ROGER ALMEIDA PEREIRA MELO

SELEÇÃO DE VARIÁVEIS EM MODELOS DE REGRESSÃO:
UMA AVALIAÇÃO DO USO DE REDES DE PROBABILIDADES CONDICIONAIS

LAVRAS – MG

2024

ROGER ALMEIDA PEREIRA MELO

SELEÇÃO DE VARIÁVEIS EM MODELOS DE REGRESSÃO:
UMA AVALIAÇÃO DO USO DE REDES DE PROBABILIDADES CONDICIONAIS

Tese apresentada à Universidade Federal de Lavras, como parte das exigências do Programa de Pós-Graduação em Estatística e Experimentação Agropecuária, para a obtenção do título de Doutor.

Prof. Dra. Izabela Regina Cardoso de Oliveira
Orientadora

Prof. Dr. Júlio Sílvio de Sousa Bueno Filho
Coorientador

LAVRAS – MG
2024

Ficha catalográfica elaborada pelo Sistema de Geração de Ficha Catalográfica da Biblioteca Universitária da UFLA, com dados informados pelo(a) próprio(a) autor(a).

Melo, Roger Almeida Pereira
Seleção de variáveis em modelos de regressão: : uma avaliação do uso de redes de probabilidades condicionais / Roger Almeida Pereira Melo. – 2024.
167 p. :

Orientador: Prof. Dra. Izabela Regina Cardoso de Oliveira.
Coorientador: Prof. Dr. Júlio Sílvio de Sousa Bueno Filho.
Tese (doutorado) – Universidade Federal de Lavras, 2024.
Bibliografia.

1. Redes bayesianas. 2. Grafo acíclico direcionado. 3. Modelo probabilístico. I. Oliveira, Izabela Regina Cardoso de. II. Filho, Júlio Sílvio de Sousa Bueno. III. Título.

ROGER ALMEIDA PEREIRA MELO

SELEÇÃO DE VARIÁVEIS EM MODELOS DE REGRESSÃO: UMA AVALIAÇÃO DO USO DE REDES DE PROBABILIDADES CONDICIONAIS

VARIABLE SELECTION IN REGRESSION MODELS: AN EVALUATION OF THE USE OF CONDITIONAL PROBABILITY NETWORKS

Tese apresentada à Universidade Federal de Lavras, como parte das exigências do Programa de Pós-Graduação em Estatística e Experimentação Agropecuária, para a obtenção do título de Doutor.

APROVADA em 20 de Junho de 2024.

Prof. Dr. Pedro Henrique Ramos Cerqueira	UEL
Prof. Dr. Tales Jesus Fernandes	UFLA
Prof. Dra. Elaine Maria Seles Dorneles	UFLA

Prof. Dra. Izabela Regina Cardoso de Oliveira
Orientadora

Prof. Dr. Júlio Sílvio de Sousa Bueno Filho
Co-Orientador

LAVRAS – MG
2024

AGRADECIMENTOS

À minha mãe Neusa e meu pai Osvaldo, que sempre se dedicaram com tanto amor e cuidado à minha educação e ao meu desenvolvimento pessoal, serei para sempre grato pelo carinho e pelos sacrifícios que fizeram por mim.

À minha estimada Vanessa, pelo apoio, amor, respeito, conselhos e carinho ao longo desta trajetória.

Aos meus irmãos Marcel e Rafael, pela parceria de sempre.

À minha orientadora, Izabela Regina Cardoso de Oliveira, e ao meu coorientador, Júlio Silvio de Sousa Bueno Filho, pelas orientações e pelo apoio fundamental ao longo do doutorado.

Aos professores membros da banca, de qualificação e defesa, pelas importantes contribuições para o desenvolvimento deste trabalho.

Ao Departamento de Estatística e à Universidade Federal de Lavras.

Aos meus amigos, especialmente ao Nicásio, pela valiosa amizade, e aos demais colegas do Departamento.

A todas as pessoas que, de algum modo, contribuíram para a conclusão deste doutorado.

O presente trabalho foi realizado com apoio da Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) - Código de Financiamento 001.

RESUMO

A Rede bayesiana é um método apresentado por Judea Pearl em 1985 que descreve um modelo probabilístico gráfico, representando um conjunto de variáveis e suas dependências condicionais por meio de um grafo acíclico direcionado. Os vértices (ou nós) representam proposições (ou variáveis), as arestas (ou arcos), quando são direcionadas, significam as dependências probabilísticas entre essas variáveis. O objetivo deste trabalho é avaliar o uso de redes bayesianas para seleção de variáveis em modelos de regressão. A técnica é comparada com métodos *stepwise* em alguns cenários de simulação que consideram diferentes tamanhos amostrais, correlações entre as variáveis (resposta e variáveis) e diferentes números de variáveis. Além do estudo de simulação, apresentamos uma aplicação prática das redes bayesianas nesse contexto. Para isso, foram usados dados de uma pesquisa realizada entre 2018 e 2019, com médicos veterinários de Minas Gerais, com o objetivo de identificar os fatores de risco mais importantes associados à exposição acidental às vacinas anti-*Brucella abortus* (Brucelose). Uma das respostas de interesse no trabalho é a prevalência de brucelose entre esses profissionais, que foi estimada a partir de um modelo de regressão logístico. Ao utilizar Rede bayesiana, as variáveis detectadas como mais importantes associadas à exposição acidental às vacinas foram o conhecimento sobre os sintomas da brucelose, se o profissional realizou procedimentos de partos prematuros ou abortos nos últimos seis meses e a frequência que o profissional usa equipamentos de proteção individual. Todas as análises foram realizadas no software *R* utilizando o pacote *bnlearn*. Recomendamos a combinação de métodos *stepwise* com redes bayesianas, pois os métodos *stepwise* são eficazes para a seleção automática de variáveis, enquanto as redes bayesianas são excelentes para visualizar e entender associações indiretas entre variáveis. Essa combinação enriquece a análise, oferecendo uma visão mais completa e detalhada dos resultados.

Palavras-chave: redes bayesianas; grafo acíclico direcionado; modelo probabilístico; métodos *stepwise*.

ABSTRACT

The Bayesian network is a method presented by Judea Pearl in 1985 that describes a probabilistic graphical model, which represents a set of variables and their conditional dependencies with a directed acyclic graph. The vertices (or nodes) represent propositions (or variables), and directed edges (or arcs) signify the probabilistic dependencies between these variables. The objective of this study is to evaluate the use of Bayesian networks for the selection of variables in regression models. This technique is compared with stepwise methods in simulation scenarios that consider different sample sizes, correlations between the variables (responses and variables) and different numbers of variables. In addition to the simulation study, we present a practical application of Bayesian networks in this context. For this purpose, data from a study conducted between 2018 and 2019 involving veterinarians in Minas Gerais were used to identify the most important risk factors associated with accidental exposure to antiviral vaccines, specifically, vaccines for Brucella abortus (Brucellosis). One of the results of interest in this study was the prevalence of brucellosis among these professionals, which was estimated using a logistic regression model. According to the Bayesian network, the most important covariates associated with accidental exposure to vaccines were knowledge about the symptoms of brucellosis, whether the professional had performed premature childbirth procedures or abortions in the previous six months and the frequency with which the professional used personal protective equipment. All analyses were performed in R software using the bnlearn package. We recommend combining stepwise methods with Bayesian Networks, as stepwise methods are effective for automatic variable selection, while Bayesian Networks excel at visualizing and understanding indirect associations between variables. This combined approach enriches the analysis, providing a more comprehensive and detailed view of the results.

Keywords: bayesian networks; directed acyclic graph; probabilistic model; stepwise methods.

INDICADORES DE IMPACTO

O trabalho apresentado possui impactos significativos nas áreas de saúde, tecnologia e produção, além de potenciais implicações sociais e econômicas. A utilização de redes bayesianas para a seleção de variáveis em modelos de regressão, aplicada no estudo sobre a exposição acidental a vacinas de Brucelose em veterinários, contribui diretamente para o campo da saúde, uma vez que possibilita a identificação de fatores de risco de forma mais precisa e eficiente, permitindo intervenções preventivas mais eficazes. Esse tipo de análise pode reduzir a prevalência de doenças como a Brucelose entre profissionais de saúde animal, além de proporcionar um entendimento mais profundo das associações entre variáveis indiretas, que dificilmente seriam capturadas por outros métodos tradicionais. O impacto econômico desse trabalho é observado na potencial redução de custos com tratamento de doenças, promovendo uma alocação mais eficiente de recursos na saúde pública. Além disso, o desenvolvimento tecnológico proporcionado pelo uso de ferramentas de análise avançada como o R e o pacote *bnlearn* fortalece o campo da estatística aplicada à saúde. A sociedade é beneficiada indiretamente, uma vez que a redução de casos de Brucelose em profissionais diminui os riscos de propagação da doença, gerando um efeito positivo na saúde pública e nas cadeias produtivas que dependem do controle dessa zoonose. O caráter extensionista do trabalho também se reflete na parceria com médicos veterinários de Minas Gerais, promovendo um diálogo entre a academia e a prática profissional, além de contribuir com a melhoria das condições de trabalho desses profissionais. Este estudo se alinha com os Objetivos de Desenvolvimento Sustentável (ODS) da ONU, especialmente no que se refere à saúde e bem-estar (ODS 3) e trabalho decente e crescimento econômico (ODS 8).

IMPACT INDICATORS

The presented work has significant impacts in the areas of health, technology, and production, in addition to potential social and economic implications. The use of Bayesian networks for variable selection in regression models, applied in the study on accidental exposure to Brucellosis vaccines in veterinarians, directly contributes to the health field, as it enables the identification of risk factors more precisely and efficiently, allowing for more effective preventive interventions. This type of analysis can reduce the prevalence of diseases such as Brucellosis among animal health professionals, in addition to providing a deeper understanding of associations between indirect variables, which are rarely captured by other traditional methods. The economic impact of this work is observed in the potential reduction of treatment costs, promoting a more efficient allocation of resources in public health. Additionally, the technological development provided by the use of advanced analysis tools such as R and the bnlearn package strengthens the field of statistics applied to health. Society is indirectly benefited, as the reduction of Brucellosis cases among professionals decreases the risk of disease spread, generating a positive effect on public health and on the production chains that rely on controlling this zoonosis. The extensionist nature of the work is also reflected in the partnership with veterinarians from Minas Gerais, fostering dialogue between academia and professional practice, while also contributing to the improvement of working conditions for these professionals. This study aligns with the United Nations' Sustainable Development Goals (SDGs), especially with regard to health and well-being (SDG 3), as well as decent work and economic growth (SDG 8).

LISTA DE FIGURAS

LISTA DE QUADROS

LISTA DE TABELAS

SUMÁRIO

1 INTRODUÇÃO

Nas análises de dados, em que o interesse é descrever a relação entre uma variável resposta e uma ou mais variáveis explicativas, os métodos de regressão tornaram-se fundamentais para esse tipo de investigação. As variáveis podem ser classificadas como quantitativas ou qualitativas (também chamadas de categóricas).

Variáveis quantitativas assumem valores numéricos e podem ser discretas, normalmente resultantes de uma contagem, ou contínuas, geralmente provenientes de mensurações, como idade e altura (MONTGOMERY et al., 2012). Em contraste, variáveis qualitativas assumem valores em uma das diversas categorias distintas. Por exemplo, a marca de um produto comprado (marca A, B ou C) ou se uma pessoa deixa de pagar uma dívida (sim ou não) (AGRESTI, 2012).

Modelos de regressão são ferramentas estatísticas utilizadas para entender e prever a relação entre variáveis preditoras e uma variável resposta. Quando a variável resposta é contínua, e podemos identificar uma escala em que o modelo é normal, utilizamos modelos de regressão linear. Esses modelos assumem que a relação linear entre as variáveis preditoras, a variável resposta e os resíduos seguem uma distribuição normal (KUTNER et al., 2005).

Por outro lado, quando a variável resposta é categórica, utilizamos modelos de regressão logística. A regressão logística é um tipo de Modelo Linear Generalizado (MLG) que usa a função de ligação logit para modelar a probabilidade de ocorrência de um dos dois resultados possíveis (HOSMER; LEMESHOW, 2000).

Os Modelos Lineares Generalizados (MLG) generalizam os modelos de regressão linear e logística, permitindo que a variável resposta tenha diferentes distribuições (não apenas normal ou binomial) e que a relação entre a variável resposta e as variáveis preditoras seja não necessariamente linear. Essa flexibilidade torna os MLG amplamente aplicáveis a uma variedade de tipos de dados e contextos, permitindo modelar eficientemente tanto variáveis contínuas quanto categóricas (NELDER; WEDDERBURN, 1972).

Existem outros modelos capazes de lidar com respostas binárias, como probit, log e log complementar, porém, apenas a regressão logística pode ser usada para estimar a razão de chances para os preditores do modelo. Este é um recurso que desempenha um papel muito importante em áreas como estatísticas de pesquisas médicas (HILBE, 2009).

Em um estudo de regressão logística, o objetivo de uma análise utilizando esse método é encontrar o modelo mais adequado, parcimonioso e biologicamente razoável para descrever

a relação entre uma variável de resultado (dependente ou resposta) e um conjunto de variáveis independentes (preditivas ou explicativas) (HOSMER; LEMESHOW, 2000)

Um dos grandes desafios na construção de modelos é selecionar, de um extenso conjunto de variáveis preditoras, aquelas que irão compor o modelo final. Na abordagem tradicional para construção de modelos estatísticos, busca-se um modelo parcimonioso, ou seja, um modelo que use o menor número possível de variáveis, mantendo uma boa capacidade preditiva e representando adequadamente o padrão dos dados. No entanto, essa abordagem pode resultar em modelos "superajustados", o que significa que eles seguem essencialmente os erros ou ruídos dos dados (JAMES et al., 2013). Além disso, métodos puramente estatísticos podem não considerar a importância prática (clínica, biológica) de certas variáveis, subestimando sua relevância. Portanto, é crucial utilizar metodologias que incorporem o conhecimento prévio do pesquisador, garantindo que variáveis importantes não sejam negligenciadas. Isso justifica o uso de técnicas como as redes bayesianas (RB), que permitem integrar informações a priori, resultando em modelos mais interpretáveis e com potencial aplicação em contextos práticos.

Assim, surgem diversos métodos de seleção de variáveis, sendo que os mais comuns são os métodos *stepwise*, amplamente aplicados para identificar covariáveis para inclusão e/ou exclusão em modelos de regressão. Nestes métodos, as variáveis são selecionadas para inclusão ou exclusão do modelo de forma sequencial, com base apenas em critérios estatísticos.

O processo de seleção de variáveis *Forward stepwise* começa com um modelo sem preditores (conhecido como *modelo nulo*) e adiciona os preditores mais significativos um de cada vez, passo a passo. A estratégia *Backward stepwise* começa com um modelo que inclui todos os preditores sob consideração (chamado de *modelo completo*) e, em seguida, remove os preditores menos significativos um de cada vez, passo a passo. É comum também se utilizar a combinação de passos de inclusão e exclusão de variáveis (*Bidirectional stepwise*).

Purposeful selection of covariates (PVS) é um método de seleção de variáveis no contexto de regressão logística proposto por Hosmer e Lemeshow (2000), ele consiste em estabelecer critérios de escolha e inclusão de variáveis preditoras em cada passo do processo de modelagem. O PVS pode ser visto como uma variação do *stepwise*, que em geral é realizado de forma automática. A diferença é incorporar a avaliação do pesquisador sobre a importância das variáveis no modelo sob ajuste, o que pode ser interessante em algumas situações práticas.

Pereira et al. (2020b) utilizaram o método PVS, com objetivo de determinar a prevalência de brucelose, de exposição às vacinas S19 e RB51 entre veterinários e identificar os fatores

de risco mais importantes associados à exposição acidental às vacinas anti *Brucella abortus*, entretanto, algumas relações esperadas entre variáveis não foram captadas. Com isso, surgiu o interesse de investigar outros métodos de seleção que incorporam o conhecimento prévio do pesquisador nessa tarefa.

Um método a ser considerado é o uso de redes bayesianas (RB, Pearl (1985)), em que um modelo probabilístico é representado em um gráfico que relaciona o conjunto de variáveis e suas dependências condicionais chamado grafo acíclico direcionado (*directed acyclic graph* - DAG). As RB têm sido amplamente utilizadas em vários campos de aplicação, como na classificação de documentos (DENOYER; GALLINARI, 2004; CALADO et al., 2003) e em sistema de apoio à decisão (KRISTENSEN; RASMUSSEN, 2002; YET et al., 2013; LTIFI et al., 2012).

Vannucci, Stingo e Berzuini (2012) realizaram uma revisão dos métodos bayesianos para seleção de variáveis para configurações lineares e para modelos de mistura. Concluíram que esses métodos oferecem uma estrutura coerente na qual a seleção de variáveis e a predição, classificação ou agrupamento das amostras são realizadas simultaneamente, e que podem lidar com um grande número de regressores e com um número de covariáveis maior que o tamanho da amostra. Note-se, no entanto, que as redes bayesianas são estruturas de probabilidade condicional que frequentemente são estimadas com métodos frequentistas ou fiduciais, e não com métodos bayesianos.

Hruschka, Hruschka e Ebecken (2004) descreveram e avaliaram uma estratégia de seleção de variáveis em problemas de classificação, em que uma rede bayesiana é gerada a partir de um conjunto de dados e, em seguida, a cobertura de Markov (*Markov Blanket*) da variável classificatória é avaliada como critério de seleção.

Segundo Lee, Abbott e Johantgen (2005), redes bayesianas possuem algumas vantagens em relação à regressão logística, quais sejam: as RB podem (a) ser usadas sem pressupostos estatísticos convencionais, como linearidade ou aditividade; (b) lidar com um número maior de preditores quando a identificação de interações entre preditores for menos complexa; (c) detectar variáveis e relações importantes que poderiam ser perdidas pelos investigadores; e (d) produzir previsões precisas mesmo em situações em que dados completos não estejam disponíveis. Os autores ressaltam que, além disso, RB são de fácil compreensão, pois representam o conhecimento por meio de diagrama gráfico.

Sendo assim, o objetivo deste trabalho é avaliar o uso de redes bayesianas para seleção de variáveis em modelos de regressão. Os objetivos específicos são i) realizar um estudo de

simulação para comparar o desempenho de métodos *stepwise* com as redes bayesianas e avaliar os principais algoritmos implementados para obtenção de redes, e ii) aplicar redes bayesianas em um problema prático que envolve a seleção de variáveis.

Este trabalho está organizado da seguinte forma. No capítulo 2 são abordados conceitos téoricos importantes sobre modelos lineares generalizados, modelos de regressão logística, métodos de seleção de variáveis, redes bayesianas e validação cruzada. No capítulo 3 são apresentadas as especificações técnicas do estudo de simulação e a descrição da base de dados utilizada. No capítulo 4 são apresentados os principais resultados do estudo de simulação. No capítulo 5 são apresentados os resultados da aplicação e no capítulo 6 são resumidas as principais conclusões e eventuais recomendações.

2 REFERENCIAL TEÓRICO

Serão abordados nesse capítulo o estado da arte nos principais tópicos metodológicos que serão utilizados na nossa análise de seleção de variáveis. Iniciaremos pelo conceito básico de tabelas de contingência que é importante para compreender os resumos em tabelas de probabilidade condicional das redes bayesianas.

2.1 Tabelas de contingência e o conceito de razão de chances

Sejam X e Y duas variáveis categóricas, X com I categorias e Y com J categorias. Uma tabela com I linhas para as categorias de X e J colunas para as categorias de Y exibe as IJ combinações possíveis de resultados. As células da tabela representam os IJ resultados possíveis. Quando as células contém contagens de frequência de resultados para uma amostra, a tabela é chamada de *tabela de contigência* (AGRESTI, 2012), como exemplificada na Tabela 2.1.

Tabela 2.1 – Representação de uma tabela de contingência para o estudo das variáveis X, com I categorias e Y, com J categorias.

	Categorias				
Tratamentos	C_1	C_2	$\dots$	C_J	Total
1	n_{11}	n_{12}	$\dots$	n_{1J}	n_1
2	n_{21}	n_{22}	$\dots$	n_{2J}	n_2
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
I	n_{I1}	n_{I2}	$\dots$	n_{IJ}	n_J

Fonte: Do autor (2024).

Em algumas aplicações, X e Y são variáveis respostas. Suponha que os indivíduos sejam escolhidos aleatoriamente de uma determinada população. Então, as respostas (X,Y) de um indivíduo escolhido aleatoriamente têm uma distribuição de probabilidade. Seja π_{ij} denotando a probabilidade de que (X,Y) ocorra na célula da linha i e coluna j. As distribuições marginais são os totais de linha e coluna que resultam da soma das probabilidades conjuntas (AGRESTI, 2012). Denotando $\{\pi_{i+}\}$ para a variável de linha e $\{\pi_{+j}\}$ para a variável de coluna, em que o subscrito "$+$" denota a soma sobre esse índice, então

$$\pi_{i+} = \sum_j \pi_{ij} \quad \text{e} \quad \pi_{+j} = \sum_i \pi_{ij},$$

que satisfazem $\sum_i \pi_{i+} = \sum_j \pi_{+j} = \sum_i \sum_j \pi_{ij} = 1$. As distribuições marginais fornecem informações das variáveis singularmente.

Na maioria das tabelas de contingência, Y é uma variável resposta e X uma variável explicativa. Para uma categoria fixa de X, a variável Y tem uma distribuição de probabilidade, e é pertinente estudar como essa distribuição muda à medida que a categoria X muda. Dado que um indivíduo esteja classificado na linha i de X, denota-se $\pi_{j|i}$ a probabilidade de classificação na coluna j de Y, $j = 1, 2, \ldots, J$. Então, $\sum_j \pi_{j|i} = 1$. As probabilidades $(\pi_{1|i}, \ldots, \pi_{J|i})$ formam a distribuição condicional de Y na categoria i de X (AGRESTI, 2012).

A Tabela 2.2 exibe a notação para distribuições conjuntas, condicionais e marginais para o caso 2×2.

Tabela 2.2 – Notação para probabilidades conjuntas, condicionais e marginais

	Colunas		
Linhas	1	2	Total
1	π_{11}	π_{12}	π_{1+}
	$(\pi_{1\|1})$	$(\pi_{2\|1})$	$(1,0)$
2	π_{21}	π_{22}	π_{2+}
	$(\pi_{1\|2})$	$(\pi_{2\|2})$	$(1,0)$
Total	π_{+1}	π_{+2}	$1,0$

Fonte: Do autor (2024).

Para uma probabilidade π de sucesso, a *chance* é definida como

$$\Omega = \frac{\pi}{1-\pi}$$

Considerando uma tabela de contingência 2×2, dentro da linha i a chance de sucesso ao invés de fracasso é $\Omega_i = \frac{\pi_i}{1-\pi_i}$. A razão entre Ω_1 e Ω_2 é chamada de *razão de chances*,

$$\theta = \frac{\Omega_1}{\Omega_2} = \frac{\dfrac{\pi_1}{(1-\pi_1)}}{\dfrac{\pi_2}{(1-\pi_2)}}.$$

2.2 Modelos lineares generalizados

Um modelo linear generalizado (MLG) descreve uma relação entre a média de uma variável resposta Y e uma variável independente X. Um MLG consiste em 3 componentes: aleatório, sistemático e função de ligação (CASELLA; BERGER, 2001):

a) **componente aleatório**: é definido pela distribuição de probabilidade da variável resposta Y_i, $i = 1, \ldots, n$; que assume independência entre si e pertencem à família exponencial (FE) de distribuições;

b) **componente sistemático**: é uma função das variáveis preditoras x_i, lineares nos parâmetros, que se relaciona com a média das observações y_i. De modo geral, o componente sistemático é descrito pela seguinte estrutura aditiva:

$$\eta_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip};$$

c) **função de ligação:** é uma função que relaciona ou descreve a relação entre a média da variável resposta Y_i, isto é, $\mu_i = E[Y_i]$ e o preditor linear η_i.

Portanto, temos que um MLG é definido como:

$$y_i \sim FE(\mu_i, \phi)$$

$$g(\mu_i) = \eta_i \iff \mu_i = g^{-1}(\eta_i),$$

em que y_i, $i = 1, \ldots, n$, é a variável resposta com distribuição pertecente a família exponencial de distribuições com média μ_i e parâmetro de dispersão ϕ. Além disso, assume-se que y_i é independente de y_j para todo $i \neq j$.

2.2.1 Modelo de regressão linear

A regressão linear envolve prever uma variável dependente (a variável resposta) com base em uma ou mais variáveis independentes. O modelo de regressão linear é considerado um modelo linear generalizado, onde o componente aleatório é definido por uma distribuição normal para a variável resposta e a função de ligação é a identidade (f(x) = x), relacionando a média normal da variável resposta y_i com o preditor linear η_i. Em resumo, temos:

$$Y_i \overset{ind}{\sim} N(\mu_i, \sigma^2)$$

$$E[Y_i] = \mu_i = \eta_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ir} \text{ para } i = 1, 2, \ldots, n. \tag{2.1}$$

Reescrevendo o modelo 2.1 na forma vetorial, temos

$$\begin{aligned} \boldsymbol{Y} &\overset{ind}{\sim} N(\boldsymbol{\mu}, \boldsymbol{\sigma^2}) \\ \boldsymbol{\mu} &= \boldsymbol{X\beta} \end{aligned} \tag{2.2}$$

em que $\boldsymbol{Y} = (Y_1, \ldots, Y_n)^\top$ é o vetor de respostas, $\boldsymbol{X}_{n\times p}$ é matriz de delineamento com $(p = r+1)$, contendo as r colunas de covariáveis mais uma coluna de uns, $\boldsymbol{\beta} = (\beta_0, \ldots, \beta_r)^\top$ é o vetor de coeficientes, $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_n)^\top$ é o vetor de médias e $\boldsymbol{\sigma}^2 = (\sigma^2, \ldots, \sigma^2)$ é o vetor de variâncias (constante). Utilizando o método de mínimos quadrados, é possível estimar $\boldsymbol{\beta}$, que é obtido através da minimização da soma dos quadrados das diferenças entre as observações Y_i e as médias $\mu_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_r x_{ir}$, em relação aos β's (STASINOPOULOS et al., 2017). Portanto,

$$\hat{\beta} = argmin(\boldsymbol{Y} - \boldsymbol{X\beta})^\top (\boldsymbol{Y} - \boldsymbol{X\beta})$$

possui solução

$$\hat{\beta} = (\boldsymbol{X}^\top \boldsymbol{X})^{-1} \boldsymbol{X}^\top \boldsymbol{Y}$$

Seja $\hat{\boldsymbol{\mu}} = \boldsymbol{X}\hat{\boldsymbol{\beta}}$ os valores ajustados do modelo e $\boldsymbol{\varepsilon} = \boldsymbol{Y} - \hat{\boldsymbol{\mu}}$ o resíduos simples (erros ajustados). Então o estimador de máxima verossimilhança para σ^2 é

$$\sigma^2 = \frac{\hat{\boldsymbol{\varepsilon}}^\top \hat{\boldsymbol{\varepsilon}}}{n},$$

que é um estimador enviesado ($E(\sigma^2 \neq \sigma^2)$. Um estimador não viesado de σ^2 é dado por

$$s^2 = \frac{\hat{\boldsymbol{\varepsilon}}^\top \hat{\boldsymbol{\varepsilon}}}{n-p}$$

. Em algumas ocasiões, s^2 é referido como estimador de máxima verossimilhança restrita de σ^2.

2.2.2 Modelo de regressão logística

Para uma variável resposta binária Y e uma variável explicativa X, seja

$$\pi(x) = P(Y = 1|X = x) = 1 - P(Y = 0|X = x).$$

Agresti (2012) define o modelo de regressão logística por:

$$\pi(x) = \frac{e^{\alpha+\beta x}}{1+e^{\alpha+\beta x}}. \tag{2.3}$$

Equivalentemente, o *logit* ($logit(x) = \log(x/(1-x)); x \in \mathbb{R}$), tem a relação linear

$$\begin{aligned}
logit[\pi(x)] &= \log\left(\frac{\pi(x)}{1-\pi(x)}\right) = \log\pi(x) - \log(1-\pi(x)) \\
&= \log\left(\frac{e^{\alpha+\beta x}}{1+e^{\alpha+\beta x}}\right) - \log\left(1 - \frac{e^{\alpha+\beta x}}{1+e^{\alpha+\beta x}}\right) \\
&= \alpha+\beta x - \log\left(1+e^{\alpha+\beta x}\right) - \log\left(\frac{1}{1+e^{\alpha+\beta x}}\right) \\
&= \alpha+\beta x - \log\left(1+e^{\alpha+\beta x}\right) + \log\left(1+e^{\alpha+\beta x}\right) \\
&= \alpha+\beta x.
\end{aligned}$$

Calculando a chance de sucesso em $x+1$, temos que:

$$\begin{aligned}
\frac{\pi(x+1)}{1-\pi(x+1)} &= \frac{e^{\alpha+\beta(x+1)}/\left(1+e^{\alpha+\beta(x+1)}\right)}{1-e^{\alpha+\beta(x+1)}/\left(1+e^{\alpha+\beta(x+1)}\right)} \\
&= \frac{e^{\alpha+\beta(x+1)}}{1+e^{\alpha+\beta(x+1)}}\frac{1+e^{\alpha+\beta(x+1)}}{1} \\
&= e^{\alpha+\beta(x+1)}.
\end{aligned}$$

Calculando a chance de sucesso em x, temos que:

$$\begin{aligned}
\frac{\pi(x)}{1-\pi(x)} &= \frac{e^{\alpha+\beta x}/\left(1+e^{\alpha+\beta x}\right)}{1-e^{\alpha+\beta x}/\left(1+e^{\alpha+\beta x}\right)} \\
&= \frac{e^{\alpha+\beta x}}{1+e^{\alpha+\beta x}}\frac{1+e^{\alpha+\beta x}}{1} \\
&= e^{\alpha+\beta x}.
\end{aligned}$$

Tomando a razão de chances comparando as chances de sucesso em $x+1$ com as chances de sucesso em x, obtemos:

$$\begin{aligned} \frac{\pi(x+1)/(1-\pi(x+1))}{\pi(x)/(1-\pi(x))} &= \frac{e^{\alpha+\beta(x+1)}}{e^{\alpha+\beta x}} = e^{\beta} \\ \frac{\pi(x+1)}{(1-\pi(x+1))} &= e^{\beta}\frac{\pi(x)}{(1-\pi(x))}. \end{aligned} \tag{2.4}$$

Isso significa que, e^{β} é a mudança multiplicativa nas chances de sucesso correspondente ao acréscimo de uma unidade em x. De (2.4), temos que o logaritmo da razão de chances em x e $x+1$ resulta, para qualquer x:

$$\log\left(\frac{\pi(x+1)/(1-\pi(x+1))}{\pi(x)/(1-\pi(x))}\right) = \beta.$$

Portanto, β é a mudança no logaritmo da razão de chances de sucesso ao acréscimo de uma unidade em x.

Para a equação (2.3) estendida a múltiplos preditores, as chances são:

$$\frac{\pi(x)}{1-\pi(x)} = e^{\alpha+\beta_1 x_1+\cdots+\beta_p x_p}$$

O logaritmo das chances têm a relação linear

$$\log\left(\frac{\pi(x)}{1-\pi(x)}\right) = \alpha+\beta_1 x_1+\cdots+\beta_p x_p$$

Portanto, o modelo de regressão logística é um MLG com componente aleatório binomial e função de ligação *logit*. Os estimadores de máxima verossimilhança dos parâmetros do modelo de regressão logística possuem distribuições normais em grandes amostras. A inferência pode utilizar métodos consagrados como as estatística de *Wald*, os testes de razão de verossimilhança e os testes de hipóteses pontuais. Mais detalhes em Agresti (2012).

2.3 Seleção de variáveis em modelos de regressão

Serão apresentados nos tópicos a seguir os métodos de seleção de variáveis que serão utilizados neste trabalho.

2.3.1 Métodos *stepwise*

Forward stepwise

Forward stepwise é um método de seleção de variáveis que se inicia com um modelo que não contém preditores (chamado *modelo nulo*), e no próximo passo deve-se adicionar os preditores mais significativos, um após o outro. Abaixo, segue o algoritmo.

Quadro 2.1 – Algoritmo *Forward stepwise*

Passo	Descrição
1	Seja M_0 o modelo nulo, que não contém preditores.
2	Para $k = 0, \ldots, p-1$; (a) Considere todos os $p-k$ modelos que aumentam os preditores em M_k com um preditor adicional. (b) Escolha o melhor entre esses $p-k$ modelos e chame-o de M_{k+1}. O melhor modelo é definido como tendo a menor soma dos quadrados dos resíduos (residual sum of squares - RSS) ou o maior R^2.
3	Selecione um único melhor modelo entre $M_0, \ldots, M_p$ usando erro de previsão de validação cruzada, $C_p\,(AIC)$, (BIC) ou R^2 ajustado.

Fonte: James et al. (2013)

Backward stepwise

Backward stepwise é um método de seleção de variáveis que, inicia-se com um modelo que contém todos os preditores em consideração (chamado *modelo completo*), e no próximo passo deve-se remover os preditores menos significativos. O algoritmo a seguir foi retirado de James et al. (2013).

Quadro 2.2 – Algoritmo *Backward stepwise*

Passo	Descrição
1	Seja M_p o modelo completo, que contém todos os p preditores.
2	Para $k = p, p-1, \ldots, 1$; (a) Considere todos os k modelos que contém todos menos um dos preditores em M_k, para um total de $k-1$ preditores. (b) Escolha o melhor entre esses k modelos e chame-o de M_{k-1}. O melhor modelo é definido como tendo o menor RSS ou o maior R^2.
3	Selecione um único melhor modelo entre $M_0, \ldots, M_p$ usando erro de previsão de validação cruzada, $C_p\,(AIC)$, (BIC) ou R^2 ajustado.

Fonte: James et al. (2013)

Bidirectional stepwise

Bidirectional stepwise é uma combinação dos métodos *Forward stepwise* e *Backward stepwise*. Na etapa na qual uma variável foi adicionada (passo *forward*), todas as variáveis candidatas no modelo são verificadas para ver se sua significância foi reduzida abaixo do nível de tolerância especificado. Se uma variável não significativa for encontrada, ela é removida do modelo (passo *backward*). Finaliza-se o algoritmo quando não há variáveis a serem adicionadas e nem retiradas do modelo.

2.3.2 *Purposeful selection of covariates - PVS*

Hosmer e Lemeshow (2000) propuseram um método de seleção de variáveis no contexto de regressão logística, chamado *Purposeful selection of covariates*, para o qual usaremos a sigla PVS e que consiste em sete passos:

a) realizar uma análise inicial univariada para cada variável independente. Para variáveis categóricas, os autores sugerem que seja realizada uma análise via tabela de contingência padrão, das respostas $(y = 0, 1)$ contra os i níveis da variável independente. Para variáveis contínuas, a melhor análise univariada envolve o ajuste de um modelo de regressão logística para obter o coeficiente estimado, o teste de razão de verossimilhanças para a significância do coeficiente e a estatística de Wald univariada. Através do uso dessas análises univariadas, serão consideradas candidatas a compor um primeiro modelo multivariado quaisquer variáveis cujo teste apresente probabilidade de significância menor que $0,25$. Além disso são adicionadas as variáveis de importância clínica suposta conhecida;

b) ajustar o modelo contendo todas as covariáveis identificadas para inclusão no primeiro passo. Seguindo o ajuste desse modelo, avalia-se a importância de cada covariável usando o valor p de sua estatística de Wald. As variáveis que não contribuírem, nos níveis tradicionais de significância estatística, devem ser eliminadas e um novo modelo deve ser ajustado. O novo modelo menor deve ser comparado ao modelo antigo maior, usando o teste de razão de verossimilhanças parcial;

c) comparar os valores dos coeficientes estimados no modelo menor com seus respectivos valores do modelo maior. Deve-se preocupar com qualquer variável cujo coeficiente mudou marcadamente em magnitude. Isso indica que uma ou mais das variáveis excluídas

são importantes no sentido de fornecer um ajuste necessário do efeito das variáveis que permaneceram no modelo. Essas variáveis devem ser adicionadas de volta ao modelo;

d) adicionar cada variável não selecionada no primeiro passo, ao modelo obtido na conclusão do ciclo do segundo e terceiro passo, uma de cada vez. Verificar a significância pelo valor *p* da estatística Wald ou pelo teste da razão de verossimilhança parcial, se for uma variável categórica com mais de dois níveis. Ao final desta etapa, o modelo obtido é chamado de *modelo preliminar de efeitos principais*;

e) obter o modelo preliminar de efeitos principais e reexaminar as variáveis do modelo. Checar se as categorias para as variáveis categóricas estão apropriadas. Para cada variável contínua, deve-se verificar a suposição de que o logit aumenta ou diminui linearmente como uma função da covariável. O modelo obtido ao final desse passo é chamado de *modelo de efeitos principais*;

f) verificar as interações entre as variáveis no modelo. A presença de interação entre duas variáveis implica que o efeito de uma não é constante ao longo dos níveis da outra;

g) avaliar a adequação e verificar o ajuste do modelo obtido.

Bursac et al. (2008), realizaram um estudo de simulação comparando os método PVS, *forward stepwise, backward stepwise* e *bidirectional stepwise*. Os autores concluem que a PVS identificou e reteve os fatores de confusão corretamente, a uma taxa maior do que os demais procedimentos de seleção.

O método PVS é de grande importância, pois permite ao pesquisador incorporar seu conhecimento no processo de seleção de variáveis, diferentemente dos métodos automáticos de decisão. Essa abordagem oferece uma maneira mais informada de construir modelos preditivos, valorizando o conhecimento do pesquisador e aumentando a relevância prática dos resultados. O uso do método PVS por Pereira et al. (2020b) nos inspirou a explorar outras abordagens com características similares, levando-nos a considerar as redes bayesianas (RB). As RB também permitem a integração de informações a priori, proporcionando um meio eficaz de combinar conhecimento prévio com dados empíricos, resultando em modelos mais interpretáveis.

2.4 Redes bayesianas

As redes bayesianas são grafos acíclicos direcionados em que os vértices (ou nós) representam proposições (ou variáveis), as arestas (ou arcos) quando são direcionadas significam a

existência de influências causais diretas entre as proposições ligadas, e as forças dessas influências são quantificadas por probabilidades condicionais (PEARL, 1985). Uma característica importante da rede bayesiana é fornecer uma representação gráfica clara para muitos relacionamentos de interdependência embutidos no modelo probabilístico subjacente.

Para compreender o conceito sobre redes bayesianas, é necessário abordar algumas definições relacionadas à teoria de grafos. As definições a seguir foram baseadas em Koski e Noble (2009), Neapolitan (2003) e Pearl (2009).

2.4.1 Grafos

Definição 2.4.1 (Vértice) *É uma coleção de componentes de um vetor aleatório* $\boldsymbol{X} = (X_1, \ldots, X_d)$.

Definição 2.4.2 (Grafo, grafo simples) *Um* ***grafo*** $\mathbb{G} = (V, E)$ *consiste em um conjunto finito V de vértices e um conjunto de arestas E, em que cada aresta está contida em* $V \times V$. *O conjunto de arestas, portanto, consiste em pares ordenados de vértices.*

Seja $V = \{\alpha_1, \alpha_2, \ldots, \alpha_d\}$. *Um grafo é considerado simples se:*

a) E não contém nenhuma aresta da forma (α_j, α_j);

b) qualquer aresta $(\alpha_j, \alpha_k) \in E$ *aparece exatamente uma vez.*

Para quaisquer dois vértices distintos α *e* $\beta \in V$, *o par ordenado* $(\alpha, \beta) \in E$, *se, e somente se, houver uma aresta direcionada para* α *e* β. *Uma aresta não direcionada será denotada por* $\langle \alpha, \beta \rangle$. *Em termos de arestas direcionadas*

$$\langle \alpha, \beta \rangle \in E \Leftrightarrow (\alpha, \beta) \in E \quad e \quad (\alpha, \beta) \in E.$$

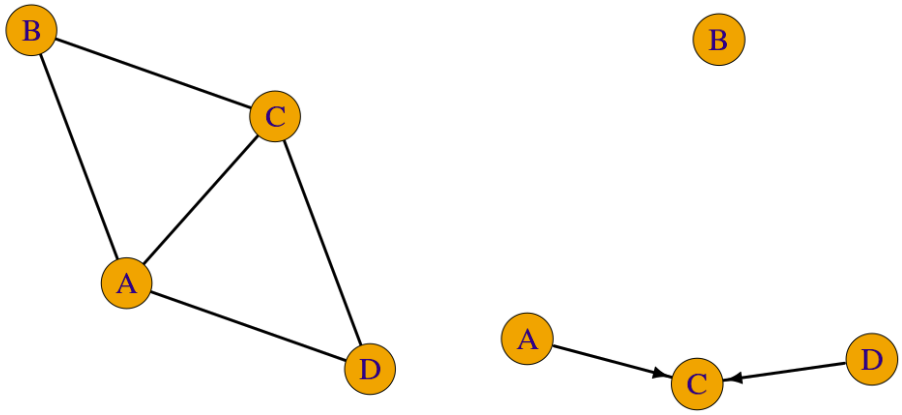
Um ***grafo simples*** *pode conter arestas direcionadas e arestas não direcionadas, e o conjunto de arestas E pode ser decomposto com* $E = D \cup U$, *em que* $D \cap U = \emptyset$. *Os conjuntos U e D são definidos por*

$$\langle \alpha, \beta \rangle \in U \Leftrightarrow (\alpha, \beta) \in E \qquad e \quad (\beta, \alpha) \in E.$$
$$(\alpha, \beta) \in D \Leftrightarrow (\alpha, \beta) \in E \qquad e \quad (\beta, \alpha) \notin E.$$

Em que D é o conjunto de arestas direcionadas e U o conjunto de arestas não direcionadas.

Para o grafo não direcionado, as arestas não indicam causa e efeito entre os vértices, diferentemente do grafo direcionado, pois nestes as arestas indicam relações de causa e efeito entre as variáveis. Portanto, se $A \rightarrow B$, temos que a variável A afeta a variável B ou que a variável B depende da variável A. Graficamente, a representação das relações entre as variáveis é feita através de setas.

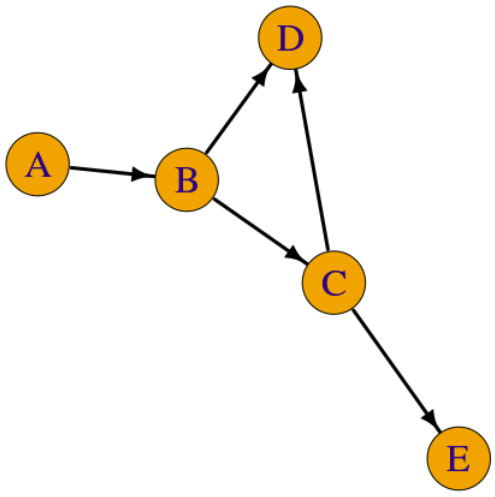
Figura 2.1 – Exemplo de grafo simples direcionado e grafo simples não direcionado respectivamente.



Fonte: Do autor (2024).

Definição 2.4.3 (Caminho, caminho direcionado) *Seja $\mathbb{G} = (V,E)$ um grafo simples, em que $E = D \cup U$. Um* ***caminho*** *de tamanho m de um vértice α até um vértice β é uma sequência de vértices distintos $(\tau_0, \ldots, \tau_m)$ tal que $\tau_0 = \alpha$ e $\tau_m = \beta$ de modo que $(\tau_{i-1}, \tau_i) \in E$ para cada $i = 1, \ldots, m$. Ou seja, para cada $i = 1, \ldots m$; $(\tau_{i-1}, \tau_i) \in D$, ou $\langle \tau_{i-1}, \tau_i \rangle \in U$. Um caminho é dito* ***direcionado*** *se $(\tau_{i-1}, \tau_i) \in D$ para cada $i = 1, \ldots, m$.*

Figura 2.2 – Exemplo de Grafo.

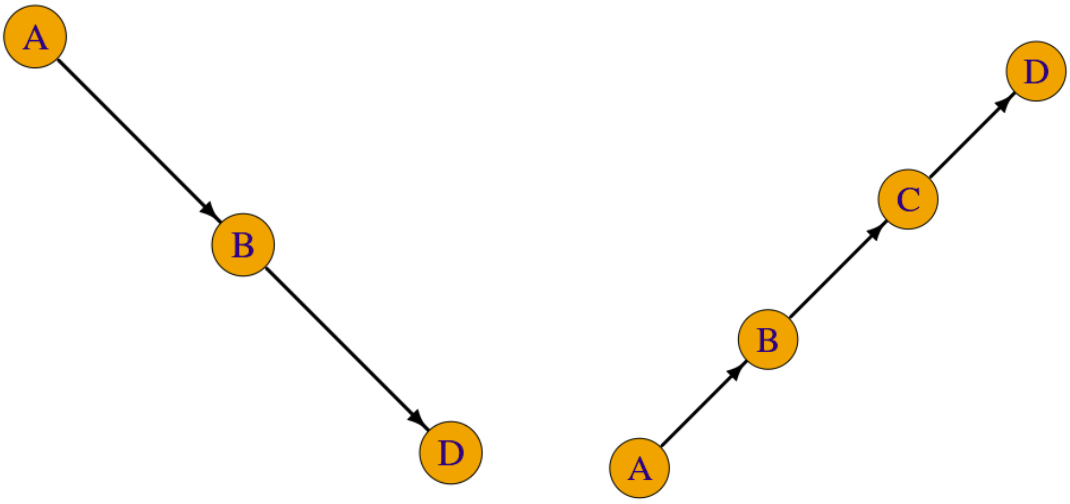


Fonte: Do autor (2024).

Um caminho é qualquer rota ininterrupta, sem interseção, traçada ao longo das arestas de um grafo, que pode ir ao longo ou contra as setas. Se cada aresta em um caminho é uma seta que aponta do primeiro ao segundo vértice do par, temos um caminho direcionado.

Considerando o grafo da figura 2.2, existem dois caminhos de A para D.

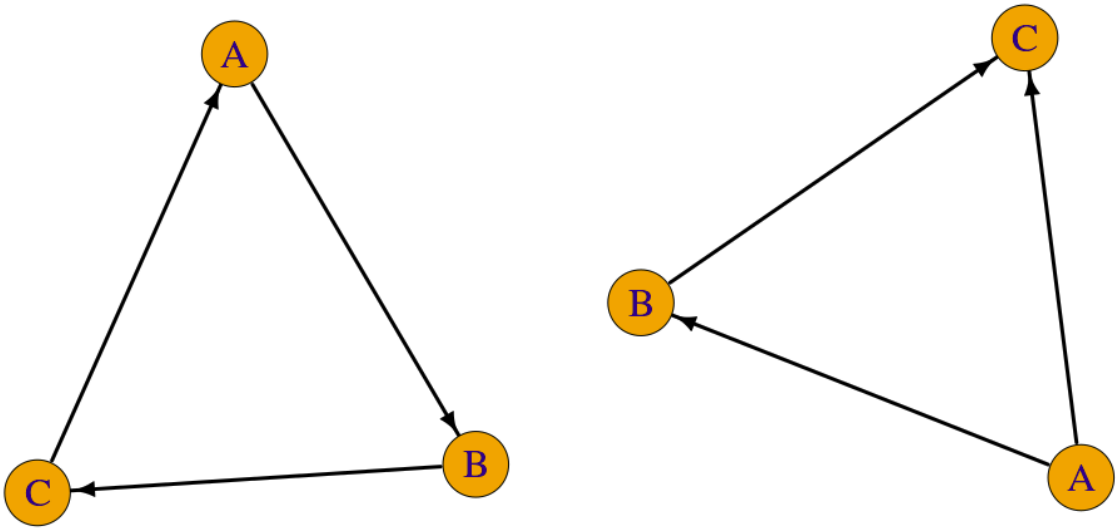
Figura 2.3 – Exemplo de caminho.



Fonte: Do autor (2024).

Definição 2.4.4 (Ciclo) *Seja* $\mathbb{G} = (V, E)$ *um grafo. Um* ***ciclo m*** *em* $\mathbb{G}$ *é uma sequência de vértices distintos* $\tau_i, \ldots, \tau_{m-1}, \tau_0$ *é um caminho.*

Figura 2.4 – Exemplo de grafo cíclico e acíclico respectivamente.



Fonte: Do autor (2024).

Definição 2.4.5 (Grafo acíclico direcionado (DAG)) *Um grafo $\mathbb{G} = (V,E)$ é considerado um **grafo acíclico direcionado** se cada aresta for direcionada (isto é, $\mathbb{G}$ é um grafo simples de modo que para cada par $(\alpha,\beta) \in V \times V$, $(\alpha,\beta) \in E \Rightarrow (\beta,\alpha) \notin E$) e para qualquer vértice $\alpha \in V$ não existe qualquer conjunto de vértices distintos $\tau_1, \ldots, \tau_m$, de tal modo que $\alpha \neq \tau_i, \forall i = 1, \ldots, m$ e $(\alpha, \tau_1, \ldots, \tau_m, \alpha)$ forma um caminho direcionado.*

Ou seja, de acordo com a definição 2.4.5, um grafo será considerado um DAG quando todos vértices são conectados na estrutura gráfica e todas as arestas direcionadas de forma que não haja ciclos.

Definição 2.4.6 (Pai, descendente, ancestral, não-descendente) *Dado um DAG $(\mathbb{G},E)$ e vértices α e $\beta \in V$, β é chamado de **pai** de α se houver uma aresta de β a α. β é chamado de **descendente** de α e α é chamado de **ancestral** de β se houver um caminho de α para β, e β é chamado de **não descendente** de α se β não for descendente de α.*

O grafo da figura 2.2 é considerado um DAG, pois ele é direcionado e não há ciclos. Podemos retirar algumas informações sobre esse DAG:

a) A é pai de B, pois A $\rightarrow$ B;
b) C é filho de B, pois B $\rightarrow$ C;
c) D é um descendente de A, pois A é pai de B e B é pai de D;
d) A é um ancestral de D;
e) D possui dois pais, B e C;
f) C é pai de E;
g) E é não descendente de D, pois não há um caminho de D para E;
h) E é descendente de A, pois A é pai de B, B é pai de C e C é pai de E;

Definição 2.4.7 (Condição de Markov) *Suponha que tenha-se uma distribuição de probabilidade conjunta P das variáveis aleatórias de algum conjunto V e um DAG $\mathbb{G} = (V,E)$. Se para cada variável $X \in V$, $\{X\}$ é condicionalmente independente de todos os seus não descendentes dado o conjunto de todos os seus pais, então diz-se que $(\mathbb{G},P)$ satisfaz a **Condição de Markov**.*

Definição 2.4.8 (Redes bayesianas) *Seja P uma distribuição de probabilidade conjunta das variáveis aleatórias de algum conjunto $\boldsymbol{V}$, e $\mathbb{G} = (V,E)$ um DAG. $(\mathbb{G},P)$ é chamado de **redes bayesianas** se $(\mathbb{G},P)$ satisfazer a condição de Markov.*

O esquema básico de decomposição oferecido por DAG pode ser ilustrado da seguinte forma. Suponha que tenha-se uma distribuição P definida em n variáveis discretas, que podem

ser ordenadas arbitrariamente como $X_1, X_2, \ldots, X_n$. A regra da cadeia do cálculo de probabilidade nos permite decompor P como um produto de n distribuições condicionais

$$P(x_1, \ldots, x_n) = \prod_j P(x_j | x_1, \ldots, x_{j-1})$$

Supondo que a probabilidade de alguma variável X_j não seja sensível a todos os predecessores de X_j, mas apenas a um subconjunto desses predecessores, ou seja, que X_j seja independente de todos os outros predecessores, uma vez que conhecemos o valor de um seleto grupo de predecessores chamado PA_j. Pode-se escrever então

$$P(x_j | x_1, \ldots, x_{j-1}) = P(x_j | pa_j)$$

O conjunto PA_j é chamado de **pais Markovianos** de X_j.

Definição 2.4.9 *Seja $V = X_1, \ldots, X_n$ um conjunto ordenado de variáveis, e seja $P(v)$ a distribuição de probabilidade conjunta dessas variáveis. Um conjunto de variáveis PA_j é chamado de pais Markovianos de X_j se PA_j é um conjunto mínimo de predecessores de X_j que torna X_j independente de todos os seus outros predecessores. Em outras palavras, PA_j é qualquer subconjunto de $X_1, \ldots, X_{j-1}$ satisfazendo*

$$P(x_j | pa_j) = P(x_j | x_1, \ldots, x_{j-1}) \tag{2.5}$$

e tal que nenhum subconjunto próprio de PA_j satisfaça a equação (2.5).

A definição 2.4.9 atribui a cada variável X_j um conjunto selecionado de variáveis precedentes que são suficientes para determinar a probabilidade de X_j. Portanto, toda distribuição que satisfaça (2.5) deve-se decompor no produto

$$P(x_1, \ldots, x_n) = \prod_i P(x_i | pa_i).$$

Por exemplo o DAG da figura 2.2 induz a decomposição

$$P(A, B, C, D, E) = P(A)P(B|A)P(C|B)P(D|B,C)P(E|C).$$

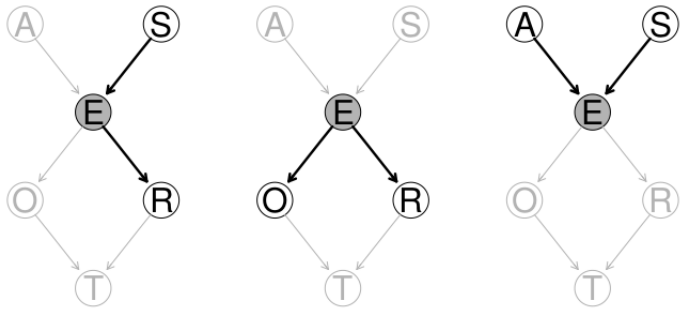
Definição 2.4.10 (d-Separação) *Um caminho p é dito d-separado (ou bloqueado) por um conjunto de vértices Z se, e somente se,*

a) p contiver uma cadeia $i \to m \to j$ ou um garfo $i \leftarrow m \to j$, tal que o vértice do meio m esteja em Z, ou

b) p contiver um garfo invertido (ou colisor) $i \to m \leftarrow j$, tal que o vértice do meio m não esteja em Z e que nenhum descendente de m esteja em Z

Diz-e que um conjunto X é d-separado de Y se, e somente se, Z bloqueia todos os caminhos de um vértice em X para um vértice em Y. Denota-se por $X \perp\!\!\!\perp_G Y|Z$.

Figura 2.5 – Exemplos de d-Separação: conexão em cadeia (esquerda), conexão em garfo (centro) e conexão de garfo invertido (direita)



Fonte: Scutari e Denis (2021)

A separação gráfica ($\perp\!\!\!\perp_G$) implica independência probabilística ($\perp\!\!\!\perp_P$) em uma RB, se todos os caminhos entre X e Y estiverem bloqueados, X e Y são (condicionalmente) independentes. A recíproca não é verdadeira, nem toda relação de independência condicional é refletida no gráfico (SCUTARI; DENIS, 2021).

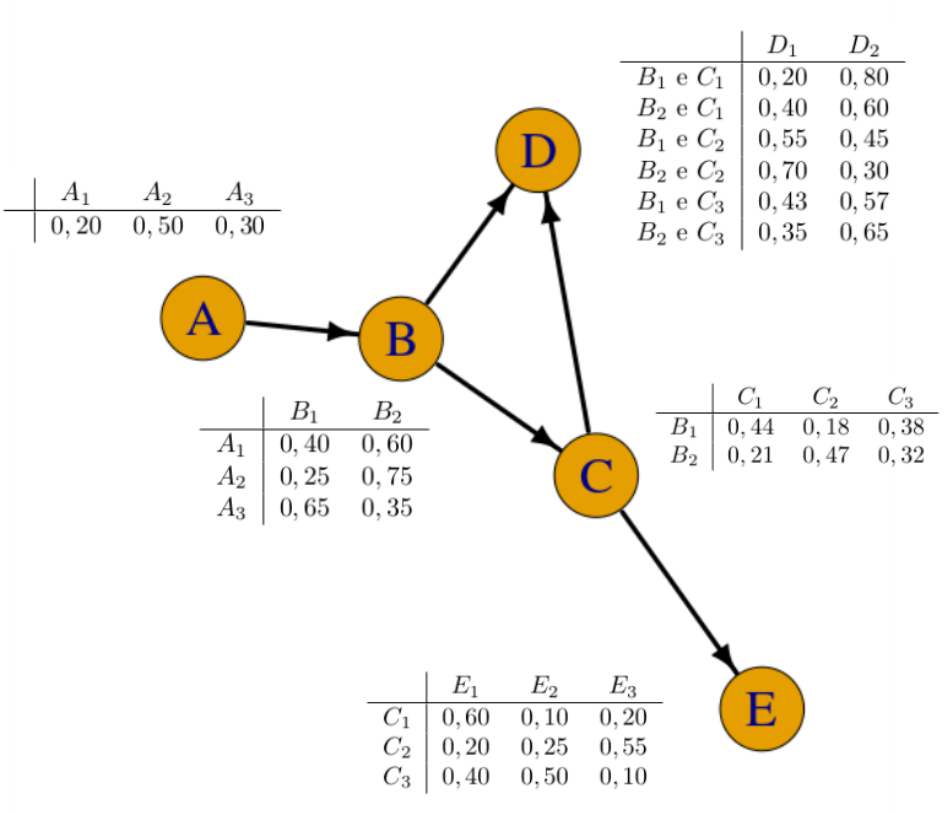
2.4.2 Rede bayesiana discreta

Para conjuntos de dados discretos ou categóricos, frequentemente referidos como casos discretos, tanto a distribuição global quanto a local são multinomiais, sendo estas últimas representadas como tabelas de probabilidades condicionais (CPT). Este caso é a suposição mais comum na literatura, e as redes bayesianas correspondentes são denominadas redes bayesianas discretas (NAGARAJAN; SCUTARI; LBRE, 2013).

As probabilidades das distribuições locais podem ser determinadas através do cálculo das distribuições condicionais.

$$P(B|A) = \frac{P(B \cap A)}{P(A)}.$$

Figura 2.6 – DAG representando as relações de dependência entre as variáveis e as tabelas de probabilidade condicional.



Fonte: Do autor (2024).

Dessa forma, considerando o DAG da Figura 2.6, podemos calcular a seguinte probabilidade

$$P(E = E_3|C = C_2) = \frac{P(E = E_3, C = C_2)}{P(C = C_2)} = 0,55$$

Da mesma maneira, é possível calcular as probabilidades de $P(A)$, $P(B|A)$, $P(C|B)$, $P(D|B,C)$ e $P(E|C)$.

Em geral, em RB discretas, os parâmetros a estimar são as probabilidades condicionais nas distribuições locais (SCUTARI; DENIS, 2021). Eles podem ser estimados, por exemplo, como as frequências empíricas correspondentes no conjunto de dados

$$\begin{aligned} \hat{P}(E = E_3|C = C_2) &= \frac{\hat{P}(E = E_3, C = C_2)}{\hat{P}(C = C_2)} \\ &= \frac{\text{número de observações para as quais } E = E_3 \text{ e } C = C_2}{\text{número de observações para as quais } C = C_2} \end{aligned}$$

Isso resulta nas estimativas clássicas frequentistas e de máxima verossimilhança.

2.4.3 Rede bayesiana gaussiana

As seguintes suposições caracterizam as redes lineares gaussianas bayesianas (RBG) (SCUTARI; DENIS, 2021):

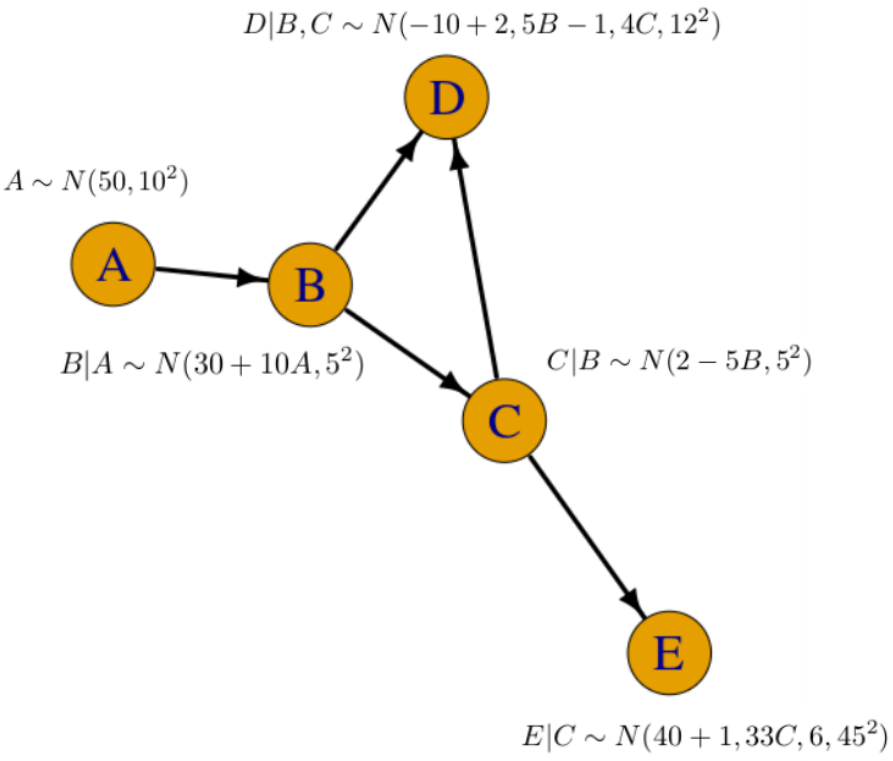
a) cada nó segue uma distribuição normal;

b) nós sem nenhum pai, conhecidos como nós raiz, são descritos pelas respectivas distribuições marginais;

c) o efeito condicionante dos nós pais é dado por um termo linear aditivo na média e não afeta a variância. Em outras palavras, cada nó possui uma variância específica daquele nó e não depende dos valores dos pais;

d) a distribuição local de cada nó pode ser expressa de forma equivalente como um modelo linear gaussiano que inclui uma interceptação e os pais do nó como variáveis explicativas, sem qualquer termo de interação.

Figura 2.7 – DAG representando as relações de dependência entre as variáveis e as distribuições de probabilidade locais são mostradas para cada nó.



Fonte: Do autor (2024).

A partir dos pressupostos paramétricos, temos que cada distribuição local pode ser expressa como um modelo clássico de regressão linear gaussiana em que o nó é a variável resposta, por exemplo (D) e seus pais são as variáveis explicativas (C, B). As contribuições dos pais são puramente aditivas; o modelo não contém nenhum termo de interação, apenas o efeito principal de cada pai $(B + C)$ e o intercepto.

2.4.4 Rede bayesiana gaussiana condicional

Para o caso em que as variáveis são tanto contínuas quanto discretas, pode-se recorrer às redes bayesianas gaussianas condicionais (RBGC), que seguem as seguintes pressuposições (SCUTARI; DENIS, 2021):

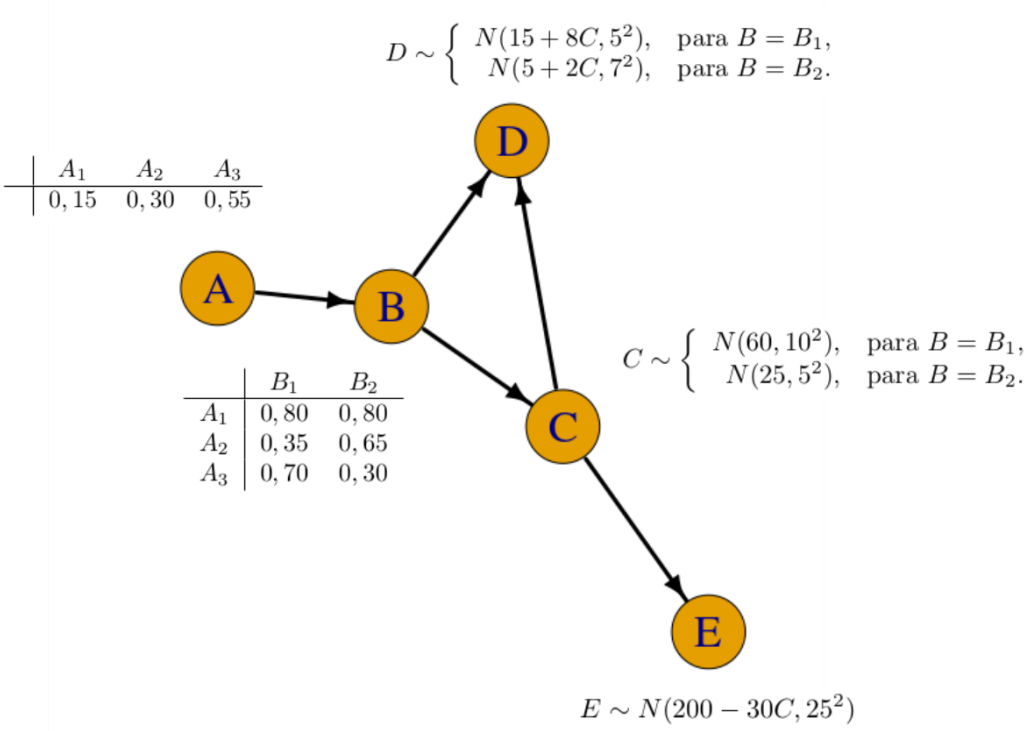
a) nós discretos seguem uma distribuição multinomial;
b) nós contínuos que não possuem nenhum nó discreto entre seus pais seguem uma distribuição normal;
c) nós contínuos que possuem um ou mais nós discretos entre seus pais seguem uma mistura de distribuições normais com um componente para cada combinação dos valores desses pais discretos;
d) cada distribuição normal em cada mistura possui média e variância separadas, ou seja, cada componente da mistura possui um conjunto independente de parâmetros;
e) os nós contínuos podem ter nós contínuos e nós discretos como pais, mas os nós discretos só podem ter outros nós discretos como pais.

Estimar os parâmetros de uma RBGC quando seu DAG é conhecido é uma extensão dos casos das Seções 2.4.2 e 2.4.3. No caso de variáveis discretas, as probabilidades condicionais de um nó (por exemplo, B) dados os seus pais são estimadas utilizando frequências empíricas.

Quanto aos nós contínuos sem pai discreto (por exemplo, E), porém com alguns pais contínuos, os parâmetros das distribuições locais são os coeficientes de regressão associados aos pais e o desvio padrão dos resíduos.

Para os nós contínuos com pelo menos um pai discreto são modelados com uma mistura de regressões lineares. Considere o nó C, possui apenas o nó B discreto como pai, pode-se calcular as estimativas dos parâmetros de máxima verossimilhança para tal mistura dividindo os dados nos subconjuntos correspondentes aos dois valores de B ("B_1", "B_2") e ajustando um modelo linear separado para cada subconjunto.

Figura 2.8 – Exemplo de Grafo.



Fonte: Do autor (2024).

2.4.5 Processo de modelagem das redes bayesianas

Segundo Scutari e Denis (2021), a seleção do modelo e a estimativa de RB são conhecidas coletivamente como aprendizado e geralmente são realizadas como um processo de duas etapas.

a) *structure learning*, a aprendizagem da estrutura do DAG.
b) *parameter learning*, a aprendizagem sobre as distribuições locais implícitas na estrutura do DAG aprendida na etapa anterior.

Ambas as etapas podem ser realizadas como aprendizado não supervisionado, usando as informações fornecidas por um conjunto de dados, ou como aprendizado supervisionado, entrevistando especialistas nas áreas relevantes para o fenômeno que está sendo modelado.

A aprendizagem da RB pode ser descrita como

$$\underbrace{P(B|D) = P(\mathbb{G}, \Theta|D)}_{\text{aprendizagem}} = \underbrace{P(\mathbb{G}|D)}_{\text{aprendizagem da estrutura}} \cdot \underbrace{P(\Theta|\mathbb{G}, D)}_{\text{aprendizagem dos parâmetros}}$$

em que:

a) D é o conjunto de dados;

b) $\mathbb{G}$ representa a estrutura gráfica;

c) Θ são os parâmetros da distribuição global;

d) a RB é dada por $B = (\mathbb{G}, \Theta)$.

O aprendizado da estrutura pode ser feito na prática encontrando o DAG $\mathbb{G}$ que maximiza

$$P(\mathbb{G}|D) \propto P(\mathbb{G})P(D|\mathbb{G}) = P(\mathbb{G}) \int P(D|\mathbb{G}, \Theta)P(\Theta|\mathbb{G})d\Theta \tag{2.6}$$

2.4.6 Algoritmos e implementações de redes bayesianas

Os algoritmos que são utilizados para a aprendizagem da estrutura das redes bayesianas podem ser de três tipos: *constraint-based*, *score-based* e *hybrid*. De acordo com Scutari e Denis (2021) todos esses algoritmos de aprendizado de estrutura operam sob um conjunto de suposições comuns:

a) deve haver uma correspondência um-para-um entre os nós no DAG e as variáveis aleatórias;

b) todas as relações entre as variáveis devem ser independências condicionais, pois, por definição, o único tipo de relação que pode ser expressa por uma RB;

c) cada combinação dos valores possíveis das variáveis devem representar um evento válido, observável (mesmo que realmente improvável). Esta suposição implica uma distribuição global estritamente positiva;

d) as observações são tratadas como realizações independentes do conjunto de nós. Se alguma forma de dependência temporal ou espacial estiver presente, ela deve ser considerada especificamente na definição da rede, como nas redes bayesianas dinâmicas.

Os algoritmos *constraint-based* aprendem a estrutura da rede, analisando as relações probabilísticas decorrentes da propriedade de Markov das redes bayesianas com testes de independência condicional e, em seguida, construindo um gráfico que satisfaça as declarações de d-Separação correspondentes. Esses algoritmos são todos baseados no algoritmo de causação indutiva (IC) de Verma e Pearl (1990), que fornece uma estrutura teórica para aprender os modelos causais de estrutura (SCUTARI, 2010).

Para esses algoritmos, pode ser empregado alguns testes de independência condicional, como: informação mútua, qui-quadrado (para dados categóricos), correlação linear e transformação de Fisher.

Os algoritmos *score-based* representam a aplicação de técnicas de otimização heurística ao problema de aprendizagem da estrutura de uma RB. Cada RB candidata, rececebe uma pontuação de rede refletindo sua qualidade de ajuste, que o algoritmo tenta maximizar (SCUTARI; DENIS, 2021).

Algumas pontuações de rede que estão implementadas no pacote *bnlearn* (SCUTARI, 2010): log-verossimilhança, log-verossimilhança preditiva, K2, AIC, BIC, diferentes verossimilhanças marginais de Dirichlet (BDe, BDs, BDJ), um Dirichlet Bayesiano modificado para misturas de dados intervencionais e observacionais, o escore BDe localmente médio (BDla), o logaritmo da máxima verossimilhança normalizada fatorada (fNML), o logaritmo da máxima verossimilhança normalizada quociente (qNML), a densidade posterior gaussiana equivalente ao escore BGe.

O *Hill-Climbing*, é um algoritmo *score-based* e está descrito abaixo:

Quadro 2.3 – Algoritmo *Hill-Climbing*

Passo	**Descrição**
1	Escolha uma estrutura de rede G sobre V, geralmente (mas não necessariamente) vazia;
2	Calcule o *score* de G, denotada como $Score_G$ = *Score*(G);
3	Defina $maxscore = score_G$;
4	Repita as etapas a seguir enquanto a pontuação máxima aumentar: (a) Para cada possível adição, exclusão ou reversão de arco que não resulte em uma rede cíclica: - Calcule o *score* da rede modificada G^*, $Score_{G^*} = Score_{G^*}$; - Se $Score_{G^*} > Score_G$, defina $G = G^*$ e $Score_G = Score_{G^*}$; (b) Atualize *maxscore* com o novo valor de $Score_G$;
5	Retorne o DAG G.

Fonte: Nagarajan, Scutari e Lbre (2013)

O algoritmo *hybrid* combina os algoritmos *constraint-based* e *score-based*. Ambos os algoritmos são baseados em duas etapas, chamadas restringir e maximizar. No primeiro, o conjunto candidato para os pais de cada nó X_i é reduzido de todo o conjunto de nós V para um

conjunto menor $C_i \in V$ de nós cujo comportamento se mostrou relacionado de alguma forma ao de X_1 . Isso, por sua vez, resulta em um espaço de pesquisa menor e mais regular. A segunda etapa busca a rede que maximiza uma determinada função de pontuação, sujeita às restrições impostas pelos conjuntos C_i (NAGARAJAN; SCUTARI; LBRE, 2013).

De acordo com Scutari e Denis (2021), pode-se destacar alguns pacotes do *software* R que tratam de RB (Quadro 2.4).

Quadro 2.4 – Pacotes do R relacionados à Rede Bayesiana.

	bnlearn	*catnet*	*deal*	*pcalg*	*abn*	*gRbase*	*gRain*	*rbmn*
Dados discretos	x	x	x	x	x	x	x	
Dados contínuos	x		x	x	x	x		x
Dados mistos	x		x		x			
Aprendizagem *constraint-based*	x			x				
Aprendizagem *score-based*	x	x	x		x			
Aprendizagem *hybrid*	x							
Manipulação de estrutura	x	x			x	x		
Estimativa de parâmetros	x	x	x	x	x			x
Predição	x	x					x	x
Inferência	x	x					x	

Fonte: Scutari e Denis (2021)

Sobre o pacote *bnlearn*, podemos destacar os algoritmos *structure learning* que estão implementados.

Quadro 2.5 – Algoritmos implementados no pacote *bnlearn*.

Algortimos	*Constraint-based*	*Score-based*	*Hybrid*
PC	x		
Grow-Shrink (GS)	x		
Incremental Association (IAMB)	x		
Fast Incremental Association (Fast-IAMB)	x		
Interleaved Incremental Association (Inter-IAMB)	x		
Incremental Association with FDR (IAMB-FDR)	x		
Max-Min Parents and Children (MMPC)	x		
Semi-Interleaved HITON-PC	x		
Hybrid Parents and Children (HPC)	x		
Hill-climbing (HC)		x	
Tabu search (TABU)		x	
Max-Min Hill Climbing (MMHC)			x
Hybrid HPC (H2PC)			x
General 2-phase Restricted Maximization (RSMAX2)			x

Fonte: Do autor (2024).

Uma vez aprendida a estrutura do BN a partir dos dados, a tarefa de estimar e atualizar os parâmetros da distribuição global é bastante simplificada pela decomposição em distribui-

ções locais. Duas abordagens são comuns na literatura: *máxima verossimilhança* e *estimativa bayesiana* (SCUTARI; DENIS, 2021).

Na prática, as distribuições locais envolvem apenas um pequeno número de variáveis. O número de parâmetros necessários para identificar exclusivamente a distribuição global, que é a soma do número de parâmetros das distribuições locais, também é reduzido porque as relações de independência condicional codificadas na estrutura da rede fixam grandes partes do espaço de parâmetros (NAGARAJAN; SCUTARI; LBRE, 2013).

2.4.6.1 Listas de arcos permitidos e arcos proibidos

As informações prévias sobre os dados, como aquelas fornecidas por especialistas nas áreas pertinentes, podem ser incorporadas em todos os algoritmos de aprendizagem utilizando os conceitos de *whitelist* e *blacklist*. Ambos os métodos permitem especificar conjuntos de conexões que devem estar presentes (*whitelist*) ou ausentes (*blacklist*) na rede bayesiana. Essa abordagem oferece uma maneira altamente flexível de representar suposições específicas sobre os dados e é capaz de lidar com estruturas de rede parcialmente direcionadas (SCUTARI, 2010):

a) qualquer arco colocado na *whitelist* em uma de suas direções possíveis (por exemplo, $A \rightarrow B$ está na lista, mas $B \rightarrow A$ não) é garantido que estará no grafico na direção especificada;
b) qualquer arco colocado na *blacklist* em uma de suas direções possíveis (por exemplo, $A \rightarrow B$ está na *blacklist*, mas $B \rightarrow A$ não está) nunca estará presente no gráfico. O mesmo vale para $A - B$, mas não para $B \rightarrow A$;
c) qualquer arco colocado na *blacklist* em ambas as direções, bem como o arco não direcionado correspondente, nunca está presente no grafo. Portanto, se $A \rightarrow B$ e $B \rightarrow A$ estiverem na *blacklist*, $A - B$ também será considerado na *blacklist*.

No nosso trabalho, destacamos os nós e arcos presentes em uma *whitelist* na RB utilizando a cor azul. Esta abordagem visual permite identificar facilmente as relações e variáveis específicas que foram previamente selecionadas. Ao realçar esses componentes, facilitamos a interpretação dos dados e ressaltamos a integração do conhecimento prévio no processo de modelagem.

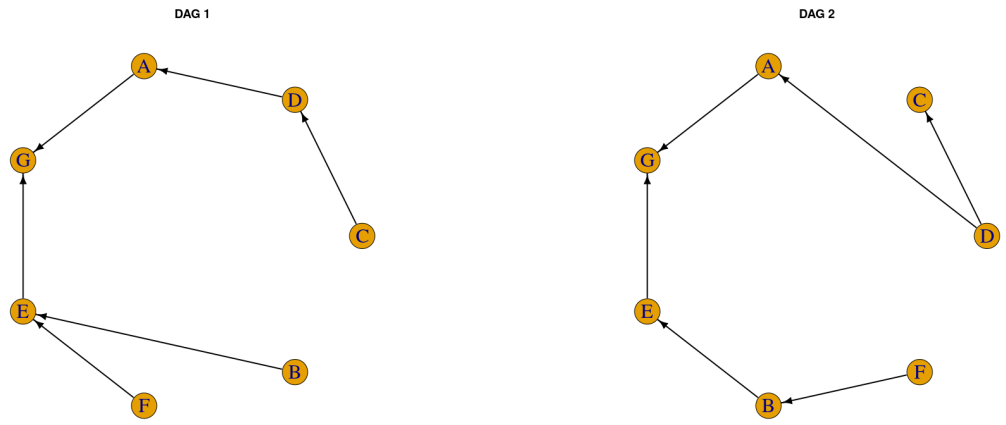
2.4.7 Comparando estruturas de rede bayesianas

A comparação entre duas estruturas de redes bayesianas pode ser feita assumindo uma delas como referência e comparando a outra em relação à primeira. Nesse processo, verifica-se a presença ou ausência de arcos em comum, assim como diferenças estruturais, identificando semelhanças e divergências entre as redes.

Considerando a Figura 2.9, podemos tomar o DAG 1 como referência e compará-lo com o DAG 2. Três características podem ser utilizadas para essa comparação:

a) arcos *verdadeiros positivos* (VP): aqueles que aparecem tanto no DAG 1 quanto no DAG 2;
b) arcos *falsos positivos* (FP): aqueles que aparecem no DAG 2, mas não no DAG 1;
c) arcos *falsos negativos* (FN): aqueles que aparecem no DAG 1, mas não no DAG 2.

Figura 2.9 – Estruturas de redes bayesianas que serão comparadas.

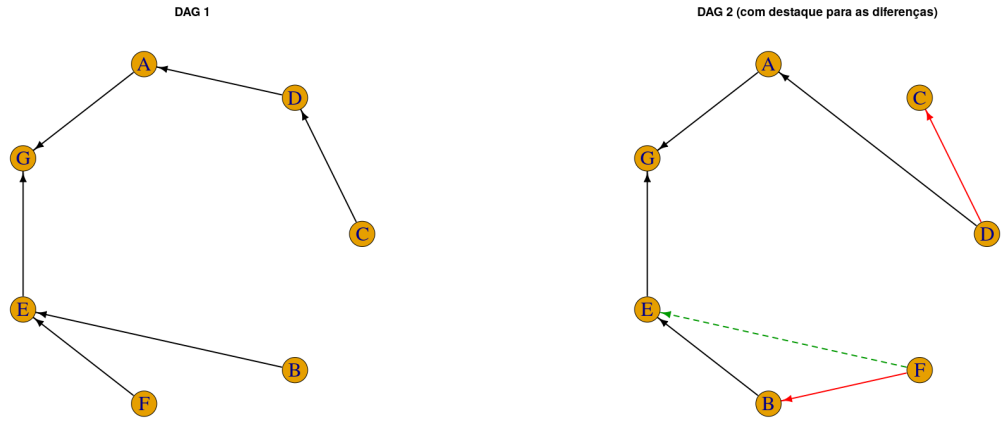


Fonte: Do autor (2024).

Nessa comparação, é possível verificar que a quantidade de arcos verdadeiros positivos (VP) é igual a 4: $D \rightarrow A$, $A \rightarrow G$, $E \rightarrow G$ e $B \rightarrow E$. A quantidade de arcos falsos positivos (FP) é igual a 2: $D \rightarrow C$ e $F \rightarrow B$. Por fim, a quantidade de arcos falsos negativos (FN) é igual a 2: $C \rightarrow D$ e $F \rightarrow E$.

Também é possível identificar as diferenças entre as estruturas por meio de comparações visuais, destacando-as com cores distintas, conforme mostrado na Figura 2.10. Os arcos em vermelho representam as relações presentes no DAG 2, mas ausentes no DAG 1, enquanto o arco pontilhado na cor verde representa a relação existente no DAG 1, mas ausente no DAG 2.

Figura 2.10 – Comparação visual das estruturas de redes bayesianas.



Fonte: Do autor (2024).

2.4.8 Inferência de rede bayesiana

As redes bayesianas, assim como outros modelos estatísticos, oferecem uma abordagem para responder a questões que ultrapassam a simples descrições dos dados amostrais. O processo de obtenção dessas respostas com base em novas evidências é comumente referido como inferência. No contexto das redes bayesianas, esse processo é também conhecido como argumento probabilístico ou atualização de crença, onde as hipóteses ou variáveis de interesse são chamadas de "questões". Essa estrutura reflete a flexibilidade das redes bayesianas em lidar tanto com incertezas quanto com o conhecimento parcial sobre o sistema modelado.

Na prática, o argumento probabilístico em redes bayesianas opera dentro do escopo da estatística bayesiana, concentrando-se no cálculo de probabilidades ou densidades posteriores com base nas evidências. Esse processo de inferência é conduzido por meio do Teorema de Bayes, permitindo a atualização das crenças (distribuições a *priori*) à luz de novas evidências (distribuições a *posteriori*). Seja uma rede bayesiana B com estrutura $\mathbb{G}$ e parâmetros Θ, o objetivo é utilizar B para investigar os efeitos de uma nova evidência E, utilizando o conhecimento codificado em B, ou seja, investigar a distribuição a *posteriori* $P(X|E,B) = P(X|E,\mathbb{G},\Theta)$ (NAGARAJAN; SCUTARI; LBRE, 2013).

As abordagens usadas para esse tipo de análise variam dependendo da natureza de E e da natureza da informação de interesse. Os dois tipos de evidência mais comuns são:

a) *Hard evidence*, uma instanciação precisa de uma ou mais variáveis na rede, onde a evidência observada é definitiva, restringindo as variáveis a valores específicos.

$$\mathbf{E} = \{X_{i_1} = e_1, X_{i_2} = e_2, \ldots, X_{i_k} = e_k\} \quad i_1, \ldots i_k \in \{1, \ldots, n\}.$$

b) *Soft evidence*,uma nova distribuição probabilística para uma ou mais variáveis, Ccmo tanto a estrutura da rede quanto as suposições distribucionais são tratadas como fixas, a *Soft evidence* é geralmente especificada como um novo conjunto de parâmetros,

$$\mathbf{E} = \{X_{i_1} \sim (\Theta_{X_{i_1}}), X_{i_2} \sim (\Theta_{X_{i_2}}), \ldots, X_{i_k} \sim (\Theta_{X_{i_k}})\} \quad i_1, \ldots i_k \in \{1, \ldots, n\}.$$

No que diz respeito às "questões"ou hipóteses a serem respondidas, as consultas de Probabilidade Condicional (*Conditional Probability Queries* - CPQ) e de Máxima a Posteriori (*Maximum a Posteriori* - MAP), também conhecidas como consulta de Explicação Mais Provável (*Most Probable Explanation* - MPE), são amplamente utilizadas. Ambas as consultas são geralmente aplicadas no contexto de *hard evidence*, mas também podem ser usadas em conjunto com *soft evidence*, permitindo uma flexibilidade maior no manejo das incertezas e variações nos dados.

Em redes bayesianas discretas, a previsão pode ser tratada como uma consulta MAP, onde se busca o valor mais provável para as variáveis de interesse. No caso de redes bayesianas gaussianas (RBG) ou redes bayesianas com variáveis contínuas (RBGC), o processo de inferência muda, sendo calculado com base nas expectativas posteriores em vez de simplesmente aplicar MAP, refletindo a complexidade adicional da modelagem de variáveis contínuas (SCUTARI; DENIS, 2021).

2.5 Validação cruzada

A validação cruzada é uma técnica crucial na avaliação de métodos de aprendizagem estatística, permitindo estimar o erro de teste associado a um determinado modelo. Esse erro de teste é uma medida crucial do desempenho do modelo e pode ser usado para avaliar sua eficácia em generalizar para dados não vistos. Além disso, a validação cruzada pode ser empregada na seleção do nível apropriado de flexibilidade do modelo.

O processo de avaliação do desempenho de um modelo por meio da validação cruzada é conhecido como avaliação de modelo (JAMES et al., 2013). Ele envolve dividir os dados disponíveis em conjuntos de treinamento e teste repetidamente, ajustando o modelo em diferentes combinações desses conjuntos e calculando o erro médio de teste ao longo dessas iterações

2.5.1 Validação cruzada *k-fold*

Nesta abordagem, as observações são divididas aleatoriamente em k grupos, ou dobras, com tamanhos aproximadamente iguais. O primeiro grupo é usado como um conjunto de validação, enquanto o modelo é ajustado aos $k-1$ grupos restantes. Por exemplo, o erro quadrático médio (EQM) é então calculado para as observações na dobra de validação. Esse processo é repetido k vezes, onde cada vez um grupo diferente de observações é selecionado como conjunto de validação. Assim, são obtidas k estimativas do erro de teste $(EQM_1, EQM_2, \ldots, EQM_k)$ (JAMES et al., 2013). A estimativa de validação cruzada *k-fold* é obtida pela média desses valores,

$$CV_{(k)} = \frac{1}{k}\sum_{i=1}^{k} EQM_i.$$

Para o caso em que a variável resposta é qualitativa, a validação cruzada é realizada da mesma forma descrita, porém altera-se a métrica utilizada para quantificar o erro de teste. Nesse caso pode-se utilizar o número de observações classificadas incorretamente, portanto, na configuração de classificação a taxa de erro de validação cruzada *k-fold* é dada por

$$CV_{(k)} = \frac{1}{k}\sum_{i=1}^{k} I(y_i \neq \hat{y}_i).$$

em que $I(y_i \neq \hat{y}_i)$ é uma variável indicadora que é igual a 1 se $y_i \neq \hat{y}_i$ e 0 se $y_i = \hat{y}_i$.

A validação cruzada é uma abordagem muito geral que pode ser aplicada a quase todos os métodos de aprendizagem estatística. Neste trabalho, a validação cruzada será uma ferramenta essencial na aplicação para avaliar a qualidade de predição de cada modelo considerado.

3 METODOLOGIA

A metodologia do trabalho consistiu em analisar uma aplicação a dados reais disponibilizados por Pereira et al. (2020b) e um estudo de simulação comparando as redes bayesianas a métodos automáticos de seleção de variáveis.

3.1 Especificação para o estudo de simulação

Foi realizado um estudo de simulação para avaliar o desempenho de quatro algoritmos de redes bayesianas (HC: *Hill Climbing*, *Tabu*, MMHC: *Max-Min Hill Climbing* e *RSMAX2*) e dos métodos *backward stepwise*, *forward stepwise* e *bidirectional stepwise* na seleção de variáveis em modelos de regressão. Para o *bidirectional stepwise*, foram considerados dois modelos iniciais: o modelo nulo (*bidirectional F*) e o modelo completo (*bidirectional B*). Foram analisadas três estruturas de associação entre variáveis, diferentes tamanhos amostrais ($n = 50, 150$ e 450), correlações variadas entre variáveis (resposta e covariáveis) e o número de variáveis. O foco foi a semelhança entre o modelo ajustado e o verdadeiro, analisando quantas e quais variáveis foram selecionadas pelos métodos.

Algumas estruturas foram projetadas para cenários que favorecem a seleção automática, como a estrutura 1. Por outro lado, outros cenários foram desenvolvidos com estruturas que podem oferecer vantagens de interpretação com as redes bayesianas, como as estruturas 2 e 3.

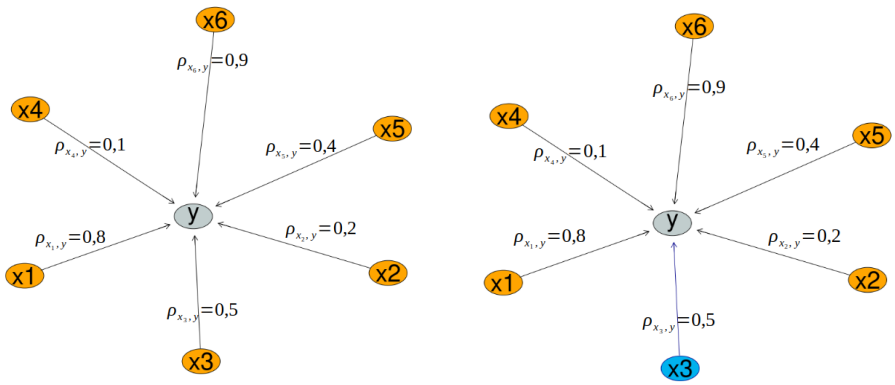
Caso de resposta contínua

Estrutura de Associação 1

a) $x_1, x_2, x_3, x_4, x_5, x_6 \sim N(0,1)$;

b) $y = 1 + 0,8x_1 + 0,2x_2 + 0,5x_4 + 0,1x_4 + 0,4x_5 + 0,9x_6$.

Figura 3.1 – Estrutura de associação 1, considerando o caso sem *whitelist* e com *whitelist* respectivamente.



Fonte: Do autor (2024).

Estrutura de Associação 2

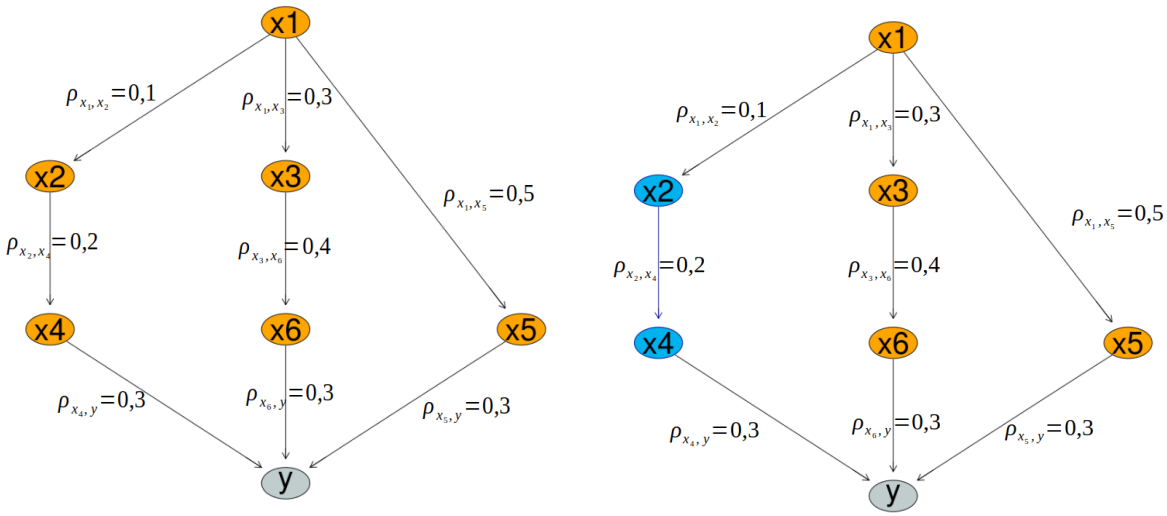
a) $x_1, x_{21}, x_{31}, x_{41}, x_{51}, x_{61} \sim N(0,1)$;

b) $x_2 = x_{21} + 0,1x_1$; $x_3 = x_{31} + 0,3x_1$; $x_5 = x_{51} + 0,5x_1$;

c) $x_4 = x_{41} + 0,2x_2$; $x_6 = x_{61} + 0,4x_3$;

d) $y = 1 + 0,3x_4 + 0,3x_5 + 0,3x_6$.

Figura 3.2 – Estrutura de associação 2, considerando o caso sem *whitelist* e com *whitelist* respectivamente.



Fonte: Do autor (2024).

Estrutura de Associação 3

a) $x_1, x_{21}, x_3, x_{41}, x_5, x_{61}, x_{71}, x_{81}, x_{91}, x_{101} \sim N(0,1)$;

b) $x_2 = x_{21} + 0,2x_3$; $\quad x_4 = x_{41} + 0,8x_3$; $\quad x_6 = x_{61} + 0,8x_1$;

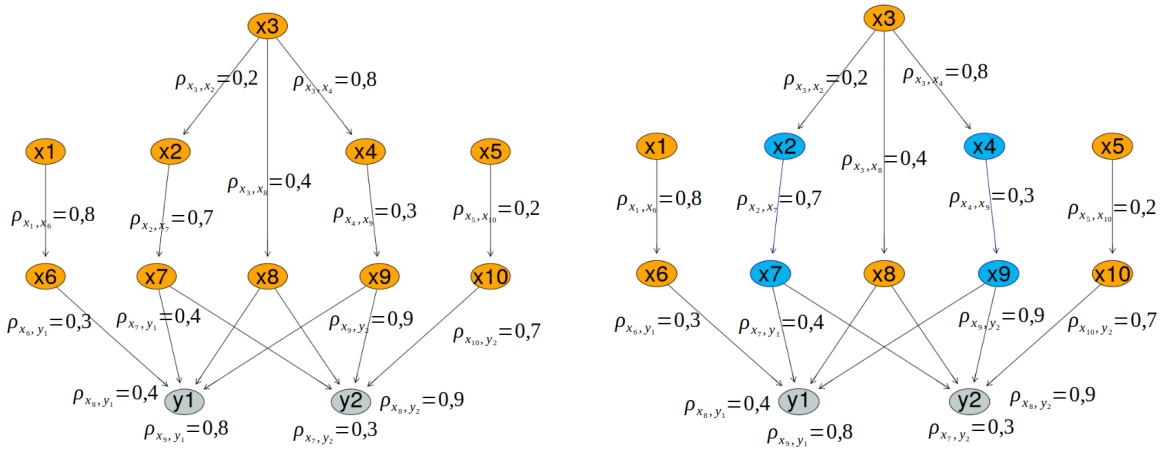
c) $x_7 = x_{71} + 0,7x_2$; $\quad x_8 = x_{81} + 0,4x_3 \quad x_9 = x_{91} + 0,3x_4$;

d) $x_{10} = x_{101} + 0,2x_5$;

e) $y_1 = 1 + 0,3x_6 + 0,4x_7 + 0,4x_8 + 0,8x_9$;

f) $y_2 = 1 + 0,3x_7 + 0,9x_8 + 0,9x_9 + 0,7x_{10}$.

Figura 3.3 – Estrutura de associação 3, considerando o caso sem *whitelist* e com *whitelist* respectivamente.



Fonte: Do autor (2024).

Caso de resposta binária

Considerando os mesmos cenários descritos acima, as variáveis foram discretizadas considerando as medidas de posição relativa.

A variável resposta foi obtida considerando valor da função de densidade cumulativa (cdf) da variável resposta y. Com a probabilidade acumulada de cada observação de y, cada um representando o resultado de uma tentativa de Bernoulli com probabilidade de sucesso de $F(y)$. O resultado da tentativa será 0 ou 1, em que, 0 representa um fracasso e 1 representa um sucesso. A categorização das variáveis explicativas, ocorreu da seguinte maneira:

a) **variáveis explicativas com duas categorias (VE2)**:

- categoria 1 - valores menores que a mediana;
- categoria 2 - valores maiores que a mediana;

b) **variáveis explicativas com três categorias (VE3)**:

- categoria 1 - valores menores que o 1º tercil;

- categoria 2 - valores entre o 1º tercil e o 2º tercil;
- categoria 3 - valores maiores que o 2º tercil;

c) **variáveis explicativas com quatro categorias (VE4)**:

- categoria 1 - valores menores que o 1º quartil;
- categoria 2 - valores entre o 1º quartil e o 2º quartil;
- categoria 3 - valores entre o 2º quartil e o 3º quartil;
- categoria 3 - valores maiores que o 3º quartil.

As simulações foram conduzidas utilizando o pacote *bnlearn* (SCUTARI, 2010) e o software R (R Core Team, 2020).

3.2 Base de dados para a aplicação

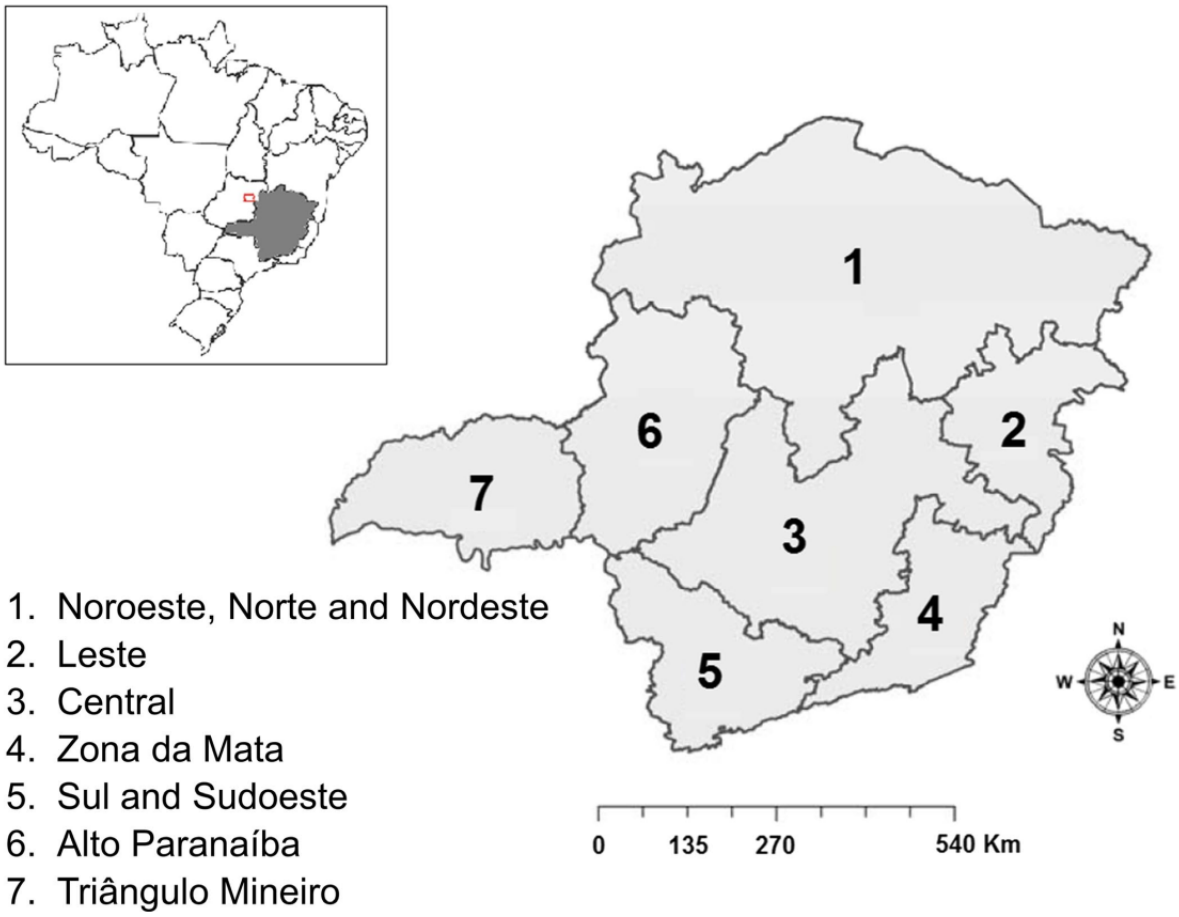
A brucelose é uma importante doença ocupacional, principalmente entre médicos veterinários, devido ao seu contato frequente com animais doentes, secreções contaminadas e vacinas anti-Brucella. Em particular, a exposição acidental às vacinas anti-Brucella representa um risco significativo, uma vez que essas vacinas contêm cepas vivas atenuadas da bactéria. Pereira et al. (2020b) realizaram um estudo com objetivo de determinar a prevalência de exposição às vacinas S19 e RB51 entre veterinários, e identificar os fatores de risco mais importantes associados à exposição acidental às vacinas anti-*Brucella abortus*. Os dados foram coletados por meio de questionário online, em que trezentos e vinte e nove veterinários foram incluídos nas análises, usando amostragem aleatória estratificada.

O questionário iniciou-se com perguntas gerais como idade, experiência profissional, área de profissionalização, percepção de doenças, práticas de controle de infecções e procedimentos de risco (administração de vacinas e cuidados veterinários relacionados à reprodução bovina). Em seguida, foram feitas perguntas específicas relacionadas ao contato acidental desprotegido com S19 e RB51 e infecção ocupacional por *B. abortus*.

Os indivíduos que relataram exposição não intencional a vacinas vivas atenuadas anti-*B. abortus* ou brucelose foram questionados sobre as prováveis causas do desfecho, tipo de exposição às cepas S19 e RB51, medidas de profilaxia adotadas, métodos diagnósticos utilizados, ocorrência e duração dos sintomas, tratamentos implementados e possíveis recaídas da doença.

Esse estudo transversal foi realizado no período de novembro de 2018 à maio de 2019 no estado de Minas Gerais. O estado foi dividido em sete regiões (estratos) de produção bovina, e está representado na Figura 3.4.

Figura 3.4 – Mapa do estado de Minas Gerais, mostrando as regiões (estratos) definidas no estudo. O estado foi dividido em sete regiões: 1. Noroeste, Norte e Nordeste; 2. Leste; 3. Centrais; 4. Zona da Mata; 5. Sul e Sudoeste; 6. Alto Paranaíba; e 7. Triângulo Mineiro



Fonte: Pereira et al. (2020b)

Os autores construíram o modelo utilizando o método *purposeful selection of variables* para a regressão logística segundo Hosmer e Lemeshow (2000). As variáveis "escore de conhecimento sobre os sintomas da brucelose humana"e "escore de uso de EPI durante as atividades laborais"foram significativamente associadas à exposição acidental a S19 e RB51 e foram incluídas no modelo final. O valor do *Receiver Operating Characteristic* (ROC) foi de $0,62$. O Quadro 3.1 apresenta as descrições das variáveis utilizadas nas redes bayesianas.

Quadro 3.1 – Descrição das variáveis utilizadas na RB.

Variável	Descrição da variável
regioes	Região em que o profissional reside; (Alto Paranaíba - 36 (10,94%); Central - 99 (30,09%); Leste - 18 (5,47%); Noroeste, Norte e Noedeste - 32 (9,73%); Sul e Sudoeste - 72 (21,88%); Triângulo Mineiro - 35 (10,64%); Zona da Mata - 37 (11,25%));
idade	Idade (anos) do profissional;
sexo	Masculino - 273 (82.98%); Feminino - 56 (17,02%);
experiencia	Experiência profissional (anos);
cadastro	Cadastro no PNCEBT (anos);
habilitado	Está habilitado no PNCEBT para realizar diagnóstico de brucelose; Sim - 144 (43,77%); Não - 185 (56,23%);
especie	Principal campo de trabalho; Gado leitero - 214 (65,05%); Gado de corte - 59 (17,93%); Outros - 56 (17,02%)
vinculo	Vículo empregatício; Autônomo - 235 (71.43%); Empresa privada - 67 (20,36%); Servidor público - 27 (8,21%);
parto	Realizou procedimento de partos nos últimos seis meses; Sim - 240 (72,95%); Não - 89 (27,05%);
placenta	Realizou procedimento de remoção de placenta nos últimos seis meses; Sim - 180 (54,71%); Não - 149 (45,29%);
vacinacao	Realizou procedimento de vacinação contra brucelose nos últimos seis meses; Sim - 309 (93,92%); Não - 20 (6,08%);
aborto	Realizou procedimento de parto prematuro ou aborto nos últimos seis meses; Sim - 178 (54,10%); Não - 151 (45,90%);
EPI	Uso de equipamentos de proteção individual (EPI); (uso de luvas, jaleco, óculos de proteção e máscaras) Pontuação entre 0 (nunca utilizou nenhum EPI) e 22 (sempre utilizou todos os EPI recomendados);
transmissao	Conhecimento sobre a transmissão da brucelose; Bom - 156 (47,42%); Médio - 171 (51,98%); Ruim - 2 (0,61%);
sintomas	Conhecimento sobre os sintomas da brucelose; Bom - 275 (83,59%); Médio - 37 (11,25%); Ruim - 17 (5,17%);
leite	Consumiu leite ou derivados sem tratamento térmico nos últimos 6 meses; Sim - 223 (67,78%); Não - 106 (32,22%);
vacinadores	Existe algum vacinador cadastrado sob sua responsabilidade; Sim - 243 (73,86%); Não - 86 (26,14%);
exposicao **(Variável Resposta)**	Foi exposto acidentalmente às vacinas $B19$ ou $RB51$; Sim - 108 (32,83%); Não - 221 (67,17%);

Fonte: Do autor (2024).

Segundo os autores, esperava-se que certas variáveis fossem incluídas no modelo final, mas isso não ocorreu. Especificamente, as variáveis habilitado, parto, placenta, vacinação,

aborto e espécie foram inicialmente consideradas importantes e relevantes para o estudo. No entanto, durante o processo de seleção, essas variáveis não foram incluídas no modelo final.

4 ESTUDO DE SIMULAÇÃO

Os resultados são apresentados em Tabelas que indicam a taxa de seleção das variáveis esperadas em um total de 5000 repetições, assim como a média dos valores AIC para cada método. Especificamente para os algoritmos RB, uma tabela adicional apresenta a média de verdadeiros positivos, falsos positivos e falsos negativos na identificação das variáveis no modelo. Algumas linhas das tabelas estarão destacadas com cores de fundo específicas: a linha com esta cor de fundo (mais escura) indica que os métodos selecionaram apenas as variáveis esperadas, a linha com esta cor de fundo (intermediária) mostra que os métodos selecionaram as variáveis esperadas e mais alguma outra, e a linha com esta cor de fundo (mais clara) destaca as variáveis que deveriam ser selecionadas.

4.1 O caso contínuo

Faremos uma análise mais aprofundada dos resultados para o cenário com tamanho amostral de $n = 50$. Os resultados para os outros tamanhos amostrais encontram-se no Apêndice A e serão discutidos na Seção 4.3, uma vez que são repetitivos e seguem um padrão semelhante.

4.1.1 Estrutura de associação 1

Considerando a estrutura de associação de variáveis da Figura 3.1, a Tabela 4.1 mostra quantas e quais variáveis foram selecionadas pelos métodos.

Tabela 4.1 – Taxa de seleção das variáveis para cada método, considerando estrutura de associação 1, $n = 50$ e variável resposta contínua.

Acertos	Backward	Forward	Bidirectional F[1]	Bidirectional B[2]	HC	Tabu	MMHC	RSMAX2
Apenas as 6 variáveis	1,00	1,00	1,00	1,00	0,00	0,00	0,00	0,00
Uma das variáveis	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00
Duas das variáveis	0,00	0,00	0,00	0,00	0,00	0,00	0,13	0,13
Três das variáveis	0,00	0,00	0,00	0,00	0,00	0,00	0,55	0,55
Quatro das variáveis	0,00	0,00	0,00	0,00	0,00	0,00	0,32	0,32
Cinco das variáveis	0,00	0,00	0,00	0,00	1,00	1,00	0,01	0,01
x1	1,00	1,00	1,00	1,00	1,00	1,00	0,99	0,99
x2	1,00	1,00	1,00	1,00	1,00	1,00	0,05	0,05
x3	1,00	1,00	1,00	1,00	1,00	1,00	0,70	0,70
x4	1,00	1,00	1,00	1,00	0,00	0,00	0,01	0,01
x5	1,00	1,00	1,00	1,00	1,00	1,00	0,45	0,45
x6	1,00	1,00	1,00	1,00	1,00	1,00	1,00	1,00

[1] Modelo inicial é o modelo nulo.
[2] Modelo inicial é o modelo completo.

Fonte: Do autor (2024).

De acordo com a Tabela 4.1, os métodos *Backward*, *Forward*, *Bidirectional F* e *Bidirectional B*, selecionaram todas as variáveis explicativas corretamente, enquanto os algoritmos *HC* e *Tabu*, selecionaram sempre apenas cinco das seis variáveis esperadas. Os algoritmos *MMHC* e *RSMAX*2 tiveram um desempenho abaixo em relação aos outros métodos, selecionando 55% das vezes, três das variáveis esperadas. A variável menos selecionada pelos algoritmos de redes bayesianas foi x_4 que possuía menor correlação com a variável resposta.

A Tabela 4.2 representa quantas e quais variáveis foram selecionadas pelos algoritmos de rede bayesiana, adicionando a informação de *whitelist* $x_3 \rightarrow y$.

Tabela 4.2 – Taxa de seleção das variáveis para cada algoritmo de rede bayesiana considerando *whitelist*, para estrutura de associação 1, $n = 50$ e variável resposta contínua.

Acertos	HC	Tabu	MMHC	RSMAX2
Apenas as 6 variáveis	0,00	0,00	0,00	0,00
Uma das variáveis	0,00	0,00	0,00	0,00
Duas das variáveis	0,00	0,00	0,00	0,00
Três das variáveis	0,00	0,00	0,52	0,52
Quatro das variáveis	0,00	0,00	0,46	0,46
Cinco das variáveis	1,00	1,00	0,01	0,01
x1	1,00	1,00	0,99	0,99
x2	1,00	1,00	0,04	0,04
x3	1,00	1,00	1,00	1,00
x4	0,00	0,00	0,01	0,01
x5	1,00	1,00	0,44	0,44
x6	1,00	1,00	1,00	1,00

Fonte: Do autor (2024).

Com a inclusão da *whitelist*, apenas os algoritmos *MMHC* e *RSMAX2* apresentaram uma melhora nos resultados, porém ainda mantiveram um desempenho inferior aos outros métodos na seleção de variáveis.

A Tabela 4.3 apresenta a média das relações corretamente identificadas (VP), a média das relações incorretamente identificadas (FP) e a média das relações não identificadas (FN) pelos algoritmos de RB. Observou-se que os algoritmos *HC* e *Tabu* tiveram um bom desempenho na identificação das relações existentes, porém foram os que mais indicaram relações inexistentes entre a estrutura de associação esperada com a estrutura de associação obtida. Por outro lado, os algoritmos *MMHC* e *RSMAX2* apresentaram um bom desempenho na média de FP, mas deixaram de identificar algumas das relações existentes. Apenas *MMHC* e *RSMAX2* apresentaram melhorias nos resultados com a inclusão de informações adicionais (*whitelist*).

Tabela 4.3 – Média de verdadeiro positivo (VP), falso positivo (FP) e falso negativo (FN) das relações dos algoritmos de RB, para estrutura de associação 1, $n = 50$ e variável resposta contínua.

Algoritmos	VP	FP	FN
HC	5,00	2,61	1,00
Tabu	5,00	2,71	1,00
MMHC	3,20	0,43	2,80
RSMAX2	3,20	0,43	2,80
HC (com wl)	5,00	2,61	1,00
Tabu (com wl)	5,00	2,71	1,00
MMHC (com wl)	3,49	0,41	2,51
RSMAX2 (com wl	3,49	0,43	2,51

Fonte: Do autor (2024).

Tabela 4.4 – Média dos valores AIC para cada método, considerando estrutura de associação 1, $n = 50$ e variável resposta contínua.

Método	AIC	Método	AIC
Backward	-3362,96	MMHC	52,34
Forward	-3368,52	RSMAX2	52,43
Bidirectional F	-3368,52	HC (com wl)	-81,87
Bidirectional B	-3362,96	Tabu (com wl)	-81,87
HC	-81,87	MMHC (com wl)	34,79
Tabu	-81,87	RSMAX2 (com wl	34,88

Fonte: Do autor (2024).

A Tabela 4.4 apresenta a média dos valores AIC para diferentes métodos, considerando uma estrutura de associação 1, com $n = 50$ e uma variável resposta contínua. Os menores valores de AIC foram obtidos pelos métodos *Forward* e *Bidirectional F* $(AIC = -3368,52)$, indicando que ambos proporcionaram o melhor ajuste para o modelo avaliado, sendo, portanto, os mais eficientes entre os métodos analisados.

4.1.2 Estrutura de associação 2

Considerando a estrutura de associação de variáveis representada na Figura 3.2, a Tabela 4.5 indica quantas e quais variáveis foram selecionadas pelos diferentes métodos.

Os métodos *Backward*, *Forward*, *Bidirectional F* e *Bidirectional B* sempre identificaram as variáveis de interesse (x_4, x_5 e x_6), enquanto os métodos *Forward* e *Bidirectional F*, em 54% das vezes, selecionaram apenas essas variáveis, demonstrando o melhor desempenho. Por outro lado, os algoritmos *HC*, *Tabu*, *MMHC* e *RSMAX2* selecionaram consistentemente apenas

duas das três variáveis esperadas. Notavelmente, *MMHC* e *RSMAX2* raramente selecionaram as variáveis x_1, x_2 e x_3, em contraste com os outros métodos.

Tabela 4.5 – Taxa de seleção das variáveis para cada método, considerando estrutura de associação 2, $n = 50$ e variável resposta contínua.

Acertos	Backward	Forward	Bidirectional F	Bidirectional B	HC	Tabu	MMHC	RSMAX2
Apenas as 3 variáveis	0,08	0,54	0,54	0,05	0,00	0,00	0,00	0,00
Uma das variáveis	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00
Duas das variáveis	0,00	0,00	0,00	0,00	1,00	1,00	1,00	1,00
Três das variáveis	1,00	1,00	1,00	1,00	0,00	0,00	0,00	0,00
x1	0,58	0,19	0,19	0,65	0,31	0,31	0,01	0,01
x2	0,59	0,19	0,19	0,67	0,33	0,34	0,00	0,00
x3	0,59	0,18	0,18	0,67	0,35	0,35	0,01	0,01
x4	1,00	1,00	1,00	1,00	0,45	0,45	0,56	0,56
x5	1,00	1,00	1,00	1,00	0,82	0,82	0,73	0,73
x6	1,00	1,00	1,00	1,00	0,73	0,73	0,71	0,71

Fonte: Do autor (2024).

Após a adição da informação da *whitelist*, não foi observada nenhuma melhoria na taxa de seleção das variáveis, como indicado na Tabela 4.6.

Tabela 4.6 – Taxa de seleção das variáveis para cada algoritmo de rede bayesiana considerando *whitelist*, para estrutura de associação 2, $n = 50$ e variável resposta contínua.

Acertos	HC	Tabu	MMHC	RSMAX2
Apenas as 3 variáveis	0,00	0,00	0,00	0,00
Uma das variáveis	0,00	0,00	0,01	0,01
Duas das variáveis	1,00	1,00	0,99	0,99
Três das variáveis	0,00	0,00	0,00	0,00
x1	0,31	0,31	0,01	0,01
x2	0,33	0,34	0,00	0,00
x3	0,35	0,35	0,01	0,01
x4	0,45	0,45	0,56	0,56
x5	0,82	0,82	0,73	0,73
x6	0,73	0,73	0,71	0,71

Fonte: Do autor (2024).

O desempenho dos algoritmos, manteve-se de acordo com o que ocorreu na estrutura de correlação anterior, os métodos *HC* e *Tabu* tiveram melhores taxas em identicar as relações existentes, mas foram os algoritmos que mais indicaram relações que não existiam na estrutura de associação entre variáveis verdadeiras.

Tabela 4.7 – Média de verdadeiro positivo (VP), falso positivo (FP) e falso negativo (FN) das relações dos algoritmos de RB, para estrutura de associação 2, $n = 50$ e variável resposta contínua.

Algoritmos	VP	FP	FN
HC	4,85	3,20	3,15
Tabu	4,12	4,00	3,88
MMHC	3,53	0,23	4,47
RSMAX2	3,53	0,23	4,47
HC (com wl)	5,42	3,11	2,58
Tabu (com wl)	4,75	3,84	3,25
MMHC (com wl)	4,38	0,21	3,62
RSMAX2 (com wl	4,39	0,21	3,61

Fonte: Do autor (2024).

Os algoritmos *MMHC* e *RSMAX2* tiveram um bom desempenho na média de FP, mas deixaram de identificar algumas das relações existentes. Adicionando a informação de *whitelist* todos os algoritmos tiveram uma melhora na taxa de *VP*.

Tabela 4.8 – Média dos valores AIC para cada método, considerando estrutura de associação 2, $n = 50$ e variável resposta contínua

Método	AIC	Método	AIC
Backward	-3452,08	MMHC	28,33
Forward	-3438,48	RSMAX2	28,24
Bidirectional F	-3438,50	HC (com wl)	19,06
Bidirectional B	-3452,44	Tabu (com wl)	19,06
HC	19,06	MMHC (com wl)	28,36
Tabu	19,06	RSMAX2 (com wl	28,27

Fonte: Do autor (2024).

A Tabela 4.8 apresenta a média dos valores AIC para cada método, considerando uma estrutura de associação 2, com $n = 50$ e uma variável resposta contínua. Os métodos *Backward* e *Bidirectional B* obtiveram os melhores resultados, com os menores valores de AIC ($-3452,08$ e $-3452,44$ respectivamente), indicando que esses métodos foram os mais eficazes na seleção de modelos.

4.1.3 Estrutura de associação 3

Considerando a estrutura de associação de variáveis da Figura 3.3, havia duas variáveis resposta (y_1 e y_2), e o objetivo era identificar quantas e quais variáveis seriam selecionadas para cada uma das variáveis resposta.

Variável resposta y_1

A Tabela 4.9 apresenta a quantidade e quais variáveis foram selecionadas pelos métodos. Os métodos *Backward*, *Forward*, *Bidirectional F* e *Bidirectional B* sempre escolheram as variáveis de interesse (x_6,x_7, x_8 e x_9), enquanto os métodos *Forward* e *Bidirectional F* selecionaram apenas as variáveis de interesse em 31% das vezes, demonstrando o melhor desempenho. Por outro lado, os algoritmos *HC* e *Tabu* sempre optaram por apenas três das quatro variáveis esperadas. Quanto aos métodos *MMHC* e *RSMAX2*, houve uma tendência a não selecionar as variáveis x_1, x_2, x_3, x_4, x_5 e x_{10}, embora tenham escolhido apenas duas das quatro variáveis esperadas na maioria das vezes.

Tabela 4.9 – Taxa de seleção das variáveis para cada método, considerando estrutura de associação 3, $n = 50$ e variável resposta contínua y_1.

Acertos	Backward	Forward	Bidirectional F	Bidirectional B	HC	Tabu	MMHC	RSMAX2
Apenas as 4 variáveis	0,01	0,31	0,31	0,00	0,00	0,00	0,00	0,00
Uma das variáveis	0,00	0,00	0,00	0,00	0,00	0,00	0,23	0,20
Duas das variáveis	0,00	0,00	0,00	0,00	0,00	0,00	0,54	0,61
Três das variáveis	0,00	0,00	0,00	0,00	1,00	1,00	0,22	0,18
Quatro das variáveis	1,00	1,00	1,00	1,00	0,00	0,00	0,00	0,00
x1	0,67	0,19	0,19	0,73	0,78	0,78	0,01	0,02
x2	0,68	0,19	0,19	0,75	0,21	0,21	0,01	0,04
x3	0,68	0,18	0,18	0,75	0,29	0,30	0,01	0,02
x4	0,69	0,17	0,17	0,75	0,17	0,18	0,01	0,03
x5	0,67	0,20	0,20	0,74	0,17	0,17	0,00	0,00
x6	1,00	1,00	1,00	1,00	0,27	0,27	0,38	0,25
x7	1,00	1,00	1,00	1,00	0,96	0,96	0,62	0,62
x8	1,00	1,00	1,00	1,00	0,78	0,78	0,00	0,11
x9	1,00	1,00	1,00	1,00	1,00	1,00	0,98	0,99
x10	0,25	0,19	0,19	0,26	0,19	0,19	0,01	0,00

Fonte: Do autor (2024).

Conforme a Tabela 4.10 e a Tabela 4.11, novamente o padrão de desempenho manteve-se, e os métodos *HC* e *Tabu* alcançaram melhores taxas para identificar as relações existentes, mas indicaram relações que não existiam na estrutura de associação entre variáveis verdadeiras com maior frequência.

Os algoritmos *MMHC* e *RSMAX2* tiveram um bom desempenho na média de FP, porém deixaram de encontrar algumas das relações existentes. Incrementando a informação de *whitelist* ($x_2 \rightarrow x_7$, $x_4 \rightarrow x_9$) todos os algoritmos tiveram uma melhora na taxa de *VP*.

Tabela 4.10 – Taxa de seleção das variáveis para cada algoritmo de rede bayesiana considerando *whitelist*, para estrutura de associação 3, $n = 50$ e variável resposta contínua y_1.

Acertos	HC	Tabu	MMHC	RSMAX2
Apenas as 4 variáveis	0,00	0,00	0,00	0,00
Uma das variáveis	0,00	0,00	0,23	0,21
Duas das variáveis	0,00	0,00	0,54	0,61
Três das variáveis	1,00	1,00	0,22	0,18
Quatro das variáveis	0,00	0,00	0,00	0,00
x1	0,78	0,78	0,01	0,02
x2	0,21	0,21	0,01	0,04
x3	0,29	0,30	0,01	0,02
x4	0,17	0,18	0,00	0,01
x5	0,17	0,17	0,00	0,00
x6	0,27	0,27	0,38	0,25
x7	0,96	0,96	0,62	0,62
x8	0,78	0,78	0,00	0,11
x9	1,00	1,00	0,98	0,99
x10	0,19	0,19	0,01	0,00

Fonte: Do autor (2024).

Tabela 4.11 – Média de verdadeiro positivo (VP), falso positivo (FP) e falso negativo (FN) das relações dos algoritmos de RB, para estrutura de associação 3, $n = 50$ e variável resposta contínua.

Algoritmos	VP	FP	FN
HC	10,46	11,21	4,54
Tabu	9,77	12,29	5,23
MMHC	7,25	0,60	7,75
RSMAX2	7,21	0,65	7,79
HC (com wl)	10,93	10,98	4,07
Tabu (com wl)	10,45	11,84	4,55
MMHC (com wl)	8,19	0,52	6,81
RSMAX2 (com wl)	8,14	0,57	6,86

Fonte: Do autor (2024).

A Tabela 4.12 apresenta a média dos valores AIC para cada método, considerando uma estrutura de associação 3, com $n = 50$ e uma variável resposta contínua $y1$. Os métodos *Backward* e *Bidirectional B* destacaram-se com os menores valores de AIC ($-3404,23$ e $-3404,77$ respectivamente), indicando que esses métodos foram os mais eficazes na seleção de modelos para esta configuração.

Tabela 4.12 – Média dos valores AIC para cada método, para estrutura de associação 3, $n = 50$ e variável resposta contínua y_1

Método	AIC	Método	AIC
Backward	-3404,23	MMHC	90,35
Forward	-3390,93	RSMAX2	89,81
Bidirectional F	-3390,98	HC (com wl)	29,10
Bidirectional B	-3404,77	Tabu (com wl)	29,08
HC	29,10	MMHC (com wl)	90,69
Tabu	29,09	RSMAX2 (com wl)	89,98

Fonte: Do autor (2024).

Variável resposta y_2

Os métodos *Backward*, *Forward*, *Bidirectional F* e *Bidirectional B* sempre escolheram as variáveis de interesse (x_6,x_7, x_8 e x_9). No entanto, os métodos *Forward* e *Bidirectional F* selecionaram apenas as variáveis de interesse em 33% das vezes, demonstrando um melhor desempenho. Em contraste, os algoritmos *HC* e *Tabu* sempre optaram por apenas três das quatro variáveis esperadas. Quanto aos métodos *MMHC* e *RSMAX2*, houve uma tendência a não selecionar as variáveis x_1, x_2, x_3, x_4, x_5 e x_6, embora tenham escolhido apenas duas das quatro variáveis esperadas na maioria das vezes. É importante observar que a variável x_7 não foi selecionada em nenhuma ocasião pelas redes bayesianas.

Tabela 4.13 – Taxa de seleção das variáveis para cada método, considerando estrutura de associação 3, $n = 50$ e variável resposta contínua y_2.

Acertos	Backward	Forward	Bidirectional F	Bidirectional B	HC	Tabu	MMHC	RSMAX2
Apenas as 4 variáveis	0,00	0,33	0,33	0,00	0,00	0,00	0,00	0,00
Uma das variáveis	0,00	0,00	0,00	0,00	0,00	0,00	0,14	0,19
Duas das variáveis	0,00	0,00	0,00	0,00	0,00	0,00	0,83	0,75
Três das variáveis	0,00	0,00	0,00	0,00	1,00	1,00	0,02	0,05
Quatro das variáveis	1,00	1,00	1,00	1,00	0,00	0,00	0,00	0,00
x1	0,70	0,18	0,18	0,76	0,16	0,17	0,00	0,00
x2	0,72	0,20	0,20	0,78	1,00	1,00	0,00	0,02
x3	0,72	0,17	0,17	0,78	0,17	0,18	0,01	0,01
x4	0,73	0,17	0,17	0,79	0,16	0,17	0,01	0,01
x5	0,71	0,19	0,19	0,76	0,17	0,17	0,00	0,00
x6	0,71	0,17	0,17	0,77	0,15	0,16	0,00	0,00
x7	1,00	1,00	1,00	1,00	0,00	0,00	0,00	0,00
x8	1,00	1,00	1,00	1,00	1,00	1,00	0,98	0,96
x9	1,00	1,00	1,00	1,00	1,00	1,00	0,03	0,07
x10	1,00	1,00	1,00	1,00	1,00	1,00	0,87	0,82

Fonte: Do autor (2024).

Tabela 4.14 – Taxa de seleção das variáveis para cada algoritmo de rede bayesiana considerando *whitelist*, para estrutura de associação 3, $n = 50$ e variável resposta contínua y_2.

Acertos	HC	Tabu	MMHC	RSMAX2
Apenas as 4 variáveis	0,00	0,00	0,00	0,00
Uma das variáveis	0,00	0,00	0,14	0,19
Duas das variáveis	0,00	0,00	0,84	0,75
Três das variáveis	1,00	1,00	0,02	0,05
Quatro das variáveis	0,00	0,00	0,00	0,00
x1	0,16	0,17	0,00	0,00
x2	1,00	1,00	0,00	0,02
x3	0,17	0,18	0,01	0,01
x4	0,16	0,17	0,00	0,00
x5	0,17	0,17	0,00	0,00
x6	0,15	0,16	0,00	0,00
x7	0,00	0,00	0,00	0,00
x8	1,00	1,00	0,98	0,96
x9	1,00	1,00	0,03	0,06
x10	1,00	1,00	0,87	0,82

Fonte: Do autor (2024).

Mesmo com a inclusão da informação da *whitelist*, as taxas de seleção dos algoritmos permaneceram inalteradas, e tanto *HC* quanto *Tabu* continuaram a escolher a variável x_2 em vez da variável x_7.

A Tabela 4.15 apresenta a média dos valores AIC obtidos para diferentes métodos, considerando uma estrutura de associação 3, com $n = 50$ e uma variável resposta contínua $y2$. Os métodos *Backward* e *Bidirectional B* alcançaram as melhores pontuações com os menores valores de AIC ($-3375,88$ e $-3404,77$ respectivamente), indicando que esses métodos foram os mais eficazes na seleção do modelo para esta configuração.

Entre os algoritmos de redes bayesianas, os métodos *Hill Climbing* (HC) e *Tabu Search* apresentaram as melhores pontuações, com valores de AIC relativamente baixos ($25,41$ e $25,43$ respectivamente).

Tabela 4.15 – Média dos valores AIC para cada método, considerando estrutura de associação 3, $n = 50$ e variável resposta contínua y_2

Método	AIC	Método	AIC
Backward	-3375,88	MMHC	148,31
Forward	-3362,80	RSMAX2	146,67
Bidirectional F	-3390,98	HC (com wl)	25,43
Bidirectional B	-3404,77	Tabu (com wl)	25,41
HC	25,43	MMHC (com wl)	148,66
Tabu	25,41	RSMAX2 (com wl)	147,70

Fonte: Do autor (2024).

4.2 O caso de resposta binária

Para o caso de resposta binária, serão apresentados detalhadamente apenas os resultados referentes à estrutura de associação 2, com variáveis explicativas de quatro categorias. Os resultados para os demais cenários encontram-se no Apêndice A e serão discutidos na Seção 4.3, uma vez que seguem um padrão semelhante e, portanto, são repetitivos.

4.2.1 Estrutura de associação 2

Considerando as variáveis explicativas com quatro categorias, os métodos *Backward*, *Forward*, *Bidirectional F* e *Bidirectional B*, a taxa de seleção foi de 30%, enquanto o algoritmo de RB *Tabu* obteve uma taxa de 23%, sendo este o melhor resultado entre os algoritmos de RB.

Tabela 4.16 – Taxa de seleção das variáveis para cada método, considerando estrutura de associação 2, $n = 50$, VE4 e variável resposta binária.

Acertos	Backward	Forward	Bidirectional F	Bidirectional B	HC	Tabu	MMHC	RSMAX2
Apenas as 3 variáveis	0,30	0,30	0,30	0,30	0,22	0,23	0,07	0,07
Uma das variáveis	0,08	0,09	0,09	0,08	0,26	0,25	0,45	0,45
Duas das variáveis	0,35	0,36	0,36	0,35	0,44	0,44	0,40	0,40
Três das variáveis	0,56	0,55	0,54	0,56	0,25	0,27	0,07	0,07
x1	0,23	0,23	0,22	0,23	0,11	0,10	0,05	0,06
x2	0,21	0,20	0,20	0,21	0,07	0,07	0,03	0,03
x3	0,22	0,21	0,21	0,22	0,08	0,09	0,04	0,04
x4	0,82	0,81	0,81	0,82	0,60	0,62	0,45	0,45
x5	0,83	0,82	0,82	0,83	0,65	0,67	0,49	0,49
x6	0,83	0,82	0,82	0,83	0,64	0,66	0,51	0,51

Fonte: Do autor (2024).

Tabela 4.17 – Taxa de seleção das variáveis para cada algoritmo de rede bayesiana considerando *whitelist*, para estrutura de associação 2, $n = 50$, VE4 e variável resposta binária.

Acertos	HC	Tabu	MMHC	RSMAX2
Apenas as 3 variáveis	0,22	0,24	0,06	0,06
Uma das variáveis	0,26	0,25	0,45	0,45
Duas das variáveis	0,44	0,44	0,40	0,40
Três das variáveis	0,25	0,27	0,06	0,06
x1	0,11	0,10	0,05	0,06
x2	0,07	0,07	0,03	0,03
x3	0,08	0,09	0,04	0,04
x4	0,60	0,62	0,44	0,44
x5	0,65	0,67	0,49	0,49
x6	0,64	0,66	0,51	0,51

Fonte: Do autor (2024).

Adicionando a informação *whitelist* ($x_2 \rightarrow x_4$), assim como nos outros cenários, a taxa de acerto das relações nas estruturas de variáveis em todos os algoritmos aumentou. No entanto, não foi observado impacto nas relações que influenciam diretamente a variável resposta y.

Tabela 4.18 – Média de verdadeiro positivo(VP), falso positivo(FP) e falso negativo(FN) das relações dos algoritmos de RB, considerando estrutura de associação 2, $n = 50$, VE4 e variável resposta binária.

Algoritmos	VP	FP	FN
HC	3,62	1,04	4,38
Tabu	3,41	1,43	4,59
MMHC	2,90	0,41	5,10
RSMAX2	2,90	0,42	5,10
HC (com wl)	4,39	1,04	3,61
Tabu (com wl)	4,21	1,38	3,79
MMHC (com wl)	3,69	0,40	4,31
RSMAX2 (com wl)	3,70	0,40	4,30

Fonte: Do autor (2024).

Os algoritmos *HC (com wl)* e *Tabu (com wl)* apresentaram os maiores valores de VP, indicando que eles foram mais eficazes na identificação de relações verdadeiras. O método *HC (com wl)* teve o melhor desempenho com um VP médio de $4,39$.

O *MMHC* e o *RSMAX2*, tanto com quanto sem *whitelist*, apresentaram os menores valores de FP ($0,40$ a $0,42$), indicando que esses métodos foram mais precisos e geraram menos falsos positivos. O *Tabu* e o *HC*, tanto com quanto sem *whitelist*, tiveram valores de FP relativamente mais altos, especialmente o *Tabu*.

Tabela 4.19 – Média dos valores AIC para cada método, considerando estrutura de associação 2, $n = 50$, VE4 e variável resposta binária.

Método	AIC	Método	AIC
Backward	55,25	MMHC	63,20
Forward	55,28	RSMAX2	63,20
Bidirectional F	55,26	HC (com wl)	60,84
Bidirectional B	55,25	Tabu (com wl)	60,68
HC	60,84	MMHC (com wl)	63,30
Tabu	60,69	RSMAX2 (com wl)	63,29

Fonte: Do autor (2024).

A Tabela 4.19 mostra a média dos valores AIC para os diferentes métodos aplicados. Os métodos *Backward*, *Forward*, *Bidirectional F* e *Bidirectional B* apresentaram as melhores pontuações, com valores de AIC mais baixos, indicando um melhor ajuste do modelo em rela-

ção ao número de variáveis. Entre os algoritmos de redes bayesianas, *HC* e *Tabu* também se destacaram, embora com valores de AIC um pouco mais altos.

Com o aumento do tamanho amostral para $n = 150$, todos os métodos obtiveram melhores taxas de seleção. Diferentemente do caso com $n = 50$, os algoritmos de RB se destacaram em relação aos métodos *stepwise*. Especificamente, os métodos *HC* e *Tabu* alcançaram a maior porcentagem de acertos, com 85%, seguidos pelos métodos *MMHC* e *RSMAX2*, ambos com 80%.

Tabela 4.20 – Taxa de seleção das variáveis para cada método, considerando estrutura de associação 2, $n = 150$, VE4 e variável resposta binária.

Acertos	Backward	Forward	Bidirectional F	Bidirectional B	HC	Tabu	MMHC	RSMAX2
Apenas as 3 variáveis	0,57	0,57	0,57	0,57	0,85	0,85	0,80	0,80
Uma das variáveis	0,00	0,00	0,00	0,00	0,00	0,00	0,01	0,01
Duas das variáveis	0,01	0,01	0,01	0,01	0,08	0,08	0,16	0,16
Três das variáveis	0,99	0,99	0,99	0,99	0,92	0,92	0,83	0,83
x1	0,18	0,18	0,18	0,18	0,04	0,04	0,03	0,03
x2	0,16	0,15	0,15	0,16	0,03	0,03	0,01	0,01
x3	0,16	0,16	0,16	0,16	0,03	0,03	0,02	0,02
x4	0,99	0,99	0,99	0,99	0,96	0,96	0,93	0,93
x5	1,00	1,00	1,00	1,00	0,98	0,98	0,95	0,95
x6	1,00	1,00	1,00	1,00	0,98	0,98	0,95	0,95

Fonte: Do autor (2024).

Adicionando a informação *whitelist* ($x_2 \rightarrow x_4$), assim como nos outros cenários, a taxa de acerto das relações nas estruturas de variáveis aumentou em todos os algoritmos. Destaca-se o algoritmo *HC (com wl)* que apresentou a maior média de verdadeiros positivos, com um valor de $6,00$, seguido pelo *MMHC (com wl)* com $5,88$. Esses valores indicam que, ao adicionar a informação da *whitelist*, esses algoritmos foram mais eficazes em identificar corretamente as relações existentes.

Em termos de falsos positivos, que representam a quantidade de relações incorretamente identificadas como verdadeiras, o *MMHC* e o *RSMAX2* tiveram as menores taxas, com valores de $0,28$ e $0,29$, respectivamente. Isso sugere que esses métodos são mais precisos em evitar a identificação incorreta de relações.

Tabela 4.21 – Taxa de seleção das variáveis para cada algoritmo de rede bayesiana considerando *whitelist* considerando estrutura de associação 2, $n = 150$, VE4 e variável resposta binária.

Acertos	HC	Tabu	MMHC	RSMAX2
Apenas as 3 variáveis	0,85	0,85	0,80	0,79
Uma das variáveis	0,00	0,00	0,01	0,01
Duas das variáveis	0,08	0,08	0,17	0,17
Três das variáveis	0,92	0,92	0,83	0,83
x1	0,04	0,04	0,03	0,03
x2	0,03	0,03	0,01	0,01
x3	0,03	0,03	0,02	0,02
x4	0,96	0,96	0,92	0,92
x5	0,98	0,98	0,95	0,95
x6	0,98	0,98	0,95	0,95

Fonte: Do autor (2024).

Tabela 4.22 – Média de verdadeiro positivo(VP), falso positivo(FP) e falso negativo(FN) das relações dos algoritmos de RB, considerando estrutura de associação 2, $n = 150$, VE4 e variável resposta binária.

Algoritmos	VP	FP	FN
HC	5,42	0,66	2,58
Tabu	5,15	1,01	2,85
MMHC	5,27	0,28	2,73
RSMAX2	5,27	0,29	2,73
HC (com wl)	6,00	0,63	2,00
Tabu (com wl)	5,74	0,96	2,26
MMHC (com wl)	5,88	0,26	2,12
RSMAX2 (com wl)	5,87	0,27	2,13

Fonte: Do autor (2024).

Conforme a Tabela 4.23 Os métodos *Backward*, *Forward*, *Bidirectional F* e *Bidirectional B* obtiveram as melhores pontuações AIC, com um valor médio uniforme de $161,78$. Esses resultados indicam que esses métodos são altamente eficazes na modelagem do problema considerado, proporcionando a melhor adequação do modelo aos dados.

Entre os algoritmos de redes bayesianas, *HC (com wl)* e *Tabu (com wl)* apresentaram os melhores resultados, com valores AIC de $172,77$ e $172,76$, respectivamente. Estes métodos se destacaram por oferecer uma performance superior em comparação com outros algoritmos de redes bayesianas, que apresentaram valores AIC mais altos, como o *MMHC* e *RSMAX2* (com *whitelist*).

Tabela 4.23 – Média dos valores AIC para cada método, considerando estrutura de associação 2, $n = 150$, VE4 e variável resposta binária.

Método	AIC	Método	AIC
Backward	161,78	MMHC	173,51
Forward	161,78	RSMAX2	173,49
Bidirectional F	161,78	HC (com wl)	172,77
Bidirectional B	161,78	Tabu (com wl)	172,76
HC	172,77	MMHC (com wl)	173,53
Tabu	172,76	RSMAX2 (com wl)	173,50

Fonte: Do autor (2024).

Com o aumento do tamanho amostral para $n = 450$, todos os métodos RB obtiveram melhores taxas de seleção e os métodos *stepwise* tiveram uma queda de 3% na taxa de seleção. Nesse cenário os métodos *MMHC* e *RSMAX2* alcançaram a maior porcentagem de acerto, com 95%, seguidos pelos métodos *HC* e *Tabu*, ambos com 93%.

Tabela 4.24 – Taxa de seleção das variáveis para cada método, considerando estrutura de associação 2, $n = 450$, VE4 e variável resposta binária.

Acertos	Backward	Forward	Bidirectional F	Bidirectional B	HC	Tabu	MMHC	RSMAX2
Apenas as 3 variáveis	0,54	0,54	0,54	0,54	0,93	0,93	0,95	0,95
Uma das variáveis	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00
Duas das variáveis	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00
Três das variáveis	1,00	1,00	1,00	1,00	1,00	1,00	1,00	1,00
x1	0,21	0,21	0,21	0,21	0,03	0,03	0,03	0,03
x2	0,16	0,16	0,16	0,16	0,01	0,01	0,01	0,01
x3	0,18	0,18	0,18	0,18	0,02	0,02	0,02	0,02
x4	1,00	1,00	1,00	1,00	1,00	1,00	1,00	1,00
x5	1,00	1,00	1,00	1,00	1,00	1,00	1,00	1,00
x6	1,00	1,00	1,00	1,00	1,00	1,00	1,00	1,00

Fonte: Do autor (2024).

A adição da informação *whitelist* ($x_2 \rightarrow x_4$) resultou em um aumento geral na taxa de acerto das relações para todos os algoritmos. Os métodos *MMHC* e *RSMAX2* se destacaram, alcançando as melhores médias de verdadeiro positivo ($6,16$ e $6,17$, respectivamente) e as menores médias de falso positivo ($0,31$ para ambos). Estes resultados indicam que, com a inclusão da *whitelist*, *MMHC* e *RSMAX2* não apenas identificaram com mais precisão as relações verdadeiras, mas também reduziram a ocorrência de falsos positivos.

Tabela 4.25 – Taxa de seleção das variáveis para cada algoritmo de rede bayesiana considerando *whitelist*, para estrutura de associação 2, $n = 450$, VE4 e variável resposta binária.

Acertos	HC	Tabu	MMHC	RSMAX2
Apenas as 3 variáveis	0,93	0,93	0,95	0,95
Uma das variáveis	0,00	0,00	0,00	0,00
Duas das variáveis	0,00	0,00	0,00	0,00
Três das variáveis	1,00	1,00	1,00	1,00
x1	0,03	0,03	0,03	0,03
x2	0,01	0,01	0,01	0,01
x3	0,02	0,02	0,02	0,02
x4	1,00	1,00	1,00	1,00
x5	1,00	1,00	1,00	1,00
x6	1,00	1,00	1,00	1,00

Fonte: Do autor (2024).

Apesar desse progresso na identificação das relações, não foi observado um impacto significativo nas relações que afetam diretamente a variável resposta y. Portanto, enquanto a inclusão da *whitelist* aprimorou a performance geral dos algoritmos na identificação de relações, a influência direta sobre a variável resposta não apresentou mudanças substanciais.

Tabela 4.26 – Média dos valores AIC para cada método, considerando estrutura de associação 2, $n = 450$, VE4 e variável resposta binária.

Método	AIC	Método	AIC
Backward	478,73	MMHC	507,48
Forward	478,73	RSMAX2	507,48
Bidirectional F	478,73	HC (com wl)	507,41
Bidirectional B	478,73	Tabu (com wl)	507,41
HC	507,41	MMHC (com wl)	507,48
Tabu	507,41	RSMAX2 (com wl)	507,48

Fonte: Do autor (2024).

Os métodos *Backward*, *Forward* e *Bidirectional F* obtiveram as melhores pontuações, com um valor médio de AIC de $478,73$. Estes métodos demonstraram superioridade em termos de ajuste do modelo, sugerindo que são mais eficazes na identificação das melhores estruturas de associação entre variáveis neste contexto específico.

Entre os algoritmos de redes bayesianas, *HC* e *Tabu* também se destacaram, com valores de AIC de $507,41$. Embora esses algoritmos apresentem um desempenho inferior comparado aos métodos mencionados acima, ainda assim se mostraram competitivos, especialmente considerando a inclusão da *whitelist*. Isso sugere que, apesar da adição da *whitelist*, esses métodos

mantêm um desempenho robusto, mas não alcançam a eficiência dos métodos tradicionais de seleção de modelos.

4.3 Discussão dos resultados

De maneira geral para o caso contínuo, conforme apresentado na Tabela 4.27, os métodos de seleção de variáveis do tipo *stepwise* destacaram-se em comparação aos algoritmos de rede bayesiana. No cenário de estrutura de associação 1, os métodos *Backward*, *Forward*, *Bidirectional F* e *Bidirectional B* consistentemente escolheram as seis variáveis esperadas $(x_1, x_2, x_3, x_4, x_5 \text{ e } x_6)$. No contexto da estrutura de associação 2, para todos os tamanhos de amostra, os métodos *Forward* e *Bidirectional F* demonstraram uma taxa de seleção mais eficiente das variáveis previstas $(x_4, x_5, \text{e}, x_6)$ em comparação com os demais métodos. Já no cenário de estrutura de associação 3, tanto para a variável resposta y_1 quanto para y_2, os métodos *Forward* e *Bidirectional F* apresentaram um desempenho superior na seleção das variáveis esperadas, incluindo $(x_6, x_7, x_8 \text{ e } x_9)$ e $(x_7, x_8, x_9 \text{ e } x_{10})$, respectivamente.

É importante ressaltar que os algoritmos de redes bayesianas, tais como *HC*, *Tabu*, *MMHC* e *RSMAX2*, não foram capazes de selecionar todas as variáveis do modelo verdadeiro em nenhum dos cenários considerados.

Tabela 4.27 – Taxa de seleção das variáveis que estão no verdadeiro modelo para cada método, considerando o caso de resposta contínua

Estrutura	n	Backward	Forward	Bidirectional F	Bidirectional B	HC	Tabu	MMHC	RSMAX2
1	50	**1,00**	**1,00**	**1,00**	**1,00**	0,00	0,00	0,00	0,00
	150	**1,00**	**1,00**	**1,00**	**1,00**	0,00	0,00	0,00	0,00
	450	**1,00**	**1,00**	**1,00**	**1,00**	0,00	0,00	0,00	0,00
2	50	0,08	**0,54**	**0,54**	0,05	0,00	0,00	0,00	0,00
	150	0,06	**0,59**	**0,59**	0,04	0,00	0,00	0,00	0,00
	450	0,06	**0,60**	**0,60**	0,04	0,00	0,00	0,00	0,00
3 (y_1)	50	0,01	**0,31**	**0,31**	0,00	0,00	0,00	0,00	0,00
	150	0,00	**0,35**	**0,35**	0,00	0,00	0,00	0,00	0,00
	450	0,00	**0,36**	**0,36**	0,00	0,00	0,00	0,00	0,00
3 (y_2)	50	0,00	**0,33**	**0,33**	0,00	0,00	0,00	0,00	0,00
	150	0,00	**0,37**	**0,37**	0,00	0,00	0,00	0,00	0,00
	450	0,08	**0,38**	**0,38**	0,05	0,00	0,00	0,00	0,00

Fonte: Do autor (2024).

Considerando o caso de resposta binária, na primeira estrutura, os métodos *Backward*, *Forward*, *Bidirectional F* e *Bidirectional B* demonstraram desempenho superior em comparação

aos algoritmos de redes bayesianas (RB). Para $n = 50$, os métodos *Backward* e *Bidirectional B* apresentaram as maiores taxas de seleção, especialmente considerando as variáveis explicativas com três categorias. Observa-se que o aumento do tamanho da amostra e o acréscimo no número de categorias das variáveis explicativas resultaram em melhorias nas taxas de seleção para todos os métodos.

Tabela 4.28 – Taxa de seleção das variáveis que estão no verdadeiro modelo para cada método, considerando o caso de resposta binária

Estrutura	n	Categorias	Backward	Forward	Bidirectional F	Bidirectional B	HC	Tabu	MMHC	RSMAX2
1	50	2	**0,01**	**0,01**	**0,01**	**0,01**	0,00	0,00	0,00	0,00
		3	**0,02**	0,01	0,01	**0,02**	0,00	0,00	0,00	0,00
		4	**0,01**	**0,01**	**0,01**	**0,01**	0,00	0,00	0,00	0,00
	150	2	**0,03**	**0,03**	**0,03**	**0,03**	0,00	0,00	0,00	0,00
		3	**0,06**	**0,06**	**0,06**	**0,06**	0,00	0,00	0,00	0,00
		4	**0,06**	**0,06**	**0,06**	**0,06**	0,00	0,00	0,00	0,00
	450	2	0,15	0,15	0,15	0,15	0,01	0,01	0,00	0,00
		3	**0,21**	**0,21**	**0,21**	**0,21**	0,01	0,01	0,00	0,00
		4	**0,23**	**0,23**	**0,23**	**0,23**	0,01	0,01	0,00	0,00
2	50	2	**0,19**	0,18	**0,19**	**0,19**	0,10	0,11	0,03	0,03
		3	**0,27**	0,26	**0,27**	**0,27**	0,18	0,19	0,05	0,05
		4	**0,30**	**0,30**	**0,30**	**0,30**	0,22	0,23	0,07	0,07
	150	2	0,46	0,46	0,46	0,46	**0,63**	**0,63**	0,54	0,54
		3	0,55	0,55	0,55	0,55	**0,81**	**0,81**	0,74	0,73
		4	0,57	0,57	0,57	0,57	**0,85**	**0,85**	0,80	0,80
	450	2	0,30	0,30	0,30	0,30	0,80	0,80	**0,81**	**0,81**
		3	0,46	0,46	0,46	0,46	0,90	0,90	**0,92**	**0,92**
		4	0,54	0,54	0,54	0,54	0,93	0,93	**0,95**	**0,95**
$3\,(y_1)$	50	2	**0,02**	**0,02**	**0,02**	**0,02**	**0,02**	**0,02**	0,00	0,00
		3	0,03	**0,04**	**0,04**	0,03	**0,04**	**0,04**	0,00	0,00
		4	0,04	0,04	0,04	0,04	**0,05**	**0,05**	0,00	0,00
	150	2	0,12	0,12	**0,13**	0,12	0,12	0,12	0,02	0,02
		3	0,19	0,19	0,19	0,19	**0,20**	0,19	0,04	0,04
		4	0,21	**0,22**	**0,22**	0,21	**0,22**	**0,22**	0,04	0,05
	450	2	0,11	0,12	0,12	0,11	0,11	0,11	**0,29**	**0,30**
		3	0,23	0,24	0,24	0,23	0,24	0,24	**0,45**	**0,45**
		4	0,28	0,30	0,30	0,28	0,30	0,30	**0,55**	**0,56**
$3\,(y_2)$	50	2	**0,02**	**0,02**	**0,02**	**0,02**	**0,02**	**0,02**	0,00	0,00
		3	0,03	0,03	0,03	0,03	**0,04**	**0,04**	0,00	0,00
		4	0,03	**0,04**	**0,04**	0,03	**0,04**	**0,04**	0,00	0,00
	150	2	0,11	**0,12**	**0,12**	0,11	**0,12**	**0,12**	0,02	0,02
		3	0,16	**0,17**	**0,17**	0,16	**0,17**	**0,17**	0,04	0,04
		4	0,18	0,19	0,19	0,18	**0,20**	**0,20**	0,04	0,04
	450	2	0,13	**0,14**	**0,14**	0,13	**0,14**	**0,14**	**0,14**	**0,14**
		3	0,24	**0,25**	**0,25**	0,23	**0,25**	**0,25**	0,18	0,18
		4	0,28	**0,31**	**0,31**	0,28	**0,31**	0,30	0,20	0,20

Fonte: Do autor (2024).

No contexto da estrutura de associação 2, os métodos *Stepwise* superaram as RB para o tamanho amostral de $n = 50$. No entanto, para $n = 150$ e $n = 450$, os algoritmos de redes bayesianas demonstraram desempenho superior em relação aos métodos *Stepwise*. Em particular, os algoritmos *HC* e *Tabu* registraram as maiores taxas de seleção para $n = 150$, enquanto *MMHC* e *RSMAX2* se destacaram para $n = 450$. Esses resultados corroboram com a pesquisa de Kitson et al. (2022), que destaca a importância de métodos híbridos, que combinam técnicas baseadas em restrições e busca de pontuação, como uma forma de equilibrar precisão e escalabilidade. Em todas as situações, o aumento nas categorias das variáveis explicativas resultou em melhorias nas taxas de seleção para todos os métodos analisados.

Quanto à estrutura de associação 3, os métodos *Stepwise*, *HC* e *Tabu* apresentaram desempenhos bastante similares, com taxas de seleção de variáveis muito próximas. Para a variável resposta y_1 com $n = 450$, os algoritmos *MMHC* e *RSMAX2* registraram as melhores taxas de seleção de variáveis. No entanto, para os demais cenários, esses dois algoritmos demonstraram um desempenho inferior.

O trabalho de Friedman e Goldszmidt (2000) aborda a utilização de redes bayesianas para a seleção de variáveis em conjuntos de dados com atributos contínuos. Uma das principais conclusões é que a discretização de variáveis contínuas, quando combinada com redes bayesianas, pode melhorar significativamente a identificação de variáveis relevantes, capturando interações complexas que os métodos convencionais de regressão não conseguem modelar adequadamente.

Os resultados obtidos, principalmente considerando as estruturas 2 e 3, destacam a eficácia das redes bayesianas na modelagem de variáveis categóricas, especialmente em cenários de resposta binária. O estudo de Campos, Zeng e Ji (2009) reforça a adequação dessa técnica em situações onde é necessário capturar relações de dependência complexas entre variáveis discretas. Essa eficiência no aprendizado de estrutura é coerente com as conclusões de Bari (2011), que enfatiza o desempenho robusto das redes bayesianas em cenários com variáveis discretas, validando sua aplicabilidade em modelos binários.

Portanto, uma abordagem eficaz seria empregar métodos *stepwise* para identificar as variáveis que têm um impacto direto na variável resposta, enquanto as redes bayesianas seriam utilizadas para avaliar e modelar o grau de associação e as interações entre as variáveis.

5 RESULTADOS DA APLICAÇÃO

Para realizar as análises, primeiramente, foi feita uma discretização das variáveis "idade", "experiencia", "cadastro"e "EPI", pois as suposições paramétricas para dados mistos têm limitações, uma vez que impõem restrições sobre quais arcos podem estar presentes no gráfico (por exemplo, um nó contínuo não pode ser pai de um nó discreto). As variáveis "idade", "experiencia"e "cadastro"foram divididas em duas categorias, considerando a mediana. Portanto as variáveis ficaram com essas categorias

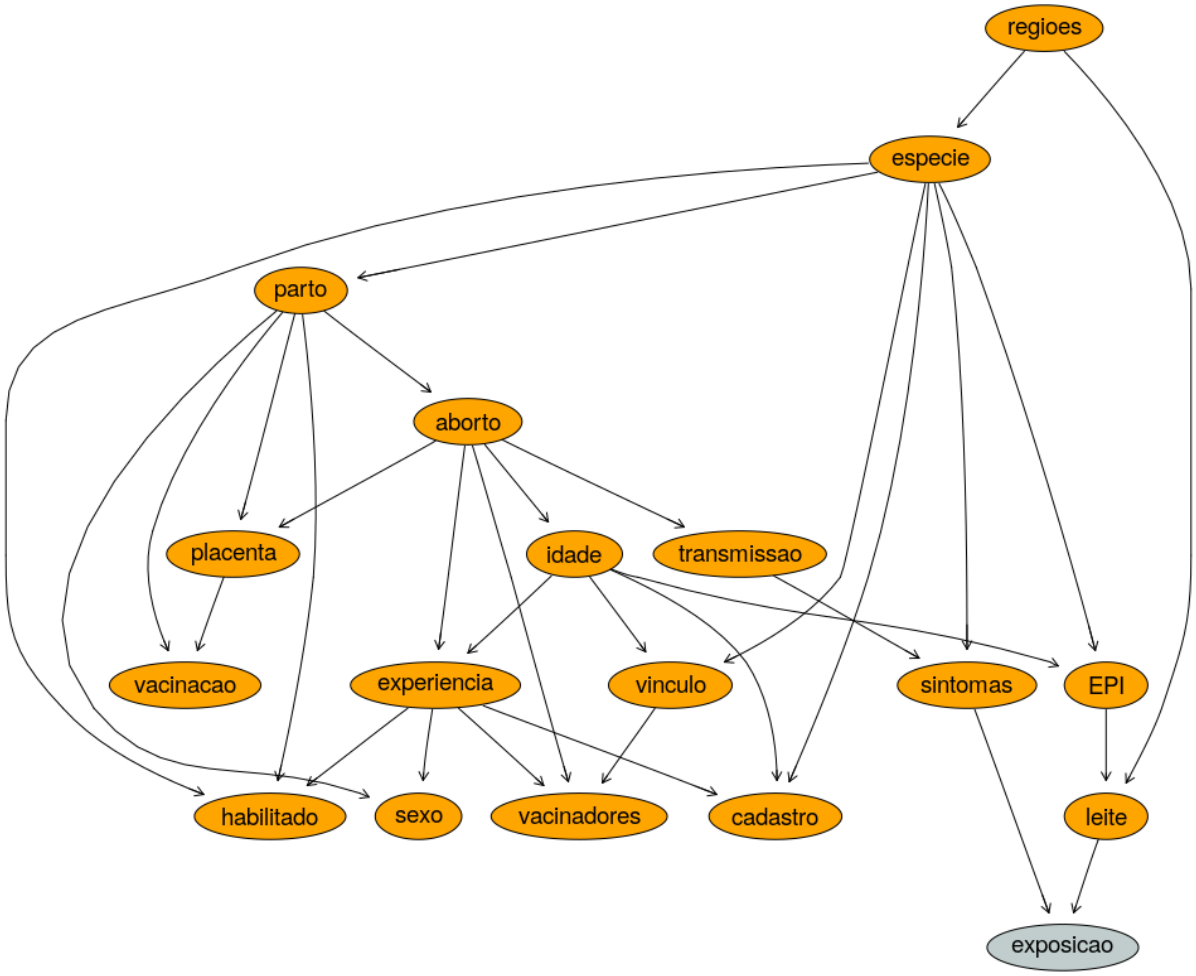
Tabela 5.1 – Tabela descrevendo as variáveis discretizadas.

Variável	Categorias
idade	até 36 anos
	acima de 36 anos
experiencia	até 10 anos
	acima de 10 anos
cadastro	até 8 anos
	acima de 8 anos
EPI	Pouco (pontuação$\leq$ 7)
	Intermediário (8 $\leq$ pontuação $\leq$ 15)
	Muito (pontuação $>$ 15)

Fonte: Do autor (2024).

Para construir a estrutura gráfica da Figura 5.1, utilizou-se o algoritmo *Hill-Climbing*, informando apenas que a variável resposta "exposicao"não deve afetar nenhuma variável explicativa. Desse modo, obtemos o seguinte modelo gráfico

Figura 5.1 – Rede encontrada pelo algoritmo HC.

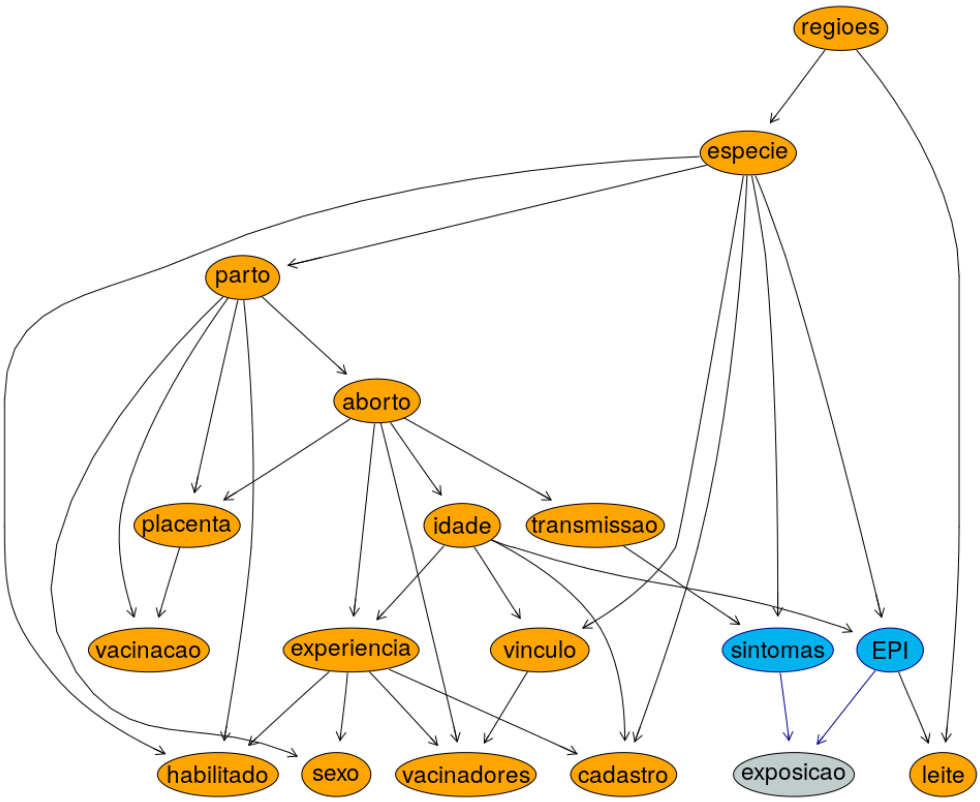


Fonte: Do autor (2024).

Foi identificado outra estrutura, conforme ilustrado na Figura 5.2, no qual foi incorporada uma *whitelist* indicando as variáveis selecionadas pelo método *stepwise* (EPI e sintomas), as mesmas variáveis selecionadas pelo método *purposeful selection of variables* no estudo de Pereira et al. (2020b).

As variáveis "habilitado", "parto", "placenta", "vacinação", "aborto"e "espécie", não escolhidas pelo método *stepwise*, foram consideradas importantes pela pesquisadora. Portanto, essas relações foram consideradas e adicionadas uma a uma. Foram incluídas apenas as variáveis que melhoraram a pontuação (AIC) da rede, conforme apresentado na Tabela 5.2.

Figura 5.2 – Rede acrescentando *whitelist*.



Fonte: Do autor (2024).

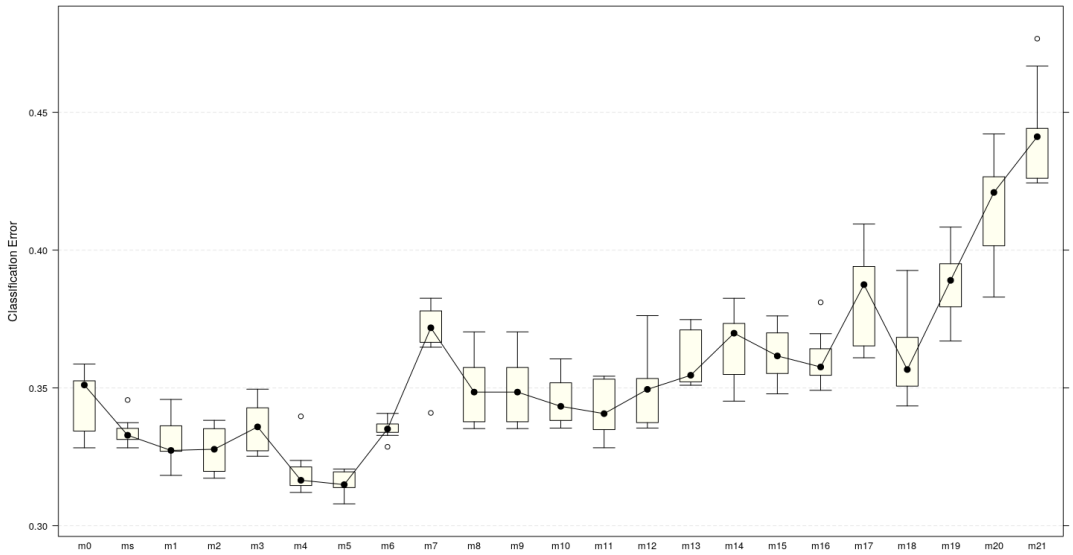
Tabela 5.2 – Variáveis que influenciam diretamente a variável resposta exposição

Modelo	Variáveis	AIC
m_0	leite + sintomas	$-3803,46$
m_s	EPI + sintomas	$-3806,6$
m_1	EPI + sintomas + habilitado	$-3812,83$
m_2	EPI + sintomas + parto	$-3811,06$
m_3	EPI + sintomas + placenta	$-3812,87$
m_4	EPI + sintomas + vacinacao	$-3811,56$
m_5	EPI + sintomas + aborto	$-3808,07$
m_6	EPI + sintomas + especie	$-3815,65$
m_7	EPI + sintomas + especie + habilitado	$-3833,61$
m_8	EPI + sintomas + especie + parto	$-3836,60$
m_9	EPI + sintomas + especie + placenta	$-3836,40$
m_{10}	EPI + sintomas + especie + vacinacao	$-3838,47$
m_{11}	EPI + sintomas + especie + aborto	$-3829,04$
m_{12}	EPI + sintomas + especie + vacinacao + habilitado	$-3879,07$
m_{13}	EPI + sintomas + especie + vacinacao + placenta	$-3884,72$
m_{14}	EPI + sintomas + especie + vacinacao + parto	$-3887,01$
m_{15}	EPI + sintomas + especie + vacinacao + aborto	$-3879,03$
m_{16}	EPI + sintomas + especie + vacinacao + parto + habilitado	$-3977,20$
m_{17}	EPI + sintomas + especie + vacinacao + parto + placenta	$-3981,67$
m_{18}	EPI + sintomas + especie + vacinacao + parto + aborto	$-3980,41$
m_{19}	EPI + sintomas + especie + vacinacao + parto + placenta + habilitado	$-4179,98$
m_{20}	EPI + sintomas + especie + vacinacao + parto + placenta + aborto	$-4182,61$
m_{21}	EPI + sintomas + especie + vacinacao + parto + placenta + aborto + habilitado	-4598.77

Fonte: Do autor (2024).

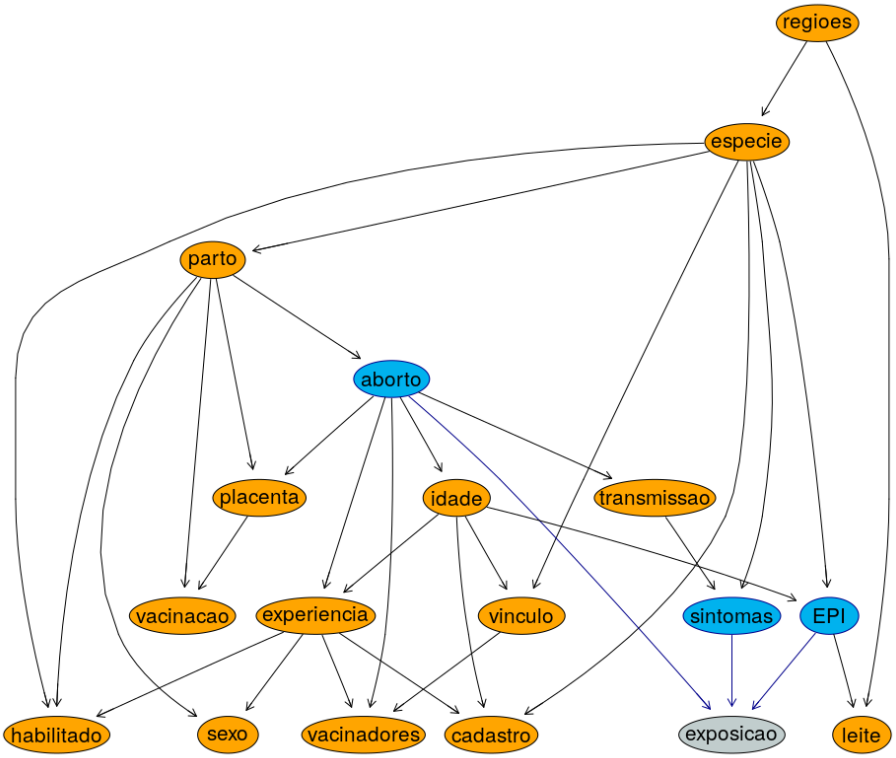
Dado os modelos representados na tabela 5.2, foi realizado 10 vezes a validação 10−*fold* para avaliar a qualidade de predição de cada modelo.

Figura 5.3 – Boxplot dos resultados obtidos da validação cruzada.



Fonte: Do autor (2024).

Figura 5.4 – Rede final.



Fonte: Do autor (2024).

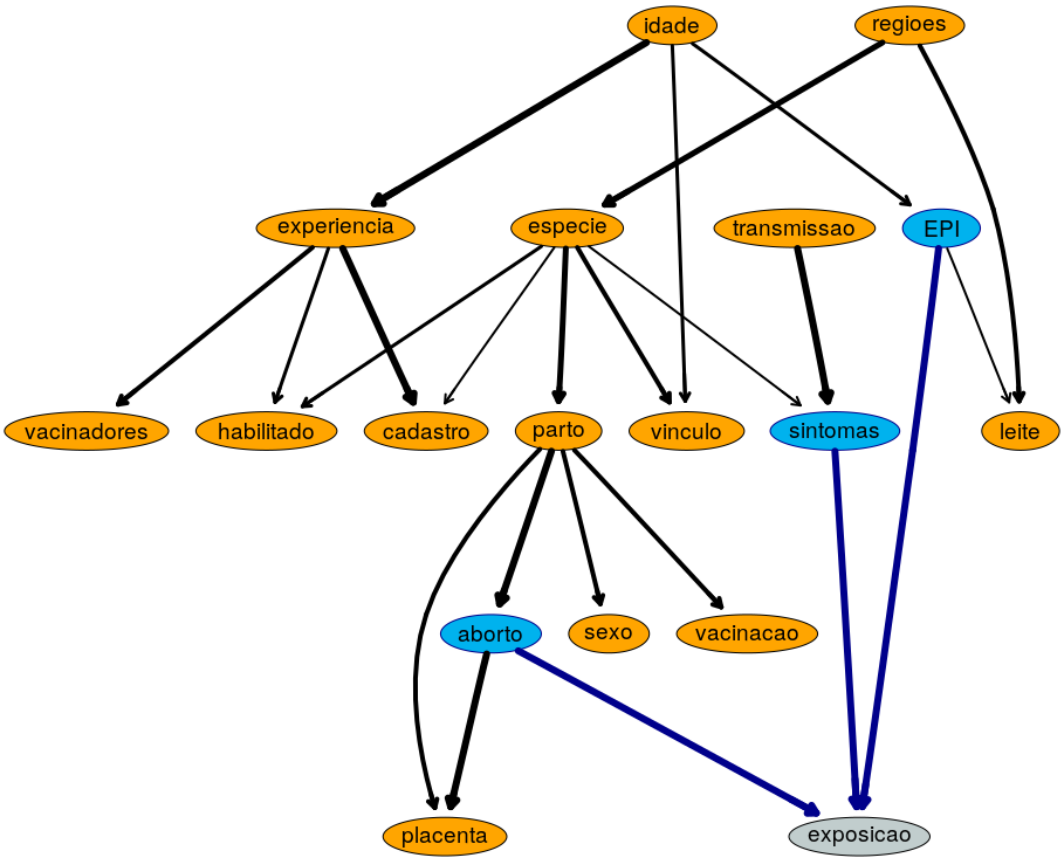
Conforme mostrado na Figura 5.3, o modelo m_5 que contém as variáveis "sintomas", "EPI"e "aborto"influenciando diretamente a variável resposta, apresentou o menor erro de predição e foi escolhido como modelo final, representado na Figura 5.4.

Para identificar quais arcos representam relacionamentos fortes, utilizou-se os seguintes passos:

a) os dados foram reamostrados utilizando *bootstrap*;
b) foi construída uma rede distinta para cada amostra de *booststrap*;
c) a frequência de ocorrência de cada possível conexão nas redes foi examinada;
d) uma rede de consenso foi formada utilizando as conexões mais frequentemente observadas.

A rede média resultante

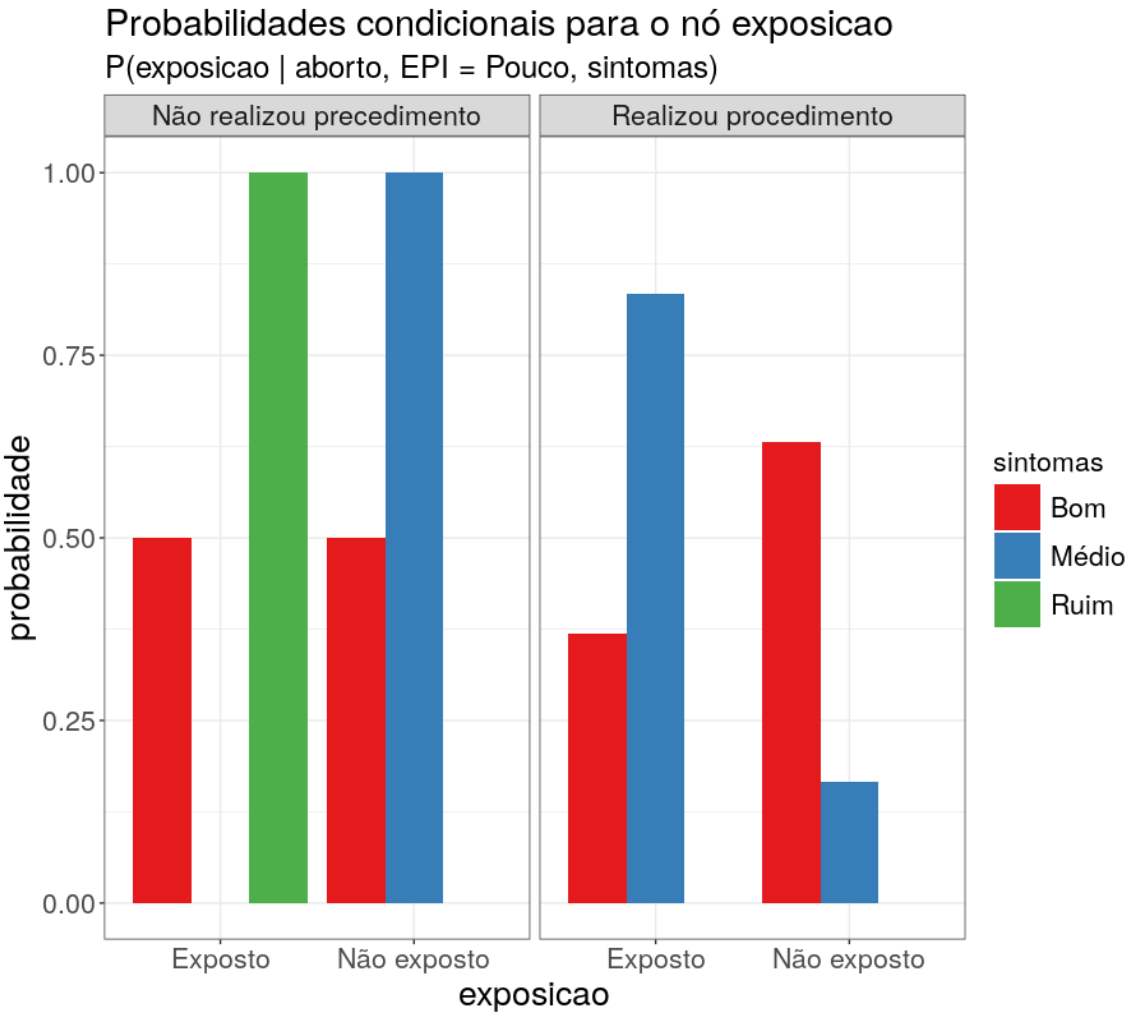
Figura 5.5 – Rede bayesiana média.



Fonte: Do autor (2024).

Considerando a rede 5.5, obtemos as tabelas de probabilidade condicional para o nó de interesse, "exposicao".

Figura 5.6 – Tabela de probabilidade condicional para o nó exposição, para a categoria pouco de EPI.

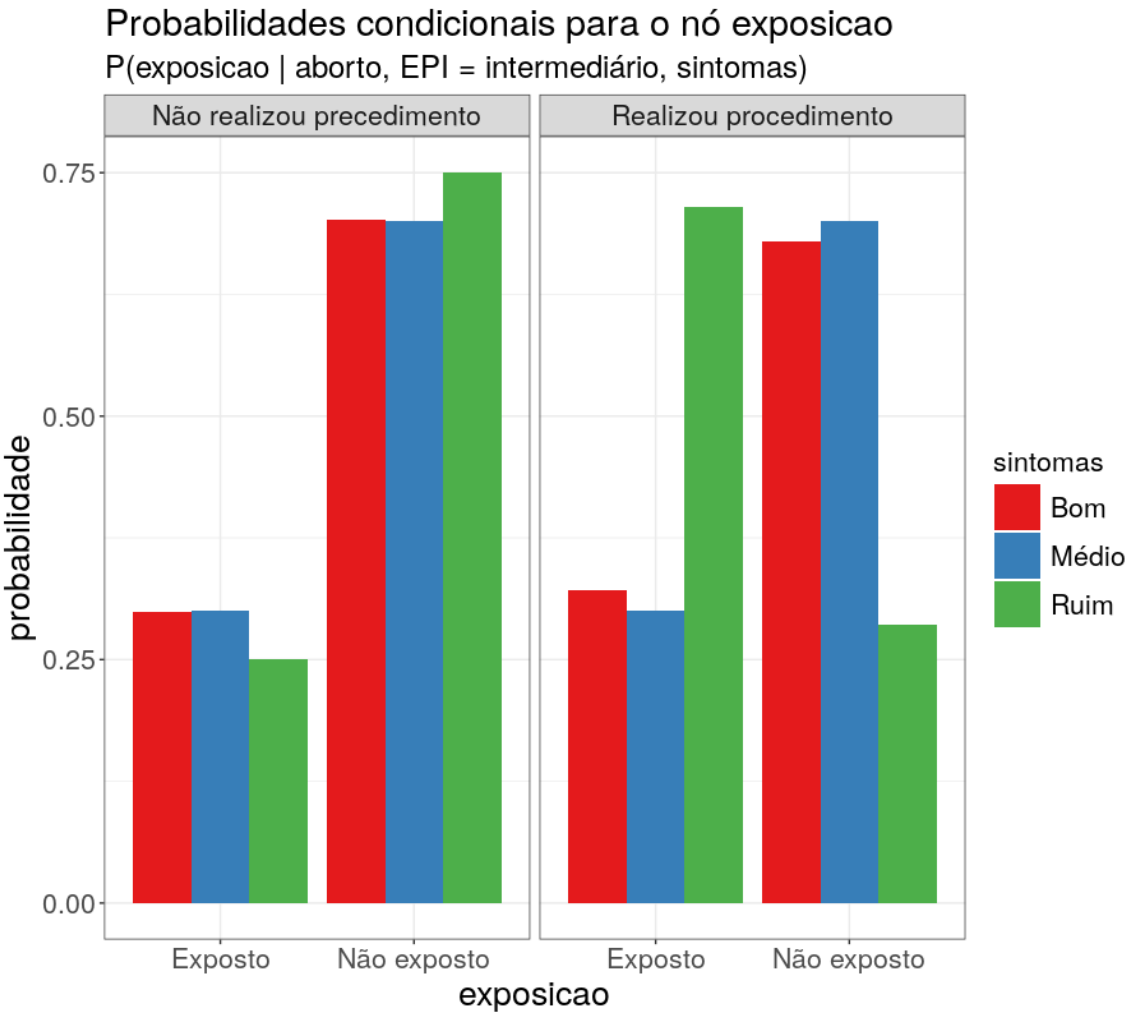


Fonte: Do autor (2024).

Entre os profissionais que utilizam pouco Equipamento de Proteção Individual (EPI) e que realizaram procedimentos de parto prematuro ou aborto, aqueles com maior conhecimento sobre os sintomas da Brucelose tiveram menos exposição às vacinas. Por outro lado, entre os profissionais que não passaram por procedimentos de parto prematuro, aqueles com conhecimento deficiente sobre os sintomas tiveram uma maior exposição.

Com base na Figura 5.7, observa-se que os profissionais que não realizaram procedimentos de aborto geralmente sofreram menos exposição, enquanto aqueles com um bom conhecimento dos sintomas têm uma probabilidade maior de evitar a exposição. Por outro lado, os profissionais que realizam partos prematuros têm uma probabilidade aumentada de exposição, especialmente quando possuem um conhecimento deficiente sobre os sintomas da brucelose.

Figura 5.7 – Tabela de probabilidade condicional para o nó exposição, para a categoria intermediário de EPI

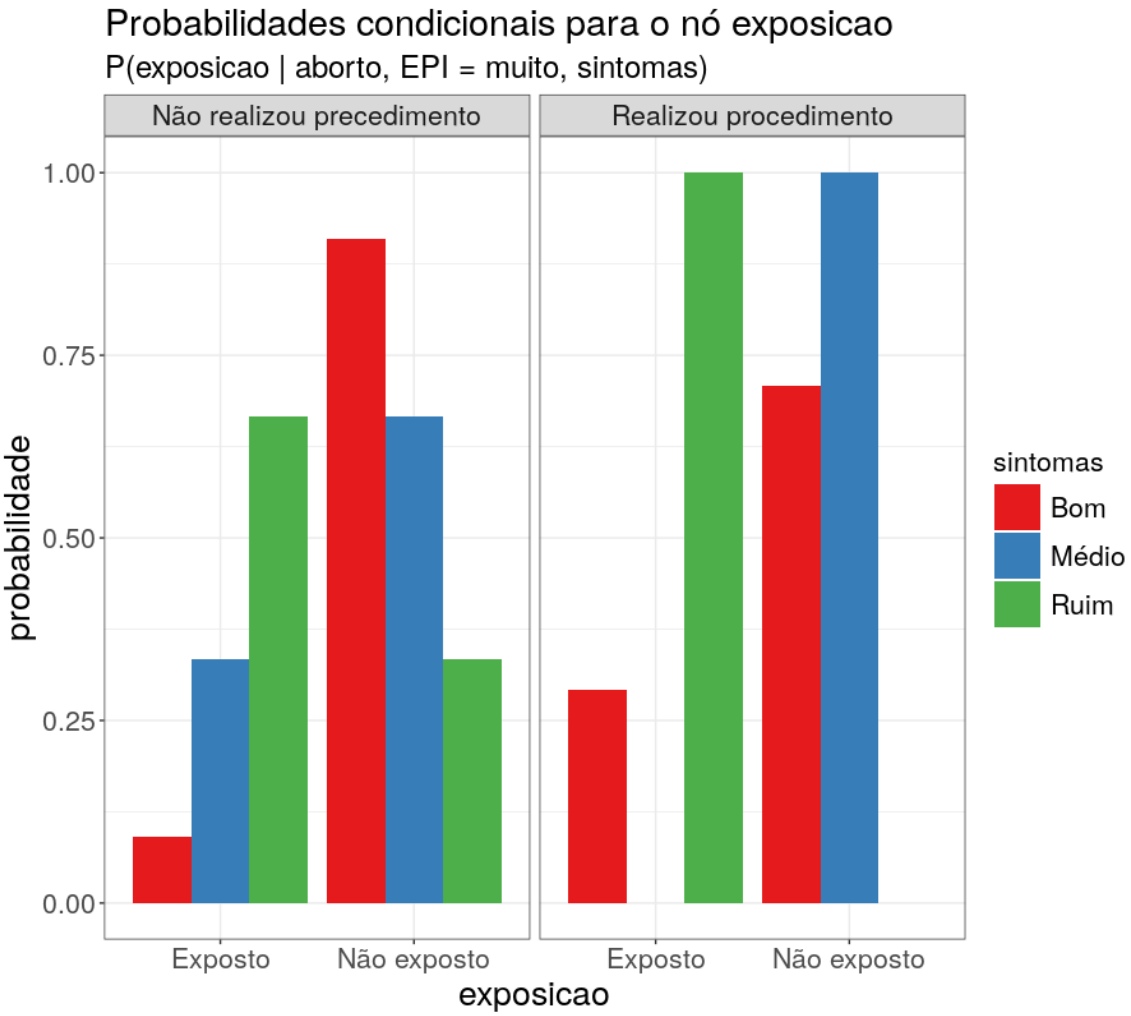


Fonte: Do autor (2024).

Entre os profissionais que usam EPI com frequência, ilustrado na Figura 5.8, aqueles com conhecimento médio ou bom dos sintomas da brucelose têm uma menor probabilidade de exposição às vacinas. Em contrapartida, profissionais com conhecimento deficiente sobre a brucelose tendem a ter uma exposição maior, especialmente se realizaram procedimentos de aborto.

Analisando as tabelas de probabilidades condicionais, observa-se que ao aumentar o uso dos Equipamentos de Proteção Individual (EPIs), juntamente com um bom conhecimento dos sintomas da brucelose, os profissionais tendem a ter uma exposição menor às vacinas. Por outro lado, a realização de procedimentos de parto prematuro ou aborto aumenta a exposição dos profissionais.

Figura 5.8 – Tabela de probabilidade condicional para o nó exposição, para a categoria muito de EPI.



Fonte: Do autor (2024).

Ao observar a rede bayesiana média resultante, mostrada na Figura 5.5, verifica-se com clareza as relações diretas e indiretas entre as variáveis explicativas e a variável resposta "exposição". Nota-se que as variáveis aborto, sintomas e EPI influenciam diretamente a variável resposta, o que está em conformidade com os estudos de Pereira et al. (2020a), que indicaram uma maior chance de infecção entre ocupações de campo que mantêm contato direto com animais e seus produtos. Veterinários e assistentes veterinários também relataram realizar atividades associadas a alto risco de infecção, como o atendimento a partos, tratamento de casos de infertilidade, e o manuseio de fetos abortados, placentas retidas e natimortos (AVDIKOU; MAIPA; ALAMANOS, 2005; CASH-GOLDWASSER et al., 2018; MAILLES et al., 2016). Além disso, o uso inadequado de EPI foi relatado em alguns casos (PROCH et al., 2018), corroborando os resultados apresentados nas Tabelas 5.6, 5.7 e 5.8.

Com base na representação da rede bayesiana final na Figura 5.5, podemos conduzir inferências sobre a exposição dos profissionais dadas algumas evidências específicas. Portanto, podemos estar interessados em determinar a probabilidade de um profissional ser exposto às vacinas, considerando a região de residência, a frequência de uso de EPIs e o nível de conhecimento dos sintomas da brucelose. Ao gerar amostras aleatórias condicionadas à evidência, estimamos a probabilidade condicional do evento "exposição = Exposto". Repetindo esse processo 2000 vezes, obtivemos um intervalo de confiança de 95%, como mostrado na Tabela 5.3.

Tabela 5.3 – Intervalo de confiança para a probabilidade de um profissional ser exposto às vacinas, considerando a região de residência, a frequência de uso de EPIs e o nível de conhecimento dos sintomas da brucelose.

Regiões	Sintomas	EPI	Exposição
			Exposto
Noroeste	Bom	Pouco	$(40,95;41,37)$
		Intermediário	$(29,22;29,39)$
		Muito	$(26,83;27,11)$
Leste	Bom	Pouco	$(41,58;42,14)$
		Intermediário	$(29,65;29,89)$
		Muito	$(26,81;27,19)$
Central	Bom	Pouco	$(42,63;42,87)$
		Intermediário	$(29,99;30,09)$
		Muito	$(27,15;27,31)$
Zona da Mata	Bom	Pouco	$(42,53;42,93)$
		Intermediário	$(29,99;30,15)$
		Muito	$(27,25;27,54)$
Sul	Bom	Pouco	$(42,51;42,78)$
		Intermediário	$(29,99;30,10)$
		Muito	$(27,15;27,35)$
Alto Paranaíba	Bom	Pouco	$(42,69;43,09)$
		Intermediário	$(30,02;30,18)$
		Muito	$(27,05;27,33)$
Triângulo Mineiro	Bom	Pouco	$(41,35;41,74)$
		Intermediário	$(29,32;29,48)$
		Muito	$(26,85;27,11)$

Fonte: Do autor (2024).

Ao analisar os resultados, percebemos que as regiões com as maiores probabilidades de exposição são Sul $(42,69\%;43,09\%)$ e Central $(42,53\%;42,93\%)$, ambas com EPI classificado como "Pouco". Por outro lado, as regiões com as menores probabilidades de exposição são

Leste $(26,81\%;27,19\%)$ e Triângulo Mineiro $(26,85\%;27,11\%)$, ambas com EPI classificado como "Muito". Esses dados evidenciam que o uso adequado de EPI parece reduzir a probabilidade de exposição à vacina, ressaltando a importância do uso correto desses equipamentos.

Esses resultados destacam a importância de considerar múltiplos fatores ao avaliar o risco de exposição à vacina de brucelose entre os profissionais de saúde, incluindo o conhecimento dos sintomas, o uso de EPI e a região geográfica de residência. Essas informações podem ser úteis para orientar políticas de prevenção e proteção desses profissionais contra a exposição a doenças infecciosas.

6 CONCLUSÕES E RECOMENDAÇÕES

Neste trabalho, alcançamos algumas conclusões significativas. Em todos os cenários, quanto maior o tamanho da amostra, maior é a taxa de seleção de variáveis. Em geral os métodos *stepwise* mostraram um desempenho superior em relação às redes bayesianas, especialmente no contexto de resposta contínua. Em determinadas situações com resposta categorizada, as redes bayesianas apresentam resultados equivalentes ou superiores.

Nos cenários com apenas uma variável resposta binária, aumenta a taxa de seleção das variáveis em todos os métodos à medida que o número de categorias nas variáveis explicativas aumentava. No entanto, a inclusão de um grande número de categorias torna a interpretação dos dados mais complexa, podendo dificultar a análise e a tomada de decisões.

Na aplicação com dados reais de brucelose as variáveis explicativas "aborto", "EPI"e "sintomas"têm uma influência direta na variável resposta "exposição". Além disso, também observa-se que as variáveis "parto", "idade", "transmissao", "regioes"e "espécie"exercem um efeito indireto sobre a variável resposta, indicando a complexidade das relações entre as variáveis no contexto estudado. Isto não se identifica com métodos automáticos de seleção.

Recomendamos, portanto, a combinação dos métodos *stepwise* com as redes bayesianas. Enquanto os métodos *stepwise* se mostraram eficazes na seleção automática de variáveis, as redes bayesianas oferecem uma ferramenta poderosa para visualizar as relações entre as variáveis e compreender associações indiretas. Essa abordagem combinada pode enriquecer a análise e interpretação dos resultados, proporcionando uma visão mais completa e detalhada do problema em estudo.

REFERÊNCIAS

AGRESTI, A. **Categorical Data Analysis**. 3. ed. [S.l.]: Wiley, 2012.

AVDIKOU, I.; MAIPA, V.; ALAMANOS, Y. Epidemiology of human brucellosis in a defined area of northwestern greece. **Epidemiology and Infection**, v. 133, n. 5, p. 905–910, 2005.

BARI, M. F. Bayesian network structure learning. In: **Proceeding of the 4th Annual**. [S.l.: s.n.], 2011.

BURSAC, Z. et al. Purposeful selection of variables in logistic regression. **Source Code for Biology and Medicine**, v. 3, n. 17, 2008. Disponível em: <https://doi.org/10.1186/1751-0473-3-17>.

CALADO, P. et al. Combining link-based and content-based methods for web document classification. In: **Proceedings of the Twelfth International Conference on Information and Knowledge Management**. New York, NY, USA: Association for Computing Machinery, 2003. (CIKM '03), p. 394–401. ISBN 1581137230. Disponível em: <https://doi.org/10.1145/956863.956938>.

CAMPOS, C. P. de; ZENG, Z.; JI, Q. Structure learning of bayesian networks using constraints. In: **Proceedings of the 26th Annual International Conference on Machine Learning**. New York, NY, USA: Association for Computing Machinery, 2009. (ICML '09), p. 113–120. ISBN 9781605585161. Disponível em: <https://doi.org/10.1145/1553374.1553389>.

CASELLA, G.; BERGER, R. **Statistical Inference**. [S.l.]: Duxbury Resource Center, 2001. Textbook Binding. ISBN 0534243126.

CASH-GOLDWASSER, S. et al. Risk factors for human brucellosis in northern tanzania. **The American Journal of Tropical Medicine and Hygiene**, The American Society of Tropical Medicine and Hygiene, Arlington VA, USA, v. 98, n. 2, p. 598 – 606, 2018. Disponível em: <https://www.ajtmh.org/view/journals/tpmd/98/2/article-p598.xml>.

DENOYER, L.; GALLINARI, P. Bayesian network model for semi-structured document classification. **Information Processing Management**, v. 40, n. 5, p. 807–827, 2004. ISSN 0306-4573. Bayesian Networks and Information Retrieval. Disponível em: <https://www.sciencedirect.com/science/article/pii/S030645730400041X>.

FRIEDMAN, N.; GOLDSZMIDT, M. Discretizing continuous attributes while learning bayesian networks. **Proceedings of the Thirteenth International Conference on Machine Learning**, 06 2000.

HILBE, J. M. **Logistic Regression Models**. [S.l.]: CRC Press, 2009.

HOSMER, D. W.; LEMESHOW, S. **Applied logistic regression**. [S.l.]: John Wiley and Sons, 2000. ISBN 0471356328, 9780471356325.

HRUSCHKA, E. R.; HRUSCHKA, E. R.; EBECKEN, N. F. F. Feature selection by bayesian networks. In: TAWFIK, A. Y.; GOODWIN, S. D. (Ed.). **Advances in Artificial Intelligence**. Berlin, Heidelberg: Springer Berlin Heidelberg, 2004. p. 370–379. ISBN 978-3-540-24840-8.

JAMES, G. et al. **An Introduction to Statistical Learning: with Applications in R**. Springer, 2013. Disponível em: <https://faculty.marshall.usc.edu/gareth-james/ISL/>.

KITSON, N. K. et al. **A survey of Bayesian Network structure learning**. 2022. Disponível em: <https://arxiv.org/abs/2109.11415>.

KOSKI, T.; NOBLE, J. M. **Bayesian Networks: An Introduction**. [S.l.]: Jonh Wiley and Sons, 2009.

KRISTENSEN, K.; RASMUSSEN, I. A. The use of a bayesian network in the design of a decision support system for growing malting barley without use of pesticides. **Computers and Electronics in Agriculture**, v. 33, n. 3, p. 197–217, 2002. ISSN 0168-1699. Disponível em: <https://www.sciencedirect.com/science/article/pii/S0168169902000078>.

KUTNER, M. H. et al. **Applied Linear Statistical Models**. [S.l.]: McGraw-Hill, 2005.

LEE, S.-M.; ABBOTT, P.; JOHANTGEN, M. Logistic regression and bayesian networks to study outcomes using large data sets. **Nursing research**, v. 54, p. 133–8, 03 2005.

LTIFI, H. et al. Dynamic decision support system based on bayesian networks application to fight against the nosocomial infections. arXiv, 2012. Disponível em: <https://arxiv.org/abs/1211.2126>.

MAILLES, A. et al. Human brucellosis in france in the 21st century: Results from national surveillance 2004–2013. **Médecine et Maladies Infectieuses**, v. 46, n. 8, p. 411–418, 2016. ISSN 0399-077X. Disponível em: <https://www.sciencedirect.com/science/article/pii/S0399077X16306667>.

MONTGOMERY, D. et al. **Introduction to Linear Regression Analysis, 5th Edition**. [S.l.]: John Wiley & Sons, 2012.

NAGARAJAN, R.; SCUTARI, M.; LBRE, S. **Bayesian Networks in R: With Applications in Systems Biology**. [S.l.]: Springer Publishing Company, Incorporated, 2013. ISBN 1461464455.

NEAPOLITAN, R. E. **Learning Bayesian Networks**. [S.l.]: Prentice Hall, 2003.

NELDER, J. A.; WEDDERBURN, R. W. M. Generalized linear models. **Journal of the Royal Statistical Society. Series A (General)**, [Royal Statistical Society, Wiley], v. 135, n. 3, p. 370–384, 1972. ISSN 00359238. Disponível em: <http://www.jstor.org/stable/2344614>.

PEARL, J. Bayesian networks: A model of self-activated memory for evidential reasoning. 1985.

PEARL, J. **Causality**: Models, reasoning, and inference. 2. ed. Cambridge, UK: Cambridge University Press, 2009. ISBN 978-0-521-89560-6.

PEREIRA, C. R. et al. Occupational exposure to brucella spp.: A systematic review and meta-analysis. **PLOS Neglected Tropical Diseases**, Public Library of Science, v. 14, n. 5, p. 1–19, 05 2020. Disponível em: <https://doi.org/10.1371/journal.pntd.0008164>.

PEREIRA, C. R. et al. Accidental exposure to brucella abortus vaccines and occupational brucellosis among veterinarians in minas gerais state, brazil. **Transbound Emerg Dis.**, p. 00:1–14, 2020. Disponível em: <https://doi.org/10.1111/tbed.13797>.

PROCH, V. et al. Risk factors for occupational brucella infection in veterinary personnel in india. **Transboundary and Emerging Diseases**, v. 65, n. 3, p. 791–798, 2018. Disponível em: <https://onlinelibrary.wiley.com/doi/abs/10.1111/tbed.12804>.

R Core Team. **R: A Language and Environment for Statistical Computing**. Vienna, Austria, 2020. Disponível em: <https://www.R-project.org/>.

SCUTARI, M. Learning bayesian networks with the bnlearn r package. **Journal of Statistical Software**, v. 35, n. 3, p. 1–22, 2010. Disponível em: <https://www.jstatsoft.org/index.php/jss/article/view/v035i03>.

SCUTARI, M.; DENIS, J. **Bayesian Networks: With Examples in R**. [S.l.]: CRC Press, 2021. (Chapman and Hall/CRC Texts in Statistical Science Series). ISBN 9780367366513.

STASINOPOULOS, M. D. et al. **Flexible regression and smoothing : using GAMLSS in R**. Chapman and Hall/CRC, 2017. (R). ISBN 9781138197909 1138197904. Disponível em: <https://www.crcpress.com/Flexible-Regression-and-Smoothing-Using-GAMLSS-in-R/Stasinopoulos-Rigby-Heller-Voudouris-Bastiani/p/book/9781138197909>.

VANNUCCI, M.; STINGO, F.; BERZUINI, C. Bayesian models for variable selection that incorporate biological information. In: ____. **Bayesian Statistics 9**. United Kingdom: Oxford University Press, 2012. v. 9780199694587. ISBN 9780199694587. Publisher Copyright: © Oxford University Press 2011. All rights reserved.

VERMA, T.; PEARL, J. Equivalence and synthesis of causal models. **Probabilistic and Causal Inference**, 1990.

YET, B. et al. Decision support system for warfarin therapy management using bayesian networks. **Decision Support Systems**, v. 55, n. 2, p. 488–498, 2013. ISSN 0167-9236. 1. Analytics and Modeling for Better HealthCare 2. Decision Making in Healthcare. Disponível em: <https://www.sciencedirect.com/science/article/pii/S016792361200262X>.

APÊNDICE A – RESULTADOS DO ESTUDO DE SIMULAÇÃO

Caso de resposta contínua

Estrutura de associação 1

n = 150

Tabela 1 – Taxa de seleção das variáveis para cada método, considerando estrutura de associação 1, $n = 150$ e variável resposta contínua.

Acertos	Backward	Forward	Bidirectional F	Bidirectional B	HC	Tabu	MMHC	RSMAX2
Apenas as 6 variáveis	1,00	1,00	1,00	1,00	0,00	0,00	0,00	0,00
Uma das variáveis	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00
Duas das variáveis	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00
Três das variáveis	0,00	0,00	0,00	0,00	0,00	0,00	0,04	0,04
Quatro das variáveis	0,00	0,00	0,00	0,00	0,00	0,00	0,69	0,69
Cinco das variáveis	0,00	0,00	0,00	0,00	1,00	1,00	0,26	0,26
x1	1,00	1,00	1,00	1,00	1,00	1,00	1,00	1,00
x2	1,00	1,00	1,00	1,00	1,00	1,00	0,27	0,27
x3	1,00	1,00	1,00	1,00	1,00	1,00	1,00	1,00
x4	1,00	1,00	1,00	1,00	0,00	0,00	0,02	0,02
x5	1,00	1,00	1,00	1,00	1,00	1,00	0,94	0,94
x6	1,00	1,00	1,00	1,00	1,00	1,00	1,00	1,00

Fonte: Do autor (2024).

Tabela 2 – Taxa de seleção das variáveis para cada algoritmo de rede bayesiana considerando *whitelist*, para estrutura de associação 1, $n = 150$ e variável resposta contínua.

Acertos	HC	Tabu	MMHC	RSMAX2
Apenas as 6 variáveis	0,00	0,00	0,00	0,00
Uma das variáveis	0,00	0,00	0,00	0,00
Duas das variáveis	0,00	0,00	0,00	0,00
Três das variáveis	0,00	0,00	0,04	0,04
Quatro das variáveis	0,00	0,00	0,70	0,70
Cinco das variáveis	1,00	1,00	0,26	0,26
x1	1,00	1,00	1,00	1,00
x2	1,00	1,00	0,27	0,27
x3	1,00	1,00	1,00	1,00
x4	0,00	0,00	0,02	0,02
x5	1,00	1,00	0,94	0,94
x6	1,00	1,00	1,00	1,00

Fonte: Do autor (2024).

Tabela 3 – Média de verdadeiro positivo (VP), falso positivo (FP) e falso negativo (FN) das relações dos algoritmos de RB, para estrutura de associação 1, $n = 150$ e variável resposta contínua.

Algoritmos	VP	FP	FN
HC	5,00	2,50	1,00
Tabu	5,00	2,55	1,00
MMHC	4,22	0,27	1,78
RSMAX2	4,22	0,27	1,78
HC (com wl)	5,00	2,50	1,00
Tabu (com wl)	5,00	2,55	1,00
MMHC (com wl)	4,22	0,26	1,78
RSMAX2 (com wl)	4,22	0,27	1,78

Fonte: Do autor (2024).

Tabela 4 – Média dos valores AIC para cada método, considerando estrutura de associação 1, $n = 150$ e variável resposta contínua.

Método	AIC	Método	AIC
Backward	-10001,33	MMHC	-65,76
Forward	-10008,64	RSMAX2	-65,71
Bidirectional F	-10008,64	HC (com wl)	-258,22
Bidirectional B	-10001,33	Tabu (com wl)	-258,22
HC	-258,22	MMHC (com wl)	-66,68
Tabu	-258,22	RSMAX2 (com wl	-66,60

Fonte: Do autor (2024).

n = 450

Tabela 5 – Taxa de seleção das variáveis para cada método, considerando estrutura de associação 1, $n = 450$ e variável resposta contínua.

Acertos	Backward	Forward	Bidirectional F	Bidirectional B	HC	Tabu	MMHC	RSMAX2
Apenas as 6 variáveis	1,00	1,00	1,00	1,00	0,00	0,00	0,00	0,00
Uma das variáveis	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00
Duas das variáveis	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00
Três das variáveis	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00
Quatro das variáveis	0,00	0,00	0,00	0,00	0,00	0,00	0,14	0,14
Cinco das variáveis	0,00	0,00	0,00	0,00	1,00	1,00	0,86	0,86
x1	1,00	1,00	1,00	1,00	1,00	1,00	1,00	1,00
x2	1,00	1,00	1,00	1,00	1,00	1,00	0,81	0,79
x3	1,00	1,00	1,00	1,00	1,00	1,00	1,00	1,00
x4	1,00	1,00	1,00	1,00	0,00	0,00	0,06	0,07
x5	1,00	1,00	1,00	1,00	1,00	1,00	1,00	1,00
x6	1,00	1,00	1,00	1,00	1,00	1,00	1,00	1,00

Fonte: Do autor (2024).

Tabela 6 – Taxa de seleção das variáveis para cada algoritmo de rede bayesiana considerando *whitelist*, para estrutura de associação 1, $n = 450$ e variável resposta contínua.

Acertos	HC	Tabu	MMHC	RSMAX2
Apenas as 6 variáveis	0,00	0,00	0,00	0,00
Uma das variáveis	0,00	0,00	0,00	0,00
Duas das variáveis	0,00	0,00	0,00	0,00
Três das variáveis	0,00	0,00	0,00	0,00
Quatro das variáveis	0,00	0,00	0,14	0,14
Cinco das variáveis	1,00	1,00	0,86	0,86
x1	1,00	1,00	1,00	1,00
x2	1,00	1,00	0,81	0,79
x3	1,00	1,00	1,00	1,00
x4	0,00	0,00	0,05	0,07
x5	1,00	1,00	1,00	1,00
x6	1,00	1,00	1,00	1,00

Fonte: Do autor (2024).

Tabela 7 – Média de verdadeiro positivo (VP), falso positivo (FP) e falso negativo (FN) das relações dos algoritmos de RB, para estrutura de associação 1, $n = 450$ e variável resposta contínua.

Algoritmos	VP	FP	FN
HC	5,00	2,42	1,00
Tabu	5,00	2,46	1,00
MMHC	4,86	0,16	1,14
RSMAX2	4,86	0,16	1,14
HC (com wl)	5,00	2,42	1,00
Tabu (com wl)	5,00	2,46	1,00
MMHC (com wl)	4,86	0,16	1,14
RSMAX2 (com wl	4,86	0,16	1,14

Fonte: Do autor (2024).

Tabela 8 – Média dos valores AIC para cada método, considerando estrutura de associação 1, $n = 450$ e variável resposta contínua.

Método	AIC	Método	AIC
Backward	-29619,37	MMHC	-652,33
Forward	-29625,96	RSMAX2	-642,00
Bidirectional F	-29625,96	HC (com wl)	-788,45
Bidirectional B	-29619,37	Tabu (com wl)	-788,45
HC	-788,45	MMHC (com wl)	-654,87
Tabu	-788,45	RSMAX2 (com wl	-642,86

Fonte: Do autor (2024).

Estrutura 2 - n =150

Tabela 9 – Taxa de seleção das variáveis para cada método, considerando estrutura de associação 2, $n = 150$ e variável resposta contínua.

Acertos	Backward	Forward	Bidirectional F	Bidirectional B	HC	Tabu	MMHC	RSMAX2
Apenas as 3 variáveis	0,06	0,59	0,59	0,04	0,00	0,00	0,00	0,00
Uma das variáveis	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00
Duas das variáveis	0,00	0,00	0,00	0,00	1,00	1,00	1,00	1,00
Três das variáveis	1,00	1,00	1,00	1,00	0,00	0,00	0,00	0,00
x_1	0,59	0,16	0,16	0,66	0,26	0,26	0,00	0,00
x_2	0,62	0,17	0,17	0,68	0,63	0,63	0,00	0,00
x_3	0,62	0,17	0,17	0,68	0,32	0,32	0,00	0,00
x_4	1,00	1,00	1,00	1,00	0,30	0,30	0,30	0,30
x_5	1,00	1,00	1,00	1,00	0,89	0,89	0,89	0,89
x_6	1,00	1,00	1,00	1,00	0,81	0,81	0,81	0,81

Fonte: Do autor (2024).

Tabela 10 – Taxa de seleção das variáveis para cada algoritmo de rede bayesiana considerando *whitelist*, para estrutura de associação 2, $n = 150$ e variável resposta contínua.

Acertos	HC	Tabu	MMHC	RSMAX2
Apenas as 3 variáveis	0,00	0,00	0,00	0,00
Uma das variáveis	0,00	0,00	0,00	0,00
Duas das variáveis	1,00	1,00	1,00	1,00
Três das variáveis	0,00	0,00	0,00	0,00
x_1	0,26	0,26	0,00	0,00
x_2	0,63	0,64	0,00	0,00
x_3	0,32	0,32	0,00	0,00
x_4	0,30	0,30	0,30	0,30
x_5	0,89	0,89	0,89	0,89
x_6	0,81	0,81	0,81	0,81

Fonte: Do autor (2024).

Tabela 11 – Média de verdadeiro positivo (VP), falso positivo (FP) e falso negativo (FN) das relações dos algoritmos de RB, para estrutura de associação 2, $n = 150$ e variável resposta contínua.

Algoritmos	VP	FP	FN
HC	5,68	3,31	2,32
Tabu	4,78	4,25	3,22
MMHC	4,99	0,31	3,01
RSMAX2	4,97	0,31	3,03
HC (com wl)	6,02	3,14	1,98
Tabu (com wl)	5,16	4,00	2,84
MMHC (com wl)	5,61	0,30	2,39
RSMAX2 (com wl	5,59	0,30	2,41

Fonte: Do autor (2024).

Tabela 12 – Média dos valores AIC para cada método, considerando estrutura de associação 2, $n = 150$ e variável resposta contínua.

Método	AIC	Método	AIC
Backward	-10245,71	MMHC	70,69
Forward	-10195,01	RSMAX2	70,69
Bidirectional F	-10195,01	HC (com wl)	59,91
Bidirectional B	-10246,03	Tabu (com wl)	59,91
HC	59,91	MMHC (com wl)	70,69
Tabu	59,91	RSMAX2 (com wl	70,69

Fonte: Do autor (2024).

n = 450

Tabela 13 – Taxa de seleção das variáveis para cada método, considerando estrutura de associação 2, $n = 450$ e variável resposta contínua.

Acertos	Backward	Forward	Bidirectional F	Bidirectional B	HC	Tabu	MMHC	RSMAX2
Apenas as 3 variáveis	0,06	0,60	0,60	0,04	0,00	0,00	0,00	0,00
Uma das variáveis	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00
Duas das variáveis	0,00	0,00	0,00	0,00	1,00	1,00	1,00	1,00
Três das variáveis	1,00	1,00	1,00	1,00	0,00	0,00	0,00	0,00
x_1	0,60	0,16	0,16	0,66	0,18	0,18	0,01	0,01
x_2	0,64	0,16	0,16	0,70	0,89	0,89	0,00	0,00
x_3	0,61	0,15	0,15	0,66	0,24	0,24	0,00	0,00
x_4	1,00	1,00	1,00	1,00	0,12	0,12	0,12	0,12
x_5	1,00	1,00	1,00	1,00	0,98	0,98	0,98	0,98
x_6	1,00	1,00	1,00	1,00	0,90	0,90	0,90	0,90

Fonte: Do autor (2024).

Tabela 14 – Taxa de seleção das variáveis para cada algoritmo de rede bayesiana considerando *whitelist*, para estrutura de associação 2, $n = 450$ e variável resposta contínua.

Acertos	HC	Tabu	MMHC	RSMAX2
Apenas as 3 variáveis	0,00	0,00	0,00	0,00
Uma das variáveis	0,00	0,00	0,00	0,00
Duas das variáveis	1,00	1,00	1,00	1,00
Três das variáveis	0,00	0,00	0,00	0,00
x_1	0,18	0,18	0,01	0,01
x_2	0,89	0,89	0,00	0,00
x_3	0,24	0,24	0,00	0,00
x_4	0,12	0,12	0,12	0,12
x_5	0,98	0,98	0,98	0,98
x_6	0,90	0,90	0,90	0,90

Fonte: Do autor (2024).

Tabela 15 – Média de verdadeiro positivo (VP), falso positivo (FP) e falso negativo (FN) das relações dos algoritmos de RB, para estrutura de associação 2, $n = 450$ e variável resposta contínua.

Algoritmos	VP	FP	FN
HC	6,23	3,30	1,77
Tabu	5,20	4,31	2,80
MMHC	5,93	0,30	2,07
RSMAX2	5,90	0,28	2,10
HC (com wl)	6,54	3,03	1,46
Tabu (com wl)	5,51	4,00	2,49
MMHC (com wl)	6,18	0,20	1,82
RSMAX2 (com wl	6,13	0,19	1,87

Fonte: Do autor (2024).

Tabela 16 – Média dos valores AIC para cada método, considerando estrutura de associação 2, $n = 450$ e variável resposta contínua.

Método	AIC	Método	AIC
Backward	-30373,30	MMHC	213,78
Forward	-30193,56	RSMAX2	213,76
Bidirectional F	-30193,56	HC (com wl)	190,49
Bidirectional B	-30373,56	Tabu (com wl)	190,49
HC	190,49	MMHC (com wl)	213,79
Tabu	190,49	RSMAX2 (com wl	213,76

Fonte: Do autor (2024).

Estrutura 3 - n = 150

Variável resposta y_1

Tabela 17 – Taxa de seleção das variáveis para cada método, considerando estrutura de associação 3, $n = 150$ e variável resposta contínua y_1.

Acertos	Backward	Forward	Bidirectional F	Bidirectional B	HC	Tabu	MMHC	RSMAX2
Apenas as 4 variáveis	0,00	0,35	0,35	0,00	0,00	0,00	0,00	0,00
Uma das variáveis	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,03
Duas das variáveis	0,00	0,00	0,00	0,00	0,00	0,00	0,16	0,52
Três das variáveis	0,00	0,00	0,00	0,00	1,00	1,00	0,84	0,44
Quatro das variáveis	1,00	1,00	1,00	1,00	0,00	0,00	0,00	0,00
x1	0,69	0,17	0,17	0,74	0,92	0,92	0,01	0,02
x2	0,68	0,17	0,17	0,73	0,16	0,17	0,00	0,01
x3	0,69	0,16	0,16	0,74	0,22	0,23	0,01	0,00
x4	0,69	0,14	0,14	0,74	0,15	0,15	0,01	0,01
x5	0,70	0,17	0,17	0,75	0,17	0,17	0,00	0,00
x6	1,00	1,00	1,00	1,00	0,10	0,10	0,83	0,13
x7	1,00	1,00	1,00	1,00	1,00	1,00	0,94	0,79
x8	1,00	1,00	1,00	1,00	0,91	0,91	0,06	0,49
x9	1,00	1,00	1,00	1,00	1,00	1,00	1,00	1,00
x10	0,18	0,16	0,16	0,26	0,18	0,18	0,00	0,00

Fonte: Do autor (2024).

Tabela 18 – Taxa de seleção das variáveis para cada algoritmo de rede bayesiana considerando *whitelist*, para estrutura de associação 3, $n = 150$ e variável resposta contínua y_1.

Acertos	HC	Tabu	MMHC	RSMAX2
Apenas as 4 variáveis	0,00	0,00	0,00	0,00
Uma das variáveis	0,00	0,00	0,00	0,03
Duas das variáveis	0,00	0,00	0,16	0,52
Três das variáveis	1,00	1,00	0,84	0,44
Quatro das variáveis	0,00	0,00	0,00	0,00
x1	0,92	0,92	0,01	0,02
x2	0,16	0,17	0,00	0,01
x3	0,22	0,23	0,01	0,00
x4	0,15	0,15	0,01	0,00
x5	0,17	0,17	0,00	0,00
x6	0,10	0,10	0,83	0,13
x7	1,00	1,00	0,94	0,79
x8	0,91	0,91	0,06	0,49
x9	1,00	1,00	1,00	1,00
x10	0,18	0,18	0,00	0,00

Fonte: Do autor (2024).

Tabela 19 – Média de verdadeiro positivo (VP), falso positivo (FP) e falso negativo (FN) das relações dos algoritmos de RB, para estrutura de associação 3, $n = 150$ e variável resposta contínua.

Algoritmos	VP	FP	FN
HC	11,19	11,03	3,81
Tabu	10,56	11,94	4,44
MMHC	9,21	0,74	5,79
RSMAX2	8,76	0,70	6,24
HC (com wl)	11,36	10,92	3,64
Tabu (com wl)	10,97	11,59	4,03
MMHC (com wl)	9,93	0,67	5,07
RSMAX2 (com wl)	9,47	0,66	5,53

Fonte: Do autor (2024).

Tabela 20 – Média dos valores AIC para cada método, para estrutura de associação 3, $n = 150$ e variável resposta contínua y_1.

Método	AIC	Método	AIC
Backward	-10125,50	MMHC	190,08
Forward	-10056,01	RSMAX2	223,12
Bidirectional F	-10056,01	HC (com wl)	74,09
Bidirectional B	-10125,93	Tabu (com wl)	74,08
HC	74,09	MMHC (com wl)	190,18
Tabu	74,09	RSMAX2 (com wl)	223,17

Fonte: Do autor (2024).

Variável resposta y_2

Tabela 21 – Taxa de seleção das variáveis para cada método, considerando estrutura de associação 3, $n = 150$ e variável resposta contínua y_2.

Acertos	Backward	Forward	Bidirectional F	Bidirectional B	HC	Tabu	MMHC	RSMAX2
Apenas as 4 variáveis	0,00	0,37	0,37	0,00	0,00	0,00	0,00	0,00
Uma das variáveis	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,02
Duas das variáveis	0,00	0,00	0,00	0,00	0,00	0,00	0,89	0,86
Três das variáveis	0,00	0,00	0,00	0,00	1,00	1,00	0,11	0,12
Quatro das variáveis	1,00	1,00	1,00	1,00	0,00	0,00	0,00	0,00
x1	0,74	0,14	0,14	0,78	0,14	0,15	0,00	0,00
x2	0,74	0,17	0,17	0,78	1,00	1,00	0,00	0,03
x3	0,73	0,16	0,16	0,78	0,15	0,16	0,01	0,02
x4	0,73	0,15	0,15	0,78	0,15	0,16	0,01	0,00
x5	0,75	0,17	0,17	0,79	0,17	0,17	0,00	0,00
x6	0,74	0,14	0,14	0,78	0,15	0,15	0,00	0,00
x7	1,00	1,00	1,00	1,00	0,00	0,00	0,05	0,00
x8	1,00	1,00	1,00	1,00	1,00	1,00	1,00	0,98
x9	1,00	1,00	1,00	1,00	1,00	1,00	0,06	0,12
x10	1,00	1,00	1,00	1,00	1,00	1,00	1,00	1,00

Fonte: Do autor (2024).

Tabela 22 – Taxa de seleção das variáveis para cada algoritmo de rede bayesiana considerando *whitelist*, para estrutura de associação 3, $n = 150$ e variável resposta contínua y_2.

Acertos	HC	Tabu	MMHC	RSMAX2
Apenas as 4 variáveis	0,00	0,00	0,00	0,00
Uma das variáveis	0,00	0,00	0,00	0,02
Duas das variáveis	0,00	0,00	0,89	0,88
Três das variáveis	1,00	1,00	0,10	0,09
Quatro das variáveis	0,00	0,00	0,00	0,00
x1	0,14	0,15	0,00	0,00
x2	1,00	1,00	0,00	0,03
x3	0,15	0,16	0,01	0,02
x4	0,15	0,16	0,01	0,00
x5	0,17	0,17	0,00	0,00
x6	0,15	0,15	0,00	0,00
x7	0,00	0,00	0,05	0,00
x8	1,00	1,00	1,00	0,98
x9	1,00	1,00	0,06	0,09
x10	1,00	1,00	1,00	1,00

Fonte: Do autor (2024).

Tabela 23 – Média dos valores AIC para cada método, para estrutura de associação 3, $n = 150$ e variável resposta contínua y_2.

Método	AIC	Método	AIC
Backward	-10058,30	MMHC	419,09
Forward	-9978,48	RSMAX2	406,93
Bidirectional F	-10056,01	HC (com wl)	68,97
Bidirectional B	-10125,93	Tabu (com wl)	68,96
HC	68,97	MMHC (com wl)	421,57
Tabu	68,96	RSMAX2 (com wl)	413,59

Fonte: Do autor (2024).

n = 450

Variável resposta y_1

Tabela 24 – Taxa de seleção das variáveis para cada método, considerando estrutura de associação 3, $n = 450$ e variável resposta contínua y_1.

Acertos	Backward	Forward	Bidirectional F	Bidirectional B	HC	Tabu	MMHC	RSMAX2
Apenas as 4 variáveis	0,00	0,36	0,36	0,00	0,00	0,00	0,00	0,00
Uma das variáveis	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00
Duas das variáveis	0,00	0,00	0,00	0,00	0,00	0,00	0,08	0,04
Três das variáveis	0,00	0,00	0,00	0,00	1,00	1,00	0,92	0,96
Quatro das variáveis	1,00	1,00	1,00	1,00	0,00	0,00	0,00	0,00
x1	0,70	0,16	0,16	0,75	1,00	1,00	0,00	0,02
x2	0,71	0,16	0,16	0,76	0,15	0,16	0,00	0,00
x3	0,72	0,14	0,14	0,76	0,14	0,15	0,01	0,00
x4	0,69	0,15	0,15	0,74	0,14	0,14	0,01	0,00
x5	0,73	0,16	0,16	0,77	0,16	0,16	0,00	0,00
x6	1,00	1,00	1,00	1,00	0,00	0,00	0,48	0,21
x7	1,00	1,00	1,00	1,00	1,00	1,00	0,82	0,84
x8	1,00	1,00	1,00	1,00	1,00	1,00	0,61	0,91
x9	1,00	1,00	1,00	1,00	1,00	1,00	1,00	1,00
x10	0,17	0,16	0,16	0,17	0,16	0,16	0,00	0,00

Fonte: Do autor (2024).

Tabela 25 – Taxa de seleção das variáveis para cada algoritmo de rede bayesiana considerando *whitelist*, para estrutura de associação 3, $n = 450$ e variável resposta contínua y_1.

Acertos	HC	Tabu	MMHC	RSMAX2
Apenas as 4 variáveis	0,00	0,00	0,00	0,00
Uma das variáveis	0,00	0,00	0,00	0,00
Duas das variáveis	0,00	0,00	0,08	0,04
Três das variáveis	1,00	1,00	0,92	0,96
Quatro das variáveis	0,00	0,00	0,00	0,00
x1	1,00	1,00	0,00	0,02
x2	0,15	0,16	0,00	0,00
x3	0,14	0,15	0,01	0,00
x4	0,14	0,14	0,01	0,00
x5	0,16	0,16	0,00	0,00
x6	0,00	0,00	0,48	0,21
x7	1,00	1,00	0,82	0,84
x8	1,00	1,00	0,61	0,91
x9	1,00	1,00	1,00	1,00
x10	0,16	0,16	0,00	0,00

Fonte: Do autor (2024).

Tabela 26 – Média de verdadeiro positivo (VP), falso positivo (FP) e falso negativo (FN) das relações dos algoritmos de RB, para estrutura de associação 3, $n = 450$ e variável resposta contínua.

Algoritmos	VP	FP	FN
HC	11,49	10,88	3,51
Tabu	10,99	11,65	4,01
MMHC	10,76	1,26	4,24
RSMAX2	10,58	1,14	4,42
HC (com wl)	11,58	10,81	3,42
Tabu (com wl)	11,28	11,36	3,72
MMHC (com wl)	11,05	1,23	3,95
RSMAX2 (com wl)	10,87	1,14	4,13

Fonte: Do autor (2024).

Tabela 27 – Média dos valores AIC para cada método, para estrutura de associação 3, $n = 450$ e variável resposta contínua y_1.

Método	AIC	Método	AIC
Backward	-30040,76	MMHC	518,68
Forward	-29764,64	RSMAX2	476,35
Bidirectional F	-29764,64	HC (com wl)	199,16
Bidirectional B	-30041,14	Tabu (com wl)	199,15
HC	199,16	MMHC (com wl)	518,72
Tabu	199,15	RSMAX2 (com wl)	476,35

Fonte: Do autor (2024).

Variável resposta y_2

Tabela 28 – Taxa de seleção das variáveis para cada método, considerando estrutura de associação 3, $n = 450$ e variável resposta contínua y_2.

Acertos	Backward	Forward	Bidirectional F	Bidirectional B	HC	Tabu	MMHC	RSMAX2
Apenas as 4 variáveis	0,00	0,38	0,38	0,00	0,00	0,00	0,00	0,00
Uma das variáveis	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00
Duas das variáveis	0,00	0,00	0,00	0,00	0,00	0,00	0,52	0,76
Três das variáveis	0,00	0,00	0,00	0,00	1,00	1,00	0,48	0,24
Quatro das variáveis	1,00	1,00	1,00	1,00	0,00	0,00	0,00	0,00
x1	0,76	0,14	0,14	0,80	0,14	0,15	0,00	0,00
x2	0,76	0,16	0,16	0,80	1,00	1,00	0,00	0,02
x3	0,74	0,15	0,15	0,78	0,14	0,14	0,01	0,01
x4	0,74	0,14	0,14	0,78	0,15	0,15	0,02	0,00
x5	0,76	0,16	0,16	0,80	0,17	0,17	0,00	0,00
x6	0,77	0,14	0,14	0,81	0,14	0,14	0,00	0,00
x7	1,00	1,00	1,00	1,00	0,00	0,00	0,39	0,03
x8	1,00	1,00	1,00	1,00	1,00	1,00	1,00	1,00
x9	1,00	1,00	1,00	1,00	1,00	1,00	0,09	0,21
x10	1,00	1,00	1,00	1,00	1,00	1,00	1,00	1,00

Fonte: Do autor (2024).

Tabela 29 – Taxa de seleção das variáveis para cada algoritmo de rede bayesiana considerando *whitelist*, para estrutura de associação 3, $n = 450$ e variável resposta contínua y_2.

Acertos	HC	Tabu	MMHC	RSMAX2
Apenas as 4 variáveis	0,00	0,00	0,00	0,00
Uma das variáveis	0,00	0,00	0,00	0,00
Duas das variáveis	0,00	0,00	0,53	0,78
Três das variáveis	1,00	1,00	0,47	0,22
Quatro das variáveis	0,00	0,00	0,00	0,00
x1	0,14	0,15	0,00	0,00
x2	1,00	1,00	0,00	0,02
x3	0,14	0,15	0,01	0,01
x4	0,15	0,15	0,02	0,00
x5	0,17	0,17	0,00	0,00
x6	0,14	0,14	0,00	0,00
x7	0,00	0,00	0,39	0,03
x8	1,00	1,00	1,00	1,00
x9	1,00	1,00	0,08	0,20
x10	1,00	1,00	1,00	1,00

Fonte: Do autor (2024).

Tabela 30 – Média dos valores AIC para cada método, para estrutura de associação 3, $n = 450$ e variável resposta contínua y_2.

Método	AIC	Método	AIC
Backward	-29864,93	MMHC	1204,82
Forward	-29531,30	RSMAX2	1112,94
Bidirectional F	-29764,64	HC (com wl)	197,41
Bidirectional B	-30041,14	Tabu (com wl)	197,40
HC	197,41	MMHC (com wl)	1210,52
Tabu	197,40	RSMAX2 (com wl)	1126,50

Fonte: Do autor (2024).

Caso de resposta binária

Estrutura de associação 1

n = 50

Váriaveis explicativas com duas categorias

Tabela 31 – Taxa de seleção das variáveis para cada método, considerando estrutura de associação 1, $n = 50$, VE2 e variável resposta binária.

Acertos	Backward	Forward	Bidirectional F	Bidirectional B	HC	Tabu	MMHC	RSMAX2
Apenas as 6 variáveis	0,01	0,01	0,01	0,01	0,00	0,00	0,00	0,00
Uma das variáveis	0,09	0,11	0,11	0,09	0,43	0,42	0,57	0,57
Duas das variáveis	0,29	0,30	0,30	0,29	0,37	0,37	0,38	0,38
Três das variáveis	0,34	0,33	0,34	0,34	0,16	0,17	0,05	0,05
Quatro das variáveis	0,21	0,20	0,20	0,21	0,03	0,04	0,00	0,00
Cinco das variáveis	0,06	0,05	0,05	0,06	0,00	0,00	0,00	0,00
x1	0,74	0,73	0,73	0,74	0,53	0,54	0,45	0,45
x2	0,25	0,24	0,24	0,25	0,09	0,10	0,06	0,06
x3	0,47	0,45	0,45	0,47	0,24	0,25	0,18	0,18
x4	0,23	0,21	0,21	0,23	0,08	0,08	0,04	0,04
x5	0,39	0,37	0,37	0,39	0,17	0,18	0,12	0,12
x6	0,82	0,80	0,80	0,82	0,61	0,63	0,54	0,54

Fonte: Do autor (2024).

Tabela 32 – Taxa de seleção das variáveis para cada algoritmo de rede bayesiana considerando *whitelist*, para estrutura de associação 1, $n = 50$, VE2 e variável resposta binária.

Acertos	HC	Tabu	MMHC	RSMAX2
Apenas as 6 variáveis	0,00	0,00	0,00	0,00
Uma das variáveis	0,12	0,11	0,17	0,17
Duas das variáveis	0,40	0,39	0,56	0,56
Três das variáveis	0,36	0,37	0,26	0,26
Quatro das variáveis	0,10	0,11	0,02	0,02
Cinco das variáveis	0,01	0,01	0,00	0,00
x1	0,52	0,53	0,41	0,42
x2	0,09	0,10	0,05	0,05
x3	1,00	1,00	1,00	1,00
x4	0,07	0,08	0,03	0,04
x5	0,18	0,19	0,11	0,10
x6	0,62	0,63	0,53	0,52

Fonte: Do autor (2024).

Tabela 33 – Média de verdadeiro positivo (VP), falso positivo (FP) e falso negativo (FN) das relações dos algoritmos de RB, considerando estrutura de associação 1, $n = 50$, VE2 e variável resposta binária

Algoritmos	VP	FP	FN
HC	1,72	1,32	4,28
Tabu	1,76	1,33	4,24
MMHC	1,39	0,84	4,61
RSMAX2	1,39	0,84	4,61
HC (com wl)	2,49	1,32	3,51
Tabu (com wl)	2,52	1,33	3,48
MMHC (com wl)	2,13	0,76	3,87
RSMAX2 (com wl	2,13	0,77	3,87

Fonte: Do autor (2024).

Tabela 34 – Média dos valores AIC para cada método, considerando estrutura de associação 1, $n = 50$, VE2 e variável resposta binária

Método	AIC	Método	AIC
Backward	59,92	MMHC	66,69
Forward	59,93	RSMAX2	66,69
Bidirectional F	59,93	HC (com wl)	66,08
Bidirectional B	59,92	Tabu (com wl)	66,00
HC	65,58	MMHC (com wl)	67,29
Tabu	65,48	RSMAX2 (com wl	67,30

Fonte: Do autor (2024).

Váriaveis explicativas com três categorias

Tabela 35 – Taxa de seleção das variáveis para cada método, considerando estrutura de associação 1, $n = 50$, VE3 e variável resposta binária.

Acertos	Backward	Forward	Bidirectional F	Bidirectional B	HC	Tabu	MMHC	RSMAX2
Apenas as 6 variáveis	0,02	0,01	0,01	0,02	0,00	0,00	0,00	0,00
Uma das variáveis	0,04	0,05	0,05	0,04	0,30	0,28	0,47	0,47
Duas das variáveis	0,22	0,24	0,24	0,22	0,41	0,41	0,45	0,45
Três das variáveis	0,36	0,36	0,36	0,36	0,23	0,24	0,08	0,08
Quatro das variáveis	0,27	0,25	0,25	0,27	0,06	0,06	0,00	0,00
Cinco das variáveis	0,09	0,08	0,08	0,09	0,01	0,01	0,00	0,00
x1	0,83	0,82	0,82	0,83	0,63	0,64	0,53	0,53
x2	0,28	0,26	0,26	0,28	0,10	0,11	0,05	0,06
x3	0,54	0,52	0,52	0,54	0,29	0,30	0,19	0,19
x4	0,23	0,21	0,21	0,23	0,07	0,08	0,04	0,04
x5	0,43	0,42	0,42	0,43	0,21	0,22	0,13	0,13
x6	0,89	0,88	0,88	0,89	0,71	0,73	0,63	0,63

Fonte: Do autor (2024).

Tabela 36 – Taxa de seleção das variáveis para cada algoritmo de rede bayesiana considerando *whitelist*, considerando estrutura de associação 1, $n = 50$, VE3 e variável resposta binária.

Acertos	HC	Tabu	MMHC	RSMAX2
Apenas as 6 variáveis	0,00	0,00	0,00	0,00
Uma das variáveis	0,00	0,06	0,10	0,10
Duas das variáveis	0,32	0,31	0,52	0,52
Três das variáveis	0,44	0,44	0,36	0,36
Quatro das variáveis	0,15	0,16	0,03	0,03
Cinco das variáveis	0,02	0,02	0,00	0,00
x1	0,64	0,65	0,50	0,50
x2	0,11	0,11	0,05	0,05
x3	1,00	1,00	1,00	1,00
x4	0,07	0,08	0,03	0,03
x5	0,21	0,22	0,12	0,12
x6	0,73	0,73	0,62	0,61

Fonte: Do autor (2024).

Tabela 37 – Média de verdadeiro positivo (VP), falso positivo (FP) e falso negativo (FN) das relações dos algoritmos de RB, considerando estrutura de associação 1, $n = 50$, VE3 e variável resposta binária.

Algoritmos	VP	FP	FN
HC	2,01	0,80	3,99
Tabu	2,08	0,87	3,92
MMHC	1,57	0,56	4,43
RSMAX2	1,57	0,56	4,43
HC (com wl)	2,75	0,80	3,25
Tabu (com wl)	2,79	0,87	3,21
MMHC (com wl)	2,32	0,54	3,68
RSMAX2 (com wl	2,32	0,55	3,68

Fonte: Do autor (2024).

Tabela 38 – Média dos valores AIC para cada método, considerando estrutura de associação 1, $n = 50$, VE3 e variável resposta binária

Método	AIC	Método	AIC
Backward	57,07	MMHC	64,47
Forward	57,08	RSMAX2	64,47
Bidirectional F	57,08	HC (com wl)	63,20
Bidirectional B	57,07	Tabu (com wl)	63,11
HC	62,88	MMHC (com wl)	64,77
Tabu	62,73	RSMAX2 (com wl	64,77

Fonte: Do autor (2024).

Váriaveis explicativas com quatro categorias

Tabela 39 – Taxa de seleção das variáveis para cada método, considerando estrutura de associação 1, $n = 50$, VE4 e variável resposta binária.

Acertos	Backward	Forward	Bidirectional F	Bidirectional B	HC	Tabu	MMHC	RSMAX2
Apenas as 6 variáveis	0,01	0,01	0,01	0,01	0,00	0,00	0,00	0,00
Uma das variáveis	0,04	0,05	0,05	0,04	0,28	0,25	0,47	0,47
Duas das variáveis	0,20	0,22	0,22	0,20	0,40	0,40	0,46	0,46
Três das variáveis	0,36	0,37	0,37	0,36	0,24	0,25	0,07	0,07
Quatro das variáveis	0,27	0,25	0,25	0,11	0,07	0,08	0,00	0,00
Cinco das variáveis	0,11	0,10	0,10	0,11	0,01	0,01	0,00	0,00
x1	0,89	0,87	0,87	0,89	0,68	0,70	0,55	0,55
x2	0,27	0,26	0,26	0,27	0,10	0,11	0,05	0,05
x3	0,52	0,49	0,49	0,52	0,25	0,27	0,11	0,11
x4	0,23	0,22	0,22	0,23	0,08	0,08	0,04	0,04
x5	0,45	0,43	0,43	0,45	0,23	0,23	0,14	0,14
x6	0,90	0,89	0,89	0,90	0,75	0,76	0,66	0,66

Fonte: Do autor (2024).

Tabela 40 – Taxa de seleção das variáveis para cada algoritmo de rede bayesiana considerando *whitelist*, para estrutura de associação 1, $n = 50$, VE4 e variável resposta binária.

Acertos	HC	Tabu	MMHC	RSMAX2
Apenas as 6 variáveis	0,00	0,00	0,00	0,00
Uma das variáveis	0,04	0,04	0,08	0,08
Duas das variáveis	0,26	0,26	0,48	0,49
Três das variáveis	0,47	0,48	0,40	0,40
Quatro das variáveis	0,19	0,19	0,03	0,03
Cinco das variáveis	0,03	0,03	0,00	0,00
x1	0,72	0,73	0,54	0,54
x2	0,11	0,11	0,05	0,05
x3	1,00	1,00	1,00	1,00
x4	0,08	0,08	0,03	0,03
x5	0,23	0,24	0,12	0,13
x6	0,76	0,77	0,64	0,64

Fonte: Do autor (2024).

Tabela 41 – Média de verdadeiro positivo (VP), falso positivo (FP) e falso negativo (FN) das relações dos algoritmos de RB, considerando estrutura de associação 1, $n = 50$, VE4 e variável resposta binária.

Algoritmos	VP	FP	FN
HC	2,09	0,95	3,91
Tabu	2,16	1,02	3,84
MMHC	1,55	0,66	4,45
RSMAX2	1,55	0,66	4,45
HC (com wl)	2,90	0,95	3,10
Tabu (com wl)	2,93	1,03	3,07
MMHC (com wl)	2,39	0,64	3,61
RSMAX2 (com wl	2,39	0,64	3,61

Fonte: Do autor (2024).

Tabela 42 – Média dos valores AIC para cada método, considerando estrutura de associação 1, $n = 50$, VE4 e variável resposta binária.

Método	AIC	Método	AIC
Backward	56,25	MMHC	64,18
Forward	56,27	RSMAX2	64,19
Bidirectional F	56,26	HC (com wl)	62,46
Bidirectional B	56,25	Tabu (com wl)	62,38
HC	62,15	MMHC (com wl)	64,42
Tabu	61,98	RSMAX2 (com wl	64,42

Fonte: Do autor (2024).

n = 150

Variáveis explicativas com duas categorias

Tabela 43 – Taxa de seleção das variáveis para cada método, considerando estrutura de associação 1, $n = 150$, VE2 e variável resposta binária.

Acertos	Backward	Forward	Bidirectional F	Bidirectional B	HC	Tabu	MMHC	RSMAX2
Apenas as 6 variáveis	0,03	0,03	0,03	0,03	0,00	0,00	0,00	0,00
Uma das variáveis	0,00	0,00	0,00	0,00	0,04	0,03	0,05	0,05
Duas das variáveis	0,04	0,04	0,04	0,04	0,32	0,32	0,42	0,42
Três das variáveis	0,24	0,25	0,25	0,24	0,45	0,44	0,42	0,42
Quatro da variáveis	0,45	0,45	0,45	0,45	0,17	0,18	0,10	0,10
Cinco das variáveis	0,23	0,23	0,23	0,23	0,02	0,02	0,01	0,01
x1	0,98	0,98	0,98	0,98	0,90	0,91	0,87	0,87
x2	0,33	0,32	0,32	0,33	0,10	0,10	0,07	0,07
x3	0,80	0,80	0,80	0,80	0,49	0,49	0,41	0,41
x4	0,21	0,20	0,20	0,21	0,04	0,04	0,03	0,03
x5	0,66	0,66	0,66	0,66	0,33	0,34	0,26	0,26
x6	0,99	0,99	0,99	0,99	0,96	0,96	0,95	0,95

Fonte: Do autor (2024).

Tabela 44 – Taxa de seleção das variáveis para cada algoritmo de rede bayesiana considerando *whitelist*, para estrutura de associação 1, $n = 150$, VE2 e variável resposta binária.

Acertos	HC	Tabu	MMHC	RSMAX2
Apenas as 6 variáveis	0,00	0,00	0,00	0,00
Uma das variáveis	0,00	0,00	0,00	0,00
Duas das variáveis	0,07	0,06	0,10	0,10
Três das variáveis	0,55	0,55	0,63	0,63
Quatro das variáveis	0,33	0,33	0,25	0,25
Cinco das variáveis	0,04	0,04	0,02	0,02
x1	0,91	0,91	0,87	0,87
x2	0,10	0,10	0,07	0,07
x3	1,00	1,00	1,00	1,00
x4	0,04	0,04	0,03	0,03
x5	0,34	0,34	0,26	0,26
x6	0,96	0,97	0,95	0,95

Fonte: Do autor (2024).

Tabela 45 – Média de verdadeiro positivo(VP), falso positivo(FP) e falso negativo(FN) das relações dos algoritmos de RB, considerando estrutura de associação 1, $n = 150$, VE2 e variável resposta binária.

Algoritmos	VP	FP	FN
HC	2,83	0,33	3,17
Tabu	2,84	0,34	3,16
MMHC	2,59	0,29	3,41
RSMAX2	2,59	0,29	3,41
HC (com wl)	3,35	0,33	2,65
Tabu (com wl)	3,36	0,34	2,64
MMHC (com wl)	3,18	0,29	2,82
RSMAX2 (com wl	3,18	0,29	2,82

Fonte: Do autor (2024).

Tabela 46 – Média dos valores AIC para cada método, considerando estrutura de associação 1, $n = 150$, VE2 e variável resposta binária.

Método	AIC	Método	AIC
Backward	177,11	MMHC	190,41
Forward	177,11	RSMAX2	190,42
Bidirectional F	177,11	HC (com wl)	189,08
Bidirectional B	177,11	Tabu (com wl)	189,06
HC	189,36	MMHC (com wl)	189,85
Tabu	189,32	RSMAX2 (com wl	189,86

Fonte: Do autor (2024).

Variáveis explicativas com três categorias

Tabela 47 – Taxa de seleção das variáveis para cada método, considerando estrutura de associação 1, $n = 150$, VE3 e variável resposta binária.

Acertos	Backward	Forward	Bidirectional F	Bidirectional B	HC	Tabu	MMHC	RSMAX2
Apenas as 6 variáveis	0,06	0,06	0,06	0,06	0,00	0,00	0,00	0,00
Uma das variáveis	0,00	0,00	0,00	0,00	0,01	0,01	0,01	0,01
Duas das variáveis	0,02	0,02	0,02	0,02	0,20	0,20	0,30	0,30
Três das variáveis	0,17	0,17	0,17	0,17	0,46	0,46	0,51	0,51
Quatro das variáveis	0,45	0,46	0,46	0,45	0,29	0,29	0,18	0,18
Cinco da variáveis	0,30	0,30	0,30	0,30	0,04	0,04	0,01	0,01
x1	1,00	1,00	1,00	1,00	0,97	0,97	0,95	0,95
x2	0,37	0,37	0,37	0,37	0,11	0,11	0,07	0,07
x3	0,88	0,88	0,88	0,88	0,62	0,62	0,51	0,51
x4	0,23	0,22	0,22	0,23	0,05	0,05	0,03	0,03
x5	0,74	0,74	0,74	0,74	0,43	0,43	0,33	0,33
x6	1,00	1,00	1,00	1,00	0,99	0,99	0,99	0,99

Fonte: Do autor (2024).

Tabela 48 – Taxa de seleção das variáveis para cada algoritmo de rede bayesiana considerando *whitelist*, para estrutura de associação 1, $n = 150$, VE3 e variável resposta binária.

Acertos	HC	Tabu	MMHC	RSMAX2
Apenas as 6 variáveis	0,00	0,00	0,00	0,00
Uma das variáveis	0,00	0,00	0,00	0,00
Duas das variáveis	0,02	0,02	0,03	0,03
Três das variáveis	0,47	0,47	0,60	0,60
Quatro das variáveis	0,44	0,44	0,35	0,35
Cinco das variáveis	0,07	0,07	0,02	0,02
x1	0,97	0,97	0,95	0,95
x2	0,12	0,12	0,07	0,07
x3	1,00	1,00	1,00	1,00
x4	0,05	0,05	0,02	0,03
x5	0,44	0,44	0,33	0,33
x6	0,99	0,99	0,99	0,99

Fonte: Do autor (2024).

Tabela 49 – Média de verdadeiro positivo(VP), falso positivo(FP) e falso negativo(FN) das relações dos algoritmos de RB, considerando estrutura de associação 1, $n = 150$, VE3 e variável resposta binária.

Algoritmos	VP	FP	FN
HC	3,17	0,35	2,83
Tabu	3,18	0,36	2,82
MMHC	2,88	0,28	3,12
RSMAX2	2,88	0,28	3,12
HC (com wl)	3,57	0,35	2,43
Tabu (com wl)	3,57	0,36	2,43
MMHC (com wl)	3,36	0,28	2,64
RSMAX2 (com wl	3,36	0,28	2,64

Fonte: Do autor (2024).

Tabela 50 – Média dos valores AIC para cada método, considerando estrutura de associação 1, $n = 150$, VE3 e variável resposta binária.

Método	AIC	Método	AIC
Backward	167,63	MMHC	181,10
Forward	167,63	RSMAX2	181,11
Bidirectional F	167,63	HC (com wl)	179,40
Bidirectional B	167,63	Tabu (com wl)	179,38
HC	179,73	MMHC (com wl)	180,36
Tabu	179,70	RSMAX2 (com wl	180,36

Fonte: Do autor (2024).

Variáveis explicativas com quatro categorias

Tabela 51 – Taxa de seleção das variáveis para cada método, considerando estrutura de associação 1, $n = 150$, VE4 e variável resposta binária.

Acertos	Backward	Forward	Bidirectional F	Bidirectional B	HC	Tabu	MMHC	RSMAX2
Apenas as 6 variáveis	0,06	0,06	0,06	0,06	0,00	0,00	0,00	0,00
Uma das variáveis	0,00	0,00	0,00	0,00	0,01	0,01	0,01	0,01
Duas das variáveis	0,01	0,01	0,01	0,01	0,19	0,19	0,40	0,40
Três das variáveis	0,16	0,16	0,16	0,16	0,45	0,45	0,46	0,46
Quatro das variáveis	0,44	0,45	0,45	0,32	0,31	0,31	0,12	0,12
Cinco das variáveis	0,32	0,32	0,32	0,32	0,05	0,05	0,01	0,01
x1	1,00	1,00	1,00	1,00	0,98	0,98	0,96	0,96
x2	0,39	0,39	0,39	0,39	0,13	0,13	0,08	0,08
x3	0,87	0,86	0,86	0,87	0,59	0,59	0,30	0,30
x4	0,23	0,22	0,22	0,23	0,05	0,05	0,03	0,03
x5	0,78	0,78	0,78	0,78	0,47	0,47	0,35	0,35
x6	1,00	1,00	1,00	1,00	1,00	1,00	0,99	0,99

Fonte: Do autor (2024).

Tabela 52 – Taxa de seleção das variáveis para cada algoritmo de rede bayesiana considerando *whitelist*, para estrutura de associação 1, $n = 150$, VE4 e variável resposta binária.

Acertos	HC	Tabu	MMHC	RSMAX2
Apenas as 6 variáveis	0,00	0,00	0,00	0,00
Uma das variáveis	0,00	0,00	0,00	0,00
Duas das variáveis	0,01	0,01	0,02	0,02
Três das variáveis	0,44	0,43	0,58	0,58
Quatro das variáveis	0,48	0,48	0,37	0,37
Cinco das variáveis	0,08	0,08	0,03	0,03
x1	0,99	0,99	0,96	0,96
x2	0,13	0,13	0,08	0,08
x3	1,00	1,00	1,00	1,00
x4	0,05	0,05	0,03	0,03
x5	0,47	0,47	0,35	0,35
x6	1,00	1,00	0,99	0,99

Fonte: Do autor (2024).

Tabela 53 – Média de verdadeiro positivo(VP), falso positivo(FP) e falso negativo(FN) das relações dos algoritmos de RB, considerando estrutura de associação 1, $n = 150$, VE4 e variável resposta binária.

Algoritmos	VP	FP	FN
HC	3,21	0,69	2,79
Tabu	3,22	0,70	2,78
MMHC	2,71	0,61	3,29
RSMAX2	2,71	0,61	3,29
HC (com wl)	3,64	0,69	2,36
Tabu (com wl)	3,64	0,70	2,36
MMHC (com wl)	3,41	0,60	2,59
RSMAX2 (com wl	3,41	0,60	2,59

Fonte: Do autor (2024).

Tabela 54 – Média dos valores AIC para cada método, considerando estrutura de associação 1, $n = 150$, VE4 e variável resposta binária.

Método	AIC	Método	AIC
Backward	164,72	MMHC	179,54
Forward	164,72	RSMAX2	179,54
Bidirectional F	164,72	HC (com wl)	176,61
Bidirectional B	164,72	Tabu (com wl)	176,61
HC	176,93	MMHC (com wl)	177,70
Tabu	176,89	RSMAX2 (com wl	177,70

Fonte: Do autor (2024).

n = 450

Variáveis explicativas com duas categorias

Tabela 55 – Taxa de seleção das variáveis para cada método, considerando estrutura de associação 1, $n = 450$, VE2 e variável resposta binária.

Acertos	Backward	Forward	Bidirectional F	Bidirectional B	HC	Tabu	MMHC	RSMAX2
Apenas as 6 variáveis	0,15	0,15	0,15	0,15	0,01	0,01	0,00	0,00
Uma das variáveis	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00
Duas das variáveis	0,00	0,00	0,00	0,00	0,02	0,02	0,03	0,03
Três das variáveis	0,01	0,01	0,01	0,01	0,23	0,23	0,28	0,28
Quatro das variáveis	0,32	0,32	0,32	0,32	0,60	0,60	0,58	0,58
Cinco	0,52	0,52	0,52	0,52	0,15	0,15	0,12	0,12
x1	1,00	1,00	1,00	1,00	1,00	1,00	1,00	1,00
x2	0,56	0,56	0,56	0,56	0,18	0,18	0,15	0,15
x3	0,99	0,99	0,99	0,99	0,92	0,92	0,90	0,90
x4	0,29	0,29	0,29	0,29	0,05	0,05	0,04	0,04
x5	0,96	0,96	0,96	0,96	0,75	0,75	0,70	0,70
x6	1,00	1,00	1,00	1,00	1,00	1,00	1,00	1,00

Fonte: Do autor (2024).

Tabela 56 – Taxa de seleção das variáveis para cada algoritmo de rede bayesiana considerando *whitelist*, para estrutura de associação 1, $n = 450$, VE2 e variável resposta binária.

Acertos	HC	Tabu	MMHC	RSMAX2
Apenas as 6 variáveis	0,01	0,01	0,00	0,00
Uma das variáveis	0,00	0,00	0,00	0,00
Duas das variáveis	0,00	0,00	0,00	0,00
Três das variáveis	0,20	0,20	0,24	0,24
Quatro da variáveis	0,63	0,63	0,63	0,63
Cinco das variáveis	0,17	0,17	0,13	0,13
x1	1,00	1,00	1,00	1,00
x2	0,18	0,18	0,15	0,15
x3	1,00	1,00	1,00	1,00
x4	0,05	0,05	0,04	0,04
x5	0,75	0,75	0,70	0,70
x6	1,00	1,00	1,00	1,00

Fonte: Do autor (2024).

Tabela 57 – Média de verdadeiro positivo(VP), falso positivo(FP) e falso negativo(FN) das relações dos algoritmos de RB, considerando estrutura de associação 1, $n = 450$, VE2 e variável resposta binária.

Algoritmos	VP	FP	FN
HC	3,90	0,22	2,10
Tabu	3,90	0,22	2,10
MMHC	3,79	0,18	2,21
RSMAX2	3,79	0,18	2,21
HC (com wl)	3,98	0,22	2,02
Tabu (com wl)	3,98	0,22	2,02
MMHC (com wl)	3,89	0,18	2,11
RSMAX2 (com wl	3,89	0,18	2,11

Fonte: Do autor (2024).

Tabela 58 – Média dos valores AIC para cada método, considerando estrutura de associação 1, $n = 450$, VE2 e variável resposta binária.

Método	AIC	Método	AIC
Backward	523,27	MMHC	554,16
Forward	523,27	RSMAX2	554,17
Bidirectional F	523,27	HC (com wl)	553,39
Bidirectional B	523,27	Tabu (com wl)	553,39
HC	553,57	MMHC (com wl)	553,87
Tabu	553,56	RSMAX2 (com wl	553,87

Fonte: Do autor (2024).

Variáveis explicativas com três categorias

Tabela 59 – Taxa de seleção das variáveis para cada método, considerando estrutura de associação 1, $n = 450$, VE3 e variável resposta binária.

Acertos	Backward	Forward	Bidirectional F	Bidirectional B	HC	Tabu	MMHC	RSMAX2
Apenas as 6 variáveis	0,21	0,21	0,21	0,21	0,01	0,01	0,00	0,00
Uma das variáveis	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00
Duas das variáveis	0,00	0,00	0,00	0,00	0,00	0,00	0,01	0,01
Três das variáveis	0,00	0,00	0,00	0,00	0,11	0,11	0,16	0,16
Quatro das variáveis	0,24	0,24	0,24	0,24	0,63	0,63	0,66	0,66
Cinco das variáveis	0,55	0,55	0,55	0,55	0,24	0,24	0,17	0,17
x1	1,00	1,00	1,00	1,00	1,00	1,00	1,00	1,00
x2	0,65	0,65	0,65	0,65	0,24	0,25	0,19	0,19
x3	1,00	1,00	1,00	1,00	0,98	0,98	0,96	0,96
x4	0,33	0,33	0,33	0,33	0,06	0,06	0,04	0,04
x5	0,99	0,99	0,99	0,99	0,86	0,86	0,81	0,81
x6	1,00	1,00	1,00	1,00	1,00	1,00	1,00	1,00

Fonte: Do autor (2024).

Tabela 60 – Taxa de seleção das variáveis para cada algoritmo de rede bayesiana considerando *whitelist*, para estrutura de associação 1, $n = 450$, VE3 e variável resposta binária.

Acertos	HC	Tabu	MMHC	RSMAX2
Apenas as 6 variáveis	0,01	0,01	0,00	0,00
Uma das variáveis	0,00	0,00	0,00	0,00
Duas das variáveis	0,00	0,00	0,00	0,00
Três das variáveis	0,10	0,10	0,14	0,14
Quatro das variáveis	0,65	0,65	0,68	0,68
Cinco da variávei	0,25	0,25	0,18	0,18
x1	1,00	1,00	1,00	1,00
x2	0,25	0,25	0,19	0,19
x3	1,00	1,00	1,00	1,00
x4	0,06	0,06	0,04	0,04
x5	0,86	0,86	0,81	0,81
x6	1,00	1,00	1,00	1,00

Fonte: Do autor (2024).

Tabela 61 – Média de verdadeiro positivo(VP), falso positivo(FP) e falso negativo(FN) das relações dos algoritmos de RB, considerando estrutura de associação 1, $n = 450$, VE3 e variável resposta binária.

Algoritmos	VP	FP	FN
HC	4,15	0,22	1,85
Tabu	4,15	0,22	1,85
MMHC	4,01	0,18	1,99
RSMAX2	4,01	0,18	1,99
HC (com wl)	4,17	0,22	1,83
Tabu (com wl)	4,17	0,22	1,83
MMHC (com wl)	4,05	0,18	1,95
RSMAX2 (com wl	4,05	0,18	1,95

Fonte: Do autor (2024).

Tabela 62 – Média dos valores AIC para cada método, considerando estrutura de associação 1, $n = 450$, VE3 e variável resposta binária.

Método	AIC	Método	AIC
Backward	493,97	MMHC	524,77
Forward	493,97	RSMAX2	524,77
Bidirectional F	493,97	HC (com wl)	523,92
Bidirectional B	493,97	Tabu (com wl)	523,92
HC	523,98	MMHC (com wl)	524,62
Tabu	523,98	RSMAX2 (com wl	524,62

Fonte: Do autor (2024).

Variáveis explicativas com quatro categorias

Tabela 63 – Taxa de seleção das variáveis para cada método, considerando estrutura de associação 1, $n = 450$, VE4 e variável resposta binária.

Acertos	Backward	Forward	Bidirectional F	Bidirectional B	HC	Tabu	MMHC	RSMAX2
Apenas as 6 variáveis	0,23	0,23	0,23	0,23	0,01	0,01	0,00	0,00
Uma das variáveis	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00
Duas das variáveis	0,00	0,00	0,00	0,00	0,00	0,00	0,03	0,03
Três das variáveis	0,00	0,00	0,00	0,00	0,09	0,09	0,26	0,26
Quatro das variáveis	0,21	0,21	0,21	0,56	0,61	0,61	0,56	0,56
Cinco da variáveis	0,56	0,56	0,56	0,56	0,28	0,28	0,15	0,15
x1	1,00	1,00	1,00	1,00	1,00	1,00	1,00	1,00
x2	0,68	0,68	0,68	0,68	0,28	0,28	0,21	0,21
x3	1,00	1,00	1,00	1,00	0,97	0,97	0,74	0,74
x4	0,34	0,34	0,34	0,34	0,07	0,07	0,04	0,04
x5	0,99	0,99	0,99	0,99	0,90	0,90	0,85	0,85
x6	1,00	1,00	1,00	1,00	1,00	1,00	1,00	1,00

Fonte: Do autor (2024).

Tabela 64 – Taxa de seleção das variáveis para cada algoritmo de rede bayesiana considerando *whitelist*, para estrutura de associação 1, $n = 450$, VE3 e variável resposta binária.

Acertos	HC	Tabu	MMHC	RSMAX2
Apenas as 6 variáveis	0,01	0,01	0,01	0,01
Uma das variáveis	0,00	0,00	0,00	0,00
Duas das variáveis	0,00	0,00	0,00	0,00
Três das variáveis	0,07	0,07	0,12	0,12
Quatro das variáveis	0,63	0,63	0,67	0,67
Cinco das variáveis	0,29	0,29	0,20	0,20
x1	1,00	1,00	1,00	1,00
x2	0,28	0,28	0,21	0,21
x3	1,00	1,00	1,00	1,00
x4	0,07	0,07	0,04	0,04
x5	0,90	0,90	0,85	0,85
x6	1,00	1,00	1,00	1,00

Fonte: Do autor (2024).

Tabela 65 – Média de verdadeiro positivo(VP), falso positivo(FP) e falso negativo(FN) das relações dos algoritmos de RB, considerando estrutura de associação 1, $n = 450$, VE4 e variável resposta binária.

Algoritmos	VP	FP	FN
HC	4,21	0,89	1,79
Tabu	4,22	0,90	1,78
MMHC	3,85	0,85	2,15
RSMAX2	3,85	0,85	2,15
HC (com wl)	4,24	0,89	1,76
Tabu (com wl)	4,24	0,90	1,76
MMHC (com wl)	4,10	0,84	1,90
RSMAX2 (com wl	4,10	0,85	1,90

Fonte: Do autor (2024).

Tabela 66 – Média dos valores AIC para cada método, considerando estrutura de associação 1, $n = 450$, VE4 e variável resposta binária.

Método	AIC	Método	AIC
Backward	485,77	MMHC	518,91
Forward	485,77	RSMAX2	518,92
Bidirectional F	485,77	HC (com wl)	515,89
Bidirectional B	485,77	Tabu (com wl)	515,89
HC	515,97	MMHC (com wl)	516,71
Tabu	515,96	RSMAX2 (com wl	516,71

Fonte: Do autor (2024).

Estrutura de associação 2

n = 50

Váriaveis explicativas com duas categorias

Tabela 67 – Taxa de seleção das variáveis para cada método, considerando estrutura de associação 2, $n = 50$, VE2 e variável resposta binária.

Acertos	Backward	Forward	Bidirectional F	Bidirectional B	HC	Tabu	MMHC	RSMAX2
Apenas as 3 variáveis	0,19	0,18	0,19	0,19	0,10	0,11	0,03	0,03
Uma das variáveis	0,18	0,20	0,20	0,18	0,41	0,40	0,52	0,52
Duas das variáveis	0,44	0,43	0,43	0,44	0,36	0,37	0,30	0,29
Três das variáveis	0,36	0,34	0,34	0,36	0,12	0,12	0,03	0,03
x1	0,26	0,26	0,25	0,26	0,13	0,12	0,08	0,08
x2	0,22	0,21	0,21	0,22	0,08	0,08	0,04	0,05
x3	0,23	0,22	0,21	0,23	0,11	0,11	0,07	0,07
x4	0,70	0,69	0,69	0,70	0,47	0,48	0,38	0,38
x5	0,72	0,70	0,70	0,72	0,52	0,53	0,42	0,42
x6	0,72	0,70	0,70	0,72	0,50	0,51	0,40	0,40

Fonte: Do autor (2024).

Tabela 68 – Taxa de seleção das variáveis para cada algoritmo de rede bayesiana considerando *whitelist*, para estrutura de associação 2, $n = 50$, VE2 e variável resposta binária.

Acertos	HC	Tabu	MMHC	RSMAX2
Apenas as 3 variáveis	0,10	0,11	0,02	0,02
Uma das variáveis	0,41	0,40	0,52	0,52
Duas das variáveis	0,36	0,37	0,28	0,28
Três das variáveis	0,12	0,13	0,03	0,03
x1	0,13	0,12	0,08	0,08
x2	0,08	0,08	0,03	0,04
x3	0,11	0,11	0,07	0,07
x4	0,47	0,48	0,34	0,34
x5	0,52	0,53	0,42	0,42
x6	0,50	0,51	0,40	0,40

Fonte: Do autor (2024).

Tabela 69 – Média de verdadeiro positivo(VP), falso positivo(FP) e falso negativo(FN) das relações dos algoritmos de RB, considerando estrutura de associação 2, $n = 50$, VE2 e variável resposta binária.

Algoritmos	VP	FP	FN
HC	3,18	1,35	4,82
Tabu	2,91	1,67	5,09
MMHC	2,40	0,62	5,60
RSMAX2	2,40	0,62	5,60
HC (com wl)	3,97	1,32	4,03
Tabu (com wl)	3,71	1,64	4,29
MMHC (com wl)	3,20	0,53	4,80
RSMAX2 (com wl)	3,20	0,53	4,80

Fonte: Do autor (2024).

Tabela 70 – Média dos valores AIC para cada método, considerando estrutura de associação 2, $n = 50$, VE2 e variável resposta binária.

Método	AIC	Método	AIC
Backward	59,21	MMHC	66,16
Forward	59,24	RSMAX2	66,15
Bidirectional F	59,23	HC (com wl)	64,71
Bidirectional B	59,21	Tabu (com wl)	64,61
HC	64,71	MMHC (com wl)	66,30
Tabu	64,61	RSMAX2 (com wl)	66,29

Fonte: Do autor (2024).

Váriaveis explicativas com três categorias

Tabela 71 – Taxa de seleção das variáveis para cada método, considerando estrutura de associação 2, $n = 50$, VE3 e variável resposta binária.

Acertos	Backward	Forward	Bidirectional F	Bidirectional B	HC	Tabu	MMHC	RSMAX2
Apenas as 3 variáveis	0,27	0,26	0,27	0,27	0,18	0,19	0,05	0,05
Uma das variáveis	0,10	0,11	0,11	0,10	0,31	0,29	0,47	0,47
Duas das variáveis	0,40	0,41	0,41	0,40	0,43	0,44	0,38	0,38
Três das variáveis	0,49	0,47	0,47	0,49	0,21	0,22	0,05	0,05
x1	0,24	0,23	0,22	0,24	0,12	0,11	0,06	0,06
x2	0,21	0,20	0,20	0,21	0,07	0,07	0,03	0,03
x3	0,22	0,21	0,21	0,22	0,09	0,09	0,05	0,05
x4	0,78	0,76	0,76	0,78	0,56	0,58	0,44	0,44
x5	0,80	0,78	0,78	0,80	0,61	0,62	0,46	0,47
x6	0,81	0,79	0,79	0,81	0,61	0,62	0,47	0,47

Fonte: Do autor (2024).

Tabela 72 – Taxa de seleção das variáveis para cada algoritmo de rede bayesiana considerando *whitelist*, para estrutura de associação 2, $n = 50$, VE3 e variável resposta binária.

Acertos	HC	Tabu	MMHC	RSMAX2
Apenas as 3 variáveis	0,18	0,19	0,05	0,05
Uma das variáveis	0,31	0,29	0,47	0,47
Duas das variáveis	0,43	0,43	0,37	0,37
Três das variáveis	0,21	0,22	0,05	0,05
x1	0,12	0,11	0,06	0,06
x2	0,07	0,07	0,03	0,03
x3	0,09	0,09	0,05	0,05
x4	0,56	0,58	0,42	0,42
x5	0,61	0,63	0,46	0,47
x6	0,61	0,63	0,47	0,47

Fonte: Do autor (2024).

Tabela 73 – Média de verdadeiro positivo(VP), falso positivo(FP) e falso negativo(FN) das relações dos algoritmos de RB, considerando estrutura de associação 2, $n = 50$, VE3 e variável resposta binária.

Algoritmos	VP	FP	FN
HC	3,60	1,06	4,40
Tabu	3,30	1,57	4,70
MMHC	2,88	0,45	5,12
RSMAX2	2,88	0,45	5,12
HC (com wl)	4,41	1,07	3,59
Tabu (com wl)	4,13	1,53	3,87
MMHC (com wl)	3,69	0,43	4,31
RSMAX2 (com wl)	3,69	0,44	4,31

Fonte: Do autor (2024).

Tabela 74 – Média dos valores AIC para cada método, considerando estrutura de associação 2, $n = 50$, VE3 e variável resposta binária.

Método	AIC	Método	AIC
Backward	56,37	MMHC	64,12
Forward	56,40	RSMAX2	64,10
Bidirectional F	56,38	HC (com wl)	61,93
Bidirectional B	56,37	Tabu (com wl)	61,79
HC	61,93	MMHC (com wl)	64,22
Tabu	61,80	RSMAX2 (com wl)	64,20

Fonte: Do autor (2024).

n = 150

Váriaveis explicativas com duas categorias

Tabela 75 – Taxa de seleção das variáveis para cada método, considerando estrutura de associação 2, $n = 150$, VE2 e variável resposta binária.

Acertos	Backward	Forward	Bidirectional F	Bidirectional B	HC	Tabu	MMHC	RSMAX2
Apenas as 3 variáveis	0,46	0,46	0,46	0,46	0,63	0,63	0,54	0,54
Uma das variáveis	0,00	0,00	0,00	0,00	0,03	0,03	0,05	0,05
Duas das variáveis	0,07	0,07	0,07	0,07	0,25	0,25	0,36	0,36
Três das variáveis	0,93	0,93	0,93	0,93	0,72	0,72	0,59	0,59
x1	0,24	0,24	0,24	0,24	0,08	0,07	0,06	0,06
x2	0,17	0,17	0,17	0,17	0,03	0,03	0,02	0,02
x3	0,23	0,23	0,23	0,23	0,07	0,07	0,06	0,06
x4	0,97	0,97	0,97	0,97	0,87	0,87	0,82	0,82
x5	0,98	0,98	0,98	0,98	0,91	0,91	0,87	0,87
x6	0,98	0,97	0,97	0,98	0,90	0,90	0,85	0,85

Fonte: Do autor (2024).

Tabela 76 – Taxa de seleção das variáveis para cada algoritmo de rede bayesiana considerando *whitelist*, para estrutura de associação 2, $n = 150$, VE2 e variável resposta binária.

Acertos	HC	Tabu	MMHC	RSMAX2
Apenas as 3 variáveis	0,63	0,63	0,54	0,53
Uma das variáveis	0,03	0,03	0,05	0,05
Duas das variáveis	0,25	0,25	0,36	0,36
Três das variáveis	0,72	0,72	0,59	0,59
x1	0,08	0,07	0,06	0,06
x2	0,03	0,03	0,02	0,02
x3	0,07	0,07	0,06	0,06
x4	0,87	0,87	0,81	0,81
x5	0,91	0,92	0,87	0,87
x6	0,90	0,90	0,85	0,85

Fonte: Do autor (2024).

Tabela 77 – Média de verdadeiro positivo(VP), falso positivo(FP) e falso negativo(FN) das relações dos algoritmos de RB, considerando estrutura de associação 2, $n = 150$, VE2 e variável resposta binária.

Algoritmos	VP	FP	FN
HC	5,02	0,65	2,98
Tabu	4,68	1,09	3,32
MMHC	4,77	0,43	3,23
RSMAX2	4,77	0,43	3,23
HC (com wl)	5,80	0,65	2,20
Tabu (com wl)	5,48	1,07	2,52
MMHC (com wl)	5,56	0,42	2,44
RSMAX2 (com wl)	5,56	0,42	2,44

Fonte: Do autor (2024).

Tabela 78 – Média dos valores AIC para cada método, considerando estrutura de associação 2, $n = 150$, VE2 e variável resposta binária.

Método	AIC	Método	AIC
Backward	174,71	MMHC	186,98
Forward	174,71	RSMAX2	186,98
Bidirectional F	174,71	HC (com wl)	186,07
Bidirectional B	174,71	Tabu (com wl)	186,05
HC	186,07	MMHC (com wl)	187,01
Tabu	186,05	RSMAX2 (com wl)	187,00

Fonte: Do autor (2024).

Váriaveis explicativas com três categorias

Tabela 79 – Taxa de seleção das variáveis para cada método, considerando estrutura de associação 2, $n = 150$, VE3 e variável resposta binária.

Acertos	Backward	Forward	Bidirectional F	Bidirectional B	HC	Tabu	MMHC	RSMAX2
Apenas as 3 variáveis	0,55	0,55	0,55	0,55	0,81	0,81	0,74	0,73
Uma das variáveis	0,00	0,00	0,00	0,00	0,01	0,00	0,01	0,01
Duas das variáveis	0,02	0,02	0,02	0,02	0,11	0,11	0,21	0,21
Três das variáveis	0,98	0,98	0,98	0,98	0,88	0,88	0,78	0,77
x1	0,19	0,18	0,18	0,19	0,05	0,05	0,03	0,04
x2	0,16	0,16	0,16	0,16	0,03	0,03	0,02	0,02
x3	0,18	0,18	0,18	0,18	0,04	0,04	0,03	0,03
x4	0,99	0,99	0,99	0,99	0,95	0,95	0,91	0,91
x5	0,99	0,99	0,99	0,99	0,97	0,97	0,93	0,93
x6	0,99	0,99	0,99	0,99	0,96	0,96	0,93	0,93

Fonte: Do autor (2024).

Tabela 80 – Taxa de seleção das variáveis para cada algoritmo de rede bayesiana considerando *whitelist*, para estrutura de associação 2, $n = 150$, VE3 e variável resposta binária.

Acertos	HC	Tabu	MMHC	RSMAX2
Apenas as 3 variáveis	0,81	0,81	0,73	0,73
Uma das variáveis	0,01	0,00	0,01	0,01
Duas das variáveis	0,11	0,11	0,22	0,22
Três das variáveis	0,88	0,88	0,77	0,77
x1	0,05	0,05	0,03	0,04
x2	0,03	0,03	0,02	0,02
x3	0,04	0,04	0,03	0,03
x4	0,95	0,95	0,90	0,90
x5	0,97	0,97	0,93	0,93
x6	0,96	0,96	0,93	0,93

Fonte: Do autor (2024).

Tabela 81 – Média de verdadeiro positivo(VP), falso positivo(FP) e falso negativo(FN) das relações dos algoritmos de RB, considerando estrutura de associação 2, $n = 150$, VE3 e variável resposta binária.

Algoritmos	VP	FP	FN
HC	5,74	0,68	2,26
Tabu	5,29	1,22	2,71
MMHC	5,54	0,38	2,46
RSMAX2	5,54	0,38	2,46
HC (com wl)	6,39	0,66	1,61
Tabu (com wl)	5,95	1,17	2,05
MMHC (com wl)	6,20	0,37	1,80
RSMAX2 (com wl)	6,20	0,38	1,80

Fonte: Do autor (2024).

Tabela 82 – Média dos valores AIC para cada método, considerando estrutura de associação 2, $n = 150$, VE3 e variável resposta binária.

Método	AIC	Método	AIC
Backward	165,39	MMHC	177,29
Forward	165,39	RSMAX2	177,28
Bidirectional F	165,39	HC (com wl)	176,42
Bidirectional B	165,39	Tabu (com wl)	176,40
HC	176,42	MMHC (com wl)	177,33
Tabu	176,41	RSMAX2 (com wl)	177,32

Fonte: Do autor (2024).

n = 450

Váriaveis explicativas com duas categorias

Tabela 83 – Taxa de seleção das variáveis para cada método, considerando estrutura de associação 2, $n = 450$, VE2 e variável resposta binária.

Acertos	Backward	Forward	Bidirectional F	Bidirectional B	HC	Tabu	MMHC	RSMAX2
Apenas as 3 variáveis	0,30	0,30	0,30	0,30	0,80	0,80	0,81	0,81
Uma das variáveis	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00
Duas das variáveis	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00
Três das variáveis	1,00	1,00	1,00	1,00	1,00	1,00	1,00	1,00
x1	0,39	0,39	0,39	0,39	0,11	0,11	0,10	0,10
x2	0,19	0,19	0,19	0,19	0,02	0,02	0,02	0,02
x3	0,33	0,33	0,33	0,33	0,08	0,08	0,07	0,07
x4	1,00	1,00	1,00	1,00	1,00	1,00	1,00	1,00
x5	1,00	1,00	1,00	1,00	1,00	1,00	1,00	1,00
x6	1,00	1,00	1,00	1,00	1,00	1,00	1,00	1,00

Fonte: Do autor (2024).

Tabela 84 – Taxa de seleção das variáveis para cada algoritmo de rede bayesiana considerando *whitelist*, para estrutura de associação 2, $n = 450$, VE2 e variável resposta binária.

Acertos	HC	Tabu	MMHC	RSMAX2
Apenas as 3 variáveis	0,80	0,80	0,81	0,81
Uma das variáveis	0,00	0,00	0,00	0,00
Duas das variáveis	0,00	0,00	0,00	0,00
Três das variáveis	1,00	1,00	1,00	1,00
x1	0,11	0,11	0,10	0,10
x2	0,02	0,02	0,02	0,02
x3	0,08	0,08	0,07	0,07
x4	1,00	1,00	1,00	1,00
x5	1,00	1,00	1,00	1,00
x6	1,00	1,00	1,00	1,00

Fonte: Do autor (2024).

Tabela 85 – Média de verdadeiro positivo(VP), falso positivo(FP) e falso negativo(FN) das relações dos algoritmos de RB, considerando estrutura de associação 2, $n = 450$, VE2 e variável resposta binária.

Algoritmos	VP	FP	FN
HC	6,43	0,64	1,57
Tabu	6,02	1,08	1,98
MMHC	6,42	0,49	1,58
RSMAX2	6,42	0,49	1,58
HC (com wl)	6,88	0,60	1,12
Tabu (com wl)	6,47	1,04	1,53
MMHC (com wl)	6,87	0,46	1,13
RSMAX2 (com wl)	6,87	0,46	1,13

Fonte: Do autor (2024).

Tabela 86 – Média dos valores AIC para cada método, considerando estrutura de associação 2, $n = 450$, VE2 e variável resposta binária.

Método	AIC	Método	AIC
Backward	517,49	MMHC	546,88
Forward	517,49	RSMAX2	546,89
Bidirectional F	517,49	HC (com wl)	546,80
Bidirectional B	517,49	Tabu (com wl)	546,80
HC	546,80	MMHC (com wl)	546,89
Tabu	546,80	RSMAX2 (com wl)	546,89

Fonte: Do autor (2024).

Váriaveis explicativas com três categorias

Tabela 87 – Taxa de seleção das variáveis para cada método, considerando estrutura de associação 2, $n = 450$, VE3 e variável resposta binária.

Acertos	Backward	Forward	Bidirectional F	Bidirectional B	HC	Tabu	MMHC	RSMAX2
Apenas as 3 variáveis	0,46	0,46	0,46	0,46	0,90	0,90	0,92	0,92
Uma das variáveis	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00
Duas das variáveis	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00
Três das variáveis	1,00	1,00	1,00	1,00	1,00	1,00	1,00	1,00
x1	0,24	0,24	0,24	0,24	0,04	0,04	0,04	0,04
x2	0,17	0,17	0,17	0,17	0,02	0,02	0,01	0,01
x3	0,24	0,24	0,24	0,24	0,04	0,04	0,03	0,03
x4	1,00	1,00	1,00	1,00	1,00	1,00	1,00	1,00
x5	1,00	1,00	1,00	1,00	1,00	1,00	1,00	1,00
x6	1,00	1,00	1,00	1,00	1,00	1,00	1,00	1,00

Fonte: Do autor (2024).

Tabela 88 – Taxa de seleção das variáveis para cada algoritmo de rede bayesiana considerando *whitelist*, para estrutura de associação 2, $n = 450$, VE3 e variável resposta binária.

Acertos	HC	Tabu	MMHC	RSMAX2
Apenas as 3 variáveis	0,90	0,90	0,92	0,92
Uma das variáveis	0,00	0,00	0,00	0,00
Duas das variáveis	0,00	0,00	0,00	0,00
Três das variáveis	1,00	1,00	1,00	1,00
x1	0,04	0,04	0,04	0,04
x2	0,02	0,02	0,01	0,01
x3	0,04	0,04	0,03	0,03
x4	1,00	1,00	1,00	1,00
x5	1,00	1,00	1,00	1,00
x6	1,00	1,00	1,00	1,00

Fonte: Do autor (2024).

Tabela 89 – Média de verdadeiro positivo(VP), falso positivo(FP) e falso negativo(FN) das relações dos algoritmos de RB, considerando estrutura de associação 2, $n = 450$, VE3 e variável resposta binária.

Algoritmos	VP	FP	FN
HC	6,79	0,53	1,21
Tabu	6,33	1,04	1,67
MMHC	6,78	0,37	1,22
RSMAX2	6,78	0,37	1,22
HC (com wl)	7,05	0,45	0,95
Tabu (com wl)	6,60	0,94	1,40
MMHC (com wl)	7,04	0,30	0,96
RSMAX2 (com wl)	7,04	0,30	0,96

Fonte: Do autor (2024).

Tabela 90 – Média dos valores AIC para cada método, considerando estrutura de associação 2, $n = 450$, VE3 e variável resposta binária.

Método	AIC	Método	AIC
Backward	489,94	MMHC	518,84
Forward	489,94	RSMAX2	518,84
Bidirectional F	489,94	HC (com wl)	518,75
Bidirectional B	489,94	Tabu (com wl)	518,75
HC	518,75	MMHC (com wl)	518,84
Tabu	518,75	RSMAX2 (com wl)	518,84

Fonte: Do autor (2024).

Estrutura de associação 3

n = 50

Variável resposta y_1 - Variável explicativa com duas categorias

Tabela 91 – Taxa de seleção das variáveis para cada método, considerando estrutura de associação 3, $n = 50$, VE2 e variável resposta binária y_1.

Acertos	Backward	Forward	Bidirectional F	Bidirectional B	HC	Tabu	MMHC	RSMAX2
Apenas as 4 variáveis	0,02	0,02	0,02	0,02	0,02	0,02	0,00	0,00
Uma das variáveis	0,18	0,22	0,22	0,18	0,23	0,23	0,54	0,54
Duas das variáveis	0,38	0,39	0,39	0,38	0,40	0,40	0,18	0,19
Três das variáveis	0,32	0,29	0,29	0,32	0,28	0,28	0,01	0,01
Quatro das variáveis	0,10	0,07	0,07	0,10	0,06	0,06	0,00	0,00
x1	0,27	0,24	0,23	0,27	0,22	0,22	0,03	0,04
x2	0,29	0,26	0,25	0,29	0,23	0,24	0,04	0,05
x3	0,26	0,23	0,22	0,27	0,21	0,21	0,05	0,05
x4	0,28	0,23	0,23	0,28	0,21	0,22	0,06	0,06
x5	0,25	0,21	0,21	0,25	0,19	0,19	0,03	0,03
x6	0,42	0,37	0,36	0,42	0,36	0,36	0,09	0,09
x7	0,53	0,49	0,49	0,53	0,47	0,48	0,17	0,18
x8	0,49	0,45	0,45	0,49	0,43	0,44	0,13	0,14
x9	0,86	0,85	0,84	1,10	0,84	0,84	0,53	0,55
x10	0,24	0,20	0,20	0,00	0,18	0,18	0,03	0,03

Fonte: Do autor (2024).

Tabela 92 – Taxa de seleção das variáveis para cada algoritmo de rede bayesiana considerando *whitelist*, para estrutura de associação 3, $n = 50$, VE2 e variável resposta binária y_1.

Acertos	HC	Tabu	MMHC	RSMAX2
Apenas as 4 variáveis	0,02	0,02	0,00	0,00
Uma das variáveis	0,23	0,23	0,53	0,54
Duas das variáveis	0,40	0,40	0,16	0,18
Três das variáveis	0,28	0,28	0,01	0,01
Quatro das variáveis	0,06	0,06	0,00	0,00
x1	0,22	0,22	0,03	0,04
x2	0,23	0,24	0,04	0,04
x3	0,21	0,21	0,05	0,05
x4	0,21	0,22	0,05	0,05
x5	0,19	0,19	0,03	0,03
x6	0,36	0,36	0,09	0,09
x7	0,47	0,48	0,16	0,17
x8	0,43	0,44	0,13	0,14
x9	0,84	0,84	0,51	0,53
x10	0,18	0,18	0,03	0,03

Fonte: Do autor (2024).

Tabela 93 – Média de verdadeiro positivo(VP), falso positivo(FP) e falso negativo(FN) das relações dos algoritmos de RB, considerando estrutura de associação 3, $n = 50$, VE2 e variável resposta binária.

Algoritmos	VP	FP	FN
HC	7,86	11,17	7,14
Tabu	7,31	11,99	7,69
MMHC	4,63	1,68	10,37
RSMAX2	4,70	1,71	10,30
HC (com wl)	8,67	10,68	6,33
Tabu (com wl)	8,32	11,36	6,68
MMHC (com wl)	5,70	1,49	9,30
RSMAX2 (com wl)	5,77	1,52	9,23

Fonte: Do autor (2024).

Tabela 94 – Média dos valores AIC para cada método, considerando estrutura de associação 3, $n = 50$, VE2 e variável resposta binária y_1.

Método	AIC	Método	AIC
Backward	56,94	MMHC	66,28
Forward	57,05	RSMAX2	66,06
Bidirectional F	57,02	HC (com wl)	61,53
Bidirectional B	56,94	Tabu (com wl)	61,49
HC	61,53	MMHC (com wl)	66,52
Tabu	61,50	RSMAX2 (com wl)	66,27

Fonte: Do autor (2024).

Variável resposta y_2 - Variável explicativa com duas categorias

Tabela 95 – Taxa de seleção das variáveis para cada método, considerando estrutura de associação 3, $n = 50$, VE2 e variável resposta binária y_2.

Acertos	Backward	Forward	Bidirectional F	Bidirectional B	HC	Tabu	MMHC	RSMAX2
Apenas as 4 variáveis	0,02	0,02	0,02	0,02	0,02	0,02	0,00	0,00
Uma das variáveis	0,15	0,18	0,19	0,15	0,20	0,19	0,51	0,52
Duas das variáveis	0,38	0,39	0,39	0,38	0,41	0,41	0,20	0,22
Três das variáveis	0,35	0,31	0,31	0,35	0,30	0,31	0,02	0,02
Quatro das variáveis	0,10	0,08	0,07	0,11	0,07	0,07	0,00	0,00
x1	0,24	0,20	0,19	0,24	0,18	0,18	0,02	0,02
x2	0,25	0,22	0,21	0,26	0,19	0,20	0,03	0,03
x3	0,28	0,24	0,23	0,28	0,22	0,22	0,06	0,06
x4	0,27	0,24	0,23	0,28	0,22	0,22	0,06	0,07
x5	0,25	0,21	0,21	0,25	0,19	0,19	0,03	0,03
x6	0,25	0,19	0,19	0,25	0,17	0,17	0,02	0,02
x7	0,33	0,28	0,27	0,33	0,26	0,27	0,05	0,06
x8	0,74	0,71	0,71	0,74	0,70	0,70	0,38	0,40
x9	0,75	0,72	0,71	0,75	0,71	0,71	0,31	0,33
x10	0,57	0,53	0,53	0,57	0,51	0,52	0,21	0,21

Fonte: Do autor (2024).

Tabela 96 – Taxa de seleção das variáveis para cada algoritmo de rede bayesiana considerando *whitelist*, considerando estrutura de associação 3, $n = 50$, VE2 e variável resposta binária y_2.

Acertos	HC	Tabu	MMHC	RSMAX2
Apenas as 4 variáveis	0,02	0,02	0,00	0,00
Uma das variáveis	0,20	0,19	0,50	0,51
Duas das variáveis	0,41	0,41	0,19	0,21
Três das variáveis	0,30	0,31	0,01	0,02
Quatro das variáveis	0,07	0,07	0,00	0,00
x1	0,18	0,18	0,02	0,02
x2	0,19	0,20	0,03	0,03
x3	0,22	0,22	0,06	0,06
x4	0,22	0,22	0,05	0,05
x5	0,19	0,19	0,03	0,03
x6	0,17	0,17	0,02	0,02
x7	0,26	0,27	0,05	0,05
x8	0,70	0,71	0,38	0,40
x9	0,71	0,71	0,30	0,31
x10	0,51	0,52	0,21	0,21

Fonte: Do autor (2024).

Tabela 97 – Média dos valores AIC para cada método, considerando estrutura de associação 3, $n = 50$, VE2 e variável resposta binária y_2.

Método	AIC	Método	AIC
Backward	57,49	MMHC	66,88
Forward	57,59	RSMAX2	66,66
Bidirectional F	57,02	HC (com wl)	62,18
Bidirectional B	56,94	Tabu (com wl)	62,15
HC	62,18	MMHC (com wl)	67,07
Tabu	62,16	RSMAX2 (com wl)	66,84

Fonte: Do autor (2024).

Variável resposta y_1 - Variável explicativa com três categorias

Tabela 98 – Taxa de seleção das variáveis para cada método, considerando estrutura de associação 3, $n = 50$, VE3 e variável resposta binária y_1.

Acertos	Backward	Forward	Bidirectional F	Bidirectional B	HC	Tabu	MMHC	RSMAX2
Apenas as 4 variáveis	0,03	0,04	0,04	0,03	0,04	0,04	0,00	0,00
Uma das variáveis	0,13	0,16	0,17	0,13	0,00	0,17	0,57	0,58
Duas das variáveis	0,35	0,37	0,37	0,35	0,17	0,38	0,20	0,22
Três das variáveis	0,37	0,34	0,34	0,37	0,39	0,33	0,02	0,02
Quatro das variáveis	0,14	0,12	0,11	0,14	0,10	0,10	0,00	0,00
x1	0,28	0,24	0,23	0,28	0,22	0,22	0,03	0,03
x2	0,30	0,26	0,25	0,30	0,23	0,24	0,04	0,04
x3	0,26	0,21	0,21	0,26	0,19	0,20	0,04	0,04
x4	0,27	0,22	0,21	0,27	0,19	0,20	0,04	0,04
x5	0,25	0,20	0,20	0,25	0,18	0,19	0,02	0,02
x6	0,47	0,42	0,42	0,47	0,40	0,41	0,09	0,09
x7	0,59	0,55	0,54	0,59	0,53	0,54	0,18	0,18
x8	0,55	0,51	0,51	0,55	0,50	0,50	0,14	0,15
x9	0,91	0,90	0,89	0,91	0,89	0,90	0,62	0,64
x10	0,24	0,21	0,20	0,24	0,18	0,18	0,02	0,02

Fonte: Do autor (2024).

Tabela 99 – Média de verdadeiro positivo(VP), falso positivo(FP) e falso negativo(FN) das relações dos algoritmos de RB, considerando estrutura de associação 3, $n = 50$, VE3 e variável resposta binária.

Algoritmos	VP	FP	FN
HC	8,78	10,22	6,22
Tabu	8,17	11,22	6,83
MMHC	5,44	1,25	9,56
RSMAX2	5,52	1,28	9,48
HC (com wl)	9,43	9,90	5,57
Tabu (com wl)	9,03	10,70	5,97
MMHC (com wl)	6,31	1,13	8,69
RSMAX2 (com wl)	6,39	1,16	8,61

Fonte: Do autor (2024).

Tabela 100 – Taxa de seleção das variáveis para cada algoritmo de rede bayesiana considerando *whitelist*, para estrutura de associação 3, $n = 50$, VE3 e variável resposta binária y_1.

Acertos	multicolumn1|c|HC	Tabu	MMHC	RSMAX2
Apenas as 4 variáveis	0,04	0,04	0,00	0,00
Uma das variáveis	0,00	0,17	0,56	0,57
Duas das variáveis	0,39	0,38	0,20	0,21
Três das variáveis	0,33	0,33	0,02	0,02
Quatro das variáveis	0,10	0,10	0,00	0,00
x1	0,22	0,22	0,03	0,03
x2	0,23	0,24	0,04	0,04
x3	0,19	0,20	0,04	0,04
x4	0,19	0,20	0,03	0,03
x5	0,18	0,19	0,02	0,02
x6	0,40	0,41	0,09	0,09
x7	0,53	0,54	0,17	0,18
x8	0,50	0,51	0,14	0,15
x9	0,89	0,90	0,60	0,62
x10	0,18	0,18	0,02	0,02

Fonte: Do autor (2024).

Tabela 101 – Média dos valores AIC para cada método, considerando estrutura de associação 3, $n = 50$, VE3 e variável resposta binária y_1

Método	AIC	Método	AIC
Backward	54,29	MMHC	64,83
Forward	54,47	RSMAX2	64,59
Bidirectional F	54,43	HC (com wl)	59,16
Bidirectional B	54,29	Tabu (com wl)	59,11
HC	59,16	MMHC (com wl)	65,07
Tabu	59,11	RSMAX2 (com wl)	64,77

Fonte: Do autor (2024).

Variável resposta y_2 - Variável explicativa com três categorias

Tabela 102 – Taxa de seleção das variáveis para cada método, considerando estrutura de associação 3, $n = 50$, VE3 e variável resposta binária y_2.

Acertos	Backward	Forward	Bidirectional F	Bidirectional B	HC	Tabu	MMHC	RSMAX2
Apenas as 4 variáveis	0,03	0,03	0,03	0,03	0,04	0,04	0,00	0,00
Uma das variáveis	0,09	0,12	0,12	0,09	0,00	0,12	0,51	0,52
Duas das variáveis	0,32	0,35	0,35	0,32	0,13	0,36	0,25	0,27
Três das variáveis	0,43	0,40	0,40	0,43	0,36	0,40	0,03	0,03
Quatro das variáveis	0,15	0,12	0,11	0,15	0,10	0,11	0,00	0,00
x1	0,26	0,19	0,19	0,26	0,17	0,17	0,02	0,02
x2	0,27	0,21	0,21	0,27	0,19	0,20	0,02	0,03
x3	0,28	0,23	0,22	0,28	0,21	0,21	0,04	0,05
x4	0,27	0,22	0,21	0,27	0,19	0,20	0,04	0,04
x5	0,25	0,21	0,21	0,25	0,19	0,19	0,02	0,03
x6	0,25	0,18	0,18	0,25	0,16	0,17	0,01	0,01
x7	0,36	0,30	0,30	0,37	0,29	0,30	0,05	0,05
x8	0,80	0,78	0,78	0,80	0,77	0,78	0,45	0,46
x9	0,82	0,80	0,79	0,82	0,79	0,79	0,37	0,38
x10	0,65	0,61	0,61	0,65	0,59	0,60	0,25	0,25

Fonte: Do autor (2024).

Tabela 103 – Média dos valores AIC para cada método, considerando estrutura de associação 3, $n = 50$, VE3 e variável resposta binária y_2.

Método	AIC	Método	AIC
Backward	54,79	MMHC	65,20
Forward	54,95	RSMAX2	65,00
Bidirectional F	54,43	HC (com wl)	59,73
Bidirectional B	54,29	Tabu (com wl)	59,67
HC	59,73	MMHC (com wl)	65,41
Tabu	59,68	RSMAX2 (com wl)	65,21

Fonte: Do autor (2024).

Tabela 104 – Taxa de seleção das variáveis para cada algoritmo de rede bayesiana considerando *whitelist, para estrutura de associação 3, $n = 50$, VE3 e variável resposta binária y_2.*

Acertos	HC	Tabu	MMHC	RSMAX2
Apenas as 4 variáveis	0,04	0,04	0,00	0,00
Uma das variáveis	0,00	0,12	0,51	0,51
Duas das variáveis	0,36	0,36	0,25	0,26
Três das variáveis	0,39	0,40	0,03	0,03
Quatro das variáveis	0,10	0,11	0,00	0,00
x1	0,17	0,17	0,02	0,02
x2	0,19	0,20	0,02	0,03
x3	0,21	0,21	0,04	0,05
x4	0,19	0,20	0,03	0,03
x5	0,19	0,19	0,02	0,03
x6	0,16	0,17	0,01	0,01
x7	0,29	0,30	0,04	0,05
x8	0,77	0,78	0,45	0,46
x9	0,79	0,80	0,35	0,35
x10	0,59	0,60	0,25	0,25

Fonte: Do autor (2024).

Variável resposta y_1 - Variável explicativa com quatro categorias

Tabela 105 – Taxa de seleção das variáveis para cada método, considerando estrutura de associação 3, $n = 50$, VE4 e variável resposta binária y_1.

Acertos	Backward	Forward	Bidirectional F	Bidirectional B	HC	Tabu	MMHC	RSMAX2
Apenas as 4 variáveis	0,04	0,04	0,04	0,04	0,05	0,05	0,00	0,00
Uma das variáveis	0,10	0,13	0,14	0,10	0,15	0,14	0,57	0,57
Duas das variáveis	0,33	0,35	0,36	0,33	0,37	0,37	0,23	0,24
Três das variáveis	0,39	0,37	0,36	0,39	0,36	0,36	0,02	0,02
Quatro das variáveis	0,17	0,14	0,13	0,17	0,11	0,12	0,00	0,00
x1	0,29	0,24	0,23	0,30	0,22	0,22	0,02	0,02
x2	0,29	0,25	0,24	0,29	0,22	0,22	0,04	0,04
x3	0,27	0,21	0,20	0,27	0,18	0,19	0,02	0,03
x4	0,27	0,22	0,21	0,27	0,19	0,20	0,04	0,05
x5	0,25	0,22	0,22	0,25	0,19	0,19	0,02	0,02
x6	0,50	0,45	0,44	0,50	0,42	0,43	0,10	0,10
x7	0,62	0,58	0,58	0,62	0,56	0,57	0,18	0,19
x8	0,58	0,54	0,54	0,58	0,52	0,52	0,15	0,16
x9	0,92	0,92	0,92	0,92	0,92	0,92	0,65	0,66
x10	0,25	0,21	0,21	0,25	0,18	0,18	0,02	0,02

Fonte: Do autor (2024).

Tabela 106 – Taxa de seleção das variáveis para cada algoritmo de rede bayesiana considerando *whitelist*, para estrutura de associação 3, $n = 50$, VE4 e variável resposta binária y_1.

Acertos	HC	Tabu	MMHC	RSMAX2
Apenas as 4 variáveis	0,05	0,05	0,00	0,00
Uma das variáveis	0,15	0,14	0,57	0,57
Duas das variáveis	0,37	0,37	0,22	0,23
Três das variáveis	0,36	0,36	0,00	0,00
Quatro das variáveis	0,11	0,12	0,00	0,00
x1	0,22	0,23	0,02	0,02
x2	0,22	0,22	0,03	0,04
x3	0,18	0,19	0,02	0,03
x4	0,19	0,20	0,03	0,03
x5	0,19	0,19	0,02	0,02
x6	0,42	0,43	0,10	0,10
x7	0,56	0,57	0,18	0,19
x8	0,52	0,52	0,15	0,16
x9	0,92	0,92	0,63	0,65
x10	0,18	0,18	0,02	0,02

Fonte: Do autor (2024).

Tabela 107 – Média de verdadeiro positivo(VP), falso positivo(FP) e falso negativo(FN) das relações dos algoritmos de RB considerando estrutura de associação 3, $n = 50$, VE4 e variável resposta binária.

Algoritmos	VP	FP	FN
HC	8,98	10,17	6,02
Tabu	8,36	11,28	6,64
MMHC	5,65	1,25	9,35
RSMAX2	5,72	1,27	9,28
HC (com wl)	9,56	9,88	5,44
Tabu (com wl)	9,17	10,76	5,83
MMHC (com wl)	6,45	1,11	8,55
RSMAX2 (com wl)	6,53	1,14	8,47

Fonte: Do autor (2024).

Tabela 108 – Média dos valores AIC para cada método, considerando estrutura de associação 3, $n = 50$, VE4 e variável resposta binária y_1.

Método	AIC	Método	AIC
Backward	53,14	MMHC	64,13
Forward	53,30	RSMAX2	63,88
Bidirectional F	53,26	HC (com wl)	58,12
Bidirectional B	53,14	Tabu (com wl)	58,08
HC	58,12	MMHC (com wl)	64,34
Tabu	58,08	RSMAX2 (com wl)	64,05

Fonte: Do autor (2024).

Variável resposta y_2 - Variável explicativa com quatro categorias

Tabela 109 – Taxa de seleção das variáveis para cada método, considerando estrutura de associação 3, $n = 50$, VE4 e variável resposta binária y_2.

Acertos	Backward	Forward	Bidirectional F	Bidirectional B	HC	Tabu	MMHC	RSMAX2
Apenas as 4 variáveis	0,03	0,04	0,04	0,03	0,04	0,04	0,00	0,00
Uma das variáveis	0,07	0,10	0,10	0,07	0,10	0,10	0,50	0,50
Duas das variáveis	0,31	0,33	0,33	0,30	0,34	0,34	0,29	0,31
Três das variáveis	0,44	0,42	0,42	0,44	0,42	0,36	0,03	0,03
Quatro das variáveis	0,18	0,14	0,13	0,18	0,12	0,13	0,00	0,00
x1	0,27	0,19	0,18	0,27	0,16	0,17	0,01	0,01
x2	0,27	0,21	0,21	0,27	0,19	0,20	0,02	0,02
x3	0,27	0,22	0,21	0,27	0,19	0,20	0,03	0,04
x4	0,27	0,22	0,21	0,27	0,19	0,20	0,04	0,05
x5	0,25	0,21	0,21	0,25	0,19	0,19	0,02	0,03
x6	0,26	0,19	0,19	0,27	0,16	0,16	0,01	0,01
x7	0,37	0,31	0,31	0,37	0,29	0,30	0,05	0,05
x8	0,83	0,81	0,81	0,83	0,80	0,81	0,48	0,50
x9	0,84	0,82	0,82	0,84	0,82	0,82	0,40	0,41
x10	0,67	0,64	0,64	0,67	0,63	0,64	0,26	0,26

Fonte: Do autor (2024).

Tabela 110 – Taxa de seleção das variáveis para cada algoritmo de rede bayesiana considerando *whitelist*, para estrutura de associação 3, $n = 50$, VE4 e variável resposta binária y_2.

Acertos	HC	Tabu	MMHC	RSXMAX2
Apenas as 4 variáveis	0,04	0,04	0,00	0,00
Uma das variáveis	0,10	0,10	0,51	0,51
Duas das variáveis	0,34	0,00	0,28	0,30
Três das variáveis	0,42	0,43	0,03	0,03
Quatro das variáveis	0,12	0,13	0,00	0,00
x1	0,16	0,17	0,01	0,01
x2	0,19	0,20	0,02	0,02
x3	0,19	0,20	0,03	0,04
x4	0,19	0,19	0,03	0,03
x5	0,19	0,19	0,02	0,03
x6	0,16	0,16	0,01	0,01
x7	0,29	0,30	0,04	0,05
x8	0,80	0,81	0,48	0,50
x9	0,82	0,82	0,37	0,38
x10	0,63	0,64	0,26	0,26

Fonte: Do autor (2024).

Tabela 111 – Média dos valores AIC para cada método, considerando estrutura de associação 3, $n = 50$, VE4 e variável resposta binária y_2

Método	AIC	Método	AIC
Backward	53,60	MMHC	64,42
Forward	53,78	RSMAX2	64,23
Bidirectional F	53,26	HC (com wl)	58,65
Bidirectional B	53,14	Tabu (com wl)	58,61
HC	58,65	MMHC (com wl)	64,65
Tabu	58,61	RSMAX2 (com wl)	64,47

Fonte: Do autor (2024).

n=150

Variável resposta y_1 - Variável explicativa com duas categorias

Tabela 112 – Taxa de seleção das variáveis para cada método, considerando estrutura de associação 3, $n = 150$, VE2 e variável resposta binária y_1.

Acertos	Backward	Forward	Bidirectional F	Bidirectional B	HC	Tabu	MMHC	RSMAX2
Apenas as 4 variáveis	0,12	0,12	0,13	0,12	0,12	0,12	0,02	0,02
Uma das variáveis	0,01	0,02	0,02	0,01	0,02	0,02	0,32	0,31
Duas das variáveis	0,14	0,14	0,14	0,14	0,15	0,15	0,47	0,48
Três das variáveis	0,44	0,44	0,44	0,44	0,45	0,45	0,17	0,18
Quatro das variáveis	0,41	0,40	0,39	0,41	0,39	0,39	0,02	0,02
x1	0,25	0,24	0,24	0,25	0,24	0,24	0,04	0,05
x2	0,25	0,25	0,25	0,25	0,24	0,24	0,07	0,07
x3	0,24	0,23	0,23	0,24	0,23	0,23	0,06	0,07
x4	0,26	0,24	0,24	0,26	0,25	0,25	0,08	0,08
x5	0,18	0,17	0,17	0,18	0,17	0,17	0,01	0,01
x6	0,64	0,63	0,63	0,64	0,63	0,63	0,23	0,23
x7	0,82	0,81	0,81	0,82	0,81	0,81	0,43	0,44
x8	0,78	0,77	0,77	0,78	0,77	0,77	0,26	0,27
x9	1,00	1,00	1,00	1,18	1,00	1,00	0,95	0,95
x10	0,18	0,17	0,17	0,00	0,17	0,17	0,01	0,01

Fonte: Do autor (2024).

Tabela 113 – Taxa de seleção das variáveis para cada algoritmo de rede bayesiana considerando *whitelist*, para estrutura de associação 3, $n = 150$, VE2 e variável resposta binária y_1.

Acertos	HC	Tabu	MMHC	RSMAX2
Apenas as 4 variáveis	0,12	0,12	0,02	0,02
Uma das variáveis	0,02	0,02	0,33	0,31
Duas das variáveis	0,15	0,15	0,47	0,48
Três das variáveis	0,45	0,45	0,17	0,18
Quatro das variáveis	0,39	0,39	0,02	0,02
x1	0,24	0,24	0,04	0,05
x2	0,24	0,24	0,06	0,07
x3	0,23	0,23	0,06	0,07
x4	0,25	0,25	0,07	0,07
x5	0,17	0,17	0,01	0,01
x6	0,63	0,63	0,23	0,23
x7	0,81	0,81	0,43	0,44
x8	0,77	0,77	0,26	0,27
x9	1,00	1,00	0,94	0,95
x10	0,17	0,17	0,01	0,01

Fonte: Do autor (2024).

Tabela 114 – Média de verdadeiro positivo(VP), falso positivo(FP) e falso negativo(FN) das relações dos algoritmos de RB, considerando estrutura de associação 3, $n = 150$, VE2 e variável resposta binária.

Algoritmos	VP	FP	FN
HC	11,03	10,37	3,97
Tabu	10,30	11,30	4,70
MMHC	8,09	1,44	6,91
RSMAX2	8,15	1,46	6,85
HC (com wl)	11,46	10,08	3,54
Tabu (com wl)	11,00	10,74	4,00
MMHC (com wl)	8,69	1,33	6,31
RSMAX2 (com wl)	8,75	1,35	6,25

Fonte: Do autor (2024).

Tabela 115 – Média dos valores AIC para cada método, considerando estrutura de associação 3, $n = 150$, VE2 e variável resposta binária y_1.

Método	AIC	Método	AIC
Backward	172,91	MMHC	189,51
Forward	172,93	RSMAX2	189,21
Bidirectional F	172,92	HC (com wl)	183,01
Bidirectional B	172,91	Tabu (com wl)	183,00
HC	183,01	MMHC (com wl)	189,63
Tabu	183,00	RSMAX2 (com wl)	189,32

Fonte: Do autor (2024).

Variável resposta y_2 - Variável explicativa com duas categorias

Tabela 116 – Taxa de seleção das variáveis para cada método, considerando estrutura de associação 3, $n = 150$, VE2 e variável resposta binária y_2.

Acertos	Backward	Forward	Bidirectional F	Bidirectional B	HC	Tabu	MMHC	RSMAX2
Apenas as 4 variáveis	0,11	0,12	0,12	0,11	0,12	0,12	0,02	0,02
Uma das variáveis	0,00	0,00	0,00	0,00	0,00	0,00	0,17	0,16
Duas das variáveis	0,07	0,08	0,08	0,07	0,08	0,08	0,46	0,46
Três das variáveis	0,53	0,54	0,54	0,53	0,54	0,54	0,33	0,34
Quatro das variáveis	0,39	0,38	0,38	0,39	0,38	0,38	0,02	0,02
x1	0,18	0,16	0,16	0,18	0,16	0,16	0,01	0,01
x2	0,23	0,22	0,22	0,23	0,22	0,22	0,03	0,04
x3	0,26	0,25	0,24	0,26	0,24	0,25	0,07	0,07
x4	0,27	0,25	0,25	0,27	0,25	0,25	0,06	0,06
x5	0,18	0,18	0,18	0,18	0,17	0,18	0,02	0,02
x6	0,17	0,16	0,16	0,18	0,15	0,16	0,01	0,01
x7	0,46	0,45	0,45	0,46	0,45	0,45	0,09	0,09
x8	0,98	0,98	0,98	0,98	0,98	0,98	0,84	0,85
x9	0,98	0,98	0,98	0,98	0,98	0,98	0,69	0,71
x10	0,89	0,89	0,89	0,89	0,89	0,89	0,56	0,56

Fonte: Do autor (2024).

Tabela 117 – Taxa de seleção das variáveis para cada algoritmo de rede bayesiana considerando *whitelist*, para estrutura de associação 3, $n = 150$, VE2 e variável resposta binária y_2.

Acertos	HC	Tabu	MMHC	RSMAX2
Apenas as 4 variáveis	0,12	0,12	0,02	0,02
Uma das variáveis	0,00	0,00	0,17	0,17
Duas das variáveis	0,08	0,08	0,47	0,47
Três das variáveis	0,54	0,54	0,33	0,33
Quatro das variáveis	0,38	0,38	0,02	0,02
x1	0,16	0,16	0,01	0,01
x2	0,22	0,22	0,03	0,04
x3	0,24	0,25	0,07	0,07
x4	0,25	0,25	0,06	0,06
x5	0,17	0,18	0,02	0,02
x6	0,15	0,16	0,01	0,01
x7	0,45	0,45	0,09	0,09
x8	0,98	0,98	0,84	0,85
x9	0,98	0,98	0,68	0,69
x10	0,89	0,89	0,56	0,56

Fonte: Do autor (2024).

Tabela 118 – Média dos valores AIC para cada método, considerando estrutura de associação 3, $n = 150$, VE2 e variável resposta binária y_2.

Método	AIC	Método	AIC
Backward	173,26	MMHC	189,35
Forward	173,27	RSMAX2	189,11
Bidirectional F	172,92	HC (com wl)	183,57
Bidirectional B	172,91	Tabu (com wl)	183,56
HC	183,57	MMHC (com wl)	189,51
Tabu	183,56	RSMAX2 (com wl)	189,31

Fonte: Do autor (2024).

Variável resposta y_1 - Variável explicativa com três categorias

Tabela 119 – Taxa de seleção das variáveis para cada método, considerando estrutura de associação 3, $n = 150$, VE3 e variável resposta binária y_1.

Acertos	Backward	Forward	Bidirectional F	Bidirectional B	HC	Tabu	MMHC	RSMAX2
Apenas as 4 variáveis	0,19	0,19	0,19	0,19	0,20	0,19	0,04	0,04
Uma das variáveis	0,00	0,01	0,01	0,00	0,00	0,01	0,24	0,23
Duas das variáveis	0,07	0,08	0,08	0,07	0,01	0,08	0,49	0,49
Três das variáveis	0,38	0,38	0,38	0,38	0,08	0,39	0,23	0,24
Quatro das variáveis	0,54	0,54	0,54	0,55	0,53	0,53	0,04	0,04
x1	0,23	0,22	0,22	0,23	0,21	0,21	0,03	0,03
x2	0,21	0,21	0,20	0,22	0,21	0,21	0,05	0,06
x3	0,22	0,20	0,20	0,22	0,20	0,20	0,03	0,04
x4	0,22	0,20	0,20	0,22	0,19	0,20	0,04	0,05
x5	0,17	0,16	0,16	0,17	0,15	0,16	0,01	0,01
x6	0,73	0,72	0,72	0,73	0,71	0,71	0,28	0,28
x7	0,88	0,88	0,88	0,88	0,88	0,88	0,51	0,51
x8	0,85	0,85	0,85	0,85	0,85	0,85	0,30	0,31
x9	1,00	1,00	1,00	1,00	1,00	1,00	0,97	0,98
x10	0,17	0,16	0,16	0,17	0,16	0,16	0,01	0,01

Fonte: Do autor (2024).

Tabela 120 – Taxa de seleção das variáveis para cada algoritmo de rede bayesiana considerando *whitelist*, para estrutura de associação 3, $n = 150$, VE3 e variável resposta binária y_1.

Acertos	HC	Tabu	MMHC	RSMAX2
Apenas as 4 variáveis	0,20	0,19	0,04	0,04
Uma das variáveis	0,00	0,01	0,25	0,23
Duas das variáveis	0,08	0,08	0,48	0,49
Três das variáveis	0,39	0,39	0,23	0,23
Quatro das variáveis	0,53	0,53	0,04	0,04
x1	0,21	0,21	0,03	0,03
x2	0,21	0,21	0,05	0,06
x3	0,20	0,20	0,03	0,04
x4	0,19	0,20	0,04	0,04
x5	0,15	0,16	0,01	0,01
x6	0,71	0,71	0,28	0,28
x7	0,88	0,88	0,50	0,51
x8	0,85	0,85	0,30	0,31
x9	1,00	1,00	0,97	0,97
x10	0,16	0,16	0,01	0,01

Fonte: Do autor (2024).

Tabela 121 – Média de verdadeiro positivo(VP), falso positivo(FP) e falso negativo(FN) das relações dos algoritmos de RB, considerando estrutura de associação 3, $n = 150$, VE3 e variável resposta binária.

Algoritmos	VP	FP	FN
HC	11,78	9,74	3,22
Tabu	11,09	10,72	3,91
MMHC	9,10	1,18	5,90
RSMAX2	9,13	1,21	5,87
HC (com wl)	12,07	9,56	2,93
Tabu (com wl)	11,64	10,26	3,36
MMHC (com wl)	9,51	1,09	5,49
RSMAX2 (com wl)	9,54	1,12	5,46

Fonte: Do autor (2024).

Tabela 122 – Média dos valores AIC para cada método, considerando estrutura de associação 3, $n = 150$, VE3 e variável resposta binária y_1.

Método	AIC	Método	AIC
Backward	164,28	MMHC	182,14
Forward	164,30	RSMAX2	181,94
Bidirectional F	164,29	HC (com wl)	174,37
Bidirectional B	164,28	Tabu (com wl)	174,36
HC	174,37	MMHC (com wl)	182,23
Tabu	174,37	RSMAX2 (com wl)	182,00

Fonte: Do autor (2024).

Variável resposta y_2 - Variável explicativa com três categorias

Tabela 123 – Taxa de seleção das variáveis para cada método, considerando estrutura de associação 3, $n = 150$, VE3 e variável resposta binária y_2.

Acertos	Backward	Forward	Bidirectional F	Bidirectional B	HC	Tabu	MMHC	RSMAX2
Apenas as 4 variáveis	0,16	0,17	0,17	0,16	0,17	0,17	0,04	0,04
Uma das variáveis	0,00	0,00	0,00	0,00	0,00	0,00	0,08	0,07
Duas das variáveis	0,03	0,03	0,03	0,03	0,00	0,03	0,39	0,39
Três das variáveis	0,46	0,47	0,48	0,46	0,03	0,48	0,49	0,49
Quatro das variáveis	0,51	0,50	0,49	0,51	0,48	0,49	0,04	0,04
x1	0,19	0,16	0,16	0,19	0,15	0,15	0,01	0,01
x2	0,22	0,21	0,20	0,22	0,20	0,21	0,03	0,03
x3	0,21	0,19	0,19	0,21	0,19	0,19	0,04	0,04
x4	0,22	0,20	0,19	0,22	0,20	0,20	0,03	0,03
x5	0,18	0,17	0,17	0,18	0,17	0,17	0,01	0,01
x6	0,18	0,15	0,15	0,18	0,14	0,15	0,01	0,01
x7	0,54	0,53	0,53	0,54	0,52	0,52	0,09	0,10
x8	0,99	0,99	0,99	0,99	0,99	0,99	0,91	0,91
x9	0,99	0,99	0,99	0,99	0,99	0,99	0,79	0,80
x10	0,95	0,95	0,95	0,95	0,95	0,95	0,69	0,69

Fonte: Do autor (2024).

Tabela 124 – Taxa de seleção das variáveis para cada algoritmo de rede bayesiana considerando *whitelist*, para estrutura de associação 3, $n = 150$, VE3 e variável resposta binária y_1.

Acertos	HC	Tabu	MMHC	RSMAX2
Apenas as 4 variáveis	0,17	0,17	0,04	0,04
Uma das variáveis	0,00	0,00	0,08	0,08
Duas das variáveis	0,03	0,03	0,40	0,40
Três das variáveis	0,49	0,48	0,48	0,49
Quatro das variáveis	0,48	0,49	0,04	0,04
x1	0,15	0,15	0,01	0,01
x2	0,20	0,21	0,03	0,03
x3	0,19	0,19	0,04	0,04
x4	0,20	0,20	0,03	0,03
x5	0,17	0,17	0,01	0,01
x6	0,14	0,15	0,01	0,01
x7	0,52	0,52	0,09	0,10
x8	0,99	0,99	0,91	0,91
x9	0,99	0,99	0,78	0,79
x10	0,95	0,95	0,69	0,69

Fonte: Do autor (2024).

Tabela 125 – Média dos valores AIC para cada método, considerando estrutura de associação 3, $n = 150$, VE3 e variável resposta binária y_2.

Método	AIC	Método	AIC
Backward	164,50	MMHC	180,64
Forward	164,51	RSMAX2	180,50
Bidirectional F	164,29	HC (com wl)	174,84
Bidirectional B	164,28	Tabu (com wl)	174,83
HC	174,84	MMHC (com wl)	180,74
Tabu	174,83	RSMAX2 (com wl)	180,63

Fonte: Do autor (2024).

Variável resposta y_1 - Variável explicativa com quatro categorias

Tabela 126 – Taxa de seleção das variáveis para cada método, considerando estrutura de associação 3, $n = 150$, VE4 e variável resposta binária y_1.

Acertos	Backward	Forward	Bidirectional F	Bidirectional B	HC	Tabu	MMHC	RSMAX2
Apenas as 4 variáveis	0,21	0,22	0,22	0,21	0,22	0,22	0,04	0,05
Uma das variáveis	0,00	0,00	0,00	0,00	0,00	0,00	0,20	0,20
Duas das variáveis	0,05	0,06	0,06	0,05	0,06	0,06	0,47	0,47
Três das variáveis	0,35	0,35	0,35	0,35	0,36	0,36	0,28	0,28
Quatro das variáveis	0,60	0,59	0,58	0,60	0,58	0,58	0,05	0,05
x1	0,22	0,21	0,21	0,22	0,20	0,21	0,03	0,03
x2	0,20	0,20	0,20	0,20	0,20	0,20	0,04	0,05
x3	0,21	0,18	0,18	0,21	0,17	0,18	0,02	0,02
x4	0,22	0,20	0,20	0,23	0,20	0,20	0,04	0,04
x5	0,18	0,17	0,17	0,18	0,16	0,16	0,01	0,01
x6	0,75	0,75	0,74	0,76	0,74	0,74	0,29	0,30
x7	0,91	0,90	0,90	0,91	0,90	0,90	0,54	0,55
x8	0,88	0,87	0,87	0,88	0,87	0,87	0,34	0,35
x9	1,00	1,00	1,00	1,00	1,00	1,00	0,98	0,98
x10	0,17	0,16	0,16	0,17	0,15	0,16	0,01	0,01

Fonte: Do autor (2024).

Tabela 127 – Taxa de seleção das variáveis para cada algoritmo de rede bayesiana considerando *whitelist*, para estrutura de associação 3, $n = 150$, VE4 e variável resposta binária y_1.

Acertos	HC	Tabu	MMHC	RSMAX2
Apenas as 4 variáveis	0,22	0,22	0,04	0,05
Uma das variáveis	0,00	0,00	0,21	0,20
Duas das variáveis	0,06	0,06	0,47	0,47
Três das variáveis	0,36	0,36	0,00	0,00
Quatro das variáveis	0,58	0,58	0,04	0,05
x1	0,20	0,21	0,03	0,03
x2	0,20	0,20	0,04	0,05
x3	0,17	0,18	0,02	0,02
x4	0,20	0,20	0,03	0,04
x5	0,16	0,16	0,01	0,01
x6	0,74	0,74	0,29	0,30
x7	0,90	0,90	0,53	0,55
x8	0,87	0,87	0,34	0,35
x9	1,00	1,00	0,98	0,98
x10	0,15	0,16	0,01	0,01

Fonte: Do autor (2024).

Tabela 128 – Média de verdadeiro positivo(VP), falso positivo(FP) e falso negativo(FN) das relações dos algoritmos de RB, considerando estrutura de associação 3, $n = 150$, VE4 e variável resposta binária.

Algoritmos	VP	FP	FN
HC	11,87	10,25	3,13
Tabu	11,20	11,23	3,80
MMHC	9,28	1,43	5,72
RSMAX2	9,31	1,45	5,69
HC (com wl)	12,13	10,08	2,87
Tabu (com wl)	11,73	10,77	3,27
MMHC (com wl)	9,63	1,33	5,37
RSMAX2 (com wl)	9,66	1,36	5,34

Fonte: Do autor (2024).

Tabela 129 – Média dos valores AIC para cada método, considerando estrutura de associação 3, $n = 150$, VE2 e variável resposta binária y_1.

Método	AIC	Método	AIC
Backward	160,96	MMHC	179,09
Forward	160,98	RSMAX2	178,90
Bidirectional F	160,97	HC (com wl)	171,08
Bidirectional B	160,96	Tabu (com wl)	171,07
HC	171,08	MMHC (com wl)	179,19
Tabu	171,07	RSMAX2 (com wl)	178,93

Fonte: Do autor (2024).

Variável resposta y_2 - Variável explicativa com quatro categorias

Tabela 130 – Taxa de seleção das variáveis para cada método, considerando estrutura de associação 3, $n = 150$, VE4 e variável resposta binária y_2.

Acertos	Backward	Forward	Bidirectional F	Bidirectional B	HC	Tabu	MMHC	RSMAX2
Apenas as 4 variáveis	0,18	0,19	0,19	0,18	0,20	0,20	0,04	0,04
Uma das variáveis	0,00	0,00	0,00	0,00	0,00	0,00	0,05	0,05
Duas das variáveis	0,02	0,02	0,02	0,02	0,02	0,02	0,36	0,36
Três das variáveis	0,44	0,46	0,46	0,44	0,46	0,36	0,54	0,54
Quatro das variáveis	0,54	0,52	0,52	0,54	0,51	0,52	0,04	0,04
x1	0,19	0,15	0,15	0,19	0,14	0,15	0,01	0,01
x2	0,22	0,20	0,20	0,22	0,20	0,20	0,03	0,03
x3	0,21	0,18	0,18	0,21	0,18	0,18	0,03	0,03
x4	0,21	0,19	0,18	0,21	0,19	0,19	0,03	0,03
x5	0,18	0,17	0,17	0,18	0,16	0,17	0,01	0,01
x6	0,18	0,15	0,15	0,18	0,14	0,15	0,01	0,01
x7	0,56	0,54	0,54	0,56	0,54	0,54	0,10	0,10
x8	1,00	1,00	1,00	1,00	1,00	1,00	0,93	0,94
x9	1,00	1,00	1,00	1,00	1,00	1,00	0,82	0,82
x10	0,96	0,96	0,96	0,96	0,96	0,96	0,72	0,72

Fonte: Do autor (2024).

Tabela 131 – Taxa de seleção das variáveis para cada algoritmo de rede bayesiana considerando *whitelist*, para estrutura de associação 3, $n = 150$, VE4 e variável resposta binária y_2.

Acertos	HC	Tabu	MMHC	RSMAX2
Apenas as 4 variáveis	0,20	0,20	0,04	0,04
Uma das variáveis	0,00	0,00	0,05	0,05
Duas das variáveis	0,02	0,00	0,37	0,37
Três das variáveis	0,46	0,46	0,53	0,54
Quatro das variáveis	0,51	0,52	0,04	0,04
x1	0,14	0,15	0,01	0,01
x2	0,20	0,20	0,03	0,03
x3	0,18	0,18	0,03	0,03
x4	0,19	0,19	0,03	0,03
x5	0,16	0,16	0,01	0,01
x6	0,14	0,15	0,01	0,01
x7	0,54	0,54	0,10	0,10
x8	1,00	1,00	0,93	0,94
x9	1,00	1,00	0,81	0,81
x10	0,96	0,96	0,72	0,72

Fonte: Do autor (2024).

Tabela 132 – Média dos valores AIC para cada método, considerando estrutura de associação 3, $n = 150$, VE4 e variável resposta binária y_2.

Método	AIC	Método	AIC
Backward	161,01	MMHC	177,15
Forward	161,03	RSMAX2	177,04
Bidirectional F	160,97	HC (com wl)	171,37
Bidirectional B	160,96	Tabu (com wl)	171,36
HC	171,37	MMHC (com wl)	177,23
Tabu	171,36	RSMAX2 (com wl)	177,13

Fonte: Do autor (2024).

n=450

Variável resposta y_1 - Variável explicativa com duas categorias

Tabela 133 – Taxa de seleção das variáveis para cada método, considerando estrutura de associação 3, $n = 450$, VE2 e variável resposta binária y_1.

Acertos	Backward	Forward	Bidirectional F	Bidirectional B	HC	Tabu	MMHC	RSMAX2
Apenas as 4 variáveis	0,11	0,12	0,12	0,11	0,11	0,11	0,29	0,30
Uma das variáveis	0,00	0,00	0,00	0,00	0,00	0,00	0,01	0,01
Duas das variáveis	0,00	0,00	0,00	0,00	0,00	0,00	0,16	0,15
Três das variáveis	0,07	0,07	0,07	0,07	0,07	0,07	0,46	0,45
Quatro das variáveis	0,93	0,93	0,93	0,93	0,92	0,92	0,38	0,38
x1	0,28	0,28	0,28	0,28	0,27	0,27	0,06	0,07
x2	0,35	0,35	0,35	0,35	0,34	0,34	0,09	0,09
x3	0,31	0,31	0,30	0,31	0,31	0,31	0,10	0,10
x4	0,37	0,36	0,36	0,37	0,37	0,37	0,11	0,11
x5	0,16	0,16	0,16	0,16	0,16	0,16	0,01	0,01
x6	0,94	0,94	0,94	0,94	0,94	0,94	0,67	0,67
x7	0,99	0,99	0,99	0,99	0,99	0,99	0,89	0,90
x8	0,99	0,99	0,99	0,99	0,99	0,99	0,63	0,64
x9	1,00	1,00	1,00	1,17	1,00	1,00	1,00	1,00
x10	0,17	0,17	0,17	0,00	0,17	0,17	0,01	0,01

Fonte: Do autor (2024).

Tabela 134 – Taxa de seleção das variáveis para cada algoritmo de rede bayesiana considerando *whitelist*, para estrutura de associação 3, $n = 450$, VE2 e variável resposta binária y_1.

Acertos	HC	Tabu	MMHC	RSMAX2
Apenas as 4 variáveis	0,11	0,11	0,29	0,30
Uma das variáveis	0,00	0,00	0,01	0,01
Duas das variáveis	0,00	0,00	0,15	0,15
Três das variáveis	0,07	0,07	0,46	0,45
Quatro das variáveis	0,92	0,92	0,38	0,38
x1	0,27	0,27	0,06	0,07
x2	0,34	0,34	0,09	0,09
x3	0,31	0,31	0,10	0,10
x4	0,37	0,37	0,11	0,11
x5	0,16	0,16	0,01	0,01
x6	0,94	0,94	0,67	0,67
x7	0,99	0,99	0,89	0,90
x8	0,99	0,99	0,63	0,64
x9	1,00	1,00	1,00	1,00
x10	0,17	0,17	0,01	0,01

Fonte: Do autor (2024).

Tabela 135 – Média de verdadeiro positivo(VP), falso positivo(FP) e falso negativo(FN) das relações dos algoritmos de RB, considerando estrutura de associação 3, $n = 450$, VE2 e variável resposta binária.

Algoritmos	VP	FP	FN
HC	12,88	11,27	2,12
Tabu	12,16	12,16	2,84
MMHC	11,57	1,78	3,43
RSMAX2	11,58	1,79	3,42
HC (com wl)	13,09	11,09	1,92
Tabu (com wl)	12,66	11,69	2,34
MMHC (com wl)	11,70	1,72	3,30
RSMAX2 (com wl)	11,72	1,74	3,28

Fonte: Do autor (2024).

Tabela 136 – Média dos valores AIC para cada método, considerando estrutura de associação 3, $n = 450$, VE2 e variável resposta binária y_1.

Método	AIC	Método	AIC
Backward	514,17	MMHC	548,58
Forward	514,17	RSMAX2	548,39
Bidirectional F	514,17	HC (com wl)	541,52
Bidirectional B	514,17	Tabu (com wl)	541,52
HC	541,52	MMHC (com wl)	548,61
Tabu	541,52	RSMAX2 (com wl)	548,40

Fonte: Do autor (2024).

Variável resposta y_2 - Variável explicativa com duas categorias

Tabela 137 – Taxa de seleção das variáveis para cada método, considerando estrutura de associação 3, $n = 450$, VE2 e variável resposta binária y_2.

Acertos	Backward	Forward	Bidirectional F	Bidirectional B	HC	Tabu	MMHC	RSMAX2
Apenas as 4 variáveis	0,13	0,14	0,14	0,13	0,14	0,14	0,14	0,14
Uma das variáveis	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00
Duas das variáveis	0,00	0,00	0,00	0,00	0,00	0,00	0,03	0,03
Três das variáveis	0,22	0,22	0,22	0,22	0,22	0,22	0,80	0,79
Quatro das variáveis	0,78	0,78	0,78	0,78	0,78	0,78	0,18	0,18
x1	0,17	0,15	0,15	0,17	0,15	0,15	0,01	0,01
x2	0,25	0,25	0,25	0,25	0,25	0,25	0,06	0,07
x3	0,36	0,36	0,36	0,36	0,36	0,36	0,11	0,11
x4	0,34	0,33	0,33	0,34	0,33	0,33	0,09	0,09
x5	0,19	0,19	0,19	0,19	0,18	0,18	0,01	0,01
x6	0,17	0,15	0,15	0,17	0,15	0,15	0,01	0,01
x7	0,78	0,78	0,78	0,78	0,78	0,78	0,18	0,19
x8	1,00	1,00	1,00	1,00	1,00	1,00	1,00	1,00
x9	1,00	1,00	1,00	1,00	1,00	1,00	0,99	0,99
x10	1,00	1,00	1,00	1,00	1,00	1,00	0,98	0,98

Fonte: Do autor (2024).

Tabela 138 – Taxa de seleção das variáveis para cada algoritmo de rede bayesiana considerando *whitelist*, para estrutura de associação 3, $n = 450$, VE2 e variável resposta binária y_2.

Acertos	HC	Tabu	MMHC	RSMAX2
Apenas as 4 variáveis	0,14	0,14	0,14	0,14
Uma das variáveis	0,00	0,00	0,00	0,00
Duas das variáveis	0,00	0,00	0,03	0,03
Três das variáveis	0,22	0,22	0,80	0,79
Quatro das variáveis	0,78	0,78	0,18	0,18
x1	0,15	0,15	0,01	0,01
x2	0,25	0,25	0,06	0,07
x3	0,36	0,36	0,11	0,11
x4	0,33	0,33	0,09	0,09
x5	0,18	0,18	0,01	0,01
x6	0,15	0,15	0,01	0,01
x7	0,78	0,78	0,18	0,19
x8	1,00	1,00	1,00	1,00
x9	1,00	1,00	0,99	0,99
x10	1,00	1,00	0,98	0,98

Fonte: Do autor (2024).

Tabela 139 – Média dos valores AIC para cada método, considerando estrutura de associação 3, $n = 450$, VE2 e variável resposta binária y_2.

Método	AIC	Método	AIC
Backward	515,33	MMHC	548,49
Forward	515,34	RSMAX2	548,42
Bidirectional F	514,17	HC (com wl)	543,29
Bidirectional B	514,17	Tabu (com wl)	543,29
HC	543,29	MMHC (com wl)	548,50
Tabu	543,29	RSMAX2 (com wl)	548,42

Fonte: Do autor (2024).

Variável resposta y_1 - Variável explicativa com três categorias

Tabela 140 – Taxa de seleção das variáveis para cada método, considerando estrutura de associação 3, $n = 450$, VE3 e variável resposta binária y_1.

Acertos	Backward	Forward	Bidirectional F	Bidirectional B	HC	Tabu	MMHC	RSMAX2
Apenas as 4 variáveis	0,23	0,24	0,24	0,23	0,24	0,24	0,45	0,45
Uma das variáveis	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00
Duas das variáveis	0,00	0,00	0,00	0,00	0,00	0,00	0,08	0,08
Três das variáveis	0,03	0,03	0,03	0,03	0,00	0,03	0,42	0,42
Quatro das variáveis	0,97	0,97	0,97	0,97	0,97	0,97	0,50	0,50
x1	0,20	0,20	0,20	0,20	0,20	0,20	0,04	0,04
x2	0,24	0,23	0,23	0,24	0,24	0,24	0,04	0,04
x3	0,23	0,21	0,21	0,23	0,21	0,22	0,04	0,04
x4	0,25	0,23	0,23	0,25	0,24	0,24	0,04	0,04
x5	0,17	0,16	0,16	0,17	0,16	0,16	0,01	0,01
x6	0,97	0,97	0,97	0,97	0,97	0,97	0,76	0,76
x7	1,00	1,00	1,00	1,00	1,00	1,00	0,94	0,94
x8	1,00	1,00	1,00	1,00	1,00	1,00	0,71	0,71
x9	1,00	1,00	1,00	1,00	1,00	1,00	1,00	1,00
x10	0,17	0,16	0,16	0,17	0,16	0,17	0,01	0,01

Fonte: Do autor (2024).

Tabela 141 – Taxa de seleção das variáveis para cada algoritmo de rede bayesiana considerando *whitelist*, para estrutura de associação 3, $n = 450$, VE3 e variável resposta binária y_1.

Acertos	HC	Tabu	MMHC	RSMAX2
Apenas as 4 variáveis	0,24	0,24	0,45	0,45
Uma das variáveis	0,00	0,00	0,00	0,00
Duas das variáveis	0,00	0,00	0,08	0,08
Três das variáveis	0,03	0,03	0,42	0,42
Quatro das variáveis	0,97	0,97	0,50	0,50
x1	0,20	0,20	0,04	0,04
x2	0,24	0,24	0,04	0,04
x3	0,21	0,22	0,04	0,04
x4	0,24	0,24	0,04	0,04
x5	0,16	0,16	0,01	0,01
x6	0,97	0,97	0,76	0,76
x7	1,00	1,00	0,94	0,94
x8	1,00	1,00	0,71	0,71
x9	1,00	1,00	1,00	1,00
x10	0,16	0,17	0,01	0,01

Fonte: Do autor (2024).

Tabela 142 – Média de verdadeiro positivo(VP), falso positivo(FP) e falso negativo(FN) das relações dos algoritmos de RB, considerando estrutura de associação 3, $n = 450$, VE3 e variável resposta binária.

Algoritmos	VP	FP	FN
HC	13,20	10,09	1,80
Tabu	12,59	10,94	2,41
MMHC	12,18	1,46	2,82
RSMAX2	12,18	1,48	2,82
HC (com wl)	13,34	9,97	1,66
Tabu (com wl)	12,97	10,57	2,03
MMHC (com wl)	12,21	1,43	2,79
RSMAX2 (com wl)	12,22	1,46	2,78

Fonte: Do autor (2024).

Tabela 143 – Média dos valores AIC para cada método, considerando estrutura de associação 3, $n = 450$, VE3 e variável resposta binária y_1.

Método	AIC	Método	AIC
Backward	487,52	MMHC	521,71
Forward	487,52	RSMAX2	521,63
Bidirectional F	487,52	HC (com wl)	514,63
Bidirectional B	487,52	Tabu (com wl)	514,63
HC	514,63	MMHC (com wl)	521,72
Tabu	514,63	RSMAX2 (com wl)	521,63

Fonte: Do autor (2024).

Variável resposta y_2 - Variável explicativa com três categorias

Tabela 144 – Taxa de seleção das variáveis para cada método, considerando estrutura de associação 3, $n = 450$, VE3 e variável resposta binária y_2.

Acertos	Backward	Forward	Bidirectional F	Bidirectional B	HC	Tabu	MMHC	RSMAX2
Apenas as 4 variáveis	0,24	0,25	0,25	0,23	0,25	0,25	0,18	0,18
Uma das variáveis	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00
Duas das variáveis	0,00	0,00	0,00	0,00	0,00	0,00	0,01	0,01
Três das variáveis	0,14	0,15	0,15	0,14	0,00	0,15	0,79	0,79
Quatro das variáveis	0,86	0,85	0,85	0,86	0,85	0,85	0,20	0,20
x1	0,17	0,14	0,14	0,17	0,14	0,15	0,00	0,01
x2	0,21	0,20	0,20	0,21	0,20	0,20	0,04	0,04
x3	0,26	0,24	0,24	0,26	0,25	0,25	0,05	0,05
x4	0,24	0,23	0,22	0,24	0,23	0,23	0,04	0,04
x5	0,17	0,17	0,17	0,17	0,17	0,17	0,01	0,01
x6	0,17	0,14	0,14	0,17	0,14	0,15	0,00	0,00
x7	0,86	0,85	0,85	0,86	0,85	0,85	0,20	0,20
x8	1,00	1,00	1,00	1,00	1,00	1,00	1,00	1,00
x9	1,00	1,00	1,00	1,00	1,00	1,00	1,00	1,00
x10	1,00	1,00	1,00	1,00	1,00	1,00	0,99	0,99

Fonte: Do autor (2024).

Tabela 145 – Taxa de seleção das variáveis para cada algoritmo de rede bayesiana considerando *whitelist*, para estrutura de associação 3, $n = 450$, VE3 e variável resposta binária y_2.

Acertos	HC	Tabu	MMHC	RSMAX2
Apenas as 4 variáveis	0,25	0,25	0,18	0,18
Uma das variáveis	0,00	0,00	0,00	0,00
Duas das variáveis	0,00	0,00	0,01	0,01
Três das variáveis	0,15	0,15	0,79	0,79
Quatro das variáveis	0,85	0,85	0,20	0,20
x1	0,14	0,15	0,01	0,01
x2	0,20	0,20	0,04	0,04
x3	0,25	0,25	0,05	0,05
x4	0,23	0,23	0,04	0,04
x5	0,17	0,17	0,01	0,01
x6	0,14	0,15	0,00	0,00
x7	0,85	0,85	0,20	0,20
x8	1,00	1,00	1,00	1,00
x9	1,00	1,00	1,00	1,00
x10	1,00	1,00	0,99	0,99

Fonte: Do autor (2024).

Tabela 146 – Média dos valores AIC para cada método, considerando estrutura de associação 3, $n = 450$, VE3 e variável resposta binária y_2.

Método	AIC	Método	AIC
Backward	488,74	MMHC	522,48
Forward	488,75	RSMAX2	522,44
Bidirectional F	487,52	HC (com wl)	516,56
Bidirectional B	487,52	Tabu (com wl)	516,56
HC	516,56	MMHC (com wl)	522,50
Tabu	516,56	RSMAX2 (com wl)	522,44

Variável resposta y_1 - Variável explicativa com quatro categorias

Tabela 147 – Taxa de seleção das variáveis para cada método, considerando estrutura de associação 3, $n = 450$, VE4 e variável resposta binária y_1.

Acertos	Backward	Forward	Bidirectional F	Bidirectional B	HC	Tabu	MMHC	RSMAX2
Apenas as 4 variáveis	0,28	0,30	0,30	0,28	0,30	0,30	0,55	0,56
Uma das variáveis	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00
Duas das variáveis	0,00	0,00	0,00	0,00	0,00	0,00	0,06	0,06
Três das variáveis	0,02	0,02	0,02	0,02	0,02	0,02	0,35	0,35
Quatro das variáveis	0,98	0,98	0,98	0,98	0,98	0,98	0,59	0,59
x1	0,20	0,19	0,19	0,20	0,19	0,19	0,03	0,03
x2	0,21	0,20	0,20	0,21	0,20	0,20	0,03	0,03
x3	0,20	0,18	0,18	0,20	0,18	0,18	0,01	0,01
x4	0,21	0,20	0,20	0,21	0,20	0,20	0,03	0,03
x5	0,16	0,16	0,16	0,16	0,16	0,16	0,01	0,01
x6	0,98	0,98	0,98	0,98	0,98	0,98	0,79	0,79
x7	1,00	1,00	1,00	1,00	1,00	1,00	0,95	0,95
x8	1,00	1,00	1,00	1,00	1,00	1,00	0,79	0,79
x9	1,00	1,00	1,00	1,00	1,00	1,00	1,00	1,00
x10	0,17	0,16	0,16	0,17	0,16	0,16	0,01	0,01

Fonte: Do autor (2024).

Tabela 148 – Taxa de seleção das variáveis para cada algoritmo de rede bayesiana considerando *whitelist*, para estrutura de associação 3, $n = 450$, VE4 e variável resposta binária y_1.

Acertos	HC	Tabu	MMHC	RSMAX2
Apenas as 4 variáveis	0,30	0,30	0,55	0,56
Uma das variáveis	0,00	0,00	0,00	0,00
Duas das variáveis	0,00	0,00	0,06	0,06
Três das variáveis	0,02	0,02	0,00	0,00
Quatro das variáveis	0,98	0,98	0,59	0,59
x1	0,19	0,19	0,03	0,03
x2	0,20	0,20	0,03	0,03
x3	0,18	0,18	0,01	0,01
x4	0,20	0,20	0,03	0,03
x5	0,16	0,16	0,01	0,01
x6	0,98	0,98	0,79	0,79
x7	1,00	1,00	0,95	0,95
x8	1,00	1,00	0,79	0,79
x9	1,00	1,00	1,00	1,00
x10	0,16	0,16	0,01	0,01

Fonte: Do autor (2024).

Tabela 149 – Média de verdadeiro positivo(VP), falso positivo(FP) e falso negativo(FN) das relações dos algoritmos de RB, considerando estrutura de associação 3, $n = 450$, VE4 e variável resposta binária.

Algoritmos	VP	FP	FN
HC	13,16	11,05	1,84
Tabu	12,56	11,80	2,44
MMHC	12,14	2,10	2,86
RSMAX2	12,15	2,13	2,85
HC (com wl)	13,31	10,95	1,69
Tabu (com wl)	12,91	11,48	2,09
MMHC (com wl)	12,18	2,06	2,82
RSMAX2 (com wl)	12,19	2,10	2,81

Fonte: Do autor (2024).

Tabela 150 – Média dos valores AIC para cada método, considerando estrutura de associação 3, $n = 450$, VE4 e variável resposta binária y_1.

Método	AIC	Método	AIC
Backward	477,14	MMHC	510,46
Forward	477,14	RSMAX2	510,31
Bidirectional F	477,14	HC (com wl)	504,26
Bidirectional B	477,14	Tabu (com wl)	504,26
HC	504,26	MMHC (com wl)	510,48
Tabu	504,26	RSMAX2 (com wl)	510,31

Fonte: Do autor (2024).

Variável resposta y_2 - Variável explicativa com quatro categorias

Tabela 151 – Taxa de seleção das variáveis para cada método, considerando estrutura de associação 3, $n = 450$, VE4 e variável resposta binária y_2.

Acertos	Backward	Forward	Bidirectional F	Bidirectional B	HC	Tabu	MMHC	RSMAX2
Apenas as 4 variáveis	0,28	0,31	0,31	0,28	0,31	0,30	0,20	0,20
Uma das variáveis	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00
Duas das variáveis	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00
Três das variáveis	0,12	0,12	0,12	0,12	0,13	0,02	0,79	0,79
Quatro das variáveis	0,88	0,88	0,88	0,88	0,87	0,87	0,21	0,21
x1	0,18	0,15	0,15	0,18	0,15	0,15	0,00	0,00
x2	0,19	0,19	0,18	0,19	0,18	0,19	0,04	0,04
x3	0,21	0,19	0,19	0,21	0,19	0,19	0,02	0,02
x4	0,21	0,18	0,18	0,21	0,19	0,19	0,03	0,03
x5	0,17	0,17	0,17	0,17	0,17	0,17	0,01	0,01
x6	0,17	0,14	0,14	0,17	0,14	0,14	0,00	0,00
x7	0,88	0,88	0,88	0,88	0,87	0,87	0,21	0,21
x8	1,00	1,00	1,00	1,00	1,00	1,00	1,00	1,00
x9	1,00	1,00	1,00	1,00	1,00	1,00	1,00	1,00
x10	1,00	1,00	1,00	1,00	1,00	1,00	1,00	1,00

Fonte: Do autor (2024).

Tabela 152 – Taxa de seleção das variáveis para cada algoritmo de rede bayesiana considerando *whitelist*, para estrutura de associação 3, $n = 450$, VE4 e variável resposta binária y_2.

Acertos	HC	Tabu	MMHC	RSMAX2
Apenas as 4 variáveis	0,31	0,30	0,20	0,20
Uma das variáveis	0,00	0,00	0,00	0,00
Duas das variáveis	0,00	0,00	0,00	0,00
Três das variáveis	0,13	0,13	0,79	0,79
Quatro das variáveis	0,87	0,87	0,21	0,21
x1	0,15	0,15	0,00	0,00
x2	0,18	0,19	0,03	0,04
x3	0,19	0,19	0,02	0,02
x4	0,19	0,19	0,03	0,03
x5	0,17	0,17	0,01	0,01
x6	0,14	0,14	0,01	0,00
x7	0,87	0,87	0,21	0,21
x8	1,00	1,00	1,00	1,00
x9	1,00	1,00	1,00	1,00
x10	1,00	1,00	1,00	1,00

Fonte: Do autor (2024).

Tabela 153 – Média dos valores AIC para cada método, considerando estrutura de associação 3, $n = 450$, VE4 e variável resposta binária y_2.

Método	AIC	Método	AIC
Backward	477,92	MMHC	511,99
Forward	477,93	RSMAX2	511,98
Bidirectional F	477,14	HC (com wl)	505,75
Bidirectional B	477,14	Tabu (com wl)	505,74
HC	505,75	MMHC (com wl)	511,99
Tabu	505,74	RSMAX2 (com wl)	511,98

Fonte: Do autor (2024).