UNIVERSIDADE DE SÃO PAULO
Instituto de Ciências Matemáticas e de Computação

Redes Bayesianas para classificação com aprendizado via *Scoring and Restrict*: método, aplicação e comparação com métodos tradicionais

Camila Sgarioni Ozelame

Dissertação de Mestrado do Programa Interinstitucional de Pós-Graduação em Estatística (PIPGEs)

ICMC USP SÃO CARLOS

Camila Sgarioni Ozelame

Bayesian networks for classification with learning via Scoring and Restrict: method, application and comparison with traditional methods

Master dissertation submitted to the Instituto de Ciências Matemáticas e de Computação – ICMC-USP and to the Department of Statistics – DEs-UFSCar, in partial fulfillment of the requirements for the degree of the Master Interagency Program Graduate in Statistics. *FINAL VERSION*

Concentration Area: Statistics

Advisor: Prof. Dr. Francisco Louzada Neto

USP – São Carlos
June 2021

Camila Sgarioni Ozelame

Redes Bayesianas para classificação com aprendizado via *Scoring and Restrict*: método, aplicação e comparação com métodos tradicionais

Dissertação apresentada ao Instituto de Ciências Matemáticas e de Computação – ICMC-USP e ao Departamento de Estatística – DEs-UFSCar, como parte dos requisitos para obtenção do título de Mestra em Estatística – Programa Interinstitucional de Pós-Graduação em Estatística. *VERSÃO REVISADA*

Área de Concentração: Estatística

Orientador: Prof. Dr. Francisco Louzada Neto

USP – São Carlos
Junho de 2021

UNIVERSIDADE FEDERAL DE SÃO CARLOS

Centro de Ciências Exatas e de Tecnologia
Programa Interinstitucional de Pós-Graduação em Estatística

Folha de Aprovação

Defesa de Dissertação de Mestrado da candidata Camila Sgarioni Ozelame, realizada em 05/04/2021.

Comissão Julgadora:

Prof. Dr. Francisco Louzada Neto (USP)

Prof. Dr. Anderson Luiz Ara Souza (UFBA)

Profa. Dra. Lilia Carolina Carneiro da Costa (UFBA)

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

O Relatório de Defesa assinado pelos membros da Comissão Julgadora encontra-se arquivado junto ao Programa Interinstitucional de Pós-Graduação em Estatística.

AGRADECIMENTOS

Agradeço, primeiramente, a minha família, em especial a minha mãe Susana e meu pai Paulo, que nunca mediram esforços e sempre me apoiaram ao longo de toda a minha jornada, principalmente quando decidi enfrentar mais este desafio que está terminando. Agradeço imensamente meu orientador, o professor Dr. Francisco Louzada Neto, que aceitou a missão de me conduzir durante esse período e que me apresentou ao professor Dr. Anderson Ara, que foi a pessoa que tanto contribuiu e se empenhou para minha evolução e para a evolução desta pesquisa.

Agradeço aos meus professores da graduação da UFMT, que nos motivaram a persistir na carreira apesar das adversidades e que, também, sempre foram entusiastas da vida acadêmica, inspirando a mim e a muitos de meus colegas. E também aos professores do PIPGEs (UFSCar/USP) que instigam seus alunos a serem cada vez melhores.

Não menos importantes foram meus amigos, que estiveram comigo desde o curso de verão, durante as disciplinas e, que nos momentos difíceis, foram alicerce para a construção desse caminho, Deborah Stern, Jadson Marcelino, Gustavo Sabillon, Luiz Cotrim, Rocio Katy.

Um agradecimento bastante especial também ao meu namorado, Bruno Turci, que desde que entrou na minha vida, acompanhou bem de perto minha trajetória. Me apoiou incessantemente nos momentos críticos, em especial nesse período de pandemia tão complicado; a companhia e parceria, junto com nossos cães, foram fundamentais para que seguisse em frente.

Mais agradecimentos aos amigos, à Fabiana Tortorelli que me motivou a iniciar o mestrado em São Carlos, à Virgínia Jangrossi que me acolheu em sua casa no início de tudo e quando eram somente incertezas, à Marina Paiva que esteve por perto com seu amparo desde a graduação. Não poderia deixar de agradecer meu chefe, Ricardo Blanco que sempre foi tão gentil e acolhedor nesse período de jornada dupla. E aos amigos de Cuiabá, aos amigos do esporte, aos amigos de convivência diária e àqueles que estiveram, mesmo que só pontualmente, fazendo parte da rede apoio durante a caminhada.

À todos, os citados ou não, que foram importantes à sua maneira, tempo e intensidade, muito obrigada.

RESUMO

OZELAME, C. S. **Redes Bayesianas para classificação com aprendizado via *Scoring and Restrict*: método, aplicação e comparação com métodos tradicionais**. 2021. 141 p. Dissertação (Mestrado em Estatística – Programa Interinstitucional de Pós-Graduação em Estatística) – Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, São Carlos – SP, 2021.

Este trabalho é uma investigação sobre o comportamento das Redes Bayesianas (RB) discretas que visam resolver problemas de classificação. Esta metodologia é baseada em teorias dos grafos e de probabilidade, sendo as RBs definidas como um modelo gráfico probabilístico que permite visualizar as relações entre as variáveis consideradas aleatórias e, em geral, simplifica o entendimento de domínios complexos. Com o intuito de compreender seu desempenho, foram selecionados os classificadores *Naïve Bayes* (*NB*), o *Tree Augmented Naïve Bayes* (*TAN*), o *K-Dependence Bayesian Network* (*KDB*), o *Bayesian Network Augmented Naïve Bayes* (*BAN*), o *General Bayesian Network* (*GBN*) e o *Averaged One-Dependence Estimator* (*AODE*) para serem comparados. Desse modo, o *AODE*, um classificador combinado, apresenta a melhor performance preditiva em relação aos demais. Aliado a isso, foi proposta uma metodologia híbrida de estimação de rede, que tem como principal objetivo a classificação de maneira mais parcimoniosa. Os estudos de simulação conduzidos apontam que o novo método atende às expectativas de acréscimo na capacidade preditiva e indicam a redução da complexidade das relações entre as variáveis. Além disso, as aplicações em bases de dados reais auxiliam a melhor compreensão em torno da nova abordagem. Por fim, foi avaliada uma combinação entre os classificadores apresentados por meio do *stacking*, que sinalizou aumento na capacidade preditiva em relação aos classificadores analisados individualmente.

Palavras-chave: Classificadores, Redes Bayesianas, Estimação de Estrutura, Comparação, *Stacking*.

ABSTRACT

OZELAME, C. S. **Bayesian networks for classification with learning via Scoring and Restrict: method, application and comparison with traditional methods**. 2021. 141 p. Dissertação (Mestrado em Estatística – Programa Interinstitucional de Pós-Graduação em Estatística) – Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, São Carlos – SP, 2021.

This work is an investigation towards the behavior of discrete Bayesian Networks (BN) which aims to solve classification problems. This methodology is based on graphs and probability theories, and it is defined to be a probabilistic graphical model that allows the relationship visualization among (random) variables and, in general, simplifies the understanding of complex domains. To understand their performance, some classifiers were selected to be compared, such as Naïve Bayes (*NB*), Tree Augmented Naïve Bayes (*TAN*), K-Dependence Bayesian Network (*KDB*), Bayesian Network Augmented Naïve Bayes (*BAN*), and Averaged One-Dependence Estimator (*AODE*). In general, the performance of the ensemble classifier *AODE* outperforms the others. In addition, a hybrid method for structure estimation is proposed and it aims the parsimonious classification. The simulation studies show the new method fits the expectation of increase the prediction performance also, balance the number of connections among variables and, the applications in real datasets support the better understanding of the new approach. Finally, a combination of the classifiers via stacking was presented and it indicated an increase in their performances when they were compared to the methods themselves.

Keywords: Classifiers, Bayesian Networks, Structure Estimation, Comparison, *Stacking*.

SUMÁRIO

CAPÍTULO

1

INTRODUÇÃO

A incerteza é um aspecto permanentemente inerente à ciência sob inúmeras óticas de seus campos diversos como, por exemplo, sociologia, medicina, engenharia, agricultura, finanças e esportes, bem como toma sua forma com base no contexto no qual está inserida, por tanto, não existe uma maneira única e generalista de defini-la (BLAU, 2011). No cenário das ciências exatas e tecnologias, a incerteza surge na imperfeição do conhecimento sobre o que determina sistemas funcionais, na aleatoriedade de fenômenos naturais ou como o domínio do universo pode ser interpretado (NASUTION, 2018).

A busca por informação para apoiar o raciocínio sobre fatos incertos requer uma paridade entre ignorar completamente o que não se conhece e listar todas as condições de seu contexto. Esse equilíbrio seria uma sumarização de aspectos que são essenciais ao raciocínio e, sua linguagem e processos precisam estar bem estabelecidos (PEARL, 1988). Segundo Halpern (2017), a representação da incerteza é iniciada com conjuntos o que são chamados de *mundos possíveis*, ou *estados*, ou *eventos* e, a partir deles são atribuídas medidas de probabilidade que descrevem a plausibilidade daqueles eventos serem os verdadeiros.

Desta forma, os termos de sumarização da incerteza são abordados, por meio do cálculo probabilístico e a generalização desses sumários são realizados por meio de ferramentas e esquemas estudados por uma ciência contemporânea, a Estatística. Ela tem como pilar de existência a modelagem da incerteza e se utiliza do raciocínio probabilístico e, mais recentemente, da computação moderna. Todo procedimento estatístico envolve metodologias que visam a garantia de estratégias que melhor se adequam à análise estatística dos dados, por meio de processos metodológicos que envolvem a contextualização do problemas, modelagem e comunicação do conhecimento.

Leo Breiman[1] em seu artigo (BREIMAN, 2001) discute duas diferentes culturas da

[1] Leo Breiman foi físico de formação e contribuiu para a ciência de diversas maneiras, principalmente nos campos de teoria da informação e teoria de apostas, além, da teoria da probabilidade, estatística e

modelagem estatística que são distintas em suas essências. De um lado a modelagem de dados, considerando que os resultados provém de uma função de variáveis preditivas, parâmetros e ruídos aleatórios; essa função é um dos diversos modelos estatísticos e, de acordo com seu *ajuste* aos dados, é considerado uma estimação do "*blackbox*", a maneira desconhecida de onde foram gerados. De outro lado, a cultura de modelagem algorítmica a qual foca seus esforços em *predição* da variável de interesse utilizando variáveis que são, supostamente, relacionadas a ela, desconsiderando, essencialmente, a suposições da natureza de como os dados foram gerados.

Por meio da incorporação dessas áreas complementares, Judea Pearl propôs o método de Redes Bayesianas (PEARL, 1988), que podem ser, resumidamente, descritas como modelos visuais de relação probabilística entre variáveis aleatórias. Primariamente, foram criadas para auxiliar sistemas de raciocínio de inteligência artificial e, ao longo do tempo, se tornaram uma poderosa ferramenta que envolve as áreas descritiva, decisória, preditiva e de causalidade (WANG; CHENG; DENG, 2018; COX, 2014).

As Redes Bayesianas, também chamadas de redes de crenças e redes causais (NEAPOLITAN, 2004), permitem a utilização de duas abordagens inferenciais, a clássica ou a Bayesiana (KORB; NICHOLSON, 2010). Sua topologia pode proceder de literatura especializada, pode ser estimada por meio de heurísticas de estimação de estrutura e também utilizar a junção do conhecimento de especialistas e informações obtidas por meio dos dados e, além disso, pode ser capaz de lidar com dados faltantes (SINGH, 1997; RIGGELSEN, 2006; ADEL; CAMPOS, 2017).

A versatilidade dessa metodologia vai além da teoria, ela tem sido aplicada em diversas situações como em: análise dados de serviços médicos (ACID *et al.*, 2004); neurociências para entender o princípio da organização estrutural no cérebro humano, (JOSHI *et al.*, 2010); análise do comportamento de estudantes em sistemas de tutoria inteligente (MULDNER *et al.*, 2011); predição do nível risco em projetos da engenharia (CANO; MARTÍNEZ, 2011); estudo de privacidade de dados (SAMET; MIRI; GRANGER, 2013); redução de dimensionalidade (PARAMESWARAN, 2016); utilização em *Bussiness Inteligence* para a elaboração de mapas estratégicos (WANG; CHENG; DENG, 2018); análise de acidentes em estradas para entender suas causas (ZOU; YUE, 2017); análise de deslizamento de terra (PHAM *et al.*, 2018); predição de sobreviventes com câncer depois de 5 anos (POURHOSEINGHOLI; KHEIRIAN; ZALI, 2017); análise de interações em redes sociais (DING; ZHUANG, 2018); diagnóstico por imagem (RUZ; ARAYA-DÍAZ, 2018) e análise esportiva para predição (RAHMAN *et al.*, 2018).

inteligência artificial. Foi professor da UCLA e teve algumas posições de visitante nas Universidades de Stanford e Yale.

1.1 Aprendizado Estatístico

O conjunto de ferramentas utilizadas no intuito de modelar e entender dados complexos por meio do reconhecimento dos padrões observados é chamado de *Aprendizado Estatístico* (BISHOP, 2006; JAMES *et al.*, 2013). Dessa maneira, a Teoria do Aprendizado Estatístico tem por objetivo a formalização dos procedimentos e critérios utilizados para a generalização do comportamento de objetos de estudo, se baseando nos registros disponíveis (BISHOP, 2006), já o Aprendizado Estatístico de Máquina se concentra em automatizar os procedimentos formais das técnicas utilizadas (BOUCKAERT, 1995). Os métodos, em geral, concentram-se em minimizar o erro controlável proveniente do modelo, uma vez que o erro aleatório é inacessível (JAMES *et al.*, 2013).

Contudo, Bousquet, Boucheron e Lugosi (2003) cita o teorema do "Sem Almoço Grátis" (*No Free Lanch*), formalizado em Wolpert e Macready (1997), o qual oficializa a inexistência de uma melhor escolha para generalizar um comportamento pois, segundo o autor, não se entende de que forma o passado está relacionado com o futuro e, portanto, não existe restrições para isso. Mas Bousquet argumenta que para contornar essa situação, é dito que a ocorrência de um fenômeno pode ser descrita por meio de um modelo, que sob determinadas condições, pode desenvolver um raciocínio que associa regularidades ao conhecimento, para qualquer que seja a finalidade.

De acordo com James *et al.* (2013), as análises com a premissa de aprendizado, ou estimação, de padrões, podem ser alocadas em duas categorias: a de aprendizado supervisionado e não-supervisionado. No aprendizado *não-supervisionado* não existe uma variável resposta, são associados grupos de variáveis, como por exemplo na técnica de análise fatorial, ou grupos de observações conforme seu comportamento similar, como no caso do *clustering*, sendo que essa similaridade pode ser mensurada de diferentes maneiras. A outra categoria é a de aprendizado *supervisionado*, essa associa o comportamento de variáveis explicativas a um, ou mais, alvos. O objetivo desse tipo de aprendizado é representar a relação entre uma variável de interesse (chamada também de variável resposta, dependente, ou *output*) e essas variáveis preditoras (também chamadas de atributos ou *inputs*), que, em geral, são independentes entre si; para isso, estima-se uma função que descreve a maneira que a informação das preditoras refletem na resposta (HASTIE; TIBSHIRANI; FRIEDMAN, 2009; JAMES *et al.*, 2013; IZBICKI; SANTOS, 2020).

Nesse contexto, existem duas principais motivações para o desenvolvimento do aprendizado dessa função, uma delas é a **inferência** que se refere à tarefa de entender a forma na qual as variáveis independentes podem afetar a quantidade ou categoria de interesse. A outra motivação é a **predição** direcionada mais puramente a acertar o valor, ou rótulo, da variável de interesse baseado nas informações refletidas pelos dados (JAMES *et al.*, 2013). Neste sentido, para que o aprendizado supervisionado seja conduzido é necessário que se tenha um conjunto de **treinamento**, que é a amostra de desenvolvimento do modelo, e um outro de **teste**, referente a

amostra de estimação, utilizada para avaliar a performance dessa generalização de maneira justa.

Neste contexto de aprendizado supervisionado e guiada pela natureza da variável de interesse, existem também dois tipos de abordagens, a regressão que trata alvos numéricos e a classificação que lida com rótulos ou categorias na variável de interesse. Essa dissertação considera os classificadores de Redes Rayesianas, os quais são metodologias exclusivas para conduzir estudos de classificação.

Classificadores

Os classificadores possuem diversas formas propostas na literatura como por exemplo, a regressão logística, análise de discriminante linear e também os k-vizinhos mais próximos para classificação; além de métodos que são mais intensivos computacionalmente como modelos aditivos generalizados, métodos baseados em árvores, redes neurais, entre outros (JAMES *et al.*, 2013).

O foco dessa dissertação está restrito ao estudo de classificadores de Redes Bayesianas discretas, e dentre as diversas metodologias apresentadas na literatura, algumas foram selecionadas para essa dissertação como: *Naïve Bayes - NB* (RISH, 2001), *Tree Augmented Naïve Bayes - TAN* (JIANG *et al.*, 2005), *k-Dependence Bayesian Network - KDB* (SAHAMI, 1996), *Bayesian Network Augmented Naïve Bayes - BAN* (ZHANG, 2004), *Averaged One-Dependence Estimator - AODE* (WEBB; BOUGHTON; WANG, 2005), *General Bayesian Network - GBN* (CHENG; GREINER, 2001).

Objetivos

Essa dissertação tem como objetivo geral investigar e comparar diferentes metodologias de estimação de estrutura em Redes Bayesianas discretas, bem como propor procedimentos alteranativos à estes métodos. Especificamente, o trabalho se visa:

- Considerar classificadores tradicionais como: NB, TAN, KDB, BAN, GBN e AODE;
- Considerar métodos tradicionais de estimação de estrutura em Redes Bayesianas, como os métodos *K2* e *PC*;
- Propor e investigar uma combinação *K2+PC*, denominada *scoring and restrict*, em comparação com os métodos de classificação tradicionais;
- Propor e investigar os modelos acima via combinação de modelos por meio de *Stacking*;
- Explorar os resultados em dados reais e artificiais.

As próximas seções deste capítulo abordam noções básicas de teoria dos grafos, teoria das probabilidades e estatística bayesiana. No Capítulo 2, a teoria que fundamenta as Redes

Bayesianas é brevemente descrita, além dos procedimentos de aprendizado de estrutura, de parâmetros e de suas respectivas métricas. O Capítulo 3 trata da descrição dos classificadores Bayesianos. No Capítulo 5 são apresentadas os estudos de simulação conduzidos para avaliação das metodologias propostas, o Capítulo 6 demonstra o comportamento dos métodos apresentados em situações reais. O Capítulo 7 expõe a combinação dos modelos via *Stacking*. Por fim, o Capítulo 8 apresenta as considerações finais desta dissertação.

1.2 Noções de Grafos

A Teoria de Grafos se originou no século XVIII quando o matemático suíço Leonard Euler desenvolveu uma resposta para o problema das Sete Pontes de Königsberg - na Prússia, que segundo Gross, Yellen e Zhang (2013) propunha o seguinte: "Partindo de qualquer uma das quatro regiões da cidade, divididas pelo Rio *Pregel* (ou Rio Prególia) e interligadas por meio das *sete* pontes; é possível achar um caminho que atravessasse cada uma das pontes apenas uma vez, terminando também, em qualquer uma das regiões?". No mapa ilustrativo da Figura 1 as quatro regiões mencionadas estão indicadas com as letras $\{A, B, C, D\}$ e as sete pontes estão demarcadas em branco.

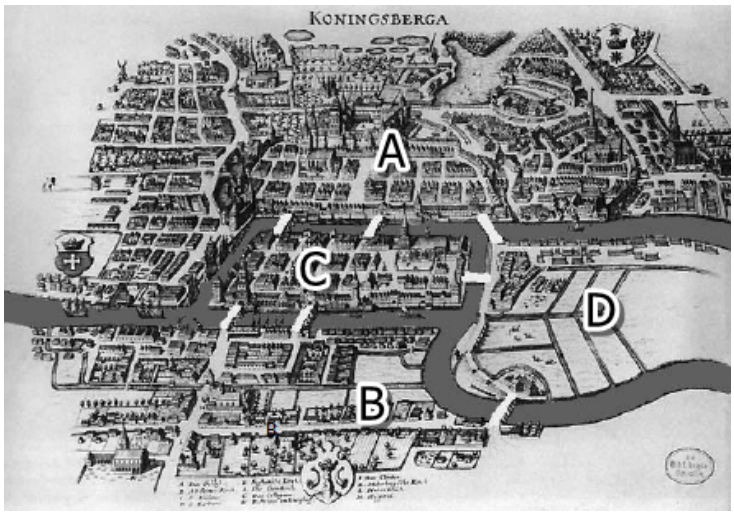


Figura 1 – Mapa ilustrativo do problema das 7 pontes de Königsberg.

Fonte: Adaptado de Handbook of Graph Theory (GROSS; YELLEN; ZHANG, 2013), página 258.

Em 1723, Euler provou que não era possível passar apenas uma vez por cada ponte e visitar todas as partes da cidade (STEEN, 2010). A representação gráfica foi apresentada apenas no século XIX por Rouse Ball que utilizou grafos para se referir a solução da questão provada por Euler.

Análogo ao mapa anterior, o esboço gráfico apresentado na Figura 2 possui as quatro regiões e as sete pontes que as ligam e foi utilizado para formalizar a teoria apresentada no século anterior. Dessa forma, o problema pode ser generalizado para outras situações, o que possibilitou que a teoria de grafos fosse desenvolvida e expandida.

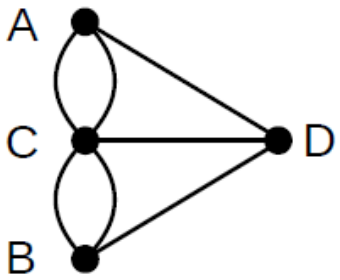


Figura 2 – Modelo gráfico para solução do problema das 7 pontes de Königsberg.

Fonte: Handbook of Graph Theory (GROSS; YELLEN, 2004), página 259.

Essa dissertação se refere ao conceito específico de grafos acíclicos e direcionados. Para explorar a teoria desse caso particular alguns conceitos serão apresentados a seguir.

1.2.1 Definição e Composição

Um grafo, ou rede, é formado por um par $\{\mathbf{V},\mathbf{A}\}$, sendo que $\mathbf{V} = \{V_1, V_2, ..., V_d\}$ é o conjunto não vazio e finito de vértices (pontos, ou nós), de tamanho d; e $\mathbf{A} = \{A_1 = (V_1, V_2), ..., A_m = (V_{d-1}, V_d)\}$ é um grupo, de tamanho m, composto por pares de elementos de $\mathbf{V}$ conectados, chamados de arcos ou arestas. Quando dois vértices compõem um par nesse conjunto de arestas, é dito que ambos são *adjacentes* e podem também ser chamados *vizinhos* (NEAPOLITAN, 2004), por exemplo, o arco A_i se estabelece como, $A_i = (V_k, V_p)$, tal definição implica que V_k e V_p são vizinhos, ou adjacentes.

Da mesma forma, arcos adjacentes são pares de vértices que possuem um elemento em comum. Para ilustrar esse conceito, consideram-se os arcos A_i e $A_j = (V_k, V_q)$ que são ditos adjacentes pois compartilham o ponto V_k. Quando, em uma aresta, os pontos inicial e final são o mesmo vértice, $A_l = (V_w, V_w)$, um *loop*, ou laço, é formado; porém, esse tipo de estrutura não é permitida nas redes que serão tratadas nesta dissertação.

As relações de vizinhança podem ser resumidas por meio de uma *matriz de adjacência* M_G, do par $\{\mathbf{V},\mathbf{A}\}$, a qual é definida como:

$$M_G(i,j) = \begin{cases} 1, & \text{se } V_i \text{ e } V_j \text{ são adjacentes,} \\ 0, & \text{caso contrário} \end{cases}$$

sendo que $1 \leq i, j \leq d$ e $\mathbf{V}$ explicitamente ordenado (GROSS; YELLEN; ZHANG, 2013).

Tanto em grafos direcionados, quanto não-direcionados, ou mesmo parcialmente orientados, sua formação é indicada pelo par $\{\mathbf{V},\mathbf{A}\}$ porém, se diferem pelos aspectos das suas conexões e das propriedades inerentes de cada formação (GROSS; YELLEN; ZHANG, 2013). Uma das características que determina a direção do grafo é a ordenação dos vértices que compõem as

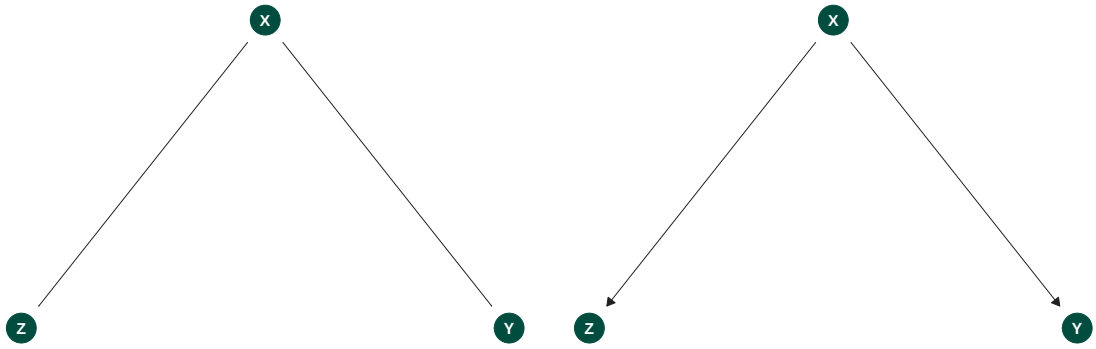
arestas. Quando cada par de nós possui uma ordem (V_i, V_j), o grafo é *direcionado*, quando não possui ordenação formam um grafo *não-direcionado* (SPRITES; GLYMOUR; SCHEINES, 2000).

O aspecto de orientação do grafo é fundamental para estabelecer qual é o tipo de relação entre os nós da rede, conforme será apresentado em seções futuras.

1.2.2 Direção

Os componentes de um grafo não direcionado são os vértices $\mathbf{V}$ e o conjunto de arestas não ordenadas $\mathbf{A}'$, convenientemente representadas da seguinte maneira, $\mathbf{A}' = \{A'_1 = \{V_1, V_2\}, \ldots, A'_m = \{V_{d-1}, V_d\}\}$, elas estabelecem relações simétricas entre os nós. Considerando, para fins de ilustração, os vértices $\{X, Y, Z\}$ e o conjunto de arcos $\mathbf{A}' = \{A'_1 = \{X, Y\}, A'_2 = \{X, Z\}\} = \{A'_1 = \{Y, X\}, A'_2 = \{Z, X\}\}$ a representação gráfica desse par é dada pela Figura 3a, um grafo não direcionado.

Por outro lado, um *grafo direcionado*, ou dígrafo, descreve relações assimétricas entre vértices e para tanto, possui a característica de possuir apenas *pares ordenados distintos* no conjunto de conexões, ou seja, $\mathbf{A} = \{A_1 = (V_1, V_2), \ldots, A_m = (V_{d-1}, V_d)\}$ (GROSS; YELLEN; ZHANG, 2013) é sua coleção de arcos. Sua direção é graficamente representada por setas, e sua matriz de adjacência não é simétrica. Por exemplo, na Figura 3b observa-se um dígrafo formado pelo mesmo conjunto de vértices do caso anterior, $\{X, Y, Z\}$, e um conjunto com dois arcos direcionados $\mathbf{A} = \{A_1 = (X, Y), A_2 = (X, Z)\}$, se a ordem dos componentes das arestas se modifica. Isso é refletido na representação gráfica com a produção de um dígrafo distinto.



(a) Não-direcionado. (b) Direcionado.

Figura 3 – Exemplos de orientação em grafos

Fonte – Elaborado pela autora.

Em grafos direcionados, os vértices são ligados partindo da chamada cauda (*tail*) para a cabeça (*head*), para onde a seta se direciona. Para se referir às relações estabelecidas entre os vértices a terminologia de *parentesco* é bastante propagada na literatura que utiliza essa

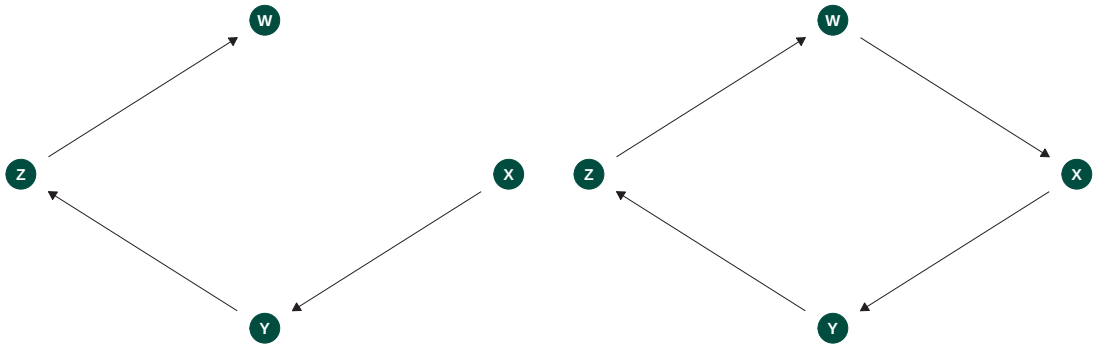
orientação como forma de interpretação da ferramenta, como os manuscritos a respeito das próprias Redes Bayesianas (PEARL, 1988; NEAPOLITAN, 2004; KORB; NICHOLSON, 2010).

Dessa forma, voltando ao exemplo da Figura 3b, se existe uma seta ligando o vértice X ao vértice Y, X é dito ser *pai* de Y pois o arco A_1 parte de X para Y, pelo mesmo motivo Y é dito *filho* de X (SPRITES; GLYMOUR; SCHEINES, 2000), analogamente, também é válido para o arco A_2. O conjunto composto pelos pais de um vértice V_i será denotado por $pa(V_i)$ e o conjunto de filhos de V_i por $ch(V_i)$. Existem ainda outras noções que são definidas por meio da ordenação dos pontos; como a de *esposos*, quando dois nós compartilham algum filho; *ancestrais* são os vértices que antecedem as conexões diretas e descendentes são aqueles ordenados após o nó de referência.

Outro conceito importante que tange o segmento de orientação em grafos é o de *esqueleto*. O esqueleto de um grafo direcionado G é um grafo não-direcionado E, que contém todos seus vértices e arcos, porém, por sua definição, não ordenados (KOLLER; FRIEDMAN, 2009).

1.2.3 Ciclos

Para iniciar a definição de um ciclo em um grafo, considera-se um trajeto W como sendo uma sequência alternada de vértices e arcos, cujo o tamanho é dado pela quantidade de arcos presentes no conjunto. Também pode ser descrito, em um grafo simples, por uma sequência de vértices que quando direcionados, formam um trajeto direcionado.



(a) Caminhos Direcionados. (b) Ciclo.

Figura 4 – Caminho.

Fonte – Elaborado pela autora.

Quando, nesse trajeto, nenhum arco se repete ele é chamado de trilha e quando, em uma trilha, nenhum vértice interno - que não é nem inicial, nem final - se repete, ela é chamada caminho (*path*); essa teoria é análoga para grafos direcionados. E além disso, quando um caminho tem o mesmo vértice inicial e final, é dito que se tem um *ciclo* (STEEN, 2010).

Os conceitos de caminho direcionado e ciclo estão ilustrados na Figura 4. O caminho direcionado, Figura 4a, é composto pelos vértices $\{X,Y,Z,W\}$ e possui 3 arestas, ele se inicia

em X e termina em W. Já na Figura 4b, com 4 arestas, mesma quantidade de vértices, logo, partindo de qualquer um dos nós é possível retornar a ele pois, se trata de um caminho fechado, um ciclo. O mesmo não acontece com os caminhos direcionados.

1.2.4 Grafos Acíclicos Direcionados

Um Grafo Acíclico Direcionado, (DAG - D*irected* A*cyclic* Graph) G, é composto de dois elementos $G = (\mathbf{V}, \mathbf{A})$, sendo $\mathbf{V}$ o conjunto de vértices e $\mathbf{A}$ o conjunto de arestas ordenadas.

Ilustrando esse conceito por meio da Figura 5, considera-se que o conjunto de vértices é constituído por seis nós, $\mathbf{V} = \{T, U, W, X, Y, Z\}$ e cinco arcos $\mathbf{A} = \{A_1 = (T, X), A_2 = (X, Y), A_3 = (U, Y), A_4 = (Y, Z), A_5 = (Z, W)\}$. Sua representação gráfica indica a orientação das ligações por meio de setas e não existem ciclos, ou seja, partindo de qualquer um dos nós, não é possível chegar a ele novamente por nenhum caminho existente.

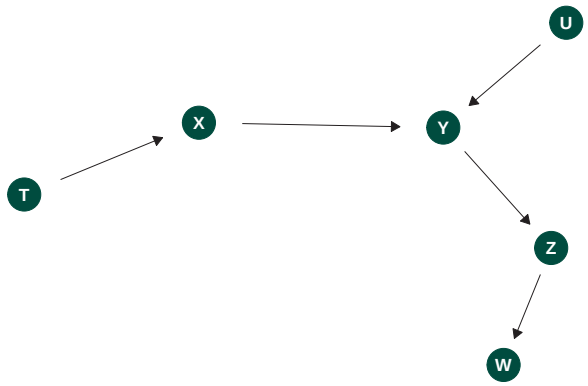


Figura 5 – Exemplo de um DAG.

Fonte – Elaborado pela autora.

Nesse contexto, a terminologia de parentesco é utilizada para entender a estrutura completa da rede. Para o exemplo apresentado anteriormente e tomando Y como referência, o conjunto de seus *ancestrais*, ou antecessores, é formado por $\mathbf{an}(Y) = \{T, X, U\}$, esses nós precedem o vértice Y em um caminho existente no DAG e do qual façam parte. Analogamente, Y é chamado de *descendente* dos componentes de $\mathbf{an}(Y)$ e também possui um conjunto de descendentes $\mathbf{des}(Y) = \{Z, W\}$ (NEAPOLITAN, 2004).

Além do mais, o conjunto de pais de Y é $\mathbf{pa}(Y) = \{X, U\}$, os vértices que compõem esse grupo são chamados de esposos já que possuem um filho em comum. E também, Y tem como filho o vértice Z, essas relações são de adjacências no grafo.

A teoria de grafos possui conceitos e definições que se adaptam a diversas circunstâncias, contudo, esse texto não aborda sua completude. Para maiores detalhes e aprofundamento nessa área existem referências como Gross, Yellen e Zhang (2013) e Steen (2010) que tratam especificamente de teoria de grafos, Lauritzen (1996) que apresenta a teoria e também aplicação

de conceitos nos modelos gráficos bem como Pearl (2013), Bishop (2006) e Sprites, Glymour e Scheines (2000) com conceitos aplicados.

1.3 Noções de Probabilidades

A filosofia, que permeia e é a origem das ciências, tem como propósito generalista a solução de problemas conceituais e teóricos sobretudo os que se referem a fundamentos de conhecimento e valores, conforme o livro '*O lugar da probabilidade na ciência* (EELLS; FETZER, 2010). Nessa obra, os autores discorrem a respeito da relevância dos conceitos de probabilidade na filosofia da ciência, em especial a ideia de explicação probabilística que é essencial em diversas áreas.

Em Jaynes (2003) o autor disserta em torno do raciocínio plausível do cérebro humano e dos processos de raciocínio dedutivo para introduzir conceitos formais de probabilidade, conforme Laplace, em 1918, resume a teoria como sendo *"...senso comum reduzido a cálculo."*. Dessa maneira, a construção da lógica probabilística é resultado de observação da natureza, construção de hipóteses e experimentos com diferentes resultados utilizados para fornecer uma explicação sobre leis da natureza, as quais, em teoria, são desconhecidas e imutáveis (SALMON, 1998; NEAPOLITAN, 2004).

O estudo da teoria da probabilidade, ao contrário de outros ramos da matemática, possui uma origem recente, estimada no renascimento com Girolame Cardano e depois com Galileu Galilei (TABAK, 2004). Porém, apenas no século XVII com Pierre de Fermat e Blaise Pascal, dois matemáticos franceses, é que o cálculo das probabilidades começa a ter relevância; eles tinham o intuito de investigar problemas de imprecisão e incerteza trazidos pelos jogos de azar (KORB; NICHOLSON, 2010; TABAK, 2004).

Mas somente em 1933 que Andrei Nickolaevitch Kolmogorov formalizou o cálculo de probabilidade desenvolvendo um conjunto de textos na teoria de probabilidade moderna (NUALART, 2004). O conjunto de princípios baseados em teoria da medida fundamentam matematicamente as aplicações de probabilidade (NEAPOLITAN, 2004). Kolmogorov define noções básicas como a de espaço probabilístico, experimento aleatório, variável aleatória, e funções de probabilidade, para nortear seus axiomas e embasar conceitos mais complexos.

Em torno desse assunto, livros e outras obras foram elaboradas detalhando a teoria de probabilidade e seus desdobramentos. Algumas delas são mais filosóficas como Eells e Fetzer (2010), Tabak (2004), outras são teóricas como Jaynes (2003), Nualart (2004), Box e Tiao (1973), DeGroot e Schervish (2012), DeGroot (2004).

Alguns conceitos mais básicos de probabilidade não serão detalhados nessa dissertação, mas podem ser encontrados em qualquer referência teórica formal do assunto de probabilidade, este texto está direcionado aos princípios utilizados para compreensão do objeto de estudo, como

fundamentos de probabilidade condicional.

As próximas seções apresentam conceitos básicos de probabilidade condicional e abordagem Bayesiana de probabilidade.

1.3.1 Probabilidade Condicional

A definição clássica de probabilidade condicional, considerando os eventos A e $B \in \mathscr{T} \subset \Omega$, tal que $P(B) > 0$, é dada por:

$$P(A|B) = \frac{P(A,B)}{P(B)}, \tag{1.1}$$

sendo $P(A,B)$ a probabilidade da ocorrência de A e B simultaneamente, em termos de teoria dos conjuntos $A \cap B$. $\mathscr{T}$ é definido como sendo o campo de Borel, ou σ-álgebra de subconjuntos de Ω, espaço amostral, e que em conjunto com a probabilidade P formam o espaço probabilístico de medida (NUALART, 2004).

Essa formulação expressa a medida de *dada a ocorrência do evento B, o evento A também ocorra* (KORB; NICHOLSON, 2010). Ou seja, é necessário considerar a ocorrência de um evento A quando uma informação adicional sobre o resultado pode ser obtida pela ocorrência de um outro evento B (DEGROOT, 2004).

Esse conceito condicional corrobora na definição de *independência*, ou independência marginal, o qual diz que dois eventos A e B podem ser considerados probabilisticamente independentes se condicionando em um deles, a probabilidade do outro permanece inalterada (KORB; NICHOLSON, 2008). Sem perda de generalidade, pode ser expresso da seguinte maneira:

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

Nesse contexto, a generalização para *independência condicional* é análoga, de forma que A e B são independentes dado C se o cálculo de sua probabilidade puder ser expresso por:

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

A igualdade $P(A|B,C) = P(A|C)$ é generalização pois no caso de C ser qualquer valor em Ω, a expressão se reduz, a expressão se reduz a independência marginal novamente (KORB; NICHOLSON, 2010).

Teorema de Bayes

Permancendo no segmento da teoria de probabilidade condicional e considerando $B_1, \ldots, B_k$ partições do espaço amostral Ω sendo $P(B_j) > 0$ para qualquer j em $\{1, \ldots, k\}$,

é possível reescrever a Equação (1.1) da seguinte maneira:

$$P(B_i|A) = \frac{P(B_i)P(A|B_i)}{\sum_{j=1}^{k} P(B_j)P(A|B_j)}. \tag{1.2}$$

O teorema de Bayes permite calcular a probabilidade condicional de cada evento em uma partição dada uma evidência, um evento A observado (NEAPOLITAN, 2004). A principal utilização de partições é para dividir o espaço amostral em parte suficientemente menores fazendo com que a coleção de eventos de interesse se torne condicionalmente independente dado cada evento da partição (DEGROOT; SCHERVISH, 2012).

Conforme visto, probabilidades condicionais se comportam da mesma forma que probabilidades, então, a seguinte extrapolação é verdadeira, considerando um outro evento C:

$$P(B_i|A \cap C) = \frac{P(B_i|C)P(A|B_i \cap C)}{\sum_{j=1}^{k} P(B_j|C)P(A|B_j \cap C)}. \tag{1.3}$$

Probabilidades associadas a ocorrência de eventos de interesse têm sido calculadas nas últimas décadas a partir de evidências - probabilidades conhecidas - graças ao teorema de Bayes (NEAPOLITAN, 2004). Supondo que existe interesse em alguns eventos disjuntos $B_1, \ldots, B_k$ acontecerão e apenas será observado um outro evento A, se a condicional de $P(A|B_i)$ estiver disponível para cada i então, esse teorema fornece as probabilidades condicionais dos B_i eventos dado A (DEGROOT; SCHERVISH, 2012).

1.4 Abordagem Bayesiana

Thomas Bayes, em foto apresentada na Figura 6, foi um pastor presbiteriano e matemático amador inglês a quem é dada a referência por ter iniciado os estudos bayesianos no século XVIII por meio da publicação póstuma de "*An essay towards solving a problem in the doctrine of chances*", em tradução livre, "Um esboço buscando resolver um problema na doutrina das probabilidades".

Conforme Dale (1999), a obra de Bayes, datada do ano de 1973, marca o nascimento do estudo de *probabilidade inversa* medidas referentes ao que o autor denomina problemas inversos, que são denominados assim por serem o oposto dos problemas ditos diretos, os quais mapeiam de um conjunto de objetos possíveis até o conjunto de todos os dados possíveis. Por problema inverso, entende-se então, que é a determinação de um objeto proveniente de um conjunto de dados.

Heckerman (2008) diz que a probabilidade clássica é uma propriedade física do mundo real, já a probabilidade Bayesiana é uma propriedade do indivíduo o qual determina essa quantidade, por isso a medida pode ser chamada de grau de crença. Segundo Korb e Nicholson (2010), o Bayesianismo pode ser encarado como uma generalização das contas físicas - também conhecidas

Figura 6 – Reverendo Thomas Bayes (1702–1761).

Fonte – Bayesian Artificial Intelligence, página 11 (KORB; NICHOLSON, 2010)

por frequentistas ou clássicas - de probabilidade, isso significa que é perfeitamente compatível com a ótica Bayesiana de probabilidade de como medir graus de crença para estabelecer uma probabilidade subjetiva a partir da frequência de um resultado.

Uma das principais diferenças entre ambas as abordagens, é que a Bayesiana, ao contrário da frequentista, não requer repetidas tentativas, exprimindo as crenças pela análise das distribuições a priori, posteriori e preditiva (PERKINS; WANG, 2004).

De acordo com Wasserman (2013), a conjectura Bayesiana se apoia em três principais postulados: (1) A probabilidade descreve grau de crença e não frequência. Dessa forma, é possível fazer afirmações probabilísticas sobre várias situações, não somente dados, o que é inerente da variação aleatória; (2) É possível fazer afirmações probabilísticas sobre parâmetros, mesmo que eles sejam constantes fixadas; e (3) A inferência de um parâmetro θ pode ser feita produzindo a distribuição de probabilidade para θ. E valores de interesse como estimativas pontuais e intervalares, são retiradas dessa distribuição.

Análise Bayesiana

O método Bayesiano pode ser sumarizado da seguinte maneira, segundo Wasserman (2013):

1. Determina-se a função de probabilidade $P(\theta)$, chamada de *distribuição a priori*, que expressa as crenças em torno de um parâmetro, ou conjunto de parâmetros, θ, sem a observação da amostra;

2. Escolhe-se um modelo $P(\mathbf{X}|\theta)$ que reflete as crenças sobre o conjunto de variáveis aleatórias $\mathbf{X}$ dado θ. Essa é a função pela qual os dados $\boldsymbol{X}$ modificam o conhecimento a respeito de θ, é chamada de *função de verossimilhança* (BOX; TIAO, 1973);

3. Depois da observação dos dados $x_1, \ldots, x_n$, as crenças são *atualizadas* e a distribuição $P(\theta|x_1, \ldots, x_n)$ é calculada. Essa distribuição é chamada *distribuição a posteriori*, ela descreve o que se tem de informação sobre θ e é uma das características mais importantes da análise Bayesiana (BICKEL; DOKSUM, 2015).

Tais passos, que serão detalhados a seguir, estão compactados na expressão:

$$\underbrace{P(\theta|\mathbf{X})}_{\text{posteriori}} \propto \underbrace{P(\mathbf{X}|\theta)}_{\text{verossimilhança}} \underbrace{P(\theta)}_{\text{priori}}. \tag{1.4}$$

Desde o Teorema de Bayes apresentado na Equação (1.2), com alguma noção de probabilidade e supondo que existe o vetor de n obervações $(x_1, \ldots, x_n)$, com distribuição discreta de probabilidade $P(\mathbf{X}|\theta)$ que depende do valor de d parâmetros $\theta = (\theta_1, \ldots, \theta_d)$, e supondo também que θ possui uma distribuição de probabilidade $P(\theta)$, então (BOX; TIAO, 1973):

$$P(\mathbf{X}|\theta)P(\theta) = P(\mathbf{X}, \theta) = P(\theta|\mathbf{X})P(\mathbf{X}). \tag{1.5}$$

Dada informação observada $\boldsymbol{X}$, a distribuição condicional de θ é:

$$P(\theta|\mathbf{X}) = \frac{P(\mathbf{X}|\theta)P(\theta)}{P(\mathbf{X})}. \tag{1.6}$$

Assim, $P(\mathbf{X})$, chamada de verossimilhança marginal, pode ser escrita como:

$$P(\mathbf{X}) = \mathbb{E}[P(\mathbf{X}|\theta)] = c^{-1} = \sum P(\mathbf{X}|\theta)P(\theta). \tag{1.7}$$

sendo que a soma é admitida para toda amplitude do espaço de θ e, portanto, pode ser reescrita como:

$$P(\theta|\mathbf{X}) = cP(\mathbf{X}|\theta)P(\theta), \tag{1.8}$$

logo, c é a constante normalizadora.

O domínio de θ pode ser explorado pela definição de distribuições de probabilidade $P(\theta)$, chamadas de prioris e os exemplos mais comuns dessas funções estão apresentadas a seguir.

1.4.1 Prioris

Para DeGroot e Schervish (2012), as prioris são definidas como distribuições atribuídas ao parâmetro, ou conjunto de parâmetros, sem a observação de outras variáveis relacionadas a ele. Elas podem ser chamadas também de distribuições marginais dos parâmetros, e são quaisquer funções de probabilidade que expressam algum conhecimento prévio no que diz respeito aos parâmetros desconhecidos de um modelo (BOX; TIAO, 1973).

As informações *a priori* podem auxiliar na avaliação do nível de plausibilidade em um objeto de estudo (JAYNES, 2003), e representam a incerteza formal a respeito dos parâmetros. São baseadas em suposições sobre o modelo ou também, são construídas a partir de informação dos dados e portanto, carregam conhecimento, ainda que pouco ou vago, a respeito dos parâmetros. Alguns tipos mais difundidos de prioris e suas respectivas construções estão apresentadas a seguir.

A primeira é a priori conjugada, e o que caracteriza a família de distribuições conjugadas é que se a priori for membro desse grupo, depois de multiplicada pela verossimilhança, a posteriori também será parte da família (BERGER, 2013), fazendo com que, tipicamente, os cálculos sejam simplificados (DEGROOT; SCHERVISH, 2012).

Para distribuições discretas de probabilidade, o que é mais frequentemente utilizado é o processo *Dirichlet-Multinomial* no qual se assume uma distribuição multinomial para os dados e uma distribuição Dirichlet para os respectivos parâmetros.

De acordo com Tu (2014), a forma da distribuição multinomial para um conjunto de variáveis aleatórias discretas não-negativas $\mathbf{X}$ de dimensão d é a seguinte:

$$f(x_1,\ldots,x_d;p_1,\ldots,p_d,n) = \frac{\Gamma(n+1)}{\prod_{i=1}^{d}\Gamma(x_i+1)}\prod_{i=1}^{d}p_i^{x_i},$$

sendo que $n=\sum_{i=1}^{d}x_i$ e é denotada por $Multi(p_1,\ldots,p_d,n)$.

Da mesma forma, a distribuição Dirichlet parametrizada por escalares positivos α_i

$$f(x_1,\ldots,x_d;\alpha_1,\ldots,\alpha_d) = \frac{\Gamma(\sum_{i=1}^{d}\alpha_i)}{\prod_{i=1}^{d}\Gamma(\alpha_i)}\prod_{i=1}^{d}x^{\alpha_i-1}. \tag{1.9}$$

É denotada por $Dir(\alpha_1,\ldots,\alpha_d)$ e definida no espaço $d-1$ uma vez que o suporte da distribuição Dirichlet é um simplex $\mathscr{S}_d$ de dimensão $d-1$, que significa que todos os d vetores foram uma distribuição de probabilidade válida (TU, 2014).

Dessa forma, sua posteriori é dada por uma distribuição também Dirichlet, conforme definição, com uma reparametrização, tal que:

$$\alpha'_j = \alpha_i + \sum_{x_i\in D} x_i^{(j)},$$

sendo que D se refere a base de dados observados.

Esses vetores α_i que definem a reparametrização da distibuição a posteriori para o método Multinomial-Dirichlet são considerados um parâmetro de concentração (PETITJEAN *et al.*, 2018) para cada uma das variáveis X_i. Para o caso das Redes Bayesianas, os vetores α são associados a cada uma das configurações de variáveis e seus respectivos pais, conforme é tratado no próximo capítulo.

Outro importante exemplar é a priori de Jeffreys, que foi proposta em Jeffreys (1946). Ela é considerada localmente uniforme, portanto, não informativa, e dita invariante $1-1$ com

respeito a reparametrizações, motivo da sua proposição. Para uma distribuição $P(X|\theta)$, sua forma é dada por $P(\theta) \propto |I(\theta)|^{\frac{1}{2}}$, sendo que, $I(\theta)$ é matriz de informação de Fisher, e portanto, é considerada uma priori objetiva uma vez que depende da verossimilhança para que sua forma seja definida.

Um aspecto importante da priori de Jeffreys é que, em grande parte das análises, seu uso leva a distribuições próprias a posteriori (YANG; BERGER, 1996). A priori de Jeffreys para a distribuição Multinomial, por exemplo, resulta na $Dirichlet(\frac{1}{k})$, sendo que k é o número de possíveis resultados da variável aleatória, ou seja, a conjugada Multinomial-Dirichlet para $alfa = \frac{1}{k}$. De acordo com Consonni *et al.* (2018) existem algumas variações da priori de Jeffreys para desempenho em situações específicas, como a de alta dimensão e a utilização para quaisquer funções de parâmetros.

Existem outros tipos de prioris propostas como as prioris vagas. Essas distribuições do tipo *flat* podem auxiliar quando nenhuma informação a respeito dos parâmetros está disponível, ou seja, quando toda a inferência deve ser feita baseada somente nos dados. O termo priori *flat*, ou priori *plana*, foi utilizado sob o senso que sua forma envolve funções com decaimento lento da cauda (SCHMITT; GILARDONI; ANDRADE, 2019) e é caracterizada por $V(\theta) \to \infty$, o que descreve que a variância é grande o suficiente para que os seus valores não estejam concentrando a informação da distribuição em uma única região. Um exemplo difundido dessa classe é a uniforme, definida por $P(\theta) \propto c$, possuindo igual probabilidade para intervalos de mesmo tamanho (CONSONNI *et al.*, 2018).

Maiores detalhes a respeito de prioris e análise Bayesiana em Yang e Berger (1996), Jaynes (2003), Box e Tiao (1973), DeGroot e Schervish (2012), Consonni *et al.* (2018).

1.4.2 Modelos Gráficos Probabilísticos

Os modelos gráficos tem sua origem em diferentes áreas da ciência, segundo Lauritzen (1996) uma dessas áreas é a física estatística com Gibbs (GIBBS, 1902). Um pouco mais tarde, o geneticista Sewall Wright iniciou o desenvolvimento de estudos gráficos a partir de 1920, denominados análise de caminho. Wright investigava propriedades genéticas hereditárias de espécies utilizando componentes causais e não causais baseados em hipóteses (SCHEINER; GUREVITCH, 2001). Nessa aplicação, os grafos determinam caminhos e para isso, devem ser direcionados (LAURITZEN, 1996).

Modelos gráficos probabilísticos utilizam uma abordagem baseada em grafos que decifram, de maneira compacta, distribuições complexas de probabilidade em um espaço de alta dimensão (KOLLER; FRIEDMAN, 2009). Bishop (2006) avalia as vantagens da sua utilização apontando que, conciliar elementos probabilísticos e elementos gráficos facilita a manipulação de modelos sofisticados uma vez que, por meio de propriedades gráficas e matemáticas, carregam suas sutilezas implicitamente.

Essa metodologia de análise é composta por uma distribuição de probabilidade e uma estrutura gráfica. Ambos os elementos se relacionam entre si, os nós do grafo são as representações das variáveis aleatórias do domínio da distribuição e suas conexões podem sugerir afinidade compartilhada ou mesmo dependência entre elas.

Nesse sentido, esses modelos podem ser divididos em duas grandes classes, uma delas se utiliza de distribuição de probabilidade associada à um *grafo não direcionado*, são as denominadas *Redes de Markov*, suas conexões traduzem o grau de afinidade que duas variáveis compartilham. De maneira geral, é um modelo paramétrico utilizado para representar uma categoria específica de distribuições de probabilidade, chamadas de distribuições de *Gibbs* (EDERA; STRAPPA; BROMBERG, 2014).

A outra classe, por sua vez, tem como componente um grafo *direcionado*, tipo de grafo utilizado em *Redes Bayesianas*. Os arcos de sua rede sugerem relações de dependência condicional entre os nós e para tanto, restringem a estrutura a ser *acíclica*, levando o nome de *DAG* designado para grafos acíclicos e direcionados.

Os DAGs foram definidos na Seção 1.2.4 e podem ser utilizados para representar relações de independência condicional dentre variáveis em uma distribuição de probabilidades (SPRITES; GLYMOUR; SCHEINES, 2000). Essa definição de independência condicional é ponto focal da *modelagem causal* (PEARL, 2013) portanto, o DAG é uma ferramenta utilizada para a visualização dessas relações de causa-efeito sugerida pela orientação das arestas de estrutura. Nesse sentido, a relação entre nós que compartilham um arco não é simétrica mas sim, de influência direta de um para outro - do pai para o filho (RUSSELL; NORVIG, 2010).

A modelagem gráfica auxilia a compreensão de sistemas complexos pois simplifica o trabalho de representação da distribuição conjunta em estruturas locais (ALPAYDIN, 2010), o que facilita a análise das variáveis. Pela sua versatilidade e flexibilidade de modelos, as Redes Bayesianas são amplamente utilizadas com objetivos diversos.

1.5 Comentários gerais

Neste capítulo foram apresentados os conceitos nos quais os componentes das Redes Bayesianas se baseiam, a teoria de grafos e a teoria de probabilidade, além de uma breve introdução ao conceito de modelos gráficos probabilísticos onde as Redes Rayesianas são fundamentadas.

O próximo capítulo descreve as propriedades relacionadas a fusão dessas técnicas que constituem as Redes Bayesianas bem como definições chave, propriedades, processos de estimação de estrutura e parâmetros desses modelos.

CAPÍTULO

2

REDES BAYESIANAS

As Redes Bayesianas surgiram em Pearl (1988), desenvolvidas por Judea Pearl, filósofo e cientista da computação, com o intuito de facilitar as tarefas de predição, indução e dedução em sistemas de inteligência artificial (PEARL, 2000). Os trabalhos de Judea Pearl têm sido referência em estudos probabilísticos nas áreas de análise contrafactual, equações estruturais, causalidade e inteligência artificial.

Conforme a teoria dos grafos, uma Rede Bayesiana (RB) é uma classe especial de modelos gráficos que representa as dependências probabilísticas entre um conjunto de variáveis na forma visual de um grafo acíclico e direcionado (NAGARAJAN; SCUTARI; LÈBRE, 2013). Elas podem performar tarefas de raciocínio diagnóstico, preditivo, e até mesmo causal (KORB; NICHOLSON, 2010), além de poderem ser estendidas para diagramas de influência (BARBER, 2012), por exemplo.

Sua topologia pode advir de *experts* no sistema estudado ou, pode ser estimada de maneira irrestrita, por meio de heurísticas (HECKERMAN; GEIGER; CHICKERING, 1995) que se utilizam de procedimentos de maximização de funções objetivos, testes de independência, ou mesmo uma combinação entre elas. Com respeito a essa estrutura, a estimação dos seus parâmetros pode ser conduzida, completando assim, os dois elementos essenciais que definem uma Rede Bayesiana.

De uma forma geral, elas recebem o nome ”*Bayesianas*” pois se utilizam da implementação do Teorema de Bayes para construir um mecanismo de análise baseado em fragmentos de evidências (DAWID *et al.*, 1999). Apesar disso, podem ser analisadas como frequentistas ou Bayesianas conforme abordagem utilizada para o aprendizado de seus elementos.

Nas próximas seções serão introduzidos os conceitos e propriedades de uma Rede Bayesiana, bem como as heurísticas utilizadas para a estimação de sua estrutura e metodologias para a estimação de seus parâmetros.

2.1 Definições e Propriedades

Formalmente, as Redes Bayesianas são modelos gráficos probabilísticos compostos, essencialmente, por dois elementos:

1. A estrutura gráfica, $\mathbb{G}$, a qual é definida como sendo um *DAG* (do inglês, ***D**irected **A**cyclic **G**raph*, um grafo acíclico e direcionado), pois esse objeto não permite ciclos, nem arcos não direcionados. O DAG, por sua vez, se estabelece por dois conjuntos: o de nós e o arcos que conectam esses nós;

2. A distribuição conjunta de probabilidades P, que é obtida por meio das variáveis aleatórias $\mathbf{X} = \{X_1, X_2, X_3, \ldots, X_d\}$, com respeito aos respectivos conjuntos de parâmetros em Θ.

Os componentes desses elementos se relacionam de modo que os nós são as representações gráficas das variáveis aleatórias e os arcos correspondem as relações de dependência entre elas. A leitura das independências condicionais pode ser explorada por meio das propriedades markovianas (KALISCH; BÜHLMANN, 2007). Estabelecidas essas correspondências, é possível referir-se ao DAG como $\mathbb{G} = (\mathbf{X}, \mathbf{A})$, sendo $\mathbf{A}$ o conjunto de arcos, e a função de probabilidade condicional pode ser denotada por $P(\mathbf{X}, \Theta)$.

Visto isso, segundo Neapolitan (2004), o par ($\mathbb{G}$,P) satisfaz a **Condição de Markov** se P puder ser escrita como o produtório das distribuições condicionais de cada uma das variáveis aleatórias, e assim, portanto, é formalmente definido como uma Rede Bayesiana. Essa premissa verbaliza o que está expresso na Equação (2.1):

$$P(\mathbf{X}) = \prod_{i=1}^{d} P(X_i | \mathbf{pa}_{\mathbb{G}}(X_i)), \tag{2.1}$$

sendo que $P(X_i|\mathbf{pa}_{\mathbb{G}}(X_i))$ é a distribuição *local* de probabilidade de X_i, ou seja, uma distribuição marginal condicionada às variáveis as quais X_i é dependente. Essas variáveis compõem o conjunto de *pais* de X_i e é representado por $\mathbf{pa}_{\mathbb{G}}(X_i)$, que está, necessariamente, sujeito às conexões em $\mathbb{G}$ (KOLLER; FRIEDMAN, 2009).

Essa compatibilidade entre os elementos gráficos e probabilísticos permite a representação, inclusive, de distribuições conjuntas complexas de probabilidade (ALMOND *et al.*, 2015) e, além disso, tal característica se mantém para variáveis aleatórias contínuas, de forma que a fatorar as funções de densidade conjunta f, expressa da seguinte forma:

$$f(\mathbf{X}) = \prod_{i=1}^{d} f(X_i | \mathbf{pa}_{\mathbb{G}}(X_i)),$$

condição assegurada também para distribuições que misturam variáveis de ambas as naturezas (NAGARAJAN; SCUTARI; LÈBRE, 2013).

Apesar da adaptabilidade da metodologia, essa dissertação trata de Redes Bayesianas *discretas*, as quais podem receber apenas variáveis que possuem uma quantidade de estados finito (NIELSEN; JENSEN, 2009), sejam eles numéricos ou categóricos.

Uma vez que em uma Rede Bayesiana, todas as variáveis são assumidamente aleatórias, consideradas como os nós na rede, e a dependência condicional entre elas é representada pelos arcos direcionados, a ausência desses arcos indica independência condicional (ABELLÁN *et al.*, 2006), essa implicação é fundamentada pelas propriedades apresentadas neste capítulo.

Para exemplificar os conceitos apresentados, toma-se uma base de dados hipotética contendo informações a respeito do contágio de uma doença respiratória e é composta das seguintes variáveis dicotômicas, as quais possuem o domínio binário da forma $\{0, 1\}$, indicando *sim* ou *não* respectivamente:

- *Diagnóstico* (D);
- *Distanciamento Social* (I);
- *Uso de Máscara* (M);
- *Presença de Sintomas* (S);
- *Trasmissão* (T).

As relações entre essas variáveis foi, supostamente, determinada por um profissional da área saúde que sugeriu a Rede Bayesiana expressa na Figura 7, estruturada da seguinte forma: a *Transmissão* tem como pais as variáveis *Diagnóstico*, *Distanciamento Social* e *Uso de Máscara*, e a variável *Diagnóstico* é filha da *Presença de Sintomas*.

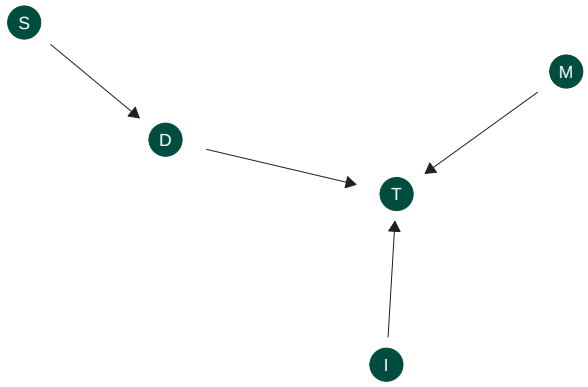


Figura 7 – Exemplo de uma Rede Bayesiana.

Fonte – Elaborado pela autora.

De acordo com a definição da Equação (2.1), é possível escrever a distribuição global de uma Rede Bayesiana como um produto das distribuições locais de suas variáveis, as quais são condicionadas aos seus respectivos pais conforme o grafo, da seguinte forma:

$$P(D,I,M,S,T) = P(S) \times P(D|S) \times P(I) \times P(M) \times P(T|D,I,M) \tag{2.2}$$

A Equação (2.2) explicita a dependência visualizada nas conexões gráficas. Então, a Transmissão é condicionada ao Uso de Máscara, ao Distanciamento Social e ao Diagnóstico da doença respiratória. Entende-se também que os sintomas influenciam o Diagnóstico, isso pode ser motivado pela suposição de que a maior parte das pessoas que fazem o teste da doença, são aquelas que apresentam sintomas.

Além disso, o exemplo anterior apresenta uma estrutura de rede que é conhecida, ou seja, é determinada por um profissional da área com uma *expertise* no assunto tratado. Contudo, as conexões entre as variáveis não são sempre familiares ao pesquisador, pelo contrário, a descoberta do esqueleto e direcionamento dos arcos de uma Rede Bayesiana é um dos aspectos relevantes e investigados em toda sua teoria.

Os próximos itens definem a *d-separação* e as chamadas propriedades Markovianas, que fundamentam as características das RB, e também as tarefas inerentes ao seu aprendizado, as de estimação de estrutura - que determina quais variáveis possuem relação entre si - e de parâmetros - que indicam como se dá essa conexão.

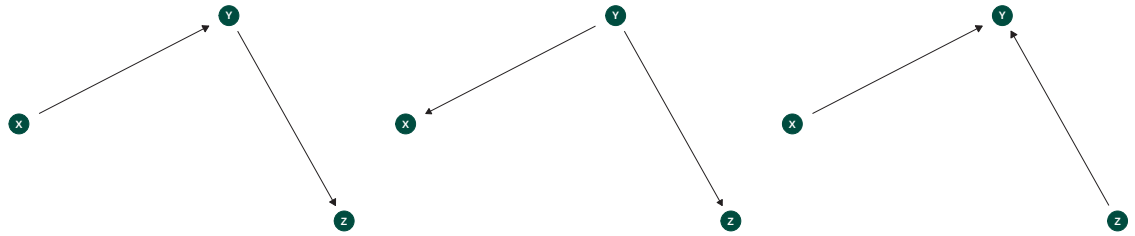
2.1.1 D-separação

De acordo com Pearl (1995), a decomposição recursiva do produto implica diretamente na indicação de independência condicional sugerida graficamente pela ausência de arcos direcionados. Ou seja, satisfazer a Condição de Markov, chamada também de Propriedade de Markov (NAGARAJAN; SCUTARI; LÈBRE, 2013), ou compatibilidade de Markov (PEARL, 2013), explicitamente, faz com que as Redes Bayesianas sejam *mapas de independência*, também chamados *I-maps* (KORB; NICHOLSON, 2010).

A relação, portanto, entre trios de variáveis auxiliam no processso de leitura das dependências, ou independências, condicionais e é interessante que a influência probabilística seja analisada como um fluxo no grafo (KOLLER; FRIEDMAN, 2009). Para ilustrar, sejam as variáveis X, Y e Z, bem como os três tipos de conexão entre elas na Figura 8.

Nas três figuras são apresentados o mesmo caminho $X \rightleftharpoons Y \rightleftharpoons Z$ contudo, o fluxo de informação contido em cada um é distinto, da seguinte forma:

- Na **Cadeia**, a relação da Figura 8a, é visto que X é pai de Y e Y é pai de Z então, X influencia Y, que por sua vez influencia em Z;
- No **Garfo**, apresentado na Figura 8b, Y é pai de X e Z e então Y influencia X e Z ao mesmo tempo.



(a) Cadeia. (b) Garfo (c) Garfo Invertido ou Colisor.

Figura 8 – Tipos de relações entre trios de variáveis.

Fonte – Elaborado pela autora, adaptado de *Bayesian Artificial Inteligence*, Korb e Nicholson (2010).

- O **Garfo Invertido ou Colisor**, na relação da Figura 8c o conjunto de variáveis pais de Y é composto por X e Z. Esse tipo de conexão expressa que X e Z influenciam Y ao mesmo tempo. Pode ser chamada também de **estrutura-v** e tem papel fundamental na interpretação probabilística das Redes Bayesianas (KORB; NICHOLSON, 2010; KOLLER; FRIEDMAN, 2009).

Para que as independências condicionais possam ser lidas, o critério gráfico de conectividade chamado de *d-separação*, uma "separação direcional" (PEARL, 2013), alicerça a interpretação da seguinte maneira:

D-separação: Sejam **X**, **Y** e **Z** três subconjuntos disjuntos de variáveis em um DAG $\mathbb{G}$. É dito que **X** e **Y** são *d-separados* dado **Z**, se, ao longo de todas as sequências de arestas - desconsiderando seus direcionamentos - entre um elemento de **X** e um elemento de **Y** existe um nó w satisfazendo uma das condições:

1. w é ponto de colisão, ou colisor; e se nem w ou seus descendentes estão em **Z**;

2. w não tem arestas convergentes, ou seja, está em cadeia ou garfo, e w está em **Z** (PEARL, 2013; NAGARAJAN; SCUTARI; LÈBRE, 2013).

Algumas regras que facilitam a leitura das d-conexões são descritas em Pearl (2000) são elas:

Primeira Regra. Dois conjuntos de variáveis **X** e **Y** são d-conectadas se existe um *caminho ativo* entre elas. Um caminho ativo é qualquer sequência de arcos, independentemente de suas direções, exceto por colisores.

Segunda Regra. Duas variáveis X e Y são d-conectadas dado um conjunto **Z** de vértices, se existir um caminho sem colisores entre X e Y que não passe por nenhum elemento de **Z**. Se o caminho não existir, X e Y são ditas d-separadas por Z, levando a conclusão que todo caminho entre X e Y é bloqueado por Z.

Terceira Regra. Se um colisor é um membro do conjunto condicionante, ou tem algum descendente em Z, então ele não bloqueia o caminho que passa por ele.

De maneira semelhante, $\mathbf{Z}$ é dito d-separar $\mathbf{X}$ de $\mathbf{Y}$, em $\mathbb{G}$, denotado por $(\mathbf{X} \perp\!\!\!\perp \mathbf{Y}|Z)_{\mathbb{G}}$, se, e somente se, $\mathbf{Z}$ bloqueia todos os caminhos de um nó em $\mathbf{X}$ para um nó em Y (PEARL, 1995). A d-separação pode ser exemplificada pela Figura 9, adaptado de Pearl (1988). O grafo direcionado e acíclico possui os conjuntos $\mathbf{X} = \{b\}$ e $\mathbf{Y} = \{c\}$ estes são d-separadas por $\mathbf{Z} = \{a\}$ uma vez que o caminho $b \leftarrow a \rightarrow c$ que conecta os conjuntos $\mathbf{X}$ e $\mathbf{Y}$ é bloqueado por $a \in \mathbf{Z}$. Da mesma forma que o caminho $b \rightarrow d \leftarrow c$ é bloqueado pois d e todos os seus descendentes não pertencem a $\mathbf{Z}$. Agora, $\mathbf{X}$ e $\mathbf{Y}$ não são d-separados pelo conjunto $\mathbf{T} = \{a, e\}$, pois no caminho $b \rightarrow d \leftarrow c$ o membro $e \in \mathbf{T}$ é descendente de d, neste caso, é dito que e ativa o caminho entre $\boldsymbol{X}$ e $\boldsymbol{Y}$ uma vez que conhecendo o valor de e, suas causas fazem com que seja dependente de $b \in \mathbf{X}$ e $c \in \mathbf{Y}$.

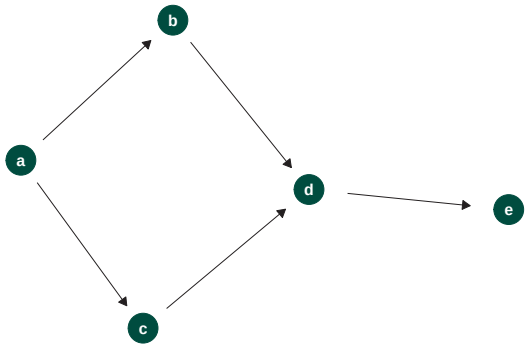


Figura 9 – DAG para exemplificar a d-separação.

Fonte – Elaborado pela autora, adaptado de *Probabilistic Reasoning in Intelligent Systems*, página 118 (PEARL, 1988).

Probabilisticamente, de acordo com o Teorema 1.2.4 em Pearl (2013), se os conjuntos $\mathbf{X}$ e $\mathbf{Y}$ são d-separados por $\mathbf{Z}$ em um DAG $\mathbb{G}$, então $\mathbf{X}$ é independente de $\mathbf{Y}$ condicionado a $\mathbf{Z}$, em todas as distribuições compatíveis com o $\mathbb{G}$. O contrário acontece se $\mathbf{X}$ e $\mathbf{Y}$ não são d-separados por $\mathbf{Z}$ em um DAG $\mathbb{G}$ logo, $\mathbf{X}$ e $\mathbf{Y}$ são condicionados a $\mathbf{Z}$, em, pelo menos, uma distribuição compatível com $\mathbb{G}$. Então, algumas relações como a de cadeia (Figura 8a) e garfo (Figura 8b) presumem a mesma condição de independência condicional (KOLLER; FRIEDMAN, 2009).

2.1.2 Cobertura de Markov

Em conformidade com a teoria de d-separação, de acordo com Neapolitan (2004), seja $\mathbf{X}$ um conjunto de variáveis aleatórias, tal que $X_i \in \mathbf{X}$ e P sua distribuição conjunta de probabilidade. A **Cobertura de Markov** de X_i, notada por $\mathbf{mb}(X_i)$, é formada por qualquer conjunto de variáveis que faz com que X_i seja condicionalmente independente de todas as outras variáveis dado $\mathbf{mb}(X_i)$, da seguinte maneira:

$$I_P(\{X_i\}, \mathbf{X} - (\mathbf{mb}(X_i) \cup \{X_i\} | \mathbf{mb}(X_i)))$$

A cobertura de Markov se utiliza da d-separação para ser identificada (PEARL, 1988), sua principal propriedade é a de bloquear o efeito das variáveis que não fazem parte desse conjunto e é constituída, portanto, pelos pais de X_i, os filhos de X_i e as variáveis que compartilham, pelo menos um filho com X_i (NIELSEN; JENSEN, 2009).

Nesses critérios, um exemplo de cobertura de Markov está apresentado na Figura 10. Para dado conjunto de dados $\mathbf{X} = \{X_1, X_3, X_4, X_5\}$, e analisando a variável X_2, sua cobertura de Markov é composta por seus pais $\{X_1, X_5\}$, seu filho $\{X_3\}$ e das que compartilham pelo menos um filho $\{X_6\}$, logo, $\mathbf{mb}(X_2) = \{X_1, X_3, X_5, X_6\}$. A variável X_4 não faz parte da cobertura pois não atende aos critérios, sendo ela pai de um pai de X_2.

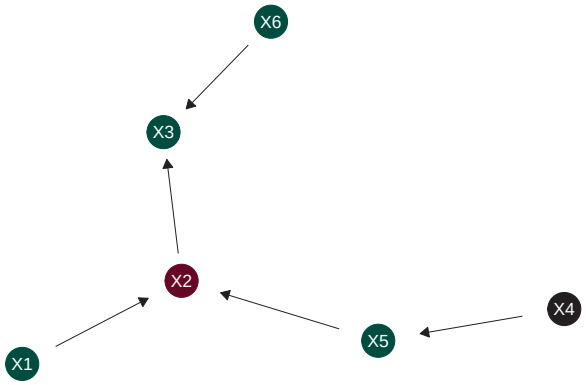


Figura 10 – Cobertura de Markov da variável X_2.

Fonte – Elaborado pela autora.

Mais especificamente, quando a cobertura de Markov de X_i, $\mathbf{mb}(X_i)$, é o único conjunto possível que d-separa X_i das variáveis que não o compõem, ou seja, quando não é possível que se remova nenhum arco contido na estrutura, ela é a cobertura de Markov *minimal* e é chamada de *Fronteira de Markov* (NEAPOLITAN, 2004). Ela pode ser definida também como um conjunto finito de coberturas de Markov (CHENG; GREINER, 1999).

Por sua definição, Bielza e Larrañaga (2014) descrevem a Cobertura de Markov de uma variável X_i como sendo o conjunto mínimo necessário para a sua predição, pode servir então, como metodologia de seleção de variáveis (FRENO, 2007). Com isso, muitos algoritmos de estimação de estrutura do grafo são baseados em procedimentos que tem por objetivo inicial o descobrimento da cobertura de Markov, ou da fronteira de Markov, como é o caso do *Incremental Association Markov Blanket - IAMB* (TSAMARDINOS *et al.*, 2003) e *Grow Shrink* que serão apresentados posteriormente.

2.1.3 Equivalência

O esqueleto de um DAG $\mathbb{G}$ é um grafo não direcionado obtido pela remoção das suas direções (KOLLER; FRIEDMAN, 2009). Dois DAGs $\mathbb{G}_1$ e $\mathbb{G}_2$ representam as mesmas estruturas

de d-separação, ou seja, as duas estruturas são equivalentes se, e somente se, ambas têm o mesmo esqueleto e codifiquem as mesmas restrições de independência condicional estabelecidas pelo conjunto de colisores, as estruturas-v (VERMA; PEARL, 1991; SCUTARI; GRAAFLAND; GUTIÉRREZ, 2019).

Logo, se duas estruturas são equivalentes, as Redes Bayesianas que elas compõem fazem parte de uma mesma *classe de equivalência* de modelos. Estatisticamente, se relacionam por terem verossimilhanças idênticas (KORB; NICHOLSON, 2010) e assim, suas distribuições de probabilidade podem ser decompostas de maneira equivalente (SCUTARI; GRAAFLAND; GUTIÉRREZ, 2019).

Essas classes de equivalências são representadas visualmente por meio do não direcionamento de seus arcos, nas estruturas chamadas *PDAGs* (do inglês, **P**artially-**D**irected **A**cylic Graph), o que significa eles serem grafos acíclios porém parcialmente direcionados (ACID; CAMPOS; CASTELLANO, 2005). Esses grafos, por não apresentarem completa orientação, não podem compor as Redes Bayesianas, o direcionamento dos arcos deve ser encontrado de alguma maneira.

Com o embasamento teórico a respeito dos fundamentos da metodologia de Redes Bayesianas, obtido com os tópicos anteriores, é importante salientar também uma das utilizações desse tipo de modelo na literatura, com a abordagem de interpretação causal (PEARL, 2013; SPRITES; GLYMOUR; SCHEINES, 2000; HITCHCOCK, 1997). Conforme os autores, as redes possuem definições que permitem, sob certas condições, que sejam analisadas de maneira causal. A teoria aprofundada a respeito de causalidade é bastante rica e complexa, podendo ser encontrada nos recursos mencionados, além de outros que fazem a ligação dos conceitos das Redes Bayesianas, probabilidade e de inferência causal, como em Zhang e Spirtes (2011), Korb e Nicholson (2010) e Neapolitan (2004). A próxima seção introduz brevemente esse contexto e como as condições se relacionam com a estrutura dos modelos que utilizam esse conceito.

2.2 Causalidade

A descoberta de relações causais a partir de dados tem intrigado filósofos desde o século XVIII (PEARL, 2000). O tema é de grande relevância na ciência de uma forma geral, mas gera discussões entre diferentes autores, como tratado em Druzdzel e Simon (1993) e Salmon (1998).

No mesmo sentido, em Pearl *et al.* (2009), os autores dissertam a respeito de causalidade na estatística e os campos dessa ciência os quais estudam tal fenômeno, que são: equações estruturais não-paramétricas, análise contrafactual, modelos gráficos e a junção entre os dois últimos. Muitas metodologias foram, e têm sido, propostas para direcionar o entendimento de raciocínios causais. E um desses métodos é o de Redes Bayesianas (ZHANG; POOLE, 1996).

De maneira formal, as definições de Redes Bayesianas apresentadas até agora se referem

especificamente a aspectos probabilísticos, como as propriedades relacionadas a dependência e independência condicional (KOLLER; FRIEDMAN, 2009). Aliado a isso, as características de classes de equivalência atuam no sentido de que, para algumas estruturas estimadas, existe uma indistinguibilidade probabilística o que pode impedir a interpretação de causa-efeito das direções dos arcos obtidos por meio de metodologias de aprendizado (NAGARAJAN; SCUTARI; LÈBRE, 2013), ou podem ser utilizadas como base para experimentos que visam identificar as direções do grafo (MEGANCK; LERAY; MANDERICK, 2006). Por outro lado, existem as estruturas baseadas em conhecimento préviamente obtido que estabelecem os arcos conforme a plausibilidade teórica da relação entre variáveis, feita de maneira determinística.

Em uma escala específica, as relações entre trios de variáveis, apresentadas na Figura 8, podem representar cenários de interações causais entre variáveis de uma estrutura de Rede Bayesiana (KORB; NICHOLSON, 2010);

- A Figura 8a, que pode ser chamada também de **Cadeia Causal**, pode ser interpretada como X sendo a causa Y e Y a causa Z;
- O garfo da Figura 8b, é chamado de **Causa Comum**, e tem que Y causa ambas X e Z.
- E o garfo invertido, consolidor ou estrutura-v, na Figura 8c, é o **Efeito Comum**, no qual existe um conjunto de casuas de Y que é composto por X e por Z.

De maneira abrangente, para interpretação de causalidade por meio de um modelo causal, Spirtes *et al.* (2000) propõem três condições que relacionam probabilidade com causalidade e elas são o que os autores denominam de:

1. A **Condição Causal de Markov**, que é semelhante a Condição de Markov discutida anteriormente, a qual determina que, cada variável X_i, que compõe uma Rede Bayesiana $(\mathbb{G}, P)$, é independente de todos os seus não-descendentes dado o conjunto de pais $\mathbf{pa}_{\mathbb{G}}(X_i)$ (PEARL, 2013), conforme expresso probabilisticamente na Equação (2.1), mas com uma interpretação causal das direções dos arcos entre variáveis (KOLLER; FRIEDMAN, 2009). Quando um modelo de Rede Bayesiana satisfaz essa condição causal de markov, é dito ser um mapa de independências (*I-map*) (KORB; NICHOLSON, 2008).

2. A **Condição Causal de Minimalidade** propõe que a cada conexão causal direta prevê alguma relação de independência ou independência condicional que poderia ter sido obtida, motivada pela análise da distribuição contrafactual da distribuição $P_{\mathbb{G}}$ (ZHANG; SPIRTES, 2011). Formalmente, é tido que uma estrutura causal $\mathbb{G}$ satisfaz essa condição se, e somente se, todo subgrafo próprio [1] da estrutura não satisfizer a condição anterior de Causalidade de Markov relacionada a distribuição de probabilidade P (SPIRTES *et al.*, 2000). Ou seja,

[1] Um subgrafo próprio é um subgrafo obtido retirando um arco ou aresta da estrutura do grafo principal.

dado que a condição de Markov é satisfeita, a estrutura causal é verdadeira somente se $\mathbb{G}$ for a estrutura minimal de P, conforme Zhang e Spirtes (2011);

3. A **Condição de Fidedignidade** os componentes de um modelo $(\mathbb{G}, P)$ são fiéis um ao outro se toda e qualquer relação de independência de P são aquelas implicadas pela d-separação em uma estrutura gráfica $\mathbb{G}$ (KOLLER; FRIEDMAN, 2009; SPIRTES *et al.*, 2000). E quando um modelo de satisfaz essa condição de fidedignidade, é dito ser um mapa de dependencias (*D-map*) (KORB; NICHOLSON, 2008).

Segundo Koller e Friedman (2009), quando as condições 1 e 3 são satisfeitas, é dito que a estrutura $\mathbb{G}$ é um *mapa perfeito* da distribuição de probabilidade P. Conforme Korb e Nicholson (2008), existe interesse em encontrar um *I-map* que seja minimal, ou seja, se algum arco for retirado, o modelo não é mais um mapa de independência, mas essa não é uma característica necessária para que o modelo componha o mapa perfeito. Portanto, de acordo com Spirtes *et al.* (2000) se um modelo $(\mathbb{G}, P)$ for um mapa perfeito minimal, seus arcos podem ter uma interpretação causal.

A literatura acima citada discute amplamente a utilização das relações entre análise probabilística e análise causal, e ainda, discorre a respeito da relevância dos aspectos que cercam essa discussão para o desenvolvimento das técnicas e heurísticas de estimação em Redes Bayesianas. Uma vez que, por definição, as Redes Bayesianas satisfazem, ao menos, as condições 1 e 3 conforme explanação na Seção 2.1.

Apesar de todas as questões que envolvem as Redes Bayesianas e interpretação causal, esta dissertação foca essencialmente em descrever metodologias que determinam as características estruturais e são aprendidas na fase de *Estimação da Estrutura* e posteriormente, com a *Estimação dos Parâmetros*, ou seja, ao estabelecimento de como as variáveis se conectam (SCUTARI, 2016).

2.3 Estimação de Redes Bayesianas

Por definição, uma Rede Bayesiana é caracterizada pela dupla gráfico-probabilística $(\mathbb{G}, P)$, e portanto, sua tarefa completa de estimação, realizada com respeito a um conjunto de dados $\mathbb{D}$, é composta de duas etapas conforme a Equação (2.3) (SCUTARI; GRAAFLAND; GUTIÉRREZ, 2019):

$$P(\mathbb{G}, \Theta|\mathbb{D}) = P(\mathbb{G}|\mathbb{D}) \times P(\Theta|\mathbb{G}, \mathbb{D}), \tag{2.3}$$

sendo que a primeira etapa, referente a $P(\mathbb{G}|\mathbb{D})$, é a de estimação de estrutura, a qual se embasa no conjunto de dados composto de d variáveis e n observações. E a segunda é referente a $P(\Theta|\mathbb{G}, \mathbb{D})$, a estimação dos parâmetros (Θ) que é condicionada a estrutura gráfica e ao conjunto de dados.

Um dos maiores desafios das Redes Bayesianas é representar a topologia da interação entre as variáveis que organizem o conhecimento probabilístico e permita seu cálculo de maneira coerente (PEARL, 2000), uma vez que o esqueleto de interação entre as variáveis de interesse é, em grande parte, desconhecido. Essa tarefa de estimar a estrutura de uma Rede Bayesiana é não trivial e complexa devido ao imenso número de possíveis soluções em $\mathscr{G}$, o espaço de grafos do domínio do problema. Tal esforço é conhecido por ser *NP-complete* e *NP-hard*, tais provas podem ser encontradas em Chickering (1996) e Chickering, Heckerman e Meek (2004), respectivamente. Contudo, o aspecto de complexidade da estimação de estrutura não será foco desta dissertação.

A seguir, serão detalhadas as heurísticas utilizadas para essa tarefa de aprendizado e as principais abordagens para essa investigação podem ser divididas em três classes: as duas principais são as de Aprendizado Restrito (*Constraint Learning*, que pode ser chamado também de *rule-based*) e a de Aprendizado Baseado em Métricas (*score-based*, ou *score-and-search*); a terceira classe, e mais recente delas, é a de Aprendizado Híbrido que associa elementos de ambos grupos anteriores.

2.3.1 Estimação Restrita de Estrutura

A principal finalidade das heurísticas dessa classe é construir uma estrutura que expresse as relações de dependência e independência correspondentes à distribuição dos dados (CAMPOS, 2006). É denominada *rule-based* ou *constraint-based* uma vez que seus arcos são resultados de testes de independência entre os pares de variáveis.

Considerando o caso discreto, foco dessa dissertação, as comparações realizadas se utilizam da hipótese nula de independência condicional, por meio da distribuição assintótica χ^2 (NAGARAJAN; SCUTARI; LÈBRE, 2013). As estatísticas utilizadas nos testes de independência condicional devem ser definidas previamente, algumas das opções estão descritas a seguir: *Informação Mútua Condicional*, o clássico χ^2 *de Pearson* e o mais recente, *Estimador Encolhido da Informação Mútua Condicional* (SCUTARI, 2020).

Todas as estatísticas são baseadas nas tabelas de probabilidade condicional (TPC) obtidas por meio da frequência observada no conjunto de dados, a base que contém n observações e d variáveis, para as variáveis X, Y condicionada às possíveis configurações $\mathbf{Z}$, que são as combinações entre os estados das componentes desse conjunto. As TPCs serão denotadas por $P(X,Y|\mathbf{Z})$, as frequências são denotadas por n_{ijk} tal que número de linhas da tabela $i = \{1,\ldots,R\}$, sendo R o número de configurações i e j possíveis das categorias de X e Y e o número de colunas a tabela $k = \{1,\ldots,q\}$, sendo q o número de configurações possíveis dos pais.

Informação Mútua

A Informação Mútua é a medida, advinda da teoria da informação, que dimensiona a dependência entre duas variáveis aleatórias sendo uma medida mais geral que a correlação linear (PETHEL; HAHS, 2014), pois mensura a quantidade de informação compartilhada por duas variáveis (BERRETT; SAMWORTH, 2017).

Para duas variáveis aleatórias X e Y a informação mútua é definida como:

$$IM(X,Y) = \sum_{X,Y} P(X,Y) log \frac{P(X,Y)}{P(X)P(Y)},$$

sendo que $P(X) > 0$ e $P(Y) > 0$. Essa medida é a divergência de Kullback-Leibler entre $P(X,Y)$ e $P(X)P(Y)$ então, pode assumir apenas valores não negativos (BERRETT; SAMWORTH, 2017). Ela pode ser interpretada como a redução na incerteza da variável. Essa redução é simétrica logo, será nula se, e somente se, X e Y forem independentes (COVER; THOMAS, 2012).

A extensão dessa medida, para o caso condicionado a um conjunto $\mathbf{Z}$, está apresentada na Equação (2.4) (CAMPOS, 2006; NAGARAJAN; SCUTARI; LÈBRE, 2013), métrica denominada como **Informação Mútua Condicional**.

$$\begin{aligned} I(X,Y|\mathbf{Z}) &= \sum_{Z} \left(P(Z) \sum_{Y} \sum_{X} P(X,Y|Z) log \left(\frac{P(X,Y|Z)}{P(X|Z)P(Y|Z)} \right) \right) \\ &= \sum_{i=1}^{n} \sum_{j=1}^{d} \sum_{k=1}^{q} \frac{n_{ijk}}{n} log \frac{n_{ijk} n_{**k}}{n_{i*k} n_{*jk}}, \end{aligned} \tag{2.4}$$

sendo que quando os índices das variáveis estão omitidos com $*$, significa que são considerados todos os seus valores, como em n_{**k}, são considerados todos os valores assumidos por X e Y; em n_{i*k} todos os valores assumidos por Y e em n_{*jk} todos os valores assumidos por X.

χ^2 de Pearson

O clássico χ^2 de Pearson para tabelas de contingência, altamente reconhecido na literatura, sendo utilizado para testar associações. É dado por:

$$\chi^2(X,Y|\mathbf{Z}) = \sum_{l=1}^{L} \frac{(O_l - E_l)^2}{E_l} = \sum_{i=1}^{n} \sum_{j=1}^{d} \sum_{k=1}^{q} \frac{(n_{ijk} - m_{ijk})^2}{m_{ijk}},$$

sendo $m_{ijk} = \frac{n_{i*k} n_{*jk}}{n_{**k}}$, O_l as quantidade de observações da célula l, E_l o valor esperado para l no caso de independência, L número de células da tabela (NAGARAJAN; SCUTARI; LÈBRE, 2013).

As alternativas de estatísticas descritas acima são utilizadas nos testes realizados iterativamente conforme os passos estabelecidos nas heurísticas adotadas, cada uma delas possui

caracterizações específicas que serão detalhadas a seguir. Os algoritmos selecionados para representarem essa classe de aprendizado são: o Algoritmo PC (COLOMBO; MAATHUIS, 2014), em sua versão modernizada, o *Incremental Association Markov Blanket* para Redes Bayesianas (IAMB) (TSAMARDINOS *et al.*, 2003) e o *Grow Shrink* (GS) (MARGARITIS, 2003).

Algoritmo *PC*

O algoritmo PC, o mais difundido entre os componentes dessa classe, é baseado no algoritmo IC (*Inductive Causation*) proposto em Pearl e Verma (1992). Ele é iniciado com um grafo completamente conectado e a cada par de variáveis, realiza um teste de independência condicional que remove os arcos não válidos, ou seja, os arcos que não rejeitam a hipótese nula de independência condicional, conforme visto anteriormente. A versão otimizada dessa metodologia de estimação será utilizada, chamada de *PC-stable* e foi proposta em Colombo e Maathuis (2014).

Em linhas gerais, seu objetivo inicial é encontrar a relação não direcionada entre variáveis, ou seja, o esqueleto do grafo e, para isso, testa conjuntos de variáveis adjacentes com cardinalidade que varia entre 0 e $d-2$, de modo a encontrar arestas que devem ser removidas do grafo completo.

Partindo do esqueleto, o grafo é estendido para a versão parcialmente direcionada, ou *CPDAG*, que representa a classe de equivalência do modelo, essas direções são induzidas, por meio de identificação de estruturas-v e posteriormente aplicando as regras (CHICKERING, 2013) descritas no Algoritmo 1 que apresenta os passos gerais da heurística. Aqui as versões tradicional e otimizada se diferem, a nova proposta faz com que, no momento da estimação do esqueleto, os conjuntos de variáveis adjacentes de mesma cardinalidade não sejam afetados pela remoção de arcos desse passo. Os autores provam ainda, que essa evolução adiciona uma propriedade de independência na ordenação das variáveis.

Essa implementação modernizada será considerada na condução dos resultados dessa dissertação e está descrita abaixo conforme apresentado em Scutari, Graafland e Gutiérrez (2019). Considerando um conjunto de dados observados D composto por n registros e de $\mathbf{X} = \{X_1, \ldots, X_d\}$, além disso, um teste de independência condicional (TCI) deve ser informado para a execução do algoritmo como segue no Algoritmo 1.

A partir do passo 2, o esqueleto do grafo é encontrado; no passo 8, começam as inserções de estruturas-v ao grafo e, posteriormente, no passo 9 são aplicadas as regras que evitam ciclos (passo 10) e evitam novos colisores (passo 11).

Além disso, o retorno da estrutura G desse algoritmo não garante que um DAG seja formado, pelo contrário, esse método é utilizado para que sejam encontrados os *PDAGs* que representam as classes de equivalência dos modelos gerados para estruturas com o esqueleto e estruturas-v de G, conforme item 2.1.3.

Algoritmo 1 – PC-*stable*

ENTRADA($\mathbb{D}$, TCI) ▷ Base de dados, Teste

1: inicie com um grafo G completo
2: **para** m em $(0,\ldots,d-2)$ **faça**
3: **repita** para todos $(X_i,X_j) \in \mathbf{X}, i \neq j$, tal que X_i possui, pelo menos, m vizinhos na estrutura G atual, excluindo X_j:
4: **escolha** um novo subconjunto $\mathbf{Z}$ de dimensão m dos vizinhos de X_i, exceto X_j;
5: **se** $(X_i \perp X_j | \mathbf{Z})$ **então**
6: **remova** $X_i - X_j$ de G e faça $\mathbf{Z}_{X_i,X_j} = \mathbf{Z}$, como um conjunto que separa X_i de X_j;
7: **se** X_i e X_j não são mais adjacentes, ou não existem mais opções para $\mathbf{Z}$ **então** volte ao passo 3
8: **substitua** para $X_i \rightarrow X_j \leftarrow X_k$, para cada trio $X_i - X_j - X_k$, tal que X_i e X_k não são adjacentes e $X_j \notin \mathbf{Z}_{X_i,X_j}$.
9: **encontre** outras direções segundo as regras:
10: **se** $X_i - X_j$ e existe um caminho direcionado de X_i par X_j **então substitua** por $X_i \rightarrow X_j$;
11: **se** X_i e X_j não adjacentes, mas $X_i \rightarrow X_k$ e $X_k - X_j$ **então substitua** por $X_k \rightarrow X_j$.

SAÍDA (G)

Algoritmo *Grow Shrink*

O *Grow Shrink* (GS) é baseado no algoritmo mais simples de detecção da Cobertura de Markov, chamado *Grow Shrink Markov Blanket - GSMB* (NAGARAJAN; SCUTARI; LÈBRE, 2013) proposto em Margaritis e Thrun (2000). Para encontrar a cobertura de Markov das variáveis, essa metodologia se utiliza da noção de Fronteira de Markov, conforme mencionado na Seção 2.1.2, ela é definida como sendo a Cobertura de Markov *minimal*, ou seja, nenhum dos subconjuntos próprios possíveis na estrutura forma outro conjunto que pode ser definido como sendo uma cobertura de Markov (LIU; LIU, 2018).

O *GSMB* começa com um conjunto vazio, em sua fase de crescimento (*grow*), buscando por variáveis que fazem parte da fronteira de Markov de um X_i testando a dependência de cada par de variáveis condicionado ao conjunto que se inicia vazio e armazena essa lista para a próxima fase. O encolhimento acontece na fase *Shrink*, quando o algoritmo identifica e remove as variáveis que foram adicionadas e que não fazem parte dessa fronteira. A ideia é investigar a violação da propriedade de Markov verificando se existe alguma variável não descendente de X_i que é dependente de X_i dado o conjunto hipotético de fronteira (MARGARITIS, 2003).

Partindo desse contexto, o algoritmo *Grow Shrink* para a descoberta de Redes Bayesianas foi proposto em Margaritis e Thrun (2000). O *GS* utiliza do conhecimento absorvido pelo aprendizado da cobertura de Markov de cada variável para compor a estrutura de Rede Bayesiana. Esse algoritmo se baseia no conceito de vizinhança pois, depois que a cobertura de Markov de cada variável foi encontrada, o procedimento reconhece cada vizinhança local $\mathbf{V}(X_i)$ - constituída de conexão direta - para resgatar o esqueleto exato em torno de cada variável. E por último, os arcos são direcionados conforme análise de v-estruturas (SU *et al.*, 2012).

Margaritis (2003) descreve os procedimentos conforme o Algoritmo 2, sendo que D é definido como o conjunto de dados para estimação e *TIC* é o teste de independência condicional previamente selecionado.

Algoritmo 2 – *Grow Shrink*

ENTRADA($\mathbb{D}$,*TIC*)
para $X_i \in \mathbf{X}$ **faça**
  encontre $\mathbf{mb}(X_i)$
Fase *Grow*
  para $X_i \in \mathbf{X}$ e $X_j \in \mathbf{mb}(X_i)$ **faça**
    se X_i e X_j são dependentes dado $\mathbf{Z}$ para todo $\mathbf{Z} \subseteq \mathbf{H}$, onde $\mathbf{H}$ é o menor entre $\mathbf{mb}(X_i) - \{X_j\}$ e $\mathbf{mb}(X_j) - \{X_i\}$ **então**
      determine $X_j - X_i$;
    para $X_i \in \mathbf{X}$ e $X_j \in \mathbf{V}(X_i)$ **faça**
      se existir uma variável $X_k \in \mathbf{V}(X_i) - \mathbf{V}(X_j) - \{X_j\}$ tal que X_j e X_k são dependentes dado $\mathbf{Z}\{X_i\}$, para todo $\mathbf{Z} \subseteq \mathbf{H}$, onde $\mathbf{H}$ é o menor entre $\mathbf{mb}(X_j) - \{X_i, T\}$ e $\mathbf{mb}(T) - \{X_i, X_j\}$ **então**
        oriente $X_j \rightarrow X_i$;
Fase *Shrink*
      enquanto existirem ciclos **faça**
        calcule o conjunto $\mathbf{E} = \{X_i \rightarrow X_j$ tal que $X_i \rightarrow X_j$ é parte de um ciclo$\}$
        remova da estrutura G atual, o arco em $\mathbf{E}$ que é parte do maior número de ciclos e adicione em $\mathbf{R}$;
        Reinsira os arcos de $\mathbf{R}$ na direção reversa e na ordem contrária a remoção;
        enquanto não se aplicar **faça**
          para $X_i \in \mathbf{X}$ e $X_j \in \mathbf{V}(X_i)$ tal que nem $X_j \rightarrow X_i$, nem $X_i \rightarrow X_j$ **faça**
            se existe um caminho direto de X_i para X_j **então**
              oriente $X_i \rightarrow X_j$;
SAÍDA (G) =0

Algoritmo *IAMB*

O *Incremental Association Markov Blanket* (IAMB) para Redes Bayesianas foi mencionado anteriormente no item 2.1.2 que trata da Cobertura de Markov, uma vez que se utiliza dessa definição para restringir o subconjunto de variáveis com as quais serão realizados os testes de independência condicional (BERETTA *et al.*, 2018).

O procedimento homônimo utilizado para descobrir a cobertura de Markov de uma variável foi proposto em Tsamardinos *et al.* (2003). Ele busca um conjunto de possíveis variáveis candidatas a comporem a Cobertura de Markov em duas fases, a primeira inclui as que possuem maior valor de medida de associação, que é uma função objetivo f. Na segunda fase as variáveis que não condizem com os critérios de independência condicional são descartadas.

Esse procedimento completo está detalhado no Algoritmo 3, conforme apresentado em Zhang *et al.* (2010), que requer, além da base de dados (D) e do teste *TIC*, uma função objetivo f, que deve ser maximizada na primeira fase do algoritmo.

Algoritmo 3 – IAMB

ENTRADA$(\mathbb{D}, f, TIC)$
1: Para cada $X_i \in \mathbf{X}$
Fase *Forward*
2: $cmb(X_i) = \varnothing$
3: **repita** até $mb(X_i)$ não se alterar:
4: encontre X_{max} em $\mathbf{X} - mb(X_i) - \{X_i\}$ que maximiza $f(X_{max}, X_i | cmb(X_i)))$
5: **se** não $I(X_{max}, X_i | cmb(X_i))$ **então**
6: $cmb(X_i) = cmb(X_i) \cup \{X_{max}\}$
Fase *Backward*
7: Para cada $V_j \in mb(X_i)$
8: **se** $I(V_j, X_i | cmb(X_i) - \{V_j\})$ **então**
9: $cmb(X_i) = cmb(X_i) - \{V_j\}$
SAÍDA ($\mathbf{mb}(X_i)$) ▷ para todo $i = 1, \ldots, d$

Mais detalhadamente, na primeira fase, chamada de *forward*, as variáveis candidatas a fazerem parte da Cobertura de Markov de X_i são incluídas no conjunto aspirante a cobertura de Markov $cmb(X_i)$. É possível que nessa fase entrem mais variáveis do que é necessário para o conjunto $\mathbf{mb}(X_i)$, apesar de serem admitidas somente aquelas que maximizam uma função da heurística, a qual deve refletir a correspondência entre variáveis e seu resultado não deve ser zero para toda variável de $\mathbf{mb}(X_i)$.

É importante que a função f a ser maximizada seja informativa, para que, o conjunto de variáveis, $cmb(X_i)$, que passa para a fase *backward* possa ser o menor possível, com o objetivo de não desperdiçar tempo com variáveis irrelevantes e não requerer um tamanho de amostra maior do que o necessário para aplicação dos testes. Nessa segunda fase os falsos positivos são identificados um a um, por meio dos testes de independência condicional, e removidos. Ao final desse processo obtém-se $cmb(X_i) = \mathbf{mb}(X_i)$ (TSAMARDINOS *et al.*, 2003).

Além dessa versão, existem outras variações desse algoritmo, que foram propostos com o intuito de otimizar seus procedimentos, como o *InterIAMB*, o *InterIAMBnPC* (TSAMARDINOS *et al.*, 2003), o *fastIAMB* (YARAMAKALA; MARGARITIS, 2005), entre outros listados em Bielza e Larrañaga (2014), Scutari (2020).

2.3.2 Estimação de Estrutura Baseado em Métricas

Os procedimentos de estimação de estrutura dessa classe, chamados de *score based* ou também de *score and search*, se utilizam de métricas, por meio da maximização de função objetivo (função *score*) ou de minimização de uma função de custo, e estão aliados a uma heurística de busca com a intenção de mensurar todas as estruturas exploradas dentro das

possíveis soluções (ACID; CAMPOS, 2003). Os algoritmos dessa classe ainda podem ser divididos em três grupos menores de *greedy search*, *simulated annealing* e algoritmos genéticos, em Russell e Norvig (2010) os autores apresentam a definição de cada um destes grupos detalhadamente.

Nesta dissertação, os estudos estão voltados aos algoritmos *greedy search* (sua tradução literal é "*busca gulosa*") pois são os mais amplamente difundidos nessa classe de estimação de estrutura para Redes Bayesianas (SCUTARI; GRAAFLAND; GUTIÉRREZ, 2019), e como representantes desse grupo serão apresentadas as heurísticas *Hill-Climbing* e *Tabu Search*, que são baseadas na ideia da busca pela *estrutura ótima*, $\mathbb{G}^*$, por meio da maximização de uma função objetivo f, conforme expresso pela Equação (2.6).

Os métodos que avaliam uma métrica recursivamente, em geral, supõem que tal medida represente uma pontuação de ajuste da estrutura aos dados observados e que essa função f possa ser *decomposta* de forma que, para uma estrutura $\mathbb{G}$, deva ser calculada pela soma da função aplicada em cada uma das variáveis condicionada a seus respectivos pais (LIU; MALONE; YUAN, 2012), de acordo com a Equação (2.5):

$$f(\mathbb{G}|D) = \sum_{i=1}^{d} f(X_i|\mathbf{pa}(X_i), D). \tag{2.5}$$

Essa expressão está balizada pela decomposição da distribuição de probabilidade condicional assegurada pela condição de Markov nas Redes Bayesianas, apresentada na Equação (2.1) do item 2.1.

Dessa forma, o problema de otimização se restringe a busca da melhor estrutura $\mathbb{G}^*$, pelo argumento máximo da função objetivo f com respeito a estrutura $\mathbb{G}$ e os dados observados D, conforme é apresentado:

$$\mathbb{G}^* = \arg\max_{\mathbb{G}\in\mathscr{G}} f(\mathbb{G}, D). \tag{2.6}$$

sendo que D é composto pelo conjunto de variáveis $\mathbf{X}$ com n observações (CAMPOS, 2006).

Visto isso, a função ou métrica a ser utilizada para atacar esse problema de otimização deve ser escolhida de forma que não apenas reflita o ajuste da estrutura aos dados, mas que também evite o *overfitting* (super-ajuste) desses dados. Então, o *score*, em geral, aproxima a probabilidade da estrutura condicionada aos dados, equilibrando entre um bom ajuste e a complexidade regulada da rede (LIU; MALONE; YUAN, 2012).

Desse modo, a função f deve assumir valores informativos que discriminam estruturas diferentes. Contudo, para os DAGs que são representados pelos mesmos grafos essenciais[2] ou

[2] os grafos essenciais são os grafos que representam as classes de equivalência, possuem mesmo esqueleto e mesmas estruturas-v (SCUTARI, 2018).

seja, para modelos de mesma classe de equivalência, a função objetivo pode assumir mesmo valor, quando isso acontece ela é chamada de *score equivalente* (CAMPOS, 2006).

Existem inúmeras maneiras de mensurar o equilíbrio ajuste-complexidade (YANG; CHANG, 1996; LIU; MALONE; YUAN, 2012) e, de acordo com Campos (2006), podem ser divididas entre medidas Bayesianas, chamadas também de verossimilhanças marginais Dirichlet-Bayesianas (SCUTARI, 2018), e medidas de informação que penalizam uma função de ajuste dos dados. Algumas foram selecionadas e estão descritas a seguir.

Por se tratarem de redes discretas, a frequência das células de cada uma das d tabelas de probabilidade condicional serão utilizadas para cálculo de modo que: $i = \{1,2,\ldots,d\}$ sendo d o número de variáveis em $\mathbf{X}$; $j = \{1,2,\ldots,q_i\}$, sendo que q_i é a quantidade de combinações de pais em $\mathbf{pa}(X_i)$ para cada X_i; $k = \{1,2,\ldots,c_i\}$ c_i o número de classes da variável X_i, por fim, N_{ijk} é o número de observações na base de dados na qual a variável X_i recebe o valor x_{ik} e $N_{ij} = \sum_{k=1}^{c_i} N_{ijk}$ (NAGARAJAN; SCUTARI; LÈBRE, 2013).

A métrica *K2*

Uma das primeiras métricas é proposta em Cooper e Herskovits (1992), na qual, suas suposições permeiam os dados serem discretos, a ausência de valores *missing* e de observações independentes.

A métrica *K2* é inspirada na densidade posteriori apresentada na Equação (2.7) conforme seu manuscrito de proposta,

$$P(D,G) = P(G)\prod_{i=1}^{d}\prod_{j=1}^{q_i}\frac{(c_i-1)!}{(N_{ij}+c_i-1)!}\prod_{k=1}^{c_i}N_{ijk}!, \tag{2.7}$$

e seu cálculo é com respeito à estrutura do grafo G, com probabilidade *a priori* dada por $P(G)$. Pela sua propriedade de decomposição, pode ser expressa conforme estabelecido na Equação (2.8).

$$f_{K2}(D,G) = log(P(G)) + \sum_{i=1}^{d}\left(\sum_{j=1}^{q_i}\left(log\left(\frac{(c_i-1)!}{(N_{ij}+c_i-1)!}\right) + \sum_{k=1}^{c_i} log\left(N_{ijk}!\right)\right)\right). \tag{2.8}$$

Essa métrica não possui a característica de atribuir o mesmo valor para distribuições na mesma classe de equivalência, (CAMPOS, 2006).

A métrica *score* equivalente *Dirichlet* Bayesiana com priori uniforme

Como uma generalização da métrica *K2*, Heckerman, Geiger e Chickering (1995) propuseram a métrica *Dirichlet Bayesiana*, f_{DB}, a qual incorpora hiperparâmetros, η_{ijk}. de uma

distribuição Dirichlet *a priori* dos parâmetros dada estrutura da rede G, conforme a expressão abaixo:

$$f_{DB}(D,G) = \\ = log(P(G)) + \sum_{i=1}^{d}\left(\sum_{j=1}^{q_i}\left(log\left(\frac{\Gamma(\eta_{ij})}{\Gamma(N_{ij}+\eta_{ij})}\right) + \sum_{k=1}^{c_i} log\left(\frac{\Gamma(N_{ijk}+\eta_{ijk})}{\Gamma(\eta_{ijk})}\right)\right)\right),$$

sendo $\Gamma(.)$ a função *Gamma*, a qual $\Gamma(r) = (r-1)!$ e $\eta_{ij} = \sum_{k=1}^{c_i}\eta_{ijk}$, quando $\eta_{ijk} = 1$, retorna-se à métrica *K2* (CAMPOS, 2006).

Para facilitar a utilização desses hiperparâmetros, Heckerman, Geiger e Chickering (1995) inserem a propriedade de equivalência da verossimilhança, construindo uma alternativa *score* equivalente à função f_{DB}; de modo que cada um dos hiperparâmetros, η_{ijk}, possa ser escrito como $\eta \times p(x_{ik}, w_{ij}|G)$, sendo que η é o parâmetro de tamanho de amostra equivalente e $p(.|G)$ é a probabilidade da distribuição com a estrutura G (CAMPOS, 2006).

Além disso, quando uma probabilidade uniforme é atribuida para cada configuração de $X_i|\mathbf{pa}(X_i)$, nasce a métrica *score equivalente Dirichlet Bayesiana com priori uniforme*, f_{DBeu}, e sua forma é dada pela Equação (2.9) conforme descrito por Campos (2006):

$$f_{DBeu}(D,G) = \\ = log(P(G)) + \sum_{i=1}^{d}\left(\sum_{j=1}^{q_i}\left(log\left(\frac{\Gamma(\frac{\eta}{q_i})}{\Gamma(N_{ij}+\frac{\eta}{q_i})}\right) + \sum_{k=1}^{c_i} log\left(\frac{\Gamma(N_{ijk}+\frac{\eta}{c_i q_i})}{\Gamma(\frac{\eta}{c_i q_i})}\right)\right)\right). \tag{2.9}$$

Os critérios de Informação

Além dessas métricas baseadas na função Dirichlet Bayesiana, outra abordagem possível são as medidas provenientes da informação que se utiliza do logaritmo da função de verossimilhança que pode ser escrita como:

$$LL(G,D) = \sum_{i=1}^{d}\sum_{j=1}^{q_i}\sum_{k=1}^{c_i} N_{ijk} log\left(\frac{N_{ijk}}{N_i j}\right).$$

Associando a ela um termo de penalização da complexidade do modelo dado por $\sum_{i=1}^{d}(c_i-1)q_i$ multiplicada por uma função não negativa de regularidade $h(n)$ relativa ao tamanho amostral assim, obtém-se a função geral dos critérios de informação (CAMPOS, 2006), dada por:

$$f_{LL}(G,D) = \sum_{i=1}^{d}\sum_{j=1}^{q_i}\sum_{k=1}^{c_i} N_{ijk} log\left(\frac{N_{ijk}}{N_i j}\right) - \left\{\sum_{i=1}^{d}(c_i-1)q_i\right\} h(n). \tag{2.10}$$

Para casos especiais da Equação (4.1.1), se $h(n) = 1$, obtém-se o critério de informação Akaike (*Akaike Information Criterion* - AIC) (AKAIKE, 1973):

$$f_{AIC}(G,D) = \sum_{i=1}^{d} \sum_{j=1}^{q_i} \sum_{k=1}^{c_i} N_{ijk} log\left(\frac{N_{ijk}}{N_ij}\right) - \sum_{i=1}^{d} (c_i - 1)q_i, \tag{2.11}$$

e para $h(n) = \frac{1}{2}log(n)$, obtém-se o critério de informação Schwarz, também chamado de critério de informação Bayesiano (BIC) (SCHWARZ, 1978):

$$f_{BIC}(G,D) = \sum_{i=1}^{d} \sum_{j=1}^{q_i} \sum_{k=1}^{c_i} N_{ijk} log\left(\frac{N_{ijk}}{N_ij}\right) - \left\{\sum_{i=1}^{d} (c_i - 1)q_i\right\} \frac{1}{2} log(n). \tag{2.12}$$

O cálculo das métricas para cada rede em um domínio de possíveis soluções é uma tarefa de grande complexidade (CHICKERING, 1995) e portanto, deve estar aliado a algoritmos computacionais que se propõem a fazer uma busca guiada no conjunto, ou subconjuntos, de estruturas candidatas.

A partir de agora, os algoritmos que serão apresentados fazem parte da classe *greedy search* de heurísticas baseadas em métricas, são eles o *Hill Climbing* e o *Tabu Search*.

Hill Climbing - HC

O *Hill Climbing* é um procedimento básico de busca por estruturas que retornam o maior valor de pontuação e utiliza o conceito de vizinhança para restringir o conjunto de possibilidades rastreadas. É uma técnica simples que parte de uma estrutura inicial, que em uma quantidade finita de passos, encontra uma solução que maximiza a métrica f previamente definida (GÁMEZ; MATEO; PUERTA, 2011).

Para cada possível solução G no espaço de busca S, a vizinhança, $V(G)$, é definida como uma função $V : S \rightarrow 2^S$ que aplica à cada uma das possibilidades um subconjunto não vazio de S (BERETTA *et al.*, 2018). A exploração dessa vizinhança é feita localmente com movimentos de adição, remoção ou alteração no sentido dos arcos da estrutura analisada, sendo que isso ocorre em cada iteração (SCUTARI; VITOLO; TUCKER, 2019).

Russell e Norvig (2010) explanam sobre a utilização do *Hill Climbing*, chamado também de *greedy local search* pois seleciona a melhor estrutura vizinha sem vislumbrar o próximo passo. O procedimento é uma simples maximização de uma função *score*, a notação G^* segue para a estrutura com maior valor da medida a ser otimizada e G e G' estruturas candidatas. Os passos descritos no Algoritmo 4 seguem conforme Scutari, Vitolo e Tucker (2019), Nagarajan, Scutari e Lèbre (2013), sendo a base de dados, D e a função objetivo a ser maximizada f.

O algoritmo segue com iterativos cálculos de estrutura e comparação de métricas calculadas recursivamente para cada modificação no grafo. Como esse algoritmo busca sua solução em um conjunto vizinho, ou seja, faz uma busca local e de acordo com Scutari, Vitolo e Tucker

Algoritmo 4 – *Hill Climbing*

ENTRADA(D,f)
1: inicie com um grafo vazio (em geral, mas não precisa ser vazio);
2: calcule o valor de $f(G,D)$;
3: faça $max(f(G^*,D)) = f(G,D)$ e $G^* = G$
4: **enquanto** $max(f(G^*,D))$ aumenta **faça**
5: **para** cada possível adição, remoção ou reversão da direção de arcos em G^*, onde G^* é um DAG **faça**
6: calcule o valor de f da modificação na estrutura G', $f(G',D)$
7: **se** $f(G',D) > f(G^*,D)$ e $f(G^*,D) > f(G,D)$ e G^* não cíclico **então**
8: $G = G'$ e $f(G,D) = f(G',D)$
9: **se** $f(G,D) > f(G^*,D)$ **então**
10: $f(G^*,D) = f(G,D)$ e $G^* = G$
SAÍDA (G^*) =0

(2019), para garantir que o método não se prende ao máximo da região, uma adaptação de *recomeços* é utilizada adicionando um passo 11 extra no Algoritmo 4.

Essa evolução do procedimento é feita de modo que existam saltos aleatórios fora da vizinhança perturbando a escolha vigente. A transição, chamada também de *recomeço aleatório*, pode ser feita em uma quantidade t de vezes, recomeçando a partir do passo 4 do algoritmo.

Tabu Search - TS

A meta-heurística *Tabu Search* (TS) proposta em Glover (1986) é um procedimento de busca guiada para explorar o espaço de soluções evitando o aprisionamento em ótimos locais, passível de acontecer com o algoritmo anterior uma vez que não trata dessa possibilidade em suas regras. Essa metodologia seleciona uma nova solução em um conjunto vizinhos à solução atual de modo que maximize a função objetivo selecionada (BOUCKAERT, 1995) e quando não encontra uma métrica mais elevada, o algoritmo termina, selecionando a última estrutura candidata.

O *Tabu Search* é uma adaptação de outros métodos tradicionais de busca, como o *Hill Climbing* mas tem um aspecto diferencial que é o uso de memória adaptativa para a ibinição de buscas em estruturas já visitadas. Esse método cria uma ordenação de estruturas recentemente visitadas em uma lista, chamada de lista *tabu*, de tamanho L e, a cada iteração, ela é atualizada podendo revisitar alguma estrutura apenas L passos depois (RUSSELL; NORVIG, 2010).

O procedimento da heurística considerando a maximização da função objetivo f está sintetizado no Algoritmo 5, conforme Scutari, Graafland e Gutiérrez (2019), sendo D e f, respectivamente, a base de dados e a função objetivo.

As condições de parada mais utilizadas são quando o número de iterações é maior que o máximo de iterações permitidas, ou se nenhuma alteração foi aplicada a melhor solução nas últimas iterações (BERETTA *et al.*, 2018).

Algoritmo 5 – *Tabu Search*

ENTRADA(D,f)
1: inicie com um grafo vazio;
2: calcule o valor de $f(G,D)$;
3: faça $max(f(G^*,D)) = f(G,D)$ e $G^* = G$
4: **enquanto** $max(f(G^*,D))$ aumenta **faça**
5: **para** cada possível adição, remoção ou reversão da direção de arcos em G^*, onde G^* é um DAG, e tal que não tenha sido visitado nas últimas L alterações **faça**
6: calcule o valor de f da modificação na estrutura G', $f(G',D)$
7: **se** $f(G',D) > f(G^*,D)$ **então**
8: $G_0 = G^* = G'$ e $f(G_0,D) = f(G,D) = f(G',D)$
9: **se** $f(G,D) > f(G^*,D)$ **então**
10: $f(G^*,D) = f(G,D)$ e $G^* = G$
11: volte ao passo 4
SAÍDA (G^*)

Ou seja, segundo Glover e Hanafi (2001) o procedimento gera uma trajetória da vizinhança incluindo um mecanismo que proíbe a busca revisitar soluções já encontradas; nesse mesmo artigo, o autor discute a convergência do algoritmo que não será tratada nesta dissertação.

Essas metodologias apresentadas para estimação são parte do conjunto de opções com o objetivo de estimar como se dão as conexões entre as variáveis aleatórias de uma Rede Bayesiana, e esse, é um dos componentes do modelo a ser estimado. Outra tarefa, que depende dessa, é a estimação dos parâmetros de cada configuração $X_i|\mathbf{pa}(X_i)$, ou seja, para cada uma das possíveis combinações entre cada uma das instâncias das variáveis X_i e os estados de seu conjunto de pais $\mathbf{pa}(X_i)$, um parâmetro deve ser estimado.

Algumas das métricas apresentadas, as baseadas em critérios de informação, derivam a complexidade da Rede Bayesiana diretamente, conforme a quantidade de parâmetros da rede a serem estimados, referentes à distribuição conjunta de probabilidade, de acordo com cada configuração (CAMPOS, 2006), a definição dessa complexidade é dada pelo termo de regularização apresentado na Equação (2.11):

$$\sum_{i=1}^{d}(c_i - 1)q_i.$$

Esse termo varia com respeito às categorias de X_i, c_i e a quantidade de configurações diferentes do conjunto $\mathbf{pa}(X_i)$, q_i. E essa estimação de parâmetros será abordada na última seção deste capítulo.

A próxima seção trata de um novo método híbrido de estimação de estrutura, o *Scoring and Restrict*.

2.4 Metodologia *Scoring and Restrict*

Além dos métodos apresentados anteriormente, esta dissertação propõe um novo método de estimação híbrida que se baseia nas ideias de máximização de uma métrica e realização de testes de independência condicional, conforme algumas abordagens metodológicas já visitadas.

Contudo, a metodologia utilizada se diferencia das demais por considerar a existência de uma variável específica que seria alvo da predição. Em geral, essa discriminação de uma variável de interesse não é encontrada nos métodos de estimação de estrutura, uma vez que em grande parte, todas as variáveis possuem um mesmo nível de interesse e não focam suas restrições em uma variável especial.

Para melhor entendimento dos procedimentos tomados pelo método, o Algoritmo 6 esboça a construção da estrutura utilizando essa metodologia e é inspirado pelos Algoritmos 5 e 1, com a exclusividade da utilização da função objetivo $K2$ na fase de busca pela maximização da métrica. Portanto, os parâmetros de entrada são: i) a base de dados ($\mathbb{D}$); ii) o teste de independência condicional (TIC) e iii) a variável de interesse (C).

Primeiramente, um método *score-based* é aplicado na base de dados maximizando a função objetivo $K2$, obtendo-se uma rede completa de relação entre as variáveis, portanto, essa fase recebe o nome de *scoring*. Verificam-se então, as conexões diretas das covariáveis com a variável específica a ser predita, que são partes fundamentais da sua Cobertura de Markov. Esse conjunto é basicamente uma lista que especifica a incidência dos arcos (direcionados) de cada um dos pais para a variável a ser predita, essa lista é essencial para o método pois compõe a entrada da próxima fase do algoritmo, e recebe o nome de *whitelist*.

Na fase de *restrict*, como o nome sugere, o algoritmo baseado em testes de independência condicional inicia sua construção normalmente, com um grafo completamente conectado. Contudo, a *whitelist* restringe o conjunto de arcos que essa metodologia busca para realizar os testes. Ou seja, as relações estabelecidas anteriormente com a variável de interesse são mantidas mesmo depois da sucessão de retiradas de arcos realizada por esse algoritmo.

Por fim, a metodologia híbrida recebe o nome de *scoring and restrict*, nesta ordem, pois primeiramente, estima-se uma estrutura de rede maximizando uma métrica ($K2$), com busca via *Tabu-Search*, G_{K2}, conforme apresentado em Russell e Norvig (2002), Scutari (2010), linhas 1 à 11 do Algoritmo 6. Posteriormente, se utiliza das conexões diretas encontradas no resultado anterior relativas à variável de interesse C, $\mathbf{pa}_{G_{K2}}(C)$, como *whitelist*, minimizando o espaço de busca de estruturas para o algoritmo baseado em testes de independência condicional (*PC-stable*), conforme Scutari (2010), linhas 11 à 22 do Algoritmo 6. Como saída, o algoritmo retorna a lista de pais $\mathbf{pa}_{\mathbb{G}_{K2+PC}}(X_i)$ para cada variável aleatória $X_i \in \mathbf{X}$, formando a estrutura completa, $\mathbb{G}_{K2+PC}$. Esse método é também denotado no texto como $K2$+PC,

Com o intuito de verificar a adequabilidade do *Scoring and Restrict* para a estimação de estrutura, será conduzido um estudo de simulação com dados artificiais para diferentes cenários,

Algoritmo 6 – *Scoring and Restrict*

SCORING
ENTRADA($\mathbb{D}$, C) ▷ Base de dados, Variável de Interesse
1: inicie com um grafo vazio;
2: calcule o valor de $f_{K2}(G,D)$;
3: faça $max(f_{K2}(G^*,\mathrm{D})) = f_{K2}(G,\mathrm{D})$ e $G^* = G$;
4: **enquanto** $max(f_{K2}(G^*,\mathrm{D}))$ aumenta **faça**
5: **para** cada possível adição, remoção ou reversão da direção de arcos em G^*, onde G^* é um DAG, e tal que não tenha sido visitado nas últimas L alterações **faça**
6: calcule o valor de f_{K2} da modificação na estrutura G', $f_{K2}(G',\mathrm{D})$
7: **se** $f_{K2}(G',\mathrm{D}) > f_{K2}(G^*,\mathrm{D})$ **então**
8: $G_0 = G^* = G'$ e $f_{K2}(G_0,\mathrm{D}) = f_{K2}(G,D) = f_{K2}(G',\mathrm{D})$
9: **se** $f_{K2}(G,D) > f_{K2}(G^*,D)$ **então**
10: $f_{K2}(G^*,\mathrm{D}) = f_{K2}(G,\mathrm{D})$ e $G^* = G$
11: volte ao passo 4
SAÍDA (G_{K2})

seja $\mathbf{pa}_{G_{K2}}(C)$ o conjunto de pais da variável de interesse C, de acordo com a estrutura G_{K2} e seja $\mathbf{W}$ o conjunto de arcos que ligam cada um dos componente de $\mathbf{pa}_{G_{K2}}(C)$ à C;

RESTRICT
ENTRADA($\mathbb{D}$, TIC , C, W) ▷ Base de dados, Teste de Independência, Variável de Interesse, *Whitelist*
12: reinicie com um grafo G completo:
13: **para** m em $(0,\ldots,d-2)$ **faça**
14: **repita** para todos $(X_i,X_j) \in \mathbf{X}, i \neq j$, tal que X_i possui, pelo menos, m vizinhos na estrutura G atual, excluindo X_j e tal que $(X_i,X_j) \notin \mathbf{W}$:
15: **escolha** um novo subconjunto $\mathbf{Z}$ de dimensão m dos vizinhos de X_i, exceto X_j;
16: **se** $(X_i \perp X_j|\mathbf{Z},\mathrm{D})$ **então**
17: **remova** $X_i - X_j$ de G e faça $\mathbf{Z}_{X_i,X_j} = \mathbf{Z}$, como um conjunto que separa X_i de X_j;
18: **se** X_i e X_j não são mais adjacentes, ou não existem mais opções para $\mathbf{Z}$ **então**volte ao passo 14
19: **substitua** para $X_i \rightarrow X_j \leftarrow X_k$, para cada trio $X_i - X_j - X_k$, tal que X_i e X_k não são adjacentes e $X_j \notin \mathbf{Z}_{X_i,X_j}$.
20: **encontre** outras direções segundo as regras:
21: **se** $X_i - X_j$ e existe um caminho direcionado de X_i par X_j **então substitua** por $X_i \rightarrow X_j$;
22: **se** X_i e X_j não adjacentes, mas $X_i \rightarrow X_k$ e $X_k - X_j$ **então substitua** por $X_k \rightarrow X_j$.
SAÍDA ($\mathbb{G}_{K2+PC}$)

apresentado na Seção 5.2, comparando sua performance à dos algoritmos que o compõe. Além disso, estudos de aplicação em conjuntos de dados reais investigando essa comparação são conduzidos e apresentados no Capítulo 6.

Por unir métodos de abordagens distintas, o *Scoring and Restrict* ou *K2+PC* resulta em um modelo de rede com uma quantidade parcimoniosa de conexões entre as variáveis de

seu domínio, uma vez que equilibra a rede densa, que é geralmente resultado do *K2* com a rede pouco conectada ajustada pelo algoritmo *PC*. Traz uma possibilidade de entendimento das relações/influências entre as variáveis de maneira mais simplificada, além de, em teoria, priorizar a predição da variável resposta na construção da rede.

Essa priorização da variável resposta ocorre pois, em geral, as metodologias de estimação de estrutura são desenvolvidas visando o entendimento das relações entre variáveis, a adaptação sugerida para o método *scoring and restrict* fixa os arcos pais encontrados pelo *K2* para variável resposta.

A seção a seguir se refere à estimação do segundo elemento que compõe as Redes Bayesianas, os parâmetros.

2.5 Estimação dos Parâmetros

A estimação de parâmetros é a segunda tarefa primordial no contexto de modelagem de Redes Bayesianas, conforme formalizado na Equação (2.3). O termo $P(\Theta|\mathbb{G},\mathbb{D})$, dessa equação, determina que os parâmetros estão condicionados à estrutura gráfica da rede $\mathbb{G}$ e aos dados $\mathbb{D}$ utilizados com a finalidade de representar o sistema, ou o raciocínio analisado. Visto isso, é possível reescrever a Equação (2.1), em completude a fatoração do componente probabilístico das Redes Bayesianas conforme sugerido em Russell e Norvig (2010):

$$P(\mathbf{X}) = \prod_{i=1}^{d} \mathbf{P}(X_i|\mathbf{pa}_{\mathbb{G}}(X_i), \theta_i), \tag{2.13}$$

sendo que cada um dos vetores θ_i representa o conjunto referente às combinações das variáveis aleatórias e seu conjunto de pais condicionado à estrutura gráfica $\mathbb{G}$; no caso das redes discretas é a combinação dos estados desses elementos, ou seja, a probabilidade condicional.

Quando os parâmetros podem ser escritos conforme Equação (2.13), são ditos serem globalmente independentes (HECKERMAN; GEIGER, 1995; NEAPOLITAN, 2004). As metodologias para estimação de parâmetros se baseiam nas suposições de que eles são global e localmente independentes. Ji, Xia e Meng (2015) fazem uma revisão de métodos utilizados para essa tarefa, lista duas formas principais de estimação de parâmetros: o primeiro, é a *estimação por máxima verossimilhança* e o segundo é o *método Bayesiano*.

2.5.1 Método Bayesiano

O Método Bayesiano se baseia na *priori*, essa distribuição com respeito a θ e é denotada por $P(\theta)$ - na filosofia bayesiana esse é o *grau de crença* nas diferentes escolhas possíveis de parâmetros (KOLLER; FRIEDMAN, 2009). Na Seção 1.4 a utilização desse conhecimento prévio é agregada aos valores observados no intuito de encontrar um equilíbrio entre ambas,

que é a atualização da informação em mãos, a chamada *distribuição a posteriori*. A posteriori é obtida da seguinte forma:

$$P(\theta|D) = \frac{P(\theta)P(D|\theta)}{P(D)}. \tag{2.14}$$

Para as distribuições consideradas discretas, $P(D) = \sum_\theta P(D|\theta)P(\theta)$, essa probabilidade de D é dita fator de normalização já que garante a restrição de valores de probabilidade, chamada também de *verossimilhança marginal*, e D é o conjunto de dados observados. Esse é o cálculo base do método Bayesiano de estimação de parâmetros (JI; XIA; MENG, 2015).

Especificamente, para as Redes Bayesianas, sabe-se que o conjunto de parâmetros é dependente da sua topologia gráfica $\mathbb{G}$, então $\theta = (\theta_{X_1|pa(X_1)}, \ldots, \theta_{X_d|pa(X_d)})$, fazendo com que a sua distribuição a priori possa ser escrita de forma:

$$P(\theta) = \prod_{i=1}^{d} P(\theta_{X_i|pa(X_i)}), \tag{2.15}$$

supondo independência entre eles. Da mesma forma, a condição própria das Redes Bayesianas, afirma que a verossimilhança global pode ser reescrita por sua decomposição em funções locais, e então, a posteriori é dada por:

$$\begin{aligned} P(\theta|D) &\propto P(\theta)L(\theta;D|\mathbb{G}) \\ &\propto P(\theta)\prod_{i=1}^{d} L_i(D;\theta_{X_i|pa(X_i)}). \end{aligned}$$

Que combinada com a Equação (2.15) resulta em:

$$P(\theta|D) = \frac{\prod_{i=1}^{d} P(\theta_{X_i|pa(X_i)})L_i(D;\theta_{X_i|pa(X_i)})}{P(D)}.$$

Duas especificidades dessa metodologia são as prioris, que devem ser selecionadas conforme prévio conhecimento a respeito dos parâmetros θ, e a forma de predição que se dá pela maximização da densidade a posteriori conforme atualização das novas observações e informações anteriores.

Prioris

As funções que compõem a distribuição a posteriori revelam o grau de crença no suporte de θ, dessa forma, quanto mais evidências a respeito dos parâmetros mais fortes são as convicções e mais informativa será a priori, quanto menos evidências, por outro lado, menos informativa será a função selecionada.

No caso Bayesiano, a priori mais comum na análise de redes discretas é a Dirichlet. Em geral, assume-se que as variáveis, em redes discretas, $X_i|pa(X_i)$ segue uma $Multinomial\ (\theta_{X_i|pa(X_i)})$,

com K_i classes para X_i; cada combinação de $pa(X_i) = \mathbf{u}$ para cada X_i é um parâmetro que possui uma distribuição a priori Dirichlet, com os hiperparâmetros $\alpha_{X_i|pa(X_i)} = (\alpha_{x_i^1|pa(X_i)}, \ldots, \alpha_{x_i^{K_i}|pa(X_i)})$ (KOLLER; FRIEDMAN, 2009), sua expressão análoga à Equação (1.9) é dada por:

$$\begin{aligned} P(\theta_{X_i|pa(X_i)}) &= Dir(\theta_{X_i|pa(X_i)}|\alpha_{x_i^1|pa(X_i)}, \ldots, \alpha_{x_i^{K_i}|pa(X_i)}) \\ &= \frac{\Gamma(\sum_{i=1}^{d} \alpha_{x_i^1|pa(X_i)})}{\prod_{i=1}^{d} \Gamma(\alpha_{x_i^1|pa(X_i)})} \prod_{i=1}^{d} x^{\alpha_{x_i^1|pa(X_i)}-1}. \end{aligned}$$

Conforme descrito no primeiro capítulo, a conjugação Multinomial-Dirichlet resulta em uma posteriori com reparametrização dada pela Equação 1.4.1. Esse vetor α da priori Dirichlet será considerado um parâmetro a ser otimizado, na Seção 5.1.

Predição

De acordo com Koller e Friedman (2009), quando o método Bayesiano é utilizado porém o estimador não possui forma fechada, ou seja, não é possível escrever a expressão analítica em termos de funções conhecidas, utiliza-se o processo de *máximo a posteriori* (*MAP*) o qual, como o nome sugere, busca por parâmetros que maximizem a probabilidade a posteriori. Parâmetros esses que de maneira geral são estimados conforme:

$$\begin{aligned} \hat{\theta} &= \arg\max_{\theta} P(\theta|D) \\ &= \arg\max_{\theta} \log P(\theta|D) \\ &= \arg\max_{\theta} \frac{P(\theta)P(D|\theta)}{P(D)} \\ &= \arg\max_{\theta} (\log P(\theta) + \log P(D|\theta)). \end{aligned}$$

A estimação *MAP* é sensível à parametrização das distribuições, o que não ocorre com o Método de Máxima Verossimilhança nem com o estimador Bayesiano de forma fechada (KOLLER; FRIEDMAN, 2009).

Vale ressaltar que as Redes Bayesianas também podem ser utilizadas para lidar com dados incompletos, para isso, utiliza-se de métodos de aproximação para fazer a estimação dos parâmetros, segundo a revisão de Ji, Xia e Meng (2015) os algoritmos mais utilizados são o algoritmo *Expectation Maximization (EM)* e métodos de Monte Carlo.

2.6 Comentários gerais

Neste capítulo foram exibidos os conceitos básicos que fundamentam a teoria das Redes Bayesianas como as propriedades de Markov, condição, cobertura e equivalência, a d-separação

e v-estruturas. Metodologias de estimação de estrutura estão apresentadas em ambas as classes de aprendizado restrito e baseado em métricas, além da nova proposta de metodologia híbrida o *Scoring and Restrict* (*K2+PC*). E técnicas de estimação dos parâmetros de forma bayesiana foram descritas. O próximo capítulo tratará da descrição dos classificadores, casos especiais de Redes Bayesianas.

CAPÍTULO

3

CLASSIFICADORES

As Redes Bayesianas foram criadas com o propósito de facilitar representações visuais de distribuições conjuntas de probabilidade e assim viabilizar meios convenientes de expressar incertezas e/ou causalidade, testar suposições e facilitar inferências por meio de observações amostrais representativas do sistema analisado (PEARL, 2000).

Além dessas finalidades, os modelos de Rede Bayesiana podem ser utilizados para predição (KORB; NICHOLSON, 2010). Em terminologia de aprendizado de máquina, tal predição pode ser realizada para variável resposta numérica, este caso denominado regressão, quanto variável resposta categórica, este caso denominado *classificação*. Para esse último grupo preditivo, os classificadores de Redes Bayesianas são tipos especiais de estruturas que possuem um propósito exclusivo para tarefas de classificação.

Neste caso de classificação as abordagens propõem determinar as categorias da variável resposta C para cada uma das n observações descritas por um conjunto de variáveis preditoras $\mathbf{X} = \{X_1, X_2, \ldots, X_d\}$ (BIELZA; LARRAÑAGA, 2014). Os dados disponíveis para aprendizado contém registros completos com covariáveis e rótulos conhecidos $D = \{(c^{(1)}, \mathbf{x}^{(1)}), (c^{(2)}, \mathbf{x}^{(2)}), \ldots, (c^{(n)}, \mathbf{x}^{(n)})\}$. O modelo resultante desse aprendizado é aplicado em novas observações de covariáveis afim de predizer a classe da variável resposta, fazendo com que a função classificadora seja dada por (SANTAFÉ, 2007):

$$\gamma : (x_1, x_2, \ldots, x_d) \rightarrow (c_1, c_2, \ldots, c_k).$$

Inúmeras propostas de classificadores estão estabelecidas na literatura, levantamentos de revisão desses classificadores são realizados desde Friedman, Goldszmidt e Lee (1998) e Cheng e Greiner (1999) os quais comparam algumas abordagens de classificação existentes na época. Mais recentemente, nesse mesmo sentido, Flores, Gámez e Martínez (2012) e posteriormente, Bielza e Larrañaga (2014) realizam uma revisão específica a respeito do assunto de classificadores de Redes Bayesianas discretas.

Essas diferentes abordagens utilizadas para classificação se diferem especialmente pelo tipo de estrutura adotada de relação entre as variáveis aleatórias. A organização dessas ligações pode ser considerada fixa, semiflexível ou flexível conforme a necessidade de estimação dessas conexões.

Os classificadores de estrutura fixa, como o nome sugere, possuem uma arquitetura rígida de acordo com seus pressupostos. Eles demandam apenas a identificação e estimação de parâmetros para que todo o modelo de classificação seja construído, um exemplo desse tipo é o *Naïve Bayes* (MARON; KUHNS, 1960) que determina que todas as covariáveis tenham uma única influência, a variável alvo, e a partir disso estimam os parâmetros (CHENG; GREINER, 1999).

Outros classificadores que possuem estrutura semiflexível, partem de fortes pressupostos como o da variável de classificação ser pai de todas as covariáveis. Contudo, flexibilizam em alguns pontos permitindo, por exemplo, que as covariáveis possam ter alguma outra variável preditiva como pai, um representante dessa classe é o *Tree-Augmented Naïve Bayes* (TAN) (FRIEDMAN; GOLDSZMIDT, 1996). Outro exemplo de classificador com estrutura semiflexível é o *k-Dependence Bayesian Network* (*k*DB) (SAHAMI, 1996) que além de permitir pais além da variável de classificação, restringe essa quantidade ao valor k. Desse modo, a flexibilização das restrição de construção do modelo, exige que parte da estrutura de ligação entre variáveis seja estimada conforme outras regras, métricas ou testes, e posteriormente, os parâmetros devem ser estimados.

Além disso, os classificadores podem ter estrutura flexível os quais não possuem, ou possuem poucas limitações de relação - e quantidade - entre variáveis. Portanto, a estimação de sua estrutura é uma etapa fundamental da modelagem, sendo sensível às regras dos algoritmos de aprendizado. Bielza e Larrañaga (2014) argumentam que para esse tipo de classificador, apenas as covariáveis que fazem parte da cobertura de Markov da variável de classificação são as que influenciam em sua predição. Além disso, a estimação de parâmetros também é um requisito.

Conforme a definição dos classificadores, sua predição é dada pelo estabelecimento de um estado que maximiza a probabilidade a posteriori de uma variável categórica C dado um conjunto de variáveis preditoras $\mathbf{X} = \{X_1, X_2, \ldots, X_d\}$ (CHENG; GREINER, 1999). Para isso, sua função é de aproximar a distribuição conjunta de probabilidade $P(C, \mathbf{X})$ por meio de uma fatoração de um modelo de Redes Bayesianas (BIELZA; LARRAÑAGA, 2014), que pode ser expressa da seguinte maneira geral:

$$P(C, \mathbf{X}) = P(C|\mathbf{pa}(C)) \prod_{i=1}^{d} P(X_i|\mathbf{pa}(X_i)), \tag{3.1}$$

sendo essa, condicionada a estrutura do grafo $\mathbb{G}$ (MIHALJEVIĆ; BIELZA; LARRAÑAGA, 2020) conforme abordagem.

Conforme comentado na Seção 2.5, a tarefa de predição se dá a partir da escolha da

classe que maximiza a probabilidade a posteriori das d variáveis preditivas, C a variável resposta com k classes, obtida pela regra de Bayes:

$$\begin{aligned} c_{pred} &= \arg\max_{c} P(C = c | X_1 = x_1, ..., X_d = x_d) \\ &= \arg\max_{c} \beta P(C = c) P(X_1 = x_1, ..., X_d = x_d | C = c), \end{aligned}$$

sendo $\beta = 1/\sum_{j=1}^{k} P(X_1 = x_1, ..., X_d = x_d | C = j)$ a constante normalizadora, a *priori* $P(C = c)$ e a *verossimilhança* $P(X_1 = x_1, ..., X_d = x_d | C = c)$ que é a distribuição conjunta condicionada ao valor da classe (RUZ; ARAYA-DÍAZ, 2018). Cada um dos classificadores possui sua própria função de predição conforme sua distribuição conjunta.

Os seguintes classificadores serão descritos e utilizados na condução dos estudos de comparação e combinação: *Naïve Bayes* (NB), *k-Dependence Bayesian Network* (*k*DB), *Tree-Augmented Naïve Bayes* (TAN), *Bayesian Network Augmented-Naïve Bayes* (BAN), *Avereged One-Dependence Estimator* (AODE) e *General Bayesian Network* (GBN).

3.1 *Naïve Bayes*

O *Naïve Bayes* (NB), proposto em Maron e Kuhns (1960), é o classificador mais simples e difundido dessa categoria. Sua arquitetura é rígida, portanto, não necessita de nenhum método de aprendizado de estrutura para modelagem (CHENG; GREINER, 1999), apenas os parâmetros são desconhecidos e devem ser estimados.

Seu pressuposto é de independência entre as variáveis preditoras $\{X_1, X_2, \cdots, X_d\}$ condicionada a variável de classificação $\{C\}$ potencializando a eficiência computacional já que uma menor quantidade de dados é necessária pois o número de parâmetros é reduzido em relação a outros métodos (CHENG; GREINER, 1999), além disso, a variável de classificação não deve possuir nenhum pai. Essa suposição de independência implica na seguinte fatoração:

$$P(C, \mathbf{X}) = P(C) \prod_{i=1}^{d} P(X_i | C). \tag{3.2}$$

Essa estrutura probabilística da Equação (3.2) é representada pelo grafo da Figura 11, para qual o valor de d é 5:

Ou seja, a variável resposta é considerada *pai* de todas as variáveis explicativas, e elas, por sua vez, não possuem nem filhos, nem outros pais. O Algoritmo 7 apresenta o aprendizado do modelo de classificador *Naïve Bayes*, sendo D a base de dados e C a variável de interesse.

A princípio, a suposição de independência pode parecer absurda, mas esse classificador produz resultado preditivo muito satisfatório até mesmo comparado a métodos mais complexos

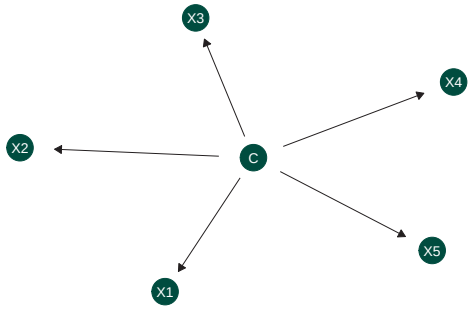


Figura 11 – Grafo da estrutura do classificador *Naïve Bayes*.

Fonte – Elaborado pela autora.

Algoritmo 7 – *Naïve Bayes*

ENTRADA(D, C)
1: determine todas as covariáveis como sendo filhos do alvo preditivo;
2: calcule as probabilidades de cada uma das combinações dos estados das variáveis.
SAÍDA ($\Theta_{\mathbb{G}_{NB}}$) =0

que dispensam um custo computacional elevado. Esse bom desempenho preditivo e simplicidade levou o *Naïve Bayes* a ser referência na sua categoria de classificador (CHENG; GREINER, 1999).

3.2 *Tree-Augmented Naïve Bayes*

A base do *Tree-Augmented Naïve Bayes* (TAN) é o Naïve Bayes, porém sua estrutura é semiflexível, uma vez que relaxa a suposição de independência condicional entre as variáveis explicativas. Esse classificador foi proposto em Friedman e Goldszmidt (1996) e possui C como a variável de interesse pressuposta a ser pai de todas as variáveis preditoras $\{X_1, X_2, \ldots, X_d\}$, as quais são permitidas depender de, no máximo, uma variável além da de classificação (BIELZA; LARRAÑAGA, 2014), além disso, a variável de classificação não deve possuir pais.

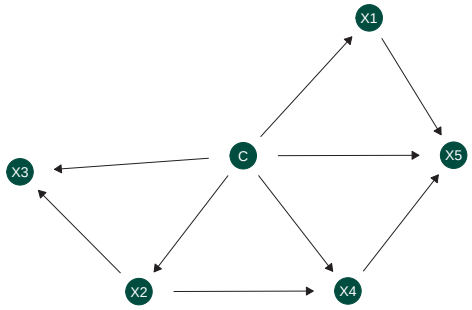


Figura 12 – Grafo da estrutura do classificador *Tree-Augmented Naïve Bayes*.

Fonte – Elaborado pela autora.

Esse classificador é inspirado no modelo de árvores Bayesianas condicionais apresentado

em Geiger (1992), que utiliza do Algoritmo de *Chow-Liu* (CHOW; LIU, 1968) para aprendizado de árvores que são, por definição, grafos que contém exato 1 pai para cada variável X_i exceto para aquela que não possui pais, chamada de nó raiz.

O Algoritmo 8 é uma adaptação do Algoritmo de *Chow-Liu* conforme descrito em Friedman e Goldszmidt (1996) que produz a estrutura do TAN, sendo que recebe os mesmos *inputs* do *Naïve Bayes*.

Algoritmo 8 – *Tree-Augmented Naïve Bayes*

ENTRADA(D, C)
1: calcule a informação mútua $IM(X_i, X_j|C)$, sendo $i \neq j$;
2: construa um grafo completo com cada variável em $\{C, \mathbf{X}\}$ sendo representada com um nó e cada arco possui o peso na informação mútua;
3: construa uma árvore de peso máximo a partir desse grafo conforme Gavril (1987);
4: transforme o resultado de uma arvore não direcionada, em uma direcionada escolhendo a variável de classificação como raiz ($|\mathbf{pa}(C)| = 0$) e direcionando seus arcos para o oposto dela.
SAÍDA ($\mathbb{G}_{TAN}$)

Sua representação probabilística pode ser dada por

$$P(C, \mathbf{X}) = P(C) \prod_{i=1}^{d} P(X_i|C, X_j),$$

que é um caso especial da Equação (3.1), onde $\mathbf{pa}(X_i) = \{C, X_j\}$, sendo X_j uma única outra covariável do conjunto $\mathbf{X}$ e $\mathbf{pa}(C) = \emptyset$, pois C é a variável raiz.

3.3 k-*Dependence Bayesian Network*

Introduzido por Sahami (1996), o *k-Dependence Bayesian Network*, Rede Bayesiana de k Dependências ou simplesmente k-DB, é flexível a ponto de permitir um limite de k pais para cada variável além da variável de classificação, e claro, por definição, não poder conter ciclos.

A flexibilização da restrição do TAN de possuir até um pai além da variável de classificação, faz com que esse parâmetro k, de *sintonização*, varie permitindo melhor generalização. Se o valor de k for grande o suficiente, espera-se que o modelo explore e capture todas as dependências que existem na base de dados (SAHAMI, 1996).

$$P(C, \mathbf{X}) = P(C) \prod_{i=1}^{d} P(X_i|\mathbf{pa}(X_i)), \tag{3.3}$$

sendo que $1 \leq |\mathbf{pa}(X_i)| \leq k$.

O Algoritmo 9 descreve as etapas de aprendizado do classificador k-DB conforme Sahami (1996), sendo que a base de dados é representada por D, a variável resposta por C e k, por definição, é a quantidade máxima de dependentes que uma covariável pode possuir.

Algoritmo 9 – k-DB

ENTRADA(D, C, k)
1: calcule a informação mútua $IM(X_i, C)$ para cada variável X_i;
2: calcule a informação mútua condicional $IM(X_i, X_j|C)$, sendo $i \neq j$;
3: seja Z uma lista vazia de variáveis utilizadas;
4: seja $\mathbb{G}$ a estrutura de Rede Bayesiana;
5: **repita** até Z incluir todas as variáveis do domínio:
6: selecione uma variável X_{max} que não esteja em Z e tenha o maior valor de $IM(X_{max}, C)$;
7: adicione um nó na estrutura $\mathbb{G}$ representando X_{max};
8: adicione um arco de C para X_{max};
9: adicione m arcos de diferentes X_j em Z com os maiores valores de $IM(X_{max}, X_j|C)$;
10: adicione X_{max} em Z;
SAÍDA ($\mathbb{G}_{KDB}$) =0

Com essa estrutura, as primeiras k variáveis terão menos de k pais, uma vez que o passo 9 seleciona o mínimo entre a cardinalidade de Z e o valor k, e por consequência, as $d-k$ variáveis restantes devem possuir os k pais (BIELZA; LARRAÑAGA, 2014). A Figura 13 apresenta um exemplo de estrutura do k-DB para qual o máximo de pais é $k = 2$, e as variáveis preditoras variam de X_1 a X_5.

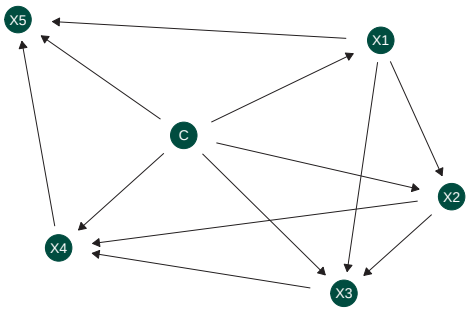


Figura 13 – Grafo da estrutura do classificador *k-Dependence Bayesian Network*, para $k = 2$.

Fonte – Elaborado pela autora.

Nesse caso, o conjunto de pais de cada uma das variáveis X_i é $pa(X_1) = \{C\}$, $pa(X_2) = \{C, X_1\}$, $pa(X_3) = \{C, X_1, X_2\}$, $pa(X_4) = \{C, X_2, X_3\}$, $pa(X_5) = \{C, X_1, X_4\}$.

Esse classificador é uma generalização de ambos os classificadores anteriores, quando define-se $k = 0$ o Naïve Bayes é obtido, ou seja, *0* dependências entre os atributos dada a classe, quando $k = 1$ retorna-se ao TAN, permissão de até uma dependência para cada covariável (BIELZA; LARRAÑAGA, 2014). Apesar essa relação entre o TAN e o 1-DB, na prática, é possível chegar a resultados distintos conforme o algoritmo de busca por essas dependências únicas é alterado.

3.4 Bayesian Network Augmented-Naïve Bayes

O *Bayesian Network Augmented-Naïve Bayes* (BAN) (FRIEDMAN; GEIGER; GOLDSZMIDT, 1997) é uma extensão do classificador Naïve Bayes e um caso especial do *k*DB. Esse classificador tem por característica a flexibilização da quantidade de pais, pois não possui limitações acerca das variáveis preditoras, mas mantém a variável de classificação sem pais (BIELZA; LARRAÑAGA, 2014) e mantém também, a restrição primária de serem acíclicos. Sua fatoração é dada por:

$$P(C,\mathbf{X}) = P(C)\prod_{i=1}^{d} P(X_i|\mathbf{pa}(X_i)), \tag{3.4}$$

sendo que $1 \leq |\mathbf{pa}(X_i)| \leq d$.

A Figura 14 apresenta um exemplo de classificador BAN, onde a variável C é pai de todas as covariáveis e não possui nenhum pai conforme estrutura do Naïve Bayes, e as preditoras não possuem um padrão estrutural, podem possuir até d pais, uma Rede Bayesiana de d dependências.

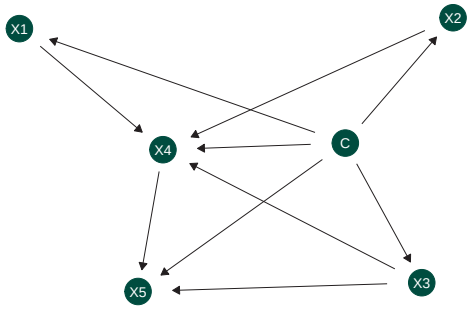


Figura 14 – Grafo da estrutura do classificador *Bayesian Network-Augmented Naïve Bayes*.

Fonte – Elaborado pela autora.

A estimação de sua estrutura possui um pressuposto fixo do Naïve Bayes e o restante do aprendizado é feito conforme o Algoritmo *KDB*, por meio de testes de independência condicional ou por meio de algum outro método de estimação estrutural, como os vistos no Capítulo anterior e descritos em Friedman, Geiger e Goldszmidt (1997) pela técnica baseada em métrica ou pela técnica baseada em testes conforme Bielza e Larrañaga (2014), uma adaptação do Algoritmo 9.

Os passos de construção desse classificador pode ser dividido em três fases, a primeira é a de rascunho, a segunda de encorpar a rede e a terceira e última, a de refinamento (BIELZA; LARRAÑAGA, 2014). O Algoritmo 10 reflete uma das possibilidades de estimação dessa estrutura.

Os três últimos classificadores apresentados, fazem parte de uma categoria de *Augmented Naïve Bayes* (ACID; CAMPOS; CASTELLANO, 2005), os quais mantém a estrutura inicial do *NB* mas, como o nome sugere, aumentam a quantidade de arcos da rede flexibilizando o número de dependências entre covariáveis.

Algoritmo 10 – BAN

ENTRADA(D, C)
Rascunho
1: determine que a variável de interesse é pai de todas as covariáveis em $\mathbf{X}$;
2: calcule uma medida de proximidade entre X_i e X_j condicionada a C, como informação mútua condicional $IM(X_i, X_j|C)$, sendo $i \neq j$;
Encorpar
3: adicione um arco quando pares de nós não possam ser d-separados;
Refinar
4: investigue cada um dos arcos por meio de testes de independência condicional, removendo os arcos de pares de variáveis que possam ser d-separadas.
SAÍDA ($\mathbb{G}_{BAN}$) =0

3.5 *Avereged One-Dependence Estimator*

O *Avereged One-Dependence Estimator* (AODE), proposto em Webb, Boughton e Wang (2005), em tradução literal é um estimador médio de uma dependência e é a estrutura mais recente de classificação. Classificadores de única dependência são do tipo TAN que permitem, além da variável de classificação, apenas um pai para cada uma das covariáveis, contudo, existem outras variações dentro dessa mesma premissa.

Geralmente, os classificadores que permitem apenas uma dependência entre as covariáveis realizam uma seleção de modelo, que se torna um processo custoso computacionalmente (ZHENG; WEBB, 2017). O AODE tenta evitar esse problema, utilizando a média de todos os SPODE (*SuperParent 1-dependence classifier*, um classificador de uma dependência super-pai).

A Figura 15 apresenta um exemplo de estrutura do classificador SPODE com X_1 sendo super-pai, as variáveis que estão nas posições centrais dos grafos são as únicas que influenciam todas as covariáveis e a variável de classificação é a única dependência da superpai.

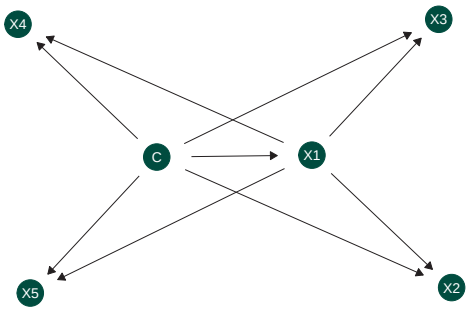


Figura 15 – Grafo da estrutura do classificador SPODE, com X_1 como super-pai.

Fonte – Elaborado pela autora.

De maneira geral, o AODE seleciona uma classe limitada de classificadores que podem ter somente uma variável como pai e, agrega as predições de todos os classificadores qualificados dentro dessa classe (WEBB; BOUGHTON; WANG, 2005). Para ser qualificado, o modelo deve

possuir uma acurada probabilidade estimada fazendo com que diminua a variância do classificador e o custo computacional (BIELZA; LARRAÑAGA, 2014). A sua forma probabilística, para cada um dos superpais, pode ser dada por:

$$P(C,\mathbf{X}) = P(C)P(X_j|C)\prod_{i=1 \text{ e } i\neq j}^{d-1} P(X_i|X_j,C),$$

sendo que X_j é a variável superpai e $i \neq j$, segundo Bielza e Larrañaga (2014) e a partir dessa equação é possível escrever uma forma geral para as predições do AODE, da seguinte maneira:

$$P(C,\mathbf{X}) = \frac{1}{u}\sum_{j=1}^{u} P(C)P(X_j|C)\prod_{i=1 \text{ e } i\neq j}^{d-1} P(X_i|X_j,C). \tag{3.5}$$

sendo u a quantidade de variáveis qualificadas como superpais e para deixar mais claro, como não é permitido haver *loops* na construção desse tipo de modelo, no produtório, define-se que $i \neq j$.

Esse classificador é chamado de *ensemble*, ou seja, uma combinação, já que agrega vários estimadores para gerar uma estrutura para o modelo final.

3.6 General Bayesian Network

O último classificador descrito possui estrutura flexível, não se prende a pressupostos relacionados à variável de classificação, nem às covariáveis. O classificador *General Bayesian Network* (GBN) tem por característica considerar a variável resposta como uma variável qualquer da rede sendo permitido que tenha um conjunto $\mathbf{pa}(C) \neq \emptyset$, mas também não determina essa condição.

Portanto, sua forma pode ser escrita como:

$$P(C,\mathbf{X}) = P(C|\mathbf{pa}(C))\prod_{i=1}^{d} P(X_i|\mathbf{pa}(X_i)),$$

Configurações desse tipo pode ter inúmeros arranjos diferentes de conexão entre as variáveis contudo, de acordo com Cheng e Greiner (1999), as variáveis necessárias para predição de C são as componentes de sua cobertura de Markov, os pais, os filhos e os pais de filhos formam esse conjunto. E a estimação dos parâmetros é feita após o estabelecimento da estrutura conforme ilustrado no capítulo anterior.

A estimação de sua estrutura se dá, portanto, de maneira irrestrita e podem utilizar qualquer técnica de estimação, seja baseada em métricas, seja baseada em testes de independência condicional, seja utilizando ambas (SCUTARI; DENIS, 2014), conforme descrito no Capítulo 2.

Portanto, sua estrutura não possui um padrão de conexão, e a Figura 16 apresenta um exemplo de GBN contendo apenas as variáveis integrantes de sua cobertura de Markov, os únicos elementos necessários para a predição de C.

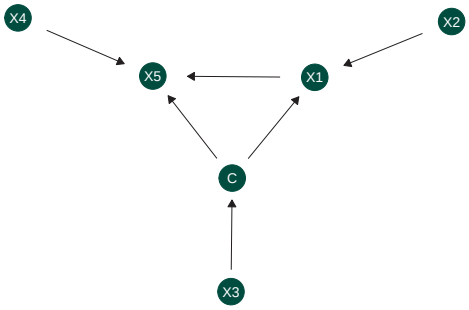


Figura 16 – Exemplo de estrutura do classificador *GBN*.

Fonte – Elaborado pela autora.

Diante de toda a teoria é possível abranger ainda mais abordagens diversas com variações dos classificadores apresentados encontradas na literatura, como a inclusão de aspectos adicionais como procedimentos de seleção de variáveis (PAZZANI; BILLSUS, 1997), combinações de modelos (JING; PAVLOVIĆ; REHG, 2008; LOUZADA; ARA, 2012), generalizações (LI *et al.*, 2007) e também introdução de variáveis latentes (KWOH; GILLIES, 1996; ANANDKUMAR *et al.*, 2013).

Segundo Bielza e Larrañaga (2014) os classificadores de Redes Bayesianas possuem algumas vantagens em relação a outros modelos utilizados para a finalidade de predição, como por exemplo a possibilidade de interpretação de seus resultados já que apresentam, em sua maioria, uma forma gráfica explícita de conexão entre variáveis. Além disso, por possuir um modelo probabilístico aliado ao grafo, diagramas de decisões podem ser aplicados, como apresentado em Barber (2012); pode também acomodar metodologias de seleção de variáveis, conforme Fu e Desmarais (2010) descreve a utilização da cobertura de Markov como uma maneira de obter uma síntese informativa a respeito das dependências (ALIFERIS *et al.*, 2010), e tratamento de observações faltantes ou dados incompletos (SINGH, 1997) entre outros importantes aspectos que enaltecem os classificadores de Redes Bayesianas frente a outras abordagens.

3.7 Comentários gerais

Neste capítulo, foram descritos os classificadores: *Naïve Bayes* (*NB*), *Tree-Augmented Naïve Bayes* (*TAN*), k-*Dependence Bayesian Network* (*KDB*), *Bayesian Network Augmented-Naïve Bayes* (*BAN*), *Avereged One-Dependence Estimator* (*AODE*) e *General Bayesian Network* (*GBN*), suas suposições e estruturas gráficas e probabilísticas primordiais para a realização da tarefa de classificação.

CAPÍTULO

4

AVALIAÇÃO E COMPARAÇÃO ENTRE CLASSIFICADORES

Para mensurar a qualidade dos modelos de Redes Bayesianas para classificação, que foram descritos no capítulo anterior, serão utilizadas algumas abordagens de avaliação, uma delas é referente ao ajuste do modelo aos dados e, a outra verifica a capacidade preditiva do modelo para observações não utilizadas para o ajuste do modelo. Para tal, medidas de capacidade preditiva são descritas na primeira seção deste capítulo. A segunda seção apresenta um estudo comparativo entre os classificadores para bases de dados frequentemente utilizadas na literatura.

4.1 Medidas de Avaliação

A avaliação das redes pode ser dada de duas maneiras: a de ajuste do modelo, a qual indica o quão confiável é o padrão de representação dos dados com finalidade, não exclusiva, de inferência, e são quantificada por meio das medidas de *plausibilidade do ajuste*. Por outro lado, o poder de predição do modelo é avaliado pelas medidas de *performance preditiva* as quais avaliam a qualidade do modelo para essa finalidade.

4.1.1 Plausibilidade do Ajuste

As métricas de plausibilidade de ajuste avaliam a capacidade do modelo de se adequar aos dados, portanto, são baseadas em uma função que os descreve completamente, a função de verossimilhança, já exibida na Seção 2.3.2. Contudo, quando essa função atribuída é demasiadamente complexa a ponto de levar ao super ajuste aos dados, ela pode inviabilizar a generalização do modelo. Para contornar essa característica adiciona-se à sua fórmula um termo de penalização da complexidade do modelo dada por $\sum_{i=1}^{d}(c_i-1)q_i$, sendo q_i é a quantidade de combinações de pais em $\mathbf{pa}(X_i)$ para cada variável X_i e c_i é o número de classes da variável X_i, associada à uma

função não negativa de regularidade $h(n)$ relacionada ao tamanho amostral, obtém-se, então, a função geral desses critérios que é dada por:

$$f = \sum_{i=1}^{d} \sum_{j=1}^{q_i} \sum_{k=1}^{c_i} N_{ijk} log\left(\frac{N_{ijk}}{N_i j}\right) - \left\{\sum_{i=1}^{d} (c_i - 1) q_i\right\} h(n).$$

Para casos especiais dessa equação, se $h(n) = 1$, obtém-se o critério de informação Akaike (*Akaike Information Criterion* - AIC) (AKAIKE, 1973):

$$AIC = \sum_{i=1}^{d} \sum_{j=1}^{q_i} \sum_{k=1}^{c_i} N_{ijk} log\left(\frac{N_{ijk}}{N_i j}\right) - \sum_{i=1}^{d} (c_i - 1) q_i,$$

e para $h(n) = \frac{1}{2} log(n)$, obtém-se o critério de informação Schwarz, também chamado de critério de informação Bayesiano (BIC) (SCHWARZ, 1978):

$$BIC = \sum_{i=1}^{d} \sum_{j=1}^{q_i} \sum_{k=1}^{c_i} N_{ijk} log\left(\frac{N_{ijk}}{N_i j}\right) - \left\{\sum_{i=1}^{d} (c_i - 1) q_i\right\} \frac{1}{2} log(n).$$

Outras medidas como as utilizadas para a estimação da estrutura pelos métodos de maximização também podem ser utilizadas para mensurar a qualidade do ajuste dos modelos aos dados.

4.1.2 Performance Preditiva

Por outro lado, para avaliar a performance preditiva dos modelos, é possível utilizar o conceito de matriz de confusão que organiza e compara os retornos do modelo com o real valor da variável de interesse, no caso de classificação, a classe a ser predita. Esse recurso auxilia de maneira natural a determinação de quantidade de classificações corretas, indicadas pela diagonal principal e todas as outras células representam as predições incorretas do modelo (JAMES *et al.*, 2013).

Um exemplo de uma matriz de confusão está apresentado na Tabela 1, para o caso de classificação binária, mas pode ser estendida para quaisquer quantidades de categorias da variável predita. Suas células representam *VP* que são os verdadeiros positivos, *VN* os verdadeiros negativos; *FP* os falsos positivos e *FN* os falsos negativos.

As medidas relativas à matriz de confusão estão apresentadas abaixo para uma quantidade k de classes a serem preditas, sendo que a extensão dessa matriz pode ser generalizada para uma matrix $k \times k$ com elementos C_{ij}, sendo que para $i = j$ a classificação está na diagonal principal representando o acerto do modelo e, caso contrário, as predições incorretas.

As medidas estão apresentadas a seguir:

Tabela 1 – Esquema da Matriz de Confusão para Classificação Binária.

	Positivo Observado (1)	Negativo Observado (0)
Positivo Predito (1)	VP	FP
Negativo Predito (0)	FN	VN

Fonte – Elaborado pela autora.

Acurácia

A acurácia (ACC) é a proporção de predições corretas dentre todos os registros da amostra de teste e é definida como:

$$ACC = \frac{\sum_{i=1}^{k} C_{ii}}{\sum_{i,j=1}^{k} C_{ij}},$$

sendo k é o número de classes as quais os resultados podem ser observados, $\sum_{i=1}^{k} C_{ii}$ é a somas das predições corretas e $\sum_{i,j=1}^{k} C_{ij}$ é o número total de observações.

Sua forma específica para predição binária é dada por:

$$ACC = \frac{(VP+VN)}{(VP+FP+FN+VN)}.$$

Coeficiente de Correlação de Matthew

O coeficiente de correlação de Matthew (MCC) é uma medida utilizada para verificar a classificação geral do modelo e é interpretada similarmente ao coeficiente de correlação de Pearson (LOUZADA; ARA, 2012), visto que é uma generalização dessa medida (GORODKIN, 2004). Sua forma multi-classe é dada pela expressão:

$$MCC = \frac{\sum_{i,j,l=1}^{k} C_{ii}C_{lj} - C_{ji}C_{il}}{\sqrt{S_{i1}}\sqrt{S_{i2}}},$$

com $S_{i1} = \sum_{i=1}^{k}\left[\left(\sum_{j=1}^{k} C_{ji}\right)\left(\sum_{f,g=1 f\neq i}^{k} C_{gf}\right)\right]$ e $S_{i2} = \sum_{i=1}^{k}\left[\left(\sum_{j=1}^{N} C_{ij}\right)\left(\sum_{f,g=1 f\neq i}^{k} C_{fg}\right)\right]$.

Quanto mais próxima a 1, mais ajustada está a predição aos resultados observados, quanto mais próxima a *zero* mais aleatória. E essa medida é apresentada de maneira normalizada com amplitude unitária variando de 0 a 1.

A forma do coeficiente para classificação binária conforme matriz de confusão definida como:

$$MCC = \frac{(VP \times VN - FP \times FN)}{\sqrt{(VP+FP)(VP+FN)(VN+FP)(VN+FN)}}.$$

Spherical Payoff

Além das medidas baseadas na matriz de confusão, uma que é muito utilizada para verificar a qualidade do modelo de Rede Bayesiana é o *Spherical Payoff* que em tradução literal do inglês significa ”recompensa esférica”. A métrica dimensiona a proximidade da probabilidade esperada pelo modelo em relação à observada e é considerada de ajuste fino pois mensura em termos da probabilidade atribuída à classe predita e não apenas à predição final. Varia de 0 a 1, sendo que 1 está relacionado à um modelo de melhor qualidade. Sua forma é definida como:

$$SP = MOAC \frac{P_c}{\sqrt{\sum_{j=1}^{n} P_j^2}},$$

sendo *MOAC* (sigla para *mean over all cases*) é a média entre todos os casos, ou seja, a média da razão para cada uma das observações, P_c é a probabilidade atribuída a real classe do indivíduo, P_j é a probabilidade de cada estado da variável categórica e n é o número de estados da variável categórica (MARCOT, 2012).

Contudo, além da escolha das medidas de capacidade preditiva, é importante que a avaliação dos modelos ocorra de maneira justa, e para isso, é necessário que as medidas de performance sejam calculadas em uma amostra dos dados que não foi utilizada para fazer sua estimação, pois uma métrica preditiva no conjunto utilizado para desenvolver a estrutura e parâmetros não garante sua performance em produção, quando o modelo for exposto à outras observações que não participaram do processo de estimação (WITTEN *et al.*, 2016).

Nesse contexto, alguns recursos que garante o rigor da avaliação dos modelos estão apresentados a seguir são chamados de métodos de reamostragem e, segundo James *et al.* (2013), são mecanismos extremamente importantes no aprendizado estatístico uma vez permitem avaliar, e selecionar, de maneira robusta modelos e hiperparâmetros utilizados para modelagem.

4.1.3 Validação Cruzada

A validação cruzada é uma ferramenta que permite avaliar Para uma certificação de que as métricas dos modelos desenvolvidos reflitam seu comportamento é necessário segregar a base de dados em duas amostras. Uma delas é onde a estimação do modelo será feita e os cálculos de plausibilidade de ajuste também, ela é chamada de amostra de desenvolvimento ou amostra de treinamento. De outro lado, a chamada amostra de teste utilizada para mensurar a capacidade de predição em conjuntos compostos por observações diferentes das utilizadas para seu desenvolvimento mas, que ainda sim, devem ser relacionadas.

Considerando que o número de registros é limitado, alguns recursos de reamostragem podem ser utilizadas para essa tarefa de estimação do poder preditivo de um modelo, retirando diferentes amostras da base disponível para modelagem e testes com o intuito de melhor estimar o erro associado ao modelo ajustado. Algumas abordagens podem ser utilizadas para esse fim,

e as que são utilizadas como ferramentas desta dissertação estão descritas a seguir, que são o *k-fold* e o *hold-out*,

Hold-out

O método *hold-out* é o mais simples dentre eles, conforme o nome sugere, esse procedimento divide, aleatoriamente, a base de dados em registros utilizados para o desenvolvimento do modelo e o restante como amostra teste para validar a sua performance preditiva. Essa abordagem simples, que é bastante utilizada em modelagem, ainda pode ser potencializada com a repetição iterativa de separação da amostra, seguida de modelagem e avaliação, sendo que a performance final é dada por alguma medida de agregação, como a média, por exemplo, este procedimento é conhecido como *hold-out repetido*.

K-fold

O *K-fold* é um método de validação cruzada que possui o parâmetro K de estratificação da base de estudo, esse parâmetro deve ser escolhido arbitraria e parcimoniosamente, a literatura não sugere um valor único mas, o que é visto, em geral, são estudos com 5 ou 10 estratos (JAMES *et al.*, 2013). O procedimento aleatoriza a base de dados e a divide em K partições de mesmo tamanho; recursivamente, $K-1$ conjuntos para o desenvolvimento e o último para teste até que se esgotem os grupos. As medidas de capacidade preditiva são calculadas em cada um dos treinamentos e, em geral, sua média é considerada como a medida final para a metodologia ajustada.

Essa abordagem também pode ser potencializada com a utilização o recurso das repetições, conforme Witten *et al.* (2016) sugere, a utilização de 10 repetições de um $10-fold$ (divisão da base em 10 partes), faz com que os dados sejam modelados 100 vezes e produz uma boa estimativa do erro, para diminuir a variância da aleatoriedade dos conjuntos, ou seja, uma boa estimação da performance final do modelo ajustado.

Bootstrap

A técnica de *boostrap* pode ser utilizada com diversas finalidades, como a de estimação de erros em coeficientes de regressão linear, por exemplo, ou para definir a quantidade de vizinhos em um *kNN*, *k*-vizinhos mais próximos. O *bootstrap* é, basicamente, um método de estimação baseado em um procedimento de reamostragem aleatória com reposição sendo que a amostra gerada é do mesmo tamanho da base de dados original.

De acordo com Witten *et al.* (2016), a probabilidade de uma observação da base de dados não ser escolhida para compor a amostra *bootstrap* é de aproximadamente $0,368$, o que significa que, em uma base razoavelmente grande, cerca de $63,5\%$ das observações fazem parte de uma replicação dessa técnica. Usualmente, a replicação *bootstrap* é feita B vezes, sendo B um número suficientemente grande mas, que não cause um alto custo computacional.

Ambiente Computacional

O software R (R Core Team, 2018) é utilizado para fazer as estimações, cálculos e vizualizações dos exercícios de análise propostos nessa dissertação. Por meio dos pacotes `bnlearn` (SCUTARI, 2010), utilizado para ajuste das redes, `infotheo` (MEYER, 2014) e `network` (BUTTS, 2008), como suportes de cálculo e visualização das redes.

Além disso, foram implementadas as medidas: Coeficiente de Correlação de Matthew (MCC), em sua forma normalizada, para variar de 0 a 1, e para avaliar análises com variáveis respostas contendo mais de duas classes, e o *Spherical Payoff* (SP), estendido também para cenários multiclasse.

A próxima seção trata de um estudo comparativo entre os classificadores apresentados no Capítulo 3, conduzido em bases de dados *bechmarks* para modelagem.

4.2 Comparação entre os Classificadores

Esse estudo visa a comparação entre os classificadores tradicionais baseados em Redes Bayesianas tanto os de estrutura fixa, semi-fixa e flexível. Para isso, foram selecionados os seguintes: *Naïve Bayes* (NB), *Tree-Augmented Naïve Bayes* (TAN), *Avereged One-Dependence Estimator* (AODE), *k-Dependence Bayesian Network* (kDB) - com $k = 1, 2, 3$, *Bayesian Network Augmented-Naïve Bayes* (BAN) e *General Bayesian Network* (GBN) - utilizando as heurísticas *Hill Climbing* com a métrica *K2* (GBN_K2) e *Tabu Search* (GBN_HC+), com a mesma medida.

Para cumprir com esse objetivo, foram selecionados 10 conjuntos de dados, com variável resposta de natureza binária, para a análise de comparação das medidas, com diversos números de covariáveis, tamanho de amostra, natureza dos dados e número de classes da discretização que, quando necessária, foi feita por igualdade das frequências. Com exceção da *db2* obtida do pacote do `R kernlab`, os arquivos das demais bases de dados são *benchmarks* disponíveis no *UCI Repository*[1]. A Tabela 2 contém uma sumarização das características das bases utilizadas, com o número de observações, número de variáveis, o número máximo de níveis (ou estados) das variáveis categóricas, ou discretizadas disponíveis para análise e a proporção de sucessos para cada uma delas.

Para avaliar as técnicas utilizou-se o método de validação cruzada, sugerido por Witten *et al.* (2016) para suavizar o viés de escolha dos dados de treinamento e teste de *k-fold* repetido, com $k = 10$. Esse procedimento de validação deve ser repetido 10 vezes e a média entre todos os ajustes é o resultado da métrica utilizada. A performance então, foi avaliada pela acurácia (ACC), Coeficiente de Correlação de Matthew (MCC) e *Spherical Payoff* (SP).

[1] HTTPS://ARCHIVE.ICS.UCI.EDU/ML/DATASETS.PHP

Tabela 2 – Descrição quantitativa dos conjuntos de dados.

ID	Nome	Número de Observações	Número de Variáveis	Quantidade Máxima de Classes	Proporção da Variável Resposta (%)
db1	banknote	1373	5	11	44,5
db2	spam	4601	58	16	39,4
db3	bank	4521	17	16	11,5
db4	german credit	1000	19	10	30,0
db5	parkinson	195	23	5	75,4
db6	audit risk	775	26	9	39,3
db7	japanese credit	666	16	5	44,9
db8	ionosphere	351	33	7	64,1
db9	tic tac toe (disc)	958	10	3	65,3
db10	shuttle	58000	10	33	21,4

Fonte – Elaborada pela autora.

Resultados

A Figura 17 apresenta um mapa de calor que quantifica a proporção de vezes que os classificadores superaram uns aos outros quando a acurácia é a medida de avaliação. Para analisar esse gráfico, seleciona-se um classificador de interesse no eixo vertical , dentro de cada uma dessas células existem os valores da proporção de bases de dados às quais o classificador escolhido obteve desempenho superior. Por exemplo, selecionando o *kDB1 (1-Dependence Estimator)* no eixo y, nota-se que em nenhuma das análises ele obteve melhores resultados que o *AODE*, em 50% das bases foi superior ao *BAN* e em 90% delas foi superior ao *Naïve Bayes*.

Para complementar, a Figura 17 mostra a comparação direta entre classificadores conforme o valor da sua acurácia. Células mais escuras refletem que o classificador que está no eixo vertical teve maior frequência de melhor performance para essa medida com relação ao classificador que está no eixo X. Assim como, cores mais claras indicam maior proporção de pior performance para a medida analisada entre os respectivos classificadores. Por esse gráfico é possível notar que o classificador com melhor desempenho em relação aos demais é o AODE e o que apresenta as menores métricas é o *Naïve Bayes*.

Para melhorar a visualização do comportamento entre classificadores, a Figura 18 apresenta a comparação entre os classificadores para o Coeficiente de Correlação de Matthew na Figura 18a, que apresenta grande variabilidade e amplitude, o comportamento entre os classificadores para essa métrica é bastante semelhante. Já para o *Spherical Payoff* apresentado na Figura 18b, por ser uma medida de predição a nível de probabilidade, já apresenta comportamento mais diverso em relação a concentração dos valores de cada um deles mas, ainda sim, não é possível afirmar que são diferentes. Observa-se valores mais baixos para o *Naïve Bayes* em relação ao seu vizinho, o *TAN* e ao *AODE* que se concentram em um patamar mais elevado.

Por meio desses resultados é possível verificar que o classificador *AODE* apresentou melhores métricas em relação aos demais, o que é esperado uma vez que se trata de um

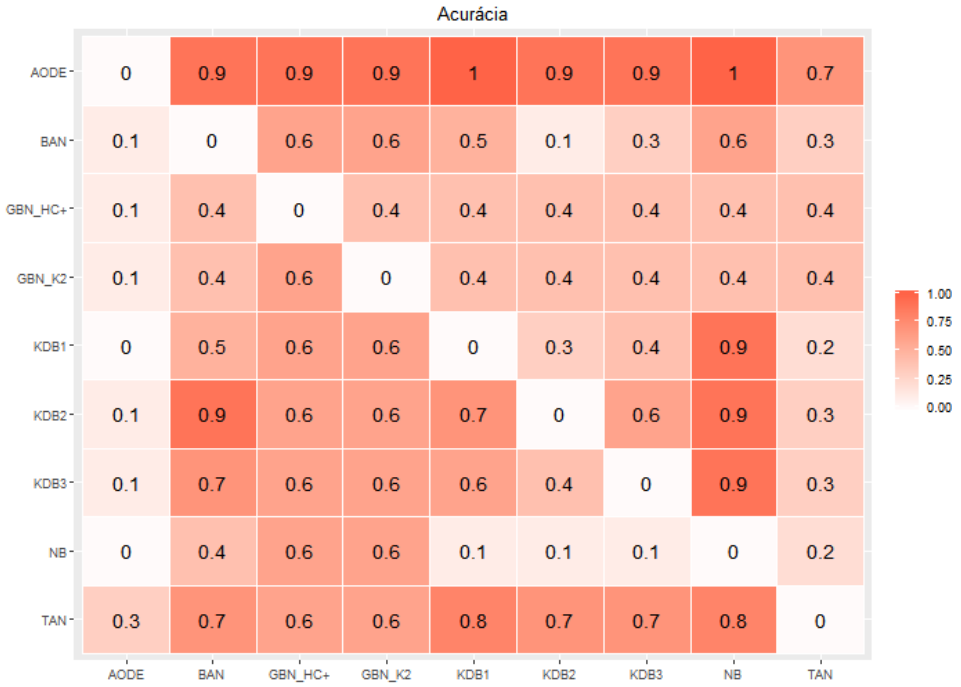


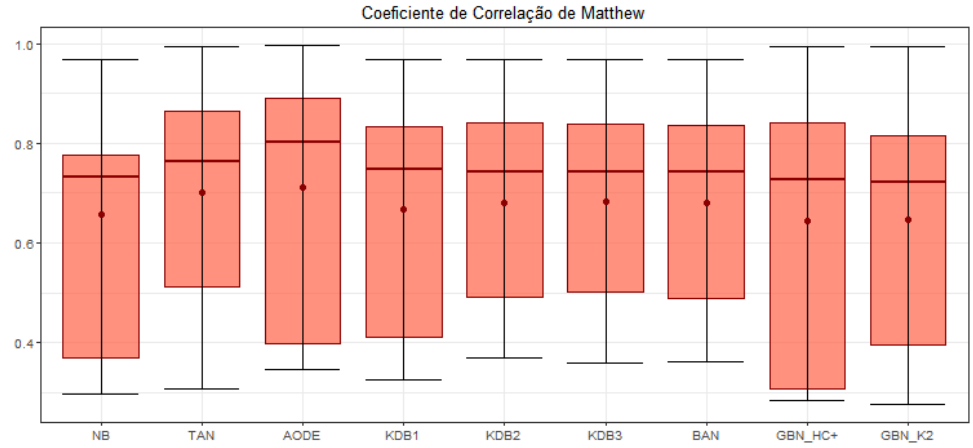
Figura 17 – Mapa de Calor da proporção de vezes que os classificadores que estão no eixo da vertical apresentaram maiores valores de acurácia comparados aos classificadores do eixo horizontal.

Fonte – Elaborado pela autora.

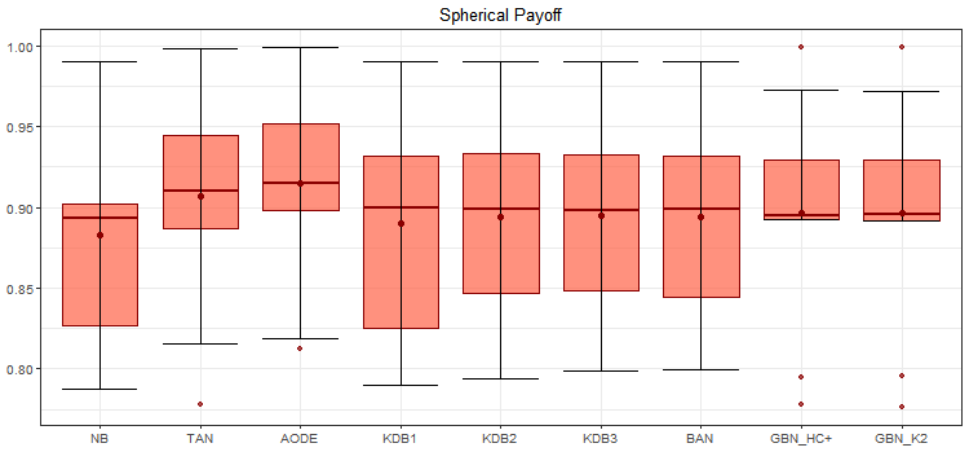
classificador *ensemble* que, por definição, visa melhor performance preditiva; o TAN fica na segunda posição do ranking de desempenho. O *Naïve Bayes* apresentou a menor capacidade preditiva, seguido de ambos classificadores GBN que tiveram comportamento semelhante e quando comparados entre si.

4.3 Comentários gerais

Este capítulo apresentou as formas de avaliação de modelos de Redes Bayesianas, métricas de plausibilidade de ajuste, capacidade preditiva, bem como as técnicas de reamostragem utilizadas para garantir uma estimação robusta da performance. O próximo capítulo apresenta estudos mais aprofundados do comportamento das Redes Bayesianas em seus dois componentes, a estrutura e os parâmetros.



(a) Coeficiente de Correlação de Matthew.



(b) *Spherical Payoff*

Figura 18 – Boxplots das Medidas de Capacidade preditiva para cada um dos classificadores.

Fonte – Elaborado pela autora.

CAPÍTULO

5

ESTUDOS DE SIMULAÇÃO

Este capítulo trata de dois estudos de simulação envolvendo a estimação dos elementos que compõem as Redes Bayesianas, sua estrutura e seus parâmetros. A estrutura, componente base desse tipo de modelo, deve ser treinada de forma a estabelecer o melhor conjunto de arcos que descreve as relações entre as variáveis disponíveis. E, baseados na estrutura, os parâmetros da rede são estimados, no caso das redes discretas, existem parâmetros a serem aprendidos para toda configuração distinta de variável condicionada a seus pais.

Conforme visto no Capítulo 2, a tarefa de estimação dos parâmetros em um modelo de Redes Bayesianas é um processo complexo e condicionado à arquitetura da rede. Visto isso, o primeiro estudo deste capítulo propõe a investigação da performance preditiva dos classificadores de Redes Bayesianas conforme os hiperparâmetros α são modificados, de acordo com a estimação bayesiana Dirichlet-Multinomial, já descrita anteriormente.

Na segunda seção é apresentado um estudo que compara métodos de estimação de estrutura tradicionais, para classificadores irrestritos, os *GBNs* descritos no Capítulo 3. Nessa investigação serão comparados métodos de cada uma das classes de heurísticas de estimação, o primeiro é o *K2* representante da classe dos algoritmos baseados em métricas, utilizando o *Tabu Search* para busca da estrutura. Representando a classe de algoritmos baseados em restrição, o *PC-stable*, versão atual do algoritmo *PC*, é utilizado para o aprendizado. Além desses, é proposto um algoritmo híbrido que combina essas metodologias anteriores de forma que suas características sejam potencializadas, gerando um classificador mais equilibrado sem perder sua capacidade preditiva.

Ambas as análises são conduzidas em bases de dados artificiais, o processo de simulação desses dados é descrito em cada uma das respectivas seções.

5.1 Estimação de Parâmetros

A condução do estudo da estimação de parâmetros é realizada no sentido de investigar a influência da modificação do valor do hiperparâmetro da priori Dirichlet quando associado à verossimilhança Multinomial conforme descrito no Capítulo 2 para a estimação das probabilidades a posteriori e, posteriormente, avaliando o ganho da performance preditiva ao se utilizar esse recurso.

Os cenários propostos para esse delineamento se diferem pelo aspecto de dependência entre as covariáveis. A simulação gera variáveis preditivas de natureza contínua seguindo uma distribuição normal multivariada, a ser discretizada de acordo com cada uma das categorias da variável dependente de natureza binária balanceada, conforme Louzada e Ara (2012).

Foram estudados cenários com 5, 10 e 15 variáveis. No primeiro caso, considerando uma situação de independência entre elas, a primeira classe da variável dependente possui variáveis preditivas seguindo uma normal com média $\mu_u = (0,0,0,0,0)$ e matriz de covariância dada por:

$$\Sigma_u = \begin{pmatrix} 4 & 0 & 0 & 0 & 0 \\ & 4 & 0 & 0 & 0 \\ & & 4 & 0 & 0 \\ & & & 4 & 0 \\ & & & & 4 \end{pmatrix},$$

para a segunda classe, o vetor de médias é $\mu_d = \left(\sqrt{\frac{1}{5}}, \sqrt{\frac{1}{5}}, \sqrt{\frac{1}{5}}, \sqrt{\frac{1}{5}}, \sqrt{\frac{1}{5}}\right)$ e matriz de covariância dada por:

$$\Sigma_d = \begin{pmatrix} 1 & 0 & 0 & 0 & 0 \\ & 1 & 0 & 0 & 0 \\ & & 1 & 0 & 0 \\ & & & 1 & 0 \\ & & & & 1 \end{pmatrix}.$$

Na situação de dependência entre as variáveis preditivas com vetor de médias dado por $\mu_l = (0,0,0,0,0,0)$ e matriz de covariância conforme:

$$\Sigma_{ij} = \begin{cases} 1, \text{ se } i = j \\ \frac{1}{4}, \text{ se } i \neq j \end{cases}$$

Para a segunda classe, o vetor de médias é dado por $\mu_r = \left(\frac{1}{2}, \frac{1}{2}, \frac{1}{2}, \frac{1}{2}, \frac{1}{2}\right)$ e matriz de covariância dada por:

$$\Sigma_{ij} = \begin{cases} 1, \text{ se } i = j \\ \frac{1}{2}, \text{ se } i \neq j \end{cases}$$

As definições são análogas para os casos de 5, 10 e 15 variáveis, sendo o vetor de média estendido para as respectivas dimensões, bem como a ordem das matrizes de covariância seguindo os escopos de independência e dependência em cada cenário.

A base de dados artificial gerada pelo procedimento descrito acima, ainda passa por um tratamento de discretização por frequência equivalente, tornando todas as variáveis explicativas em variáveis categóricas com experimentos de 2 e 10 classes. Além disso, para cada um dos cenários foram geradas amostra de tamanhos $\{500, 1000, 3000, 50000\}$, verificando possíveis padrões na diferença do número de observações.

Os valores do parâmetro α a ser otimizado na investigação estão no intervalo $[0,3]$ com uma variação $\delta = 0,1$, sendo o 0 uma comparação com o método de máxima verossimilhança uma vez que $\alpha > 0$. Portanto, foram explorados 30 pontos de observação do comportamento da capacidade preditiva conforme propostas, em cada um dos diferentes cenários. Para a avaliação das medidas preditivas foram conduzidas 30 reamostragens *bootstrap* e a mediana foi utilizada para agregar os resultados de cada uma dessas amostras, para os respectivos cenários de independência e configuração de tamanho amostral, número de variáveis e quantidade de classes.

As estruturas investigadas são as dos classificadores apresentados no Capítulo 3: *Naïve Bayes* (NB), *Tree-Augmented Naïve Bayes* (TAN), *Avereged One-Dependence Estimator* (AODE), *k-Dependence Bayesian Network* (KDB), com $k = 1, 2, 3$, *Bayesian Network Augmented-Naïve Bayes* (BAN), somente os que possuem alguma restrição em sua estrutura. Os *GBNs*, os classificadores que possuem a estimação da rede de forma irrestritas, estão presentem no estudo da próxima seção.

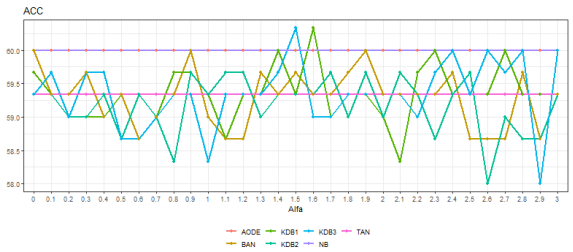
Nesse contexto, espera-se que os resultados indiquem tendência de aumento ou diminuição da capacidade preditiva conforme a modificação dos valores dos hiperparâmetros da priori Dirichlet quando associada à distribuição Multinomial, com o objetivo de estimação dos parâmetros da estrutura.

Resultados

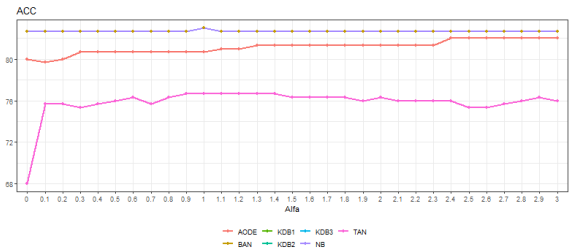
Este estudo foi conduzido utilizando o software R e o pacote especial de classificadores `bnclassify`, que contém as funções para o *Naïve Bayes*, o *Tree-Augmented Naïve Bayes* para o qual foi utilizado o algoritmo *Chow-Liu*, o *Avereged One-Dependence Estimator* (AODE), a implementação do *k-Dependence Bayesian Network* para $k = 1, 2, 3$ e, para o caso do *Bayesian Network Augmented-Naïve Bayes*, $k = n - 2$.

Inicialmente, os dados gerados para cada um dos cenários de dependência entre variáveis foram discretizados conforme as configurações apresentadas e então, os gráficos para a avaliação desses estudos estão dispostos da seguinte maneira: a primeira coluna é referente à discretização em 2 classes, a segunda coluna à discretização em 10 classes. Cada linha da matriz de gráficos apresenta as configurações de 5, 10 e 15 variáveis, respectivamente.

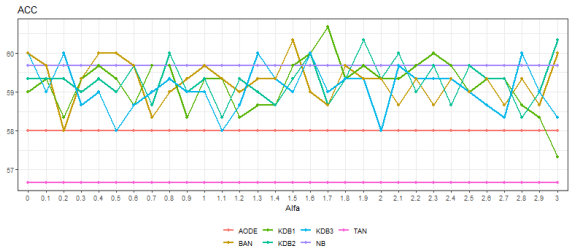
Os gráficos estão expostos em blocos para cada um dos tamanhos amostrais em 500, 1000, 3000 e 5000 observações. Os primeiros gráficos de cada um deles apresentam os resultados do cenário de independência entre as covariáveis e o segundo, o cenário de dependência.



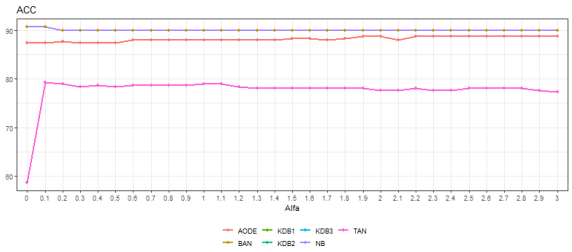
(a) 5 variáveis e 2 classes.



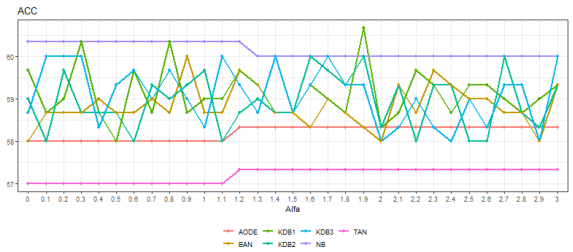
(b) 5 variáveis e 10 classes.



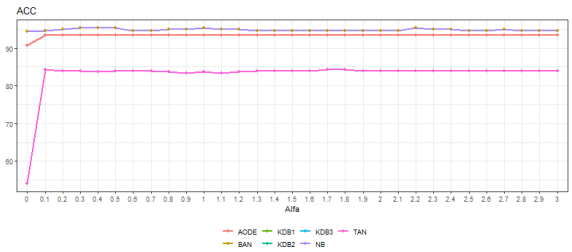
(c) 10 variáveis e 2 classes.



(d) 10 variáveis e 10 classes.



(e) 15 variáveis e 2 classes.



(f) 15 variáveis e 10 classes.

Figura 19 – Acurácia das 6 configurações para o cenário de independência entre variáveis em uma amostra de 500 registros.

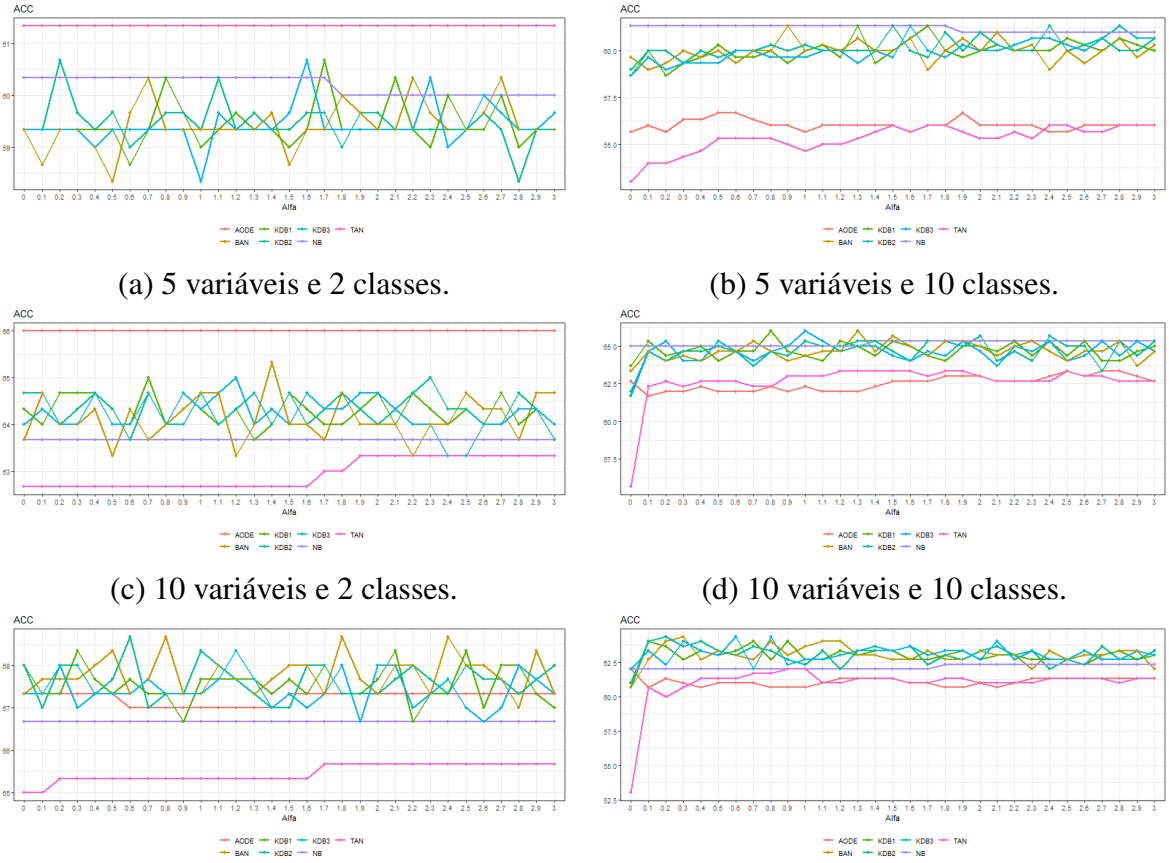
Fonte – Elaborado pela autora.

Conforme apresentado nas Figuras 19, 20, 21 e 22 é possível notar que, tanto para o cenário de dependência quanto para o cenário de independência entre as covariáveis existe uma instabilidade no comportamento dos classificadores com as alterações de α, em especial para os gráficos referentes aos menores tamanhos amostrais.

Apesar da inconstância das medidas, elas não aparentam refletir uma tendência de aumento, ou diminuição, da capacidade preditiva. De maneira geral, eles permanecem em um mesmo patamar ao longo do intervalo de valores do hiperparâmetro.

No mesmo sentido, para as discretizações com 10 classes, as linhas apresentam maior estabilidade especialmente no cenário de independência entre as variáveis explicativas, nos itens (b), (d) e (f) das Figuras 19 e 21.

Essa configuração de 10 classes exibe um comportamento peculiar do classificador *TAN*



(a) 5 variáveis e 2 classes.

(b) 5 variáveis e 10 classes.

(c) 10 variáveis e 2 classes.

(d) 10 variáveis e 10 classes.

(e) 15 variáveis e 2 classes.

(f) 15 variáveis e 10 classes.

Figura 20 – Acurácia das 6 configurações para o cenário de dependência entre variáveis, amostra de 500 registros.

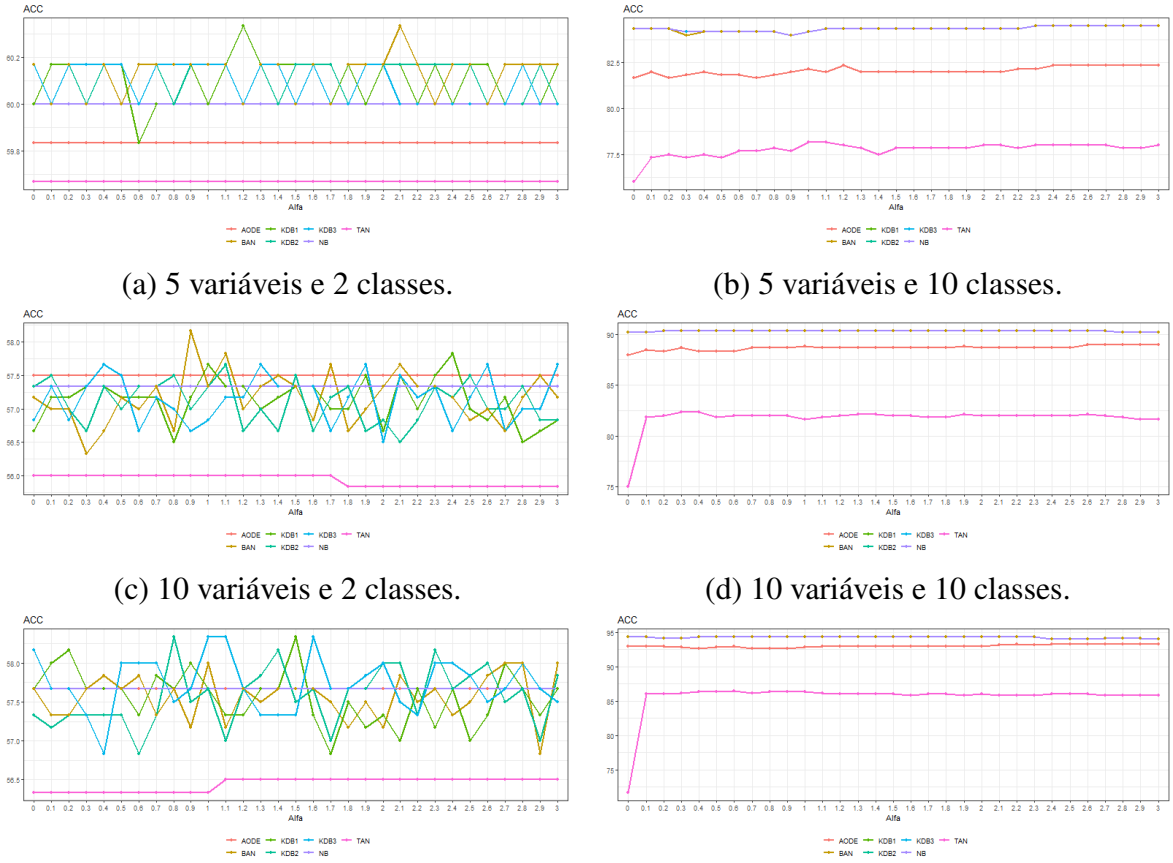
Fonte – Elaborado pela autora.

em relação aos demais. Para os menores tamanhos amostrais, em algumas configurações é possível verificar um salto crescente da medida preditiva para os menores valores de α, de 0 pra $0,1$, visualizada nos itens (b), (d) e (f) da Figura 19 , itens (d) e (f) das Figuras 20 21, e item (f), da Figura 22.

Mais suavemente, esse acréscimo também ocorre nos itens (d) e (f) da Figura 23, e (f) da Figura 25; esse comportamento está restrito ao classificador *TAN*, o restante dos classificadores se mantém com patamar constante de maneira estável, ou não, a depender do cenário.

De acordo com o aumento do tamanho amostral, Figuras 23, 24, 25 e 25, em geral, os classificadores se comportam de forma mais semelhante a proporção que se modifica o parâmetro estudado. Nos casos de dependências entre as variáveis é possível notar maior estabilidade o número de classes igual a 2, nas Figuras 24 e 25.

Além disso, para configurações com o número de variáveis reduzido, itens (a) e (b) das Figuras de 19 à 25, é possível notar sutil estabilidade dos classificadores, e sendo mais evidente quando o tamanho amostral é grande. Outros padrões relacionados ao número de variáveis não



(a) 5 variáveis e 2 classes.

(b) 5 variáveis e 10 classes.

(c) 10 variáveis e 2 classes.

(d) 10 variáveis e 10 classes.

(e) 15 variáveis e 2 classes.

(f) 15 variáveis e 10 classes.

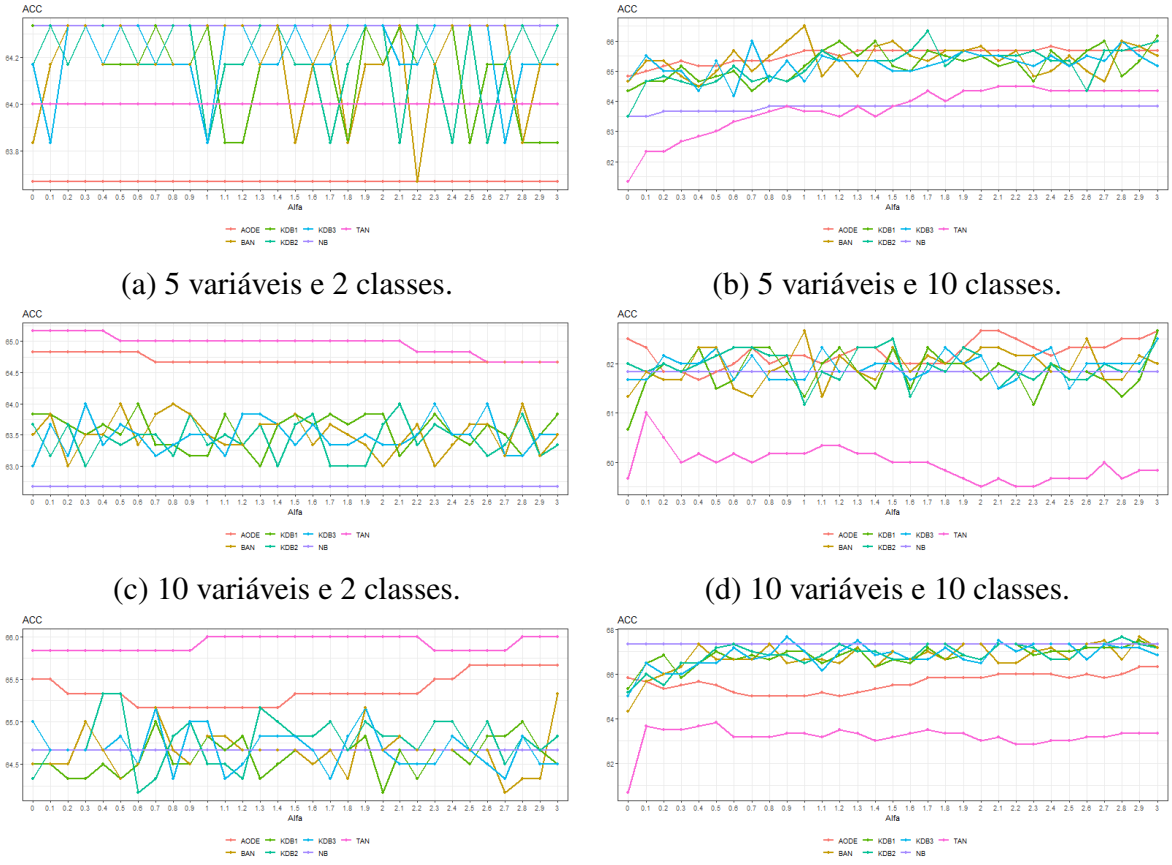
Figura 21 – Acurácia das 6 configurações para o cenário de independência entre variáveis em uma amostra de 1000 registros.

Fonte – Elaborado pela autora.

são perceptíveis. Com essas informações, por meio da simulação realizada, não é possível afirmar que o valor de α influencia a capacidade preditiva dos classificadores de uma maneira geral.

Vale ressaltar também, que em grande parte das configurações relativas ao cenário de covariáveis independentes, o *Naïve Bayes* é o destaque dentre os classificadores. Nesse mesmo contexto, os classificadores *BAN*, *KDB1*, *KDB2* e *KDB3*, seguem a mesma métrica do *NB*. Dadas as respectivas definições, a melhor performance do *NB* para o cenário de independência entre variáveis é completamente intuitivo, ao contrário do que se espera para os demais classificadores que o acompanham. Além disso, em geral, o *TAN* é o classificador que menos se destaca nessse contexto, apesar de ser, de alguma forma, sensível à mudança do hiperparâmetro.

Já para o cenário de dependência entre as variáveis explicativas, a performance do *AODE* é mais expressiva pois, na maioria dos casos, é um dos classificadores com os maiores valores de acurácia. Para as menores amostras, não existe uma clara discriminação entre os patamares de cada um dos classificadores, o comportamento geral de capacidade preditiva dos métodos é melhor estimado conforme o tamanho amostral aumenta, evidenciado a distinção entre o *AODE*,



(a) 5 variáveis e 2 classes.

(b) 5 variáveis e 10 classes.

(c) 10 variáveis e 2 classes.

(d) 10 variáveis e 10 classes.

(e) 15 variáveis e 2 classes.

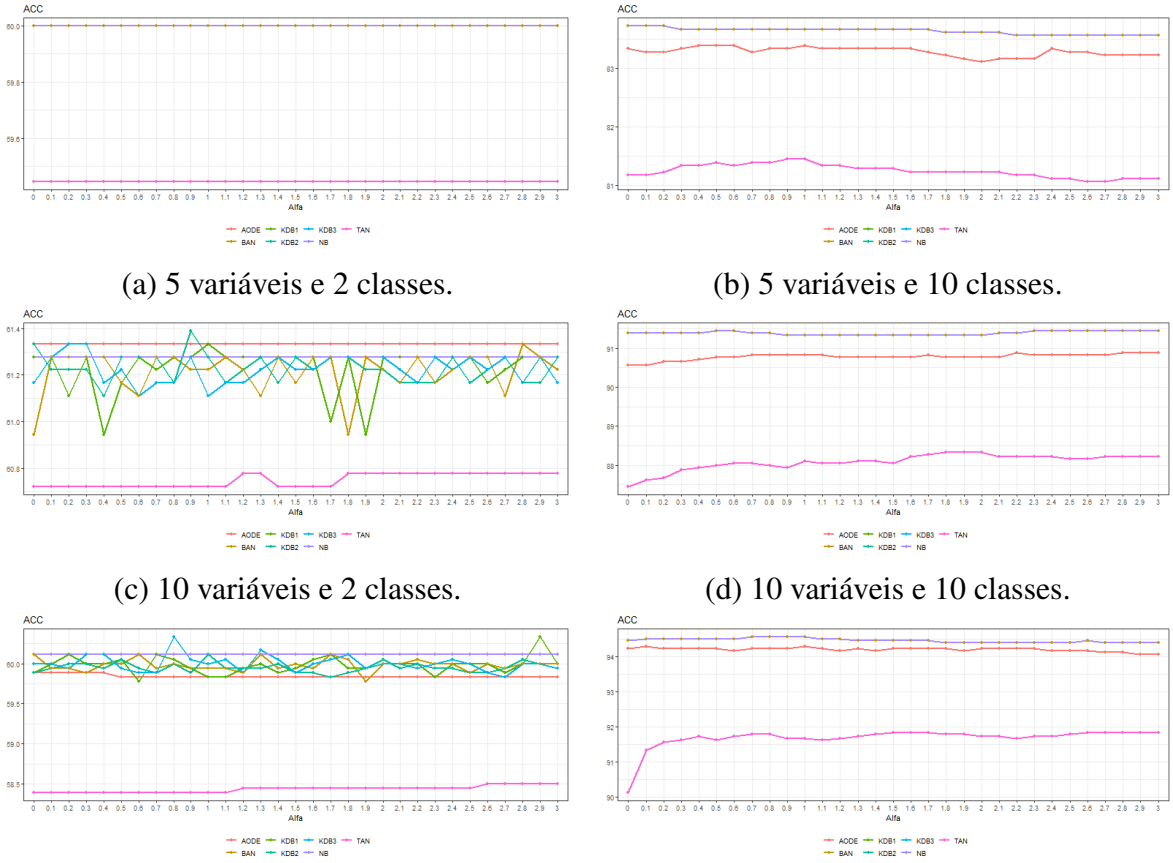
(f) 15 variáveis e 10 classes.

Figura 22 – Acurácia das 6 configurações para o cenário de dependência entre variáveis, amostra de 1000 registros.

Fonte – Elaborado pela autora.

TAN e os demais. Além do mais, nesses cenários de dependência, o *Naïve Bayes*, por sua vez, apresenta os menores valores de acurácia de forma geral. Os demais classificadores, seguem um comportamento mais semelhante entre si e mediano em relação à estes destaques mencionados.

Portanto, entende-se que os fatores determinantes para seu comportamento é o tamanho amostral, o número de variáveis e a quantidade de categorias que possuem, essas características estão relacionadas à quantidade de configurações distintas de rede e do número de pais que cada uma delas possui. Quando o número amostral é reduzido, não é uma regra que existam exemplares para todos os parâmetros, o que pode ocorrer é a atribuição desses valores como 0, sendo assim, esse estado é impossível de ser acessado em novas observações. O que o α traz é uma probabilidade não nula que, ainda que seja baixa, possibilita a atribuição de todos os rótulos e suas configurações. Então, estudos com classes desbalanceadas e/ou eventos raros possam proporcionar conclusões diferentes.



(a) 5 variáveis e 2 classes.

(b) 5 variáveis e 10 classes.

(c) 10 variáveis e 2 classes.

(d) 10 variáveis e 10 classes.

(e) 15 variáveis e 2 classes.

(f) 15 variáveis e 10 classes.

Figura 23 – Acurácia das 6 configurações para o cenário de independência entre variáveis em uma amostra de 3000 registros.

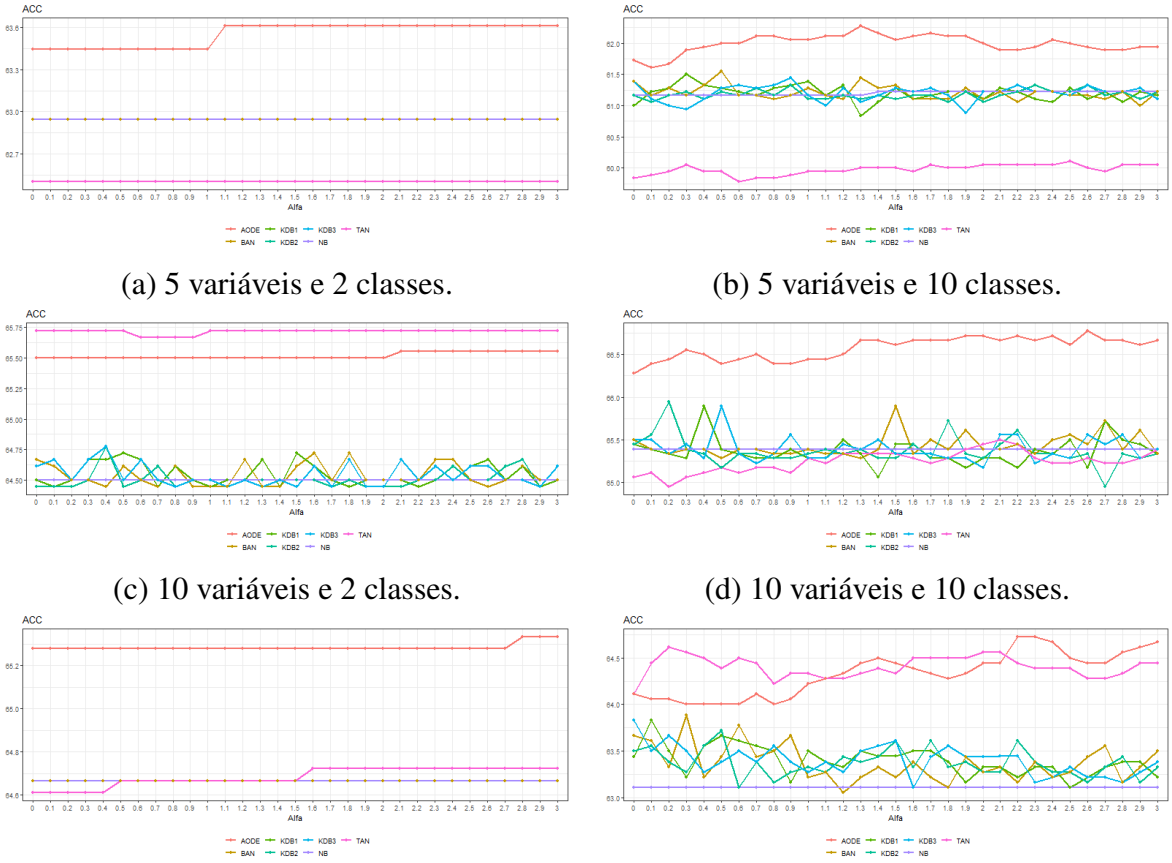
Fonte – Elaborado pela autora.

5.2 Estimação de Estrutura

O segundo estudo de simulação apresentado considera as possibilidades em torno dos procedimentos utilizados para cumprir a tarefa de estimação de estrutura. Nesse caso, foi proposto uma combinação de metodologias com o intuito de aumentar a capacidade preditiva de um modelo que possa ser graficamente visualizado.

Conforme foi elucidado em capítulos anteriores, os caminhos de predição e inferência, ou interpretação causal, podem, e tomam, rumos distintos no que tange os procedimentos para essas finalidades (JAMES *et al.*, 2013), apesar de estarem atrelados em alguns contextos conforme explorado em Izbicki e Santos (2020). Nesse sentido, os classificadores de Redes Bayesianas, que tem por objetivo a predição de classes em uma variável categórica, podem também, ter suas estruturas interpretáveis, ou ao menos, visuais (BIELZA; LARRAÑAGA, 2014).

Nesse contexto, o estudo investiga o método de classificação dado pela estimação da rede de maneira irrestrita, ou seja, não faz nenhum pressuposto de relação entre as variáveis disponíveis. Foram apresentados no Capítulo 3 como os *GBNs*, esse tipo de classificador se



(a) 5 variáveis e 2 classes.

(b) 5 variáveis e 10 classes.

(c) 10 variáveis e 2 classes.

(d) 10 variáveis e 10 classes.

(e) 15 variáveis e 2 classes.

(f) 15 variáveis e 10 classes.

Figura 24 – Acurácia das 6 configurações para o cenário de dependência entre variáveis, amostra de 3000 registros.

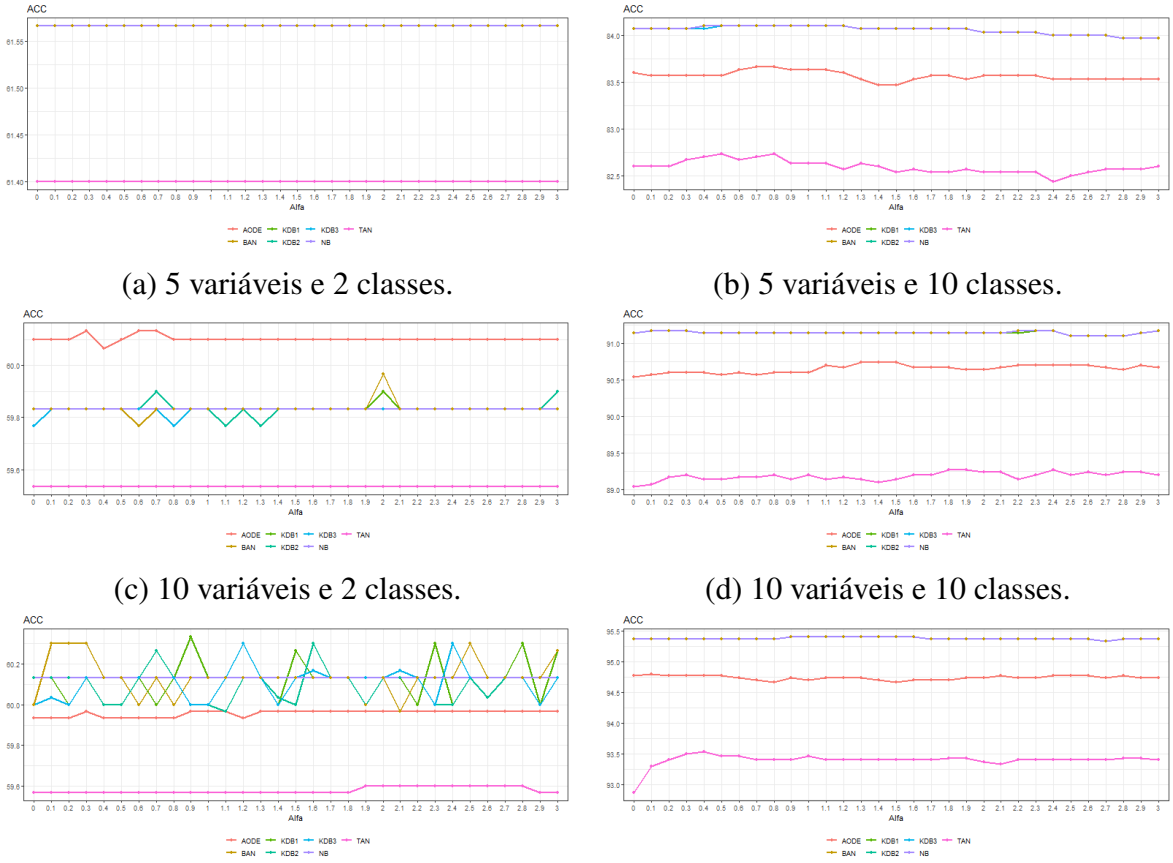
Fonte – Elaborado pela autora.

difere pela forma de estimação, que pode ser baseada em maximização de uma métrica ou por meio de testes de independência condicional, e até mesmo por métodos que combinem ambas abordagens.

Cada um dos grupos de estimação será representado por um método em específico, selecionado por ser amplamente utilizado na literatura. O algoritmo *K2* representa os baseados em métricas e o *PC*, os baseados em testes, além disso, é proposto uma metodologia híbrida que, especificamente, une esses dois algoritmos, o *scoring and restrict*.

Esse novo método de estimação híbrida é baseado na existência de uma variável específica a ser predita, visto que, em geral, os processos de estimação de estrutura em Redes Bayesianas consideram que todas as variáveis possuem um mesmo nível de interesse e não focam suas restrições em uma variável especial, mesmo que, a predição final tenha alguma delas como alvo, e está detalhado na Seção 2.4.

O experimento de simulação é projetado para responder as seguintes perguntas: i) Adequabilidade do método *scoring and restrict* para estimação de estrutura de uma base de dados



(a) 5 variáveis e 2 classes.

(b) 5 variáveis e 10 classes.

(c) 10 variáveis e 2 classes.

(d) 10 variáveis e 10 classes.

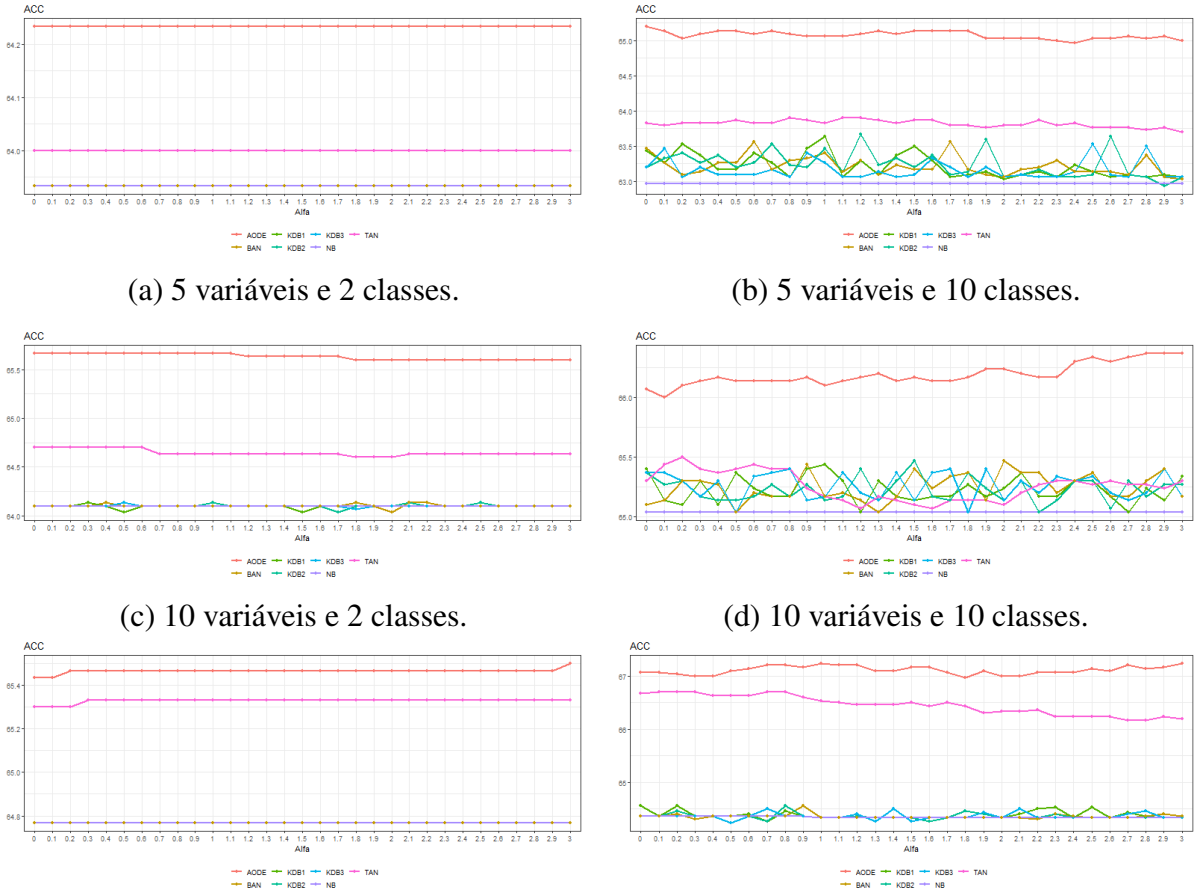
(e) 15 variáveis e 2 classes.

(f) 15 variáveis e 10 classes.

Figura 25 – Acurácia das 6 configurações para o cenário de independência entre variáveis para uma amostra com 5000 registros.

Fonte – Elaborado pela autora.

artificial. ii) Comparações com os métodos tradicionais, o algoritmo *K2* e o algoritmo *PC*. Para responder essas perguntas, utiliza-se conjuntos de dados sintéticos, além de *baselines* para comparação. A seguir, tais configurações são descritas.



(a) 5 variáveis e 2 classes.

(b) 5 variáveis e 10 classes.

(c) 10 variáveis e 2 classes.

(d) 10 variáveis e 10 classes.

(e) 15 variáveis e 2 classes.

(f) 15 variáveis e 10 classes.

Figura 25 – Acurácia das 6 configurações para o cenário de dependência entre variáveis, amostra de 5000 registros.

Fonte – Elaborado pela autora.

Conjunto de dados artificiais

A base de dados corresponde a três quantidades distintas de observações $\{100, 500, 1000\}$. Foram criadas 5 variáveis explicativas $\mathbf{X} = \{X_1, X_2, \ldots, X_5\}$ e uma variável resposta Y. Suas dependências probabilísticas foram simuladas através das relações apresentadas a seguir:

$$
\begin{array}{rll}
Y \mid X_3, X_4, X_5 \sim \text{Gamma}(\mu, \sigma) & \log(\mu) = 1.2X_3 + 2.5X_4 - 0.2X_4^2 + 1.5X_5 & \sigma = 0.65 \\
X_3 \mid X_1, X_2 \sim \text{Beta}(\nu, \phi) & \text{logit}(\nu) = 0.1X_1 + 0.05X_2 & \phi = 0.7 \\
X_1 \sim \text{Bernoulli}(\pi) & \pi = 0.3 & \\
X_2 \sim \text{Poisson}(\lambda) & \lambda = 25 & \\
X_4 \sim \text{Normal}(\eta, \tau) & \eta = 0.5 & \tau = 0.1 \\
X_5 \sim \text{Bernoulli}(\theta) & \theta = 0.65 &
\end{array}
$$

O grafo teórico dos dados simulados possui a estrutura apresentada na Figura 26; Y tem

como pais X_3, X_4, X_5, enquanto X_3, por sua vez, é filha de X_1 e X_2. Esta forma de controle da natureza de geração das variáveis utilizadas visa entender o padrão de desempenho dos métodos da estimação de estrutura utilizados nesse estudo.

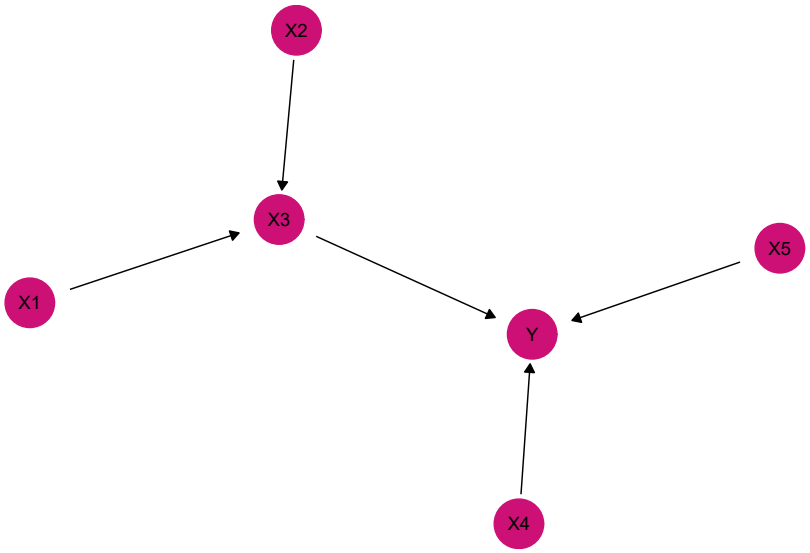


Figura 26 – Grafo da estrutura teórica simulada.

Fonte – Elaborado pela autora.

Dada a estrutura teórica gráfica, o número de parâmetros para cada configuração aumenta exponencialmente conforme discutido na Seção 2.3. A quantidade de parâmetros para o grafo da Figura 26, com L_{Xi} com $i = 1,\ldots,5$ as classes de discretização das variáveis explicativas e L_Y as classes da variável resposta, é dado por $2+L_{X_2}\left(1+L_{X3}L_{X1}\right)+2L_{X4}\left(L_YL_{X3}+0,5L_{X5}\right)$. Detalhes em (KORB; NICHOLSON, 2010), pág. 33. Assim, sendo o número mínino de parâmetros igual a 32, todas as variáveis binárias, e o número máximo igual a 182, com $L_{Xi} = 4$ e $L_Y = 3$.

O método proposto requer que as variáveis aleatórias sejam discretas e, então, a base de dados deve possuir apenas variáveis categóricas. Para atender essa restrição, quando necessário, foi aplicado um método de discretização de variáveis numéricas em classes de mesma frequência. A descrição das categorias das variáveis relevantes ao resultado, estão apresentadas ao final desta seção. A escolha da quantidade de classes é um dos fatores de mudança na configuração da rede e portanto, do tempo computacional necessário para cálculo das estimações de estrutura e parâmetros.

A metodologia descrita no Capítulo 2 é integralmente aplicada em cada um das bases de dados para configurações de 2 ou 3 classes de discretização da variável resposta e 2, 3 ou 4 classes de discretização para as covariáveis que possuem natureza numérica e precisam, portanto, serem categorizadas.

As análises das alterações de performance da rede estimada com relação a quantidade de classes escolhidas para as variáveis discretizadas, tanto as explicativas quanto a resposta, devem

ser testadas por meio de configurações distintas para avaliação de performance preditiva e para interpretação de causalidade, levando-se em consideração a escolha de modelos parcimoniosos.

Na estimação de estrutura, a estatística definida para os testes de independência condicional foi a *Informação Mútua Condicional* (CAMPOS, 2006), análoga à apresentada na Equação (2.4). Da mesma forma, na estimação baseada em métricas, a função objetivo utilizada para maximização foi a *K2* (LERNER; MALKA*, 2011). Por ser uma medida de ajuste de toda a rede é calculada com respeito ao conjunto de observações para cada variável contida em $\{\mathbf{X}, Y\}$ denotada por D, a dimensão de D é igual a $d+1$. O cálculo é condicionado a estrutura do grafo e a probabilidade da estrutura G é dada por $P(G)$, sendo expressa por,

$$f_{K_2}(D|G) = log(P(G)) + \\ + \sum_{i=1}^{d+1} \left(\sum_{j=1}^{q_i} \left(log\left(\frac{(c_i-1)!}{(N_{ij}+c_i-1)!} \right) + \sum_{k=1}^{c_i} log\left(N_{ijk}!\right) \right) \right),$$

sendo que, o número de classes da variável X_i é dado por c_i, o número de combinações das variáveis em $\mathbf{pa}(X_i)$ de X_i é q_i, N_{ijk} é o número de observações na base de dados na qual a variável X_i recebe o valor x_{jk} e $N_{ij} = \sum_{k=1}^{ri} N_{ijk}$. Essa medida é baseada em diversas suposições como as de multinomialidade e independência de parâmetros (CAMPOS, 2006), e o algoritmo de busca utilizado é o *Tabu Search*.

A estimação de parâmetros foi feita por meio da metodologia de *hold out* repetido, considerando 100 repetições de amostras distintas 70/30 - a média desse *loop* é utilizada como medida de performance comparativa. De modo que 30% da amostra de cada repetição foi utilizada para cálculo das medidas preditivas. O aprendizado das Redes Bayesianas, tanto da estrutura quanto o das tabelas de probabilidade condicional, que são os principais componentes da metodologia (RUZ; ARAYA-DÍAZ, 2018) tiveram que ser estimados.

Resultados

A Tabela 11 contém os resultados numéricos para cada um dos 36 modelos gerados divididos entre metodologias, número de classes de discretização de X e Y e quantidade de registros. Os valores estão apresentados multiplicados por 100, tanto para média quanto para o desvio padrão na estrutura de {média $\pm$ desvio padrão}; a ausência de valores indica que o algoritmo não foi capaz de gerar uma rede final que fosse direcionada para realizar a estimativa dos parâmetros.

Naturalmente, quando a quantidade de classes a serem preditas é menor, os valores das medidas preditivas são mais elevados. Portanto, a comparação deve levar em consideração a comparação da capacidade preditiva apenas entre os métodos para mesma discretização da variável resposta. Em geral, existe um ganho médio de acurácia na comparação da metodologia *K2* vs. *K2+PC* que é de 4%, o coeficiente de correlação de Mathews tem acréscimo de $3,17\%$ e

o *Spherical Payoff* apresenta um aumento médio de $2,25\%$. É visto que, conforme definição da metodologia *K2+PC*, se não houverem variáveis conectadas à resposta no método *K2* então, a rede final é uma rede *PC* isolada.

Para análise de ajuste, estão apresentadas nas Figuras 27 à 32, nas quais possuem a quantidade de registros nas colunas então coluna um corresponde ao tamanho 100, coluna 2 ao 500 e, coluna 3 ao de 1000. Semelhantemente, as linhas correspondem a quantidade de classes as quais as covariáveis foram discretizadas, de forma que a linha 1 corresponde a 2, a linha 2 corresponde a 3 e a linha 3 corresponde a 4 categorias. Cada uma das figuras faz referência ao método e a quantidade de classes de discretização da variável resposta.

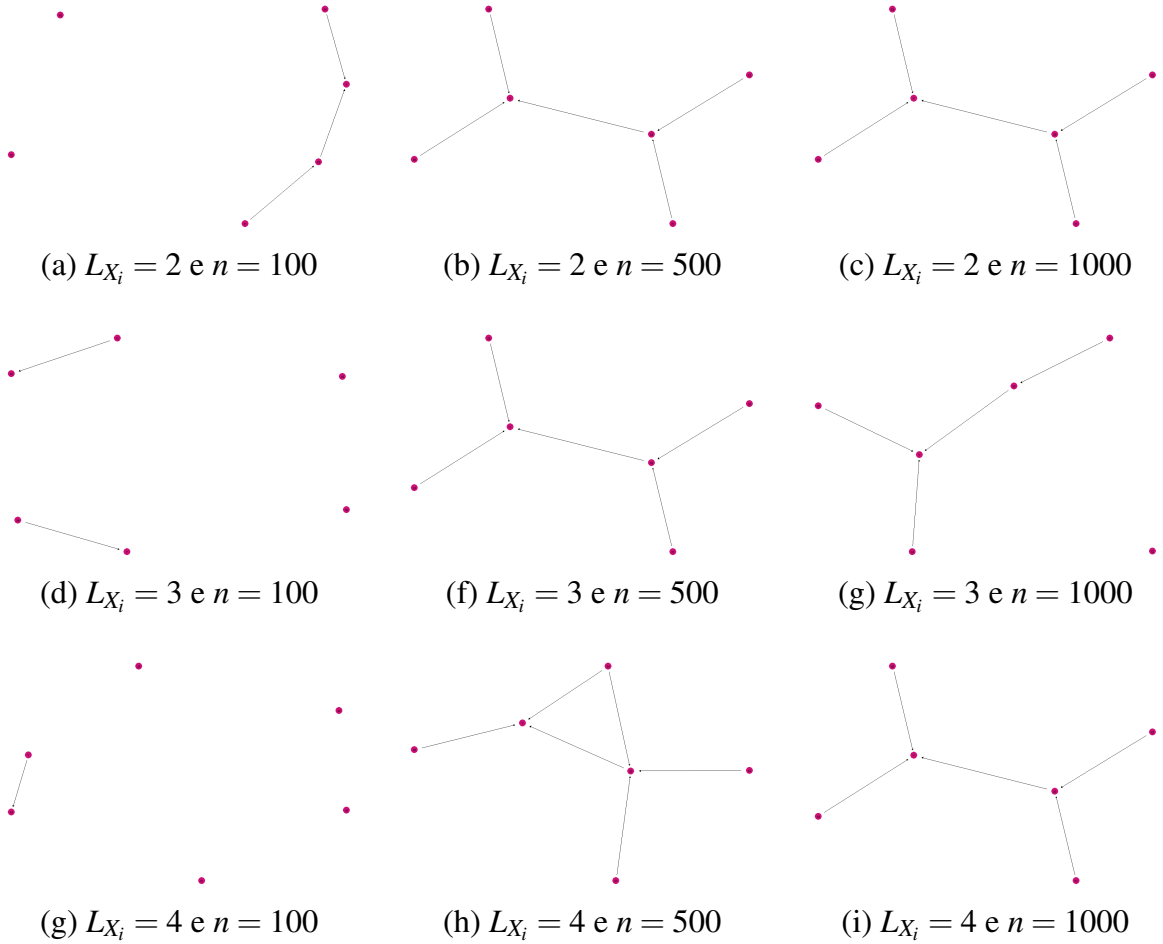
Nas Figuras 27, 28 e 29, nos quais Y é dicotomizado, correspondem respectivamente às metodologias *PC*, *K2* e *K2+PC*. Na Figura 27, os grafos são pouco ou razoavelmente conectados, na Figura 28 observa-se grafos bastante conectados quando comparados a Figura 29. Ainda, na Figura 27 o esqueleto original do grafo é obtido em 4 dos 9 grafos nos itens (b), (c), (f), (i), por outro lado, há acerto do esqueleto em 3 dos 9 grafos da Figura 28, dos itens (c), (f) e (g) e em 4 dos 9 grafos da Figura 29, itens (b), (c), (f), (i).

Ademais, o direcionamento dos arcos também apresenta congruência com o grafo teórico para todos os itens em que a metodologia acerta o esqueleto.

Analogamente, para as Figuras 30, 31 e 32 que correspondem ao número de categorias de Y igual a 3 e respectivamente para as metodologias *PC*, *K2* e *K2+PC*. No primeiro caso, da Figura 27, 5 dos 9 grafos apresentamo mesmo esqueleto do original, nos itens (b), (c), (f), (g) e (h). Na Figura 31, nenhuma das configurações resultou em um esqueleto similar ao da estrutura teórica, por outro lado, na Figura 32, de 9 grafos apresentaram não apenas o esqueleto, mas também o direcionamento dos arcos idêntico ao da estrutura simulada e os valores de suas medidas preditivas estão destacados na tabela de resultados.

Assim como nas medidas preditivas apresentadas na Tabela 11, houve um acréscimo médio de $3,14\%$ quando as metodologias *PC*, *K2* e *K2+PC* são comparadas, Sugerindo maior poder preditivos quando utiliza-se o método híbrido de estimação de estrutura. Essa conclusão é reforçada pela análise gráfica de esqueletos e direcionamento de arestas.

Em termos de tempo computacional, para todas as iterações de amostragem *bootstrap* e ajuste do modelo, obteve-se um mínimo de $0,61$ segundos e máximo de $2,04$ segundos para o ajuste da rede pelo método *K2*. Quando o método *PC* é inserido para o método *K2+PC*, o tempo aumenta em média 57% passando, para um tempo mínimo de $1,05$ segundos e máximo de $2,89$ segundos. Ou seja, a combinação entre ambos os métodos não gera, a menos que empiricamente, um procedimento exaustivo.



(a) $L_{X_i} = 2$ e $n = 100$ (b) $L_{X_i} = 2$ e $n = 500$ (c) $L_{X_i} = 2$ e $n = 1000$

(d) $L_{X_i} = 3$ e $n = 100$ (f) $L_{X_i} = 3$ e $n = 500$ (g) $L_{X_i} = 3$ e $n = 1000$

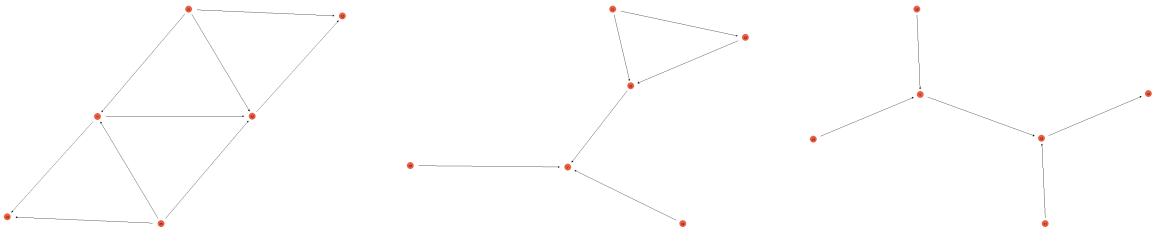
(g) $L_{X_i} = 4$ e $n = 100$ (h) $L_{X_i} = 4$ e $n = 500$ (i) $L_{X_i} = 4$ e $n = 1000$

Figura 27 – Estruturas obtidas por meio do método *PC*, para número de classes da variável resposta igual a 2.

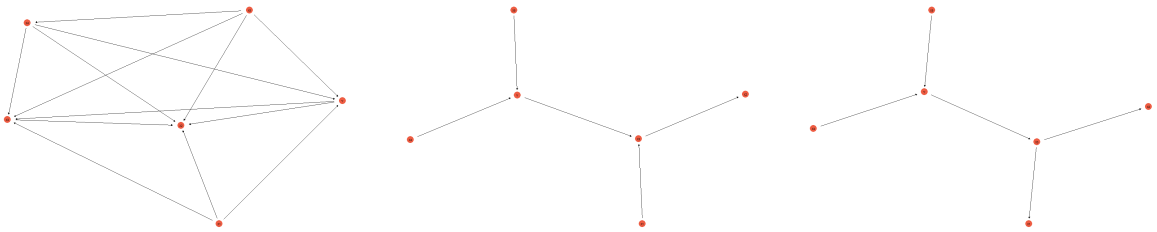
Fonte – Elaborado pela autora.

5.3 Comentários gerais

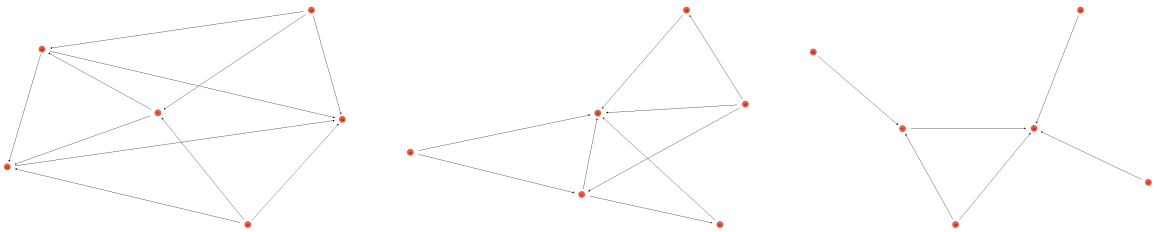
Esse capítulo de estudos de simulação apresentou a investigação de dois componentes das Redes Bayesianas. Na estimação de parâmetros, concluiu-se que, com o estudo conduzido, não há evidências para afirmar que a modificação nos valores dos hiperparâmetro afeta a capacidade preditiva dos modelos. Por outro lado, na estimação de estrutura, o método híbrido proposto pela combinação dos algoritmos *K2* e *PC* mostrou-se útil e eficiente na tarefa de melhorar o modelo final evidenciado pelo ajuste das Redes Bayesianas irrestritas, uma vez que potencializa a capacidade preditiva e de ajuste aos dados.



(a) $L_{X_i} = 2$ e $n = 100$ (b) $L_{X_i} = 2$ e $n = 500$ (c) $L_{X_i} = 2$ e $n = 1000$



(d) $L_{X_i} = 3$ e $n = 100$ (f) $L_{X_i} = 3$ e $n = 500$ (g) $L_{X_i} = 3$ e $n = 1000$



(g) $L_{X_i} = 4$ e $n = 100$ (h) $L_{X_i} = 4$ e $n = 500$ (i) $L_{X_i} = 4$ e $n = 1000$

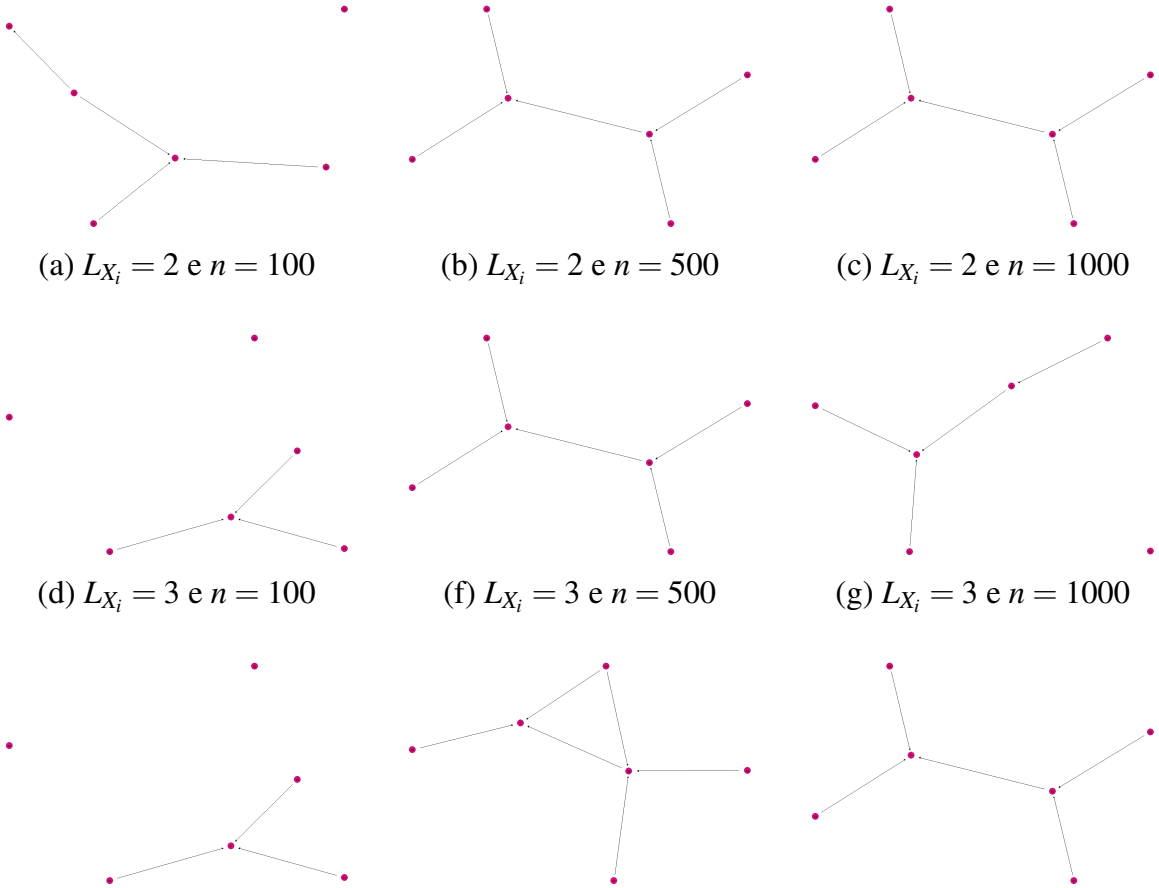
Figura 28 – Estruturas obtidas por meio do método *K2*, para número de classes da variável resposta igual a 2.

Fonte – Elaborado pela autora.

Tabela 11 – Resultados dos ajustes para os dados simulados (valores das medidas preditivas multiplicados por 100), com destaque em negrito para os maiores valores pontuais para cada tamanho amostral e número de classes da variável resposta.

Número de Registros	Classes Y	Classes X	Algoritmo	Arcos	Acurácia	Correlação de Mathews	*Spherical Payoff*
100	2	2	PC	3	70,7 ± 7,0	71,8 ± 6,3	79,5 ± 4,1
			K2	9	61,0 ± 7,1	63,7 ± 7,0	73,2 ± 3,5
			K2+PC	4	79,5 ± 6,8	80,6 ± 6,7	82,6 ± 4,2
		3	PC	2	70,7 ± 6,8	72,7 ± 6,5	79,4 ± 4,2
			K2	13	77,9 ± 6,5	78,9 ± 6,2	82,3 ± 3,8
			K2+PC	3	77,4 ± 5,8	78,4 ± 5,6	82,3 ± 3,8
		4	PC	1	45,2 ± 6,4	-	70,1 ± 0,9
			K2	11	60,0 ± 6,3	62,7 ± 6,6	73,0 ± 3,3
			K2+PC	3	**80,1** ± 6,6	**80,7** ± 6,3	**83,0** ± 5,0
	3	2	PC	4	67,5 ± 7,3	76,5 ± 5,4	72,6 ± 5,0
			K2	10	-	-	-
			K2+PC	4	63,4 ± 8,1	73,4 ± 5,8	71,6 ± 5,1
		3	PC	4	**73,5** ± 8,1	**80,7** ± 5,9	**78,6** ± 5,3
			K2	12	59,9 ± 7,9	70,9 ± 5,8	70,2 ± 5,0
			K2+PC	3	60,1 ± 7,8	71,1 ± 5,8	70,2 ± 5,0
		4	PC	1	56,6 ± 7,0	68,7 ± 5,1	67,3 ± 4,2
			K2	14	48,5 ± 7,9	63,2 ± 5,7	61,7 ± 4,4
			K2+PC	2	48,7 ± 7,9	63,3 ± 5,7	61,7 ± 4,4
500	2	2	PC	5	79,6 ± 2,7	80,1 ± 2,6	84,9 ± 1,8
			K2	6	79,6 ± 2,7	80,1 ± 2,6	84,9 ± 1,8
			K2+PC	5	79,6 ± 2,7	80,1 ± 2,6	84,9 ± 1,8
		3	PC	5	78,9 ± 2,8	79,2 ± 2,8	84,9 ± 1,7
			K2	5	80,1 ± 2,6	80,3 ± 2,6	83,7 ± 1,7
			K2+PC	5	78,9 ± 2,8	79,2 ± 2,8	84,9 ± 1,7
		4	PC	6	**81,0** ± 2,8	**81,1** ± 2,7	**85,6** ± 1,8
			K2	9	78,8 ± 2,3	79,4 ± 2,3	84,6 ± 1,6
			K2+PC	6	**81,0** ± 2,8	**81,1** ± 2,7	**85,6** ± 1,8
	3	2	PC	4	64,2 ± 3,5	73,4 ± 2,6	72,6 ± 1,9
			K2	5	64,2 ± 3,4	73,4 ± 2,5	72,6 ± 1,9
			K2+PC	4	64,2 ± 3,5	73,4 ± 2,6	72,6 ± 1,9
		3	PC	5	65,3 ± 3,7	74,2 ± 2,7	73,7 ± 2,0
			K2	7	60,1 ± 3,3	70,9 ± 2,3	72,2 ± 1,7
			K2+PC	5	65,3 ± 3,7	74,2 ± 2,7	73,7 ± 2,0
		4	PC	5	**66,1** ± 3,0	**74,7** ± 2,2	**74,6** ± 2,0
			K2	10	65,1 ± 3,2	74,1 ± 2,4	73,2 ± 1,7
			K2+PC	5	**66,1** ± 3,0	**74,7** ± 2,2	**74,6** ± 2,0
1000	2	2	PC	5	78,8 ± 1,9	79,1 ± 1,9	84,3 ± 1,1
			K2	5	78,8 ± 1,9	79,1 ± 1,8	84,3 ± 1,1
			K2+PC	5	78,8 ± 1,9	79,1 ± 1,8	84,3 ± 1,1
		3	PC	5	82,7 ± 1,9	82,8 ± 1,8	87,3 ± 1,1
			K2	4	82,7 ± 1,8	82,9 ± 1,8	86,1 ± 1,8
			K2+PC	4	82,7 ± 1,9	82,8 ± 1,8	87,3 ± 1,1
		4	PC	5	**84,9** ± 1,5	**85,0** ± 1,5	**87,7** ± 1,0
			K2	7	83,4 ± 1,7	84,0 ± 1,6	86,6 ± 1,1
			K2+PC	5	**84,9** ± 1,5	**85,0** ± 1,5	**87,8** ± 1,0
	3	2	PC	5	63,2 ± 2,7	73,4 ± 2,0	71,7 ± 1,4
			K2	5	61,9 ± 2,2	71,9 ± 1,8	73,0 ± 1,3
			K2+PC	5	61,9 ± 2,3	71,9 ± 1,8	73,0 ± 1,3
		3	PC	5	62,4 ± 2,3	72,9 ± 1,9	73,6 ± 1,2
			K2	7	69,9 ± 2,1	77,6 ± 1,5	76,4 ± 1,3
			K2+PC	4	**82,5** ± 1,7	**82,7** ± 1,7	**87,3** ± 1,1
		4	PC	6	71,0 ± 2,3	78,3 ± 1,7	77,4 ± 1,3
			K2	7	67,4 ± 2,3	75,9 ± 1,7	74,8 ± 1,2
			K2+PC	6	71,0 ± 2,3	78,3 ± 1,7	77,4 ± 1,3

Fonte – Elaborada pela autora.



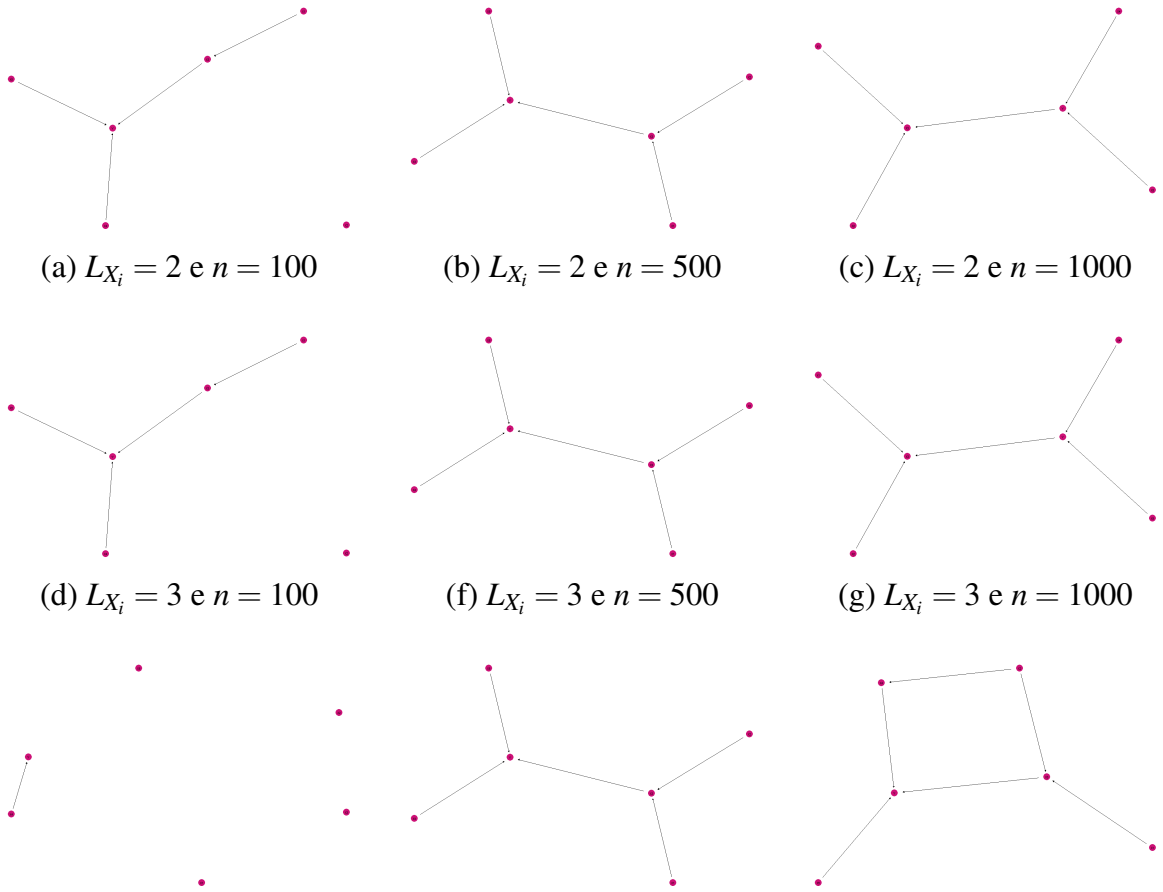
(a) $L_{X_i} = 2$ e $n = 100$ (b) $L_{X_i} = 2$ e $n = 500$ (c) $L_{X_i} = 2$ e $n = 1000$

(d) $L_{X_i} = 3$ e $n = 100$ (f) $L_{X_i} = 3$ e $n = 500$ (g) $L_{X_i} = 3$ e $n = 1000$

(g) $L_{X_i} = 4$ e $n = 100$ (h) $L_{X_i} = 4$ e $n = 500$ (i) $L_{X_i} = 4$ e $n = 1000$

Figura 29 – Estruturas obtidas por meio do método K2+PC, para número de classes da variável resposta igual a 2.

Fonte – Elaborado pela autora.



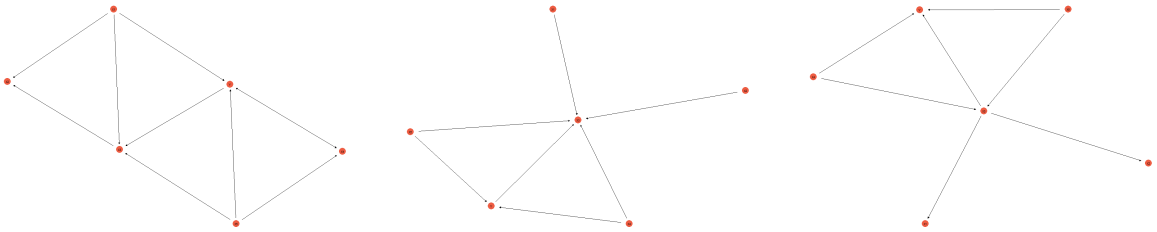
(a) $L_{X_i} = 2$ e $n = 100$ (b) $L_{X_i} = 2$ e $n = 500$ (c) $L_{X_i} = 2$ e $n = 1000$

(d) $L_{X_i} = 3$ e $n = 100$ (f) $L_{X_i} = 3$ e $n = 500$ (g) $L_{X_i} = 3$ e $n = 1000$

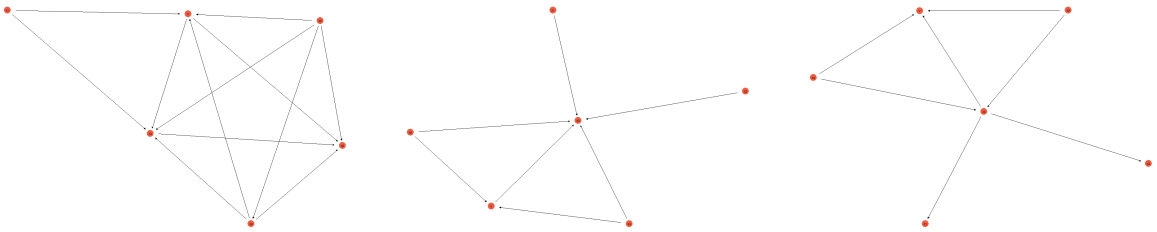
(g) $L_{X_i} = 4$ e $n = 100$ (h) $L_{X_i} = 4$ e $n = 500$ (i) $L_{X_i} = 4$ e $n = 1000$

Figura 30 – Estruturas obtidas por meio do método *PC*, para número de classes da variável resposta igual a 3.

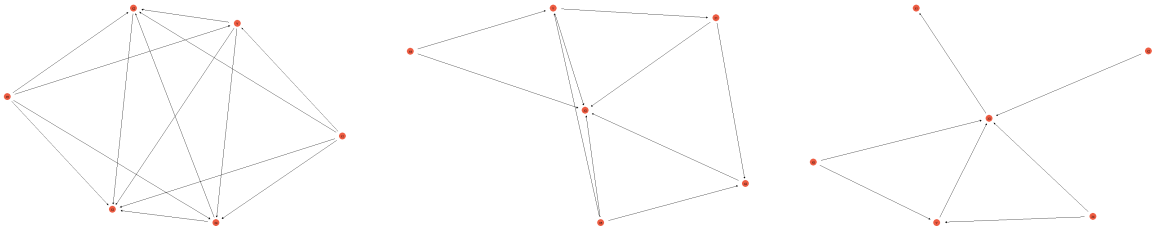
Fonte – Elaborado pela autora.



(a) $L_{X_i} = 2$ e $n = 100$ (b) $L_{X_i} = 2$ e $n = 500$ (c) $L_{X_i} = 2$ e $n = 1000$



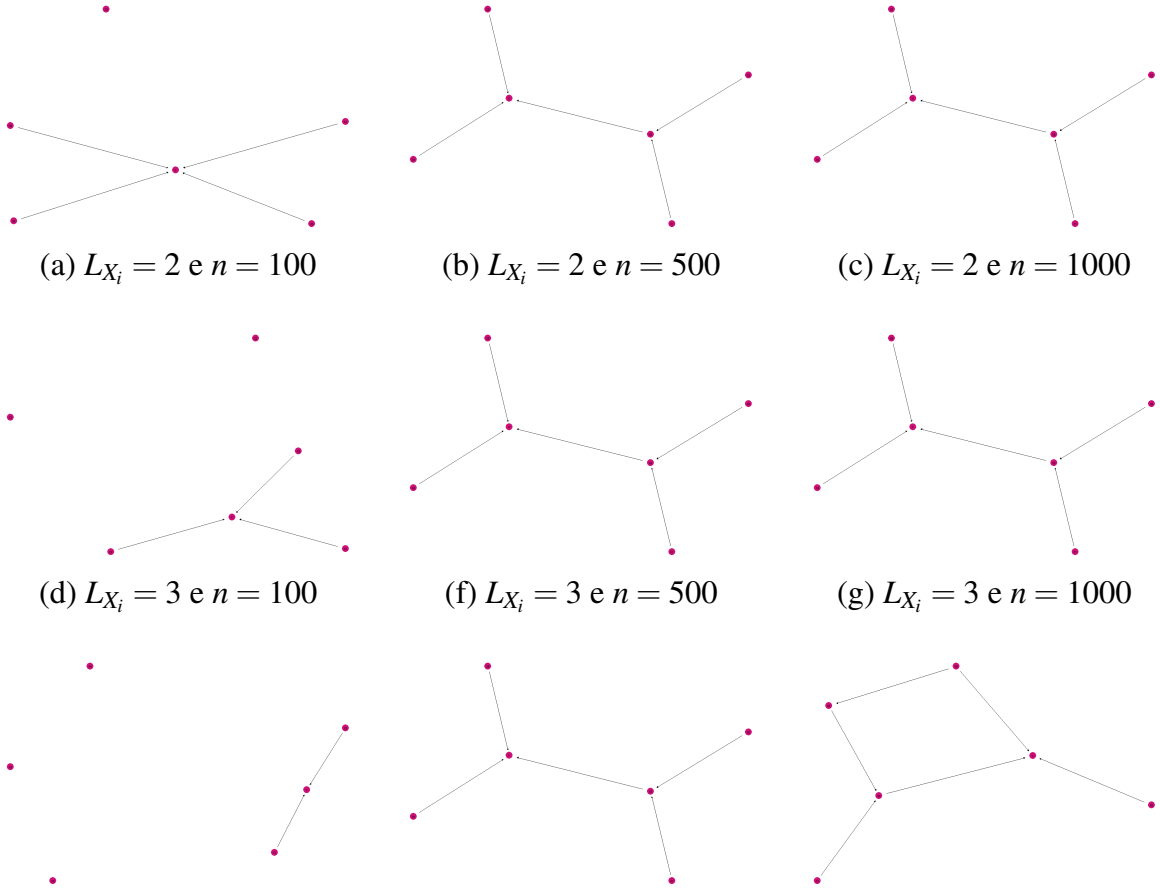
(d) $L_{X_i} = 3$ e $n = 100$ (f) $L_{X_i} = 3$ e $n = 500$ (g) $L_{X_i} = 3$ e $n = 1000$



(g) $L_{X_i} = 4$ e $n = 100$ (h) $L_{X_i} = 4$ e $n = 500$ (i) $L_{X_i} = 4$ e $n = 1000$

Figura 31 – Estruturas obtidas por meio do método K2, para número de classes da variável resposta igual a 3.

Fonte – Elaborado pela autora.



(a) $L_{X_i} = 2$ e $n = 100$ (b) $L_{X_i} = 2$ e $n = 500$ (c) $L_{X_i} = 2$ e $n = 1000$

(d) $L_{X_i} = 3$ e $n = 100$ (f) $L_{X_i} = 3$ e $n = 500$ (g) $L_{X_i} = 3$ e $n = 1000$

(g) $L_{X_i} = 4$ e $n = 100$ (h) $L_{X_i} = 4$ e $n = 500$ (i) $L_{X_i} = 4$ e $n = 1000$

Figura 32 – Estruturas obtidas por meio do método K2+PC, para número de classes da variável resposta igual a 3.

Fonte – Elaborado pela autora.

CAPÍTULO

6

APLICAÇÕES EM BASES DE DADOS REAIS

Este capítulo trata da análise das Redes Bayesianas Discretas para Classificação em contextos de três áreas dististintas de aplicação: análise agronômica, com um problema que envolve a falha em lotes de cana de açúcar no qual é necessário entender o que a provoca para tentar contornar essa dificuldade que traz prejuízos à produtores; uma análise de concessão de crédito de acordo com variáveis a respeito do comportamento de mercado dos indivíduos e, for fim, uma análise esportiva tratando de dados relacionados à jogadores de rúgbi durante uma temporada de jogos, em 2018, com o intuito de entender as relações entre as ações de cada um dos jogadores com sua posse de bola. Tais aplicações são representadas, respectivamente, nas próximas seções.

6.1 Análise de Dados Agronômicos

O conjunto de dados utilizado é composto de variáveis explicativas sobre a variedade da muda (Variedade), o solo, como seu tipo e textura (Ambiente), pH, a quantidade de matéria orgânica presente (Mo), fósforo (P), potássio (K), cálcio (Ca), magnésio (Mg), hidrogênio+alumínio (H_al), alumínio (Al), enxofre (S), soma de bases (Sb), capacidade de troca catiônica (Ctc), corretivo de solo (CORR_T) e saturação básica; sobre o clima como o mínimo, máximo e a média de temperatura (Temp_min, Temp_Med, Temp_max) e umidade do ar na região (UR_min, UR_med, UR_max), a média da radiação solar (Red_Solar_MJm2), velocidade do vento (Vel_vento_ms), quantidade de chuvas (Chuva) evapotranspiração (ETP_mm). Além dessas variáveis, há a quantidade de fertilizantes utilizada, e seu tipo, fungicida (FUNG_L), herbicida (HERB_KG, HERB_L), inseticida(INSET_KG, INSET_L) e maturador (V). A variável de interesse, a ser predita, é a Percentual de Falha (Perc_Falha) a qual se encontra em uma amplitude contínua, tal qual a maioria das anteriores. Ainda, dados a respeito de alguns dos lotes continham informações faltantes os quais foram, portanto, desconsiderados da análise, contabilizando um total de 2385 registros para o desenvolvimento e teste conforme metodologia apresentada a seguir.

Nas Figuras 33 e 34 estão apresentadas as distribuições para as variáveis numéricas contidas no conjunto de dados, sendo que o primeiro histograma se refere à variável a ser predita.

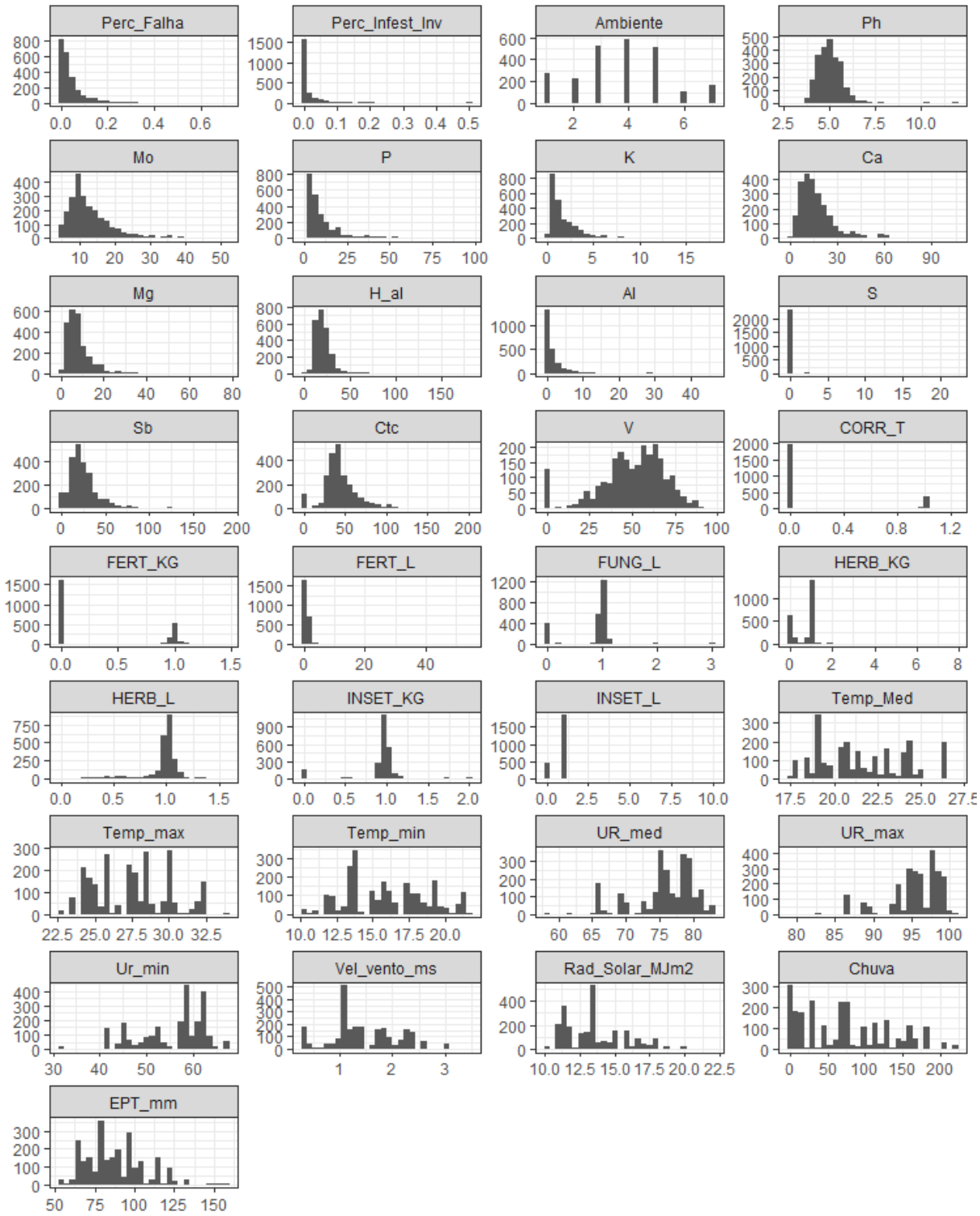


Figura 33 – Histogramas das variáveis numéricas componentes do conjunto de dados agronômicos.

Fonte – Elaborado pela autora.

Para atender à restrição de variáveis categóricas, quando necessário, foi aplicado um método de discretização de variáveis numéricas em classes de mesma frequência. A escolha da

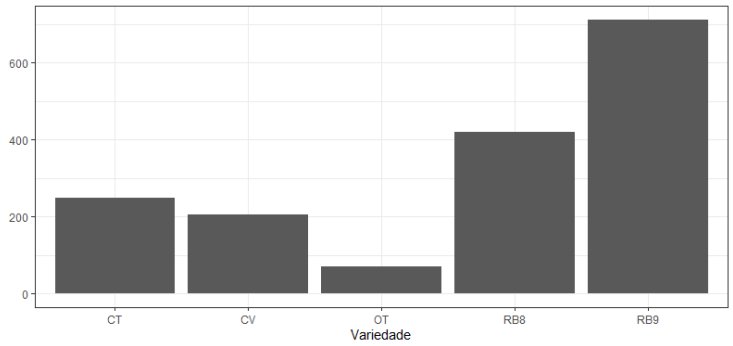


Figura 34 – Gráfico da distribuição da variável Variedade.

Fonte – Elaborado pela autora.

quantidade de classes é um dos fatores de mudança na configuração da rede e portanto, do tempo computacional necessário para cálculo das estimações de estrutura e parâmetros.

As análises das alterações de performance da rede estimada com relação a quantidade de classes escolhidas para as variáveis discretizadas, tanto as explicativas quanto a resposta, devem ser testadas por meio de configurações distintas para avaliação de performance preditiva e explicação, levando-se em consideração a escolha de modelos parcimoniosos.

O problema em torno da base de dados se dá por encontrar as *causas* das falhas em lotes de plantas para fins industriais e comerciais, mais especificamente nos lotes onde a cana-de-açúcar é plantada. As Redes Bayesianas foram escolhidas para tal tarefa por conta da sugestão de causalidade dada pelo sentido dos arcos no grafo acíclico e direcionado (PEARL, 2000). Então, uma vez que a estrutura de rede é estabelecida, se satisfizer certas condições discutidas anteriormente, é possível definir os aspectos que influenciam/causam, direta ou indiretamente, uma característica de interesse, no caso, o *Percentual de Falha* em lotes de áreas plantadas.

Após a discretização das variáveis presentes na base de dados, de modo que o processo e tempo computacional sejam otimizados, é utilizado um método de seleção de variáveis baseado na métrica de Informação Mútua, análoga à apresentada na Equação (2.3.1), a qual quantifica a informação compartilhada entre as duas variáveis X_i e Y. Os valores são calculados para cada uma das covariáveis em relação a variável resposta, de modo que quanto maior for o seu valor, mais relevante é o atributo para a explicação/predição da variável de interesse. Embora que, para o cálculo da estimação de estrutura, não seja necessária a discriminação dessa variável de resposta.

A Tabela 12 contém os resultados das métricas de avaliação das redes, as quais estão apresentadas pela média $\pm$ desvio-padrão - valores multiplicados por 100 - das 100 amostras $70/30$ utilizadas para cálculo dos parâmetros e avaliação da predição, respectivamente. Por meio dessas métricas é possível comparar as seis configurações distintas testadas, as quais são definidas pela combinação da quantidade de classes que as covariáveis e a variável resposta foram discretizadas.

Tabela 12 – Tabela de resultados da Estimação de Rede por quantidade de categorias das variáveis discretizadas (valores das medidas preditivas multiplicados por 100). Em negrito, destacam-se os maiores valores pontuais para cada uma das medidas de performance.

Classes Y	Classes X	Algoritmo	Variáveis	Arcos	Acurácia	Coeficiente Correlacão de Mathews	*Spherical Payoff*
2	2	K2	13	71	-	-	-
		K2+PC	13	27	**70,6** $\pm$ 1,8	**70,7** $\pm$ 1,8	77,1 $\pm$ 1,2
	3	K2	6	30	-	-	-
		K2+PC	6	28	-	-	-
	4	K2	10	41	70,0 $\pm$ 1,7	70,0 $\pm$ 1,7	**77,3** $\pm$ 1,0
		K2+PC	10	13	70,5 $\pm$ 2,0	70,5 $\pm$ 2,0	**77,3** $\pm$ 1,0
3	2	K2	8	26	45,6 $\pm$ 1,9	59,5 $\pm$ 1,4	61,4 $\pm$ 0,9
		K2+PC	8	16	51,0 $\pm$ 1,8	63,4 $\pm$ 13,7	62,6 $\pm$ 1,1
	3	K2	9	36	-	-	-
		K2+PC	9	13	53,2 $\pm$ 1,9	65,0 $\pm$ 1,4	63,2 $\pm$ 1,3
	4	K2	9	32	-	-	-
		K2+PC	9	12	**53,3** $\pm$ 2,0	**65,0** $\pm$ 1,5	**63,8** $\pm$ 1,5

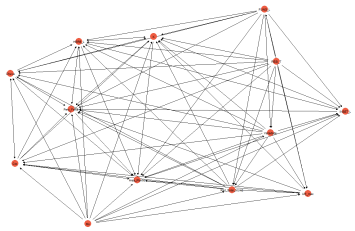
Fonte – Elaborada pela autora.

Nesse resumo de resultados, é possível notar a ausência de algumas medidas, isso se dá por conta da dificuldade do algoritmo para atribuir uma direção a aresta que liga duas variáveis. Esse não direcionamento dos arcos não permite que os parâmetros da Rede Bayesiana sejam calculados, uma vez que, ferem um dos aspectos que definem esse tipo de rede, que é o *DAG* - um grafo acíclico e direcionado.

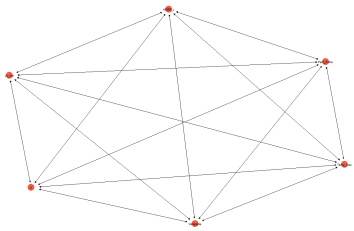
Além disso, quando o algoritmo *PC* é aplicado recebendo como *whitelist*, os pais resultantes da rede obtida pelo método *K2*, o número de arcos cai, em média, em 50%. Com a redução dessas conexões, a explicação das relações entre as variáveis se torna mais viável. Adicionalmente, o acréscimo médio entre as medidas da qualidade preditiva dos métodos é de $3,3\%$, chegando a $5,6\%$ de ganho na acurácia (quando essa comparação é possível); porém, em cerca de 67% dos modelos obtidos por meio da metodologia *K2* não é possível estimar os parâmetros devido ao não direcionamento completo dos arcos, diminuindo para 17% quando combinado ao método *PC* no método híbrido.

Para ilustrar o ajuste gráfico, as Figuras 35 e 36 mostram todas as redes finais obtidas com cada uma das configurações e, respectivamente, correspondem ao método *K2* e *K2+PC*. As colunas das figuras correspondem a quantidade de categorias as quais as covariáveis discretizadas foram submetidas $\{2,3,4\}$, as linhas, por sua vez, a quantidade de categorias as quais a variável resposta foi discretizada $\{2,3\}$.

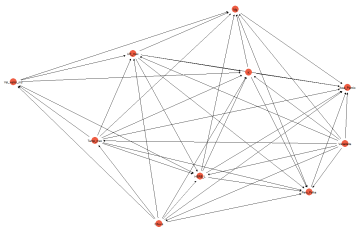
Comparando a conectividade entre metodologias, os grafos obtidos pelo método *K2* independentemente da configuração, apresentam redes bastante conectadas, porém, quando o *PC* é aplicado e seus arcos reduzidos, a viabilidade de interpretação causal ou meramente explicativa das suas conexões está aliada ao aumento ou manutenção de sua capacidade preditiva.



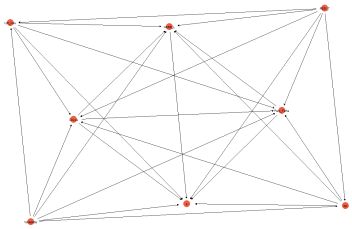
(a) $L_Y = 2$ e $L_{X_i} = 2$



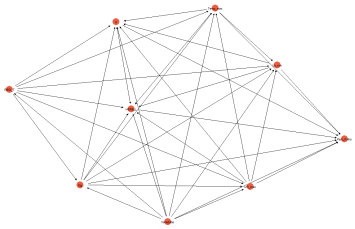
(b) $L_Y = 2$ e $L_{X_i} = 3$



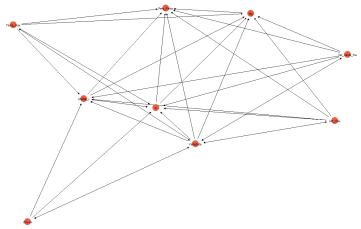
(c) $L_Y = 2$ e $L_{X_i} = 4$



(d) $L_Y = 3$ e $L_{X_i} = 2$

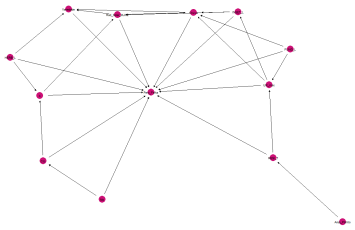


(f) $L_Y = 3$ e $L_{X_i} = 3$

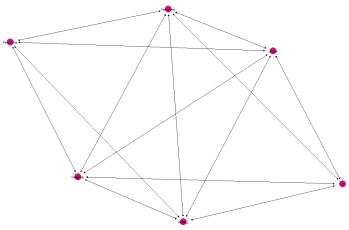


(g) $L_Y = 3$ e $L_{X_i} = 4$

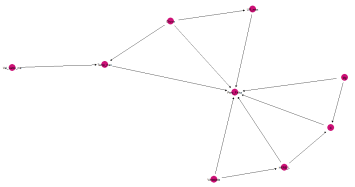
Figura 35 – Estruturas obtidas por meio do método de estimação K2.



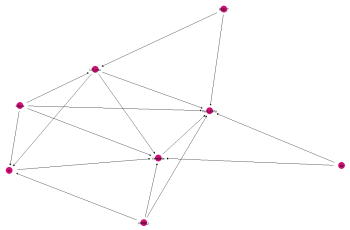
(a) $L_Y = 2$ e $L_{X_i} = 2$



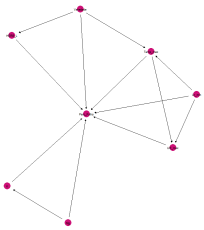
(b) $L_Y = 2$ e $L_{X_i} = 3$



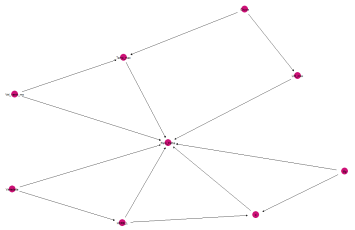
(c) $L_Y = 2$ e $L_{X_i} = 4$



(d) $L_Y = 3$ e $L_{X_i} = 2$



(f) $L_Y = 3$ e $L_{X_i} = 3$



(g) $L_Y = 3$ e $L_{X_i} = 4$

Figura 36 – Estruturas obtidas por meio do método de estimação *K2+PC*.

Fonte – Elaborado pela autora.

Em uma análise de tempo computacional, pelo aumento no número de variáveis e, consequentemente, no número de parâmetros, houve um aumento no tempo investido nas iterações de amostras *bootstrap* para ajuste de um modelo de Redes Bayesianas. Na metodologia *K2* de $2,7$ segundo até $3,7$ minutos, em consequência disso, houve um aumento médio de $63,2\%$, bastante semelhante ao aumento obtido com os dados do estudo de simulação, o aumento absoluto variou de $1,29$ a $48,1$ segundos.

Baseado nas estruturas acima, as redes que apresentaram melhor performance aliada a

menor quantidade de conexões estão apresentadas na Figura 37 e Figura 38, respectivamente para 2 e 3 classes de discretização na variável resposta, já as covariáveis possuem 2 classes no primeiro caso, e 4 classes no segundo caso. Além de que seus arcos completamente direcionados fazem com que os modelos sejam Redes Bayesianas por definição.

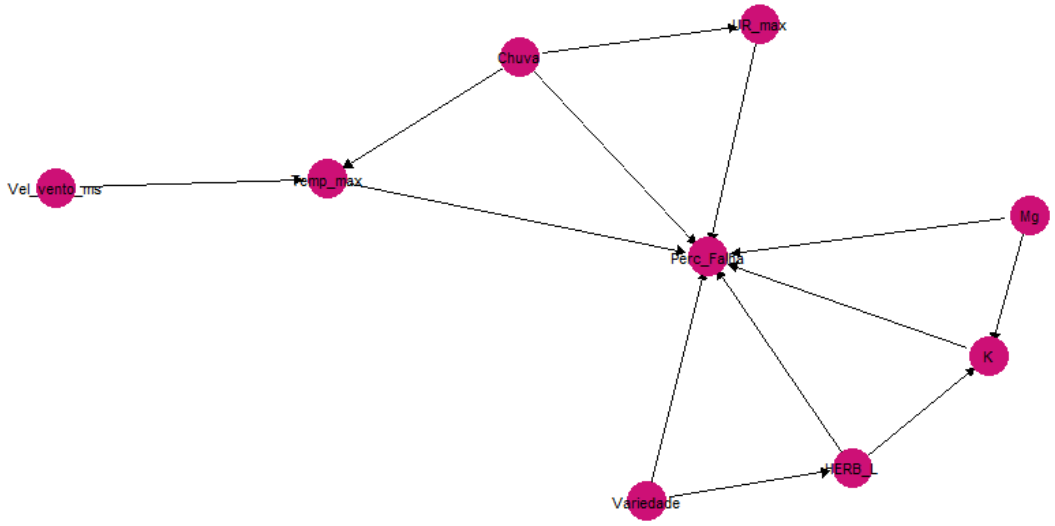


Figura 37 – Estrutura estimada pelo algoritmo híbrido *K2+PC* com maior capacidade preditiva para a variável resposta dicotomizada.

Fonte – Elaborado pela autora.

Contudo, ponderando ganho relativo em relação de uma estrutura aleatória de classificação, a Figura 38, com a resposta sendo discretizada em 3 categorias, obteve melhor desempenho do que a anterior e seu desempenho está destacado na Tabela 12.

As variáveis que estão presentes no grafo, da Figura 38, e foram submetidas ao procedimento de categorização, apresentam os seguintes pontos de corte: *Perc_Falha* $\{1,15;4,17\}$; *K* $\{0,71;1,26;2,50\}$; *Mg* $\{4,08;6,50;9,35\}$; *HERB_L* $\{0,97;1,01;1,04\}$; *Vel_vento_ms* $\{1,08;1,29;1,86\}$; *Temp_max* $\{25,86;27,73;29,89\}$; *UR_max* $\{94,75;95,91;97,75\}$; *Chuva* $\{12,95;68,90;132,4\}$; a variável *Variedade*, já de natureza categórica, recebe os seguintes valores: *CT*, *CV*, *OT*, *RB8* e *RB9*.

Nesta Figura 38, o grafo, que é uma Rede Bayesiana, permite a análise dos aspectos que influenciam o *Percentual de Falhas*, nesse cenário de variáveis. Ele possui sete pais, ou seja, existem sete variáveis que influenciam diretamente a variável resposta. Além de uma que não está ligada ao Percentual de Falhas diretamente, mas que se conecta a alguns de seus pais.

Em termos de causalidade/influência, analisando as conexões, é possível perceber que a velocidade do vento (*Vel_vento_ms*), a temperatura máxima do local, em graus Celsius, (*Temp_max*), a umidade relativa máxima (*UR_max*), a quantidade de Magnésio (*Mg*), e Potássio (*K*)

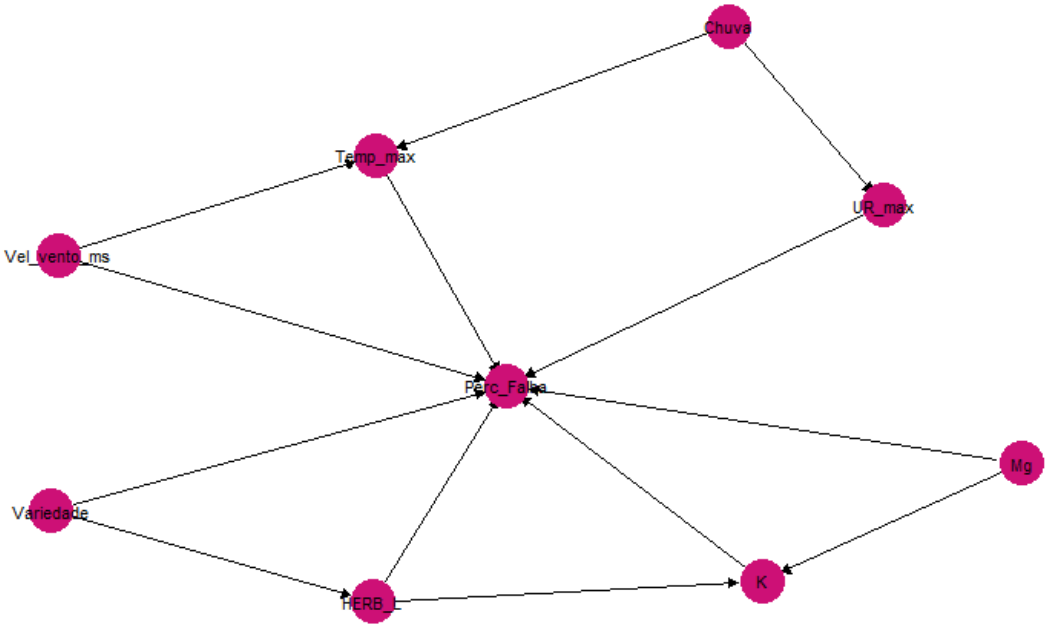


Figura 38 – Estrutura estimada pelo algoritmo híbrido *K2+PC* com maior capacidade preditiva para a variável resposta discretizada em 3 categorias.

Fonte – Elaborado pela autora.

do solo, a quantidade de herbicida, em litros, (*HERB_L*) utilizado e a variedade influenciam diretamente a incerteza da variável resposta, bem como a chuva causa a variação na temperatura máxima e na umidade relativa. Resultado que é bastante pertinente uma vez que as variáveis climáticas estão mais agrupadas umas com as outras, bem como variáveis relativas ao solo formam outro grupo.

6.2 Análise de Risco de Crédito

Outro contexto de aplicação das Redes Bayesianas, e em especial do método de que combina duas abordagens distintas de estimação de estrutura, a análise de crédito voltada à concessão. Nesse caso, os dados serão mantidos em sigilo desde a fonte até os componentes da rede. A base contém registros de meses de coleta de informação para chegar ao prazo da marcação do conceito, então, trata-se de uma base temporal. A temporalidade foi considerada para a separação de desenvolvimento e teste, de modo que os registros mais recentes estejam na segunda amostra, e além disso, para esse caso, a categorização foi realizada de maneira que suas classes tivessem volume e taxa de inadimplência semelhante ao longo do tempo, com a finalidade de garantir a estabilidade do modelo. As variáveis não estão com sua denominação original para fins de sigilo e a base de dados é particular.

Em geral, a base de dados é segmentada de maneira que cada segmento tenha um número suficiente de registros e, principalmente, que o público dentro de cada um deles seja minimamente homogêneo para que as estimações, por meio de qualquer metodologia utilizada, sejam mais

precisas. A divisão é feita de modo que maximize a discriminação e, que faça sentido para o objetivo da análise. Nesse caso, a segregação é feita de forma hierárquica e no primeiro grupo estão os registros que possuem negativações ativas, ou seja, dívidas vencidas, o segundo grupo é composto por quem possui apenas consultas de financeiras portanto, possuem histórico restritivo limpo, serão denotados, respectivamente, por A e B. O conjunto de dados referente ao Segmento A possui 8834 registros, e o Segmento B contém 32149 registros, ambos os conjuntos possuem um balanço de, aproximadamente, $70\%/30\%$ para as amostras de desenvolvimento e teste.

Dessa maneira, os estudos são feitos baseados em dados referentes à públicos-alvo com um perfil de mercado específico e sua marcação de conceito *'bom'* ou *'mau'* determina a natureza binária da variável de interesse que está sendo denotada por Y nesse estudo. Além disso, foram avaliados dois cenários para cada um dos grupos, o primeiro cenário utiliza para o desenvolvimento a amostra com o balanço entre classes da variável resposta original, sendo de $64,24\%$ no primeiro grupo e $11,93\%$ no segundo grupo. As covariáveis são referentes ao comportamento de mercado mediante à busca por crédito e negativações.

Além disso, a seleção de variáveis é feita de maneira criteriosa, uma vez que a quantidade de variáveis disponíveis é imensa, e claro, muitas não atendem às necessidades do modelo. Aliado ao procedimento de seleção de variáveis, a categorização de cada uma delas é realizada de modo que sua taxa de inadimplência seja estável ao longo das safras disponíveis. Essa estabilidade se refere tanto à frequência do perfil quanto à taxa observada, de modo que seja esperado que se mantenha estável em produção.

A seleção de variáveis finais, reduziu a base para 15 covariáveis, foi feita de maneira análoga ao método da informação mútua condicional, já apresentado em seções anteriores. As Figuras 39 e 40 contém os gráficos das distribuições das categorias dentro de cada uma das variáveis selecionadas para ajuste das Redes Bayesianas, respectivamente apresentadas para o Segmento A e Segmento B.

Contudo, a medida utilizada foi o *Information Value*, bastante comum em análise de crédito no setor financeiro, sendo que quantifica o poder preditivo de uma variável independente em relação à binária de interesse (ZENG, 2013). A separação das bases de desenvolvimento e teste foram realizadas de maneira inata à base sendo que a amostra de teste é considerada como a que possui registros mais recentes. Para que a estrutura do modelo ajustado fosse robusta, foi conduzida uma reamostragem *bootstrap* na base de desenvolvimento com 100 diferentes amostras. Para cada uma delas foi ajustado um modelo e a versão final da rede é composta apenas dos arcos, e direções, que compuseram, pelo menos, 50% dos ajustes parciais.

Como resultados, são apresentadas quatro medidas de capacidade preditiva, uma delas ainda não foi apresentada e trata-se da *estatística KS* que é amplamente utilizada na modelagem de risco de crédito para se referir ao poder de discriminação do modelo, ou seja, ela mensura o quão diferentes são os públicos que serão preditos, e é dada da seguinte forma:

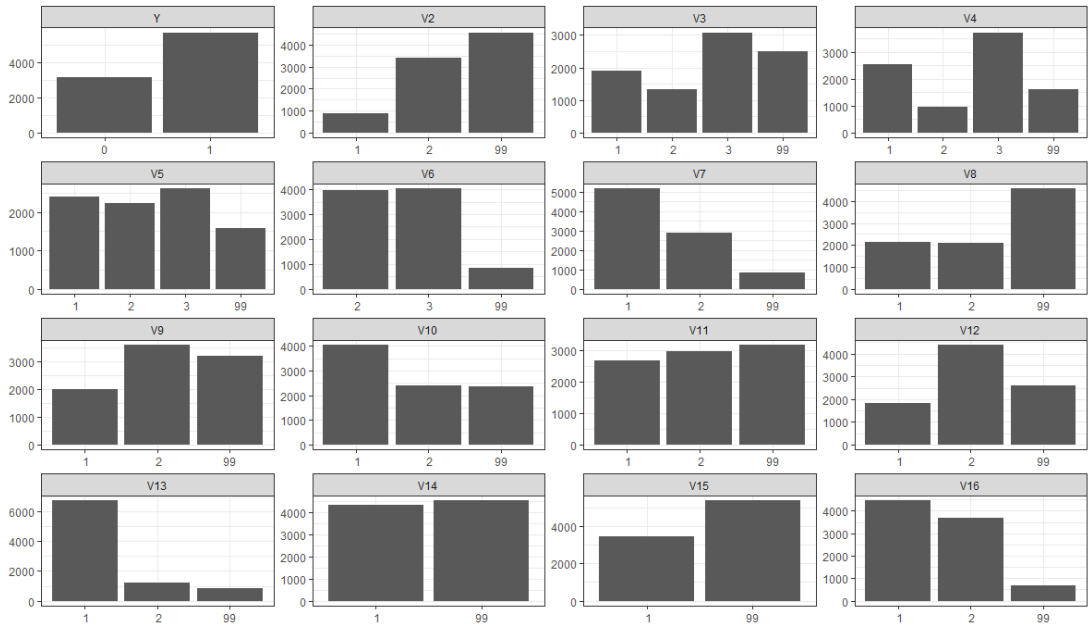


Figura 39 – Distribuições originais das classes das variáveis do Segmento A selecionadas para ajuste das redes.

Fonte – Elaborado pela autora.

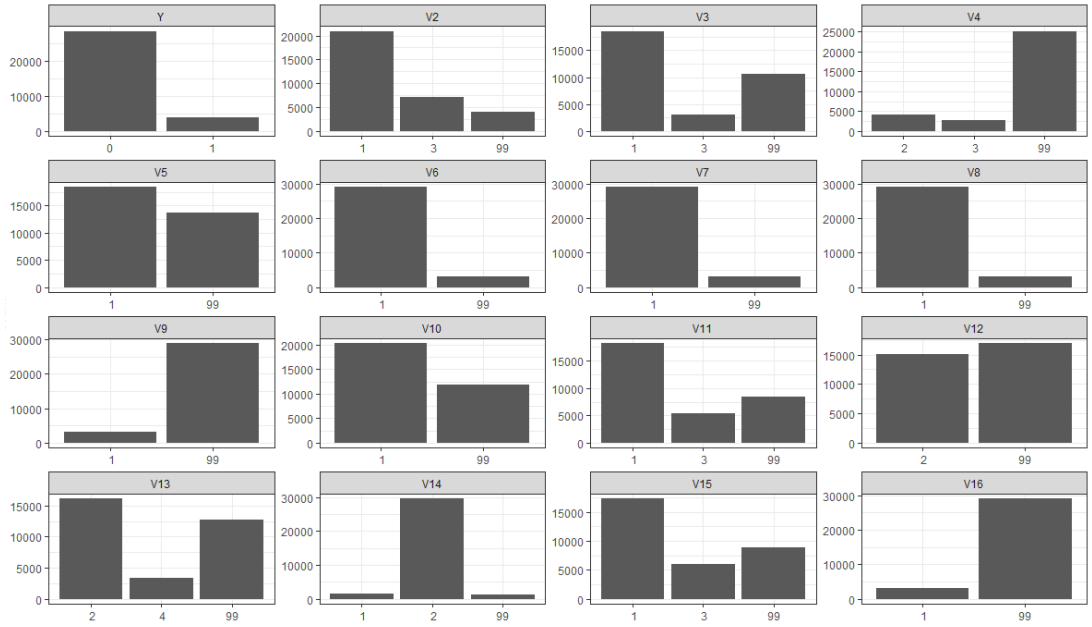


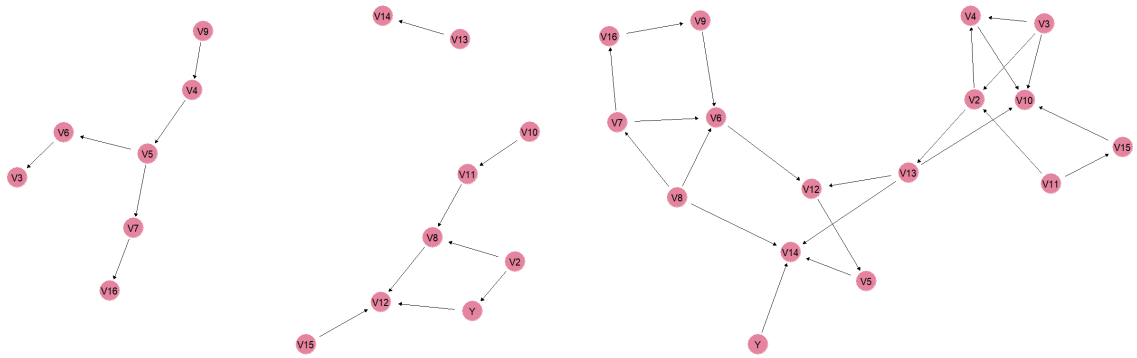
Figura 40 – Distribuições originais das classes das variáveis do Segmento B selecionadas para ajuste das redes.

Fonte – Elaborado pela autora.

$$KS = max|F_{bom}(a) - F_{mau}(a)|,$$

sendo que $F_{bom}(a)$ se refere à distribuição acumulada do público marcado como *bom* e $F_{mau}(a)$ é a distribuição acumulada do público marcado com *mau* (ŘEZÁČ; ŘEZÁČ, 2011).

Ela é interpretada como a maior diferença da distribuição cumulativa das populações de bons e maus. Essa medida é, geralmente, apresentada de maneira percentual variando de 0 a 100% e quanto maior, melhor é a qualidade dos modelos em termos discriminatórios pois as curvas possuem maior distância uma da outra.

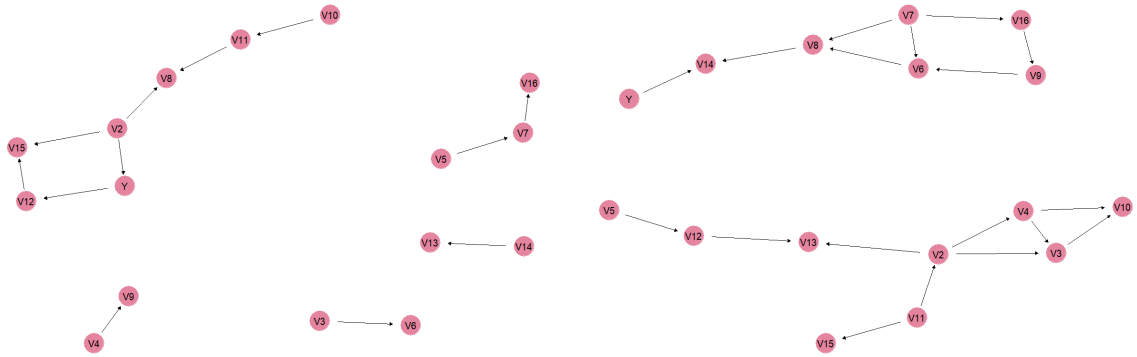


(a) Segmento A. (b) Segmento B.

Figura 41 – Estruturas estimadas pelo método PC, com a amostra de desenvolvimento original.

Fonte – Elaborado pela autora.

A estimação de estrutura pelo método *PC* está apresentado graficamente na Figura 41, na qual observa-se que existe apenas uma variável influenciando a resposta no Segmento A e no Segmento B, Y não possui pais.



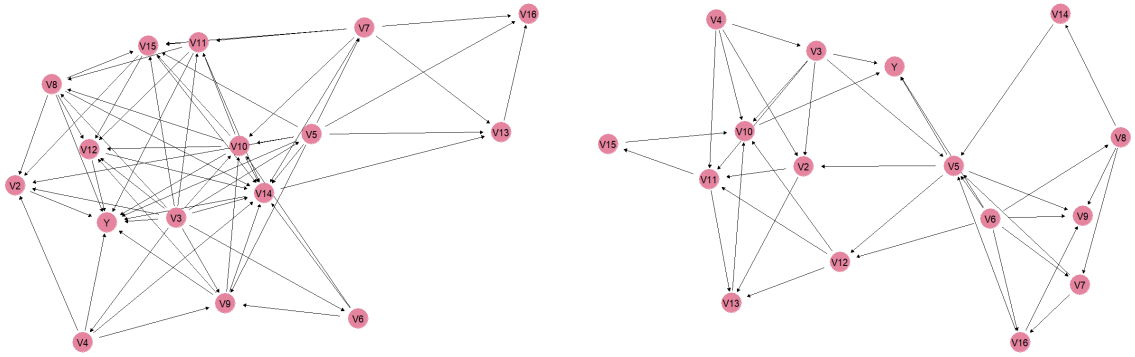
(a) Segmento A. (b) Segmento B.

Figura 42 – Estruturas estimadas pelo método PC, com a amostra de desenvolvimento balanceada.

Fonte – Elaborado pela autora.

Quando a estrutura da rede é aprendida em uma amostra balanceada, apresentada na Figura 42, é possível notar algumas diferenças nas conexões quando comparadas ao desenvolvimento apresentado na figura anterior, contudo, as variáveis adjacentes à resposta - pais e filhos - permanecem as mesmas.

Utilizando uma metodologia de aprendizado alternativa, por meio do algoritmo *Tabu-Search* e maximizando a função *K2* uma estrutura bastante diferente é gerada. Os grafos resultan-

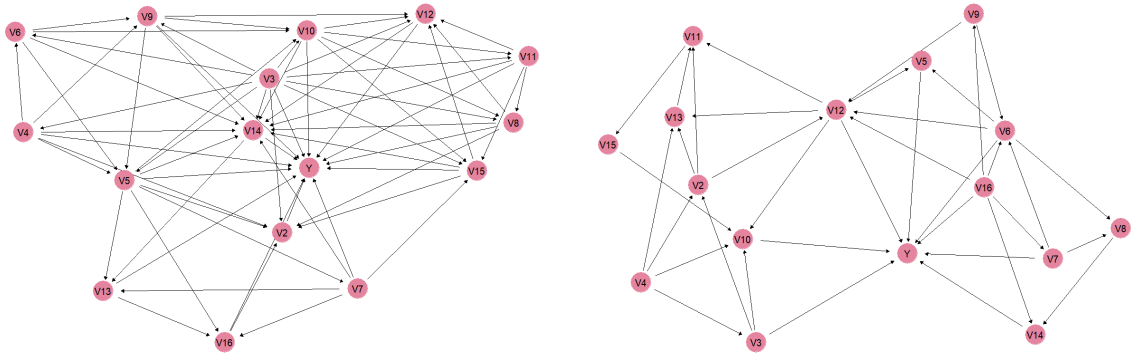


(a) Segmento A. (b) Segmento B.

Figura 43 – Estruturas estimadas pelo método K2, com a amostra de desenvolvimento original.

Fonte – Elaborado pela autora.

tes desse processo estão na Figura A outra metodologia, é a estimação da rede pelo Algoritmo *K2*. O grafo gerado está na Figura 43 e para o Segmento A, a variável resposta possui 10 pais e no Segmento B possui 4 pais.



(a) Segmento A. (b) Segmento B.

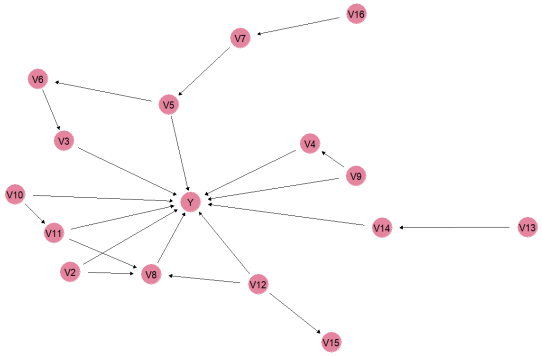
Figura 44 – Estruturas estimadas pelo método K2, com a amostra de desenvolvimento balanceada.

Fonte – Elaborado pela autora.

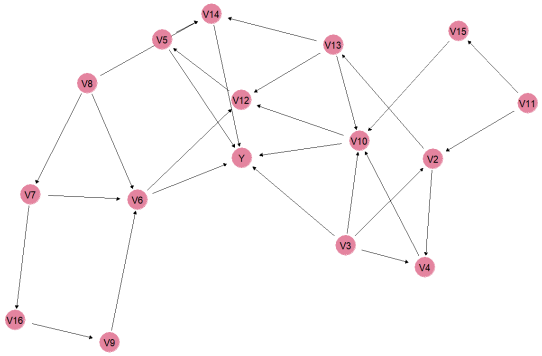
Para o ajuste no caso da amostra balanceada, o número de pais aumenta para 14 no Segmento A e no Segmento B para 8, conforme visualização gráfica na Figura 44.

De acordo com o Capítulo 5, a combinação dos métodos apresentados tem potencial para melhorar a interpretação da desse modelo gráfico, aumentando ou, ao menos, mantendo a capacidade preditiva. Os grafos da Figura 45, equilibrando a quantidade de arcos na rede e permanecendo com a influência dos 9 pais sugeridos pelo *K2*.

A Tabela 13 contém uma sumarização das medidas obtidas por meio do ajuste de cada um dos métodos, *PC*, *K2* e o hídrido *K2+PC* e, de acordo com as informações de medidas de capacidade preditiva e número de arcos, é possível notar que o equilíbrio do número de arcos da



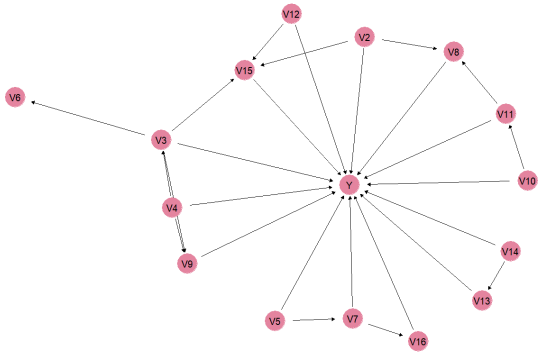
(a) Segmento A.



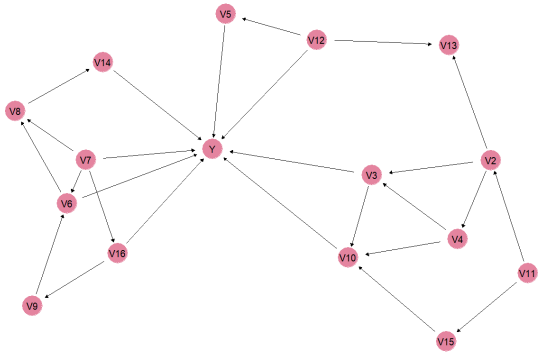
(b) Segmento B.

Figura 45 – Estruturas estimadas pelo método K2+PC, com a amostra de desenvolvimento original.

Fonte – Elaborado pela autora.



(a) Segmento A.



(b) Segmento B.

Figura 46 – Estruturas estimadas pelo método K2+PC, com a amostra de desenvolvimento balanceada.

Fonte – Elaborado pela autora.

metodologia híbrida se apresenta em todos os casos apresentados, nos segmentos A e B e para os casos de desenvolvimento com a amostra original e a balanceada.

No entanto, quando a capacidade preditiva é analisada, no geral, há um equilíbrio entre as medidas de Acurácia, Coeficiente de Correlação de Matthew e *Spherical Payoff*, para o KS alguns valores são bastante discrepantes e não apresentam um padrão de comportamento. Por exemplo, para o desenvolvimento com a amostra original, no segmento A, o modelo híbrido apresentou melhor discriminação, já no segmento B, as medidas ficaram mais similares contudo, o *PC* se sobressaiu. No desenvolvimento com amostra balanceada, para o segmento A, o *PC* se sobressai, e no segmento B, as medidas estão bastante próximas.

Modelos de análise de risco de crédito para concessão, em sua maioria, priorizam terem uma explicação simples e direta, para que o valor da probabilidade atribuída ao mau pagador seja de fácil compreensão aos que estão no meio analítico, mas são consumidores do produto gerado por essas análises. Isso ocorre quando a análise de crédito é feita de maneira generalista ou

Tabela 13 – Tabela de Avaliação Preditiva dos Modelos de Redes Bayesianas (valores das medidas preditivas multiplicados por 100), maiores valores de cada medida destacados em negrito.

Amostra de Desenvolvimento	Segmento	Método	Acurácia	Correlação de Matthew	*Spherical Payoff*	Kolmogorov Smirnov	Número de Arcos
Original	A	PC	**69,70**	**63,06**	**76,96**	14,75	14
		K2	64,01	60,43	73,60	20,96	63
		K2+PC	64,90	61,27	73,60	**22,58**	21
	B	PC	**89,78**	-	90,32	**10,22**	23
		K2	89,64	54,79	**90,34**	09,38	38
		K2+PC	89,48	**55,35**	90,31	08,93	28
Balanceada	A	PC	**62,65**	**62,81**	**74,79**	**26,92**	12
		K2	57,16	56,61	69,75	13,87	67
		K2+PC	57,51	57,23	69,75	15,20	27
	B	PC	50,18	49,58	70,71	-	18
		K2	**64,57**	**55,37**	**72,62**	**16,97**	37
		K2+PC	64,52	55,32	**72,62**	16,82	26

As caselas faltantes refletem a dificuldade da metodologia em classificar as observações entre suas classes.

Fonte – Elaborada pela autora.

para públicos mais específicos que consomem determinado produto ou possuem características mais exclusivas como por exemplo, para compradores de automóveis e, também para análise de concessão de cartão de crédito.

Nos últimos anos, os modelos de Aprendizado de Máquina vem tomando espaço significativo no que diz respeito a análise de crédito e outras frentes do setor financeiro como cobrança, por exemplo. Contudo, conforme foi elucidado anteriormente, a importância da explicação nesse setor, em especial para concessão de crédito, ainda é ponto crucial dessa modalidade de modelagem. Por isso, a utilização das Redes Bayesianas que é um modelo de aprendizado de máquina e que apresenta interpretação e razoável capacidade preditiva, pode trazer uma alternativa eficaz para a área.

6.3 Análise Esportiva

Para a análise esportiva, os dados utilizados são do rúgbi *union*, um esporte de contato que mais cresceu no país nos últimos anos, em especial na variedade do rúgbi 7's (*sevens*) com o time feminino, que vem conquistando campeonatos e visibilidade a nível mundial. O rúgbi XV (quinze), que é uma das modalidades do rúgbi *union*, e é a mais antigas dentre todas. Essa prática ainda não é tão popular quanto o futebol no Brasil e é jogado de forma distinta.

O rúgbi XV recebe esse nome pois o número de jogadores em cada time é 15, e sua

essência é a mesma em todas as variações. O objetivo do jogo é levar a bola até a linha final do campo adversário e colocá-la na área chamada de *in-goal*, marcando 5 pontos com o *try* e tendo o direito de chutar a bola para o *H*, que são as traves no rúgbi, para tentar mais 2 pontos com a *conversão*. Para marcar o *try*, os jogadores devem portar a bola com as mãos, mas só podem passar a bola para companheiros que estão para trás da sua linha de impedimento a linha da bola, pra frente é permitido chutar, mas só pode recepcionar a bola quem estava atrás do pé do chutador quando a bola foi lançada. A maneira usual de tentar reverter a posse de bola é levando seu portador ao solo, com o chamado *tackle*. Esse jogador com a bola possui o direito de disponibilizá-la para seu time, contudo, ambos disputam a posse no que é chamado de *ruck*. No *ruck* participam, pelo menos, um atleta de cada um dos times mas, em geral, costuma-se atribuir três jogadores para essa ação.

É um jogo, basicamente, territorialista e de contato, portanto, podem acontecer algumas infrações e penalidades que são cobradas conforme seu tipo e sua localização no campo. Na modalidade XV, as posições dos jogadores refletem o seu papel tático em campo. De maneira macro, existe o grupo dos *backwards* que possuem características mais voltadas para velocidade e agilidade, já o grupo dos *forwards*, possuem atributos direcionado à força e resistência. Ainda, dentro de cada um dos grupos, existem especialidades ainda mais específicas, por exemplo, posição ocupada nas formações fixas, como em *scrums*, quando se disputa a posse da bola, ou mesmo posicionamento em campo para determinada ação. Uma peculiaridade desse esporte, é que em campo os times não costumam se misturar, estão sempre disputando o território, e portanto, a bola, já que é o que demarca onde o jogo está acontecendo.

Visto isso, obteve-se uma base de dados contendo informação referente a 42 jogadores ao longo de uma temporada de jogos, com características analisadas pela comissão técnica como posse de bola, passe bom, passe ruim, *tackles*, número de chutes, ações e posição de ações, entre outras ações dos jogadores durante os jogos.

Para a realização do estudo foram conduzidos tratamentos no conjunto de dados, afim de minimizar a concentração das variáveis levando a resultados mais robustos. O primeiro tratamento agregou informações de modo que refletissem a proporção de ações de sucesso em relação às ações tomadas, foram elas: lançamento no *line*, chute, posição em formação de ataque e defesa, recepção, conversão, duelo, passe, *tackle* assistente. Posteriormente, esses valores foram relativizados pela quantidade de minutos jogados.

Além dos tratamentos conduzidos, os valores finais foram discretizados em 2 classes, tanto a variáveis resposta definida como a *posse de bola por jogo*, denotada como *bola*, quanto para as covariáveis.

Sua estrutura foi gerada por meio de 100 reamostragens *bootstrap*, considerando a base completa, a rede média foi considerada para os cálculos de estimação dos parâmetros, que foi conduzida por meio do método *hold-out* $70/30$ repetido 100 vezes para garantir boa estimação de suas medidas preditivas.

Tabela 14 – Valores das medidas preditivas relativas à posse de bola por minuto jogado (valores das medidas preditivas multiplicados por 100), maiores valores pontuais das medidas destacados em negrito.

Algoritmo	**Acurácia**	**Coeficiente de Correlação de Matthew**	**Spherical Payoff**
PC	39,5 $\pm$ 4,4	-	69,2 $\pm$ 0,8
K2	40,5 $\pm$ 6,9	-	68,4 $\pm$ 1,2
K2+PC	**63,1** $\pm$ 5,1	**59,7** $\pm$ 6,2	**72,7** $\pm$ 2,0

Fonte – Elaborada pela autora.

A Tabela 14 apresenta os resultados referentes às medidas preditivas para os modelos de classificadores via algoritmo *K2* e via *K2+PC*. Os valores indicam comportamentos bastante semelhantes em termos de performance. Da mesma forma, o grafo resultante da estimação via algoritmo *K2* está apresentado na Figura 47. Com exceção do atributo *try*, todas as variáveis possuem ao menos uma conexão entre si. A vantagem dessa rede conectada é a de refletir, como um todo, qual é a dinâmica das relações entre as ações em campo.

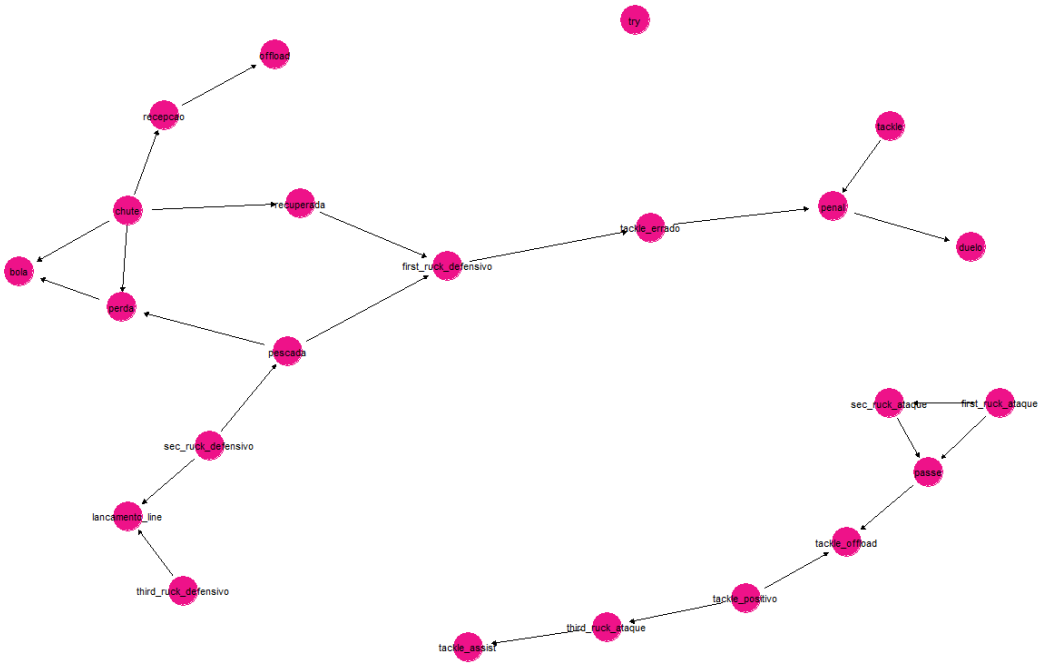


Figura 47 – Estrutura estimada pelo método *K2*.

Fonte – Elaborado pela autora.

Contudo, quando a predição, aliada a explicação da conectividade da rede são os objetivos da modelagem, um número equilibrado de arcos simplifica a interpretação de seus resultados, sem perder esse poder preditivo. Na Figura 48 a representação da estrutura da rede para o método hibrido, considerando os pais estimados em *K2* em uma *whitelist* da variável resposta.

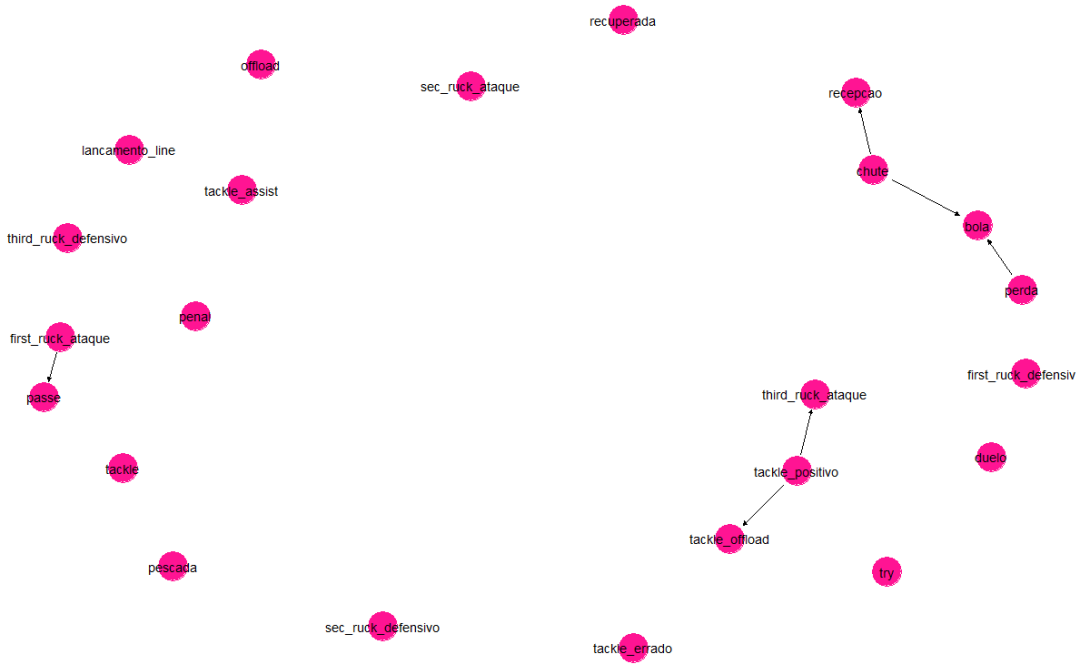


Figura 48 – Estrutura estimada pelo método *K2+PC*.

Fonte – Elaborado pela autora.

A estrutura final do classificador *K2+PC* foi reduzida para 3 além da resposta, conforme a Figura 49. Além disso, reflete reduzida complexidade ao modelo, com valores das medidas preditivas semelhantes ao modelo estimado pelo algoritmo *K2*. As variáveis que compõem essa rede são: *bola*, com ponto de corte de $8,75$ minutos de posse por jogo, pouco mais de 10%; *recepção*, indicadora de, pelo menos, uma recepção por jogo; *chute*, indicadora de, pelo menos, um chute por jogo; *perda*, aproximadamente 0.9 perdas por jogo.

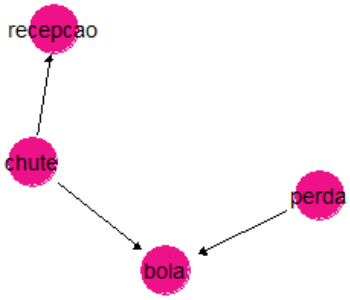


Figura 49 – Estrutura reduzida estimada pelo método *K2+PC*.

Fonte – Elaborado pela autora.

6.4 Comentários gerais

Este capítulo apresentou as análises dos procedimentos de estimação de estrutura para bases de dados reais para duas origens distintas. A primeira aplicação foi em dados agronômicos considerando diferentes configurações de discretização dos dados. A segunda base do setor financeiro, mais especificamente, da análise de crédito, com quantidade de categorias já preexistentes garantindo sua estabilidade temporal, além disso, considerando a estimação baseadas em amostras com e sem balanceamento. E a terceira referente à jogadores de rúgbi durante uma temporada de jogos analisando a posse de bola por jogo e quais as outras ações que podem influenciar esse valor. A metodologia proposta teve uma boa performance geral, em especial no primeiro e no terceiro estudo nos quais o *K2+PC* (*score and restrict*) se sobressaiu em relação aos outros métodos e de maneira bastante expressiva no último estudo que contém um número de observações bastante reduzido. Em especial, para todos os casos, essa metodologia equilibra o número de arcos quando é comparada com os ajustes dos métodos individualmente.

CAPÍTULO

7

COMBINAÇÃO VIA *STACKING*

Os estudos conduzidos e apresentados nos capítulos anteriores trataram da teoria da metodologia de Redes Bayesianas, bem como os casos particulares de classificadores, e uma série de estudos de aplicação das técnicas apresentadas em dados simulados e aplicações em bases reais.

A capacidade preditiva de um método de modelagem está atrelada à qualidade com a qual a generalização capta os padrões dos dados para prever outras observações. É descrita como uma minimização do risco preditivo de funções dos dados, que no caso de classificação, é a rotulagem correta das classes da variável de interesse.

Em James *et al.* (2013) os autores descrevem o procedimento de estimação como sendo a minimização do risco de predição que é a diferença entre o valor observado e o valor predito, sendo composta em viés e variância. Analogamente para classificação, de acordo com Izbicki e Santos (2020), a função de risco mais adequada, e mais intuitiva, é a de *perda 0-1*, a qual descreve que $L(h,(\mathbf{X},C)) = \mathbb{I}(C \neq h(\mathbf{X}))$, para uma função h qualquer das variáveis preditivas $\mathbf{X}$. Então, o risco, R, associado à sua predição é dado por (IZBICKI; SANTOS, 2020):

$$R(g) := \mathbb{E}(\mathbb{I}(C \neq h(\mathbf{X}))) = P(C \neq h(\mathbf{X})).$$

Portanto, a estimação de $h(\mathbf{X})$ é definida como sendo o argumento que maximiza $P(C = c|\mathbf{X})$, tal como utilizada por todos os métodos apresentados nesta dissertação.

A minimização da função de risco para tarefas preditivas, pode ainda contar com um recurso extra bastante difundido em procedimentos de estimação que é a combinação de modelos, de baixo poder preditivo, com o intuito de maximizar sua acurácia. Das abordagens clássicas de combinação de modelos, chamadas na literatura de modelos *ensemble*, as mais conhecidas e utilizadas são o *bagging* e o *boosting*, definidos a seguir.

Bagging

O procedimento de combinação chamado *bagging*, ou *bootstrap aggregating* foi proposto em Breiman (1996). Esse método, como o nome sugere, é caracterizado pelo ajuste de múltiplos modelos preditores distintos para a mesma variável de interesse com o mesmo tamanho da amostra de dados de desenvolvimento, contudo, são selecionados registros com reposição, ou seja, utilizando a técnica de *bootstrap* são construídas bases distintas para produção de modelos simples que serão agregados por média ou por voto, se tratando de regressão ou classificação respectivamente, para o retorno final da predição (RIBEIRO; COELHO, 2020).

Essa metodologia é bastante utilizada para redução da variabilidade, e instabilidade, do estimador preditivo e já foi proposta para utilização com Redes Bayesianas em Louzada e Ara (2012).

Boosting

O *boosting* também é um método de combinação de sub-modelos, que, em teoria, possuem baixo poder preditivo. Porém, ao contrário do *bagging* os sub-modelos são ajustados de maneira sequencial, a cada ajuste seleciona as observações mais difíceis de serem generalizada (as que obtiveram uma classe de predição diferente da classe observada) e as ajusta novamente em um novo modelo (FREUND; SCHAPIRE; ABE, 1999). Essa metodologia redefine os pesos dos modelos conforme são construídos (FRIEDMAN *et al.*, 2000).

De acordo com Ribeiro e Coelho (2020), suas versões mais utilizadas são o *AdaBoost*, de *Adapatative Boosting*, (FREUND; SCHAPIRE; ABE, 1999) e o *gradient Boosting* (FRIEDMAN *et al.*, 2000), elas se diferem pela forma que agregam os sub-modelos construindo um modelo final robusto. Além disso, uma proposta de utilização em Redes Bayesianas pode ser encontrada em Jing, Pavlović e Rehg (2008).

Apesar da grande relevância dos métodos apresentados anteriormente, e por já terem sido aplicados em classificadores bayesianos como em Jing, Pavlović e Rehg (2008) e Elidan (2011), o texto e o estudo conduzido tratarão do método *stacking* que será brevemente descrito na próxima Seção.

7.1 Método *Stacking*

A metodologia, nessa abordagem de combinação de modelos, também carrega a ideia focal de ajuste de sub-modelos baseados na amostra de desenvolvimento e posterior construção de um modelo final que agrega esses preditores. Contudo, segundo Ribeiro e Coelho (2020), o *stacking* possui dois pontos cruciais de distinção em relação aos dois métodos apresentados anteriormente, são eles:

- O primeiro é que no *stacking* a classe de sub-modelos pode ser heterogênea, ou seja,

permite a utilização de várias metodologias diferentes de predição, por exemplo, pode agregar modelos de regressão logística, *Naïve Bayes* e análise de discriminante linear, todos em um único modelo empilhado;

- A segunda é que a maneira de combinação dos sub-modelos também é realizada de maneira probabilística, ou seja, um outro modelo pode ser utilizado para empilhar os modelos preditivos, no caso de classificação, *meta-classificador*, chamado também de modelo de segundo nível. Ao contrário dos outros métodos que agregam os submodelos de maneira determinística por meio de média ou moda, por exemplo, o *stacking* utiliza as predições de cada um dos submodelos como covariável para o modelo de segundo nível, portanto, pode também ser chamado de *blending*.

O esquema da Figura 50 exemplifica as etapas de construção de um modelo combinado via *stacking*. A partir da base de desenvolvimento, são ajustados, no caso, os 5 sub-modelos de diferentes metodologias para a construção um modelo de segundo nível, um meta-classificador que também pode ter a mesma metodologia de algum modelo de primeiro nível, ou qualquer outra que permita combinar as predições retornadas pelos primeiros.

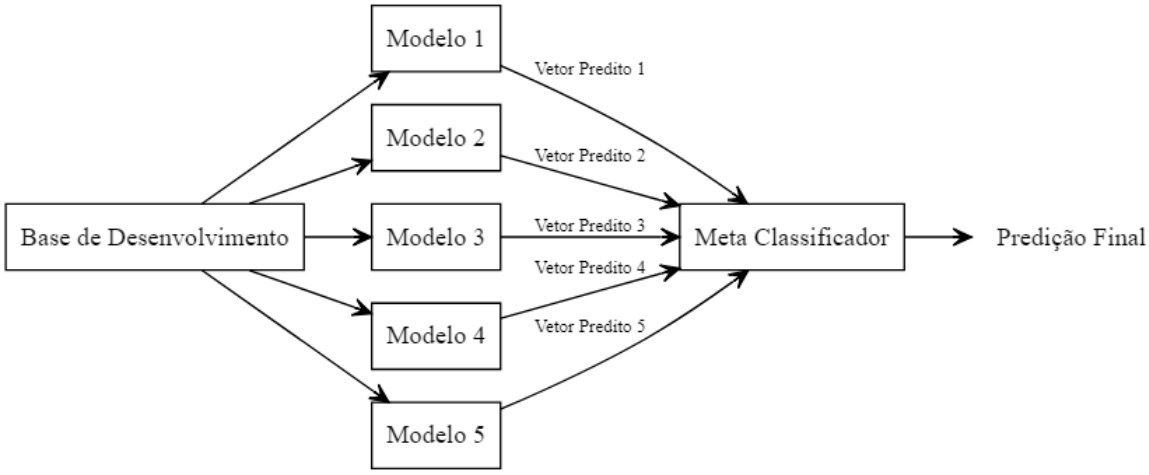


Figura 50 – Esquema do processo de ajuste do método *stacking*.

Fonte – Elaborado pela autora.

Todos os modelos ajustados tanto no primeiro, quanto no segundo nível, possuem a mesma variável de interesse como exemplo de rotulação das classes e, portanto, a predição final retorna a classe mais provável segundo o meta-classificador que é construído com predições dos sub-modelos.

7.2 *Stacking* com Classificadores de Redes Bayesianas

O procedimento utilizado para a confecção do modelo combinado com essa família heterogênea de classificadores gerando um meta-classificador (obtido por meio de um modelo de segundo nível escolhido para fazer a predição final). O meta-modelo possui como covariáveis o vetor predito por de cada um dos modelos de primeiro nível que, por fim, prediz a variável de interesse baseada nas predições individuais dos sub-modelos.

Em busca de melhorar a performance preditiva dos classificadores de Redes Bayesianas combinou-se as metodologias apresentados no Capítulo 3 e Capítulo 5 por meio da abordagem de *stacking* e, para agregar os modelos de classificadores de Redes Bayesianas, essas metodologias serão testadas como modelo de segundo nível, fazendo as predições finais. Então, os modelos utilizados tanto para primeiro como para segundo nível foram: *Naïve Bayes* (NB), *Tree-Augmented Naïve Bayes* (TAN), k-*Dependence Bayesian Network* (kDB), *Bayesian Network Augmented-Naïve Bayes* (BAN), *Avereged One-Dependence Estimator* (AODE) e, adicionando a estimação irrestrita de rede representada por um modelo baseado em restrição, o *PC*, um baseado em métricas, o *K2*, e o último híbrido, o *K2+PC* apresentado no Capítulo 5.

O estudo foi conduzido mediante bases de dados reais que foram utilizadas nas aplicações do Capítulo anterior, a base agronômica e a de risco de crédito. Para ambos os casos estudados, a variável resposta possui apenas duas categorias, na primeira base as covariáveis também são binárias mas na segunda aplicação, podem variam entre 2 e 4 classes.

Os resultados apresentados são dos diferentes modelos de segundo nível utilizados como meta-classificadores para a predição da variável resposta.

Base de Agronômica

Essa base de dados utilizada possui 15 variáveis que foram discretizadas de forma que, tanto as covariáveis quanto a variável explicativa, sejam balanceadas. A Figura 51 apresenta a comparação das metodologias que combinaram as mesmas predições dos sub-modelos avaliados.

Nota-se que colunas de coloração mais escura indicam uma performance menos expressiva daquele método de combinação comparado com os demais modelos utilizados para o mesmo fim. Do mesmo modo que colunas mais claras indicam que os modelos não foram superados com frequência. Nesse gráfico, a coluna que possui menores valores é a do *stacking* via *K2+PC*, com uma proporção média de ganho de 67%, seguida da utilização do *K2*, que ganha, em média, em 66% dos casos, e depois do *AODE* empatado com o *BAN*, com a proporção de 60% em ganho. E com pior performance global do *stacking* é via método *Naïve Bayes* que foi superado em cerca de 87% das amostras.

O resultado é congruente quando a comparação é feita pelo gráfico da Figura 52 porém, de maneira menos expressiva. O *stacking* via *Naïve Bayes* apresenta maior concentração em

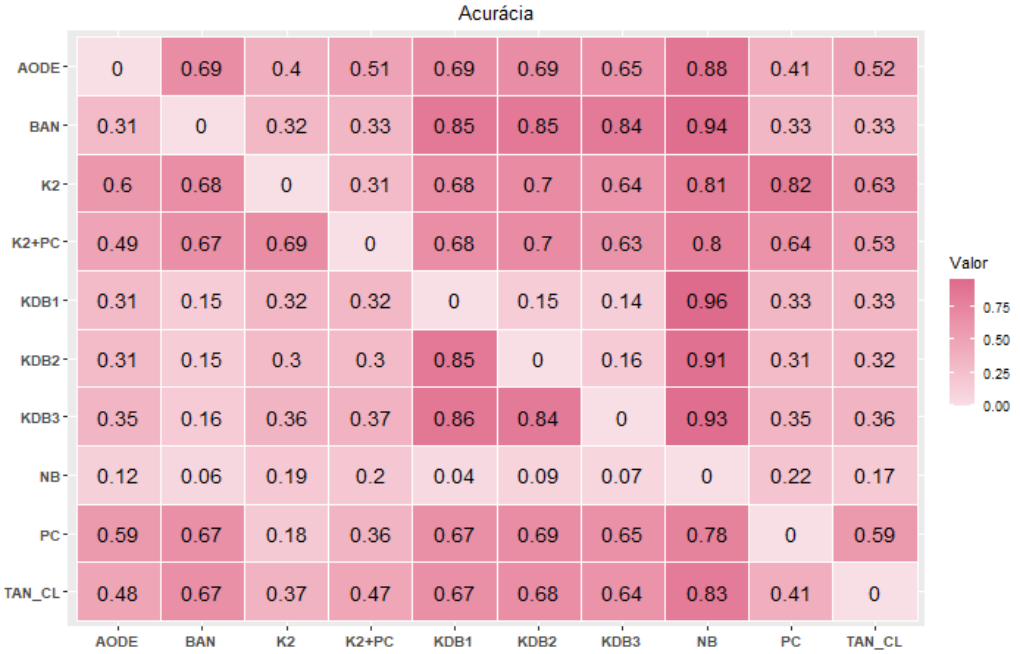


Figura 51 – Mapa de Calor da proporção de vezes que as combinações via métodos do eixo Y apresentaram melhores valores de acurácia comparados aos métodos do eixo X.

Fonte – Elaborado pela autora.

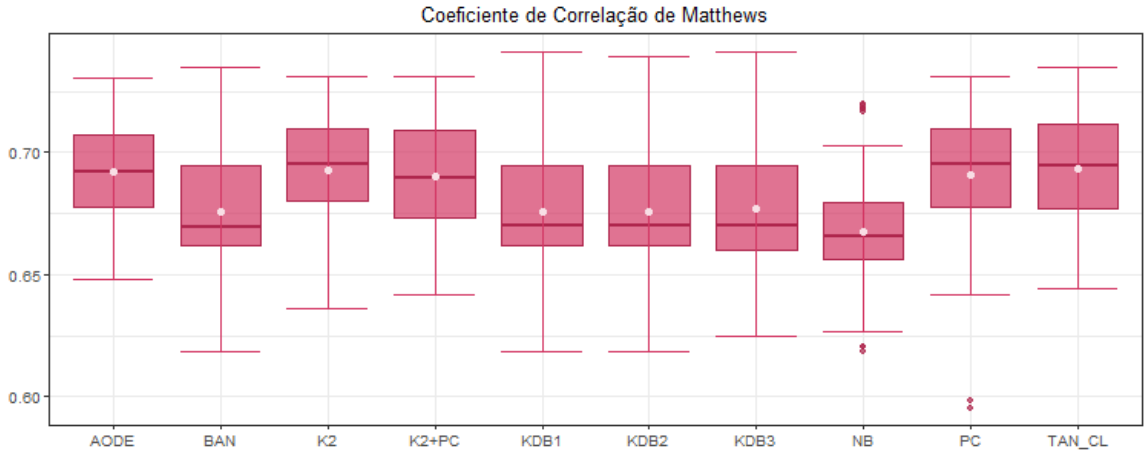


Figura 52 – Boxplot do Coeficiente de Correlação de Matthews para as medidas de cada uma dos métodos utilizados como meta-classificadores.

Fonte – Elaborado pela autora.

valores mais baixos mas não há uma clara distinção entre as demais combinações. Nota-se, contudo, que a combinação por *stacking* via *AODE*, *K2*, *K2+PC*, *PC* e *TAN* apresentam um deslocamento superior da caixa do gráfico, em relação aos demais.

Base de crédito

Para essa base, foram ajustados os seguintes sub-modelos para a utilização nos meta-classificadores *stacking*. Na Tabela 15 a performance dos modelos individualmente. A base de

ajuste desses modelos, possui um desbalanceamento original dos dados

Tabela 15 – Sub-modelos ajustados na base original.

Medidas	NB	TAN	AODE	kDB1	kDB2	kDB3	BAN	PC	K2	K2+PC
ACC	65,6 ± 0,3	69,7 ± 0,6	69,3 ± 0,2	65,6 ± 0,3	65,6 ± 0,3	65,6 ± 0,3	65,6 ± 0,3	70,3 ± 0,9	65,7 ± 1,6	66,6 ± 3,2
MCC	63,9 ± 0,1	65 ± 0,7	65,2 ± 0,2	63,9 ± 0,1	63,9 ± 0,1	63,9 ± 0,1	63,9 ± 0,1	64,2 ± 0,9	61,3 ± 1,2	62,1 ± 2,4
SP	68,3 ± 0,3	68,3 ± 0,3	77,2 ± 0,2	68,3 ± 0,3	68,3 ± 0,3	68,3 ± 0,3	68,3 ± 0,3	74,9 ± 1,8	74,7 ± 2,6	74,6 ± 2,5

Fonte – Elaborada pela autora.

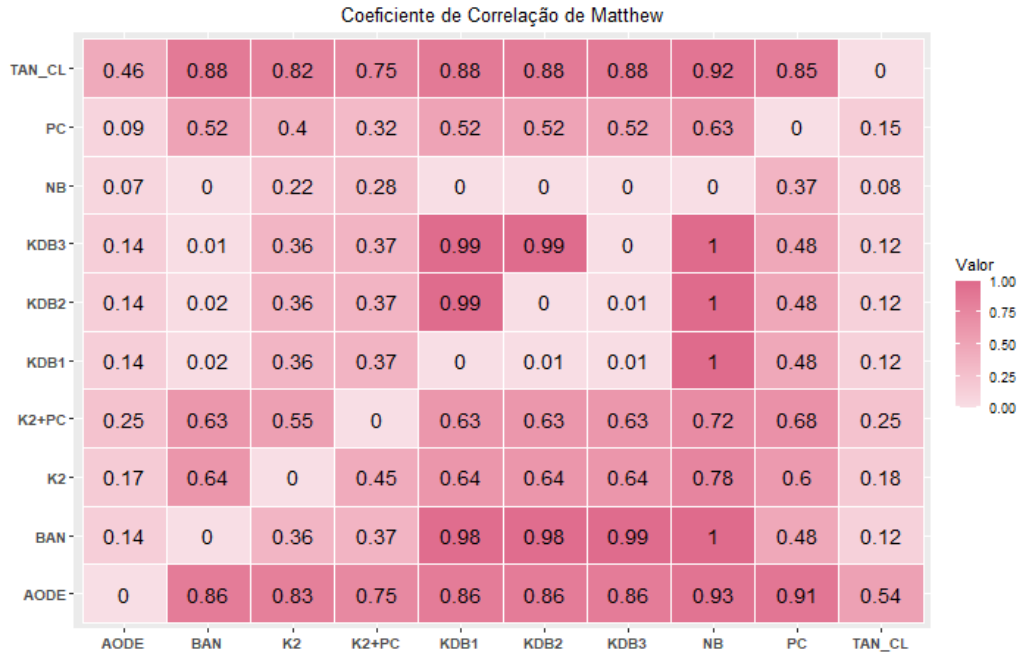


Figura 53 – Mapa de Calor da proporção de vezes que as combinações via métodos do eixo Y apresentaram melhores valores de MCC comparados aos métodos do eixo X.

Fonte – Elaborado pela autora.

Na Figura 53 a mesma análise de comparação entre modelos é realizada utilizando uma medida que reflete o desbalanceamento da variável resposta. Os classificadores *TAN* e *AODE* utilizados para a combinação apresentam o melhor resultado nesse caso, com proporções de ganho de 86% e 82%, respectivamente, graficamente nota-se pelas colunas claras ou pelas linhas escuras. Com performance um pouco inferior estão as combinações *stacking* via *BAN* e *K2+PC* com ganhos respectivos em 66% e 59% dos casos. E, novamente, a combinação via *Naïve Bayes* é superada em 86% das voltas.

No sentido de refinamento das atribuições de classes do modelo, a nível de probabilidade, o *Spherical Payoff* mostra um comportamento parcialmente similar. Na Figura 54 dos métodos utilizados na combinação os que se destacam são o *AODE*, *K2*, *K2+PC* e *PC*, apresentando valores significativamente maiores que os demais, para esse conjunto, apesar da escala se alterar apenas na segunda casa decimal, os valores entre classificadores são diferentes visto que os *boxes* são possuem intersecção.

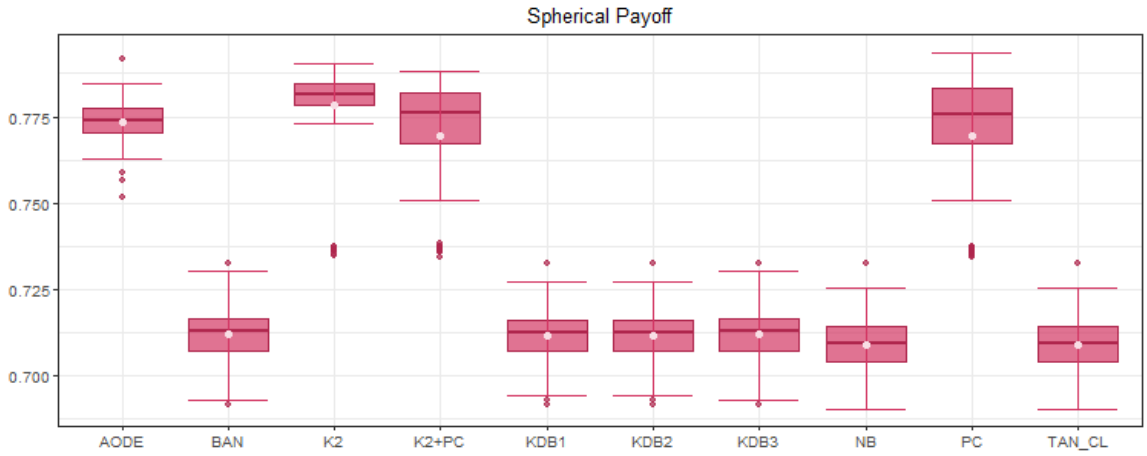


Figura 54 – Boxplot do *Spherical Payoff* para as medidas de cada uma dos métodos utilizados como meta-classificadores.

Fonte – Elaborado pela autora.

Além disso, um estudo análogo a esses dois apresentados, e complementar, da base de dados esportiva, não foi conduzido devido ao número de registros.

7.3 Comentários gerais

Esse capítulo apresentou as metodologias distintas de combinação de modelos, conhecidas como *ensemble*. Além disso, apresentou um resultado de análise de combinação de classificadores de Redes bayesianas por meio do *stacking* com o intuito de aumentar a capacidade preditiva de classificadores dessa natureza. O estudo se mostrou satisfatório, uma vez que os valores globais das estimações dos modelos individuais não supera os valores globais da metodologia combinada.

CAPÍTULO

8

CONSIDERAÇÕES FINAIS

O Aprendizado Estatístico engloba conceitos e metodologias utilizadas para a modelagem da incerteza por meio de dados. Os avanços computacionais viabilizaram o desenvolvimento de técnicas estatísticas mais complexas teoricamente e análises mais eficientes, no sentido de adequabilidade aos problemas do mundo. Portanto, a união dos dois mundos, das duas culturas de Leo Breiman, abrange soluções mais poderosas para a exploração dos dados.

Nesse sentido, as Redes Bayesianas podem ser consideradas como um método que interliga as áreas da ciência da computação e estatística, unindo as noções de grafos e de probabilidade para a representação visual da distribuição probabilística de um conjunto de variáveis. Além disso, proporciona, de maneira intuitiva, o entendimento das relações de independência condicional em modelos versáteis e flexíveis. Elas são consideradas flexíveis pois podem representar visualmente inúmeras distribuições de probabilidade, decodificando visualmente suas relações de independência e dependência, e, são consideradas versáteis porque as aplicações transitam em áreas diversificadas do conhecimento, como neurociências, biologia, medicina e sistemas de saúde, educação, engenharia, ontologia, segurança e privacidade de dados, linguagem, esporte e indústria.

Por sua vez, os classificadores que contêm uma tarefa específica possuem topologias distintamente preestabelecidas, e, suposições que direcionam os modelos em seu objetivo preditivo, mais especificamente, o classificatório. Contudo, alguns deles permitem a visualização gráfica da relação entre as variáveis, auxiliando a explicação de seus resultados, mas sem se aprofundar com a interpretação das relações.

Os objetivos dessa dissertação exploraram os seguintes pontos de análise: a comparação entre classificadores de Redes Bayesianas como o *Naïve Bayes* (*NB*), o *Tree Augmented Naïve Bayes* (*TAN*), o *K-Dependence Bayesian Network* (*k-DB*), o *Bayesian Network Augmented Naïve Bayes* (*BAN*) e o *Averaged One-Dependence Estimators* (*AODE*). Além disso, investigou-se a eficiência das Redes Bayesianas Gerais (*GBN*), estimadas por meio de dois algoritmos conhecidos

por *K2* e *PC*, e propôs-se a combinação de ambos em um método híbrido, denominado *scoring and restrict*. Acrescido a isso, uma combinação destes classificadores anteriores foi proposta por meio do *stacking* e, para cumprir com os objetivos, essas ferramentas foram aplicadas em bases de dados simulados e reais.

Inicialmente, esta dissertação apresentou os modelos de Redes Bayesianas e seu embasamento nas teorias dos grafos e de probabilidades, passando pelos fundamentos das Redes Bayesianas e, posteriormente, pelas definições dos classificadores. Após isso, seguiu-se com os estudos de comparação, simulação e combinação.

O primeiro estudo verificou o comportamento dos classificadores selecionados. Para isso, foram comparados entre si os métodos *NB*, *TAN*, *k-DB* $(k = 1,2,3)$, *BAN*, (*GBN*) e *AODE*. As bases de dados utilizadas para essa investigação foram reconhecidas na literatura por servirem a avaliações de metodologias, e são chamadas de *benchmarks*. Como as redes requerem variáveis categóricas, houve uma dicotomização das que eram de natureza contínua e as demais seguiram com suas categorias originais. Os resultados revelam que o *AODE* se sobressaiu em termos preditivos quando comparado aos demais. Este era um desfecho esperado uma vez que esse classificador é uma evolução, visto que ele foi proposto para atender especificamente à tarefa de predição, que é um método que combina estruturas intermediárias de Redes Bayesianas. Por outro lado, a estrutura simplista do *NB* refletiu um baixo desempenho nesta análise.

Na exploração do método de estimação de parâmetros via *Dirichlet-Multinomial* foram geradas amostras em cenários de dependência e independência entre covariáveis. Com o intuito de avaliar o impacto do incremento no valor do hiperparâmetro a da priori *Dirichlet*, foram considerados 30 pontos de análise, além do método tradicional de máxima verossimilhança, em que foram variados o tamanho amostral, número de variáveis e quantidade de classes dediscretização. Entretanto, é possível perceber alguns comportamentos distintos entre os classificadores. Nos cenários de independência entre as covariáveis, o *NB* apresentou melhor desempenho em todas, ou quase todas, as configurações. E, conforme discutido, esse comportamento já era esperado dada sua definição.

Entretanto, é possível perceber alguns comportamentos distintos entre os classificadores. Nos cenários de independência entre as covariáveis, o *NB* apresentou melhor desempenho em todas, ou quase todas as configurações e, conforme discutido, esse comportamento é esperado dada sua definição. Por outro lado, nos cenários de dependência entre as covariáveis, os classificadores com melhor performance foram o *ensemble AODE* seguido do *TAN*, para tamanhos amostrais maiores. Essas análises de comparação entre os classificadores sugerem que as mudanças que mais impactam no desempenho do comportamento preditivo dos classificadores está associado ao tamanho amostral e número de parâmetros - definidos pelo número de variáveis, pela quantidade de classes de discretização ou estados da variável e pelo número de conexões.

A outra simulação foi executada em um único cenário, no qual as variáveis explicativas poderiam ser dependentes ou independentes entre si. Nesse contexto, foi possível observar que o

método híbrido atende às expectativas de ter um número de conexões equilibrado entre redes muito conectadas, que são as geradas pelo método *K2*, e modelos pouco conectados, geradas pelo método *PC*. Para essa base de dados artificiais houve ainda um acréscimo nos valores preditivos das redes geradas pelo método denominado *Scoring and Restrict*, e denotado no texto como *K2+PC*.

Tendo em vista essa poderosa ferramenta e que, a estimação da estrutura tem papel fundamental na capacidade preditiva dos modelos, foram conduzidos estudos de avaliação em bases de dados reais para três áreas distintas. A primeira considerou os dados agronômicos para a avaliação das influências nas falhas de plantio da cultura de cana de açúcar. A segunda realizou uma análise financeira de concessão de crédito. Já a terceira se deu no contexto esportivo, em que foi efetuada a investigação das ações relacionadas à posse de bola de jogares de rúgbi ao longo de uma temporada de jogos. Para essas situações, os resultados elucidaram que a técnica híbrida proposta respondeu positivamente às questões de equilíbrio no número de arcos do modelo de Redes Bayesianas, sem perder em capacidade preditiva, gerando uma rede que simplifica sua explicação. Contudo, na análise dos dados financeiros, ao contrário de todas as outras, indicou melhor performance preditiva para o algoritmo *PC*. Ainda sim, houve a manutenção do equilíbrio no número de conexões.

Por fim, percorrendo o último objetivo dessa dissertação, foi conduzido um estudo de combinação entre os classificadores via *stacking* que permitiu que esses modelos heterogêneos fossem combinados de forma que, como o nome sugere, pudessem ser empilhados utilizando uma metodologia pré-selecionada com o intuito de agregar a informação de todas as técnicas utilizadas. Esses estudos foram conduzidos nas bases de dados reais nos dois primeiros campos apresentados, os dados agronômicos e os dados do setor financeiro.

Em uma fase intermediária, as mesmas técnicas foram utilizadas. No entanto, dessa vez, as covariáveis foram os resultados obtidos pelos ajustes anteriores e, posteriormente, predizendo a variável resposta original. Portanto, a avaliação foi realizada de acordo com os modelos que agregaram os ajuste, que foram chamados de meta-classificadores. Os resultados apontaram que as performances do *AODE*, *TAN*, *K2*, *K2+PC* e *PC* possuíam maior poder para agregar os modelos a serem combinados, e o *NB* apareceu novamente com um desempenho abaixo dos demais classificadores.

Com a união desses estudos, todos os objetivos desta dissertação foram integralmente contemplados, gerando resultados que agregam no que diz respeito ao comportamento dos classificadores de Redes Bayesianas, nos métodos de estimação de estrutura e na combinação dessas metodologias investigadas.

Além disso, essas análises direcionam uma gama de oportunidades de estudos e exploração de outros classificadores, outras metodologias de estimação de estrutura, de estimação de parâmetros, além da utilização de modelos diferentes como meta-classificadores no *stacking*, ou seleção de modelos para a utilização nos modelos intermediários. Acrescido a isso, podem-se

levantar outros questionamentos, dentre eles, qual seria o impacto geral da discretização no ajuste dessas redes, ou qual a ordenação das categorias ou, neste caso, como a ordenação natural das categorias das variáveis disponíveis pode alterar a capacidade de estimação da rede. Poderia ser feita uma proposição de classificadores que realizasse uma prévia seleção de variáveis antes de estimar sua estrutura completa. Tais provocações visam potencializar ainda mais o desempenho do método de Redes Bayesianas para classificação no contexto de dados discretos ou discretizados.

REFERÊNCIAS

ABELLÁN, J.; GÓMEZ-OLMEDO, M.; MORAL, S. *et al.* Some variations on the pc algorithm. In: **Probabilistic Graphical Models**. [S.l.: s.n.], 2006. p. 1–8. Citado na página 33.

ACID, S.; CAMPOS, L. M. de. Searching for bayesian network structures in the space of restricted acyclic partially directed graphs. **Journal of Artificial Intelligence Research**, v. 18, p. 445–490, 2003. Citado na página 47.

ACID, S.; CAMPOS, L. M. de; CASTELLANO, J. G. Learning bayesian network classifiers: Searching in a space of partially directed acyclic graphs. **Machine learning**, Springer, v. 59, n. 3, p. 213–235, 2005. Citado nas páginas 38 e 65.

ACID, S.; CAMPOS, L. M. de; FERNÁNDEZ-LUNA, J. M.; RODRIGUEZ, S.; RODRIGUEZ, J. M.; SALCEDO, J. L. A comparison of learning algorithms for bayesian networks: a case study based on data from an emergency medical service. **Artificial intelligence in medicine**, Elsevier, v. 30, n. 3, p. 215–232, 2004. Citado na página 14.

ADEL, T.; CAMPOS, C. P. de. Learning bayesian networks with incomplete data by augmentation. In: **Thirty-First AAAI Conference on Artificial Intelligence**. [S.l.: s.n.], 2017. Citado na página 14.

AKAIKE, H. Information Theory and an Extension of the Maximum Likelihood Principle. In: PETROV, B. N.; CSAKI, F. (Ed.). **2nd International Symposium on Information Theory**. Budapest: Akademia Kiado, 1973. p. 267–281. Citado nas páginas 50 e 70.

ALIFERIS, C. F.; STATNIKOV, A.; TSAMARDINOS, I.; MANI, S.; KOUTSOUKOS, X. D. Local causal and markov blanket induction for causal discovery and feature selection for classification part ii: analysis and extensions. **Journal of Machine Learning Research**, v. 11, n. 1, 2010. Citado na página 68.

ALMOND, R. G.; MISLEVY, R. J.; STEINBERG, L. S.; YAN, D.; WILLIAMSON, D. M. **Bayesian networks in educational assessment**. [S.l.]: Springer, 2015. Citado na página 32.

ALPAYDIN, E. **Introduction to machine learning/Ethem Alpaydin**. [S.l.]: Cambridge, MA: The MIT Press, 2010. Citado na página 29.

ANANDKUMAR, A.; HSU, D.; JAVANMARD, A.; KAKADE, S. Learning linear bayesian networks with latent variables. In: **International Conference on Machine Learning**. [S.l.: s.n.], 2013. p. 249–257. Citado na página 68.

BARBER, D. **Bayesian reasoning and machine learning**. [S.l.]: Cambridge University Press, 2012. Citado nas páginas 31 e 68.

BERETTA, S.; CASTELLI, M.; GONÇALVES, I.; HENRIQUES, R.; RAMAZZOTTI, D. Learning the structure of bayesian networks: A quantitative assessment of the effect of different algorithmic schemes. **Complexity**, Hindawi, v. 2018, 2018. Citado nas páginas 45, 50 e 51.

BERGER, J. O. **Statistical decision theory and Bayesian analysis**. [S.l.]: Springer Science & Business Media, 2013. Citado na página 27.

BERRETT, T. B.; SAMWORTH, R. J. Nonparametric independence testing via mutual information. **arXiv preprint arXiv:1711.06642**, 2017. Citado na página 42.

BICKEL, P. J.; DOKSUM, K. A. **Mathematical Statistics: Basic Ideas and Selected Topics, Volumes I-II Package**. [S.l.]: Chapman and Hall/CRC, 2015. Citado na página 26.

BIELZA, C.; LARRAÑAGA, P. Discrete bayesian network classifiers: a survey. **ACM Computing Surveys (CSUR)**, ACM, v. 47, n. 1, p. 5, 2014. Citado nas páginas 37, 46, 59, 60, 62, 64, 65, 67, 68 e 86.

BISHOP, C. M. **Pattern recognition and machine learning**. [S.l.]: springer, 2006. Citado nas páginas 15, 22 e 28.

BLAU, A. Uncertainty and the history of ideas. **History and Theory**, Wiley Online Library, v. 50, n. 3, p. 358–372, 2011. Citado na página 13.

BOUCKAERT, R. R. **Bayesian belief networks: from construction to inference**. Tese (Doutorado), 1995. Citado nas páginas 15 e 51.

BOUSQUET, O.; BOUCHERON, S.; LUGOSI, G. Introduction to statistical learning theory. In: SPRINGER. **Summer School on Machine Learning**. [S.l.], 2003. p. 169–207. Citado na página 15.

BOX, G. E.; TIAO, G. C. **Bayesian Inference in Statistical Analysis**. [S.l.]: ADDISON-WESLEY PUBLISHING COMPANY, 1973. v. 40. Citado nas páginas 22, 25, 26 e 28.

BREIMAN, L. Bagging predictors. **Machine learning**, Springer, v. 24, n. 2, p. 123–140, 1996. Citado na página 120.

____. Statistical modeling: The two cultures. **Statistical Science**, Institute of Mathematical Statistics, v. 16, n. 3, p. 199–215, 2001. ISSN 08834237. Disponível em: <http://www.jstor.org/stable/2676681>. Citado na página 13.

BUTTS, C. T. network: a package for managing relational data in r. **Journal of Statistical Software**, v. 24, n. 2, 2008. Disponível em: <http://www.jstatsoft.org/v24/i02/paper>. Citado na página 74.

CAMPOS, L. M. d. A scoring function for learning bayesian networks based on mutual information and conditional independence tests. **Journal of Machine Learning Research**, v. 7, n. Oct, p. 2149–2187, 2006. Citado nas páginas 41, 42, 47, 48, 49, 52 e 91.

CANO, I. M. D. Á.; MARTÍNEZ, J. d. S. Requirement risk level forecast using bayesian networks classifiers. Wordl Scientific, 2011. Citado na página 14.

CHENG, J.; GREINER, R. Comparing bayesian network classifiers. In: MORGAN KAUFMANN PUBLISHERS INC. **Proceedings of the Fifteenth conference on Uncertainty in artificial intelligence**. [S.l.], 1999. p. 101–108. Citado nas páginas 37, 59, 60, 61, 62 e 67.

____. Learning bayesian belief network classifiers: Algorithms and system. In: SPRINGER. **Conference of the Canadian Society for Computational Studies of Intelligence**. [S.l.], 2001. p. 141–151. Citado na página 16.

CHICKERING, D. M. A transformational characterization of equivalent bayesian network structures. In: MORGAN KAUFMANN PUBLISHERS INC. **Proceedings of the Eleventh conference on Uncertainty in artificial intelligence**. [S.l.], 1995. p. 87–98. Citado na página 50.

____. Learning bayesian networks is np-complete. In: **Learning from data**. [S.l.]: Springer, 1996. p. 121–130. Citado na página 41.

____. A transformational characterization of equivalent bayesian network structures. **arXiv preprint arXiv:1302.4938**, 2013. Citado na página 43.

CHICKERING, D. M.; HECKERMAN, D.; MEEK, C. Large-sample learning of bayesian networks is np-hard. **Journal of Machine Learning Research**, v. 5, n. Oct, p. 1287–1330, 2004. Citado na página 41.

CHOW, C.; LIU, C. Approximating discrete probability distributions with dependence trees. **IEEE transactions on Information Theory**, IEEE, v. 14, n. 3, p. 462–467, 1968. Citado na página 63.

COLOMBO, D.; MAATHUIS, M. H. Order-independent constraint-based causal structure learning. **The Journal of Machine Learning Research**, JMLR. org, v. 15, n. 1, p. 3741–3782, 2014. Citado na página 43.

CONSONNI, G.; FOUSKAKIS, D.; LISEO, B.; NTZOUFRAS, I. *et al.* Prior distributions for objective bayesian analysis. **Bayesian Analysis**, International Society for Bayesian Analysis, v. 13, n. 2, p. 627–679, 2018. Citado na página 28.

COOPER, G. F.; HERSKOVITS, E. A bayesian method for the induction of probabilistic networks from data. **Machine learning**, Springer, v. 9, n. 4, p. 309–347, 1992. Citado na página 48.

COVER, T. M.; THOMAS, J. A. **Elements of information theory**. [S.l.]: John Wiley & Sons, 2012. Citado na página 42.

COX, L. A. Causal graph models for predictive and prescriptive analytics. **Wiley StatsRef: Statistics Reference Online**, Wiley Online Library, p. 1–10, 2014. Citado na página 14.

DALE, A. I. **A History of Inverse Probability: From Thomas Bayes to Karl Pearson**. [S.l.]: Springer Science & Business Media, 1999. v. 16. Citado na página 24.

DAWID, A.; COWELL, R.; LAURITZEN, S.; SPIEGELHALTER, D. Probabilistic networks and expert systems. In: . [S.l.]: Springer-Verlag, 1999. Citado na página 31.

DEGROOT, M. H. **Optimal statistical decisions**. [S.l.]: John Wiley & Sons, 2004. v. 82. Citado nas páginas 22 e 23.

DEGROOT, M. H.; SCHERVISH, M. J. **Probability and statistics**. [S.l.]: Pearson Education, 2012. Citado nas páginas 22, 24, 26, 27 e 28.

DING, F.; ZHUANG, Y. Ego-network probabilistic graphical model for discovering on-line communities. **Applied Intelligence**, Springer, v. 48, n. 9, p. 3038–3052, 2018. Citado na página 14.

DRUZDZEL, M. J.; SIMON, H. A. Causality in bayesian belief networks. In: MORGAN KAUFMANN PUBLISHERS INC. **Proceedings of the Ninth international conference on Uncertainty in artificial intelligence**. [S.l.], 1993. p. 3–11. Citado na página 38.

EDERA, A.; STRAPPA, Y.; BROMBERG, F. The grow-shrink strategy for learning markov network structures constrained by context-specific independences. In: SPRINGER. **Ibero-American Conference on Artificial Intelligence**. [S.l.], 2014. p. 283–294. Citado na página 29.

EELLS, E.; FETZER, J. H. **The Place of Probability in Science: In Honor of Ellery Eells (1953-2006)**. [S.l.]: Springer Science & Business Media, 2010. v. 284. Citado na página 22.

ELIDAN, G. Bagged structure learning of bayesian network. In: JMLR WORKSHOP AND CONFERENCE PROCEEDINGS. **Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics**. [S.l.], 2011. p. 251–259. Citado na página 120.

FLORES, M. J.; GÁMEZ, J. A.; MARTÍNEZ, A. M. Supervised classification with bayesian networks: A review on models and applications. In: **Intelligent data analysis for real-life applications: theory and practice**. [S.l.]: IGI Global, 2012. p. 72–102. Citado na página 59.

FRENO, A. Selecting features by learning markov blankets. In: SPRINGER. **International Conference on Knowledge-Based and Intelligent Information and Engineering Systems**. [S.l.], 2007. p. 69–76. Citado na página 37.

FREUND, Y.; SCHAPIRE, R.; ABE, N. A short introduction to boosting. **Journal-Japanese Society For Artificial Intelligence**, JAPANESE SOC ARTIFICIAL INTELL, v. 14, n. 771-780, p. 1612, 1999. Citado na página 120.

FRIEDMAN, J.; HASTIE, T.; TIBSHIRANI, R. *et al.* Additive logistic regression: a statistical view of boosting (with discussion and a rejoinder by the authors). **The annals of statistics**, Institute of Mathematical Statistics, v. 28, n. 2, p. 337–407, 2000. Citado na página 120.

FRIEDMAN, N.; GEIGER, D.; GOLDSZMIDT, M. Bayesian network classifiers. **Machine learning**, Springer, v. 29, n. 2-3, p. 131–163, 1997. Citado na página 65.

FRIEDMAN, N.; GOLDSZMIDT, M. Building classifiers using bayesian networks. In: **Proceedings of the national conference on artificial intelligence**. [S.l.: s.n.], 1996. p. 1277–1284. Citado nas páginas 60, 62 e 63.

FRIEDMAN, N.; GOLDSZMIDT, M.; LEE, T. J. Bayesian network classification with continuous attributes: Getting the best of both discretization and parametric fitting. In: **ICML**. [S.l.: s.n.], 1998. v. 98, p. 179–187. Citado na página 59.

FU, S.; DESMARAIS, M. C. Markov blanket based feature selection: a review of past decade. In: NEWSWOOD LTD. **Proceedings of the world congress on engineering**. [S.l.], 2010. v. 1, p. 321–328. Citado na página 68.

GÁMEZ, J. A.; MATEO, J. L.; PUERTA, J. M. Learning bayesian networks by hill climbing: efficient methods based on progressive restriction of the neighborhood. **Data Mining and Knowledge Discovery**, Springer, v. 22, n. 1-2, p. 106–148, 2011. Citado na página 50.

GAVRIL, F. Generating the maximum spanning trees of a weighted graph. **Journal of Algorithms**, Elsevier, v. 8, n. 4, p. 592–597, 1987. Citado na página 63.

GEIGER, D. An entropy-based learning algorithm of bayesian conditional trees. In: ELSEVIER. **Uncertainty in Artificial Intelligence**. [S.l.], 1992. p. 92–97. Citado na página 63.

GIBBS, J. W. **Elementary principles in statistical mechanics: developed with especial reference to the rational foundations of thermodynamics**. [S.l.]: C. Scribner's sons, 1902. Citado na página 28.

GLOVER, F. Future paths for integer programming and links to ar tifi cial intelli g en ce. **Computers operations research**, v. 13, n. 5, p. 533–549, 1986. Citado na página 51.

GLOVER, F.; HANAFI, S. Finite convergence of tabu search. In: CITESEER. **Proc. MIC**. [S.l.], 2001. p. 333–336. Citado na página 52.

GORODKIN, J. Comparing two k-category assignments by a k-category correlation coefficient. **Computational biology and chemistry**, Elsevier, v. 28, n. 5-6, p. 367–374, 2004. Citado na página 71.

GROSS, J. L.; YELLEN, J. **Handbook of graph theory**. [S.l.]: CRC press, 2004. Citado na página 18.

GROSS, J. L.; YELLEN, J.; ZHANG, P. **Handbook of graph theory**. [S.l.]: CRC press, 2013. Citado nas páginas 17, 18, 19 e 21.

HALPERN, J. Y. **Reasoning about uncertainty**. [S.l.]: MIT press, 2017. Citado na página 13.

HASTIE, T.; TIBSHIRANI, R.; FRIEDMAN, J. **The elements of statistical learning: data mining, inference, and prediction**. [S.l.]: Springer Science & Business Media, 2009. Citado na página 15.

HECKERMAN, D. A tutorial on learning with bayesian networks. In: **Innovations in Bayesian networks**. [S.l.]: Springer, 2008. p. 33–82. Citado na página 24.

HECKERMAN, D.; GEIGER, D. Likelihoods and parameter priors for bayesian networks. **Tech. MSRTR-95-54. Microsoft Research**, 1995. Citado na página 55.

HECKERMAN, D.; GEIGER, D.; CHICKERING, D. M. Learning bayesian networks: The combination of knowledge and statistical data. **Machine learning**, Springer, v. 20, n. 3, p. 197–243, 1995. Citado nas páginas 31, 48 e 49.

HITCHCOCK, C. Probabilistic causation. 1997. Citado na página 38.

IZBICKI, R.; SANTOS, T. M. dos. **Aprendizado de máquina: uma abordagem estatística**. [S.l.: s.n.], 2020. ISBN 978-65-00-02410-4. Citado nas páginas 15, 86 e 119.

JAMES, G.; WITTEN, D.; HASTIE, T.; TIBSHIRANI, R. **An introduction to statistical learning**. [S.l.]: Springer, 2013. v. 112. Citado nas páginas 15, 16, 70, 72, 73, 86 e 119.

JAYNES, E. T. **Probability theory: The logic of science**. [S.l.]: Cambridge university press, 2003. Citado nas páginas 22, 27 e 28.

JEFFREYS, H. An invariant form for the prior probability in estimation problems. **Proceedings of the Royal Society of London. Series A. Mathematical and Physical Sciences**, The Royal Society London, v. 186, n. 1007, p. 453–461, 1946. Citado na página 27.

JI, Z.; XIA, Q.; MENG, G. A review of parameter learning methods in bayesian network. In: SPRINGER. **International Conference on Intelligent Computing**. [S.l.], 2015. p. 3–12. Citado nas páginas 55, 56 e 57.

JIANG, L.; ZHANG, H.; CAI, Z.; SU, J. Learning tree augmented naive bayes for ranking. In: SPRINGER. **International Conference on Database Systems for Advanced Applications**. [S.l.], 2005. p. 688–698. Citado na página 16.

JING, Y.; PAVLOVIĆ, V.; REHG, J. M. Boosted bayesian network classifiers. **Machine Learning**, Springer, v. 73, n. 2, p. 155–184, 2008. Citado nas páginas 68 e 120.

JOSHI, A. A.; JOSHI, S. H.; LEAHY, R. M.; SHATTUCK, D. W.; DINOV, I.; TOGA, A. W. Bayesian approach for network modeling of brain structural features. In: INTERNATIONAL SOCIETY FOR OPTICS AND PHOTONICS. **Medical Imaging 2010: Biomedical Applications in Molecular, Structural, and Functional Imaging**. [S.l.], 2010. v. 7626, p. 762607. Citado na página 14.

KALISCH, M.; BÜHLMANN, P. Estimating high-dimensional directed acyclic graphs with the pc-algorithm. **Journal of Machine Learning Research**, v. 8, n. Mar, p. 613–636, 2007. Citado na página 32.

KOLLER, D.; FRIEDMAN, N. **Probabilistic graphical models: principles and techniques**. [S.l.]: MIT press, 2009. Citado nas páginas 20, 28, 32, 34, 35, 36, 37, 39, 40, 55 e 57.

KORB, K. B.; NICHOLSON, A. E. The causal interpretation of bayesian networks. In: **Innovations in Bayesian Networks**. [S.l.]: Springer, 2008. p. 83–116. Citado nas páginas 23, 39 e 40.

____. **Bayesian artificial intelligence**. [S.l.]: CRC press, 2010. Citado nas páginas 14, 20, 22, 23, 24, 25, 31, 34, 35, 38, 39, 59 e 90.

KWOH, C.-K.; GILLIES, D. F. Using hidden nodes in bayesian networks. **Artificial intelligence**, Elsevier, v. 88, n. 1-2, p. 1–38, 1996. Citado na página 68.

LAURITZEN, S. L. **Graphical models**. [S.l.]: Clarendon Press, 1996. v. 17. Citado nas páginas 21 e 28.

LERNER, B.; MALKA*, R. Investigation of the k2 algorithm in learning bayesian network classifiers. **Applied Artificial Intelligence**, Taylor & Francis, v. 25, n. 1, p. 74–96, 2011. Citado na página 91.

LI, J.; ZHANG, C.; WANG, T.; ZHANG, Y. Generalized additive bayesian network classifiers. In: **IJCAI**. [S.l.: s.n.], 2007. p. 913–918. Citado na página 68.

LIU, X.-Q.; LIU, X.-S. Markov blanket and markov boundary of multiple variables. **The Journal of Machine Learning Research**, JMLR. org, v. 19, n. 1, p. 1658–1707, 2018. Citado na página 44.

LIU, Z.; MALONE, B.; YUAN, C. Empirical evaluation of scoring functions for bayesian network model selection. In: SPRINGER. **BMC bioinformatics**. [S.l.], 2012. v. 13, n. S15, p. S14. Citado nas páginas 47 e 48.

LOUZADA, F.; ARA, A. Bagging k-dependence probabilistic networks: An alternative powerful fraud detection tool. **Expert Systems with Applications**, Elsevier, v. 39, n. 14, p. 11583–11592, 2012. Citado nas páginas 68, 71, 80 e 120.

MARCOT, B. G. Metrics for evaluating performance and uncertainty of bayesian network models. **Ecological modelling**, Elsevier, v. 230, p. 50–62, 2012. Citado na página 72.

MARGARITIS, D. **Learning Bayesian network model structure from data**. [S.l.], 2003. Citado nas páginas 43, 44 e 45.

MARGARITIS, D.; THRUN, S. Bayesian network induction via local neighborhoods. In: **Advances in neural information processing systems**. [S.l.: s.n.], 2000. p. 505–511. Citado na página 44.

MARON, M. E.; KUHNS, J. L. On relevance, probabilistic indexing and information retrieval. **Journal of the ACM (JACM)**, ACM New York, NY, USA, v. 7, n. 3, p. 216–244, 1960. Citado nas páginas 60 e 61.

MEGANCK, S.; LERAY, P.; MANDERICK, B. Learning causal bayesian networks from observations and experiments: A decision theoretic approach. In: SPRINGER. **International Conference on Modeling Decisions for Artificial Intelligence**. [S.l.], 2006. p. 58–69. Citado na página 39.

MEYER, P. E. **infotheo: Information-Theoretic Measures**. [S.l.], 2014. R package version 1.2.0. Disponível em: <https://CRAN.R-project.org/package=infotheo>. Citado na página 74.

MIHALJEVIĆ, B.; BIELZA, C.; LARRAÑAGA, P. bnclassify: Learning bayesian network classifiers. 2020. Citado na página 60.

MULDNER, K.; BURLESON, W.; SANDE, B. Van de; VANLEHN, K. An analysis of students' gaming behaviors in an intelligent tutoring system: predictors and impacts. **User modeling and user-adapted interaction**, Springer, v. 21, n. 1-2, p. 99–135, 2011. Citado na página 14.

NAGARAJAN, R.; SCUTARI, M.; LÈBRE, S. Bayesian networks in r. **Springer**, Springer, v. 122, p. 125–127, 2013. Citado nas páginas 31, 32, 34, 35, 39, 41, 42, 44, 48 e 50.

NASUTION, M. The uncertainty: A history in mathematics. In: IOP PUBLISHING. **Journal of Physics: Conference Series**. [S.l.], 2018. v. 1116, n. 2, p. 022031. Citado na página 13.

NEAPOLITAN, R. E. **Learning Bayesian Networks**. [S.l.]: Upper Saddle River: Pearson, 2004. Citado nas páginas 14, 18, 20, 21, 22, 24, 32, 36, 37, 38 e 55.

NIELSEN, T. D.; JENSEN, F. V. **Bayesian networks and decision graphs**. [S.l.]: Springer Science & Business Media, 2009. Citado nas páginas 33 e 37.

NUALART, D. Kolmogorov and probability theory. **Arbor**, v. 178, n. 704, p. 607–619, 2004. Citado nas páginas 22 e 23.

PARAMESWARAN, Y. G. A. **On Optimally Squishing Large Datasets**. 2016. Disponível em: <https://towardsdatascience.com/on-optimally-squishing-large-datasets-9276776cf0cb>. Citado na página 14.

PAZZANI, M.; BILLSUS, D. Learning and revising user profiles: The identification of interesting web sites. **Machine learning**, Springer, v. 27, n. 3, p. 313–331, 1997. Citado na página 68.

PEARL, J. **Probabilistic Reasoning in Intelligent Systems**. [S.l.]: Morgan Kaufmann, San Mateo, CA, 1988. Citado nas páginas 13, 14, 20, 31, 36 e 37.

____. Causal diagrams for empirical research. **Biometrika**, Oxford University Press, v. 82, n. 4, p. 669–688, 1995. Citado nas páginas 34 e 36.

____. **Causality: models, reasoning and inference**. [S.l.]: Springer, 2000. v. 29. Citado nas páginas 31, 35, 38, 41, 59 e 103.

____. **Causality: models, reasoning and inference**. [S.l.]: Cambridge University Press, 2013. Citado nas páginas 22, 29, 34, 35, 36, 38 e 39.

PEARL, J. *et al.* Causal inference in statistics: An overview. **Statistics surveys**, The author, under a Creative Commons Attribution License, v. 3, p. 96–146, 2009. Citado na página 38.

PEARL, J.; VERMA, T. S. A statistical semantics for causation. **Statistics and Computing**, Springer, v. 2, n. 2, p. 91–95, 1992. Citado na página 43.

PERKINS, J.; WANG, D. A comparison of bayesian and frequentist statistics as applied in a simple repeated measures example. **Journal of Modern Applied Statistical Methods**, v. 3, n. 1, p. 24, 2004. Citado na página 25.

PETHEL, S.; HAHS, D. Exact test of independence using mutual information. **Entropy**, Multidisciplinary Digital Publishing Institute, v. 16, n. 5, p. 2839–2849, 2014. Citado na página 42.

PETITJEAN, F.; BUNTINE, W.; WEBB, G. I.; ZAIDI, N. Accurate parameter estimation for bayesian network classifiers using hierarchical dirichlet processes. **Machine Learning**, Springer, v. 107, n. 8, p. 1303–1331, 2018. Citado na página 27.

PHAM, B. T.; PRAKASH, I.; KHOSRAVI, K.; CHAPI, K.; TRONG, P.; TRINH, T. Q. N.; HOSSEINI, S. V.; BUI, D. T. A comparison of support vector machines and bayesian algorithms for landslide susceptibility modeling. **Geocarto International**, Taylor & Francis, p. 1–23, 2018. Citado na página 14.

POURHOSEINGHOLI, M. A.; KHEIRIAN, S.; ZALI, M. R. Comparison of basic and ensemble data mining methods in predicting 5-year survival of colorectal cancer patients. **Acta Informatica Medica**, The Academy of Medical Sciences of Bosnia and Herzegovina, v. 25, n. 4, p. 254, 2017. Citado na página 14.

R Core Team. **R: A Language and Environment for Statistical Computing**. Vienna, Austria, 2018. Disponível em: <https://www.R-project.org/>. Citado na página 74.

RAHMAN, M. H. A. A.; MUSTAPHA, A.; RAZALI, N.; FAUZI, R. Bayesian approach to classification of football match outcome. **International Journal of Integrated Engineering**, v. 10, n. 6, 2018. Citado na página 14.

ŘEZÁČ, M.; ŘEZÁČ, F. How to measure the quality of credit scoring models. **Finance a úvěr: Czech Journal of Economics and Finance**, v. 61, n. 5, p. 486–507, 2011. Citado na página 109.

RIBEIRO, M. H. D. M.; COELHO, L. dos S. Ensemble approach based on bagging, boosting and stacking for short-term prediction in agribusiness time series. **Applied Soft Computing**, Elsevier, v. 86, p. 105837, 2020. Citado na página 120.

RIGGELSEN, C. Learning bayesian networks from incomplete data: An efficient method for generating approximate predictive distributions. In: SIAM. **Proceedings of the 2006 SIAM International Conference on Data Mining**. [S.l.], 2006. p. 130–140. Citado na página 14.

RISH, I. An empirical study of the naive bayes classifier. In: IBM. **IJCAI 2001 Workshop on Empirical Methods in Artificial Intelligence**. [S.l.], 2001. v. 3, n. 22, p. 41–46. Citado na página 16.

RUSSELL, S.; NORVIG, P. Artificial intelligence: a modern approach. 2002. Citado na página 53.

____. Artificial intelligence: a modern approach. 2010. Citado nas páginas 29, 47, 50, 51 e 55.

RUZ, G. A.; ARAYA-DÍAZ, P. Predicting facial biotypes using continuous bayesian network classifiers. **Complexity**, Hindawi, v. 2018, 2018. Citado nas páginas 14, 61 e 91.

SAHAMI, M. Learning limited dependence bayesian classifiers. In: **Proceedings of the Second International Conference on Knowledge Discovery and Data Mining (KDD-96)**. [S.l.: s.n.], 1996. v. 96, n. 1, p. 335–338. Citado nas páginas 16, 60 e 63.

SALMON, W. C. **Causality and explanation**. [S.l.]: Oxford University Press, 1998. Citado nas páginas 22 e 38.

SAMET, S.; MIRI, A.; GRANGER, E. Incremental learning of privacy-preserving bayesian networks. **Applied Soft Computing**, Elsevier, v. 13, n. 8, p. 3657–3667, 2013. Citado na página 14.

SANTAFÉ, G. **Advances on Supervised and Unsupervised Learning of Bayesian Network Models. Application to Population Genetics**. Tese (Doutorado) — Ph. D. thesis, University of the Basque Country, 2007. Citado na página 59.

SCHEINER, S. M.; GUREVITCH, J. **Design and analysis of ecological experiments**. [S.l.]: Oxford University Press, 2001. Citado na página 28.

SCHMITT, F. O.; GILARDONI, G. L.; ANDRADE, J. A class of flat prior distributions for the poisson-gamma hierarchical model. **Statistica Neerlandica**, Wiley Online Library, 2019. Citado na página 28.

SCHWARZ, G. Estimating the Dimension of a Model. **Annals of Statistics**, v. 6, p. 461–464, 1978. Citado nas páginas 50 e 70.

SCUTARI, M. Learning bayesian networks with the bnlearn R package. **Journal of Statistical Software**, v. 35, n. 3, p. 1–22, 2010. Citado nas páginas 53 e 74.

____. An empirical-bayes score for discrete bayesian networks. In: **Conference on probabilistic graphical models**. [S.l.: s.n.], 2016. p. 438–448. Citado na página 40.

____. Dirichlet bayesian network scores and the maximum relative entropy principle. **Behaviormetrika**, Springer, v. 45, n. 2, p. 337–362, 2018. Citado nas páginas 47 e 48.

____. **Package 'bnlearn'**. 2020. Citado nas páginas 41 e 46.

SCUTARI, M.; DENIS, J.-B. **Bayesian networks: with examples in R**. [S.l.]: Chapman and Hall/CRC, 2014. Citado na página 67.

SCUTARI, M.; GRAAFLAND, C. E.; GUTIÉRREZ, J. M. Who learns better bayesian network structures: Accuracy and speed of structure learning algorithms. **International Journal of Approximate Reasoning**, Elsevier, v. 115, p. 235–253, 2019. Citado nas páginas 38, 40, 43, 47 e 51.

SCUTARI, M.; VITOLO, C.; TUCKER, A. Learning bayesian networks from big data with greedy search: computational complexity and efficient implementation. **Statistics and Computing**, Springer, v. 29, n. 5, p. 1095–1108, 2019. Citado nas páginas 50 e 51.

SINGH, M. Learning bayesian networks from incomplete data. In: **AAAI/IAAI**. [S.l.: s.n.], 1997. p. 534–539. Citado nas páginas 14 e 68.

SPIRTES, P.; GLYMOUR, C. N.; SCHEINES, R.; HECKERMAN, D. **Causation, prediction, and search**. [S.l.]: MIT press, 2000. Citado nas páginas 39 e 40.

SPRITES, P.; GLYMOUR, C.; SCHEINES, R. **Causation, prediction and search, Cambridge**. [S.l.]: The MIT press, 2000. Citado nas páginas 19, 20, 22, 29 e 38.

STEEN, M. V. Graph theory and complex networks. **An introduction**, v. 144, 2010. Citado nas páginas 17, 20 e 21.

SU, C.; ANDREW, A.; KARAGAS, M.; BORSUK, M. Overview of bayesian network approaches to model gene-environment interactions and cancer susceptibility. 2012. Citado na página 44.

TABAK, J. Probability and statistics: The science of uncertainty. facts on file. **Inc. New York**, 2004. Citado na página 22.

TSAMARDINOS, I.; ALIFERIS, C. F.; STATNIKOV, A. R.; STATNIKOV, E. Algorithms for large scale markov blanket discovery. In: **FLAIRS conference**. [S.l.: s.n.], 2003. v. 2, p. 376–380. Citado nas páginas 37, 43, 45 e 46.

TU, S. The dirichlet-multinomial and dirichlet-categorical models for bayesian inference. **Computer Science Division, UC Berkeley**, 2014. Citado na página 27.

VERMA, T.; PEARL, J. **Equivalence and synthesis of causal models**. [S.l.]: UCLA, Computer Science Department, 1991. Citado na página 38.

WANG, C.-H.; CHENG, H.-Y.; DENG, Y.-T. Using bayesian belief network and time-series model to conduct prescriptive and predictive analytics for computer industries. **Computers & Industrial Engineering**, Elsevier, v. 115, p. 486–494, 2018. Citado na página 14.

WASSERMAN, L. **All of statistics: a concise course in statistical inference**. [S.l.]: Springer Science & Business Media, 2013. Citado na página 25.

WEBB, G. I.; BOUGHTON, J. R.; WANG, Z. Not so naive bayes: aggregating one-dependence estimators. **Machine learning**, Springer, v. 58, n. 1, p. 5–24, 2005. Citado nas páginas 16 e 66.

WITTEN, I. H.; FRANK, E.; HALL, M. A.; PAL, C. J. **Data Mining: Practical machine learning tools and techniques**. [S.l.]: Morgan Kaufmann, 2016. Citado nas páginas 72, 73 e 74.

WOLPERT, D. H.; MACREADY, W. G. No free lunch theorems for optimization. **IEEE transactions on evolutionary computation**, IEEE, v. 1, n. 1, p. 67–82, 1997. Citado na página 15.

YANG, R.; BERGER, J. O. **A catalog of noninformative priors**. [S.l.]: Institute of Statistics and Decision Sciences, Duke University, 1996. Citado na página 28.

YANG, S.; CHANG, K.-C. Comparison of score metrics for bayesian network learning. In: IEEE. **1996 IEEE International Conference on Systems, Man and Cybernetics. Information Intelligence and Systems (Cat. No. 96CH35929)**. [S.l.], 1996. v. 3, p. 2155–2160. Citado na página 48.

YARAMAKALA, S.; MARGARITIS, D. Speculative markov blanket discovery for optimal feature selection. In: IEEE. **Fifth IEEE International Conference on Data Mining (ICDM'05)**. [S.l.], 2005. p. 4–pp. Citado na página 46.

ZENG, G. Metric divergence measures and information value in credit scoring. **Journal of Mathematics**, Hindawi, v. 2013, 2013. Citado na página 108.

ZHANG, H. The optimality of naive bayes. In: **Proceeding of the Seventeenth International Florida Artificial Intelligence Research Society Conference**. [S.l.: s.n.], 2004. v. 1, n. 2, p. 73–76. Citado na página 16.

ZHANG, J.; SPIRTES, P. Intervention, determinism, and the causal minimality condition. **Synthese**, Springer, v. 182, n. 3, p. 335–347, 2011. Citado nas páginas 38, 39 e 40.

ZHANG, N. L.; POOLE, D. Exploiting causal independence in bayesian network inference. **Journal of Artificial Intelligence Research**, v. 5, p. 301–328, 1996. Citado na página 38.

ZHANG, Y.; ZHANG, Z.; LIU, K.; QIAN, G. An improved iamb algorithm for markov blanket discovery. **JCP**, Citeseer, v. 5, n. 11, p. 1755–1761, 2010. Citado na página 46.

ZHENG, F.; WEBB, G. I. Averaged one-dependence estimators. In: ____. **Encyclopedia of Machine Learning and Data Mining**. Boston, MA: Springer US, 2017. p. 85–87. ISBN 978-1-4899-7687-1. Disponível em: <https://doi.org/10.1007/978-1-4899-7687-1_48>. Citado na página 66.

ZOU, X.; YUE, W. L. A bayesian network approach to causation analysis of road accidents using netica. **Journal of Advanced Transportation**, Hindawi, v. 2017, 2017. Citado na página 14.


ICMC
USP
SÃO CARLOS