, a queda nessa métrica revela uma limitação importante na detecção de eventos de precipitação. A discrepância mais severa, no entanto, ocorreu na especificidade, cuja queda de mais de 10 pontos percentuais indica que o modelo reproduzido tem dificuldade acentuada em distinguir corretamente os dias sem chuva.

4.4 Reprodutibilidade das Inferências

Além da análise estrutural das RBs e da avaliação do desempenho preditivo, esta pesquisa também se dedicou à reprodução das inferências probabilísticas apresentadas por (PUTRI; WIJAYANTO, 2021). O objetivo desta etapa foi verificar a consistência das probabilidades condicionais estimadas nos dois estudos.

A Tabela 6 apresenta uma comparação entre as probabilidades de ocorrência de chuva sob diferentes combinações de condições atmosféricas. As estimativas do estudo de referência são comparadas com aquelas obtidas na presente pesquisa, que incluem também o desvio-padrão e o intervalo de confiança (95%) para cada inferência.

No caso da condição “temperatura alta”, a probabilidade de ocorrência de chuva estimada nesta pesquisa (65,8%) foi superior àquela reportada por (PUTRI; WIJAYANTO, 2021) (57,9%). Essa diferença, embora relativamente pequena, é estatisticamente significativa, como pode ser observado pelo intervalo de confiança estreito (65,69% a 66,08%) e pelo baixo desvio-padrão associado (0,007).

Para a condição combinada de “temperatura alta e umidade relativa baixa”, também se observou um aumento na estimativa da probabilidade de chuva, de 16,2% no estudo original para 19,9% na reprodução. Embora a reprodução tenha indicado aumento da probabilidade estimada de chuva na condição combinada de temperatura alta e umidade relativa baixa, tal configuração não corresponde, em termos físicos, a um cenário classicamente favorável à precipitação. Esse resultado deve, portanto, ser interpretado com cautela, podendo refletir particularidades da discretização, da estrutura aprendida.

A maior discrepância foi observada na combinação “temperatura alta e cobertura de nuvens alta”, cuja probabilidade de precipitação passou de 81,3% no estudo de referência para 90,9% nesta pesquisa. O intervalo de confiança mais amplo (85,55% a 96,16%) e o maior desvio-padrão (0,190) indicam uma maior variabilidade nessa estimativa. Por fim, na condição “temperatura baixa e umidade relativa baixa”, a reprodução apresentou uma estimativa ligeiramente inferior (13,3%).

Tabela 6: Comparação das inferências probabilísticas sobre a ocorrência de chuva entre o estudo de referência e a reprodução.

Cenário	**Referência**	**Reprodução**	**DP**	**IC 95% (%)**
Temp. Alta	57,99%	65,8%	0,007	[65,69 ; 66,08]
Temp. Alta e UR Baixa	16,2%.	19,9%	0,01	[19,57 ; 20,41]
Temp. Alta e Cob. de Nuvens Alta	81,3%.	90,86%	0,19	[85,55 ; 96,16]
Temp. Baixa e UR Baixa	14,2%.	13,3%	0,008	[13,07 ; 13,53]

Ainda que algumas estimativas tenham se aproximado, divergências significativas foram observadas em certos cenários. Tais discrepâncias podem ser atribuídas a diferenças na discretização das variáveis, na amostragem dos dados ou na estrutura aprendida pelas redes. A inclusão dos intervalos de confiança, ausentes no estudo original, permitiu não apenas quantificar a precisão das estimativas obtidas, mas também demonstrar a variabilidade inerente a cada inferência.

5 Design Experimental e Metodologia

Esta Seção descreve o arcabouço metodológico adotado na pesquisa. O *pipeline* proposto foi concebido para analisar se a modelagem por RBs é capaz de representar, de forma interpretável, as relações probabilísticas entre variáveis atmosféricas e a ocorrência de precipitação no município do Rio de Janeiro.

A metodologia abrange a aquisição dos dados observacionais, o pré-processamento e a avaliação do conjunto final, incluindo a comparação de abordagens de aprendizado estrutural em RBs, a fim de verificar sua capacidade de representar essas relações e seu desempenho preditivo em relação à ocorrência de eventos de precipitação.

Adicionalmente, foi realizada a inferência probabilística, permitindo mensurar a influência das variáveis meteorológicas na ocorrência de chuva, seguida de validação por especialistas em Meteorologia. Por fim, com base nos tempos de treinamento e na arquitetura computacional empregada, foram quantificados o consumo de energia, as emissões equivalentes de CO_2 e a pegada hídrica, a fim de conduzir uma avaliação de sustentabilidade ambiental dos algoritmos.

A Figura 8 apresenta uma síntese do *pipeline* metodológico deste estudo, detalhado nas seções subsequentes.

5.1 Coleta e Pré-Processamento dos Dados

Neste estudo, foram utilizadas informações provenientes de estações meteorológicas automáticas. Esta Seção apresenta a origem desses dados, bem como suas principais características e os procedimentos de pré-processamento utilizados.

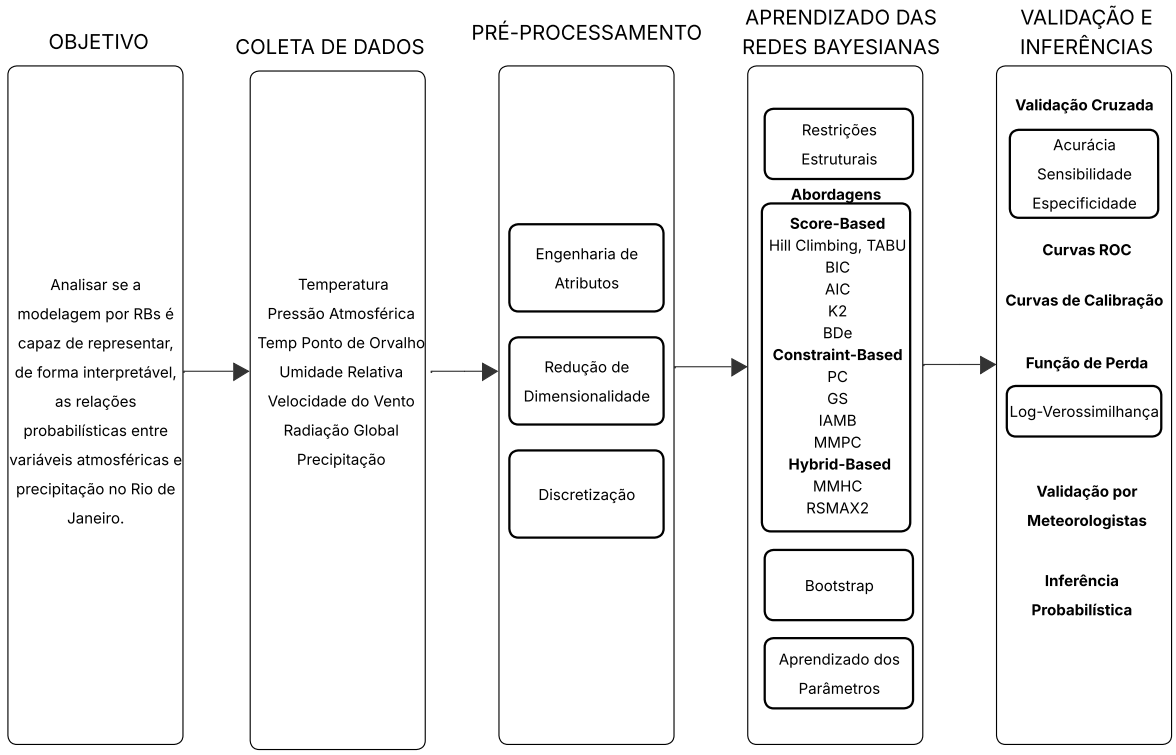

Figura 8: Fluxo Metodológico.

5.1.1 Aquisição, Descrição e Integração dos Dados

O Instituto Nacional de Meteorologia (INMET) é um órgão federal, vinculado ao Ministério da Agricultura e Pecuária, com mais de um século de história, tendo sido criado em 1909. Suas principais atribuições englobam o monitoramento e a previsão do tempo e do clima, a emissão de alertas de eventos meteorológicos severos e a produção de diversos recursos e pesquisas em apoio à agricultura.

O conjunto de dados utilizado neste estudo é composto por variáveis meteorológicas coletadas em frequência horária por estações automáticas do INMET, disponibilizadas em portal do INMET, situadas no município do Rio de Janeiro, cuja seleção contemplou as regiões de Copacabana, Marambaia, Jacarepaguá e Vila Militar.

A Tabela 7 apresenta as variáveis coletadas e suas respectivas unidades de medida. A Tabela 8 apresenta o período de coleta de dados correspondente a cada estação considerada nesta pesquisa, bem como as respectivas coordenadas geográficas de latitude e longitude.

A construção da base consolidada envolveu etapas de organização e integração das informações provenientes das diferentes estações. Inicialmente, os dados foram coletados e estruturados individualmente por estação meteorológica. Em seguida, os arquivos correspondentes a cada localidade foram organizados em tabelas de dados padronizadas,

Tabela 7: Dados das estações do INMET.

Variável	**Unidade**
Temperatura do Ar	°C
Pressão Atmosférica	mB
Temperatura do Ponto de Orvalho	°C
Umidade Relativa	%
Velocidade do Vento	m/s
Radiação Solar	kJ/m^2
Precipitação	mm

Tabela 8: Estações e período de coleta.

Estação	**Latitude**	**Longitude**	**Início**	**Fim**
Forte de Copacabana	-22,98	-43,19	2007-05-18	2024-11-30
Jacarepaguá	-22,94	-43,40	2017-08-10	2024-11-30
Marambaia	-23,05	-43,59	2002-11-08	2024-11-30
Vila Militar	-22,86	-43,41	2007-04-13	2024-11-30

garantindo consistência no formato das variáveis.

Posteriormente, as bases foram integradas em um único conjunto de dados, unificando todas as observações em uma estrutura comum. Para preservar a dimensão espacial da informação, foi adicionada uma variável indicadora da estação de origem de cada registro, possibilitando o rastreamento da procedência dos dados ao longo das análises.

O fluxo dessas etapas é sintetizado na Figura 9, que apresenta de forma esquemática o processo de coleta, padronização, integração e consolidação da base final. A base consolidada encontra-se disponível em repositório eletrônico.

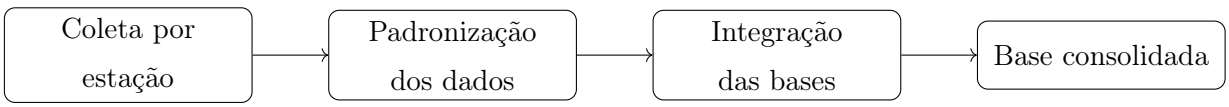

Figura 9: Fluxo resumido de construção e integração da base de dados.

5.1.2 Engenharia de Atributos, Redução de Dimensionalidade e Discretização

Após a integração dos dados provenientes de todas as estações meteorológicas, realizou-se a engenharia de atributos, com o objetivo de enriquecer o conjunto de dados e capturar informações temporais e contextuais. Foram criadas variáveis quantitativas baseadas

em defasagens e variações, definidas com base em recomendações de meteorologistas colaboradores do projeto, além de atributos qualitativos associados a padrões sazonais e diários.

Para cada variável meteorológica considerada, foram criadas variáveis que representam os valores observados nas três horas anteriores à medição, bem como as respectivas variações horárias.

Os atributos baseados em atraso, denominados *lags*, correspondem aos valores assumidos por cada variável meteorológica em instantes anteriores ao tempo de referência da observação. Formalmente, para uma variável X_t medida no instante t, foram criadas as variáveis X_{t-1}, X_{t-2} e X_{t-3}, representando, respectivamente, os valores observados uma, duas e três horas antes.

Além dos *lags*, foram construídas variáveis de variação horária, definidas como a diferença entre o valor atual e o valor observado no instante anterior, isto é,

$$\Delta_1 X_t = X_t - X_{t-1}, \quad \Delta_2 X_{t-1} = X_{t-1} - X_{t-2}, \quad \Delta_3 X_{t-2} = X_{t-2} - X_{t-3}.$$

Com isso, essas variáveis capturam a dinâmica de mudança das condições meteorológicas, permitindo verificar o aumento ou a redução dos valores ao longo do tempo recente. Além desses atributos temporais, foram criadas variáveis qualitativas com o objetivo de capturar padrões associados à variação das condições meteorológicas ao longo do dia e do ano. Essas variáveis incluíram a "Estação do Ano", categorizada de acordo com os meses correspondentes às estações no município do Rio de Janeiro, sendo verão para os meses de dezembro, janeiro e fevereiro; outono para março, abril e maio; inverno para junho, julho e agosto; e primavera para setembro, outubro e novembro. Também foi criada a variável "Turno do Dia", definida a partir da hora da observação, sendo classificada como manhã para registros entre 5h e 11h59, tarde para registros entre 12h e 17h59, e noite para registros entre 18h e 4h59.

A Tabela 9 apresenta um resumo das variáveis criadas durante a etapa de engenharia de atributos e evidencia a alta dimensionalidade do conjunto de dados resultante. Para cada uma das seis variáveis meteorológicas originais, foram geradas duas categorias de atributos derivados. As defasagens temporais representam os valores observados 1, 2 e 3 horas antes de cada registro, enquanto as variações temporais correspondem às mudanças ocorridas nesses mesmos intervalos. Esse processo resultou na criação de 36 novas variáveis numéricas. Além disso, foram incluídas duas variáveis qualitativas, "Estação do Ano" e

"Turno do Dia", totalizando 38 atributos.

Tabela 9: Variáveis criadas durante a fase de pré-processamento.

Tipo de Variável	**Variáveis Associadas**
Defasagens Temporais (1–3 horas)	Temperatura do Ar, Pressão Atmosférica, Umidade Relativa, Temperatura do Ponto de Orvalho, Velocidade do Vento, Radiação Solar
Variações Temporais (1–3 horas)	Temperatura do Ar, Pressão Atmosférica, Umidade Relativa, Temperatura do Ponto de Orvalho, Velocidade do Vento, Radiação Solar
Variáveis Qualitativas	Estação do Ano Período do Dia

Inicialmente, procedeu-se à análise de dados faltantes nas variáveis meteorológicas originais. Optou-se pela remoção dos registros que apresentavam qualquer valor ausente nessa etapa, em virtude da elevada dimensão amostral remanescente. Somente após esse procedimento foram criados os atributos temporais, como as variáveis defasadas e as variações horárias. Como a construção desses atributos depende de observações anteriores.

Essa etapa de remoção resultou na redução do conjunto original de 536.051 para 468.811 observações, correspondendo à exclusão de 12,54% dos dados.

A Tabela 10 apresenta o impacto da remoção de registros incompletos sobre a distribuição da variável precipitação. A alteração nas contagens de registros com e sem chuva ocorre porque a exclusão foi realizada considerando a presença de valores ausentes em qualquer variável meteorológica do registro. Assim, registros com precipitação observada puderam ser removidos quando alguma variável preditora associada ao mesmo instante apresentava valor ausente. Observa-se que a proporção de registros com ocorrência de chuva foi reduzida de 8,15% para 7,98%, indicando que a remoção de registros incompletos alterou pouco a distribuição relativa da variável-alvo.

Tabela 10: Impacto da remoção de registros incompletos na distribuição de precipitação.

Conjunto	**Sem chuva**	**Chuva**	**Total**	**% Chuva**	**% do total**
Com dados ausentes	492.347	43.704	536.051	8,15	100
Sem dados ausentes	431.380	37.431	468.811	7,98	87,46

A relevância das variáveis qualitativas foi avaliada por meio do teste qui-quadrado (χ^2) de Pearson, que permitiu verificar a existência de associação estatisticamente significativa entre a distribuição da precipitação, previamente discretizada, e as variáveis qualitativas

criadas (BALAKRISHNAN; VOINOV; NIKULIN, 2013).

Para essa finalidade, formulou-se o teste de independência entre as variáveis. As hipóteses avaliadas foram definidas da seguinte forma:

H_0 : A precipitação é independente das categorias da variável qualitativa.

H_1 : A precipitação não é independente das categorias da variável qualitativa.

Os resultados indicaram valores de p significativamente inferiores a 0,01 para ambas as variáveis categóricas. Dessa forma, rejeita-se a hipótese nula (H_0), evidenciando que a distribuição do evento de interesse varia de forma significativa entre as categorias analisadas, justificando a inclusão das variáveis qualitativas no conjunto de dados.

Posteriormente, devido à elevada dimensionalidade do conjunto de dados e à possível redundância informacional entre variáveis altamente correlacionadas, aplicou-se a Análise de Componentes Principais às variáveis contínuas previamente normalizadas (JOLLIFFE, 2011) para reduzir a complexidade estrutural, preservando a interpretabilidade das RBs.

A ACP é útil para sintetizar a informação contida em variáveis correlacionadas e para reduzir redundâncias no conjunto de dados. Contudo, como os componentes principais são combinações lineares das variáveis originais, eles não possuem interpretação física direta. No contexto deste trabalho, esse aspecto exige cautela, uma vez que a utilização direta desses componentes como variáveis da RB poderia dificultar a interpretação meteorológica da estrutura aprendida.

Considerando que um dos objetivos centrais desta dissertação é avaliar se a modelagem por RBs permite uma representação interpretável, do ponto de vista meteorológico, das relações probabilísticas investigadas, os componentes principais não foram empregados como nós da rede.

O gráfico de barras da variância explicada (Figura 10) mostra a contribuição individual de cada componente principal para a variância total dos dados. Observa-se que os cinco primeiros componentes principais explicam, em conjunto, 92,7% da variância total dos dados. Adotou-se como critério de retenção a explicação de pelo menos 90% da variância acumulada, um valor considerado adequado para preservação da estrutura estatística do conjunto.

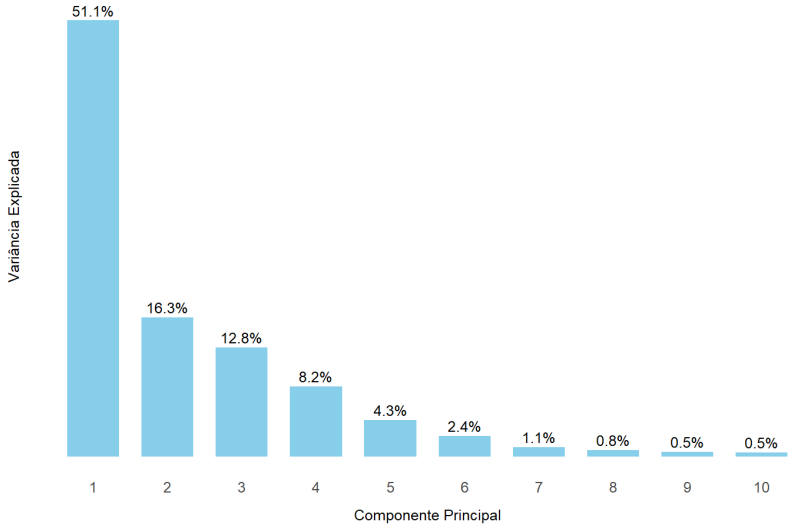

Figura 10: Variância explicada pelos dez primeiros componentes.

A Tabela 11 apresenta os desvios padrão associados, bem como as proporções de variância explicada e acumulada para os dez primeiros componentes.

Tabela 11: Resumo da variância explicada pelos dez primeiros componentes principais.

Componente	**DP**	**Var. Explicada (%)**	**Var. Acumulada (%)**
PC1	0,4406	51,1	51,1
PC2	0,2490	16,3	67,4
PC3	0,2202	12,8	80,2
PC4	0,1768	8,2	88,4
PC5	0,1275	4,3	92,7
PC6	0,0959	2,4	95,1
PC7	0,0641	1,1	96,2
PC8	0,0536	0,8	97,0
PC9	0,0447	0,5	97,5
PC10	0,0424	0,5	98,0

A análise das cargas fatoriais associadas aos cinco primeiros componentes revelou concentração de maiores contribuições absolutas nas variáveis correspondentes às defasagens temporais das medições meteorológicas. Em contrapartida, as variáveis relacionadas às variações horárias apresentaram menor relevância estatística e indícios de sobreposição informacional com as defasagens correspondentes. Diante disso, optou-se por excluir as

variáveis de variação temporal, mantendo-se apenas as variáveis de defasagem e os atributos qualitativos previamente definidos.

Essa estratégia possibilitou preservar a maior parte da variabilidade estatística do conjunto original, simultaneamente à redução de sua complexidade estrutural, facilitando a construção e interpretação das RBs. Porém, embora a ACP seja uma técnica eficaz para reduzir a dimensionalidade mantendo a maior parte da variância dos dados, é importante destacar algumas de suas limitações. A análise considera apenas relações lineares entre variáveis, podendo não capturar estruturas mais complexas. Além disso, por se tratar de um método não supervisionado, a ACP não leva em conta a variável-alvo, o que pode resultar na seleção de atributos que explicam variação geral, mas que não são necessariamente relevantes para o fenômeno de interesse (JOLLIFFE, 2011).

Para a construção das RBs, somente variáveis discretas foram consideradas. Para isso, todas as variáveis contínuas do conjunto de dados, especificamente aquelas que representam as defasagens temporais, foram discretizadas. No contexto das RBs, essa abordagem viabiliza a criação de tabelas de probabilidade condicional, otimiza o processamento computacional e aprimora a interpretabilidade dos resultados.

Entretanto, a discretização constitui uma escolha metodológica crítica, pois transforma variáveis originalmente contínuas em um número finito de categorias, implicando perda de granularidade e possível redução de variações mais sutis presentes nos dados. Em contrapartida, discretizações excessivamente detalhadas tenderiam a produzir tabelas de probabilidades condicionais muito esparsas, dificultando a estimação estável dos parâmetros e aumentando a complexidade estrutural da rede. Dessa forma, a discretização adotada neste trabalho buscou preservar a interpretabilidade e viabilidade computacional.

A discretização foi realizada utilizando o algoritmo de agrupamento K-Means (MACQUEEN, 1967), no qual cada valor é atribuído à categoria cujo centroide apresenta menor distância euclidiana. O número de categorias foi mantido deliberadamente reduzido, com o objetivo de evitar granularidade excessiva, minimizar a esparsidade nas tabelas de probabilidades condicionais e facilitar a interpretação das dependências entre variáveis.

No caso específico da variável que representa a Radiação Solar, observações com valor igual a zero foram desconsideradas temporariamente durante o processo de discretização, pois influenciavam fortemente o processo de agrupamento, sendo posteriormente reintegradas na categoria “Noite/Zero”. A Tabela 12 apresenta as categorias resultantes e os valores dos centroides obtidos no processo de discretização.

Tabela 12: Discretização das variáveis contínuas.

Variáveis (t–1h, t–2h, t–3h)	**Categoria**	**Valor do Centróide**
Temperatura	Alta	31,2
	Médio	25,9
	Médio/Baixo	22,2
	Baixo	18,1
Pressão Atmosférica	Alta	1019,3
	Médio	1012,9
	Baixo	1007,3
Temp. Ponto de Orvalho	Alta	21,5
	Médio	18,0
	Baixo	13,7
Umidade Relativa	Alta	87,8
	Médio	69,8
	Baixo	45,9
Velocidade do Vento	Alta	6,3
	Médio	2,7
	Baixo	0,7
Radiação Solar	Alta	2922,9
	Médio	2014,9
	Baixo	1543,0

A variável de precipitação foi categorizada com base nas regras apresentadas na Tabela 13, sendo definidas as categorias *Sem chuva* para valores de precipitação iguais a 0 mm/h, *Chuva* para valores no intervalo (0, 25) mm/h e *Chuva Intensa* para valores iguais ou superiores a 25 mm/h. A escolha do limiar de 25 mm/h para a categoria *Chuva Intensa* baseia-se no fato de que esse valor representa aproximadamente o percentil 99,5 da distribuição de precipitação no município do Rio de Janeiro. Ressalta-se que essa classificação difere da utilizada operacionalmente pelo Sistema Alerta Rio, que define as intensidades de precipitação horária da seguinte forma: chuva fraca (menor que 5 mm/h), chuva moderada (entre 5 e 25 mm/h), chuva forte (entre 25,1 e 50 mm/h) e chuva muito forte (acima de 50 mm/h). Essas definições são empregadas pelo sistema municipal de monitoramento meteorológico da cidade do Rio de Janeiro e encontram-se disponíveis em https://www.sistema-alerta-rio.com.br/previsao-do-tempo/termosmet/.

Tabela 13: Discretização para a variável precipitação.

Categoria	**Intervalo de Precipitação (mm/h)**
Sem chuva	$= 0$
Chuva	$(0, 25)$
Chuva Intensa	≥ 25

5.1.3 Conjunto Final

Após o *pipeline* de pré-processamento e engenharia de atributos, o conjunto de dados final utilizado para o aprendizado das RBs foi composto por 468.811 instâncias e 20 atributos. Esse conjunto inclui seis variáveis meteorológicas, cada uma com defasagens temporais de 1, 2 e 3 horas, totalizando 18 variáveis defasadas, além de duas variáveis qualitativas ("Estação do Ano" e "Período do Dia"). Os dados foram coletados com resolução horária, o que define a granularidade temporal do conjunto de dados.

A variável-alvo, precipitação, foi incluída em sua forma discretizada. Após o pré-processamento, 431.380 instâncias (92,02%) foram classificadas como "Sem Chuva", 37.248 instâncias (7,95%) como "Chuva" e 183 instâncias (0,03%) como "Chuva Intensa".

Exclusivamente para a tarefa de classificação, as classes "Chuva" e "Chuva Intensa" foram agregadas, resultando em 37.431 instâncias (7,98%) na classe "Chuva".

A Tabela 14 resume o conjunto final de atributos, juntamente com suas respectivas abreviações utilizadas ao longo dos experimentos, incluindo a variável-alvo. Para as variáveis meteorológicas, a notação "_kh" indica que cada atributo é representado com defasagens temporais, em que $k \in \{1,2,3\}$ corresponde às medições realizadas 1, 2 e 3 horas antes do instante de previsão. A última linha da tabela também apresenta o número total de instâncias do conjunto de dados.

O conjunto de dados final está disponível publicamente em https://github.com/RioNowcast-Green/rio-rainfall-bayesian-networks-dataset.

5.2 Aprendizado, Validação e Inferências das Redes Bayesianas

Nesta Seção, descreve-se o processo de construção das estruturas e dos parâmetros das RBs a partir do conjunto de dados previamente pré-processados. Esse processo segue um *pipeline* analítico projetado para assegurar interpretabilidade física e robustez, conforme

Tabela 14: Conjunto de dados final.

Variáveis	**Abreviações**
Temperatura do Ar	T_kh
Pressão Atmosférica	P_kh
Temperatura do Ponto de Orvalho	DT_kh
Umidade Relativa	H_kh
Velocidade do Vento	W_kh
Radiação Global	R_kh
Estação do Ano	S
Período do Dia	Day_p
Precipitação	Rain
Instâncias	468.811

Nota: Para as variáveis meteorológicas, $k \in \{1,2,3\}$ indica medições realizadas k horas antes da variável-alvo (Precipitação).

resumido na Figura 11, que ilustra o fluxo interno adotado durante essa fase, abrangendo a definição de restrições estruturais, o aprendizado da estrutura, a estimação dos parâmetros, a avaliação do modelo e a realização de inferências.

Neste estudo optou-se pela utilização de RBs estáticas para modelar as relações probabilísticas entre variáveis meteorológicas e a ocorrência de precipitação. Embora RBs Dinâmicas permitam modelar explicitamente a evolução temporal de sistemas probabilísticos, a presente abordagem prioriza a interpretabilidade da estrutura da rede e a análise das dependências probabilísticas entre as variáveis.

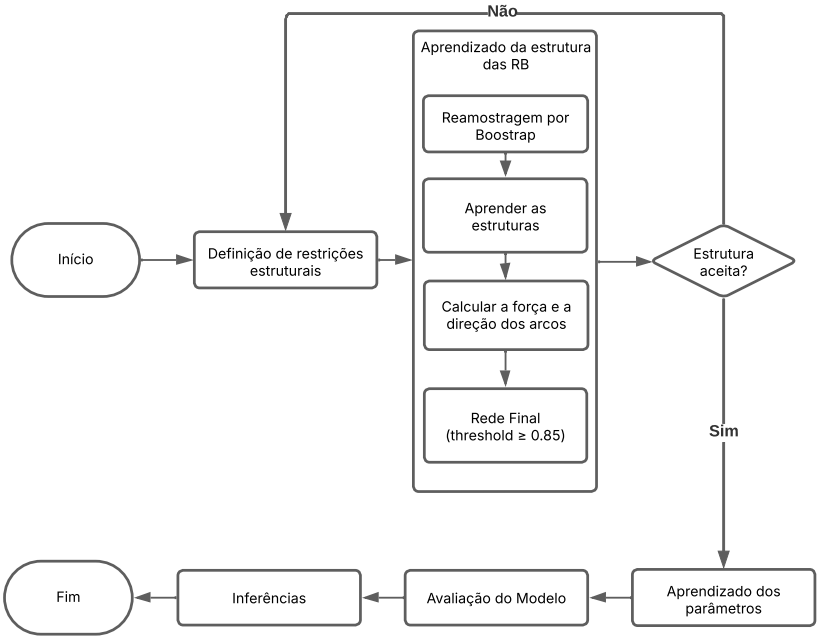

Figura 11: *Pipeline* analítico para geração de resultados.

Inicialmente, foi definido um conjunto de restrições estruturais com base no conhecimento de domínio sobre fenômenos meteorológicos. Essas restrições tiveram como objetivo garantir que a estrutura aprendida fosse fisicamente consistente e interpretável. Especificamente, a variável de interesse, precipitação, foi impedida de atuar como nó pai de qualquer outra variável, refletindo seu papel como variável resposta. Restrições temporais também foram impostas, assegurando que medições futuras não pudessem influenciar observações passadas. Além disso, variáveis sazonais, como o período do dia e a estação do ano, foram mantidas como nós raiz, uma vez que não são causadas por outros fatores meteorológicos.

A incorporação desse conhecimento prévio ao processo de aprendizado estrutural é formalmente suportada na literatura de RBs. Conforme discutido por (SCUTARI, 2010), é possível orientar os algoritmos de aprendizado por meio de restrições estruturais explícitas. As restrições de inclusão (*whitelists*) especificam arcos que devem obrigatoriamente estar presentes na rede final, enquanto as restrições de proibição (*blacklists*) definem arcos que não podem ser formados. Esse mecanismo reduz o espaço de busca, orienta o processo de otimização estrutural e contribui para a validade estrutural e interpretativa do modelo final.

A lista completa de restrições estruturais aplicadas durante o processo de modelagem é apresentada no Apêndice A, onde cada linha da tabela representa uma aresta explicitamente proibida na estrutura das RBs.

Para a construção das RBs, foram empregados ao todo 14 algoritmos de aprendizado estrutural. A Tabela 15 apresenta os métodos utilizados neste estudo, agrupados de acordo com a natureza de sua abordagem metodológica: oito algoritmos baseados em *score*, quatro baseados em *constraint* e dois de natureza híbrida. Essa diversidade de estratégias foi adotada com o objetivo de explorar uma ampla gama de possibilidades na identificação das dependências probabilísticas entre as variáveis.

A seleção dos algoritmos fundamentou-se em uma revisão da literatura científica, que evidencia a aplicação recorrente dessas técnicas em problemas de aprendizado estrutural em RBs. Detalhes conceituais e operacionais adicionais sobre cada método encontram-se descritos na Seção 2.1.7.

Após a definição do conjunto de algoritmos a serem utilizados, o aprendizado da estrutura foi conduzido em conjunto com o procedimento de reamostragem *bootstrap*, para aumentar a robustez das dependências inferidas e avaliar a estabilidade dos arcos sob variações amostrais (EFRON, 1992; FRIEDMAN; GOLDSZMIDT; WYNER, 2013).

Tabela 15: Algoritmos de aprendizado estrutural utilizados.

Categoria	**Algoritmos**
Score-Based	Hill Climbing and TABU Search (BIC, AIC, K2, BDeu)
Constraint-Based	PC, GS, IAMB and MMPC
Hybrid-Based	MMHC and RSMAX2

Para cada uma das combinações dos 14 algoritmos considerados, foram realizadas 300 reamostragens. Em cada amostra gerada, uma nova estrutura de RB foi aprendida, possibilitando estimar, para cada arco candidato, sua força, medida pela frequência de ocorrência, e sua direção predominante.

Com base nessas métricas, foi construída uma rede de consenso, na qual foram retidos apenas os arcos com frequência igual ou superior a 85%. Esse critério buscou preservar exclusivamente dependências estruturalmente estáveis, reduzindo a influência de flutuações amostrais e minimizando a incorporação de relações espúrias.

A estrutura resultante foi então submetida a uma validação qualitativa quanto à coerência temporal e física. Quando foram identificados arcos incompatíveis com o conhecimento meteorológico ou com a ordenação temporal, a estrutura foi descartada, as restrições foram revisadas e o aprendizado foi repetido. Essa estratégia está alinhada a recomendações recentes da literatura, segundo as quais abordagens baseadas em reamostragem contribuem para aumentar a robustez do aprendizado estrutural em RBs e mitigar instabilidades na identificação de arestas (CHOBTHAM; CONSTANTINOU, 2024).

Uma vez definida a estrutura da RB, a etapa seguinte consistiu no aprendizado dos parâmetros que quantificam as relações probabilísticas entre as variáveis. Para esse fim, foi empregado o método de Máxima Verossimilhança (*Maximum Likelihood Estimation* – MLE), que busca determinar os valores dos parâmetros que maximizam a verossimilhança dos dados, dada a estrutura da rede (KOLLER; FRIEDMAN, 2009).

O desempenho do modelo foi avaliado por meio de métricas quantitativas, complementadas por uma validação qualitativa conduzida por especialistas em meteorologia, com o objetivo de verificar a coerência física e a plausibilidade das relações inferidas.

Em virtude do desbalanceamento observado na variável de interesse, na qual a categoria Chuva Intensa representava apenas 0,03% do conjunto de dados, optou-se pelo agrupamento das categorias originais, uma vez que não seria possível obter estimativas estatisticamente robustas para esse tipo de evento. A variável binária, denominada Chuva,

passou a corresponder à união das categorias Chuva e Chuva Intensa. Dessa forma, a tarefa de predição foi formalizada como um problema de classificação binária, cujo objetivo consistiu em prever a ocorrência ou não de precipitação a partir das demais variáveis meteorológicas disponíveis.

Ainda assim, o conjunto de dados permaneceu fortemente desbalanceado, com 431.380 registros classificados como "Sem Chuva" e 37.431 como "Chuva". Para mitigar os efeitos desse desbalanceamento durante o treinamento, adotou-se a técnica de *undersampling* nos subconjuntos de treinamento no âmbito da validação cruzada *k-fold* com $k = 5$ (HASTIE; TIBSHIRANI; FRIEDMAN, 2009). Especificamente, selecionou-se aleatoriamente um subconjunto de instâncias da classe majoritária de modo a igualar o número de observações da classe minoritária, reduzindo o viés no processo de aprendizado e possibilitando comparações mais robustas entre os modelos avaliados.

As RBs resultantes foram avaliadas sob critérios quantitativos e qualitativos. Sob a perspectiva preditiva, empregaram-se validação cruzada, análise de curvas *Receiver Operating Characteristic* (ROC), curvas de calibração e a função de perda baseada na log-verossimilhança negativa (BRADLEY, 1997; NICULESCU-MIZIL; CARUANA, 2005; KOLLER; FRIEDMAN, 2009; HASTIE; TIBSHIRANI; FRIEDMAN, 2009; MURPHY, 2012; VAN CALSTER et al., 2019; HUANG et al., 2020).

Por fim, foram conduzidas inferências probabilísticas nas redes aprendidas com o objetivo de estimar a distribuição condicional da variável alvo (precipitação) sob diferentes cenários. Essas inferências possibilitaram uma análise detalhada de como distintas variáveis meteorológicas influenciam a probabilidade de ocorrência de eventos de precipitação em condições específicas.

A inferência probabilística foi realizada utilizando a função `cpquery` do pacote `bnlearn`, que estima probabilidades condicionais por meio de simulação Monte Carlo. Nesse procedimento, são geradas amostras aleatórias da RB parametrizada, sendo a probabilidade de interesse estimada a partir da proporção de amostras que satisfazem simultaneamente o evento e a evidência especificados.

As estimativas pontuais das probabilidades foram obtidas por meio de múltiplas execuções da função *cpquery*. Para cada probabilidade, foram realizadas 3 000 estimativas independentes, a partir das quais foi calculada a média das probabilidades estimadas.

5.3 Configuração Experimental

Os experimentos foram realizados em uma máquina equipada com processador AMD Ryzen 5 5600G e 16 GB de memória RAM.

Para a implementação dos modelos, avaliação e visualização das estruturas geradas, utilizou-se a linguagem de programação *R* (versão 4.5.2), juntamente com as seguintes bibliotecas: *bnlearn* (versão 5.1) para aprendizado estrutural e inferência, *igraph* (versão 2.2.1) para manipulação de grafos, *dplyr* (versão 1.1.4) para manipulação de dados, *ggplot2* (versão 3.3.0) para geração de gráficos e visualizações, e *Rgraphviz* (versão 2.54.0) para renderização das estruturas gráficas.

Para avaliar o impacto ambiental dos experimentos computacionais, foram estimados o consumo de energia e a pegada hídrica para cada procedimento de aprendizado estrutural. Os cálculos foram realizados utilizando o wAIter calculator, ferramenta desenvolvida para estimar a pegada ambiental de modelos de inteligência artificial, conforme proposto por (BREDER; FERRO, 2025). A pegada hídrica representa o volume de água doce utilizado (direta ou indiretamente) para sustentar o consumo energético dos processos computacionais, especialmente nos sistemas de resfriamento de *data centers*. Apesar desta pesquisa não utilizar infraestrutura computacional em *data centers* a metodologia do wAIter permite estimar o impacto indireto na água devido ao uso de energia elétrica proveniente de hidroelétricas.

Os códigos utilizados para o aprendizado estrutural das RBs e geração dos resultados estão disponíveis publicamente no repositório do projeto em github.com.

6 Resultados: Análise Exploratória

A análise exploratória dos dados foi realizada com o objetivo de examinar a distribuição das variáveis atmosféricas e a correlação entre a precipitação e diversas variáveis meteorológicas.

Observa-se um desbalanceamento considerável no conjunto de dados em relação à variável "precipitação". Assim como é comum em séries temporais meteorológicas, a maior parte das observações corresponde a períodos sem ocorrência de chuva. Esse desbalanceamento entre as classes "chuva" e "sem chuva" pode impactar negativamente o desempenho de modelos preditivos, especialmente aqueles sensíveis à distribuição das classes. A Figura 12 ilustra essa discrepância na frequência de registros, reforçando a necessidade de considerar estratégias específicas para lidar com o desbalanceamento.

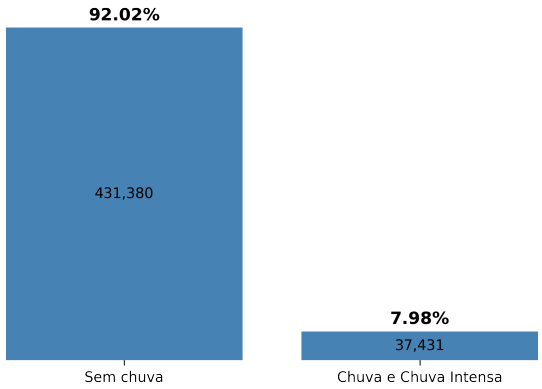

Figura 12: Distribuição das observações da variável precipitação.

A análise da correlação linear de Pearson entre a variável precipitação e os registros meteorológicos das três horas anteriores foi realizada considerando as variáveis em sua forma contínua, antes de qualquer processo de discretização. Os resultados mostram associações de baixo magnitude, em que as correlações mais expressivas, embora ainda fracas, foram observadas para a radiação solar e a umidade relativa do ar, enquanto variáveis como temperatura, pressão atmosférica e velocidade do vento apresentaram valores ainda mais próximos de zero, conforme apresentado na Tabela 16. Esses resultados evidenciam uma relação linear pouco expressiva entre a precipitação e as demais variáveis

meteorológicas, o que tem implicações importantes na modelagem com RBs.

Em especial, para as abordagens baseadas no método *Constraint-Based*, que se apoiam em testes de independência condicional, essa baixo associação linear pode dificultar a identificação de vínculos significativos entre as variáveis meteorológicas e a precipitação, impactando a estrutura aprendida pela rede.

Tabela 16: Coeficientes de correlação de Pearson entre precipitação e variáveis meteorológicas contínuas.

Variável	**Correlação**
Temperatura (t-1)	-0,030
Temperatura (t-2)	-0,010
Temperatura (t-3)	-0,0008
Velocidade do vento (t-1)	0,045
Velocidade do vento (t-2)	0,031
Velocidade do vento (t-3)	0,028
Radiação (t-1)	-0,092
Radiação (t-2)	-0,077
Radiação (t-3)	-0,064
Pressão atmosférica (t-1)	-0,020
Pressão atmosférica (t-2)	-0,020
Pressão atmosférica (t-3)	-0,034
Umidade relativa (t-1)	0,080
Umidade relativa (t-2)	0,069
Umidade relativa (t-3)	0,057
Ponto de orvalho (t-1)	0,060
Ponto de orvalho (t-2)	0,062
Ponto de orvalho (t-3)	0,063

Complementando a análise exploratória, foi investigada a relação entre a ocorrência de chuvas e a variável categórica "Estação do Ano". A Figura 13 apresenta a distribuição da quantidade de registros com precipitação observada em cada estação do ano. Observa-se que a maior incidência de chuvas ocorre na primavera, seguida pelo verão, enquanto outono e inverno apresentam menores frequências. Esses resultados sugerem uma influência da sazonalidade na ocorrência do fenômeno de interesse, reforçando a importância da inclusão dessas sazonais no conjunto de dados.

Outra informação relevante sobre a distribuição da precipitação ao longo das estações surge ao se considerar exclusivamente os eventos classificados como chuvas intensas. A Figura 14 revela que, diferentemente do padrão observado para a chuva em geral, o verão passa a apresentar a maior quantidade de registros de chuva intensa, seguido pelo outono, que ocupa apenas a terceira posição quando todos os eventos de chuva são considerados.

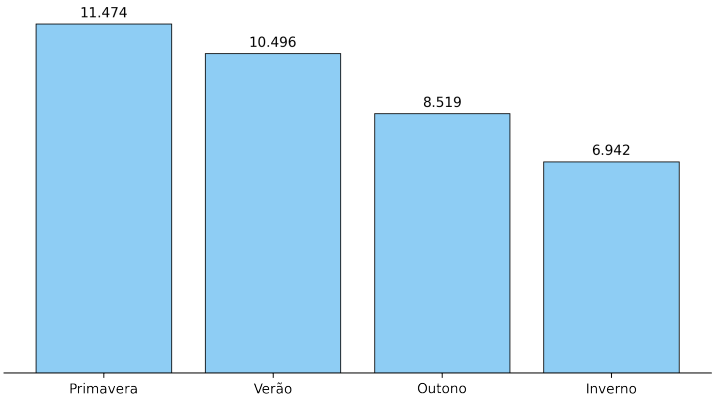

Figura 13: Distribuição da quantidade de chuvas e chuvas intensas observadas por estação do ano.

Essa inversão no padrão evidencia uma dinâmica distinta para os eventos extremos de precipitação, indicando que a frequência de chuvas intensas não segue necessariamente a mesma distribuição das chuvas mais leves.

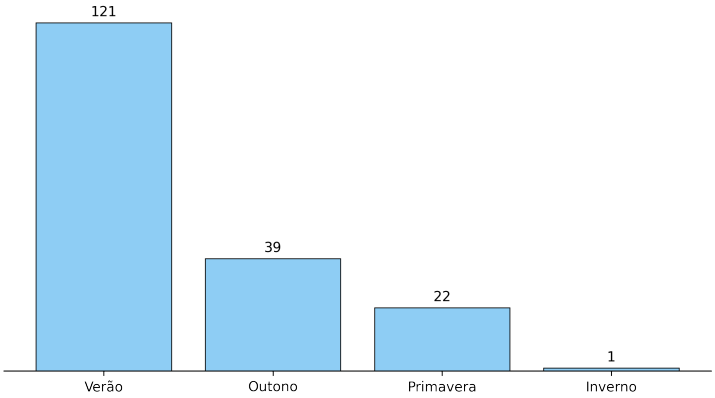

Figura 14: Distribuição da quantidade de chuvas intensas observadas por estação do ano.

Além disso, também foi analisada a distribuição das ocorrências de chuva ao longo dos diferentes turnos do dia. A Figura 15 apresenta a frequência de registros de precipitação categorizados por “Manhã”, “Tarde” e “Noite”. Os resultados indicam uma maior concentração de chuvas durante o período noturno, seguida pela manhã e, por fim, pela tarde. Esse padrão mostra uma tendência de maior incidência de precipitação durante o período classificado como noite.

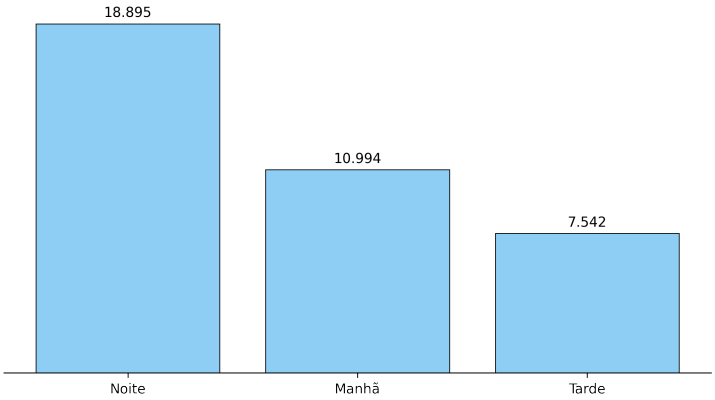

Figura 15: Distribuição da quantidade de chuvas e chuvas intensas observadas por turno do dia.

Para realizar a análise espacial da precipitação, ou seja, investigar se a ocorrência de chuvas varia significativamente entre as diferentes estações meteorológicas, foi necessário filtrar os dados a partir de 10 de agosto de 2017. Essa data marca o início das operações da estação de Jacarepaguá, conforme indicado na Tabela 8. A adoção desse recorte temporal garante a comparabilidade entre as estações, uma vez que a inclusão de períodos em que algumas estações ainda não estavam em funcionamento poderia introduzir vieses na análise, comprometendo a validade das conclusões.

A Figura 16 apresenta a frequência total de registros com ocorrência de chuva para cada uma das quatro estações meteorológicas consideradas no estudo, no período posterior a 10 de agosto de 2017. A distribuição geral de chuvas entre as estações é bastante semelhante, o que dificulta a identificação de um padrão espacial claro ou a conclusão definitiva sobre a existência de diferenças significativas na ocorrência de precipitação entre os locais analisados.

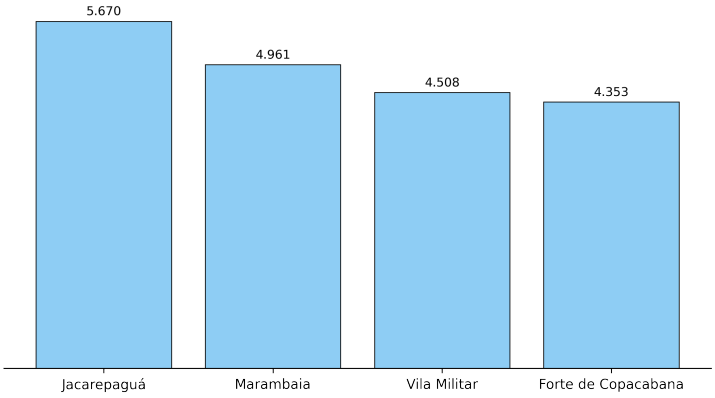

Figura 16: Distribuição da quantidade de chuvas mais chuvas intensas observadas por estação meteorológica (após 10/08/2017).

As tabelas de contingência desempenham um papel importante na análise de resultados em RBs, sobretudo na verificação dos resultados probabilísticos obtidos. Elas permitem examinar empiricamente a distribuição conjunta entre variáveis categóricas, evidenciando padrões de coocorrência e possíveis associações.

A Tabela 17 apresenta a distribuição conjunta entre as categorias discretizadas da variável temperatura (t–1h), que representa o valor da temperatura com atraso de 1 hora, e a ocorrência de precipitação. Nota-se que a maior parte dos registros de chuva e chuva intensa ocorre nas faixas de temperatura classificadas como "Médio/Baixo" e "Médio". Por outro lado, a categoria "Alta" de temperatura apresenta uma frequência muito reduzida de precipitação. Uma possível explicação para esse comportamento é que, nos momentos que antecedem a ocorrência de chuva, a maior presença de nuvens tende a reduzir a radiação solar incidente na superfície, limitando o aquecimento do ar.

Tabela 17: Tabela de contingência entre categorias de temperatura (t-1h) e ocorrência de precipitação.

Precipitação	**Temperatura (t–1h)**			
	Alta	**Médio**	**Média/Baixo**	**Baixo**
Chuva	768 (1,5%)	6.593 (4,8%)	20.071 (10,4%)	9.816 (11,3%)
Chuva Intensa	12 (0,0%)	79 (0,1%)	90 (0,0%)	2 (0,0%)
Sem Chuva	50.750 (98,5%)	131.259 (95,2%)	172.577 (89,6%)	76.794 (88,7%)

A Tabela 18 apresenta a distribuição conjunta entre a variável "vento_valor1", que representa a categoria da velocidade do vento com atraso de 1 hora, e os diferentes tipos de ocorrência de chuva. Observa-se que a maioria dos eventos de chuva ocorre quando a velocidade do vento está na categoria "baixo", seguida por "alto" e, por último, "médio". Ao considerar as proporções dentro de cada categoria de vento, observa-se que a maior frequência relativa de chuva ocorre na categoria "Médio", na qual 12,6% dos registros apresentam chuva. Em comparação, as categorias "Alto" e "Baixo" apresentam proporções menores, com 7,1% e 7,6%, respectivamente.

Tabela 18: Tabela de contingência entre categorias de velocidade do vento (t-1h) e ocorrência de precipitação.

Precipitação	**Velocidade do Vento (t–1h)**		
	Alto	**Médio**	**Baixo**
Chuva	10.006 (7,1%)	5.764 (12,6%)	21.478 (7,6%)
Chuva Intensa	66 (0,0%)	39 (0,1%)	78 (0,0%)
Sem Chuva	131.454 (92,9%)	39.946 (87,3%)	259.980 (92,4%)

A Tabela 19 apresenta a distribuição conjunta entre a variável "radiacao_valor1", que representa a intensidade da radiação solar registrada com 1 hora de atraso, e os diferentes tipos de ocorrência de chuva. Observa-se que a maioria dos eventos de chuva e de chuva intensa ocorre na categoria "Baixo". Por outro lado, categorias como "alto" e "médio" concentram menos registros de chuva.

Tabela 19: Tabela de contingência entre categorias de radiação solar (t-1h) e ocorrência de precipitação.

Precipitação	**Radiação Solar (t–1h)**			
	Noite/Zero	**Alto**	**Médio**	**Baixo**
Chuva	12.401 (7,2%)	779 (1,3%)	2.370 (3,2%)	21.698 (13,6%)
Chuva Intensa	61 (0,0%)	3 (0,0%)	8 (0,0%)	111 (0,1%)
Sem Chuva	160.230 (92,8%)	60.784 (98,7%)	72.669 (96,8%)	137.697 (86,3%)

A Tabela 20 apresenta a distribuição conjunta entre a variável "pressao_valor1", que representa a categoria da pressão atmosférica com atraso de 1 hora, e os diferentes tipos de ocorrência de chuva. Observa-se que a maior parte dos eventos de chuva ocorre quando a pressão está na categoria "Médio", seguida da categoria "Baixo".

Tabela 20: Tabela de contingência entre categorias de pressão atmosférica (t-1h) e ocorrência de precipitação.

Precipitação	**Pressão Atmosférica (t–1h)**		
	Alto	**Médio**	**Baixo**
Chuva	9.254 (8,2%)	16.309 (8,0%)	11.685 (7,7%)
Chuva Intensa	2 (0,0%)	93 (0,0%)	88 (0,1%)
Sem Chuva	103.214 (91,8%)	188.631 (92,0%)	139.535 (92,2%)

A Tabela 21 apresenta a distribuição conjunta entre a variável "umid_valor1", que representa a categoria da umidade relativa do ar com atraso de 1 hora, e os diferentes tipos de ocorrência de chuva. Observa-se que a maior parte dos eventos de chuva, incluindo eventos de chuva intensa, ocorre quando a umidade se encontra na categoria "alta", seguida pela categoria "médio". Por outro lado, a categoria "baixo" concentra proporcionalmente menos registros de precipitação, o que está em consonância com o comportamento atmosférico esperado, uma vez que níveis mais elevados de umidade favorecem a formação de nuvens e o desenvolvimento de precipitação.

Tabela 21: Tabela de contingência entre categorias de umidade relativa do ar (t-1h) e ocorrência de precipitação.

Precipitação	**Umidade Relativa do Ar (t–1h)**		
	Alta	**Médio**	**Baixo**
Chuva	31.593 (13,5%)	4.930 (2,9%)	725 (1,1%)
Chuva Intensa	153 (0,1%)	22 (0,0%)	8 (0,0%)
Sem Chuva	202.419 (86,4%)	163.801 (97,1%)	65.160 (98,9%)

A Tabela 22 apresenta a distribuição conjunta entre a variável "orvalho_valor1", que representa o valor discretizado do ponto de orvalho com atraso de 1 hora, e os diferentes tipos de ocorrência de chuva. Observa-se que a maior parte dos registros de chuva comum ocorre nas categorias "alto" e "médio" enquanto a categoria "baixo" concentra proporcionalmente menos eventos de precipitação.

Tabela 22: Tabela de contingência entre categorias de ponto de orvalho (t-1h) e ocorrência de precipitação.

Precipitação	**Ponto de Orvalho (t–1h)**		
	Alto	**Médio**	**Baixo**
Chuva	18.764 (9,8%)	15.684 (8,0%)	2.800 (3,5%)
Chuva Intensa	170 (0,1%)	13 (0,0%)	0 (0,0%)
Sem Chuva	173.202 (90,1%)	179.854 (92,0%)	78.324 (96,5%)

Com base nos resultados apresentados, a análise exploratória permitiu identificar padrões relevantes e limitações do conjunto de dados que foram considerados na etapa de modelagem. O forte desbalanceamento da variável precipitação evidenciou a necessidade de atenção à distribuição das classes durante o treinamento e na escolha das métricas

de avaliação. As baixas correlações lineares observadas entre a precipitação e as variáveis meteorológicas sugeriram que a dependência entre essas variáveis poderia não ser adequadamente descrita apenas por associações lineares simples. Além disso, os padrões observados para “Estação do Ano” e “Turno do Dia” indicaram a relevância de manter essas variáveis categóricas no conjunto de atributos, enquanto as tabelas de contingência mostraram que a discretização preservou relações empíricas plausíveis entre categorias meteorológicas e ocorrência de precipitação.

7 Resultados: Redes Bayesianas

Esta Seção apresenta os resultados obtidos a partir da metodologia descrita na Seção 5, com foco na construção e avaliação das RBs. A Figura 11 sintetiza o *pipeline* analítico adotado.

Inicialmente, são definidas restrições estruturais baseadas em conhecimento prévio, garantindo interpretabilidade física e coerência temporal no aprendizado da rede. Em seguida, a estrutura da RB é aprendida por meio de *bootstrap*, no qual múltiplas reamostragens são geradas e, para cada uma, uma rede é estimada. A rede final é construída retendo apenas os arcos com frequência igual ou superior a 85%, sendo posteriormente validada quanto à coerência física e temporal.

Após a validação estrutural, procede-se à estimação dos parâmetros e à avaliação do desempenho preditivo. Por fim, realizam-se inferências probabilísticas permitindo analisar como diferentes condições atmosféricas influenciam a probabilidade de precipitação.

7.1 Construção das Redes Bayesianas

As RBs foram construídas por meio de aprendizado estrutural automatizado, empregando múltiplos algoritmos conforme descrito na Tabela 15. A definição das estruturas foi orientada por conhecimento prévio do domínio, incorporado por restrições estruturais que guiaram a formação das dependências entre as variáveis. Para fins de representação nos grafos, adotaram-se as abreviações apresentadas na Tabela 14.

Os resultados evidenciaram uma limitação na aplicação de algoritmos *constraint-based* e *hybrid* para a modelagem da precipitação. Conforme descrito na Seção 2.1.7, esses métodos dependem de testes de independência condicional para inferir as conexões entre os nós. Entretanto, a análise exploratória (Capítulo 6) indicou correlações lineares fracas entre a precipitação e as demais variáveis meteorológicas (Tabela 16), comprometendo a sensibilidade desses testes.

Esse cenário comprometeu a capacidade dos algoritmos baseados em *constraint* e *hybrids* de identificar relações diretas com a variável de interesse. Como consequência, verificou-se o isolamento do nó de precipitação nas estruturas aprendidas, sem arestas conectando-o às demais variáveis da rede.

Tal limitação é observada nas Figuras 17 e 18, que apresentam as redes geradas pelos algoritmos baseados em *constraint* e *hybrids*. As versões ampliadas, em melhor resolução e com a visualização individual de cada estrutura, encontram-se nos Apêndices B a G.

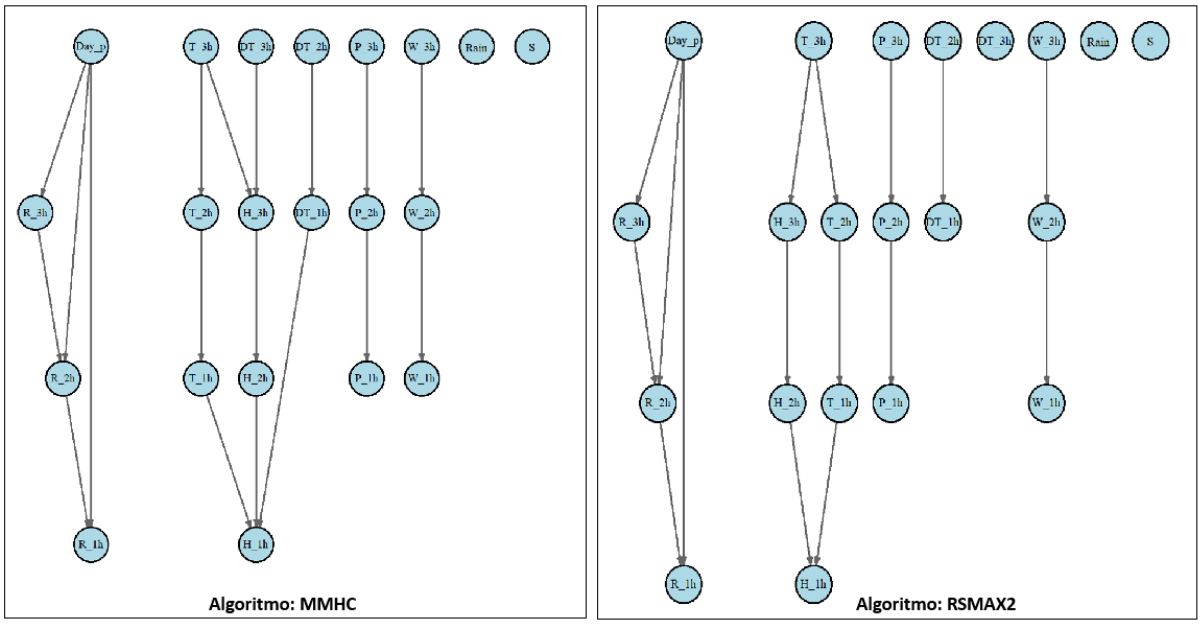

Figura 17: Redes Bayesianas geradas pelos algoritmos *hybrids*.

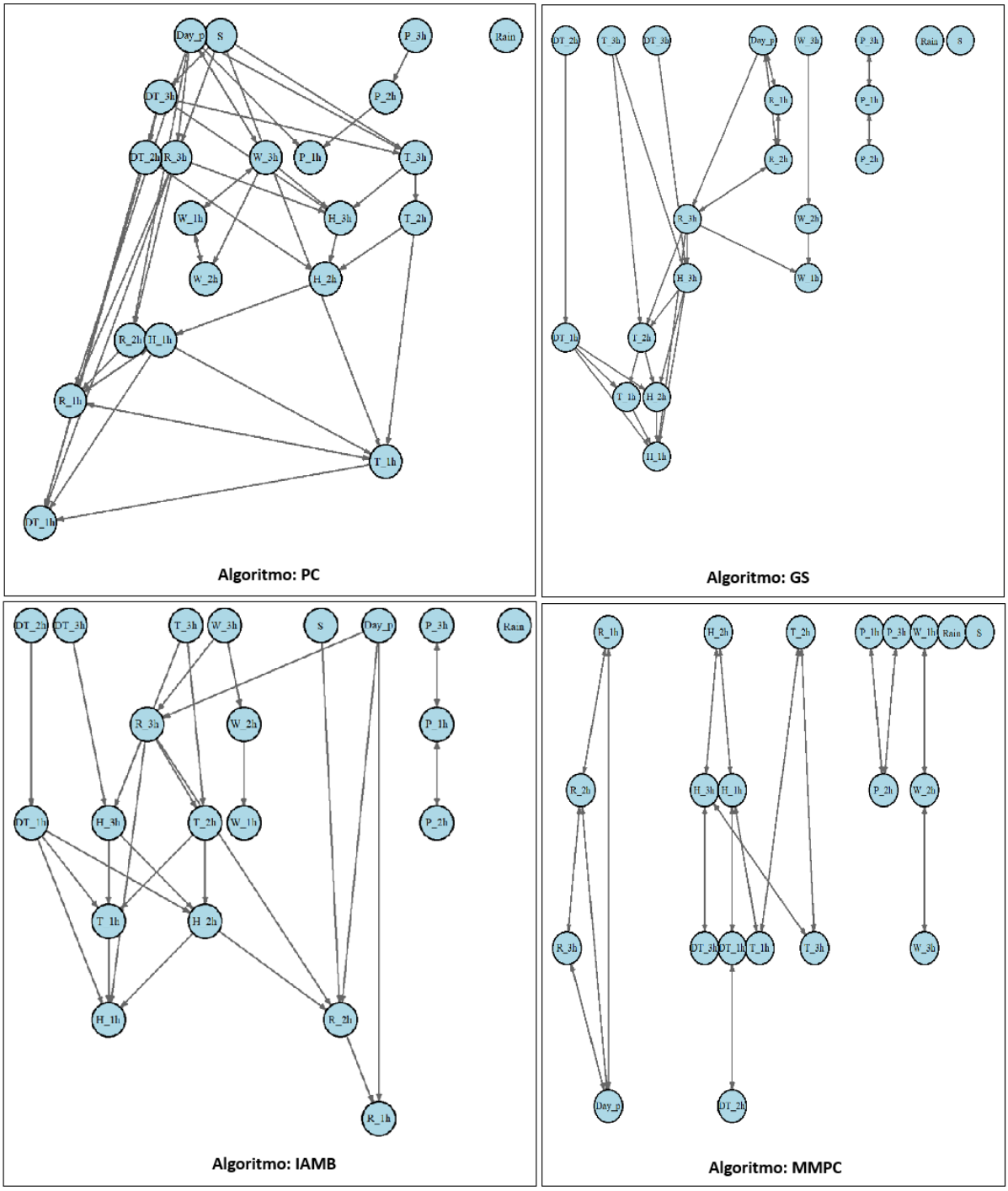

Figura 18: Redes Bayesianas geradas pelos algoritmos baseados em *constraint*.

Em contraste, os algoritmos baseados em pontuação identificaram associações relevantes ao otimizar o ajuste global aos dados. HC e *TABU Search* convergiram para o mesmo ótimo local, produzindo estruturas idênticas, as visualizações individuais das estruturas geradas por esses algoritmos são apresentadas nos Apêndices H a J. Conforme a Tabela 23, os critérios AIC e K2 apresentaram maiores valores médios de MB, Tamanho da Vizinhança e Fator de Ramificação, indicando redes mais conectadas. O MB corres-

ponde ao conjunto mínimo de variáveis que torna um nó condicionalmente independente das demais; o Tamanho da Vizinhança representa a média de conexões diretas por nó; e o Fator de Ramificação expressa a média de arestas direcionadas de saída. Tais diferenças evidenciam o efeito dos critérios de pontuação sobre a complexidade estrutural. Os critérios AIC e K2 aplicam penalizações menos severas, favorecendo estruturas mais densas, enquanto o BIC impõe maior penalização, resultando em redes mais simples. O critério *Bayesian Dirichlet equivalent* (BDe) apresentou complexidade estrutural superior à obtida com BIC e inferior à observada com K2 e AIC.

Tabela 23: Características estruturais das Redes Bayesianas aprendidas.

Score	**Arcos**	**Tamanho médio do Markov Blanket**	**Tamanho médio da Vizinhança**	**Fator médio de Ramificação**
BIC	75	9,4	7,0	3,5
AIC	91	11,6	8,6	4,3
K2	90	11,4	8,5	4,2
BDe	81	10,5	7,7	3,8

Dentre as redes geradas, a estrutura aprendida pelo algoritmo HC sob o critério BIC foi selecionada para a análise interpretativa por apresentar maior parcimônia. Esse atributo corresponde à representação das dependências entre variáveis com menor número de conexões, favorecendo uma estrutura mais simples e interpretável. A adoção de modelos parcimoniosos é recomendada quando a clareza das relações probabilísticas é fundamental para a compreensão do fenômeno analisado. A Figura 19 apresenta a estrutura selecionada. No grafo, as cores dos nós representam a categoria meteorológica de cada variável, enquanto as arestas direcionadas representam as dependências condicionais retidas na estrutura aprendida. Uma visualização interativa online da rede também está disponível na página do projeto (visualização do grafo da RB). O significado das abreviações utilizadas encontra-se na Tabela 14. As demais estruturas aprendidas pelos algoritmos HC-AIC, HC-K2 e HC-BDe estão apresentadas nos Apêndices H a J.

A estrutura resultante é composta por nós que representam variáveis meteorológicas observadas em diferentes defasagens temporais, bem como fatores sazonais associados ao período do dia e à estação do ano. A variável de interesse, Precipitação ("Rain"), localiza-se na parte inferior da rede e possui como nós pais "R_1h", "H_1h" e "W_1h", correspondentes à radiação global, umidade relativa e velocidade do vento com defasagem de uma hora. Tal configuração indica que a ocorrência de chuva está associada predominantemente a condições atmosféricas recentes, indicando a relevância de variáveis observadas no curto prazo.

A presença de "R_1h" como nó pai de "Rain" é coerente com o papel da radiação solar nos processos convectivos, especialmente em regiões tropicais, nas quais o aquecimento da superfície intensifica a instabilidade atmosférica e favorece a formação de nuvens e precipitação. A variável "W_1h" está associada ao transporte de umidade e à dinâmica de sistemas atmosféricos que podem contribuir para a organização e manutenção de eventos de chuva. Por sua vez, "H_1h" representa a disponibilidade de vapor de água na atmosfera, elemento essencial para a condensação e o desenvolvimento de nuvens.

A rede também captura dependências coerentes com o comportamento físico da atmosfera, como conexões consistentes entre temperatura, umidade e ponto de orvalho ao longo das defasagens temporais, além de modelar adequadamente fatores sazonais, como o período do dia e a estação do ano, como nós raiz.

Verifica-se também a conectividade entre níveis temporais consecutivos de uma mesma variável, caracterizando a persistência temporal. Tal padrão é típico de séries meteorológicas, nas quais estados passados influenciam diretamente os estados futuros imediatos, reforçando a consistência física e a interpretabilidade da dinâmica atmosférica local. De modo geral, a estrutura aprendida apresenta representação interpretável e fisicamente consistente da dinâmica atmosférica local.

Embora a estrutura seja coerente com o conhecimento físico e a dinâmica atmosférica local, as direções das arestas representam apenas dependências estatísticas e não implicam causalidade. A existência de conexões indica uma associação probabilística, mas não uma relação de causa e efeito. Conforme discutido por (PEARL et al., 2000), a inferência causal requer a consideração de intervenções e modelos que distingam observação de manipulação, o que não foi contemplado na abordagem adotada, conforme detalhado na Seção 2.1.1. Assim, a estrutura inferida deve ser classificada como uma RB não causal, adequada à modelagem de incertezas e à realização de inferências probabilísticas.

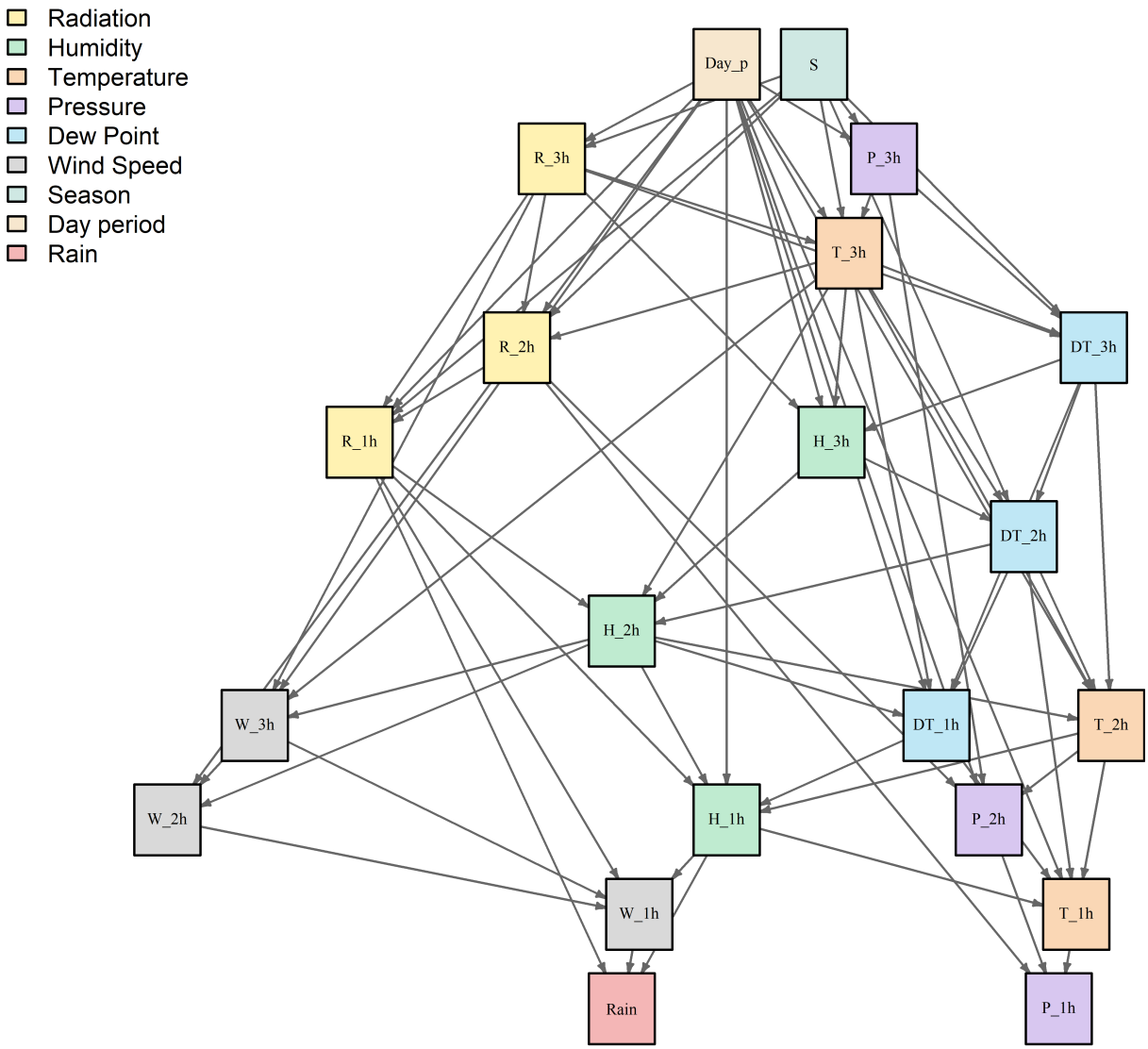

Figura 19: Rede Bayesiana aprendida utilizando o algoritmo Hill Climbing e o critério BIC para modelagem probabilística da precipitação.

A interpretabilidade da RB apresentada na Figura 19 decorre da possibilidade de analisar explicitamente a estrutura de dependências condicionais entre as variáveis meteorológicas. Nesse grafo, cada nó representa uma variável do sistema atmosférico, enquanto cada arco direcionado indica uma relação de dependência probabilística condicionada às demais variáveis da rede. Assim, a presença de uma aresta entre duas variáveis sugere que a informação contida em uma delas contribui para atualizar a distribuição de probabilidade da outra, sem que isso implique, necessariamente, uma relação causal direta.

Do ponto de vista interpretativo, a análise da rede pode ser realizada em três níveis complementares. Primeiro, a presença de arcos permite identificar quais variáveis apresentam associação probabilística direta com a precipitação. Segundo, a ausência de determinadas conexões sugere independências condicionais, isto é, indica que, dadas as

demais variáveis consideradas na rede, algumas variáveis não acrescentam informação direta relevante para explicar o nó de precipitação. Terceiro, a vizinhança local da variável de interesse, especialmente seu MB, permite identificar o conjunto mínimo de variáveis necessário para atualizar probabilisticamente a ocorrência de chuva.

Em uma aplicação prática, como um sistema de apoio à decisão para monitoramento meteorológico, a RB poderia ser utilizada para atualizar a probabilidade de precipitação à medida que novas observações fossem disponibilizadas pelas estações telemétricas. Por exemplo, ao observar um cenário com umidade elevada, baixa radiação solar e vento moderado na hora anterior, o modelo permite recalcular a probabilidade de chuva e fornecer uma explicação transparente para esse aumento de risco.

7.2 Avaliação Quantitativa dos Modelos

A presente Seção apresenta os resultados da avaliação quantitativa das RBs construídas, com base em métricas que permitem comparar o desempenho preditivo e a adequação probabilística das estruturas aprendidas.

A Tabela 24 apresenta os resultados da validação cruzada *k-fold* com $k = 5$. Valores em negrito indicam o melhor desempenho médio em cada métrica. Foram utilizadas três métricas para avaliar o desempenho dos modelos, acurácia, sensibilidade e especificidade. A acurácia corresponde à proporção total de previsões corretas. A sensibilidade representa a proporção de chuva corretamente identificadas. A especificidade indica a proporção de momentos sem chuva corretamente classificados. Mais detalhes sobre essas técnicas podem ser consultadas na Seção 2.2.6.1.

Entre os modelos avaliados, HC K2 apresentou a maior acurácia média, igual a 0,7416, e a maior especificidade média, igual a 0,7481, indicando melhor desempenho e maior capacidade de identificar momentos sem precipitação. O modelo HC AIC apresentou a maior sensibilidade média, igual a 0,7530, evidenciando maior capacidade de detecção de horas com chuva, com redução na identificação de horas sem precipitação.

O modelo baseado no critério BIC, embora mais parcimonioso, apresentou sensibilidade inferior ao modelo baseado em AIC e especificidade ligeiramente inferior ao modelo baseado em K2. Achados semelhantes foram reportados por (PUTRI; WIJAYANTO, 2021), que também observaram alto desempenho do critério BIC na estruturação de RB para previsão de chuva.

Os modelos HC BIC e HC BDe apresentaram valores idênticos em todas as métricas. Essa equivalência é coerente com a parametrização adotada para o escore BDe, sem incorporação de priori informativa, o que reduz a influência do termo de priori na pontuação.

Tabela 24: Resultados da validação cruzada.

Rede Bayesiana	Métrica	Fold 1	Fold 2	Fold 3	Fold 4	Fold 5	Média (DP)
HC - BIC	Acurácia	0,7275	0,7174	0,7194	0,7192	0,7166	0,7186 (0,001)
	Sensibilidade	0,6921	0,6914	0,6915	0,6945	0,6934	0,6926 (0,001)
	Especificidade	0,7230	0,7196	0,7219	0,7214	0,7186	0,7209 (0,001)
HC - AIC	Acurácia	0,6828	0,6820	0,6750	0,6748	0,6676	0,6764 (0,006)
	Sensibilidade	**0,7439**	**0,7458**	**0,7552**	**0,7533**	**0,7672**	**0,7530** (0,009)
	Especificidade	0,6775	0,6764	0,6680	0,6680	0,6589	0,6697 (0,007)
HC - K2	Acurácia	**0,7419**	**0,7481**	**0,7499**	**0,7463**	**0,7218**	**0,7416** (0,011)
	Sensibilidade	0,6635	0,6563	0,6553	0,6693	0,6923	0,6673 (0,015)
	Especificidade	**0,7487**	**0,7561**	**0,7581**	**0,7530**	**0,7244**	**0,7481** (0,013)
HC - BDe	Acurácia	0,7275	0,7174	0,7194	0,7192	0,7166	0,7186 (0,001)
	Sensibilidade	0,6921	0,6914	0,6915	0,6945	0,6934	0,6926 (0,001)
	Especificidade	0,7230	0,7196	0,7219	0,7214	0,7186	0,7209 (0,001)

A Figura 20 apresenta as curvas ROC dos quatro modelos avaliados. Conforme discutido na Seção 2.2.6.3, a curva ROC permite avaliar a capacidade discriminativa dos modelos ao considerar diferentes limiares de classificação. Nota-se que todas as curvas situam-se acima da reta identidade, o que evidencia a capacidade discriminativa dos modelos em prever com desempenho superior ao acaso.

O modelo baseado no critério K2 apresentou o melhor desempenho, com área sob a curva igual a 0,790, indicando maior capacidade de discriminação entre eventos de chuva e de não chuva. O modelo HC AIC também apresentou desempenho satisfatório, com área sob a curva igual a 0,785, valor ligeiramente inferior ao obtido pelo K2.

Já os modelos HC-BIC e HC-BDe exibiram desempenho idêntico, com área sob a curva de 0,766, compatível com os resultados anteriormente observados na validação cruzada.

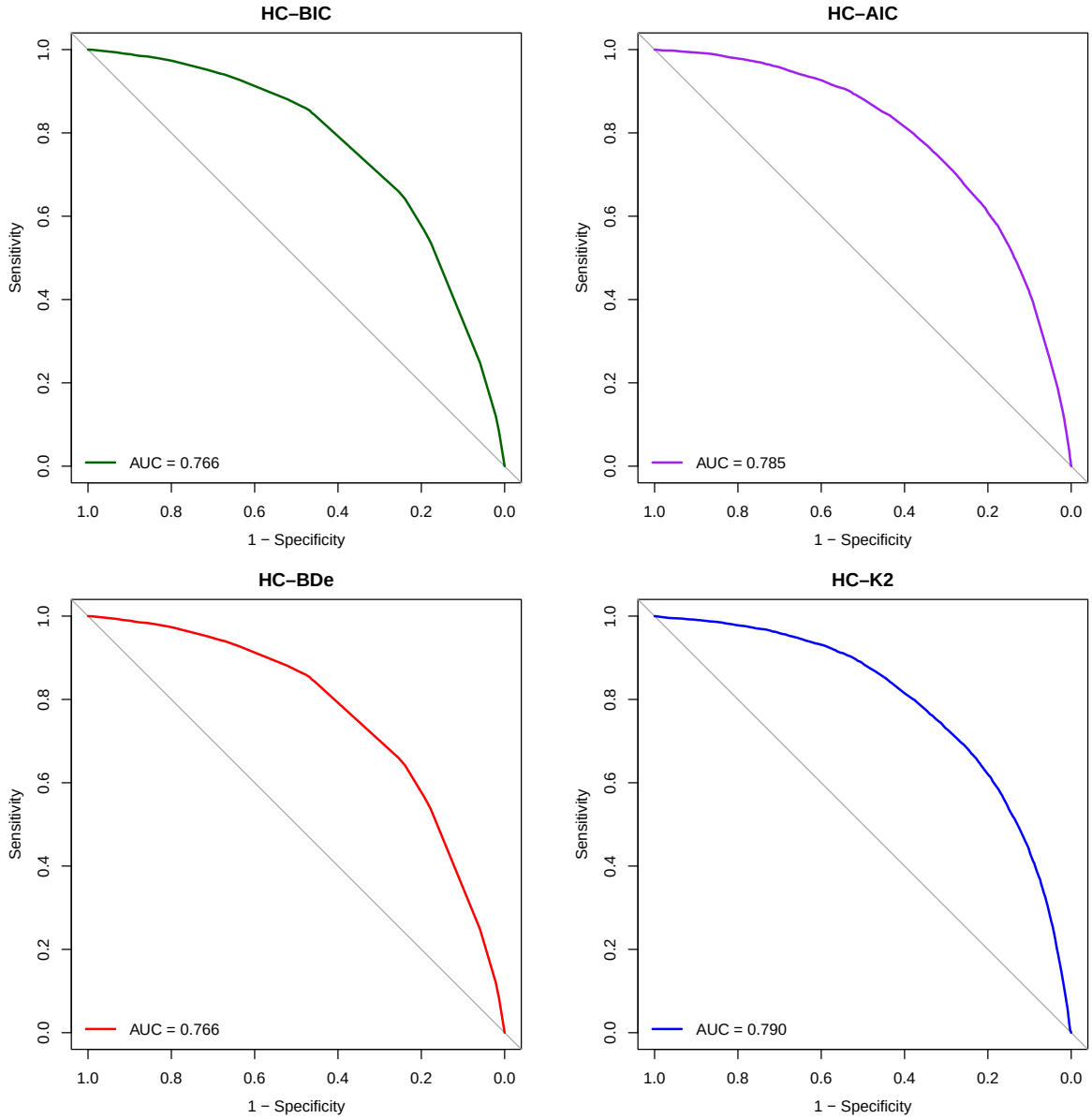

Figura 20: Curvas ROC dos modelos de Redes Bayesianas (BDe, K2, BIC, AIC).

A Figura 21 apresenta as curvas de calibração dos modelos avaliados. Essas curvas comparam as probabilidades previstas de ocorrência de chuva com as frequências efetivamente observadas, permitindo avaliar se os modelos estão superestimando ou subestimando o risco de precipitação. Idealmente, os pontos da curva devem alinhar-se à reta identidade, indicando perfeita correspondência entre predição e observação.

Nas curvas de calibração, a interceptação "A" representa o viés geral das previsões: valores negativos indicam subestimação das probabilidades de chuva, enquanto valores positivos sugerem superestimação. O coeficiente angular "B" expressa o grau de dispersão das previsões em relação ao ideal; valores próximos de 1 indicam calibração perfeita, enquanto valores menores refletem previsões menos sensíveis às variações reais de ocor-

rência.

De modo geral, todos os modelos apresentaram calibração semelhante, com curvas abaixo da reta identidade, indicando tendência à subestimação das probabilidades de chuva. Os modelos BIC e BDe exibiram resultados idênticos, com A = -0,063 e B = 0,377. O modelo K2 apresentou A = -0,058 e B = 0,360, enquanto o modelo AIC apresentou A = -0,049 e B = 0,336, com leve dispersão nos extremos. Apesar das pequenas diferenças nos valores de intercepto e inclinação, os quatro modelos apresentaram padrões próximos de calibração.

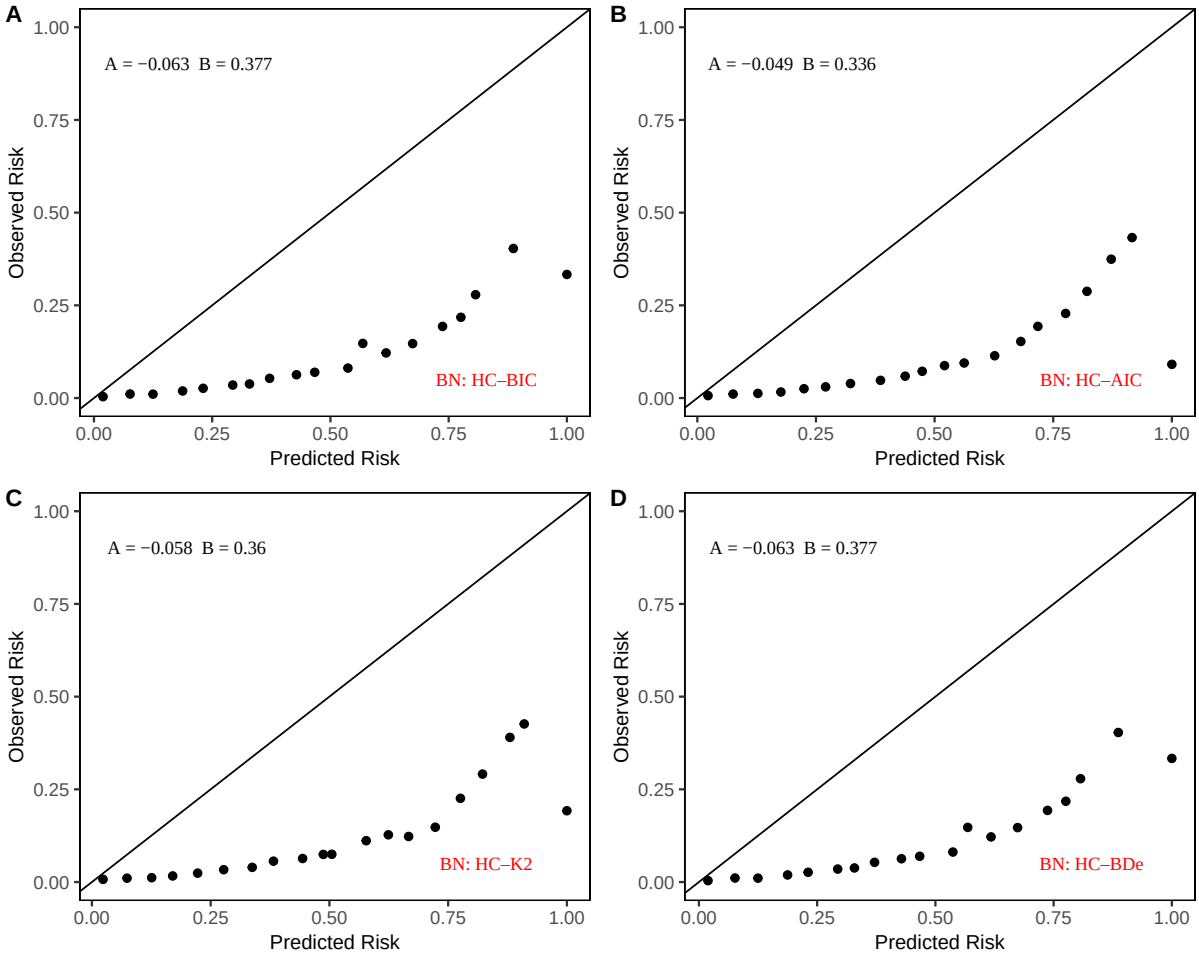

Figura 21: Curvas de calibração dos modelos de Redes Bayesianas (BIC, AIC, K2, BDe).

A Figura 22 apresenta a comparação da perda por log-verossimilhança negativa dos modelos aprendidos pelo algoritmo HC sob diferentes critérios de pontuação. O modelo otimizado pelo critério AIC apresentou o menor valor de perda, igual a 12,24, seguido por K2, com 12,25, BDe, com 12,29, e BIC, com 12,33.

Embora o critério AIC tenha apresentado o menor valor de perda, as diferenças observadas foram pequenas, indicando desempenho semelhante entre os modelos quanto ao

ajuste probabilístico aos dados. Essa proximidade sugere consistência entre as estruturas aprendidas e indica que as variações decorrentes dos diferentes critérios de pontuação não impactaram de forma relevante a qualidade do ajuste.

Observa-se que, apesar de os modelos HC-BIC e HC-BDe terem produzido resultados idênticos nas métricas preditivas, os valores de log-verossimilhança negativa apresentaram discreta diferença. Esse resultado decorre do fato de que as estruturas aprendidas não são idênticas, ainda que apresentem desempenho preditivo equivalente. Assim, a função de perda mostra-se sensível às diferenças estruturais induzidas pelo critério de pontuação adotado no aprendizado, mesmo quando tais variações não se refletem diretamente nas métricas de classificação.

Ao analisar os dados apresentados na Tabela 23, observa-se que o modelo HC-BIC resultou na estrutura mais parcimoniosa, com 75 arcos, enquanto os modelos HC-BDe, HC-K2 e HC-AIC apresentaram, respectivamente, 81, 90 e 91 arcos. Essa variação no número de conexões reflete diretamente na densidade estrutural das redes, sendo que modelos mais densos, ou seja, com maior número de arcos, tendem a proporcionar melhor ajuste probabilístico aos dados, como evidenciado pela redução dos valores da função de perda por log-verossimilhança.

Entretanto, o ganho em ajuste deve ser analisado em conjunto com a complexidade estrutural. O modelo HC-AIC apresentou a menor perda, igual a 12,24, com 91 arcos, enquanto HC-BIC obteve perda de 12,33 com 75 arcos. A diferença de 0,09 unidades é numericamente baixa frente ao acréscimo de 16 arcos, o que indica que a melhoria marginal no ajuste não compensa o aumento da complexidade, sobretudo quando se priorizam interpretabilidade e parcimônia na modelagem por RB. Além disso, as diferenças observadas nas demais métricas foram relativamente pequenas. Na análise das curvas ROC, por exemplo, os modelos HC-AIC e HC-K2 apresentaram AUCs de 0,785 e 0,790, respectivamente, enquanto HC-BIC e HC-BDe apresentaram AUC de 0,766. De forma semelhante, as curvas de calibração indicaram comportamentos próximos entre os diferentes critérios de pontuação, sem evidência suficiente para destacar um modelo como claramente superior aos demais nesse aspecto.

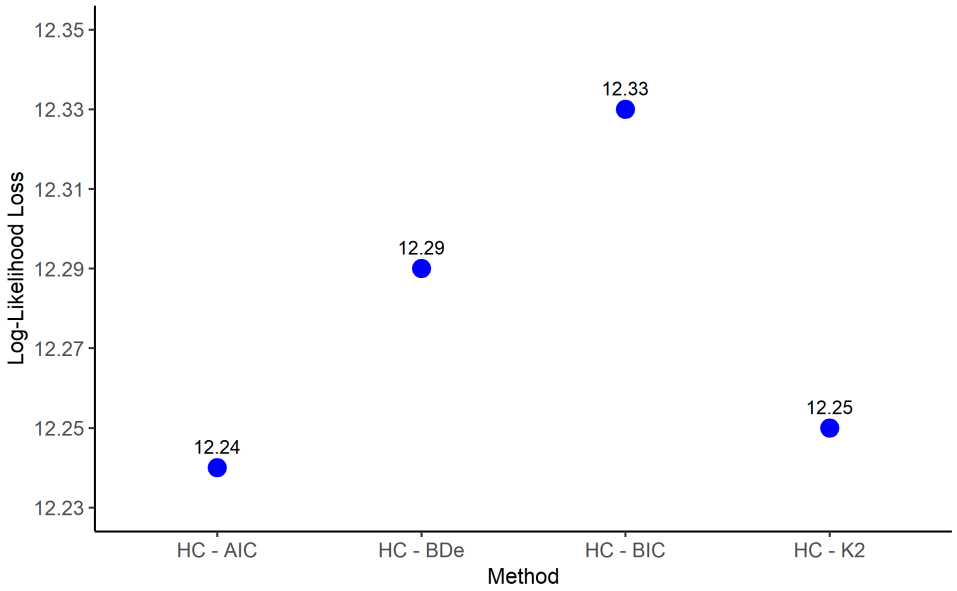

Figura 22: Perda da log-verossimilhança dos modelos de Redes Bayesianas (BIC, AIC, K2, BDe).

7.3 Inferências

A presente Seção teve como objetivo quantificar a influência conjunta das variáveis meteorológicas na ocorrência de precipitação com base na RB estruturada e parametrizada. Essa análise permitiu avaliar as probabilidades condicionais de ocorrência de chuva e de chuva intensa em função dos níveis das variáveis meteorológicas sob diferentes cenários atmosféricos. Embora a categoria de chuva intensa tenha sido excluída da fase de avaliação em razão de sua baixa frequência no conjunto de dados, ela foi incluída na etapa de inferência com o objetivo de explorar o comportamento probabilístico do modelo sob determinadas condições atmosféricas.

A Figura 23 apresenta um resumo das probabilidades condicionais estimadas para as variáveis que compõem o MB da variável "precipitação", a saber "Umidade", "Radiação" e "Vento", todas consideradas com defasagem de 1 hora. Mais detalhes sobre o conceito de MB podem ser encontrados na Seção 2.1.6.

Para a leitura do gráfico, cada coluna representa uma variável meteorológica e cada linha corresponde a um nível da variável "precipitação", isto é, "Sem Chuva", "Chuva" e "Chuva Intensa". As barras indicam a probabilidade condicional da variável de precipitação dado um determinado nível da variável meteorológica no instante anterior $(t-1)$. Por exemplo, observa-se que a probabilidade de ocorrência de "Chuva" quando a "Umidade" estava no nível alto na hora anterior é de aproximadamente 13,32%, enquanto a proba-

bilidade de "Sem Chuva" nessa mesma condição é de cerca de 86,61%. Nota-se que as maiores probabilidades de ocorrência de chuva e de chuva intensa estão associadas a condições observadas na hora anterior, caracterizadas por alta umidade, baixa radiação e vento moderado. Esse resultado é consistente com a dinâmica da formação de precipitação, uma vez que elevados níveis de umidade indicam maior disponibilidade de vapor d'água na atmosfera, condição necessária para a formação de nuvens e subsequente precipitação.

Adicionalmente, baixos níveis de radiação solar estão associados à maior cobertura de nuvens, frequentemente relacionados à ocorrência de chuva (WALLACE; HOBBS, 2006; TANG et al., 2022; LYU, 2025). No caso da velocidade do vento, observa-se que valores moderados estão associados às maiores probabilidades de chuva. Esse comportamento pode estar relacionado ao transporte de umidade e à organização de sistemas atmosféricos que favorecem a formação de precipitação (PARLANGE; KATZ, 2000; GIMENO et al., 2012).

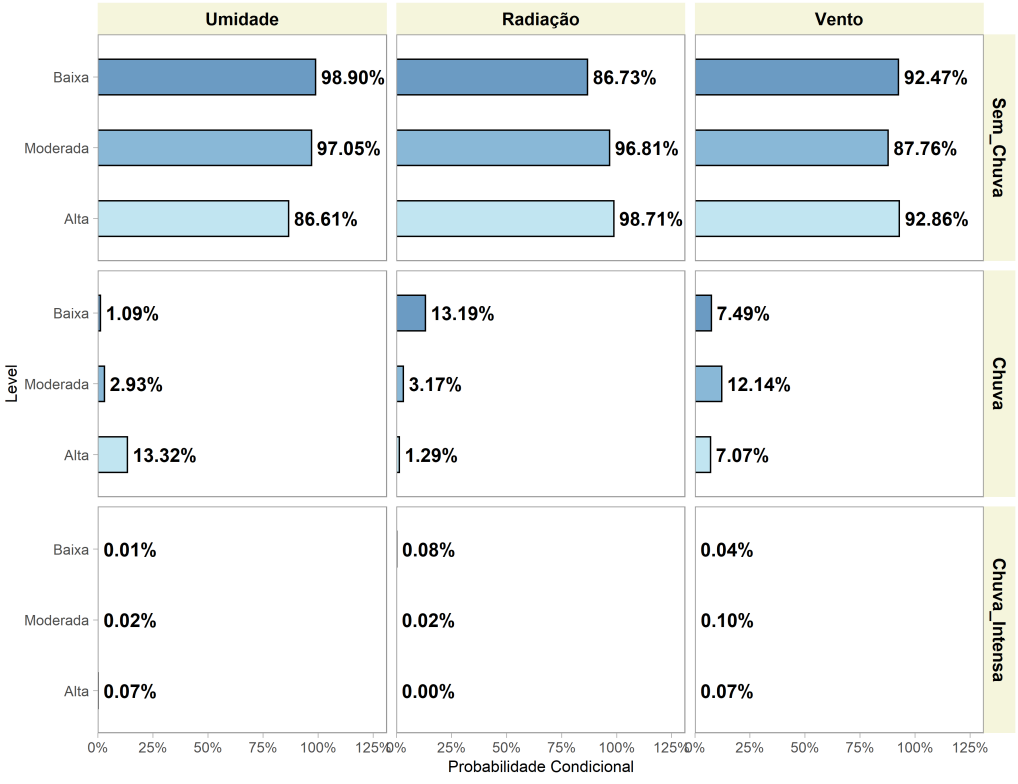

Figura 23: Distribuição de probabilidade condicional da precipitação considerando defasagem de 1 h.

Para detalhar o efeito combinado entre as variáveis, a Figura 24 ilustra as probabilidades condicionais de "precipitação" dadas as variáveis "Vento" (defasagem de 1 h) e "Umidade" (defasagem de 1 h), considerando a variável "Vento" fixada na categoria mo-

derada. Nessa análise, assume-se que o vento já foi observado em nível moderado, e o objetivo é examinar como a umidade influencia a ocorrência de precipitação sob esse regime específico de vento. Ao condicionar a variável "Vento", as probabilidades resultantes refletem o efeito combinado da "Umidade" dado que o vento se encontra em condições moderadas.

Observa-se que a probabilidade de ocorrência de "chuva" aumenta significativamente com o aumento da "Umidade" quando a velocidade do "Vento" é moderada. Sob condições de "Vento" moderado e "Umidade" alta, a probabilidade de "chuva" atinge aproximadamente 26,24%, enquanto sob "Umidade" baixa permanece inferior a 1%.

Esse padrão indica que a saturação de vapor d'água na atmosfera constitui um fator fundamental para o início da precipitação. O "Vento" moderado favorece o transporte de umidade, sem, contudo, dificultar a formação de convecção, como poderia ocorrer em condições de vento muito intenso.

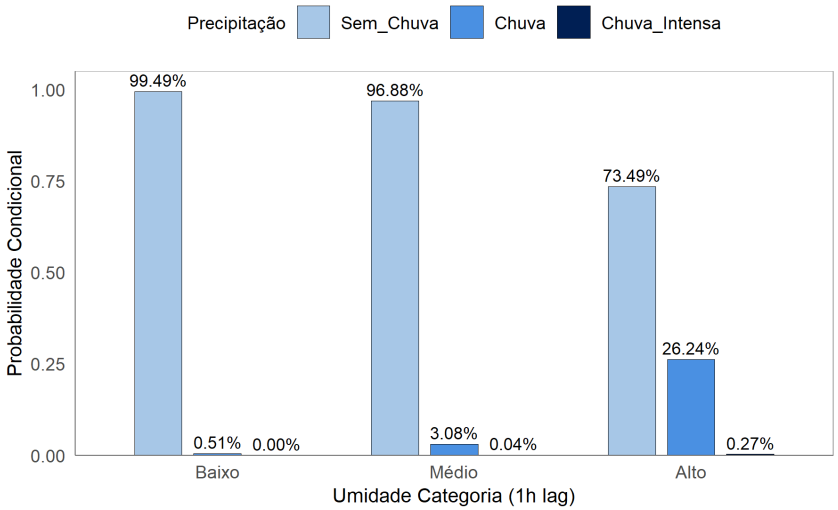

Figura 24: Probabilidade condicional de precipitação por nível de umidade (defasagem de 1 h), dado vento moderado (defasagem de 1 h).

De forma análoga, a Figura 25 apresenta as probabilidades condicionais de "precipitação" considerando as variáveis "Radiação" (defasagem de 1 h) e "Umidade" (defasagem de 1 h), quando os níveis de "Radiação" são baixos. Observa-se que, sob condições de baixa radiação solar, níveis mais elevados de "Umidade" aumentam significativamente a probabilidade de ocorrência de "chuva" e "chuva intensa". Sob condições de baixa radiação solar e "Umidade" alta, a probabilidade de "chuva" atinge aproximadamente 19,25%. Em contraste, quando a "Umidade" e a radiação solar são baixas, a probabilidade de "chuva" permanece inferior a 2%.

A combinação de baixa radiação e maior "Umidade" observada uma hora antes indica que o processo de formação de nuvens provavelmente já estava em andamento, reduzindo a quantidade de radiação que atinge a superfície. Esse intervalo de uma hora é possivelmente suficiente para que as nuvens se desenvolvam ainda mais e produzam precipitação.

Em conjunto, essas inferências reforçam a consistência física e a interpretabilidade da RB construída. O modelo reproduz padrões atmosféricos coerentes, indicando que a combinação de "Umidade" elevada, "Radiação" baixa e "Vento" moderado representa o cenário mais provável para a ocorrência de chuva. Dessa forma, a rede demonstra seu potencial como ferramenta de apoio para a análise e a previsão probabilística de eventos de precipitação, preservando a transparência das relações condicionais entre os principais fatores meteorológicos.

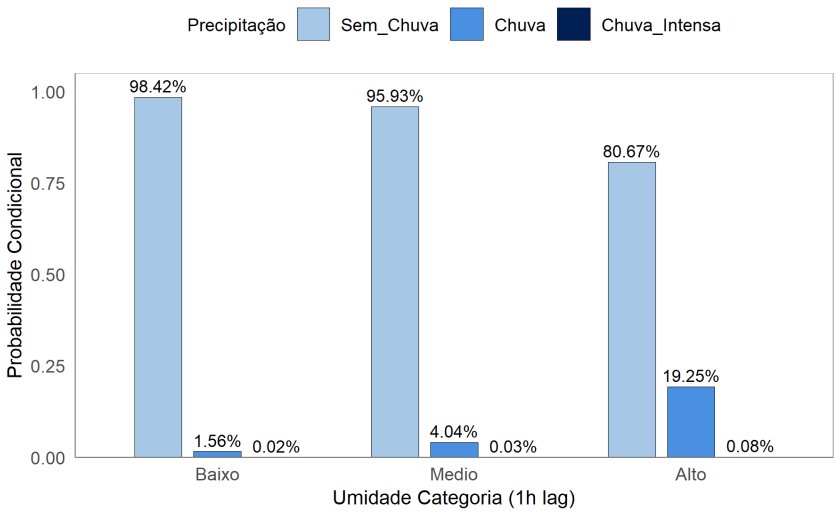

Figura 25: Probabilidade condicional de precipitação por nível de umidade (defasagem de 1 h), dada radiação baixa (defasagem de 1 h).

7.4 Consumo de Energia e Impacto Ambiental

Esta Seção apresenta as estimativas de consumo de energia e pegada hídrica associadas aos experimentos de aprendizado estrutural. As métricas foram obtidas com o auxílio do wAIter calculator, ferramenta proposta por (BREDER; FERRO, 2025) para estimar o impacto ambiental de experimentos computacionais. Conforme descrito na Seção 5.3, o cálculo dessas métricas considera o consumo energético dos experimentos e estima, de forma indireta, o volume de água associado à geração dessa energia.

A Tabela 25 apresenta o tempo de execução, o consumo estimado de energia e a

correspondente pegada hídrica para cada modelo obtido pelo algoritmo HC sob diferentes critérios de *score* avaliados neste estudo. Essas medições referem-se especificamente à etapa de construção das RBs.

Os resultados indicaram um consumo de energia aproximadamente igual entre os algoritmos, em torno de 0,05 kWh para cada execução. Embora os valores de consumo total de energia tenham sido próximos entre os experimentos, observaram-se pequenas, porém mensuráveis, variações na pegada hídrica estimada.

Este experimento foi executado durante 3 horas e 1 minuto em um processador AMD Ryzen 5 5600G, com consumo de 0,20 kWh, resultando em uma pegada estimada de 0,02 kgCO_2e e 3,65 L de água no Brasil, conforme calculado com o *wAIter* (versão beta) (BREDER; FERRO, 2025). Em termos ilustrativos, esse consumo energético é aproximadamente equivalente a manter uma lâmpada de 60 W acesa por cerca de 3 horas e 20 minutos, enquanto a pegada hídrica corresponde a aproximadamente sete garrafas de água.

Tabela 25: Consumo estimado de energia e pegada hídrica dos experimentos de aprendizado da estrutura da Rede Bayesiana.

HC - Score	**Tempo de execução (s)**	**Energia (kWh)**	**Pegada hídrica (L)**
BIC	2522	0,05	0,85
AIC	2964	0,05	0,99
K2	3030	0,05	1,02
BDe	2354	0,05	0,79
Total	10870	0,20	3,65

A inclusão dessa análise é compromisso fundamental do projeto RioNowcast+Green, onde esta pesquisa se insere, pois em projetos como estes que visam diminuir os impactos das mudanças climáticas, o próprio processo de desenvolvimento de modelos de IA pode significar uma pegada ambiental significativa. Além disso, a utilização de modelos mais leves, como os baseados em RBs, reduz a necessidade de recursos computacionais e torna sua aplicação mais viável em cidades e regiões com infraestrutura limitada. Isso contribui para diminuir desigualdades e injustiças climáticas, especialmente em países ou prefeituras que não têm acesso a equipamentos de alto desempenho. Portanto, torna-se cada vez mais urgente adotar práticas e metodologias que considerem não apenas a acurácia dos modelos, mas também sua sustentabilidade destes modelos.

8 Conclusão

Este trabalho teve como objetivo analisar se a modelagem por RBs é capaz de representar, de forma interpretável, as relações probabilísticas entre variáveis meteorológicas e a ocorrência de precipitação no município do Rio de Janeiro. À luz dos resultados obtidos, conclui-se que essa abordagem se mostrou capaz de cumprir esse propósito, ao produzir estruturas probabilísticas coerentes do ponto de vista físico, interpretáveis em termos meteorológicos e úteis para a quantificação das dependências entre os fatores atmosféricos associados à chuva. A abordagem adotada integrou dados observacionais de longa duração, técnicas de pré-processamento e métodos de aprendizado estrutural para avaliar essa capacidade de representação em um contexto urbano tropical.

Os resultados obtidos demonstraram que as RBs constituem modelos eficientes e interpretáveis para representar relações probabilísticas entre variáveis meteorológicas associadas à precipitação. Esses resultados evidenciam o potencial das RBs como ferramentas adequadas para a modelagem de sistemas atmosféricos caracterizados por múltiplas interações e elevados níveis de incerteza.

A comparação entre diferentes algoritmos de aprendizado estrutural evidenciou diferenças relevantes tanto na topologia das estruturas aprendidas quanto no desempenho preditivo dos modelos. Entre os modelos avaliados, o algoritmo HC com o critério K2 apresentou o melhor desempenho preditivo nas métricas de acurácia e especificidade. Entretanto, a estrutura obtida pelo algoritmo HC com o critério BIC foi selecionada para análise interpretativa devido à sua maior parcimônia estrutural. Essa relação entre simplicidade estrutural e otimização preditiva, ilustrada pelo contraste entre os modelos HC-BIC e HC-K2, é consistente com achados comparativos na literatura de aprendizado de estrutura de RBs, que indicam que diferentes critérios de pontuação podem produzir modelos com comportamento preditivo semelhante, porém com diferentes níveis de complexidade gráfica (SCUTARI; GRAAFLAND; GUTIÉRREZ, 2019; KITSON et al., 2023).

As avaliações quantitativas confirmaram a robustez das estruturas aprendidas, com

todas as redes apresentando desempenho estável na validação cruzada. O modelo K2 e AIC apresentaram o melhor desempenho preditivo entre os modelos avaliados, enquanto as curvas de calibração e curvas ROC indicaram comportamentos semelhantes entre os diferentes critérios de pontuação. Além disso, os valores próximos de perda por log-verossimilhança sugerem que as estruturas aprendidas apresentaram desempenho probabilístico globalmente comparável.

Em termos de robustez frente à incerteza, a estratégia de consenso baseada em *bootstrap* mostrou-se consistente com estudos anteriores que demonstram que procedimentos de reamostragem aumentam a confiabilidade das arestas e reduzem dependências espúrias em grafos aprendidos (CARAVAGNA; RAMAZZOTTI, 2021; BARTH; SERRÃO; MACIEL, 2024). A estabilidade das métricas obtidas na validação cruzada reforça a consistência das dependências probabilísticas inferidas pelo modelo.

A análise da RB aprendida indicou que a ocorrência de precipitação está principalmente associada a condições atmosféricas de curto prazo, especialmente níveis elevados de umidade relativa, baixos níveis de radiação solar e velocidades moderadas do vento com defasagem de uma hora. Essas relações mostraram-se consistentes com o conhecimento meteorológico estabelecido, sugerindo que a RB capturou padrões fisicamente plausíveis presentes nos dados observacionais.

Do ponto de vista físico, o modelo reproduziu padrões coerentes com a instabilidade atmosférica típica de ambientes tropicais úmidos. A identificação da umidade relativa como principal fator associado à ocorrência de precipitação, combinada com ventos moderados que favorecem o transporte de umidade e com a redução da radiação solar, indicando maior cobertura de nuvens e resfriamento da superfície, configura um padrão característico de instabilidade atmosférica em regiões costeiras.

A etapa de inferência probabilística permitiu quantificar como diferentes combinações de estados das variáveis meteorológicas influenciam a probabilidade de ocorrência de precipitação. Essa análise demonstrou o potencial das RBs para apoiar a interpretação probabilística de fenômenos meteorológicos e contribuir para a compreensão dos mecanismos associados à formação de chuva em ambientes urbanos tropicais.

Em termos de interpretabilidade, as RBs apresentam vantagens claras em relação a modelos determinísticos ou "*black-box*", ao representar explicitamente dependências condicionais e permitir compreensão transparente da influência probabilística de cada fator meteorológico na probabilidade de precipitação. Os resultados obtidos são consistentes com estudos anteriores baseados em RB que identificaram fatores meteorológicos associ-

ados à ocorrência de precipitação (DAS; CHANDA, 2020; PUTRI; WIJAYANTO, 2021). No presente estudo, elevada umidade relativa, baixa radiação solar e ventos moderados em curtas defasagens formaram o padrão mais informativo, ampliando evidências anteriores ao contextualizar essas relações para o município do Rio de Janeiro com base em observações horárias de longo prazo obtidas em estações meteorológicas.

Embora a discretização das variáveis implique alguma perda de granularidade na representação dos dados, os resultados indicam que as RBs constituem uma abordagem promissora e transparente para a estimação probabilística da precipitação. Ao integrar conhecimento de domínio com aprendizado orientado por dados, esses modelos oferecem suporte a processos de tomada de decisão operacional por meio de previsões interpretáveis.

Do ponto de vista metodológico, observou-se que métodos de aprendizado de estrutura baseados em restrições e híbridos apresentaram dificuldades em estabelecer ligações diretas entre preditores meteorológicos e precipitação, contrastando com resultados relatados por (PUTRI; WIJAYANTO, 2021). Uma possível explicação reside na presença de associações relativamente fracas entre os preditores e a variável alvo, além do desbalanceamento de classes no conjunto de dados, fatores que podem reduzir a estabilidade dos testes locais de independência condicional e favorecer, nesse contexto, abordagens baseadas em pontuação com otimização global.

Adicionalmente, este estudo introduziu uma dimensão relevante de sustentabilidade na avaliação de modelos meteorológicos. A baixa pegada computacional das RBs, evidenciada pelo reduzido consumo de energia e impacto ambiental nos experimentos, posiciona esses modelos como uma alternativa de *Green AI* em comparação com modelos de aprendizado profundo computacionalmente intensivos. Essa eficiência torna-se particularmente relevante para aplicações operacionais de previsão em regiões nas quais os recursos computacionais podem ser limitados.

Diante dos resultados apresentados, as principais **contribuições** deste trabalho podem ser sintetizadas nos seguintes pontos:

- Construção de uma estrutura gráfica que representa explicitamente as dependências probabilísticas entre variáveis meteorológicas associadas à ocorrência de precipitação, permitindo visualizar e interpretar as relações condicionais entre os fatores atmosféricos.

- Identificação de padrões atmosféricos associados à ocorrência de precipitação, destacando a combinação de elevada umidade relativa, baixos níveis de radiação solar

e velocidades moderadas do vento em curtas defasagens temporais.

- Quantificação probabilística das relações entre variáveis meteorológicas por meio de inferência em RBs, contribuindo para a interpretação probabilística dos mecanismos associados à formação de precipitação em ambientes urbanos tropicais.
- Evidência empírica sobre o desempenho de diferentes algoritmos de aprendizado estrutural de RBs em dados meteorológicos de longa duração do Rio de Janeiro, permitindo identificar abordagens mais adequadas para representar relações probabilísticas associadas à precipitação.

Apesar dessas contribuições, o estudo apresenta algumas **limitações**:

- A baixa frequência de eventos de precipitação intensa no conjunto de dados dificultou a caracterização adequada dessa classe, o que limita a robustez das inferências probabilísticas em cenários de maior impacto.
- Muitos dos padrões atmosféricos identificados confirmam conhecimentos meteorológicos já estabelecidos na literatura, de modo que a principal contribuição da RB consiste na formalização e quantificação probabilística desse conhecimento especializado.
- A discretização das variáveis meteorológicas implica perda parcial de granularidade na representação dos dados, podendo reduzir a capacidade do modelo de capturar variações contínuas mais sutis das condições atmosféricas.
- A estrutura proposta não modela explicitamente a evolução temporal das dependências probabilísticas entre variáveis atmosféricas. Embora defasagens temporais tenham sido incorporadas como atributos, a rede não constitui uma Rede Bayesiana Dinâmica, o que limita a representação de dependências temporais mais complexas ao longo do tempo.

Como perspectivas para **pesquisas futuras**:

- Integração de dados provenientes de diferentes fontes observacionais.
- Exploração de modelos híbridos que combinem RBs com outras técnicas de aprendizado de máquina, buscando integrar a interpretabilidade dos modelos probabilísticos com a capacidade preditiva de abordagens mais complexas.

- Desenvolvimento de estruturas Bayesianas espaço-temporais capazes de representar dependências entre diferentes estações meteorológicas e múltiplos horizontes de previsão.

- Incorporação de restrições estruturais fisicamente fundamentadas e *priors* informativos em RBs, visando melhorar a estimação probabilística de eventos de precipitação intensa preservando simultaneamente a interpretabilidade e a coerência física do modelo, em linha com abordagens recentes de *physics-informed machine learning* aplicadas a sistemas atmosféricos (KASHINATH et al., 2021; KODRA et al., 2020).

- Investigação de estratégias metodológicas específicas para a modelagem de eventos raros, com o objetivo de aprimorar a representação probabilística de episódios de precipitação extrema.

- Desenvolvimento de uma aplicação operacional para apoio à Defesa Civil e a sistemas de monitoramento meteorológico, com integração do modelo a uma API capaz de receber dados observacionais em tempo real e retornar estimativas da probabilidade de precipitação para diferentes cenários atmosféricos. Essa implementação permitiria avaliar, em contexto aplicado, como as RBs poderiam ser incorporadas a fluxos de apoio à decisão, especialmente em rotinas de vigilância e emissão de alertas.

- Desenvolvimento de estratégias de representação visual dos resultados voltadas a usuários não especialistas, de modo a traduzir as inferências probabilísticas da RB em interfaces mais acessíveis para uso institucional. Nesse contexto, podem ser explorados painéis visuais, mapas temáticos, escalas graduais de risco e indicadores que comuniquem, de forma objetiva, a probabilidade de ocorrência de precipitação e os fatores meteorológicos mais associados ao cenário previsto.

Em síntese, os resultados obtidos indicam que a modelagem por RBs foi capaz de integrar, de forma coerente, interações atmosféricas complexas em uma estrutura probabilística interpretável, combinando consistência física, rigor estatístico e eficiência computacional.Seu potencial de aplicação em sistemas de alerta precoce no município do Rio de Janeiro evidencia a viabilidade de abordagens probabilísticas transparentes e sustentáveis para a análise meteorológica.

8.1 Declaração Ética

Durante a elaboração deste trabalho, a ferramenta de inteligência artificial ChatGPT (OpenAI) foi utilizada como apoio às atividades de escrita e revisão textual. Em particular, o sistema foi empregado como auxílio na correção ortográfica e gramatical de trechos do texto, bem como no apoio à tradução e compreensão de artigos científicos originalmente publicados em língua inglesa.

O uso dessas ferramentas ocorreu unicamente como apoio, preservando a integridade e a autoria intelectual da pesquisa, em conformidade com os princípios de honestidade acadêmica da Universidade Federal Fluminense e da comunidade científica.

Nenhum conteúdo técnico ou analítico foi produzido automaticamente; todo o material científico foi elaborado pelo autor, com as sugestões das ferramentas passando por revisão, adaptação e validação criteriosa.

REFERÊNCIAS

ABDULNASSAR, AA; NAIR, Latha R. Performance analysis of Kmeans with modified initial centroid selection algorithms and developed Kmeans9+ model. **Measurement: Sensors**, Elsevier, v. 25, p. 100666, 2023.

AKAIKE, Hirotugu. A new look at the statistical model identification. **IEEE transactions on automatic control**, Ieee, v. 19, n. 6, p. 716–723, 2003.

BALAKRISHNAN, Narayanaswamy; VOINOV, Vassilly; NIKULIN, Mikhail Stepanovich. **Chi-squared goodness of fit tests with applications**. [S. l.]: Academic Press, 2013.

BANG-JENSEN, Jørgen; GUTIN, Gregory Z. **Digraphs: theory, algorithms and applications**. [S. l.]: Springer Science & Business Media, 2008.

BARTH, Vitor; SERRÃO, Fábio; MACIEL, Carlos. Assessing Credibility in Bayesian Networks Structure Learning. **Entropy**, MDPI, v. 26, n. 10, p. 829, 2024. Disponível em: <https://doi.org/10.3390/e26100829>.

BATES, Bryson C; DOWDY, Andrew J. Discerning the influence of climate variability modes, regional weather features and time series persistence on streamflow using Bayesian networks and multiple linear regression. **International Journal of Climatology**, Wiley Online Library, 2024.

BISHOP, Christopher M; NASRABADI, Nasser M. **Pattern recognition and machine learning**. [S. l.]: Springer, 2006. v. 4.

BRADLEY, Andrew P. The use of the area under the ROC curve in the evaluation of machine learning algorithms. **Pattern recognition**, Elsevier, v. 30, n. 7, p. 1145–1159, 1997.

BREDER, G.; FERRO, M. **wAIter: Serving Awareness - Environmental Footprint Estimator for Artificial Intelligence Models**. [S. l.]: Zenodo, 2025. DOI: 10.5281/zenodo.17247278.

BUNTINE, Wray. Theory refinement on Bayesian networks. In: ELSEVIER. UNCERTAINTY in artificial intelligence. [S. l.: s. n.], 1991. P. 52–60.

BUNTINE, Wray L. Operations for learning with graphical models. **Journal of artificial intelligence research**, v. 2, p. 159–225, 1994.

BUSSAB, Wilton de O; MORETTIN, Pedro A. **Estatística básica**. [S. l.: s. n.], 2010. P. xvi–540.

CANO, Rafael; SORDO, Carmen; GUTIÉRREZ, José M. Applications of Bayesian networks in meteorology. In: ADVANCES in Bayesian networks. [S. l.]: Springer, 2004. P. 309–328.

CARAVAGNA, Giulio; RAMAZZOTTI, Daniele. Learning the structure of Bayesian Networks via the bootstrap. **Neurocomputing**, v. 448, p. 48–59, 2021. ISSN 0925-2312. DOI: https://doi.org/10.1016/j.neucom.2021.03.071. Disponível em: <https://www.sciencedirect.com/science/article/pii/S0925231221004574>.

CARVALHO, Celso Santos; GALVÃO, Thiago. **Prevenção de riscos de deslizamentos em encostas em áreas urbanas**. Brasília: Instituto de Pesquisa Econômica Aplicada (IPEA), 2016. Disponível em: <http://repositorio.ipea.gov.br/handle/11058/9613>.

CASELLA, George; BERGER, Roger L. **Statistical Inference**. 2. ed. Pacific Grove, CA: Duxbury, 2002.

CHEN, Mengxuan et al. An interpretable weather forecasting model with separately-learned dynamics and physics neural networks. **Geophysical Research Letters**, Wiley Online Library, v. 52, n. 13, e2024gl114310, 2025.

CHOBTHAM, Kiattikun; CONSTANTINOU, Anthony C. Tuning structure learning algorithms with out-of-sample and resampling strategies. **Knowledge and Information Systems**, Springer, v. 66, n. 8, p. 4927–4955, 2024. Disponível em: <https://doi.org/10.1007/s10115-024-02111-9>.

CHRISTENSEN, Hannah; DELAUNAY, Antoine. Interpretable Deep Learning for Probabilistic MJO Prediction. In: EGU General Assembly Conference Abstracts. [S. l.: s. n.], 2022. egu22–12720.

COFINO, Antonio S et al. Bayesian networks for probabilistic weather prediction. In: CITESEER. 15TH Eureopean Conference on Artificial Intelligence (ECAI). [S. l.: s. n.], 2002.

CONTALDI, Carlo. **Bayesian Network Hybrid Learning Using a Parent Reducing Site-specific Mutation Rate Genetic Algorithm**. 2016. Tese (Doutorado) – University of Illinois at Chicago.

COOPER, Gregory F; HERSKOVITS, Edward. A Bayesian method for the induction of probabilistic networks from data. **Machine learning**, Springer, v. 9, p. 309–347, 1992.

ÇORBACIOĞLU, Şeref Kerem; AKSEL, Gökhan. Receiver operating characteristic curve analysis in diagnostic accuracy studies: A guide to interpreting the area under the curve value. **Turkish journal of emergency medicine**, Medknow, v. 23, n. 4, p. 195–198, 2023.

COSTA, Alexander Josef Sá Tobias da; SILVA CONCEIÇÃO, Rodrigo da; OLIVEIRA AMANTE, Fernanda de. As enchentes urbanas e o crescimento da cidade do Rio de Janeiro: estudos em direção a uma cartografia das enchentes urbanas. **Geo UERJ**, n. 32, p. 25685, 2018.

CUNHA, Anderson Mayrink da. Tese de Doutorado Movimentos de Cabeça guiados pela Voz, 2009.

DAS, Monidipa; GHOSH, Soumya K. FB-STEP: a fuzzy Bayesian network based data-driven framework for spatio-temporal prediction of climatological time series data. **Expert Systems with Applications**, Elsevier, v. 117, p. 211–227, 2019.

DAS, Prabal; CHANDA, Kironmala. Bayesian Network based modeling of regional rainfall from multiple local meteorological drivers. **Journal of Hydrology**, Elsevier, v. 591, p. 125563, 2020.

DE CAMPOS, Luis M; FRIEDMAN, Nir. A scoring function for learning Bayesian networks based on mutual information and conditional independence tests. **Journal of Machine Learning Research**, v. 7, n. 10, 2006.

DERECZYNSKI, Claudine; SILVA, Wanderson Luiz; MARENGO, Jose. Detection and projections of climate change in Rio de Janeiro, Brazil. Scientific Research Publishing, 2013.

EFRON, Bradley. Bootstrap methods: another look at the jackknife. In: BREAKTHROUGHS in statistics: Methodology and distribution. [S. l.]: Springer, 1992. P. 569–593.

FRIEDMAN, Nir; GOLDSZMIDT, Moises; WYNER, Abraham. Data analysis with Bayesian networks: A bootstrap approach. **arXiv preprint arXiv:1301.6695**, 2013.

FRIEDMAN, Nir; NACHMAN, Iftach; PE'ER, Dana. Learning Bayesian network structure from massive datasets: The"sparse candidate"algorithm. **arXiv preprint arXiv:1301.6696**, 2013.

GARROTE, Luis; MOLINA, Martin; MEDIERO, Luis. Probabilistic forecasts using Bayesian networks calibrated with deterministic rainfall-runoff models. **Extreme hydrological events: new concepts for security**, Springer, p. 173–183, 2007.

GIMENO, Luis et al. Oceanic and terrestrial sources of continental precipitation. **Reviews of Geophysics**, Wiley Online Library, v. 50, n. 4, 2012.

GLOVER, Fred. Tabu search: A tutorial. **Interfaces**, INFORMS, v. 20, n. 4, p. 74–94, 1990.

HASTIE, Trevor; TIBSHIRANI, Robert; FRIEDMAN, Jerome. An introduction to statistical learning, 2009.

HE, Haibo; GARCIA, Edwardo A. Learning from Imbalanced Data. **IEEE Transactions on Knowledge and Data Engineering**, v. 21, n. 9, p. 1263–1284, 2009. DOI: 10.1109/TKDE.2008.239.

HECKERMAN, David. A Bayesian approach to learning causal networks. **arXiv preprint arXiv:1302.4958**, 2013.

HECKERMAN, David; GEIGER, Dan; CHICKERING, David M. Learning Bayesian networks: The combination of knowledge and statistical data. **Machine learning**, Springer, v. 20, p. 197–243, 1995.

HERRING, Stephanie C et al. Explaining extreme events of 2019 from a climate perspective. **Bulletin of the American Meteorological Society**, v. 102, n. 1, s1–s115, 2021. Disponível em: <https://doi.org/10.1175/BAMS-ExplainingExtremeEvents2019.1>.

HUANG, Yingxiang et al. A tutorial on calibration measurements and calibration models for clinical prediction models. **Journal of the American Medical Informatics Association**, Oxford University Press, v. 27, n. 4, p. 621–633, 2020.

IZBICKI, Rafael; SANTOS, Tiago Mendonça dos. **Aprendizado de máquina: uma abordagem estatística**. [S. l.]: Rafael Izbicki, 2020.

JOLLIFFE, Ian. Principal Component Analysis. In: **International Encyclopedia of Statistical Science**. Edição: Miodrag Lovric. Berlin, Heidelberg: Springer Berlin Heidelberg, 2011. P. 1094–1096. ISBN 978-3-642-04898-2. DOI: 10.1007/978-3-642-04898-2_455.

KAIKKONEN, Laura et al. Bayesian networks in environmental risk assessment: A review. **Integrated environmental assessment and management**, Wiley Online Library, v. 17, n. 1, p. 62–78, 2021.

KASHINATH, K. et al. Physics-informed machine learning: case studies for weather and climate modelling. **Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences**, v. 379, n. 2194, p. 20200093, fev. 2021. ISSN 1364-503X. DOI: 10.1098/rsta.2020.0093. eprint: https://royalsocietypublishing.org/rsta/article-pdf/doi/10.1098/rsta.2020.0093/249110/rsta.2020.0093.pdf. Disponível em: <https://doi.org/10.1098/rsta.2020.0093>.

KITSON, Neville Kenneth et al. A survey of Bayesian Network structure learning. **Artificial Intelligence Review**, Springer, v. 56, n. 8, p. 8721–8814, 2023.

KODRA, Evan et al. Physics-guided probabilistic modeling of extreme precipitation under climate change. **Scientific reports**, Nature Publishing Group UK London, v. 10, n. 1, p. 10299, 2020. DOI: 10.1038/s41598-020-67088-1.

KOHAVI, Ron et al. A study of cross-validation and bootstrap for accuracy estimation and model selection. In: MONTREAL, CANADA, 2. IJCAI. [S. l.: s. n.], 1995. v. 14, p. 1137–1145.

KOLLER, Daphne; FRIEDMAN, Nir. **Probabilistic graphical models: principles and techniques**. [S. l.]: MIT press, 2009.

LEE, Hoesung et al. Climate change 2023 synthesis report summary for policymakers. **CLIMATE CHANGE 2023 synthesis report: summary for policymakers**, Intergovernmental Panel on Climate Change, 2024.

LIU, Xu-Ying; WU, Jianxin; ZHOU, Zhi-Hua. Exploratory undersampling for class-imbalance learning. **IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)**, IEEE, v. 39, n. 2, p. 539–550, 2008.

LIU, Zhifa; MALONE, Brandon; YUAN, Changhe. Empirical evaluation of scoring functions for Bayesian network model selection. **BMC bioinformatics**, Springer, v. 13, Suppl 15, s14, 2012.

LOPES, Marcel Rodrigues. **Estimação de parâmetros de populações de plantas daninhas usando inferência Bayesiana**. 2007. Tese (Doutorado) – Dissertaçao de Mestrado, EESC-USP, Sao Carlos.

LYU, Pinjie. Integrating Traditional Statistical Methods and Machine Learning for Enhanced Precipitation Forecasting. **Science and Technology of Engineering, Chemistry and Environmental Protection**, v. 1, n. 3, 2025.

MACHADO, Itallo Guilherme et al. Aprendizagem estrutural de redes bayesianas utilizando algoritmo genético multi-agente. Universidade Federal de Minas Gerais, 2021.

MACQUEEN, J. Some methods for classification and analysis of multivariate observations. In: PROCEEDINGS of 5-th Berkeley Symposium on Mathematical Statistics and Probability/University of California Press. [S. l.: s. n.], 1967.

MADIGAN, David; RAFTERY, Adrian E. Model selection and accounting for model uncertainty in graphical models using Occam's window. **Journal of the American Statistical Association**, Taylor & Francis, v. 89, n. 428, p. 1535–1546, 1994.

MARCHETE FILHO, João Rubens. **Utilização de árvores PQR para redução de cruzamentos em grafos acíclicos direcionados**. 2013. Tese (Doutorado) – [sn].

MARCOT, Bruce G. Metrics for evaluating performance and uncertainty of Bayesian network models. **Ecological Modelling**, v. 230, p. 50–62, 2012. ISSN 0304-3800. DOI: https://doi.org/10.1016/j.ecolmodel.2012.01.013. Disponível em: <https://www.sciencedirect.com/science/article/pii/S0304380012000245>.

MARGARITIS, Dimitris. **Learning Bayesian network model structure from data**. [S. l.], 2003.

MARGARITIS, Dimitris; THRUN, Sebastian. Bayesian network induction via local neighborhoods. **Advances in neural information processing systems**, v. 12, 1999.

MARQUES, Roberto Ligeiro; DUTRA, Inês. Redes Bayesianas: o que são, para que servem, algoritmos e exemplos de aplicações. **Coppe Sistemas–Universidade Federal do Rio de Janeiro, Rio de Janeiro, Brasil**, 2002.

MINING, What Is Data. Data mining: Concepts and techniques. **Morgan Kaufinann**, v. 10, n. 559-569, p. 4, 2006.

MONTGOMERY, Douglas C; RUNGER, George C. **Applied statistics and probability for engineers**. [S. l.]: John wiley & sons, 2010.

MORETTIN, Pedro A; BUSSAB, Wilton O. **Estatística básica**. [S. l.]: Saraiva Educação SA, 2017.

MURPHY, Kevin P. **Machine learning: a probabilistic perspective**. [S. l.]: MIT press, 2012.

NANDAR, Aye. Bayesian network probability model for weather prediction. In: IEEE. 2009 International Conference on the Current Trends in Information Technology (CTIT). [S. l.: s. n.], 2009. P. 1–5.

NATALIA, Yessika Adelwin et al. A Bayesian network analysis of aviation terrorism attack risks. **Journal of Transportation Security**, Springer, v. 18, n. 1, p. 1–17, 2025.

NICULESCU-MIZIL, Alexandru; CARUANA, Rich. Predicting good probabilities with supervised learning. In: PROCEEDINGS of the 22nd international conference on Machine learning. [S. l.: s. n.], 2005. P. 625–632.

OLIVEIRA BARTH, Vitor Bruno de. Um método Bayesiano orientado a dados para o aprendizado estrutural de Redes Bayesianas, 2023.

OZELAME, Camila Sgarioni et al. Estimação de estrutura em Redes Bayesianas via método scoring and restrict. **Sigmae**, v. 10, n. 1, p. 12–33, 2021.

PARLANGE, Marc B; KATZ, Richard W. An extended version of the Richardson model for simulating daily weather variables. **Journal of Applied Meteorology**, v. 39, n. 5, p. 610–622, 2000.

PEARL, Judea et al. Models, reasoning and inference. **Cambridge, UK: CambridgeUniversityPress**, v. 19, n. 2, p. 3, 2000.

PEARL, Judea. **Probabilistic reasoning in intelligent systems: networks of plausible inference**. [S. l.]: Elsevier, 2014.

PEARL, Judea; GLYMOUR, Madelyn; JEWELL, Nicholas P. **Causal inference in statistics: A primer**. [S. l.]: John Wiley & Sons, 2016.

PEARSON, Karl. X. On the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling. **The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science**, Taylor & Francis, v. 50, n. 302, p. 157–175, 1900.

PUTRI, Salwa Rizqina; WIJAYANTO, Arie Wahyu. Learning Bayesian network for rainfall prediction modeling in urban area using remote sensing satellite data (case study: Jakarta, Indonesia). In: 1. PROCEEDINGS of The International Conference on Data Science and Official Statistics. [S. l.: s. n.], 2021. v. 2021, p. 77–90.

SCHWARZ, Gideon. Estimating the dimension of a model. **The annals of statistics**, JSTOR, p. 461–464, 1978.

SCUTARI, Marco. Learning Bayesian networks with the bnlearn R package. **Journal of statistical software**, v. 35, p. 1–22, 2010.

SCUTARI, Marco; GRAAFLAND, Catharina Elisabeth; GUTIÉRREZ, José Manuel. Who learns better Bayesian network structures: Accuracy and speed of structure learning algorithms. **International Journal of Approximate Reasoning**, Elsevier, v. 115, p. 235–253, 2019. Disponível em: <https://doi.org/10.48550/arXiv.1805.11908>.

SCUTARI, Marco; SCUTARI, Maintainer Marco; MMPC, Hiton-PC. Package 'bnlearn'. **Bayesian network structure learning, parameter learning and inference, R package version**, v. 4, n. 1, 2019.

SELUCHI, Marcelo Enrique; BEU, Cássia Maria Leme; ANDRADE, Kelen. Características das frentes frias com potencial para provocar chuvas intensas na região serrana de Rio de Janeiro. **Revista Brasileira de Climatologia**, v. 18, 2016. Disponível em: <https://ojs.ufgd.edu.br/rbclima/article/view/13890>.

SHARMA, Ashutosh; GOYAL, Manish Kumar. Bayesian network for monthly rainfall forecast: a comparison of K2 and MCMC algorithm. **International Journal of Computers and Applications**, Taylor & Francis, v. 38, n. 4, p. 199–206, 2016.

SILVA, Fabricio Polifke da; SILVA, Alfredo Silveira da;
SILVA, Maria Gertrudes Alvarez Justi da. Extreme rainfall events in the Rio de Janeiro city (Brazil): description and a numerical sensitivity case study. **Meteorology and Atmospheric Physics**, Springer Nature, v. 134, p. 1–20, 77 2022. DOI: 10.1007/s00703-022-00909-2. Disponível em: <https://doi.org/10.1007/s00703-022-00909-2>.

SPIEGELHALTER, David J; DAWID, A Philip et al. Bayesian analysis in expert systems. **Statistical science**, JSTOR, p. 219–247, 1993.

SPIEGELHALTER, David J; LAURITZEN, Steffen L. Sequential updating of conditional probabilities on directed graphical structures. **Networks**, Wiley Online Library, v. 20, n. 5, p. 579–605, 1990.

SPIRTES, Peter; GLYMOUR, Clark N; SCHEINES, Richard. **Causation, prediction, and search**. [S. l.]: MIT press, 2000.

TANG, Chaoli et al. Spatiotemporal characteristics and influencing factors of sunshine duration in China from 1970 to 2019. **Atmosphere**, MDPI, v. 13, n. 12, p. 2015, 2022.

TSAMARDINOS, Ioannis; ALIFERIS, Constantin F; STATNIKOV, Alexander. Time and sample efficient discovery of Markov blankets and direct causal relations. In: PROCEEDINGS of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining. [S. l.: s. n.], 2003. P. 673–678.

TSAMARDINOS, Ioannis; ALIFERIS, Constantin F; STATNIKOV, Alexander R; STATNIKOV, Er. Algorithms for large scale Markov blanket discovery. In: FLAIRS. [S. l.: s. n.], 2003. v. 2, p. 376–81.

TSAMARDINOS, Ioannis; BROWN, Laura E; ALIFERIS, Constantin F. The max-min hill-climbing Bayesian network structure learning algorithm. **Machine learning**, Springer, v. 65, n. 1, p. 31–78, 2006.

VAN CALSTER, Ben et al. Calibration: the Achilles heel of predictive analytics. **BMC medicine**, Springer, v. 17, n. 1, p. 230, 2019.

WALLACE, John M; HOBBS, Peter V. **Atmospheric science: an introductory survey**. [S. l.]: Elsevier, 2006. v. 92.

YU, Yifan et al. A bayesian network model for neurocognitive disorders digital screening in Chinese population: development and validation study. **BMC psychiatry**, Springer, v. 25, n. 1, p. 760, 2025.

APÊNDICE A – LISTA DE RESTRIÇÕES

A tabela a seguir apresenta todas as restrições estruturais utilizadas durante a modelagem das RB. Cada linha representa uma aresta proibida, ou seja, uma dependência vetada entre duas variáveis no processo de aprendizado estrutural.

Tabela 26: Lista completa de restrições estruturais utilizadas na modelagem das Redes Bayesianas.

De (from)	**Para (to)**
chuva	umid_valor2
chuva	umid_valor3
chuva	umid_valor1
chuva	temp_valor2
chuva	temp_valor1
chuva	temp_valor3
chuva	pressao_valor1
chuva	pressao_valor3
chuva	pressao_valor2
chuva	orvalho_valor1
chuva	orvalho_valor2
chuva	orvalho_valor3
chuva	estacao_ano
chuva	turno
radiacao_atraso1h	radiacao_atraso2h
radiacao_atraso1h	radiacao_atraso3h
radiacao_atraso2h	radiacao_atraso3h
umidade_atraso1h	umidade_atraso3h

Continua na próxima página

Tabela 26 – Continuação

De (from)	**Para (to)**
umidade_atraso1h	umidade_atraso3h
umidade_atraso2h	umidade_atraso3h
temperatura_atraso1h	temperatura_atraso2h
temperatura_atraso1h	temperatura_atraso3h
temperatura_atraso2h	temperatura_atraso3h
pressao_atraso1h	pressao_atraso2h
pressao_atraso1h	pressao_atraso3h
pressao_atraso2h	pressao_atraso3h
temp_orvalho_atraso1h	temp_orvalho_atraso2h
temp_orvalho_atraso1h	temp_orvalho_atraso3h
temp_orvalho_atraso2h	temp_orvalho_atraso3h
vento_atraso1h	vento_atraso2h
vento_atraso1h	vento_atraso3h
vento_atraso2h	vento_atraso3h
estacao_do_ano	turno
turno	estacao_do_ano
radiacao_atraso1h	turno
radiacao_atraso2h	turno
radiacao_atraso3h	turno
umidade_atraso1h	turno
umidade_atraso3h	turno
umidade_atraso3h	turno
temperatura_atraso1h	turno
temperatura_atraso2h	turno
temperatura_atraso3h	turno
pressao_atraso1h	turno
pressao_atraso2h	turno
pressao_atraso3h	turno
temp_orvalho_atraso1h	turno
temp_orvalho_atraso2h	turno
temp_orvalho_atraso3h	turno
vento_atraso1h	turno

Continua na próxima página

Tabela 26 – Continuação

De (from)	**Para (to)**
vento_atraso1h	turno
vento_atraso1h	turno
radiacao_atraso1h	estacao_do_ano
radiacao_atraso2h	estacao_do_ano
radiacao_atraso3h	estacao_do_ano
umidade_atraso1h	estacao_do_ano
umidade_atraso3h	estacao_do_ano
umidade_atraso3h	estacao_do_ano
temperatura_atraso1h	estacao_do_ano
temperatura_atraso2h	estacao_do_ano
temperatura_atraso3h	estacao_do_ano
pressao_atraso1h	estacao_do_ano
pressao_atraso2h	estacao_do_ano
pressao_atraso3h	estacao_do_ano
temp_orvalho_atraso1h	estacao_do_ano
temp_orvalho_atraso2h	estacao_do_ano
temp_orvalho_atraso3h	estacao_do_ano
vento_atraso1h	estacao_do_ano
vento_atraso1h	estacao_do_ano
vento_atraso1h	estacao_do_ano
radiacao_atraso1h	umidade_atraso3h
radiacao_atraso1h	umidade_atraso3h
radiacao_atraso1h	temperatura_atraso2h
radiacao_atraso1h	temperatura_atraso3h
radiacao_atraso1h	pressao_atraso2h
radiacao_atraso1h	pressao_atraso3h
radiacao_atraso1h	temp_orvalho_atraso2h
radiacao_atraso1h	temp_orvalho_atraso3h
radiacao_atraso1h	vento_atraso2h
radiacao_atraso1h	vento_atraso3h
radiacao_atraso2h	umidade_atraso3h
radiacao_atraso2h	temperatura_atraso3h

Continua na próxima página

Tabela 26 – Continuação

De (from)	**Para (to)**
radiacao_atraso2h	pressao_atraso3h
radiacao_atraso2h	temp_orvalho_atraso3h
radiacao_atraso2h	vento_atraso3h
umidade_atraso1h	radiacao_atraso2h
umidade_atraso1h	radiacao_atraso3h
umidade_atraso1h	temperatura_atraso2h
umidade_atraso1h	temperatura_atraso3h
umidade_atraso1h	pressao_atraso2h
umidade_atraso1h	pressao_atraso3h
umidade_atraso1h	temp_orvalho_atraso2h
umidade_atraso1h	temp_orvalho_atraso3h
umidade_atraso1h	vento_atraso2h
umidade_atraso1h	vento_atraso3h
umidade_atraso3h	radiacao_atraso3h
umidade_atraso3h	temperatura_atraso3h
umidade_atraso3h	pressao_atraso3h
umidade_atraso3h	temp_orvalho_atraso3h
umidade_atraso3h	vento_atraso3h
temperatura_atraso1h	radiacao_atraso2h
temperatura_atraso1h	radiacao_atraso3h
temperatura_atraso1h	umidade_atraso3h
temperatura_atraso1h	umidade_atraso3h
temperatura_atraso1h	pressao_atraso2h
temperatura_atraso1h	pressao_atraso3h
temperatura_atraso1h	temp_orvalho_atraso2h
temperatura_atraso1h	temp_orvalho_atraso3h
temperatura_atraso1h	vento_atraso2h
temperatura_atraso1h	vento_atraso3h
temperatura_atraso2h	radiacao_atraso3h
temperatura_atraso2h	umidade_atraso3h
temperatura_atraso2h	pressao_atraso3h
temperatura_atraso2h	temp_orvalho_atraso3h

Continua na próxima página

Tabela 26 – Continuação

De (from)	**Para (to)**
temperatura_atraso2h	vento_atraso3h
pressao_atraso1h	radiacao_atraso2h
pressao_atraso1h	radiacao_atraso3h
pressao_atraso1h	umidade_atraso3h
pressao_atraso1h	umidade_atraso3h
pressao_atraso1h	temperatura_atraso2h
pressao_atraso1h	temperatura_atraso3h
pressao_atraso1h	temp_orvalho_atraso2h
pressao_atraso1h	temp_orvalho_atraso3h
pressao_atraso1h	vento_atraso2h
pressao_atraso1h	vento_atraso3h
pressao_atraso2h	radiacao_atraso3h
pressao_atraso2h	umidade_atraso3h
pressao_atraso2h	temperatura_atraso3h
pressao_atraso2h	temp_orvalho_atraso3h
pressao_atraso2h	vento_atraso3h
temp_orvalho_atraso1h	radiacao_atraso2h
temp_orvalho_atraso1h	radiacao_atraso3h
temp_orvalho_atraso1h	umidade_atraso3h
temp_orvalho_atraso1h	umidade_atraso3h
temp_orvalho_atraso1h	temperatura_atraso2h
temp_orvalho_atraso1h	temperatura_atraso3h
temp_orvalho_atraso1h	pressao_atraso2h
temp_orvalho_atraso1h	pressao_atraso3h
temp_orvalho_atraso1h	vento_atraso2h
temp_orvalho_atraso1h	vento_atraso3h
temp_orvalho_atraso2h	radiacao_atraso3h
temp_orvalho_atraso2h	umidade_atraso3h
temp_orvalho_atraso2h	temperatura_atraso3h
temp_orvalho_atraso2h	pressao_atraso3h
temp_orvalho_atraso2h	vento_atraso3h
vento_atraso1h	radiacao_atraso2h

Continua na próxima página

Tabela 26 – Continuação

De (from)	**Para (to)**
vento_atraso1h	radiacao_atraso3h
vento_atraso1h	umidade_atraso3h
vento_atraso1h	umidade_atraso3h
vento_atraso1h	temperatura_atraso2h
vento_atraso1h	temperatura_atraso3h
vento_atraso1h	pressao_atraso2h
vento_atraso1h	pressao_atraso3h
vento_atraso1h	temp_orvalho_atraso2h
vento_atraso1h	temp_orvalho_atraso3h
vento_atraso2h	radiacao_atraso3h
vento_atraso2h	umidade_atraso3h
vento_atraso2h	temperatura_atraso3h
vento_atraso2h	temp_orvalho_atraso

APÊNDICE B – ESTRUTURA APRENDIDA PELO ALGORITMO PC

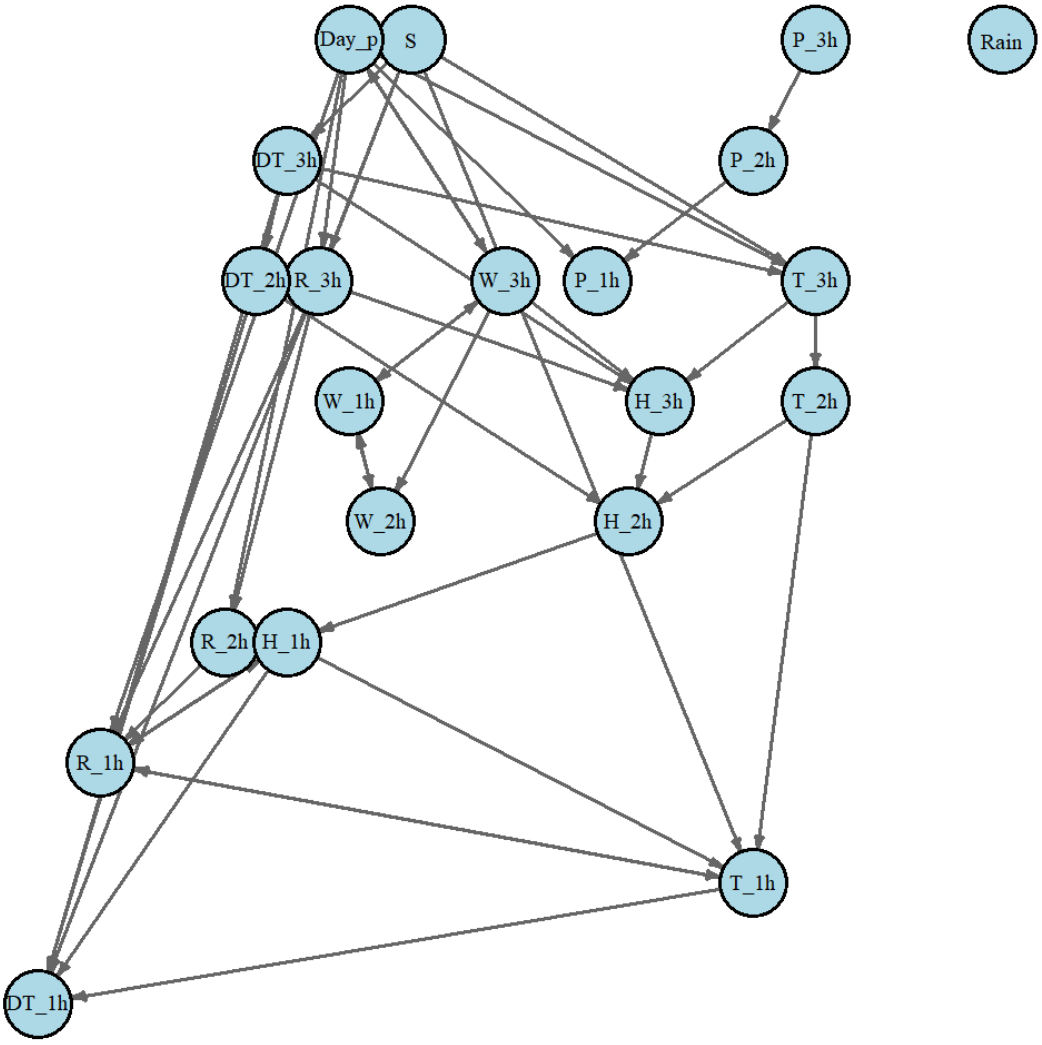

Figura 26: Representação da Rede Bayesiana aprendida pelo algoritmo PC.

APÊNDICE C – ESTRUTURA APRENDIDA PELO ALGORITMO GS

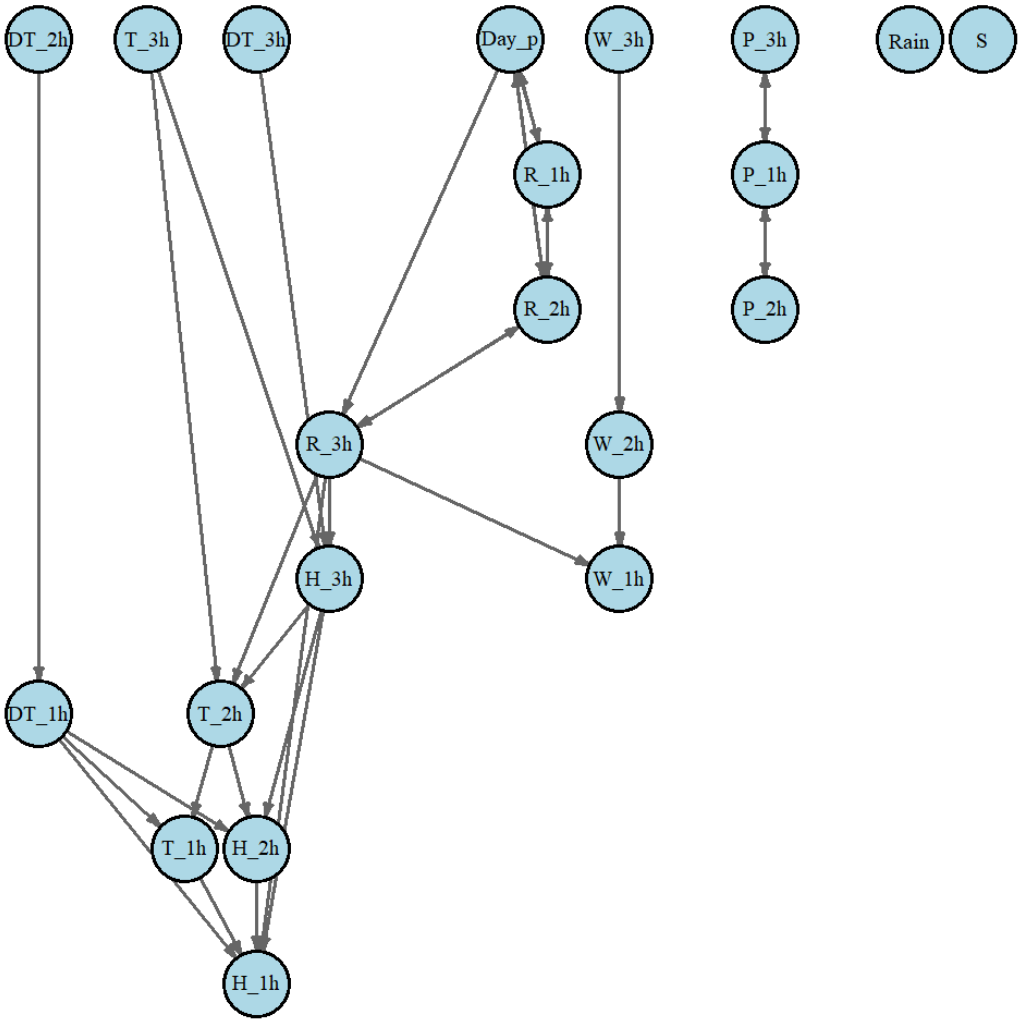

Figura 27: Representação da Rede Bayesiana aprendida pelo algoritmo GS.

APÊNDICE D – ESTRUTURA APRENDIDA PELO ALGORITMO IAMB

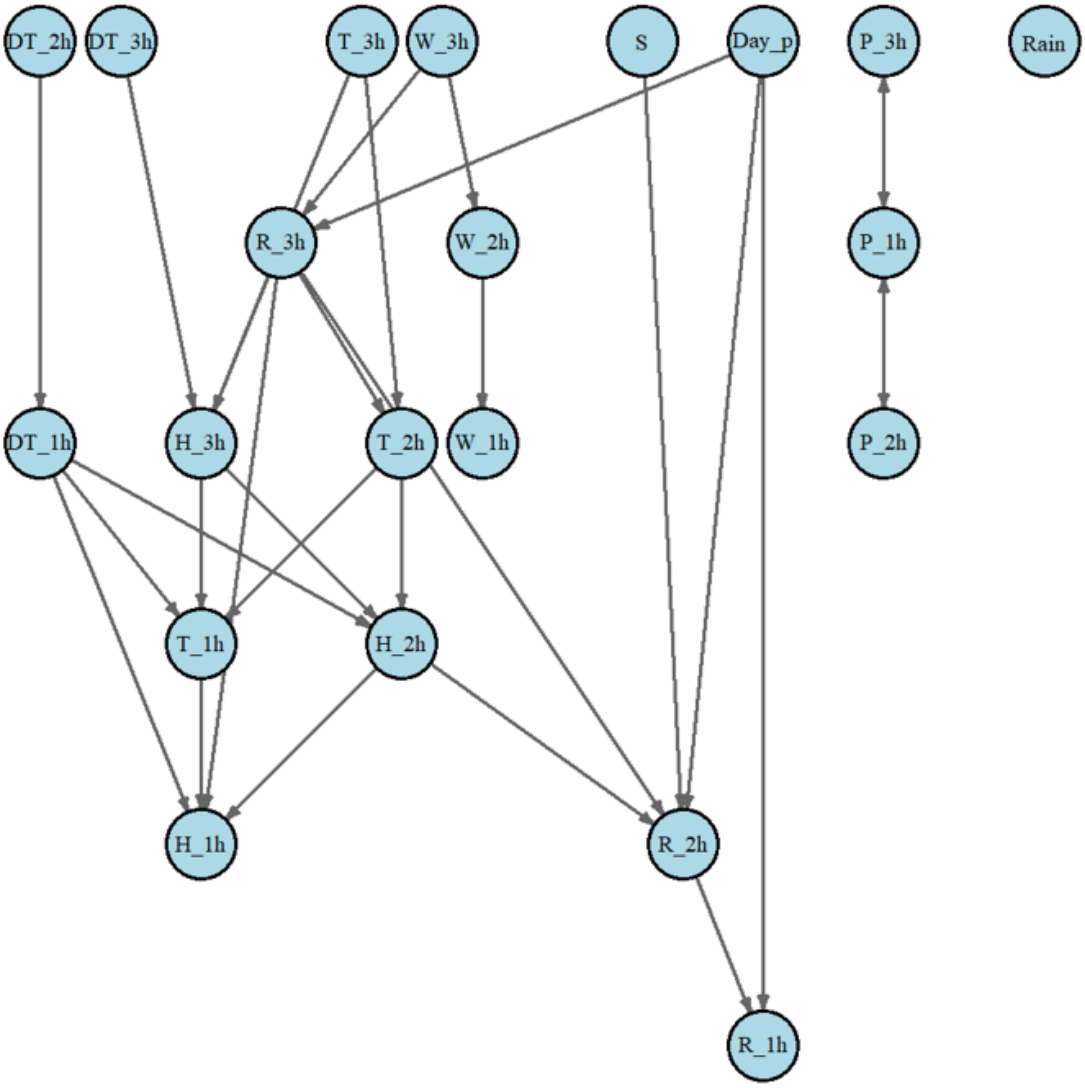

Figura 28: Representação da Rede Bayesiana aprendida pelo algoritmo IAMB.

APÊNDICE E – ESTRUTURA APRENDIDA PELO ALGORITMO MMPC

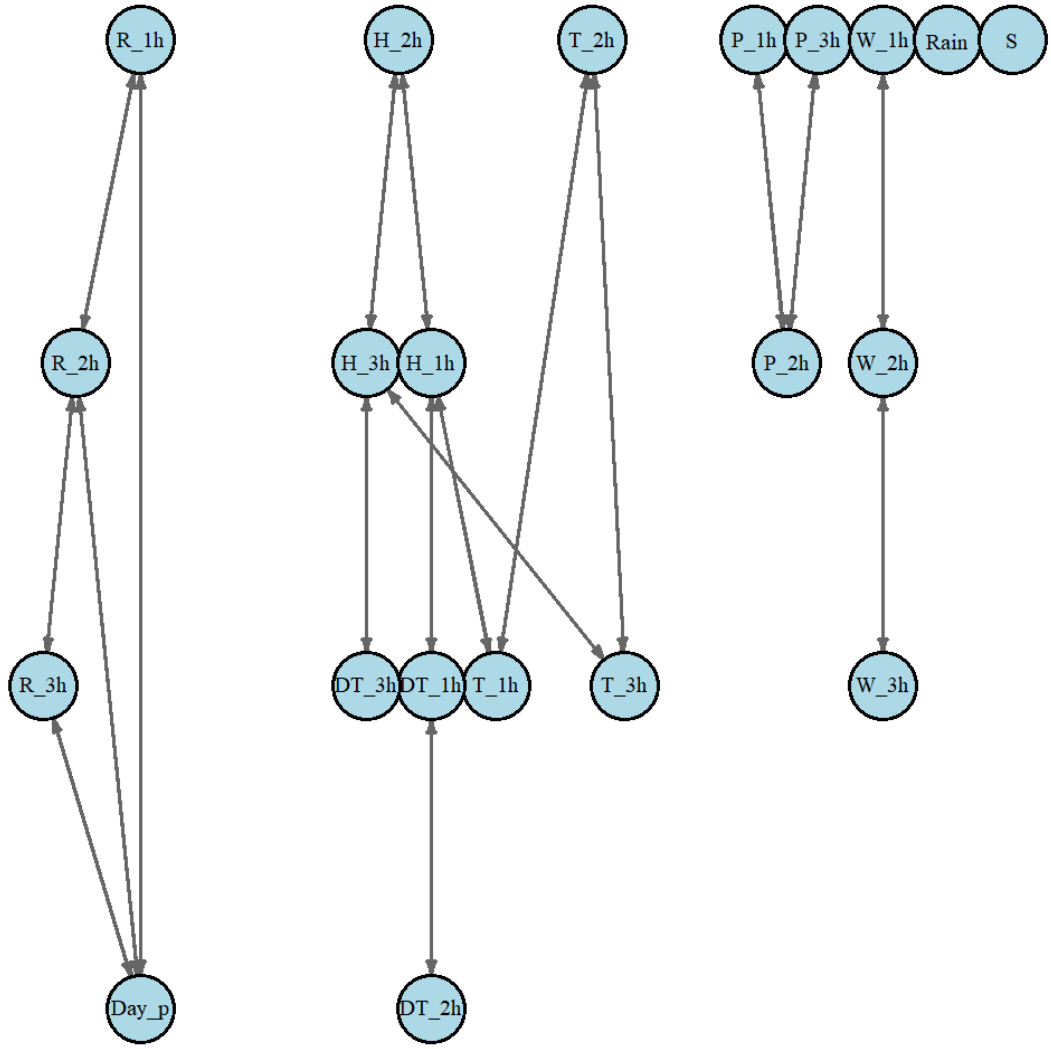

Figura 29: Representação da Rede Bayesiana aprendida pelo algoritmo MMPC.

APÊNDICE F – ESTRUTURA APRENDIDA PELO ALGORITMO MMHC

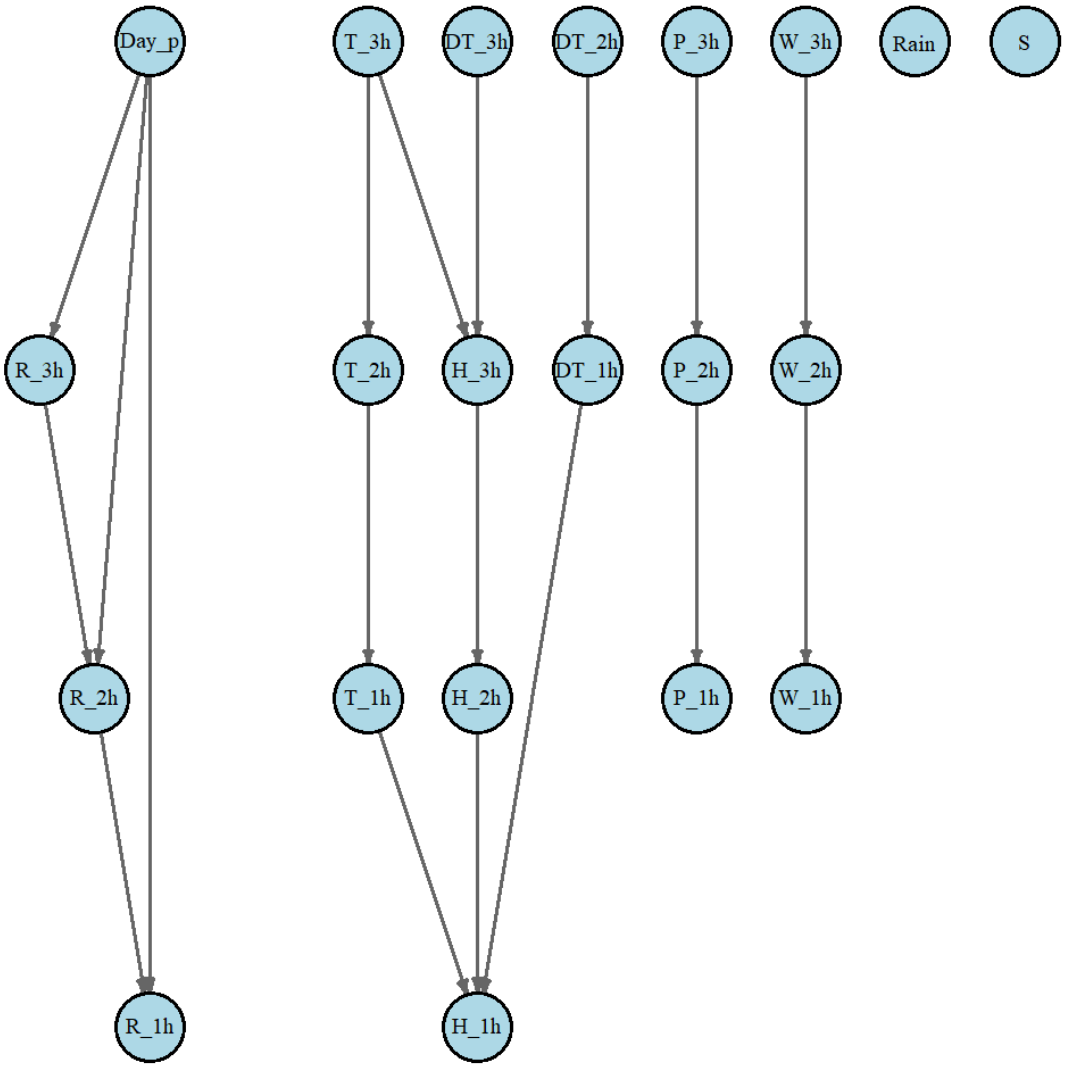

Figura 30: Representação da Rede Bayesiana aprendida pelo algoritmo MMHC.

APÊNDICE G – ESTRUTURA APRENDIDA PELO ALGORITMO RSMAX2

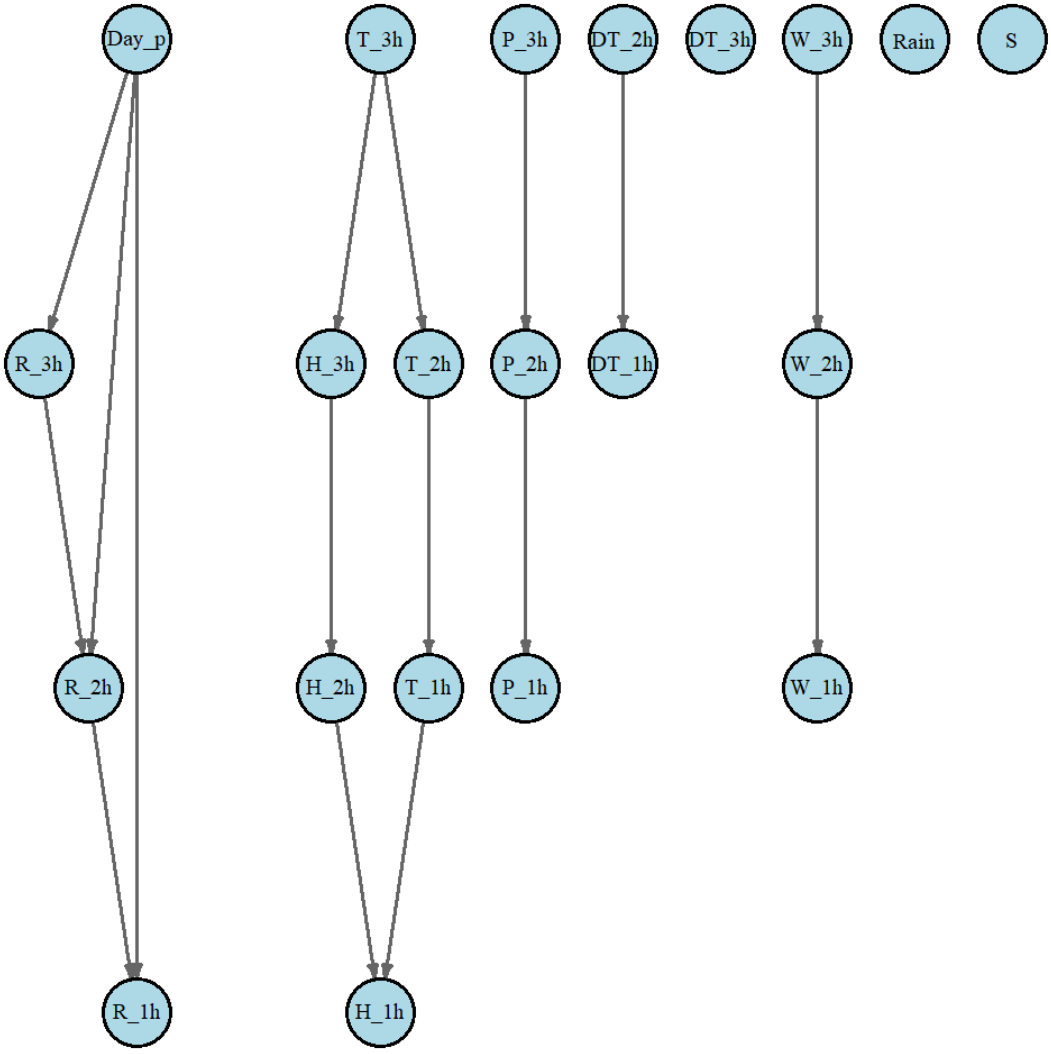

Figura 31: Representação da Rede Bayesiana aprendida pelo algoritmo RSMAX2.

APÊNDICE H – ESTRUTURA APRENDIDA PELO ALGORITMO HC AIC

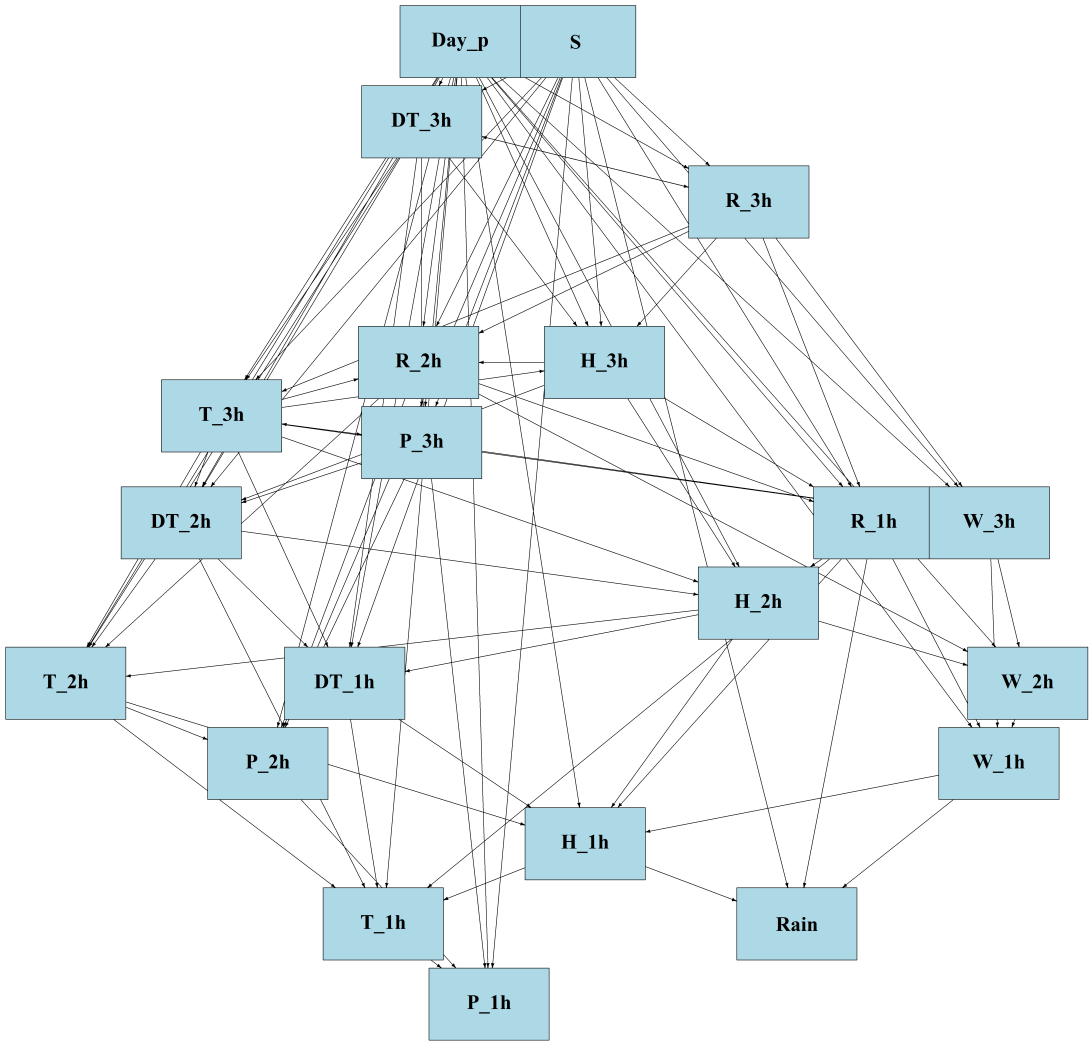

Figura 32: Representação da Rede Bayesiana aprendida pelo algoritmo HC utilizando como *score* a função AIC.

APÊNDICE I – ESTRUTURA APRENDIDA PELO ALGORITMO HC K2

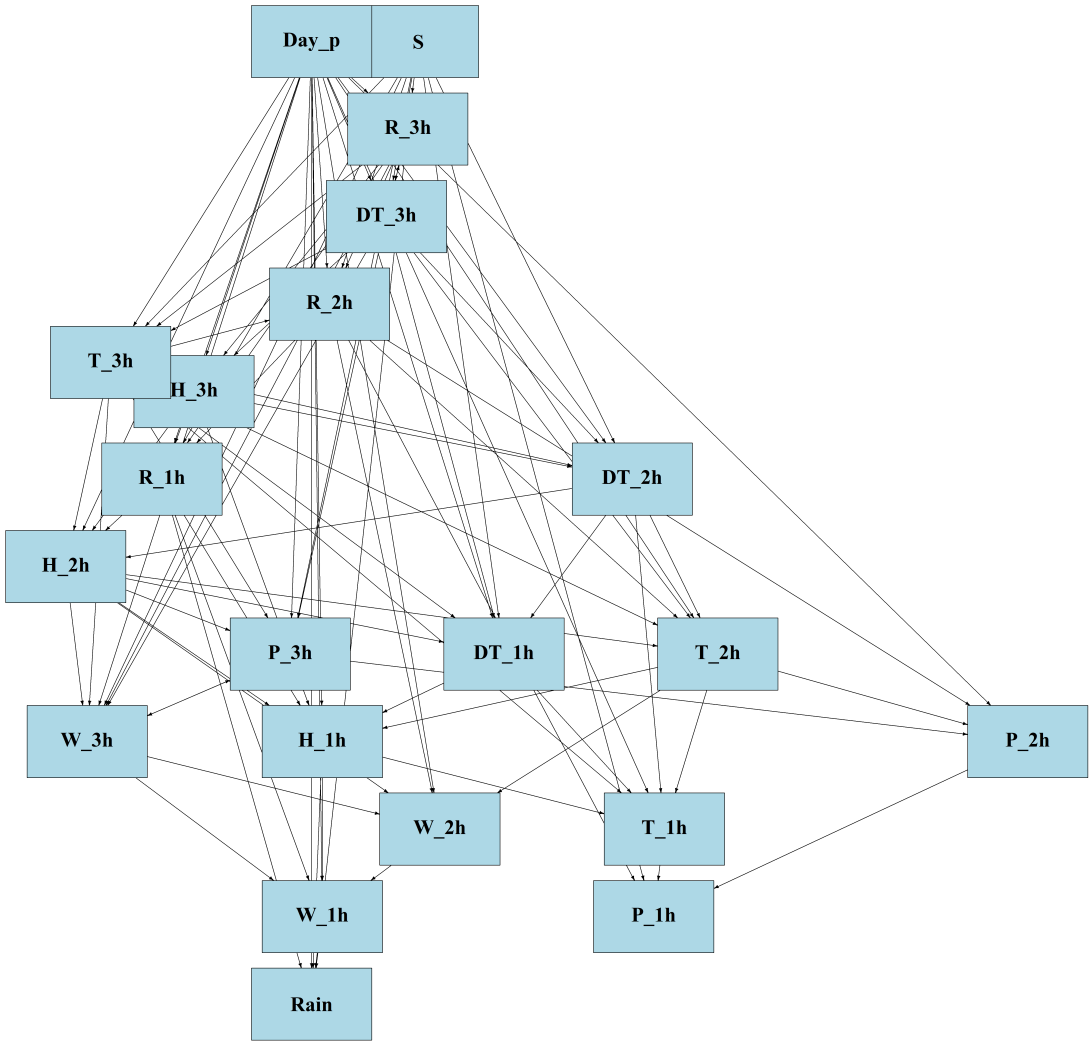

Figura 33: Representação da RB aprendida pelo algoritmo HC utilizando como *score* a função K2.

APÊNDICE J – ESTRUTURA APRENDIDA PELO ALGORITMO HC BDE

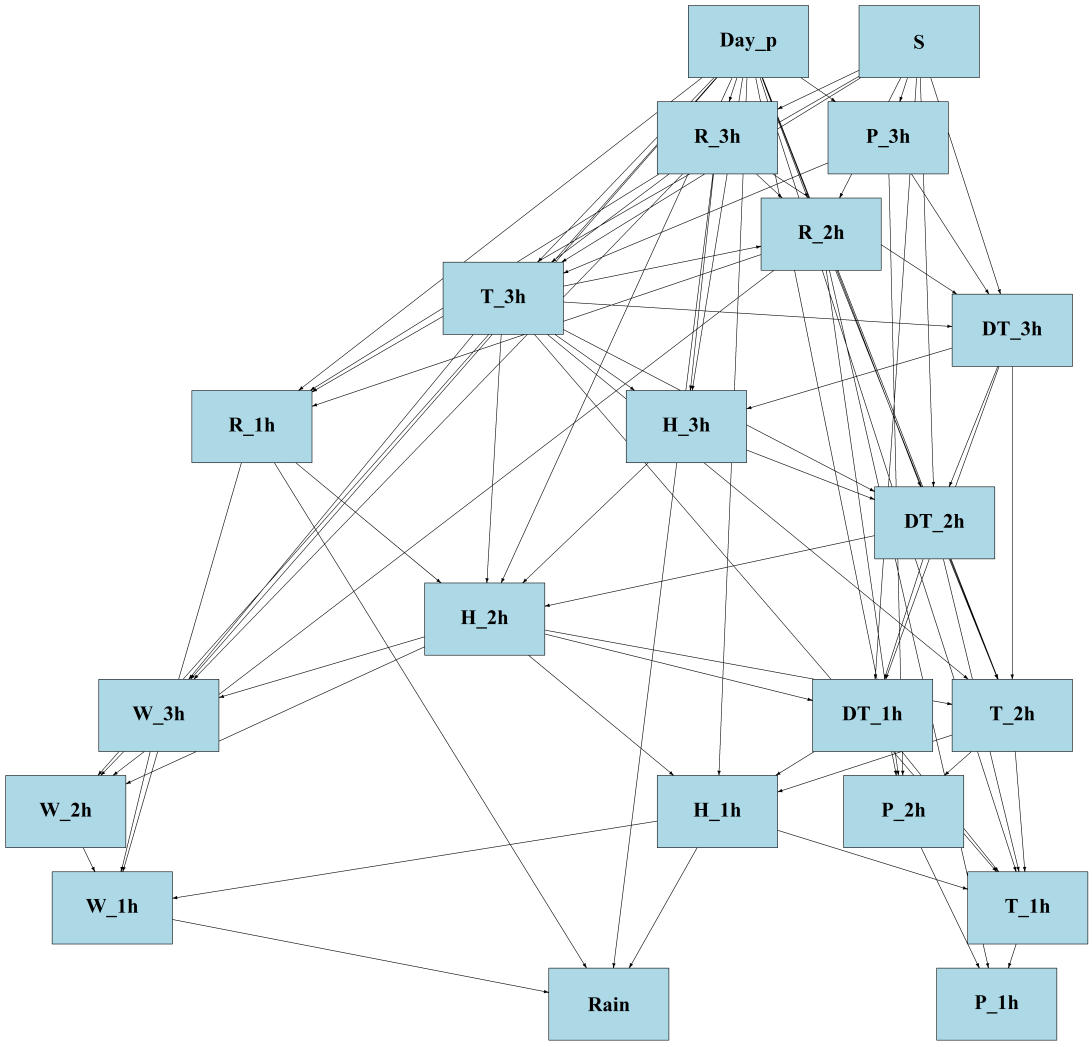

Figura 34: Representação da RB aprendida pelo algoritmo HC utilizando como *score* a função Bde.