UNIVERSIDADE FEDERAL DE SERGIPE

CENTRO DE CIENCIAS EXATAS E TECNOLOGIA

DEPARTAMENTO DE ESTATISTICA E CIENCIAS ATUARIAIS

Audir Bispo Santos

ESTUDO DE CASO COM ABORDAGEM DAS REDES BAYESIANAS NO RAMO DE SEGUROS AUTOMOTIVOS

São Cristóvão – SE

2020

Audir Bispo Santos

ESTUDO DE CASO COM ABORDAGEM DAS REDES BAYESIANAS NO RAMO DE SEGUROS AUTOMOTIVOS

Trabalho de Conclusão de Curso apresentado ao Departamento de Estatística e Ciências Atuariais da Universidade Federal de Sergipe, como parte dos requisitos para obtenção do grau de Bacharel em Ciências Atuariais.

Orientador (a): Amanda da Silva Lira

São Cristóvão – SE

2020

Audir Bispo Santos

ESTUDO DE CASO COM ABORDAGEM DAS REDES BAYESIANAS NO RAMO DE SEGUROS AUTOMOTIVOS

Trabalho de Conclusão de Curso apresentado ao Departamento de Estatística e Ciências Atuariais da Universidade Federal de Sergipe, como um dos pré-requisitos para obtenção do grau de Bacharel em Ciências Atuariais.

____/____/____

Banca Examinadora:

_____

Prof.ª Dr.a Amanda da Silva Lira

Orientadora

_____

Prof.ª Me. Elielma Santana de Jesus

1° Examinador

_____

Prof.ª Bel. Gislaine Santana Góis

2° Examinador

AGRADECIMENTOS

Primeiramente agradeço a todos que direta e indiretamente contribuíram para a conclusão deste trabalho e principalmente aqueles que sempre me incentivaram a superar os problemas e seguir em frente para concluir o curso. Agradeço a Deus por cada batalha vencida, algumas com facilidade outras não, mas sempre com saúde para seguir em frente.

Agradeço a meus familiares, estes últimos com o interesse explícito em que o curso fosse terminado com brevidade, porém foi cursado com a maior serenidade possível. Como poderia não agradecer aos colegas das turmas, que no início eram só conhecidos e que findaram pessoas próximas e com vaga garantida nas minhas lembranças.

A minha orientadora Amanda Lira aquele agradecimento especial pela paciência, oportunidade e toda a experiência que tem me passado nesses anos de estudo.

Por fim a todos os professores tanto do departamento de Estatística e Ciências Atuariais e dos outros departamentos que contribuíram para a formação.

“Mudam-se os tempos, mudam-se as vontades,
Muda-se o ser, muda-se a confiança;
Todo o mundo é composto de mudança,
Tomando sempre novas qualidades.”

Luís Camões

RESUMO

A tomada de decisão diante de informações parciais ou imprecisas que precisam de técnicas para tratar incertezas sugere o uso das Redes Bayesianas, que tem estudos relativamente bem disseminados em várias áreas do conhecimento humano, mas ainda pouco explorando a área de seguros. Existem no mercado seguros de vários tipos e a população atual demanda tecnologia, novos produtos que reduzam o risco e cobertura de acordo com suas necessidades. Os mais comuns são os seguros de automóvel, de imóveis, de vida, acidentes pessoais, de saúde, dentre outros. O seguro é um serviço de proteção muito importante para o automóvel e que muitas pessoas ainda o evitam, relacionado na maioria das vezes ao seu preço. O presente trabalho tem como objetivo verificar o grau de conhecimento e importância do seguro de veículo automotor em uma população de discentes dos cursos de exatas da Universidade Federal de Sergipe (UFS) e irá aplicar de forma pioneira as técnicas de Redes Bayesianas para análises. Para a aquisição dos dados foi elaborado um projeto piloto empregando amostragem estratificada, aplicando um questionário aos discentes do Centro de Ciências Exatas e Tecnologia (CCET) da UFS. Estes dados foram explorados com o auxílio do software R Project com ênfase nos pacotes de associações entre variáveis e a formação da estrutura das Redes Bayesianas. Utilizando a técnica de mineração de regras de associação e algoritmos de pontuação, a rede inicialmente gerada não incluiu as variáveis importância e conhecimento sobre seguro automotivo, mesmo observando que 91% do total dos entrevistados atribuíram nota entre 7 e 10 para importância do seguro. Essas variáveis foram inseridas de forma manual tendo como resultado uma rede com associações contrarias com a realidade condicional de cada variável. Deste modo, apesar das alterações e atendendo aos objetivos, a rede foi capaz de externar a importância do seguro veicular relacionado ao ramo de automóveis através da variável renda, mesmo com um conhecimento mediano entre os entrevistados, por meio das probabilidades obtidas.

Palavras-Chave: Probabilidade. Seguro de automóveis. Importância de seguros. Amostragem.

ABSTRACT

Decision making in the face of partial or inaccurate information that needs techniques to deal with uncertainties suggests the use of Bayesian Networks, which have relatively well-disseminated studies in various areas of human knowledge, but still little exploring the insurance area. There are various types of insurance on the market and the current population demands technology, new products that reduce risk and coverage according to their needs. The most common are auto, property, life, personal accident and health insurance, among others. Insurance is a very important protection service for the car and many people still avoid it, most often related to its price. The present work aims to verify the degree of knowledge and importance of motor vehicle insurance in a population of students of the exact courses of the Federal University of Sergipe (UFS) and will apply in a pioneering way the techniques of Bayesian Networks for analysis. For the acquisition of the data a pilot project was elaborated using stratified sampling, applying a questionnaire to the students of the Center of Exact Sciences and Technology (CCET) of UFS. These data were explored with the aid of the R Project software with an emphasis on the association packages between variables and the formation of the structure of the Bayesian Networks. Using the association rules mining technique and scoring algorithms, the network initially generated did not include the variables importance and knowledge about automotive insurance, even noting that 91% of the total respondents rated between 7 and 10 for the importance of insurance. These variables were entered manually resulting in a network with associations contrary to the conditional reality of each variable. Thus, despite the changes and meeting the objectives, the network was able to express the importance of vehicle insurance related to the automobile industry through the income variable, even with a median knowledge among the interviewees, through the obtained probabilities.

Key words: Probability. Car insurance. Importance of insurance. Sampling.

LISTA DE ILUSTRAÇÕES

LISTA DE TABELAS

LISTA DE ABREVIATURAS E SIGLAS

AAE	Amostra Aleatória Estratificada
ANTSEGURO	Já teve seguro antes
CCET	Centro de Ciência Exatas e Tecnologia
CNH	Carteira Nacional de Habilitação
CONHECIMENTO	Variável Conhecimento sobre Seguros Automotivos
COR	Variável raça ou cor da pele
DAG	Directed Acyclic Graph (Grafo Acíclico Orientado)
DEPTO	Departamento do curso do CCET da UFS
DPC	Distribuição de Probabilidade Conjunta
DPVAT	Seguro de Danos Pessoais Causados por Veículos Automotores de Vias Terrestres
HC	Algoritmo Hill-Climbing
IDADE	Variável Idade separada por classes
IMPORTANCIA	Importância ao seguro automotivo
IRB	Instituto de Resseguros do Brasil
MOTIVONTS	Motivo Não ter Seguro
MPV	Medida Provisória
PERIODO	Período que o discente está cursando
PVEICULO	Possui veículo
RB	Rede Bayesiana
RENDA	Renda familiar
SEXO	Variável Sexo
SNSP	Sistema Nacional de Seguros Privados
SUSEP	Superintendência de Seguros Privados
SVEICULO	Veículo possui seguro
TEMPCNH	Tempo que possui CNH em anos
TEMPSEGURO	Tempo que possui contrato de seguro em meses
TPC	Tabela de Probabilidade Condicional
TURNO	Turno que estuda em diurno ou noturno
TSINISTRO	Tipo de Sinistro
TVEICULO	Tipo do Veículo
UFS	Universidade Federal de Sergipe
USOSERG	Usou ou acionou o Seguro

SUMÁRIO

1.INTRODUÇÃO

A aplicação de habilidades quantitativas para solucionar problemas que envolvem riscos ou incertezas é cada vez mais exigida. Principalmente para as tomadas de decisões financeiras mais adequadas e ao desenvolvimento de modelos que avaliam impactos financeiros de eventos futuros e incertos, sobretudo os relacionados às operações de seguro, previdência e gestão de risco (FILHO, 2000).

O segmento dos seguros depende fortemente de dados históricos para avaliar riscos e modelar o preço para seus clientes, porém a qualidade e quantidade desses dados podem determinar o sucesso ou o insucesso em um processo de precificação. Esses dados dependem de quantos segurados existem e de quanto sinistros acontecem em um determinado período de tempo.

Em 2018 o segmento Pessoas superou o segmento Automóveis pelo segundo ano consecutivo no número de novos contratos de seguro. O segmento Automóveis apresentou uma queda na arrecadação entre 2018 e 2019, mas não menos importante em volumes (SUSEP, 2019). Desta forma, pôde-se chegar ao objetivo questionando qual o conhecimento que as pessoas tem sobre o seguro automotivo e que importância dão para possivelmente poder adquiri-lo. Dessa necessidade, a aplicação do raciocínio probabilístico se faz necessária como a técnica para tratar a incerteza. Esta técnica através da teoria da probabilidade oferece maneira quantitativa de codificar a incerteza, se utilizando de uma semântica clara, com probabilidades que podem ser obtidas através dos dados e incorporando novas evidências de forma direta (MARGARITIS, 2003).

Uma importante ferramenta matemática, com aplicação em várias áreas de conhecimento, a Teoria dos Grafos é um conjunto de vértices conectado por arestas, que oferece soluções para diversos tipos de problemas de áreas como redes tecnológicas, biológicas, sociais e de informações (JEQUESSENE, 2010). Embasada nesta teoria está a rede Bayesiana, que trata o raciocínio probabilístico e pode gerar a agilidade esperada vinda da teoria dos grafos no trato das incertezas existentes, pois os caminhos mais curtos de um grafo podem ser definidos por uma menor distância entre os vértices.

No tocante do mercado de seguro, a importância e o conhecimento dado pelos usuários vêm a ser testada neste trabalho, pois nos conturbados dias atuais, existe a evidência de que os seguros são importantes para qualquer eventualidade que possa acontecer, roubos, furtos e morte, etc. No entanto deve-se assegurar os riscos elencados como seguráveis para as seguradoras e as necessidades de cada usuário ou segurado.

Existem no mercado, seguros de vários tipos. Os mais comuns são os seguros de automóvel, de imóveis, de vida e acidentes pessoais, de saúde, dentre outros. O trabalho procura identificar qual o nível de conhecimento e importância relacionado ao seguro de veículo automotivo aplicando técnicas de redes bayesianas através do software R Project.

Pretende-se dessa forma analisar os dados, coletados a partir de um projeto piloto com aplicação de um questionário a população dos alunos dos cursos de exatas da UFS, para verificar a ideia que se tem de seguro de automóveis com a melhor utilização dessa nova técnica, no intuito de entender qual o comportamento dos possíveis novos segurados para novos ou antigos tipos de seguros e contratos, observando qual o condicionante ou variável que melhor os representam.

Este trabalho está estruturado em sete seções, começando pela introdução. Na segunda apresentam-se os objetivos, seguidos da justificativa, da revisão literária, metodologia na quinta parte. A seção seis apresenta os resultados obtidos e a discussão, finalizando com as conclusões apresentadas na seção sete. Os apêndices completam o trabalho e explanam os resultados não incorporados no texto.

2. OBJETIVOS

2.1 GERAL

Este trabalho tem como propósito um estudo de caso para aplicação de Redes Bayesianas para avaliar o conhecimento e a importância sobre o seguro automotivo dos discentes do Centro de Ciências Exatas e Tecnologia (CCET) da Universidade Federal de Sergipe (UFS).

2.2 ESPECÍFICOS

Pode-se elencar como objetivos específicos do presente trabalho os seguintes:

- Elaboração e aplicação de questionário para coletar informações sobre a importância e conhecimento dos discentes dos cursos de exatas da UFS;
- Apresentar a aplicabilidade da Rede Bayesiana com a análise na Ciência Atuarial em relação ao seguro automotivo;
- Aplicar através dos comandos no software R Project as Redes Bayesianas.

3. JUSTIFICATIVA

Na vida selvagem os leões expulsam seus filhos quando chegam a uma certa idade para não ter o risco de seu reino ameaçado. Os seres humanos têm consciência dos riscos a que estão sujeitos e por isso sentem a necessidade de também tentar mantê-los sob controle. Desta forma, não se pode evitar que o risco se concretize, mas sim uma maneira de reparar os danos gerados por essa concretização, pois para ser segurado o mesmo precisa ser incerto (FUNENSEG, 2013).

O conhecimento de uma atividade do ramo da economia, neste caso o de seguros, é essencial para a prática de processos de precificação realizado pelas empresas seguradoras para que cobrem um preço justo, e que garanta a saúde financeira da organização. Para os usuários, ou segurados, serve para que se torne claro que riscos podem ser colocados a cargo de uma seguradora. Este entendimento ou conhecimento das pessoas em relação aos seguros, principalmente de veículos automotores, pode significar um alto grau de preocupação com os bens adquiridos conforme demonstra os dados históricos.

Dessa forma, o uso da rede bayesiana pode levar a respostas mais rápidas e uma melhora no sentido de que novos integrantes da sociedade possam contribuir com dados para as suas reais necessidades futuras como segurado. Pois os produtos atuais de seguros podem ser que não sirvam como padrão atual onde as seguradoras determinam de forma unilateral as condições para o contrato de seguro.

O presente trabalho consiste na aplicação das Redes Bayesianas em áreas das ciências atuariais, mais especificamente na área de seguros automotivos, através de dados sobre o conhecimento e importância para aqueles que são segurados ou não no ramo de automóveis. Busca-se com isso verificar a agilidade de resposta dada por essa técnica estatística em questão, que tem cada vez mais relevância no mercado de trabalho com aplicações em diversas áreas, principalmente na área de gestão do risco que envolve a parte de previdência e mercado de crédito.

A incerteza é um dos principais fatores das atividades das Ciências Atuariais e as Redes Bayesianas, através da atividade do raciocínio probabilístico, tendem a minimizá-la, para uma melhor segurança nas decisões a serem tomadas avaliando o comportamento da população.

4. REVISÃO LITERARIA

4.1 A Incerteza

A incerteza é considerada como uma das três características básicas do seguro, as outras são a previdência e o mutualismo, e tem uma melhor definição colocada por Rodrigues (2008, p. 18) como sendo “a possibilidade de evento sobre a qual o gestor da decisão não dispõe de informações para inferir de forma prospectiva o curso das chances favorecendo a tomada de decisões segundo informações de forma subjetiva ou percepções pessoais”. Em outras palavras a incerteza é considerada a dúvida quanto a ocorrência do evento que provoca perdas de cunho econômico, que abrange dois aspectos, a própria ocorrência e quando ela acontecerá.

As incertezas ainda podem levar a riscos indesejáveis, o que se obtém uma maior quantidade possível de informações, transformando as incertezas em riscos. A partir de então, o que se busca são ações que possam mitigar os riscos, ou seja, não adianta apenas identifica-los, é necessário gerenciá-los (RODRIGUES, 2008).

Segundo Castro e Zuben (2008) a incerteza origina-se de alguma deficiência de informação que pode ser incompleta, vaga, imprecisa e contraditória. Técnicas como o raciocínio lógico, a lógica nebulosa e o raciocínio probabilístico são algumas das alternativas utilizadas para tratar a incerteza.

4.2 Teoria da Probabilidade

Encontra-se na natureza dois tipos de fenômenos: determinísticos e aleatórios. Os fenômenos determinísticos são aqueles em que os resultados são sempre os mesmos, qualquer que seja o número de ocorrência dos mesmos. Nos fenômenos aleatórios, os resultados não serão previsíveis, mesmo que haja um grande número de repetições do mesmo fenômeno.

Em termos gerais Orlandeli (2005) considera que as probabilidades são associadas a eventos, baseadas em experimentos e que podem ser estimadas, mas quando estas não são possíveis se torna difícil à utilização da forma clássica da probabilidade como, por exemplo: o cálculo da frequência relativa.

Alguns conceitos básicos conforme Morenttin (2010) são necessários para prosseguir a explanação:

- O experimento aleatório é um experimento cujo o resultado não pode ser previsto, ou seja, será diferente e não previsível a cada ocorrência;

- O espaço amostral (Ω) é o conjunto de valores que a variável aleatória pode assumir, como por exemplo {cara, coroa};
- O evento é um subconjunto do espaço amostral, que pode ser denotado por qualquer letra do alfabeto na forma maiúscula, tendo como exemplo o lançamento de duas moedas aparecer duas caras;
- Frequência relativa é a razão entre um evento e o número de repetições do experimento.
- A variável aleatória é aquela que assume valores num espaço amostral, pode ser entendida como uma variável quantitativa, cujo resultado (valor) depende de fatores aleatórios.

Conforme Castro e Zuben (2008) se o espaço amostral consiste de N elementos igualmente prováveis e um evento A corresponde a um subconjunto de k elementos do espaço amostral, temos que a probabilidade de ocorrer A é dada por:

$$P(A) = k / N$$

Caso o espaço amostral seja contínuo é necessário conhecer a função de distribuição de probabilidade para caracterizar a variável, caso seja discreto a função de massa de probabilidade é necessária (CASTRO; ZUBEN, 2008).

Com isso a probabilidade é uma função P que associa a cada evento de E um número real pertencente ao intervalo [0,1], ou seja, $0 \leq P(X) \leq 1$, para todo X, $X \subset \Omega$ (MORENTTIN, 2010). Para isto, axiomas abaixo são necessários:

- $P(\Omega) = 1$;
- $P(A \cup B) = P(A) + P(B)$, se A e B forem mutuamente exclusivos;
- $P(\bigcup_{i=1}^{n} A_i) = \sum_{i=1}^{n} P(A_i)$,

Se A_1, A_2, ..., A_n, forem, dois a dois, eventos mutuamente exclusivos.

Os eventos podem ser independentes (podem ocorrer ao mesmo tempo) ou mutuamente exclusivos (não ocorrem ao mesmo tempo).

4.2.1 Probabilidade Conjunta

A probabilidade conjunta tem importância relevada para o estudo em questão, pois elucida o conceito de eventos independentes, podendo se dizer que a ocorrência de um não tem nenhuma influência sobre o outro. Ela fornece a chance de dois ou mais eventos ocorrerem ao mesmo tempo (RODRIGUES, 2008).

De forma geral para ocorrer na probabilidade conjunta os eventos precisam ser independentes. Como, por exemplo, a quantidade de sinistros automotivos ocorridos (A) com o valor do sinistro (B), ou seja, P(A∩B) =P(A) x P(B).

4.2.2 Probabilidade Incondicional

Mais conhecida como a probabilidade a priori, a probabilidade incondicional é utilizada para anunciar o grau de certeza que uma proposição é verdadeira quando não há nenhuma informação adicional sobre ela (ORLANDELI, 2005). Esta é denotada por P(A), por exemplo a probabilidade da produção de uma máquina é de 0,6. Pode ser representada da seguinte forma; P(A) = 0,6.

4.2.3 Probabilidade Condicional

Como o nome já diz teremos uma condição para o resultado ser alcançado. Segundo Orlandeli (2005), a probabilidade de que o evento A ocorra, dado que o evento B ocorre, é chamada probabilidade condicional de A dado B, denotada por P(A|B), e definida como:

$$P(A|B) = \frac{P(A \cap B)}{P(B)} \quad , \qquad \text{Se } P(B) > 0$$

Logo, verifica-se que P(A|B) é uma restrição de A ao novo espaço amostral B, já P(B) é a probabilidade a priori de B, tendo que essa probabilidade seja diferente de zero (ORLANDELI, 2005).

Fazendo algumas substituições matemáticas pode-se reescrever a fórmula e responder à pergunta, dado que ocorreu o evento B, qual a probabilidade de ocorrer o evento A?

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

Com P(B) > 0, temos a regra de Bayes ou teorema de Bayes.

4.3 Teorema de Bayes

Nascido em Londres em 1702, Thomas Bayes foi matemático, reverendo e tornou-se um importante personagem por formular o teorema da probabilidade ao qual deu seu nome. Bayes formulou um teorema capaz de lidar com incertezas e atualizar nossa crença em um determinado evento à medida que novas informações chegam (CASTRO; ZUBEN, 2008).

Segundo Souza (2010), o teorema de Bayes é uma junção do teorema de probabilidade condicional e da fórmula de probabilidades totais. Já conforme Morenttin (2010) o teorema de Bayes também é chamado de teorema da probabilidade a posteriori, pois relaciona uma das parcelas de probabilidade total com a própria probabilidade total, ou ainda conhecido como teorema da probabilidade das causas.

$$P(Ai|B) = \frac{P(B|Ai)P(B)}{\sum_{j=1}^{n} P(B|Aj)P(Aj)}$$

Com isso chega-se à equação de Bayes, acima citada. Uma ferramenta útil para inferir a probabilidade a posteriori de um evento baseado na evidência e em um conhecimento a priori de outros eventos (CASTRO; ZUBEN, 2008). Tem-se o exemplo utilizado por Castro e Zuben (2010), porém com alguns dados alterados:

As máquinas A e B são responsáveis por 60% e 40%, respectivamente, da produção de uma empresa. Os índices de peças defeituosas na produção destas máquinas valem 3% e 7% respectivamente. Se uma peça defeituosa foi selecionada da produção desta empresa, qual é a probabilidade de que tenha sido produzida pela máquina B?

Solução:

Definindo os eventos

A: peça produzida por A *P(d|A) = 0,03*

B: peça produzido por B *P(d|B) = 0,07*

d: peça defeituosa *P(A) = 0,60 e P(B)=0,40*

Como se quer a P(B|d) temos que:

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

Essa é a probabilidade de ser produzida por B dado que é defeituosa!

$$P(B|d) = \frac{0{,}07.0{,}4}{0{,}03.0{,}6 + 0{,}07.0{,}4} = 0{,}6087\ ou\ 60{,}87\%$$

A regra de Bayes e o campo resultante chamada análise bayesiana formam a base da maioria das abordagens modernas para raciocínio incerto em sistemas de IA.

4.4 Cadeia de Markov

Existem modelos de probabilidade para processos que evoluem no tempo de maneira probabilística que são denominados processos estocásticos. A Cadeia de Markov é tida como uma técnica de análise adequada a certos casos especiais de problemas probabilísticos. O processo Markoviano é considerado um dos tipos de processos estocásticos, nele a probabilidade condicional de qualquer evento futuro, dado qualquer evento passado e o estado presente X(tk) = xk, é independente do evento passado e depende somente do estado presente (ORLANDELI, 2005). Matematicamente:

P {X (t k+1) < x k+1|X (t k) = xk, X (tk−1) = xk−1, ..., X(t1) = x1, X(t0) = x0} =

= P {X (t k+1) < xk+1 X(tk) = x k}.

O processo começa em um desses estados e de maneira sucessiva move-se de um estado para outro. Cada movimento é chamado de passo. Dessa forma o processo estocástico é dito Markoviano, quando o evento futuro depende do evento presente e o evento passado é desprezado. Segundo Orlandeli (2005) este processo pode ser denominado processo sem memória.

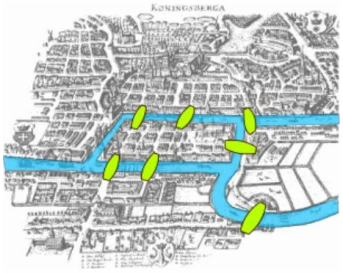
As probabilidades condicionais são denominadas Probabilidades de Transição e de forma geral representam a probabilidade do estado $X_{(tk+1)}$ ser xk+1 no instante tk+1 dado que o estado $X_{(tk)}$ no instante tk. O processo ainda pode permanecer no estado que se encontra e isso ocorre com certo valor de probabilidade.

4.5 Teoria dos Grafos

A teoria dos grafos é um ramo da matemática, em parte, novo que pode ser utilizada para resolver diversos problemas em diversas áreas de conhecimento, como as redes tecnológicas, biológicas, sociais, física, a química, a psicologia, e as engenharias, principalmente as de computação. Redes de computadores, redes de telefonia e a internet são os exemplos mais claros (BOAVENTURA NETTO; JURKIEWICZ, 2017).

O precursor desta teoria é Leonhard Paul Euler (1707-1783), matemático suíço que levantou a questão em que se relacionava ordenamento urbano e arquitetura da cidade de Konigsberg, hoje Kaliningrado na Rússia (Figura 1).

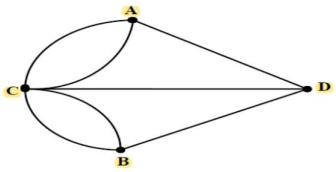
Figura 1 - Pontes de Konigsberg sobre o Rio Pregel.



Fonte: Favaro (2017)

Segundo Jequessene (2010), a pergunta era: Seria possível iniciar o percurso numa das quatro zonas e percorrer todas as pontes sem repetir nenhuma? Euler desenhou o gráfico abaixo, figura 2, e provou mais tarde que o caso não havia solução e somente teria se o número de saídas e chegadas de cada ponto fosse par.

Figura 2 - Diagrama dos sete pontos sobre o rio.



Fonte: Jequessene (2010).

Esta questão foi o começo da teoria dos grafos e mesmo sendo esquecida durante alguns anos, por pouco ter sido realizado nos anos posteriores ao trabalho de Euler, Boaventura Netto e Jurkiewicz (2017) coloca que ocorreu a utilização de modelos de grafos no estudo de circuitos elétricos, em 1847, por Gustav Robert Kirchhoff (1824-

1887) e, dez anos mais tarde, na enumeração dos isômeros dos hidrocarbonetos alifáticos saturados, por Arthur Cayley (1821-1895).

Jequessene (2010), afirma que os grafos podem ser usados para observar a informação relacionada com a estrutura de independência condicional existente entre as variáveis ou objetos de estudo. Tanto a dependência quanto a independência condicional são os pilares teóricos para os modelos grafos, combinados com as propriedades de Markov, que determinam um conjunto de regras explicitas para interpretar os grafos de independência.

Com isso, conclui-se que a teoria dos grafos disponibiliza uma estrutura unificada para a análise de estatísticas tanto para dados contínuos quanto para dados discretos e que para a representação de redes probabilísticas encontra-se nos grafos uma considerável citação.

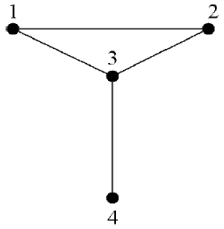
4.5.1 Conceito de Grafo

Segundo Prestes (2016) a definição de um grafo simples G = (V, A) é uma estrutura discreta formada por um conjunto finito não vazio de vértices V e um conjunto de arestas $A \subseteq P(V)$, onde $P(V) = \{\{x, y\}: x, y \in V\}$ é o conjunto de todos os pares não ordenados não necessariamente distintos gerados a partir de V. Cada aresta $\{x, y\} \in A$ é formada por um par de vértices distintos, i. e., $x \neq y$. Para cada par de vértices existe no máximo uma aresta associada.

Em outras palavras, grafo é um modelo matemático que representa relações entre objetos. Onde V representa número de vértices, ou variáveis em estatística, de G que são unidos por um conjunto de arestas de G denotadas por A.

Na figura 3 observa-se um grafo simples. Nele as linhas são as arestas e os pontos são os vértices. Cada aresta une um par de vértices distintos representados pelos pontos V(G) = {1, 2, 3, 4}. As arestas, A(G) = {(1,2); (1,3); (2,3); (3,4)}, são a representação retirada do gráfico gerando um grafo G= (4,4).

Figura 3 – Exemplo de grafo simples



Fonte: Próprio Autor

Existem ainda grafos onde o vértice vem isolado, ou seja, sem nenhuma aresta chegando ou saindo dele G= (V, Ø).

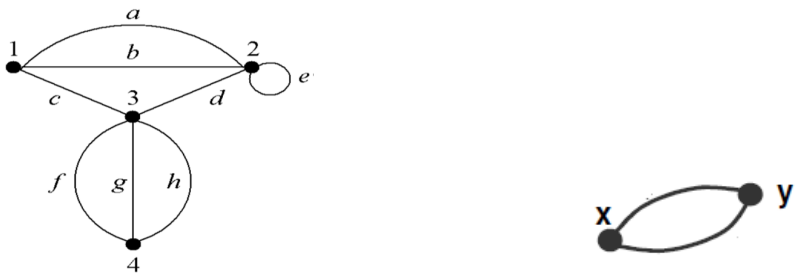
Para um grafo G = (V, A) não ser considerado simples e ser chamado de multigrafo é necessário que existam múltiplas arestas (ou arestas paralelas) entre os mesmos pares de vértices de G.

A função *f* de A em {{x, y} | x, y ∈ V, x ≠ y} representa essa ideia. Conforme figura 4(b) temos um exemplo de multigrafo. Já a figura 4(a) apresenta um pseudografo, ou seja, quando o multigrafo possui um laço. Laço ou loop é uma aresta com um só vértice, conforme a aresta "*e*" na figura 4(a) do tipo A= {x, x}, ou A={x}.

Na figura 4 (b) tem-se a representação do multigrafo por V= {x, y}, A= {(x, y), (y, x)}, onde se observa dois vértices e duas arestas paralelas ou múltiplas, V=2 e A=2. Para a figura 4 (a) tem-se um pseudografo, onde as arestas estão simbolizadas pelas letras de a até h, melhor representado pela da função de mapeamento, a chamada de função de incidência, onde f: A→P(V) é a função que representa a incidência.

Com isso f({x, y}) = {x, y}, ∀ {x, y}Є A, temos o conjunto de vértices V= {1,2,3,4} e o conjunto de arestas A = {a,b,c,d,e,f,g,h}, onde f(a)={1,2}, f(b)={1,2}, f(c)={1,3}, f(d)={2,3}, f(e)={2}, f(f)={3,4}, f(g)={3,4}, f(h)={3,4}. Como f (a) = f(b) = {1,2} e f(f)=f(g)=f(h) = {3,4}, apesar que a ≠ b e f ≠ g ≠ h, diz-se que o par de vértices {1,2} tem multiplicidade 2 e o par de vértices {3,4} tem multiplicidade 3. Multiplicidade significa os possíveis caminhos num grafo.

Figura 4 – Exemplo de pseudografo e multigrafo



a) b)

Fonte: Prestes (2016).

De acordo com Prestes (2016) ao se retirar os laços e todas as arestas paralelas de cada par de vértice, mantendo apenas uma, o pseudografo da figura 5(a) se tornará um grafo simples. Esse novo grafo formado é chamado de grafo subjacente ao pseudografo,

representado na figura 5 (b). O mesmo pode ser feito em relação a um multigrafo e ser chamado de grafo subjacente ao multigrafo, caso a figura 5(a) não possuísse laços.

Figura 5 - Exemplo de pseudografo e seu subjacente.

a) b)

Fonte: Prestes (2016).

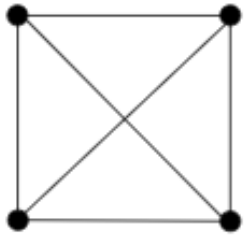
Grafo completo é o grafo no qual quaisquer dois vértices distintos são adjacentes, isto é, todo vértice é adjacente a todos os outros vértices, pois todo vértice tem ligação com todos os outros.

Como pode-se observar na figura 6 todos os vértices se conectam entre si. Na figura 6, ainda é possível atribuir o conceito de grafo conexo, pois é possível ir de um vértice a qualquer outro usando algumas de suas arestas.

4.5.2 Ordem e Tamanho de um Grafo

A ordem dos grafos é definida pelo número de elementos em determinado conjunto, ou seja, pelo número de vértices de G dito cardinalidade por Boaventura Netto e Jurkiewicz (2017). Já o tamanho de um grafo se define pela cardinalidade de seu conjunto de Arestas. Na figura 6 temos um exemplo de grafo de ordem V(G)=4 e tamanho A(G)=6.

Figura 6 – Ilustração de um Grafo Completo



Fonte: Jequessene (2010)

4.5.3 Vértices e Arestas Adjacentes de um Grafo

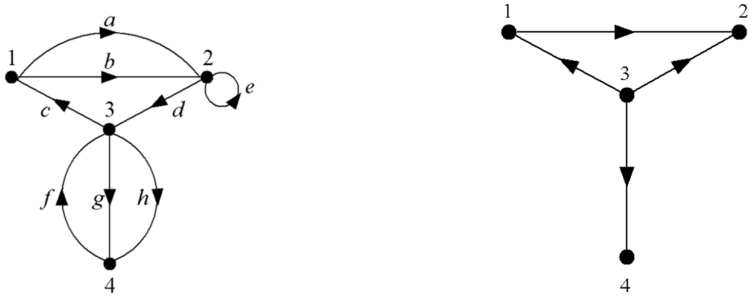
O vértice é dito adjacente quando esse está ligado diretamente por uma aresta a outro vértice, ou seja, são vértices vizinhos ligados por uma aresta. Na figura 7(a) os vértices 1 e 3 são adjacentes.

As arestas são consideradas adjacentes quando duas arestas têm um dos vértices extremos em comum. Na figura 7(a) as arestas *b* e *c* ilustram essa característica.

4.5.4 Grau de um Vértice

O grau de um vértice é o número de arestas que incidem sobre ele. Qualquer vértice de grau zero é dito isolado, os de um grau são ditos pendentes, os que tem número ímpar de arestas são chamados vértices impares e os que tiverem número de arestas pares são chamados de vértices pares. Nestes termos pode-se chegar ao chamado grafo regular, onde todos os vértices têm o mesmo grau. Segundo Boaventura Netto e Jurkiewicz (2017), a soma do grau dos vértices é sempre um número par.

Figura 7 – Pseudografo Orientado e Multigrafo Orientado



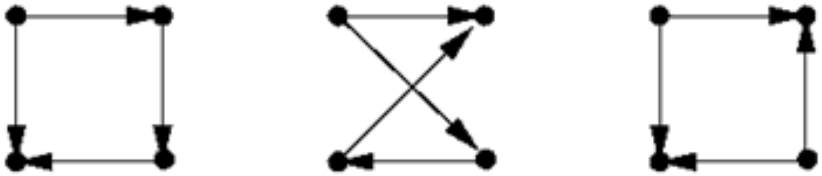
a) b)

Fonte: Prestes (2016)

4.5.5 Grafo Orientado (Dígrafo)

As arestas são responsáveis por expressar a função dos grafos de representar a relação entre objetos ou elementos. Segundo Prestes (2016), existem relações com situações não simétricas, como o fluxo de veículo nas ruas de uma cidade por exemplo, onde se existe a necessidade de se dá uma orientação. Essa informação de orientação faz com que os grafos se tornem orientados, ou seja, tem-se pares ordenados de vértices ordenados conceituando assim de grafo orientado ou dígrafo. A figura 8 ilustra três exemplos de grafos orientados.

Figura 8 – Exemplos de Grafos Orientados



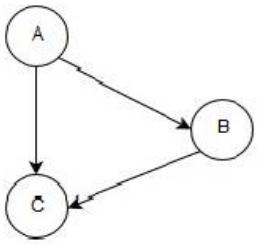
a) b) c)

Fonte: Jequessene (2010)

Os grafos orientados demonstrados acima são simples ou chamados dígrafos simples, porém existem também os multigrafos orientados e os pseudografos orientados.

Contanto, conforme explica Prestes (2016), um grafo, pseudografo ou multigrafo, orientado (ou dígrafo) D = (V, A) consiste de um conjunto V(vértices) e de um conjunto de A(arestas) de pares ordenados de vértices distintos. Em um grafo orientado, cada aresta A= (x, y) possui uma única direção de *x* para *y*. Com isso pode-se afirmar que em um grafo orientado a aresta (*x*, *y*) é divergente de *x* e convergente a *y*. Observando a figura 9, tem-se a representação de um dígrafo D=(V,A), onde *V*(G) = {vA, vB,vC} é o conjunto de vértices e A(*G*) = {(vA, vB); (vA, vC); (vB, vC)} é o conjunto das arestas. A aresta (vA, vB) é divergente ao vértice A e convergente ao vértice B.

Figura 9 – Grafo orientado ou dígrafo



Fonte: Próprio Autor.

No caso dos dígrafos percebe-se que existem graus de saída e de entrada. Os grafos que não possuem entrada são chamados de fonte, já os que não possuem saída são chamados de sumidouro, ou seja, apenas recebe informações. Um laço em um hipotético vértice *x* é contado uma única vez no grau de entrada de *x* e uma única vez no grau de saída de *x* (PRESTES, 2016).

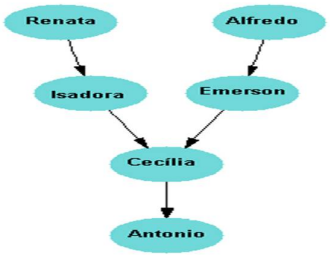
Considerando o exemplo abaixo, para um melhor entendimento do exposto, temos um grafo definido por:

V = {p | p é uma pessoa da família Santos}

A = {(x, y) | x é pai/mãe de y}

O conjunto dos vértices será V= {Renata, Alfredo, Isadora, Emerson, Cecília, Antônio}. No caso das arestas temos o conjunto A= {(Renata, Isadora), (Alfredo, Emerson), (Isadora, Cecília), (Emerson, Cecília), (Cecília, Antônio)}. Observando os nós, tem-se que o vértice Renata é pai (ou mãe) de Isadora e o mesmo acontece com Alfredo para com Emerson. Esta relação não é simétrica, pois há uma orientação e conforme a figura 10. Emerson e Isadora são descendentes de Renata e Alfredo respectivamente.

Figura 10 – Exemplo de grafo orientado e rotulado.



Fonte: https://slideplayer.com.br/slide/67281/

Com isso chega-se à definição de sucessor e antecessor em um grafo. Sucessor é indicado pela aresta que parte de *x* e chega a *y*, como no exemplo Emerson é sucessor de Alfredo. Um vértice *x* é antecessor de *y* se há um arco que parte de *x* e chega em *y* (PRESTES, 2016). No exemplo anterior diz-se que Isadora e Emerson são antecessores de Cecília.

Em relação ao grau de entrada e saída tem-se no exemplo Antônio como sumidouro, Renata e Alfredo como fonte e os demais com grau de entrada e saída. Confirma-se então que um vértice é uma fonte se o grau de entrada é zero e é um sumidouro se o valor do grau de saída é zero.

4.5.6 Grafos Valorados

Um grafo valorado G (V, A) consiste de um conjunto finito não vazio de vértices, ligados por um conjunto A de arestas (ou arcos) com pesos. Este conjunto de arestas, consiste de triplas distintas da forma (x, y, valor), em que x e y são vértices pertencentes a V e valor é um número real.

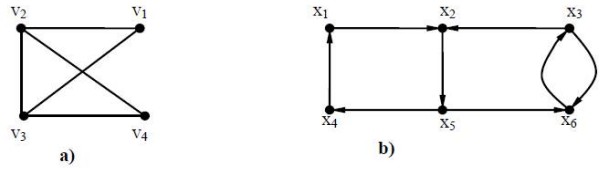
4.5.7 Cadeia, Caminho e Ciclo

Jequessene (2010) define uma cadeia sendo uma sequência qualquer de arestas adjacentes que ligam dois vértices. Por exemplo v1, v2, v3, v4 (figura 11a). Este conceito também serve para grafos orientados, desde que não se considerem as orientações. As cadeias podem ser elementares, sendo que não se repete vértices (v1, v3, v4), e simples, desde que não se repitam arestas (v1, v3, v4, v2, v3).

Um caminho é uma cadeia na qual todos os arcos possuem a mesma orientação. Ou seja, é uma sequência de vértices adjacentes em que a extremidade final de uma aresta (arco) é extremidade inicial da aresta seguinte (JEQUESSENE, 2010).

Ciclo é uma cadeia simples e fechada, onde o primeiro vértice também será o último. Pode-se observar os vértices x1, x2, x5, x4, x1 e considerar que é um grafo não acíclico, ou seja, tem um ciclo (figura 11b).

Figura 11 – Ilustração De Cadeia E Caminho Em Um Grafo



Fonte: Jequessene (2010)

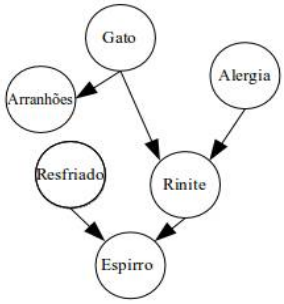
4.5.8 Grafo Acíclico

Tem como definição ser um grafo que não possui um ciclo, ou seja, o caminho que começa em um vértice e não termina nesse mesmo vértice.

4.5.8.1 Grafo Acíclico Orientado

O grafo acíclico orientado (chamado DAG do termo inglês Directed Acyclic Graph) é um grafo orientado que não possui ciclo, como o exemplo da figura 12, onde pode-se observar que o grafo orientado começa e não termina no mesmo vértice. Estes grafos podem ser utilizados, como por exemplo, em pré-requisitos de um curso de graduação, como também de ferramenta para as redes Bayesianas usadas nas inferências probabilísticas da Inteligência Artificial. O exemplo da figura 12 é uma rede Bayesiana.

Figura 12 – Grafo Acíclico Direcionado (DAG)



Fonte: https://madoc.univ-nantes.fr/pluginfile.php/342418/mod_resource/content/0/ECD/BN-student-modelling.pdf

4.6 Mineração de Dados e Regras de Associação

Com o crescimento das informações de grande conteúdo das bases de dados, o processo de descoberta de conhecimento em bancos de dados (*Knowledge Discovery in Database - KDD*) coloca a mineração de dados como a etapa principal da busca pelo conhecimento (JESUS, 2019). Diversos são os conceitos dados as técnicas de mineração de dados (data mining), e alguns deles se popularizaram para um melhor entendimento com a ideia de analisar dados com a premissa de a que resultado se quer chegar. Com isso, o que melhor define a mineração de dados é o que a trata como a exploração e análise de dados por meio de um processo computacional com o objetivo de descobrir padrões em grandes quantidades de dados. Na mineração de dados utiliza-se técnicas e métodos de diversas áreas e nelas se destacam: inteligência artificial, aprendizado de máquina, estatísticas e sistemas de banco de dados.

As regras de associação para mineração de dados (*association rules mining)* estão no trabalho de descobrir quais atributos combinam entre si. Segundo Carvalho e Vasconcelos (2004), a associação tem como premissa básica encontrar elementos que implicam na presença de outros elementos em uma mesma transação, ou seja, encontrar relacionamentos ou padrões frequentes entre conjuntos de dados. Conforme Jesus (2019), uma base de dados D formada por um conjunto de itens $I = \{i_1, i_2 \ldots, in\}$ e um conjunto de transações $T = \{t_1, t_2 \ldots, t_m\}$, sendo que cada transação contém itens pertencentes a I.

As regras de associação representam padrões existentes em transações armazenadas a partir de uma base de dados, na qual registram-se os itens adquiridos por clientes, com o uso de regras de associação, poderia gerar a seguinte regra: {cinto, bolsa} → {sapato}, a qual indica que o cliente que compra cinto e bolsa, com um certo grau de

certeza, compra também sapato. Este grau de certeza em uma regra é definido por dois índices: os fatores de suporte e de confiança (CARVALHO; VASCONCELOS, 2004).

Segundo Carvalho e Vasconcelos (2004), "o problema de mineração por regras de associação está em gerar todas as regras que contenham o suporte e confiança iguais ou maiores do que os valores mínimos determinados pelo usuário, referenciados como suporte mínimo e confiança mínima, respectivamente".

O suporte de uma regra X ⇒ Y, onde X e Y são conjuntos de itens, é dado pela seguinte fórmula:

$$Suporte = P(A \cap B) = \frac{Frequencia\ de\ X\ e\ Y}{Total\ T}$$

O numerador se refere ao número de transações em que X e Y ocorrem simultaneamente e o denominador ao total de transações.

A sua confiança é dada pela seguinte fórmula:

$$Confiança = \frac{P(A \cap B)}{P(A)} = \frac{Frequencia\ de\ X\ e\ Y}{Total\ T}$$

O numerador se refere ao número de transações em que X e Y ocorrem simultaneamente. O denominador se refere à quantidade de transações em que o item X ocorre. O suporte pode ser descrito como a probabilidade de que uma transação qualquer satisfaça tanto X quanto Y, ao passo que a confiança é a probabilidade de que uma transação satisfaça Y, dado que ela satisfaz X (CARVALHO; VASCONCELOS, 2004).

Segundo Jesus (2019), "existem vantagens por trás do uso da mineração de regras de associação que estão na descoberta de informações implícitas e a que níveis de suporte e confiança os itens de interesse estão ocorrendo junto a outros". Desta forma, não seria possível realizar inferências a respeito dos dados, já que para isso os eles devem ser consistentes para poder permitir a transformação em transações para extração de conhecimentos que possam ser absorvidos (JESUS, 2019).

4.7 Redes Bayesianas

A tomada de decisão diante de informações parciais ou imprecisas que precisam de técnicas para tratar incertezas, é um dos fatores, ou o principal, que levam ao o uso de redes Bayesianas (RBs). A utilização deste método acarreta na otimização da decisão, previsão e diagnósticos utilizando as variáveis em estudo.

As redes Bayesianas ou redes causais, como também são conhecidas, podem ser utilizadas em diversas áreas como modelar relações de causa e efeito, hábitos e eventos

de incerteza (SOBERANIS, 2010). Contudo, segundo Jequessene (2010) as redes Bayesianas não se referem apenas a casualidade, e não há exigência de que sempre as ligações representem um impacto causal, concluindo que as redes Bayesianas são estruturas de modelagens versáteis, adequadas para muitos domínios problemáticos.

4.7.1 Independência Condicional

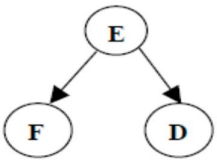
Se X e Y são independentes, então Y não é informativo para X. Portanto significa que conhecer Y, não altera em nada a probabilidade de X.

Segundo Santos (2015) uma variável de uma RB é dita condicionalmente independente de outra se o seu valor probabilístico não influencia em nenhuma maneira o valor da outra variável. Formalmente, uma variável X é dita condicionalmente independente da variável Y dada uma variável Z se P(X|Y, Z) = P(X|Z). No exemplo extraído de Orlandeli (2005), a informação estrutural do domínio, representada pela rede de Bayes, traduz-se em certas suposições de independência condicional. No exemplo, isto significa que considerando que E é uma causa primária e, F e D são efeitos de E:

$$P(D|E, F) = P(D|E)$$

Ou seja, a probabilidade de ocorrência de dor de cabeça (D) e Fobia (evidência), associadas à enxaqueca (E) que é a causa, são independentes desde que a causa seja confirmada (figura 13).

Figura 13 - Rede Bayesiana Simples



Fonte: Orlandeli (2005)

4.7.2 Conceito de Redes Bayesianas

Como se pode prevê, o termo de RB deriva da utilização do estudo de Thomas Bayes para o cálculo de probabilidades entre variáveis que apresentam relação.

Segundo Russel e Norvig (2004), a regra bayesiana observa o problema de construir hipóteses a partir de dados como um subproblema do problema mais fundamental de fazer previsões. A ideia é usar hipóteses como intermediários entre dados e previsões (RUSSEL; NORVIG, 2004).

Desenvolvidas inicialmente no fim dos anos 1970 as RBs tiveram um rápido surgimento devido a sua capacidade para inferências bidirecionais (relações citadas por Bayes) em conjunto com uma base probabilística rigorosa (CORANDY; JOUFFE, 2010). Além disso, colocam-na como uma alternativa para a Inteligência Artificial (IA) nos sistemas inteligentes que precisam de raciocínio probabilístico, tornando-se o método de escolha para o raciocínio incerto. As RBs podem ser desenvolvidas de uma combinação de inteligência humana e artificial.

As RBs combinam princípios da teoria dos grafos, teoria das probabilidades, ciência computacional e estatística. Por conseguinte, o conceito de RBs está atrelado a teoria dos grafos desde que sejam direcionados, acíclicos e conectados (DAG). Segundo Corandy e Jouffe (2010) existem modelos probabilísticos baseados em grafos acíclicos dirigidos (DAG) que têm uma longa e rica tradição, iniciada nos primeiros registros da década de 1920.

Desse modo chega-se a definição que uma RB é um par (G, Θ), em que G é um grafo orientado acíclico (DAG) e "Θ" é um conjunto particular de parâmetros. Este conjunto de parâmetros que especifica as distribuições de probabilidade condicional associadas às variáveis representadas em "G" (JEQUESSENE, 2010).

Como na teoria dos grafos, utiliza-se nas RBs os vértices ou nós, só que agora eles representam as variáveis estatísticas de interesse que possuem informações de probabilidade. Estas podem ser discretas ou contínuas, apesar que a maior parte dos exemplos explanados na literatura apresentarem variáveis discretas ou caracterizando matematicamente as variáveis contínuas como discretas. As arestas direcionadas caracterizam-se como as relações que representam as dependências estatísticas ou de casualidade entre as variáveis, ou seja, links direcionados que são usados para indicar as relações de parentesco pai e filho (CORANDY; JOUFFE, 2010).

Nesse sentido Souza (2010) explana a hierarquia utilizada nas RBs aplicando os termos pai e filho. O pai é designado a variável de onde parte a aresta e filho a variável onde ela chega. Em complemento classifica que "um nó que não possui filhos é chamado de folha e um nó que origina a rede, ou seja, que não possui pais, é chamado de raiz" (SOUZA, 2010).

Segundo Karcher (2009), uma RB possui a premissa da independência condicional, onde, cada variável é independente das variáveis que não são suas descendentes no grafo dada a observação de seus pais.

4.7.3 Tipos de Conexões em uma Rede Causal

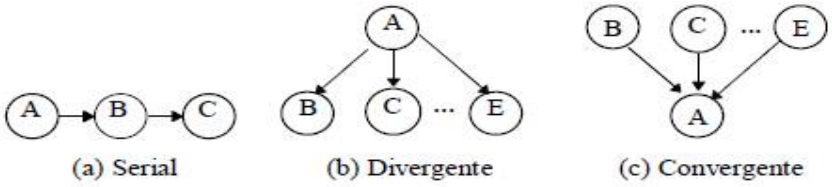
Ao se refletir com incerteza, é relevante ter a informação de que algum evento atua na crença, veracidade de uma proposição, em outros devido a criação e eliminação de relacionamentos. As formas de propagação são tidas como os tipos de conexão de uma rede causal sendo elas a conexão serial, a convergente e a divergente conforme a figura 14 (LADEIRA et al., 1999).

Na conexão serial, conforme mostrado na Figura 14 (a), uma evidência em A influencia a crença em B que influencia a crença em C. E uma evidência em C se propaga para A. Nesses casos não há propagação de influência se B está instanciado, porque o canal entre A e C fica bloqueado, tornando-os condicionalmente independentes. Com isso I(A,B,C) é válido e A e C são ditos d-separados, dado B (LADEIRA et al, 1999).

Ainda segundo Ladeira et al, (1999) na conexão divergente, uma evidência em um ascendente de A influencia a crença sobre os filhos de A, exceto se A é instanciado (figura 14 (b)).

Na conexão convergente, a evidência em A ou em um dos seus descendentes influencia a crença nos pais de A, tornando-os condicionalmente dependentes (figura 14 (c)).

Figura 14 – Tipos de Conexão de uma rede causal.



Fonte: Ladeira et all. (1999)

4.7.4 Tabela de Probabilidade Condicional

Como existe uma relação hierárquica entre as variáveis, existe um outro elemento essencial para a resolução de problemas utilizando as RBs. Segundo Souza (2010) a tabela de probabilidade condicional (TPC) é a exibição dos parâmetros de probabilidade condicional da variável condicionada a seu pai. Conforme exemplo utilizado por Souza (2010) representado na tabela 1, tem-se três variáveis assumindo valores binários, onde A e B são pais da variável C.

Tabela 1 – Tabela de Probabilidade Condicional P (C|A, B)

C	A	B	P (C|A, B)
1	1	1	θ_1
1	1	0	θ_2
1	0	1	θ_3
1	0	0	θ_4
0	1	1	θ_5
0	1	0	θ_6
0	0	1	θ_7
0	0	0	θ_8

Fonte: Souza (2010)

A distribuição de probabilidade condicional é descrita por P(X_i| pais (X_i)) e cada linha contém a probabilidade condicional relacionada a seus pais. Segundo Souza (2010) para se obter as TPCs, são utilizadas as probabilidades conjuntas de duas ou mais variáveis, onde no caso de valores contínuos são usadas funções de densidade de probabilidade, ou os valores categorizados e ainda se trabalha com intervalos.

4.7.5 Semântica das Redes Bayesianas

Segundo Souza (2010) a dinâmica de uma RB é controlada pela propriedade de Markov, a qual indica que não existem dependências diretas entre as variáveis de uma RB que não estejam explícitas através da apresentação orientada dos arcos, ou seja, cada variável possui dependência direta apenas de suas variáveis pais. A partir daí, tem-se que uma Rede Bayesiana é um par (G, Θ) definido sobre um conjunto de variáveis aleatórias $X=\{X_1, X_2, ..., X_K\}$, onde cada X_i corresponde a uma variável da rede, satisfazendo a propriedade de Markov de que a variável só depende de sua variável pai (SOUZA, 2010).

$$P\left[X_i \middle| X_j, pais(X_i)\right] = P[X_i \mid pais(X_i)] \qquad (1)$$

Fica notório que cada entrada da distribuição conjunta é caracterizada pelo produto dos elementos indicados das TPCs na RB. Desta forma em uma RB, a distribuição conjunta de probabilidades de um conjunto de variáveis discretas, é igual ao produtório das distribuições condicionais de todos os nós, dados os valores dos seus pais, ou seja, é dada pela regra da cadeia, onde a probabilidade conjunta de toda a rede se apresenta da expressão:

$$P(X_1, \dots, X_n) = \prod_{i=1}^{n} P(X_i \mid X_{pais}(X_i)) \qquad (2),$$

em que X_{pais} é o parente de que depende a variável X_i na rede em que a estrutura é G. Margaritis (2003) explana que as probabilidades condicionais $P(X_i \mid X_{pais}(X_i))$ definem a função de distribuição de probabilidade da variável estudada dado um valor atribuído a seus pais, X_{pais}. Portanto, no grafo desta equação são exatamente aquelas funções de distribuição de probabilidade locais especificadas para cada variável no domínio.

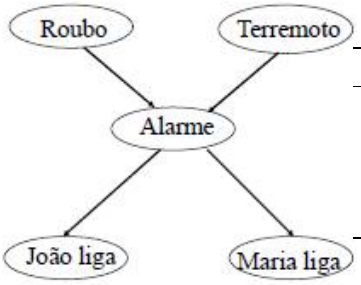
Com isso, uma vez totalmente especificada, uma RB representa compactamente a distribuição de probabilidade conjunta (DPC) e, portanto, pode ser usada para calcular as probabilidades posteriores de qualquer subconjunto de variáveis dadas evidências sobre qualquer outro subconjunto (SOBERANIS, 2010).

Segundo Margaritis (2003), usando a equação (2) toda combinação de atribuições para as variáveis $X_1 \ldots X_n$ pode ser calculada e isso pode ser ilustrado no exemplo abaixo retirado do texto de Russell e Norvig (2004):

Exemplo: *Você instalou um alarme contra roubos na sua casa, que dispara em caso de invasão. Infelizmente, o alarme é sensível a terremotos. Quando o alarme disparar, seus 2 vizinhos, João e Maria, disseram que vão te ligar. João, às vezes, confunde o alarme com a sirene do carro de bombeiro. Maria ouve música num volume alto e nem sempre escuta o alarme. (S= sim; N=não)*

P(R=S)
0,001

P(T=s)
0,002



R	T	P(A=S)
S	S	0,95
S	N	0,94
N	S	0,29
N	N	0,001

A	P(J=S)
S	0,9
N	0,05

A	P(M=S)
S	0,7
N	0,01

Dessa forma, já com as tabelas de probabilidade condicional definidas, qual a probabilidade de não haver roubo, nem terremoto, o alarme tocar, João ligar e Maria ligar?

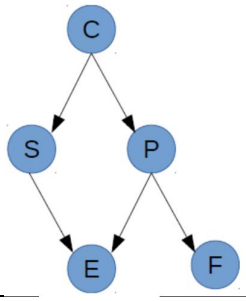
$P(\neg R \cap \neg T \cap A \cap J \cap M) = P(J \mid A) * P(M \mid A) * P(A \mid \neg R \cap \neg T) * P(\neg R) * P(\neg T)$

$= 0,9 * 0,7 * 0,001 * 0,999 * 0,998$

$= 0,00062$ ou 0,062 %

Portanto a probabilidade de não haver roubo, nem terremoto, dado que o alarme tocou João e Maria ligarem é de 0,062%.

Em um outro exemplo tem-se que a evidência é ingressar na faculdade, e considera-se as hipóteses divertir-se e ser bem-sucedido nos estudos. As informações são apresentadas nas tabelas de probabilidade condicionais e fornecem os dados necessários para explicar a vida na faculdade.



C = Ingressar na faculdade
S = Estudar
P = Frequentar festas
E = Ser bem-sucedido nos exames
F = Divertir-se

C	P(S)
V	0,8
F	0,2

P(C)
0,2

S	P	P(E)
V	V	0,6
V	F	0,9
F	V	0,1
F	F	0,2

C	P(P)
V	0,6
F	0,5

P	P(F)
V	0,9
F	0,7

Assim, observa-se a possibilidade de poder responder questões do tipo: Qual a probabilidade de ingressar na faculdade, estudar e ser bem-sucedido nos exames?

Necessita-se da probabilidade de entrar na faculdade, a de estudar dado que passou, juntamente com a de não frequentar festas dados que ingressou e a de ser bem-sucedido nos exames dado que estudou e não frequentou festa.

$P(C, S, \neg P, E, \neg F) = P(C) \bullet P(S|C) \bullet P(\neg P|C) \bullet P(E|S \cap \neg P) \bullet P(\neg F|\neg P)$

$= 0,2 \bullet 0,8 \bullet 0,4 \bullet 0,9 \bullet 0,3$

$= 0,01728$

Ou seja, a probabilidade de ingressar na faculdade, estudar e ser bem-sucedido nos exames é de 1,73%.

Nos dias atuais, técnicas de Redes Bayesianas estão presentes em vários processos englobando diversos tipos de pesquisa, tanto para o desenvolvimento de algoritmos para

aprendizado de estrutura, como em técnicas de classificação. Isso não significa um maior uso ou não pela comunidade de estatística em relação a outras áreas e segmentos acadêmicos.

As redes que podem ser obtidas por meio de aprendizagem chegam a criar situações de relações que contrariam a veracidade da análise considerada na fase inicial do estudo. Com isso:

> quando não se conhece o relacionamento entre as variáveis envolvidas no estudo, deve-se aceitar a rede resultante como verdadeira, sendo que conhecendo parcialmente como as variáveis se relacionam utiliza-se deste conhecimento em forma de blacklist e whitelist para descobrir as demais relações por meio da aprendizagem. No caso de o pesquisador ter total conhecimento da topologia da rede, esta seria definida previamente sem realização de aprendizagem, impossibilitando a descoberta de novas e interessantes relações causais. (JESUS, 2019, p. 32)

No entanto as RBs com poucas variáveis podem diretamente dificultar a compreensão sobre o completo funcionamento do processo.

4.7.6 O Algoritmo Hill-Climbing

A recente explosão de conjuntos de dados de alta dimensionalidade no campo biomédico, estatístico e em outros domínios, produzem conjuntos de dados com dezenas ou centenas de milhares de variáveis e constituiu um sério desafio para os existentes algoritmos de montagem de rede bayesiana.

A partir do problema de inferir modelos completos causais em aplicativos de mineração de dados com milhares de variáveis em larga escala, surgiu a necessidade de um plano sólido para executar os testes de independência condicional que retornam apenas as redes estatisticamente consistentes aos testes (ALIFERIS et all.,2006). O algoritmo Hill-Climbing (HC) surgiu como capaz de superar o problema sendo confiável em termos de qualidade e tempo em domínios de representação.

Segundo Aliferis et all. (2006), o HC tenta identificar a rede que maximiza uma função de pontuação indicando o quanto ela se ajusta aos dados, como também em uma segunda abordagem baseada em restrições onde utiliza estatísticas ou medidas teóricas da informação para estimar, a partir dos dados, independências condicionais entre as variáveis. Ou seja, algoritmos baseados em pontuação tem como funcionamento básico a adição, remoção ou inversão de arestas de acordo com o que proporciona uma pontuação maior após cada ação, sempre com a ajuda de uma função de pontuação (JESUS, 2019).

4.7.7 Critério de Informação Bayesiano (BIC)

Utilizando os métodos como o HC para induzir RBs a partir de dados, especialmente para fins de estimativa de distribuição de probabilidade, para a abordagem baseada em pontuação, o processo atribui uma a cada modelo de RB (MARGARITIS, 2003). Geralmente aquele que mede quão bem a RB descreve o conjunto de dados. A seleção de modelos refere-se ao problema de usar os dados para selecionar um modelo da lista de modelos concorrentes (MARGARITIS, 2003).

O BIC é um dos critérios empregados para definir qual rede ou modelo melhor se adequa aos dados, ou seja, foi projetado para encontrar o modelo mais ajustado aos dados, penalizando aqueles com muitas variáveis, sendo o de menor valor o preferido (EMILIANO, 2009).

De acordo com Emiliano (2009) o BIC é dado por:

$$BIC = -2logf(x_n|\theta) + plog\ n,$$

onde $f(x_n|\theta)$ é o modelo escolhido, p é o número de parâmetros a serem estimados e n é o número de observações da amostra.

O BIC serve para comparar quaisquer quantidades de modelos, fundamenta-se na verossimilhança, impondo, entretanto, diferentes penalizações sendo um critério de avaliação de modelos definido em termos da probabilidade a posteriori (EMILIANO, 2009).

4.8 O Seguro

4.8.1 Conceito e Evolução Histórica

Os fatos históricos apontam que o surgimento do seguro se deu a partir do instante que o homem tomou ciência da necessidade de prevenir a si e as suas criações. Segundo a FUNENSEG (2013) a espera por acontecimentos de determinados riscos tornou intensa a atitude de prevenção, ou seja, diante da incerteza, da precariedade da vida e da destruição dos bens o seguro surgiu para a cobertura dos riscos que poderiam gerar danos ao indivíduo ou as empresas.

No século XII em decorrência das viagens marítimas surgiu o primeiro contrato de dinheiro a risco marítimo. Condenado pelo Papa Gregório IX, logo foi extinto e passou a ser chamado de Feliz Destino. Para Filho (2000) o seguro moderno, como o mais

próximo ao que conhecemos hoje, teve início no século XIV com o segmento marítimo. No século XVI, implantado pelos ingleses, surgiu o seguro de vida e a partir do século XVII surgiu o seguro terrestre, por influência do grande incêndio de Londres de 1666.

O Feliz Destino consistia em que um banqueiro ou financiador comprava a embarcação com a previsão de recompra pelo vendedor. Se a embarcação chegasse ao destino sem sofrer nenhum dano, a cláusula de recompra era acionada e o banqueiro revendia a embarcação ao proprietário original por um valor maior. Se a embarcação e/ou carga se perdesse, o dinheiro adiantado pelo banqueiro corresponderia à indenização pelo dano (MANICA, 2010). Nota-se com isso o início da transferência de risco para o financiador.

Entretanto, não há certeza ou não se pode afirmar ao certo quando o seguro surgiu. Atrelado ao fato citado sobre as embarcações, supõem-se que os fenícios foram os primeiros a utilizar este tipo de atividade devido ao comércio marítimo ser a principal fonte de seu sustento, a qual estava exposta ao risco dos mares navegados. Caso alguma embarcação fosse perdida outra seria construída custeada pelos participantes da mesma viagem (MANICA, 2010).

Uma outra frente coloca que o seguro surgiu com os cameleiros nômades na época babilônica, que passavam pelo deserto para comercializar animais. Quando algum animal era perdido ou morria durante a travessia, existia a garantia de receber um outro devido aos pactos de cooperação entre todos os cameleiros (MANICA, 2010). Esses pactos originaram o mutualismo entre esses comerciantes.

No Brasil por meio da abertura dos portos ao comércio internacional, por volta de 1808 teve início o mercado de seguros, sendo a “Companhia de Seguros BOA-FÉ” a primeira companhia a funcionar no Brasil, onde a principal atividade estava a operar o seguro marítimo. Em meados dos anos 1800 surgiram inúmeras seguradoras que passaram a operar não só com o seguro marítimo, mas também com o seguro terrestre e o seguro de vida.

Segundo FUNENSEG (2013) em 1966, por meio do Decreto-lei n° 73, de 21 de novembro, foram reguladas todas as operações de seguros e resseguros e instituído o Sistema Nacional de Seguros Privados (SNSP), constituído pelo Conselho Nacional de Seguros Privados (CNSP), Superintendência de Seguros Privados (SUSEP); Instituto de Resseguros do Brasil (IRB); sociedades autorizadas a operar em seguros privados; e corretores habilitados.

Com isso, pode-se elucidar o conceito do seguro, como sendo o contrato pelo qual o segurador se obriga perante o segurado, mediante pagamento de uma certa quantia, a garantir o pagamento de um certo valor que lhe possa garantir a diminuição dos prejuízos provocados por ricos previstos.

O valor a ser ressarcido pelo segurador, não é de responsabilidade somente sua, ou seja, não é só a seguradora responsável pelo pagamento da chamada indenização devido aos danos gerados. Segundo Filho (2000) o seguro não se apresenta como um contrato único, mas sim um mecanismo de alto interesse social e humano. Desta forma a seguradora apenas intermedia as quantias pagas pelos grupos segurados, sendo todos sujeitos aos mesmos riscos, usando parte deste montante para pagar eventuais indenizações causadas por danos ocorridos.

Nesse contexto fica explicito a aplicação do mutualismo, que como explana Filho (2000) consiste em que não só um segurado pagará a indenização e sim todos do grupo de segurados daquela carteira de clientes.

De forma geral a conclusão sobre seguros é que seu objetivo principal é somente a indenização, sendo proibido o uso do seguro somente para obtê-la.

4.8.2 O Mutualismo

Segundo Filho (2000), o princípio do mutualismo junto com a probabilidade são os dois que norteiam os seguros. O mutualismo surge como a repartição do prejuízo entre os segurados, sendo assim possível a repartição do risco entre todos segurados diminuindo o prejuízo por ele trazido.

Com isso, têm-se no mutualismo a principal forma de se originar o capital necessário para o pagamento das perdas que atingem os segurados, ou de forma mais clara, sem a cooperação de uma coletividade seria praticamente impossível o indivíduo suportar os prejuízos sem ajuda (FUNENSEG, 2013).

4.8.3 Elementos Básicos e Essenciais do Seguro

Mesmo sendo um segmento fundamentado há vários anos, de acordo com FUNENSEG (2013), os elementos básicos do seguro são o segurado, o segurador, o prêmio, o risco, e a indenização. Embora deva-se considerar os sinistros como uma das partes básicas, este apenas responde como sendo a concretização do risco.

O risco é tido como um evento que pode perturbar o equilíbrio econômico. Ele tem a característica de ser incerto, aleatório, possível, concreto, lícito e fortuito

independente da vontade das partes, ou seja, de uma maneira geral o risco é a possibilidade de ocorrência de evento aleatório que cause danos de ordem pessoal, material, ou de responsabilidades (FUNENSEG, 2013).

Rodrigues (2008) caracteriza o risco como a “possibilidade de um evento sobre o qual o gestor tem base probabilística para inferir um determinado comportamento, sendo capaz de tomadas de decisões com base em dados históricos que posam mitigar perdas”. Desta forma, nem todos os riscos são objeto de seguro. Somente os seguráveis o são e devem ser indispensavelmente possíveis, futuros, incertos, que independem da vontade das partes (de forma não intencional), que não obedeçam a nenhuma lei conhecida e principalmente serem mensuráveis (FUNENSEG, 2013).

A classificação dos riscos nas operações de seguro segundo FUNENSEG (2013) se dá pelo;

- risco puro, onde só há possibilidade de perder e não perder;
- risco especulativo, que não é segurável e tem as possibilidades: perder, não perder ou ganhar;
- riscos fundamentais, que são riscos impessoais e afetam a coletividade. O tratamento desses riscos compete ao Estado, como por exemplo perdas decorrentes de guerras, e;
- riscos particulares que são aqueles que somente afetam os indivíduos ou empresas em particular, como um incêndio de uma casa, o roubo de um banco, etc.

Os riscos puros podem ser objeto de apólices de seguros, já os riscos especulativos não.

Com relação aos tipos riscos que a seguradora está exposta, os principais são:

- risco biométrico: risco que envolve desvio nos dados da demografia de uma região;
- risco de mercado: está relacionada a mudanças da taxa de juros e no valor do ativo;
- risco operacional: é o risco que envolve fraudes, erros no cálculo de prêmios;

- risco atuarial: é o risco que está presente em benefícios definidos e contribuição variada. Esse risco pode gerar a perda do equilíbrio atuarial, isto é, poderá ocorrer déficit ou superávit atuarial.

- risco moral: O indivíduo passa a agir de forma menos cautelosa, aumentando o risco de se envolver em um sinistro. O risco moral pode ser considerado um problema “pós-contratual”.

- risco de subscrição: o processo pelo qual aceitam ou rejeitam novos contratos e renovações de contratos antigos, com o objetivo de manter a lucratividade da seguradora. Porém o ato da geração do prêmio exige-se que se cumpra as normas que regulam o mercado.

Esses dois últimos tipos de risco citados são os mais presentes no mundo do seguro de automóveis.

O segurado é a pessoa ou empresa que transfere o risco, desde que este seja segurável, para uma seguradora mediante o pagamento de uma certa quantia denominada prêmio.

O segurador nada mais é que a empresa constituída na forma de sociedade anônima a qual tem como função assumir os riscos contratados com o segurado. A atribuição primordial é a de indenizar o segurado ou seu beneficiário, em caso de seguro de vida, quando ocorre a efetivação de um risco coberto por meio das condições estabelecidas em um contrato de seguro.

No Brasil, as seguradoras segundo o artigo 757 do código civil, devem ser legalmente autorizadas e são reguladas pela Superintendência de Seguros Privados (SUSEP), podendo assim assumir seguros de quaisquer valores utilizando-se de mecanismo de pulverização de riscos como a franquia, o cosseguro e o resseguro (LUCCAS FILHO, 2011). A franquia é a participação do segurado na indenização, ou seja, valores de um sinistro que não ultrapasse o valor referido definido em contrato, fica a cargo do segurado, já se o valor ultrapassar a seguradora assume o valor acima do definido. O cosseguro é a divisão do risco com outra seguradora e o resseguro é o repasse de riscos de maiores portes para uma resseguradora, podendo ser chamado de seguro das seguradoras.

Os indivíduos compram seguro para se protegerem contra as perdas ocasionadas por determinados riscos. O sinistro é tido como a realização do risco previsto no contrato

de seguro, resultando em perdas que podem atingir o segurado, seus beneficiários ou a seguradora. Quando o segurado efetua um aviso de sinistro, a seguradora é convocada a honrar a promessa feita quando emitiu o contrato de indenizá-lo pelas perdas financeiras decorrentes do sinistro relacionadas ao risco previsto no contrato (LUCCAS FILHO, 2011). Ela é total quando causa a destruição ou desaparecimento por completo do objeto segurado e parcial quando atinge somente uma parte do objeto segurado.

Em geral, o prêmio, como já apresentado, é o valor definido pela seguradora através de cálculos atuariais que o segurado tem que pagar para poder concretizar a transferência do risco para o segurador.

O prêmio pago pelo segurado para transferência do risco para o segurador é calculado com base na hipótese de ocorrência do sinistro, considerando todas as variantes que possam interferir na sua probabilidade.

Ferreira (2010) indica que o prêmio no processo de precificação do custo de um seguro possui três tipos.

- Prêmio de Risco: O prêmio de risco tem por objetivo cobrir o risco médio (E[S]). $P = E[S] = \sum_1^i S_i$,

  ou seja, é a expectativa de sinistros ocorridos (em valor, inclusive despesas de regulação de sinistros). Nele, S representa a variável aleatória "valor total das indenizações ocorridas em uma carteira de seguros" em um determinado tempo (FERREIRA, 2010).
- Prêmio Puro: O prêmio puro é o prêmio de risco associado a um carregamento de segurança estatístico (θ). Segundo Luccas Filho (2011, p. 14) "é o prêmio necessário para cobrir, com determinada probabilidade, os sinistros futuros".

  $P = E[S] (1+\theta)$

  O carregamento de segurança serve como uma margem de segurança para cobrir as flutuações estatísticas do risco, de modo que exista uma probabilidade pequena dos sinistros superarem o prêmio puro (FERREIRA, 2010).
- Prêmio Comercial: O prêmio comercial (π) corresponde ao prêmio puro com a adição do carregamento para as demais despesas da seguradora (α), incluindo a margem de lucro (FERREIRA, 2010);

$$\pi = \frac{P}{1-\alpha} = \frac{\mathrm{E}[S]\,(1+\theta)}{1-\alpha}.$$

A definição da indenização é apresentada como a contraprestação do segurador ao segurado com a efetivação do risco. Se por um lado o segurado tem por obrigação pagar um prêmio, pelo outro, a seguradora quando aceita um seguro, tem por obrigação efetuar o pagamento de uma indenização ao segurado quando ocorre um risco coberto pelo contrato de seguro (sinistro) (FUNENSEG, 2013).

4.8.4 O contrato de Seguro

A finalidade do seguro é restabelecer o equilíbrio econômico e segundo Luccas Filho (2013), o contrato de seguro é onde o mesmo se formaliza. Mais conhecido como apólice, o contrato é a formalização em que a seguradora assume o risco pré-determinado e é obrigada a garantir o interesse do segurado, mediante recebimento de do prêmio (FILHO, 2000).

Os instrumentos essenciais do contrato de seguros são a proposta e a apólice. A proposta é o instrumento formal da manifestação de vontade de quem quer efetivar um contrato de seguro. A apólice é o documento emitido pelo segurador a partir da proposta que formaliza o contrato de seguro, contendo as cláusulas e as condições gerais, especiais e particulares (FUNENSEG, 2013). As apólices, abrangem diversos campos de escolhas de como deve ser o contrato para cada indivíduo e possui algumas, regras particulares.

4.8.5 O Seguro de Veículos

O seguro de veículos no presente trabalho é o ponto de partida para entendermos o comportamento das pessoas em relação a ideia que tem sobre o assunto e sobre qual a importância dada a ele. Para se obter um contrato de seguro de veículos, e como todos os outros, é necessário o contato com um corretor, sendo esse o representante do segurado junto à seguradora. Em 2017, Segundo FENSEG (2019), no Brasil havia 17 milhões de veículos com seguro de automóvel, cerca de 30% do total. Isso ilustra uma frota de veículos com imenso potencial para a expansão do setor no mercado de seguros.

Este tipo de seguro é um produto que possibilita um leque variado de opções e tipos de coberturas no ato da contratação, atrelado a satisfação do cliente, pois cada indivíduo tem seu próprio perfil. No caso do ramo de automóveis, duas são as coberturas do risco comumente contratadas. A “cobertura de casco”, que se refere ao risco de danos

ao próprio bem segurado, que dentre outros, podem ser danos parciais, totais e de roubo; e a “cobertura de responsabilidade civil”, que se refere a danos materiais e corporais provocados pelo bem segurado a terceiros (TUDO SOBRE SEGUROS, 2019).

Desta forma, existem seguros pessoais e de passageiros, seguros para terceiros, rastreadores, bloqueadores garantia de vidros, roubos, carro reserva e assistências 24 horas (TUDO SOBRE SEGUROS, 2019). Como coberturas complementares estão “danos materiais”, relacionado a bem de terceiros e “danos corporais” relacionado a danos corporais a terceiros pelo veículo segurado.

Diversos são os riscos e prejuízos não indenizáveis, como por exemplo, o condutor sob efeito de álcool, competições, transporte de passageiros além da capacidade do veículo, excesso de carga, sinistro ocorrido pela agravação do risco, desgaste por falta de manutenção, danos a pinturas pneus e outros (DOMINGUEZ; PITA, 2011).

Segundo Dominguez e Pita (2011) o cálculo do prêmio é baseado no risco ao qual o bem está exposto, sendo suas características explicadas pela região domiciliar, modelo e ano do veículo, gênero do condutor e bônus. Para SUSEP (2019) diversas são as técnicas para o cálculo de prêmio de seguro, envolvendo vários parâmetros estatísticos. A SUSEP não estabelece a forma para a elaboração, dessa forma as Seguradoras tem liberdade de estabelecer a forma de fixação do prêmio, que deve ser apresentada à SUSEP por meio da Nota Técnica Atuarial (SUSEP, 2019).

No Brasil existe o seguro de Danos Pessoais Causados por Veículos Automotores de Vias Terrestres, ou por sua Carga, a pessoas transportadas ou não (DPVAT), que tem por finalidade de amparar as vítimas de acidentes de trânsito em todo o território nacional, independente de quem seja a culpa dos acidentes (SUSEP, 2019). Nele as coberturas são relacionadas a morte, invalidez e despesas de assistência médica e suplementares. O DPVAT é um seguro obrigatório criado pela Lei 6.194/74 que determina o pagamento por todo e qualquer veículo terrestre. As indenizações são pagas independentemente de quem seja a culpa, desde que haja vítima transportadas ou não (DOMINGUEZ; PITA, 2011).

Os recursos arrecadados, devido ao caráter social, destinam-se em 45% a Fundação Nacional de Saúde, para o custeio médico-hospitalar às vítimas de acidentes de trânsito, 5% ao Departamento Nacional de Trânsito (DENATRAN), para realização de campanha de prevenção de acidentes, SUSEP e FUNENSEG. O que sobra, 50%, vai para o pagamento das indenizações dos seguros (DOMINGUEZ; PITA, 2011). A abrangência

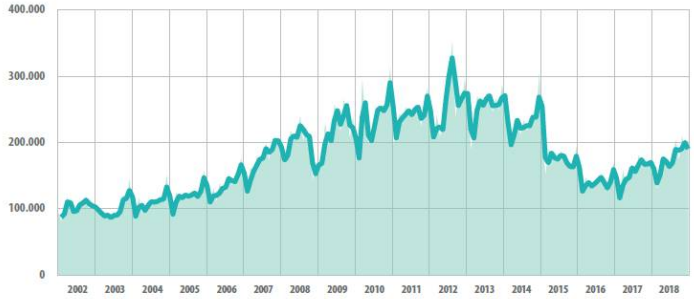
do DPVAT se dá apenas em todo território nacional. Na América do Norte e Europa existem seguros obrigatórios que cobrem danos pessoais e materiais.

As coberturas se restringem a morte, a invalidez permanente e despesas médico-hospitalares e de acordo com o código civil de 11/01/2003 o prazo para reclamar é de 3 anos. O seguro público, hoje é administrado pela Seguradora Líder DPVAT, que representa um grupo de seguradoras (DOMINGUEZ; PITA, 2011).

A Medida Provisória nº 904 de 11 de novembro de 2019, propunha a extinção do DPVAT e do Seguro Obrigatório de Danos Pessoais Causados por Embarcações ou por suas Cargas – DPEM. A MPV partia da premissa de que todos os possuidores de veículos automotores no Brasil detinham condições de adquirir um seguro em seguradora, que cobrisse a responsabilidade por danos pessoais. A MPV foi barrada por ação direta de inconstitucionalidade sendo necessária lei complementar para ser implementada (BRASIL, 2019).

Os emplacamentos realizados a cada ano têm relação direta com o mercado de seguros devido ao DPVAT e sua natureza obrigatória. Ao analisar os veículos emplacados com relação aos prêmios arrecadados, incluindo os privados e o DPVAT, entre os anos 2002 e 2018 (figura 15), observou-se uma variação com tendência ao crescimento e um desempenho parecido entre eles e o momento da economia do país no decorrer dos anos com um visível alinhamento até no período das crises econômica e política no país. Esta variação pode ser vista no mercado de seguros privados e obrigatório no Brasil como uma tendência de crescimento na evolução do mercado de seguros relacionado a um maior valor dos prêmios arrecadados pelas seguradoras.

Figura 15 – Série Histórica dos Emplacamentos, Mês a Mês – 2002 a 2018

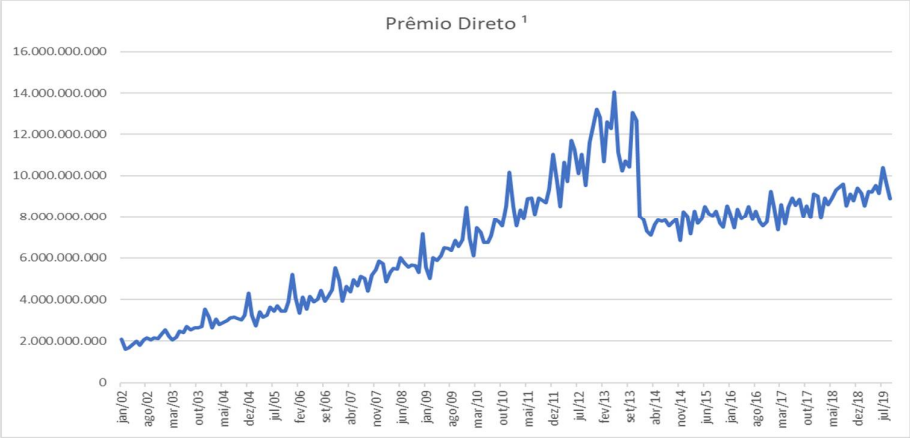


Fonte: FENABRAVE (2018)

Na figura 16 observa-se a variação da série dos prêmios diretos arrecadados. Essa tendência de crescimento acompanha a curva da série de emplacamentos, que pode ter influência desta última e de outros como por exemplo a renda e a importância dada ao

bem através do seguro. Apesar disto, o seguro de veículos teve o segundo pior desempenho na arrecadação no mercado de seguro entre 2018 e 2019 (Figura 2 no Apêndice C).

Figura 16 - Variação da Série dos Prêmios Diretos Arrecadados 2002 - 2019



Fonte: SUSEP

O seguro de veículos automotores está atrelado ao ramo de danos. O objetivo do seguro de danos é garantir ao segurado, até o limite máximo de garantia e de acordo com as condições do contrato, o pagamento de indenização por prejuízos devidamente comprovados, diretamente decorrentes de perdas e/ou danos causados aos bens segurados, ocorridos no local segurado, em consequência de risco coberto.

A partir desses pontos o seguro de veículos Automotores é definido como um contrato entre o segurado e a seguradora, onde esta última fica responsável por cobrir os riscos contratados (que podem ser colisões, roubo, furto e etc.), enquanto o segurado é responsável pelo pagamento de um prêmio por essa cobertura, além de outras obrigações.

Segundo Martins et. All (2008), além do cenário de incerteza, as seguradoras quando ofertam seus produtos (seguros) aos agentes econômicos enfrentam dois problemas relacionados à assimetria de informação, também peculiares ao mercado de seguros, pois os componentes não têm total conhecimento das ações um do outro. O primeiro é identificado na fase pré-contratual (antes da assinatura do contrato) conhecido como seleção adversa e o segundo na fase pós-contratual chamado de risco moral ou perigo moral.

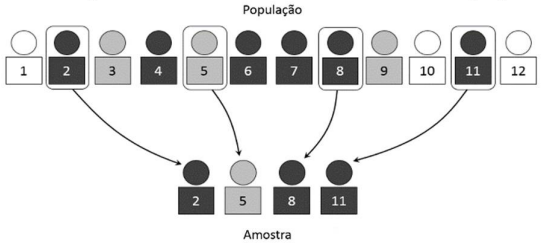
Os números no Brasil em relação a acidentes de trânsito normalmente são considerados altos em demasia e chegam a ser considerados um grave problema social, principalmente devido ao número de sinistros.

4.9 Teoria da Amostragem

A teoria da amostragem estuda as relações existentes entre uma população e as amostras extraídas dessa população. A amostragem está conceituada como o processo de determinação de uma amostra a ser pesquisada. Pois o objetivo principal da amostragem é obter informações sobre o todo tendo como base uma amostra.

Conforme Cochran (1977), a amostragem estatística consiste em uma técnica especializada usada na realização de pesquisas científicas em qualquer ramo da ação humana com o objetivo de reduzir o custo operacional e o tempo, analisando uma pequena parte da população estudada que seria a amostra (Figura 17).

Figura 17 – Exemplo de uma amostra retirada da população



Fonte: https://www.netquest.com/blog/br/blog/br/amostra-sistematica

A impossibilidade de se analisar toda a população (censo) e para que isso seja feito com o menor tempo, o menor custo e com uma maior precisão, são os motivos pelos quais se levam a utilização da amostragem. Com isso temos que uma amostra estatística é um subconjunto da população, em geral com dimensão bem menor, que também possui a característica de interesse.

A população em si, depende do interesse da pesquisa e dessa forma deve-se ter o tamanho da amostra. Normalmente é denotado pela letra “n”, que será analisado para possibilitar fazer inferência, ou seja, concluir sobre as características retiradas da amostra e aplicar a população (COCHRAN, 1977).

4.9.1 Conceitos Básicos

Alguns conceitos básicos podem elucidar as características de uma amostragem:

- População Objeto: É a população total de interesse sobre a qual desejamos obter informações.

- Característica Populacional: de acordo com a população objeto se tem a característica como o aspecto da população que precisa ser medido.
- Unidade Amostral: é definida conforme o interesse de estudo.
- Amostra: se dá como o subconjunto da população a ser analisada.
- Erro Amostral: é a diferença entre o resultado da realidade e o calculado na pesquisa, ou seja, entre a população na integral e o resultado amostral.
- Censo: a pesquisa aplicada a toda população.
- Variável de interesse: uma característica de interesse, relativa a cada elemento da população amostrada, mas que observada apenas na amostra.
- Variável observacional: é o elemento no qual a medição da(s) variável (eis) de interesse é feita.
- Unidade Observacional: é o elemento que é, de fato, selecionado para compor a amostra, ou seja,
- Unidade Amostral: é o elemento que é de fato selecionado para compor a amostra.

4.9.2 Tipos de Amostragem

A amostragem pode ser dividida em dois tipos, a amostragem não probabilística e a amostragem probabilística.

4.9.2.1 Amostragem Não-Probabilística

São amostragens em que há uma escolha deliberada dos elementos da amostra. Depende dos critérios e julgamento do pesquisador. A amostragem não-probabilística se divide em:

- Amostragem acidental: a qual a amostra é formada com quem vai aparecendo.

- Amostragem Intencional: a qual a escolha dos elementos que formarão a amostra está sujeita a uma intenção.

4.9.2.2 Amostragem Probabilística

São amostragens em que a seleção é aleatória de uma maneira em que cada elemento da população tem uma probabilidade conhecida de fazer parte da amostra. São considerados métodos extremamente científicos.

Conforme Cochran (1977), a amostragem probabilística está dividida em quatro tipos, mas abordaremos apenas os de interesse do presente trabalho:

- Amostragem Aleatória Simples: este é o processo mais elementar e se fundamenta no princípio de que todos tem a mesma probabilidade de serem incluídos na amostra. Os elementos a participarem da amostra são rotulados (ordenados) e sorteados.

- Amostragem Aleatória Estratificada (AAE): Consiste em separar a população em subgrupos mais homogêneos (estratos), de tal forma que haja uma homogeneidade dentro dos estratos e uma heterogeneidade entre os estratos. A retirada das amostras nos estratos é realizada através de uma amostra aleatória simples.

4.9.3 Amostragem Aleatória Estratificada (AAE)

A estratificação de uma população faz sentido quando é possível identificar subpopulações que variam muito entre si no que diz respeito à variável em estudo, mas que variam pouco dentro de si. Nestas condições, uma amostra estratificada pode fornecer resultados mais precisos do que uma amostra simples extraída do conjunto da população (COCHRAN, 1977).

Diante disso, como já exposto, a amostragem aleatória estratificada consiste em separar a população em subgrupos mais homogêneos (estratos). A ideia principal com isso é combater a variabilidade dos dados deixando a população em estratos mais homogêneos possíveis, pois se existe uma variância grande cresce a necessidade de uma amostra maior.

A AAE tem característica de dividir a população de forma homogênea e as vantagens de aplicá-la ficam evidentes e claras, sendo elas:

- Maior homogeneidade dentro de cada estrato, o que significa uma menor variância;
- Menor custo;
- Possibilidade de estimativas de parâmetros separados para cada estrato;
- Rapidez.

Como desvantagem podemos citar que os resultados obtidos estão sujeitos a uma margem de erro.

Em uma AAE com L estratos, temos que N_i o número de unidades amostrais no estrato i, e N o número de unidades amostrais na população. Então,

$$N = N_1+N_2+...+N_L, \text{ ou seja, } \quad N = \sum_{i=1}^{L} N_i$$

Para identificar os demais componentes da AAE pode-se colocar como:

n_i número de unidades da amostra;

x_i o valor obtido na i-ésima unidade;

W_i o chamado de peso do estrato, ou seja, a parte do estrato que deve ser alocada. Sua obtenção se dá pela razão $W_{i=} \frac{N_i}{N}$ (no caso específico do presente trabalho em que utilizou a alocação proporcional);

$\bar{x}_i$ é a média da amostra, obtida a partir de $\bar{x}_i = \frac{\sum_{i=1}^{n_i} x_i}{n_i}$;

$\bar{X}_{st}$ é a estimação média da população, advinda de $\bar{X}_{st} = \frac{\sum_{i=1}^{L} N_i \bar{x}_i}{N}$.

Nesse caso as médias e variâncias dos estratos (ou subpopulações) são designados por

$$\overline{X_1}, \dots, \overline{X_L}$$

e

$$\sigma_1^2, \dots, \sigma_L^2$$

4.9.3.1 Alocação Proporcional

Um problema na amostragem estratificada é determinar como dividir as n unidades da amostra total em cada estrato, de modo que numa amostragem aleatória simples a amostra é definida e retirada da população por meio de um sorteio em que todos tem a mesma probabilidade de ser escolhido. O tamanho da amostra é definido após o estabelecimento do erro amostral e do nível de confiança desejado. Já no caso da AAE ao subdividirmos a população em estratos não se tem apenas um único valor para o tamanho da amostra e sim um conjunto de valores determinados pela quantidade de estratos observados (COCHRAN, 1977).

O n é obtido através da fórmula (3), onde já conhecemos todos os termos exceto p que é a proporção estimada da populacional e q o seu complemento. B é o erro amostral.

$$n = \frac{\sum_{i=1}^{L} \frac{Ni^2 \widehat{q_i} \widehat{p_i}}{W_i}}{\frac{N^2 B^2}{Z_{\alpha/2}^2} + \sum_{i=1}^{L} N_i \widehat{\boldsymbol{p}}_i \, \widehat{\boldsymbol{q}}_i} \qquad (3)$$

Esta fórmula indica o número total da amostra que deve ser alocada para melhor representar a população em si. Após definido o número da amostra os estratos deverão ser montados atrelados ao W_i calculado.

5. METODOLOGIA

Este trabalho foi elaborado e caracterizado por ser uma pesquisa de campo na qual procura-se entender o conhecimento e a importância da população alvo sobre o seguro automotivo com aplicação de redes Bayesianas.

Para fundamentar e aplicar o tema estudado, foi realizada uma pesquisa bibliográfica em diversas fontes como livros, artigos, dissertações e sítios na internet que tratavam os assuntos direta ou indiretamente.

Os dados partiram de um projeto piloto aplicado diretamente a comunidade dos alunos do Centro de Ciências Exatas e Tecnologia (CCET) da Universidade Federal de Sergipe (UFS), através de um questionário com perguntas simples, não sendo necessário nenhum conhecimento específico prévio.

Após a consolidação preliminar do instrumento, este foi submetido a um pré-teste junto a uma amostra de 5 respondentes. Nesta etapa alguns ajustes foram procedidos para a adequação de dificuldades identificadas no questionário. Os alunos responderam diretamente sem intervenção do avaliador. O questionário consistiu em questões categóricas cujas informações foram agrupadas segundo cada item aplicado, podendo ser consultado no apêndice A.

Os entrevistados foram separados por estratos, sendo estes os Departamentos dos cursos de graduação ligados ao CCET, o que indica a utilização do método da Amostragem Aleatória Estratificada (AAE). Após a separação dos estratos, os dados foram coletados de maneira aleatória através de abordagem direta nas dependências da UFS, principalmente nos Departamentos, Centros Acadêmicos e preenchimento na internet através do Google Forms, com o link disponibilizado para os alunos no site do CCET, durante o mês de julho de 2019.

Com essa técnica de amostragem foi possível definir o tamanho da amostra, retirando do total da população dos alunos do CCET que no período 2019.1 era de 6205 discentes. Para determinar o tamanho da amostra foi definido o nível de significância de 10%, um erro amostral definido em 9% e uma proporção populacional de $\hat{p}$=50%, já que temos uma amostra piloto.

Após as definições das condições citadas, foi possível calcular o tamanho da amostra sendo utilizada como base a fórmula (3), encontrada no tópico 4.8.3.1 (Alocação Proporcional). Em seguida, definiu-se os tamanhos dos estratos também através da alocação proporcional utilizando a fórmula $W_{i=}\ \frac{N_i}{N}$ e $n_i=W_i*n$.

Desta forma, definidos os tamanhos da amostra e dos estratos os discentes foram entrevistados de acordo com os cursos conforme estão dispostos nos resultados. Essas entrevistas se deram de forma aleatória com visitas aos departamentos, abordagem em locais de estudo e centros acadêmicos.

Para a análise descritiva dos dados, foi realizada a verificação de todas as variáveis considerando a frequência, média e desvio de acordo com as características de cada uma, dando uma maior atenção as mais relevantes para o estudo.

As varáveis a serem testadas foram definidas como:

- SEXO (1=Masculino, 2=Feminino)
- IDADE (Classes em anos: 1=15-19; 2=20-24;3=25-29;4=30-34;5=35-40)
- DEPTO (Departamento ao qual o curso está atrelado)
- PERIODO (Qual período está cursando)
- TURNO (Turno o qual estuda)
- COR (Branco=1, Negro=2, Pardo =3, Índio=4)
- RENDA (Separados em 5 níveis, 1=até R$1.000,00, 2=R$1.000,00 a R$2.000,00, 3=R$ 2.000,00 a R$3.000,00, 4= R$3.000,00 a R$4.000,00, 5= Acima de R$4.000,00)
- CNH (possui CNH, 1=sim, 2=não)
- TEMPCNH (tempo que possui a CNH em anos)
- CONHECIMENTO (nota de 0 a 10 de conhecimento sobre seguros)
- PVEICULO (possui veículo. 1=sim, 2=não)
- TVEICULO (tipo de veículo que possui - 1=Moto, 2=Carro, 3=Motoneta, 4=Micro-ônibus,5=Outros)
- SVEICULO (possui seguro, 1=sim, 2=não)
- USOSERG (usou o seguro, 1=sim, 2=não)
- TSINISTRO (tipo de sinistro,1=Furto,2=Roubo, 3=Carroceria, 4= Outros)
- TEMPSEGURO (tempo que possui o seguro em meses)
- ANTSEGURO (já teve seguro antes, 1=sim, 2=não)
- MOTIVONTS (motivo de não ter seguro, 1=Acha que não é importante, 2=Acha caro, 3=Gostaria de ter, mas não tem condições, 4=Outros)
- IMPORTANCIA (importância ao seguro automotivo, nota de 0 a 10)

A partir dessas variáveis foi elaborado o processo através dos dados coletados no software R-Project, para verificar a existência de interação entre as variáveis por meio de seus níveis e força de relações probabilísticas.

A aprendizagem de estrutura em uma RB está atrelada a junção, retirada ou mudança no sentido das arestas que conectam cada nó, ou variável, da rede de modo que a RB final, de acordo com os parâmetros corretos, represente os dados da melhor maneira possível (JESUS, 2019).

Conforme utilizado por Jesus (2019), aplicou-se a técnica de mineração de regras de associação no software R, através do pacote *arules* para verificar a associação entre as variáveis para a formação da rede, definindo o valor do *support* em 50% e *confidence* em 90%. Com isso, somente foram considerados os valores em que as probabilidades das relações entre as variáveis estavam acima desses valores chegando no número de regras. Também foi utilizada a eliminação de regras redundantes e não significativas que foram avaliadas com o teste não paramétrico o exato de Fisher, que é utilizado para comparar dois grupos de duas amostras independentes (FONTELES, 2012). Isso resultou em regras e variáveis associadas para cada combinação de suporte (*support*) e confiança (*confidence*) para com isso realizar a montagem das redes Bayesianas. Notou-se que as regras de associação utilizadas para a verificação da ligação entre as variáveis tiveram inicialmente um total de 19. Ao variar o *confidence*, por exemplo, para 99% ou 10%, o número de regras diminui ou aumenta respectivamente. Como foi considerado um nível de significância nos dados coletados em 90%, as 19 regras foram analisadas verificando que 7 foram redundantes e poderiam ser eliminadas. Das 12 que restaram, ao realizar teste de significância percebeu-se que todas eram significativas.

Para a aprendizagem, ou montagem, da RB foram utilizados no software R os pacotes *bnlearn, gRain* e o *Rgraphviz,* sendo esse último para a geração de gráficos.

O *bnlearn* é um pacote da linguagem do software R cujo foco é o aprendizado de estrutura de RBs, estimação de parâmetros e realização de inferências. Os parâmetros utilizados na biblioteca *bnlearn* são *restart, perturb, whitelist* e *blacklist*. O parâmetro *restart* se refere ao número de vezes que o algoritmo vai reiniciar a busca a partir de um novo ponto de partida aleatório, a fim de que não ocorra o problema que a solução encontrada seja a ideal, e o parâmetro *perturb* explora a ideia de modificar brevemente os dados para proporcionar ao algoritmo a possibilidade de explorar outra região do conjunto de soluções. Os valores de *pertub* e *restart* foram alterados e essas variações fizeram com que fossem geradas várias novas redes. O *pertub* foi alterado de 1 a 50 e o

restart colocou-se com valores entre 0 e 1000. *Whitelist* e *blacklist*, são listas de arestas a serem incluídas ou recusadas no grafo resultante, respectivamente (JESUS, 2019). Essas foram utilizadas para acrescentar ou retirar variáveis na rede e verificar de acordo a formação de novas redes. Apesar de inicialmente adotar os valores padrões, pôde-se ver a variável IMPORTANCIA, que a princípio não foi relacionada na rede, posteriormente incluída através do *whitelist.*

Utilizou-se o algoritmo de pontuação Hill Climbing (HC) (ALIFERIS et all.,2006). Realizou-se combinações de parâmetros com o intuito de encontrar a rede com menor número do Critério de Informação Bayesiano, ou o *Bayesian Information Criterion* (BIC), para continuar com a montagem da rede e com isso aplicá-la (ALIFERIS et all.,2006). O BIC descreve a relação entre a variável dependente e as diversas variáveis explanatórias entre os diversos modelos sob seleção.

Encontradas com e sem o auxílio da mineração das regras de associação, as redes se apresentaram com 10 ou 11 variáveis. Escolhidas como as melhores com o algoritmo HC, analisou-se o quanto cada uma rede recuperaria as informações contidas na base de dados, através das probabilidades marginais. No entanto, na formação da rede utilizou-se todas as variáveis e não somente as que estavam presentes nas Regras de Associação, devido a ter encontrado relações que não condiziam com a realidade.

Em algumas das tabelas de probabilidade condicional pode-se observar que fazia parte delas um item chamado “Ausência”, que se define como a ausência de dados para aquela variável, pela não possibilidade de responder o item devido ao não preenchimento conforme opções do questionário.

6. RESULTADOS E DISCUSSÕES

A pesquisa foi baseada na AAE e chegou-se ao tamanho amostral total de 83 discentes. Conforme a tabela 2, tem-se os números dos tamanhos dos estratos que foram os departamentos ligados ao CCET da UFS no período 2019-1. Como foi utilizada alocação proporcional, os alunos foram separados por departamentos e assim formaram o número de entrevistados conforme pode ser observado na tabela 2.

Tabela 2 - Tabela com tamanho da amostra proporcional e separada por estratos

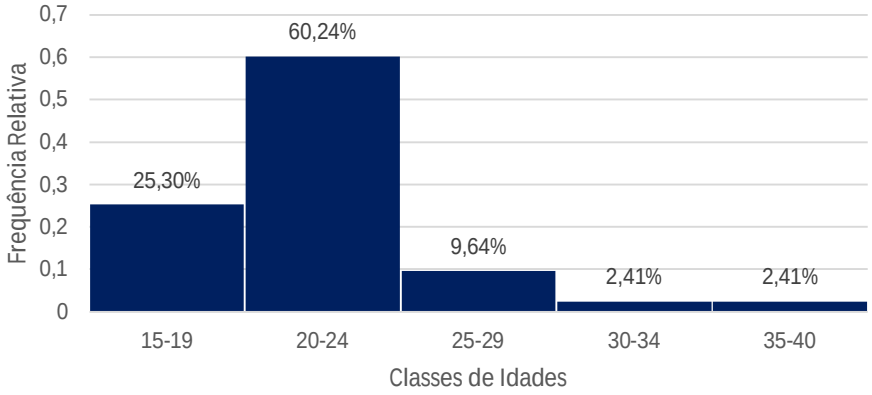
ESTRATOS	N
1-DEPARTAMENTO DE CIÊNCIA E ENGENHARIA DE MATERIAIS	2
2-DEPARTAMENTO DE COMPUTAÇÃO	13
3-DEPARTAMENTO DE ENGENHARIA AMBIENTAL	3
4-DEPARTAMENTO DE ENGENHARIA CIVIL	7
5-DEPARTAMENTO DE ENGENHARIA DE PRODUÇÃO	3
6-DEPARTAMENTO DE ENGENHARIA ELÉTRICA	7
7-DEPARTAMENTO DE ENGENHARIA MECÂNICA	4
8-DEPARTAMENTO DE ENGENHARIA QUÍMICA	6
9-DEPARTAMENTO DE ESTATÍSTICA E CIÊNCIAS ATUARIAIS	5
10-DEPARTAMENTO DE FÍSICA	11
11-DEPARTAMENTO DE GEOLOGIA	3
12-DEPARTAMENTO DE MATEMÁTICA	8
13-DEPARTAMENTO DE QUÍMICA	6
14-DEPARTAMENTO DE TECNOLOGIA DE ALIMENTOS	2
15-NÚCLEO DE GRADUAÇÃO EM ENGENHARIA DE PETRÓLEO	3
TOTAL	83

Fonte: Próprio autor

Os questionários foram aplicados em sua maioria na cidade universitária via entrevista direta aos discentes do CCET, com cerca de 90% do total (74 discentes). Os demais 10% foram coletados via questionário on-line (9 discentes). Observou-se que 81,2% dos entrevistados estudavam no turno diurno e somente 18,8% estudavam no turno noturno.

A maioria dos respondentes indicou ser jovem, ocorrendo uma variação mínima de 18 anos e máxima de 40 anos. Isso refletiu uma média de idade em torno de 22 anos, com desvio padrão de 0,81 e mediana em 21 anos. Nas médias das idades encontradas em cada estrato, observou-se uma variação com o mínimo de 20 e máximo de 26 anos. Desta forma, foi analisado que 60,2% dos entrevistados estavam no intervalo de 20 até 24 anos e as demais classes ficaram distribuídas conforme figura 18.

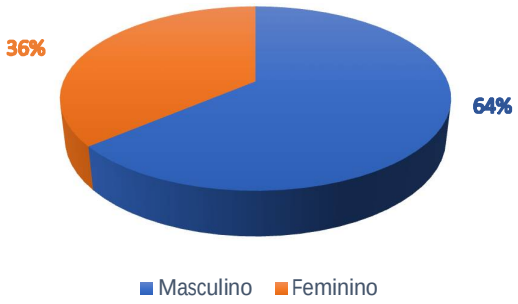
Figura 18 - Classes das idades da amostra dos alunos do CCET da UFS



Fonte: Próprio autor

Em relação ao gênero, obteve-se como resultado que 64% dos entrevistados eram do sexo masculino e 36% do sexo feminino, conforme figura 19. Esse resultado associado a importância dada ao seguro chegou a ser relevante no sentido que 40% dos entrevistados atribuíram nota 10, onde 25% destes eram do sexo masculino.

Figura 19 – Gênero dos entrevistados

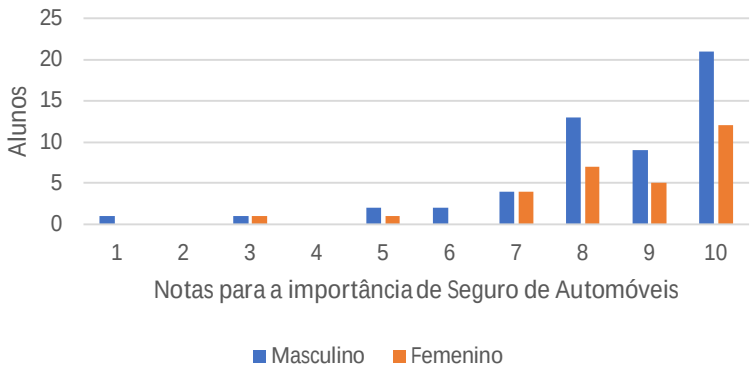


Fonte: Próprio autor

Diante das notas atribuídas a variável IMPORTANCIA (importância ao seguro automotivo) e separadas por sexo, ou seja, analisando cada um dos gêneros separadamente, verificou-se que nas médias encontradas das notas atribuídas a esta variável, que ficou em torno de 8,506 para cada uma, não foi possível observar diferença nos resultados obtidos entre elas. Existiu uma diferença na quantidade de respondentes por sexo e uma concentração nas notas mais altas atribuídas por ambos (figura 20). Utilizando o teste estatístico T-student para comparar as médias calculadas da variável IMPORTANCIA por sexo, com nível de confiança em 90%, foi possível verificar se existia igualdade entre elas. Observou-se que estatisticamente falando não houve

diferenças significativas entre os valores (médias) das notas dadas por homens e mulheres. Com isso, neste estudo, as notas altas atribuídas foram dadas independentemente do sexo e a avaliação da importância ao seguro veicular pôde ser considerada próxima ou igual entre eles, ou seja, independentemente de ser homem ou mulher o seguro foi considerado importante.

Figura 20 - Gráfico sexo e importância a seguros

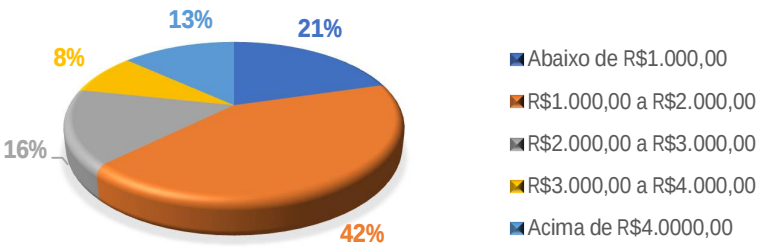


Fonte: Próprio autor

No tocante dos períodos em que os discentes estavam matriculados foi observado uma variedade grande tendo certa concentração tanto nos primeiros quanto nos últimos períodos (ver gráfico no apêndice C).

A característica da variável RENDA tem um aspecto em que a faixa que obteve a maior participação dos respondentes foi a segunda, na qual a renda na residência do indivíduo estava entre R$ 1000,00 e R$ 2000,00, como ilustrado no gráfico da figura 21. As demais faixas tiveram uma participação menor, mas não menos importante, pois cerca de 37% dos entrevistados estavam na faixa dos que tinham uma renda igual ou superior a R$ 2.000,00.

Figura 21 – Distribuição da Renda dos Discentes Entrevistados.



Fonte: Próprio autor

Após a análises descritivas partiu-se para a análise das redes Bayesianas (RBs), onde se obteve diversas destas (redes) que poderiam explanar a relação de cada variável e se chegar ao objeto de interesse deste trabalho. Em alguns casos a ordem das variáveis não explanou a sequência natural de uma relação de pais e filhos ficando fora da realidade, sendo essa a necessidade do pesquisador para verificar a correta orientação e relação entre elas, inclusive nas redes que foram formadas com a utilização das regras de associação de mineração de dados. As variáveis selecionadas com a utilização das regras associação de mineração não formaram redes condizentes com uma relação causal entre elas, não sendo assim aproveitadas. Com isso, todas variáveis foram incluídas para a solução dada pelo algoritmo na geração das redes.

Nas diversas redes alternativas geradas (alguns exemplos no apêndice D), as variáveis não se alteraram em relação a RB dada como definitiva que fora escolhida devido a possuir o menor valor de BIC, que ficou em torno de -1770,37, e as associações corretas condizentes com a realidade do estudo analisado. Portanto, a escolha da rede foi possível devido a intervenção do pesquisador que evitou a presença de associações indevidas entre as variáveis, como por exemplo, uma rede onde a variável MOTIVONTS (motivo não ter seguro) ser pai da variável USOSERG (usou o seguro), pois esta foi a rede com menor BIC encontrada, mas com o detalhe da associação lhe invalidando.

Nestas condições as associações entre as variáveis foram geradas resultando na rede ilustrada na figura 22, onde CONHECIMENTO e IMPORTÂNCIA, e a variável IDADE não fizeram parte no resultado encontrado.

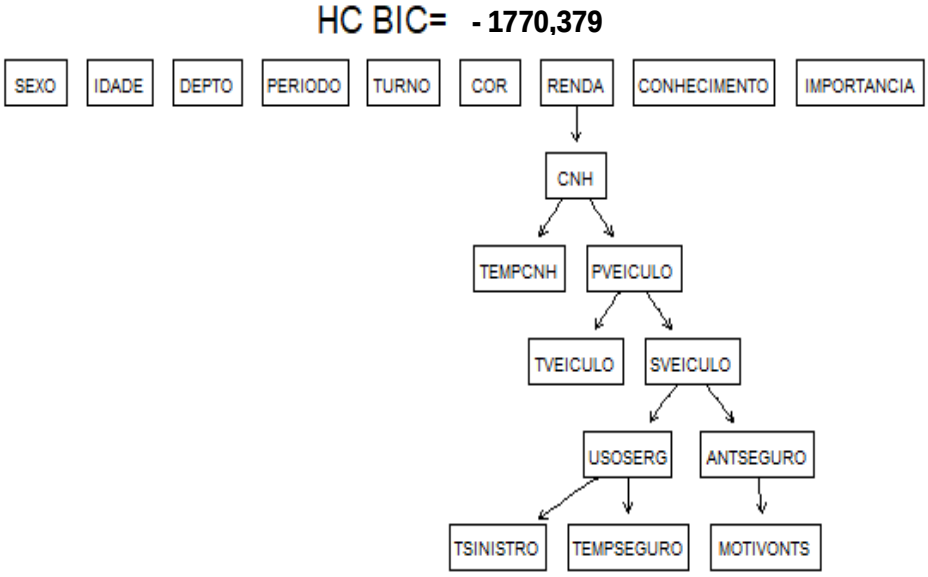
Vê-se um grafo acíclico e direcionado explanado de forma hierarquizada onde as arestas direcionadas caracterizam-se como as relações que representam as dependências estatísticas ou de casualidade entre as variáveis, ou seja, links direcionados que são usados para indicar as relações de parentesco pai e filho (figura 22).

Devido a formatação da rede as hipóteses a serem questionadas diante das evidências já encontradas, terão sempre como resultados os valores de probabilidade que estão nas tabelas de probabilidade condicional, sendo visível a característica da cadeia de Markov.

Na rede escolhida pôde-se notar que as varáveis pais encontradas condizem em influência e justifica estabelecer relações de causa e efeito nas evidências encontradas. Como a variável RENDA é pai da variável CNH, o que se explica devido ao poder aquisitivo necessário para obtê-la, podendo ser claramente aceita essa associação de causa e efeito. De acordo com os dados observados e inferindo na RB escolhida, foi de 90% dos

que tinham renda superior aos R$2.000,00 possuíam CNH. A Variável SVEICULO (possui seguro) tem como variável pai PVEICULO (possui veículo) e como filhas USOSERG (usou o seguro) e ANTSEGURO (já possuiu contrato de seguro antes).

Figura 22 – Rede Bayesiana Gerada Pelo Software R Com a Amostra e BIC



Fonte: Próprio autor

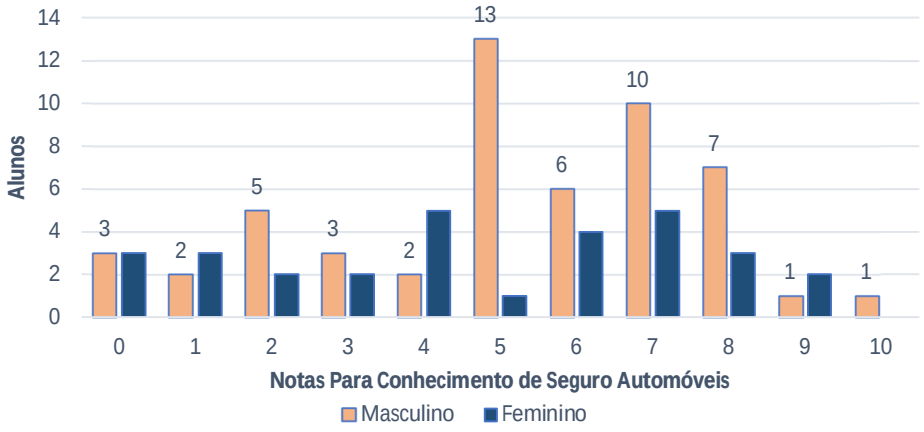
Outro aspecto analisado foram as suposições de independência, pois as variáveis TEMPCNH e PVEICULO são independentes condicionalmente considerando a evidência CNH. Encontra-se independência condicional entre causas e evidências dados os nós intermediários como, por exemplo, com PVEICULO conhecido, então SVEICULO é independente de CNH. Assim, P(SVEICULO|PVEICULO, CNH) = P(SVEICULO|PVEICULO) que ficou em 62%.

Independente das variáveis em análise não estarem compostas na rede, os comandos gerados através do algoritmo HC revelaram as tabelas de todas as variáveis da amostra, o que pode ser comparado e comprovado com alguns resultados fazendo a análise de acordo com a técnica estatística sobre a amostra estratificada exemplificada no início dessa sessão.

Analisando as variáveis “Conhecimento” e “Importância” ao seguro automotivo, notou-se que o conhecimento sobre seguros é considerado relevante entre os respondentes, pois 58% do total deram notas entre 5 e 8. As notas entre 0 e 1 representaram 35% do total, o que gerou uma média baixa e desvio padrão considerável para o conhecimento.

Agora fazendo uma análise em relação as notas dadas a variável CONHECIMENTO separadas por sexo (figura 23), observou-se uma relação de semelhança entre elas com uma variação também semelhante, diferenciada apenas pela quantidade de respondentes, como ocorreu na variável IMPORTANCIA.

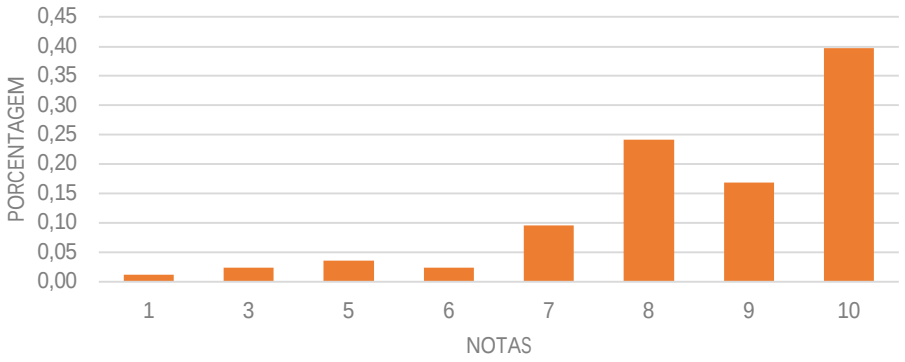
Figura 23 - Gráfico Sexo e Conhecimento sobre Seguros



Fonte: Próprio autor

No caso da importância dada ao seguro veicular, 91% do total dos entrevistados atribuíram nota entre 7 e 10, sendo que a nota 10 ficou com 40% do total (figura 24). A média estimada da amostra estratificada obtida para a importância, já citada, ficou mais alta que a do conhecimento que ficou em torno de 5 de pontos. Com isso, um desvio de 0,1889, e uma mediana de 9, observou-se que a importância dada ao seguro veicular pode ser considerada alta levando em consideração a média e mediana da amostra, mesmo com um desvio considerável que demonstra a variação nas notas atribuídas.

Figura 24 – Probabilidade da Variável "IMPORTANCIA" ao seguro automotivo



Fonte: Próprio autor

As tabelas de probabilidades condicionais (TPC) de cada variável foram geradas para então poder verificar os resultados de probabilidades com as associações geradas na rede.

Os resultados nas tabelas da figura 25 mostram as possíveis chances de ocorrer determinado evento, pois ao se obter a RB e as tabelas de probabilidades condicionais de cada variável, pode-se inferir sobre a mesma e assim concluir as hipóteses sobre as variáveis criadas através da amostra obtida. Como podemos ver a probabilidade de o respondente ter CHN dado que ele possui uma renda superior a R$ 4.000,00 é de 91%. Entretanto para renda abaixo de R$1.000,00 fica em torno de 40%. Estes valores chegam a ser relevantes, pois à medida que o indivíduo tem uma maior renda a aquisição de bens passa a ser considerada. Outra associação relevante seria o fato que a probabilidade de se ter veículo dado que possua CNH (P(PVEICULO|CNH)) é de 84% e para quem não possui veículo nem CNH fica em torno de 60%.

Variáveis como IDADE, COR, SEXO e DEPTO (departamento) que teoricamente poderiam estar na rede não obtiveram relação, de acordo com os critérios dos comandos e algoritmos utilizados. Em testes de correlação entre essas variáveis e as escolhidas na rede, a variável RENDA por exemplo, apresentaram fraca correlação.

Diante deste fato foram realizadas inclusões e retiradas de algumas variáveis na formação da rede, sendo constatado que as variáveis que não fazem parte não alteram as probabilidades na inferência.

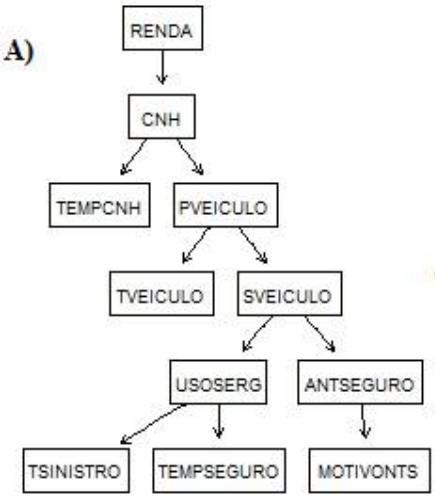
Analisando as variáveis nas tabelas que geraram a ausência de dados, em alguns casos como por exemplo na figura 25-E, a variável TEMPCNH, que é a probabilidade de tempo de carteira em anos, dado que não possui a mesma, é um evento certo para os que responderam “não” neste item, pois se não tem CNH não existe tempo de que a possui. Igualmente na probabilidade de o veículo ter seguro dado que o entrevistado não o possui. Portanto, em todas as tabelas de probabilidades condicionais da rede que apresentaram essa característica estão relacionadas a ausência de dados referentes aos itens que não possuem respostas, por motivo da sequência do questionário.

Tabelas relacionadas ao tempo tem cada uma das suas particularidades, sendo as variáveis TEMPCNH e TEMPSEGURO compostas em anos e meses respectivamente. Com isso, na figura 25-E verificou-se que 73% era a probabilidade dos entrevistados que possuíam CNH estarem no primeiro ano de curso.

Figura 25 - Rede Bayesiana e suas Tabelas de Probabilidade das Variáveis.

(continua)

A)



B)

RENDA	P(RENDA)
Abaixo R$1.000,00	0,20
R$ 1.000,00 a R$ 2.000,00	0,42
R$ 2.000,00 a R$ 3.000,00	0,16
R$3.000,00 a R$4.000,00	0,08
Acima de R$ 4.000,00	0,13

C)

RENDA	P(CNH|RENDA)	
	S	N
P(Abaixo R$1.000,00)	0,41	0,59
P(R$ 1.000,00 a R$ 2.000,00)	0,74	0,26
P(R$ 2.000,00 a R$ 3.000,00)	1,00	0,00
P(R$3.000,00 a R$4.000,00)	1,00	0,00
P(Acima de R$ 4.000,00)	0,91	0,09

C)

CNH	P(PVEICULO|CNH)	
	S	N
S	0,84	0,16
N	0,40	0,60

D)

	P(TEMPCNH | CNH)			
CNH	1	2	3	Ausência
SIM	0,73	0,16	0,11	0
NÃO	0	0	0	1

E)

	P(TVEICULO|PVEICULO)			
PVEICULO	Moto	Carro	M Onibus	99
S	0,25	0,72	0,03	0,00
N	0,00	0,05	0,00	0,95

F)

	P(SVEICULO|PVEICULO)		
PVEICULO	S	N	Ausência
S	0,62	0,36	0,02
N	0,00	0,00	1,00

G)

	P(USOSERG|SVEICULO)		
SVEICULO	S	N	Ausência
1	0,37	0,63	0,00
2	0,00	0,05	0,95
Ausência	0,00	0,00	1,00

H)

	P(ANTSEGURO|SVEICULO)		
SVEICULO	S	N	Ausência
S	0,00	0,05	0,95
N	0,09	0,91	0,00
Ausência	0,00	0,04	0,96

I)

	P(TSINISTRO|USOSERG)			
USOSERG	Roubo	Carroceria	Outros	Ausência
S	0,14	0,50	0,36	0,00
N	0,00	0,00	0,04	0,96
Ausência	0	0	0	1,00

J)

	P(TEMPSEGURO|USOSERG)										
USOSERG	0	1	2	3	4	5	6	7	10	11	Ausência
S	0,21	0,00	0,21	0,07	0,07	0,14	0,07	0,07	0,00	0,07	0,07
N	0,28	0,12	0,16	0,20	0,08	0,08	0,00	0,00	0,04	0,00	0,04
Ausência	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	1,00

Figura 25 - Rede Bayesiana e suas Tabelas de Probabilidade das Variáveis

(conclusão)

K)

	P(MOTIVONTS|ANTSEGURO)				
ANTSEGURO	**Não Importante**	**Caro**	**Sem Condições**	**Outros**	**Ausência**
S	**0,00**	**0,00**	**0,00**	**1,00**	**0**
N	**0,13**	**0,30**	**0,30**	**0,22**	**0,04**
Ausência	**0,00**	**0,00**	**0,00**	**0,00**	**1**

Fonte: Próprio autor

As tabelas elucidam a possibilidade de se obter respostas para as hipóteses apresentadas associadas a uma RB, como por exemplo a probabilidade de possuir um veículo dado que possui CNH ou a de possuir um veículo dado que sua renda seja acima de R$ 4.000,00.

No trabalho em questão as hipóteses, a princípio, seriam relacionadas com as variáveis IMPORTANCIA e CONHECIMENTO, porém com os resultados obtidos e aplicada a técnica da RB, a variável RENDA pode explanar mais a importância que o conhecimento do seguro no cotidiano das pessoas. Isso se torna claro ao se analisar a variável MOTIVONTS (motivo não ter seguro), pois mesmo não tendo contrato de seguro num momento anterior (ANTSEGURO= “N”), a probabilidade dos que acham que não é importante ficou em torno de 13%, e a dos que não possuem por acharem caro é 30%. Ainda na mesma análise os que gostariam de ter o seguro veicular, mas não tem condições, a probabilidade também ficou em 30%. Estas questões tem a renda como principal indicador de sua justificativa. Portanto, a análise da variável MOTIVONTS revela indícios que pode gerar nas companhias de seguro um interesse de se definir qual a probabilidade das pessoas não obterem seguro por acharem caro, para com isso melhorar e observar o que pode ser feito para atrair esses que não possuem seguro de automóveis. Com as evidências de não possuir seguro anteriormente, possuir veículo, CNH e renda entre R$2.000,00 e R$3.000,00 a probabilidade resulta:

P(MOTIVONTS|ANTSEGURO, SVEICULO, PVEICULO, CNH, RENDA) = 0,30.

Ainda sobre a variável RENDA, ela é a que inicia a rede e assim a análise. Portanto, pode-se questionar: Qual a probabilidade, posteriori, de um veículo possuir seguro dado que a renda está entre R$ 2.000,00 e R$ 3.000,00, possuir veículo e CNH? Como se pode observar na figura 25-G, essa probabilidade se dá pela tabela da probabilidade de o veículo ter seguro, dado que possui veículo, ficou em torno de 62%.

Em relação a variável de TSINISTRO (tipo do sinistro), verificou-se que a opção de furto não foi citada por nenhum dos respondentes, apesar que em dados do anuário das estatísticas de segurança pública indicar um aumento deste tipo sinistro em Sergipe de 2014 a 2018. A probabilidade de ocorrer sinistros na Carroceria foi o mais destacado entre elas como pode ser observado na tabela deste item figura 25-L. Vê-se que a P(TSINISTRO, Carroceria|USOSERG) ficou em torno de 50%. Mesmo com o item furto não sendo citado, o item roubo ficou com uma probabilidade de ocorrer em 14%.

Em posse da RB o uso da variável USOSERG pode revelar aproximadamente o número de sinistros através da inferência sobre a probabilidade de ter usado o seguro dado que o veículo possui seguro. Com isso, pode-se chegar à probabilidade de ocorrência dos sinistros.

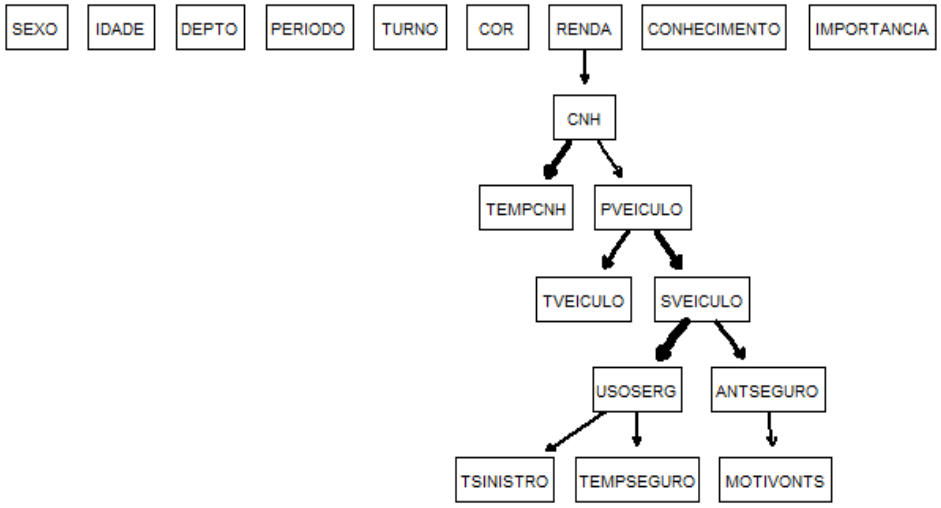
Santos (2015) coloca que a teoria de grafos é utilizada para criar uma estrutura qualitativa de um modelo e a teoria da probabilidade é usada para caracterizar a natureza e força das relações do modelo. Neste contexto, de acordo com a Tabela 3, a força das relações probabilísticas expressa pelos arcos mostra que a de maior intensidade está entre as variáveis SVEICULO e USOSERG (figura 26). Com esta característica na montagem da rede, só os arcos significativos foram utilizados. Apesar do sinal negativo, a força entre as variáveis na tabela fica mais bem ilustrada na figura 26, que representa a rede com os arcos indicando a força das relações probabilísticas.

Tabela 3 - Força das Relações Probabilísticas Expressas Pelos Arcos

Item	De	Para	Força
1	SVEICULO	USOSERG	-44,926905
2	PVEICULO	SVEICULO	-39,465257
3	PVEICULO	TVEICULO	-36,573802
4	ANTSEGURO	MOTIVONTS	-30,823150
5	SVEICULO	ANTSEGURO	-30,267008
6	USOSERG	TSINISTRO	-22,731951
7	USOSERG	TEMPSEGURO	-10,764664
8	CNH	TEMPCNH	-6,062072
9	CNH	PVEICULO	-4,761712
10	RENDA	CNH	-2,173903

Fonte: Próprio autor

Figura 26- Gráfico da Força das Relações Probabilísticas Expressas Pelos Arcos

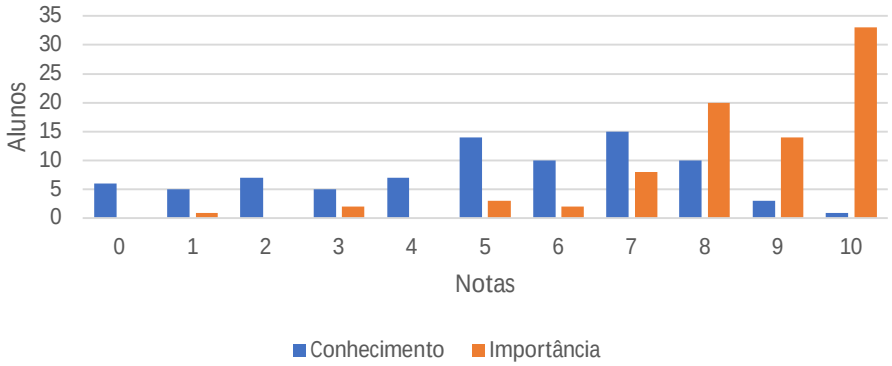


Fonte: Próprio autor

Todos os tipos de seguros têm a sua característica específica para manter o segurado ou seus beneficiários protegidos de eventuais concretizações dos riscos. O seguro de veículos automotores é extremamente importante para a prevenção de prejuízos provenientes de riscos futuros e incertos. No Brasil, por exemplo, as estatísticas de acidentes no trânsito são bem alarmantes e de certa forma favoráveis a aquisição deste tipo de seguro, devido ao alto índice de mortalidade, colisões e furtos, porém apenas 30% da frota o possui (FENSEG, 2019). A existência de um seguro obrigatório de Danos Pessoais Causados por Veículos Automotores de Vias Terrestres, ou por sua Carga, a pessoas transportadas ou não (DPVAT) tem o intuito de auxiliar, principalmente, as famílias de baixa renda após acidentes, porém não prevê riscos de danos.

O seguro de automóveis de certa forma não é uma unanimidade, mas é considerado importante no tocante da aquisição de bens e manutenção do patrimônio caso ocorra o sinistro, pois o item “acha que não é importante” nos motivos para não ter seguros foi o que teve o menor número de respostas dos que possuem veículo e não tinham seguro. Isso pode ser observado nos resultados obtidos que apesar de um conhecimento mediano e desvio de 2,5, as notas dadas a importância para seguro de automóveis para os indivíduos foram consideradas significativas (figura 27), apesar de a maior parte dos respondentes terem menos que 25 anos.

Figura 27 – Gráfico Notas para Importância e Conhecimento de Seguros dos discentes do CCET/UFS

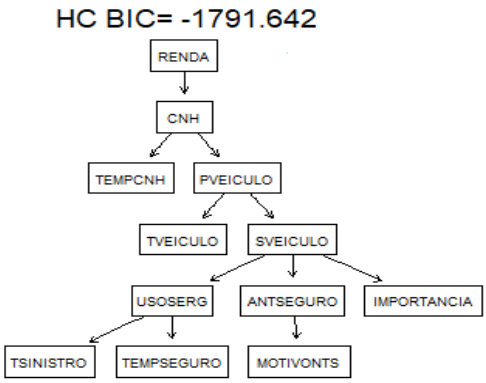


Fonte: Próprio autor

A rede resultante deixou as variáveis (importância e conhecimento) de fora, de forma a não gerar associação entre elas, apesar de ter montado a rede sem relações que não condizem com a realidade do estudo analisado. Isso resultou numa aceitação da rede como verdadeira, sendo escolhida a com o menor BIC gerado apesar das diversas variações encontradas verificadas na ordem das variáveis.

Na RB as variáveis analisadas trouxeram a renda dos entrevistados como a principal evidência para a aquisição ou a dificuldade em obter um contrato de seguro. Porém incluindo a variável IMPORTANCIA, por meio do comando *whitelist* (Figura 28), pôde-se observar a possibilidade dessa operação ser feita e externar a dependência condicional dessa variável. A tabela de probabilidade condicional, agora relacionada a variável SVEICULO, foi alterada acrescentou-se uma nova condição. Analisando a força da relação probabilística dessa nova variável na rede, observou-se ser a mais fraca devido a inserção via o parâmetro *whitelist*.

Figura 28 – Rede com a variável IMPORTANCIA



Fonte: Próprio autor

Os números encontrados mostram que se pode aplicar ações para uma maior disseminação que venha a aumentar o entendimento sobre seguros, para dele verificar a real necessidade de se obtê-lo apesar das barreiras que são impostas pela renda. As novas gerações exigem mudanças no comportamento dos seguros e podem causar mais impactos na sua composição, porém sempre mantendo a solvência das operadoras de seguro (SUSEP, 2006).

7. CONCLUSÕES

No estudo realizado foi montada uma rede bayesiana entre variáveis criadas para analisar o comportamento dos discentes do CCET relacionado a importância e o conhecimento da atividade de seguro veicular. O modelo apresentado como resultado, apesar da não utilização do gerado através das regras de associação de mineração de dados, alocou somente a variável renda (RENDA) como evidência principal e mostrou que as variáveis carteira nacional de habilitação (CNH), possui veículo (PVEICULO), veículo possui seguro (SVEICULO), usou ou acionou o Seguro (USOSERG), motivo não ter seguro (MOTIVONTS) foram as outras em que suas probabilidades a posteriori explanaram indiretamente a importância do seguro de veículos automotores.

As alterações nos itens do algoritmo HC na geração da rede (*restart* e *pertub*), não modificaram as variáveis elencadas, mas sim a ordem delas em relação a variável raiz. Ao verificar todas essas relações entre as probabilidades produzidas pela rede, devido ao escopo formado, qualquer que viesse a ser questionada estaria sendo obtida diretamente condicionada a sua variável pai, portanto diretamente na tabela de probabilidade condicional.

Os resultados obtidos atrelados ao conhecimento e importância dada ao seguro de veículo automotivo dos alunos do CCET da UFS, ficou evidente que o entendimento não se revelou muito significativo e pode ser aprimorado através de campanhas por parte das seguradoras e um aumento por parte dos órgãos governamentais responsáveis devido a este último já realiza-las. A característica jovem da amostra, como também outros fatores, pode ter dado sua participação na influência do conhecimento mediano dos alunos apesar que a maior parte deles possuíam CNH.

No tocante das altas notas dadas a importância ao seguro de automóveis, estas possibilitaram explanar, apesar do conhecimento não ser considerado alto, que os jovens discentes acham importante ter um contrato de seguro, e pôde elucidar um entendimento para que as companhias seguradoras tenham uma melhor política de abordagem e orientação ao público em geral, principalmente para os iniciantes no mercado sobre seus produtos e como funcionam, visto que de forma geral a renda dos pesquisados influencia na consolidação desses contratos. As startups chamadas Insurtechs estão no mercado para facilitar essa ligação entre as seguradoras, de todos os tipos de seguros, e os jovens.

Com isso, o resultado obtido em relação a RB, a variável que indica o motivo para não ter seguro está atrelada a variável RENDA, pois os avaliados que não o possuem, acham caro, mas gostariam de ter apesar de não terem condições financeiras para isso.

Contudo, o trabalho apresentou uma ligação em que os entrevistados entendem em sua maioria o que é o seguro de automóveis e dá a ele uma importância significativa mesmo que alguns não tenham condições financeiras de tê-lo. A RB aparece como uma possibilidade para análise de risco através de análise com as probabilidades das variáveis em estudo para ajudar a explorar esse mercado que está de forma geral subaproveitado.

REFERÊNCIAS

ALIFERIS, Constantin; TSAMARDINOS, Ioannis; BROWN, Laura. **The Max-Min Hill-Climbing Bayesian Network Structure Learning Algorithm. Machine Learning**. 2006.

BOAVENTURA NETTO, Paulo Oswaldo; JURKIEWICZ, Samuel. **Grafos: Introdução e prática.** São Paulo: Blucher, 2017. Segunda edição.

BRASIL. Presidência da República. Secretaria-Geral. Subchefia para Assuntos Jurídicos. Medida Provisória nº 904, de 11 de novembro de 2019, que dispõe sobre a extinção do Seguro Obrigatório de Danos Pessoais causados por Veículos Automotores de Vias Terrestres - DPVAT e do Seguro Obrigatório de Danos Pessoais Causados por Embarcações ou por suas Cargas - DPEM. **Portal da Legislação,** Brasília dez. 2019. Disponível em: <http://www.planalto.gov.br/ccivil_03/_ato2019-2022/2019/Mpv/mpv904impressao.htm> . Acesso em 1/03/2020.

CASTRO, Pablo A. de; ZUBEN, Fernando J. Von. **Raciocínio Probabilístico.** Campinas, SP, 2008.

CARVALHO, C. L. de; VASCONCELOS, L. M. R. de. **Aplicação de Regras de Associação para Mineração de Dados na Web.** Goiás, 2004. Disponível em: <http://ww2.inf.ufg.br/sites/default/files/uploads/relatorios-tecnicos/RT-INF_004-04.pdf>. Acessado em 10/02/2020.

COCHRAN, W. G. **Sampling Techniques.** 3º edition. Wiley. 1977.

CONRADY, Stefan; JOUFFE Lionel. **Bayesian Networks and BayesiaLab - A Practical Introduction for Researchers**. Bayesia USA. 2015.

DOMINGUEZ, A.; PITA, R. **Seguro de Automóvel.** Escola Nacional de Seguros, Rio de Janeiro: FUNENSEG. RJ. 2011. 224p.

EMILIANO, P. C. **Fundamentos e Aplicações dos Critérios de Informação: Akaike e Bayesiano.** Dissertação (mestrado) Universidade Federal de Lavras. Lavras, MG, 2009

FAVARO, Flavia F. **A Teoria dos Grafos e sua Abordagem na Sala de Aula com Recursos Educacionais Digitais**. Dissertação (mestrado) - Universidade Estadual Paulista, Instituto de Geociências e Ciências Exatas, Rio Claro, SP. 2017.

FENABRAVE. **Anuário, o Desempenho da Distribuição Automotiva no Brasil.** 2018**.**

FERREIRA, Paulo Pereira. **Modelos de Precificação e Ruína para Seguros de Curto Prazo.** 2ª reimpressão. Rio de Janeiro: FUNENSEG, 2010. 224 p.

FENSEG. **O Setor de Seguros Brasileiro.** Rio de Janeiro, 2019. Disponível em: <http://cnseg.org.br/publicacoes/o-setor-de-seguros-brasileiro.html>. Acessado em 15/02/2020.

FILHO, Domingos Afonso Kriger. **O Contrato de Seguro no Direito Brasileiro**. 1. ed. RJ: Labor Juris, 2000.

FONTELES, Mauro José. **Bioestatística aplicada à pesquisa experimental**. Volume 2. São Paulo: Editora Livraria da Física, 2012.

FUNENSEG. Diretoria de Ensino Técnico. **Teoria geral do seguro I /Supervisão e Coordenação metodológica da Diretoria de Ensino Técnico.** Assessoria técnica de Marco Aurélio de Paiva Fonseca. – 12. ed. – Rio de Janeiro: FUNENSEG, 2013.

JEQUESSENE, Plácido Mateus. **Modelos De Grafos em Estatística**. Dissertação (mestrado) Universidade Federal do Rio de Janeiro. Rio de Janeiro, RJ, 2010.

JESUS, Elielma Santana de. **Apoio De Regras De Descoberta De Associação Na Elaboração De Redes Bayesianas Aplicadas Ao Sistema De Notificações De Doenças Vesiculares No Brasil.** Dissertação (pós-graduação) Universidade Federal Rural de Pernambuco. Recife, PE, 2019.

KARCHER, Cristiane. **Redes Bayesianas Aplicadas à Análise do Risco de Crédito.** Dissertação (Mestrado) Escola Politécnica da Universidade de São Paulo. SP, 2009.

LADEIRA, M; VICARI, Rosa Maria; COELHO, Helder. **Redes Bayesianas Multiagentes.** Rio de Janeiro, 1999.

LUCCAS FILHO, Olívio. **Seguros: Fundamentos; Formação de Preço provisões e Funções Biométricas.** São Paulo, Atlas 2011.

MANICA, Lais. **O Contrato do Seguro de Vida.** Dissertação (Graduação) Universidade Federal do Rio Grande do Sul, Faculdade de Ciências Jurídicas e Sociais. Porto Alegre, RS, 2010.

MARGARITIS, Dimitris. **Learning Bayesian Network Model Structure from Data**. Tese (Doutorado em Filosofia) Universidade de Pittsburgh. Pittsburgh. PA. 2003

MARTINS, Guilherme Nunes; JUSTO, Wellington Ribeiro; PEREIRA, Wolney, **Estimação Do Risco Moral No Mercado De Seguros De Automóveis Do Estado De Pernambuco.** Revista Economia e Desenvolvimento, n. 20, 2008. Disponível em <https://periodicos.ufsm.br/eed/article/view/3467>. Acesso em 20/02/2019.

MORENTTIN, Luiz Gonzaga. **Estatística Básica: probabilidade e inferência.** Volume único**.** São Paulo: Pearson Prentice Hall**,** 2010.

PRESTES, E. **Introdução à Teoria dos Grafos**. Universidade Federal do Rio Grande do Sul. Porto Alegre: 2016.

ORLANDELI, Rogério. **Um Modelo Markoviano-Bayesiano de Inteligência Artificial para Avaliação Dinâmica do Aprendizado: Aplicação À Logística**. Tese (Doutorado em Engenharia) Universidade de Santa Catarina, Florianópolis, SC. 2005.

RODRIGUES, José Ângelo. Gestão de Risco Atuarial. São Paulo: Ed. Saraiva, 2008.

RUSSELL S.J.; NORVIG, P**. Inteligência Artificial**. Editora Campus, 2004.

SANTOS, Alberto M. O. **Aplicação de Redes Bayesianas na Ciência Forense.** Dissertação (Mestrado) Universidade de Lisboa, Faculdade de Ciências, Lisboa, 2015.

SOBERANIS, I. **An Extended Bayesian Network Approach for analyzing Supply Chain Disruptions.** College The University of Iowa, Iowa City, 2010.

SOUZA, Anderson Luiz Ara. **Redes Bayesianas: Uma Introdução Aplicada A Credit Scoring.** São Carlos, SP. 2010.

SUSEP. **DPVAT.** Rio de Janeiro, 2019. Disponível em : <http://www.susep.gov.br/menu/informacoes-ao-publico/planos-e-produtos/seguros/dpvat> Acesso em 20/12/2019

SUSEP. **Guia de orientação e defesa do segurado** / Superintendência de Seguros Privados. 2ed. Rio de Janeiro: SUSEP, 2006. <https://www2.susep.gov.br/download/cartilha/cartilha_susep2e.pdf>. Acessado em 15/02/2020.

TUDO SOBRE SEGUROS. **Tipos de Cobertura – Automóveis.** São Paulo, 2019. Disponível em: <https://www.tudosobreseguros.org.br/tss-individuo-automoveis-tipo-de-cobertura/>. Acesso em 22/02/2019.

APENDICE A

QUESTIONÁRIO

UNIVERSIDADE FEDERAL DE SERGIPE
CENTRO DE CIÊNCIAS EXATAS E TECNOLOGIA
DEPARTAMENTO DE ESTATÍSTICA E CIÊNCIAS ATUARIAIS
PROF.ª: AMANDA LIRA

TEMA: Seguro de Veículos Automotores

O presente questionário visa verificar o perfil dos alunos vinculados ao CCET em relação ao seu conhecimento sobre seguro de veículos automotores e se o utilizam.

Dados Pessoais:

1- Sexo? 1- ( ) Masculino 2- ( ) Feminino
2- Data de nascimento? _____/_____/_____
3- Qual o Departamento do curso que está fazendo? _____ (Ver lista no verso e marque).
4- Em qual período você está?
1-( ) 2-( ) 3-( ) 4-( ) 5-( ) 6-( ) 7-( ) 8-( ) 9-( ) 10-( )
5- Qual o turno você estuda? 1- ( ) Diurno 2- ( ) Noturno
6 – Como você considera sua Cor ou Raça?
1- ( ) Branco 2- ( )Negro 3- ( ) Índio 4- ( )Pardo
7- Qual a renda familiar da sua residência? 1-( ) Abaixo de R$1.000,00 2-( ) De R$ 1000,00 a R$ 2.000,00 3-( ) De R$ 2.000,00 a R$ 3.000,00 4-( ) De 3.000,00 a R$ 4.000,00 5-( ) Acima de R$4.000,00
8-Você ou alguém da família possui Carteira Nacional de Habilitação-CNH?
1- ( ) Sim 2- ( ) Não, vá para o item 10
9- Há quanto tempo, aproximadamente em meses ou anos, possui a CNH? (caso você não tenha pode ser um familiar na residência).
_____Anos ou ____ Meses

Avaliações:

10- Em relação ao seu entendimento sobre seguros, qual nota de 0 a 10 você daria?
0-( ) 1-( ) 2-( ) 3-( ) 4-( ) 5-( ) 6-( ) 7-( ) 8-( ) 9-( ) 10-( )
11- Onde você reside, algum familiar ou você possui veículo motor?
1- ( ) Sim 2- ( ) Não, vá para o item 19
12- Qual o tipo de Veículo?
1- ( ) Moto 2- ( ) Carro 3-( ) Motoneta 4-( ) Micro-ônibus 5- ( ) Outros
13- O veículo em questão tem seguro?
1- ( ) Sim 2- ( ) Não, vá para o item 17
14 - Alguma vez utilizou o seguro?
1- ( ) Sim 2- ( ) Não, vá para o item 16
15- Qual foi o tipo de avaria?
1- ( ) Furto 2- Roubo ( ) 3-( ) Carroceria(Portas, para-choque, etc.) 4- ( ) Outros
16- Há quanto tempo aproximadamente o veículo possui o seguro? Após resposta vá para o item 19.
_____Anos ou _____Meses
17- Se o veículo não tem seguro, tem ciência que já teve algum contrato antes?
1- ( ) Sim 2- ( ) Não
18- Por qual motivo não possui seguro?
1- ( ) Acha que não é importante 2-( ) Acha caro 3-( ) Gostaria de ter, mas não tem condições 4-( ) Outros
19- Atribua uma nota de 0 a 10 de como você vê a importância do seguro de veículos?
0-( ) 1-( ) 2-( ) 3-( ) 4-( ) 5-( ) 6-( ) 7-( ) 8-( ) 9-( ) 10-( )

CENTRO DE CIÊNCIAS EXATAS E TECNOLOGIA - CCET

- [ ] 1- DEPARTAMENTO DE CIÊNCIA E ENGENHARIA DE MATERIAIS
- [ ] 2- DEPARTAMENTO DE COMPUTAÇÃO
- [ ] 3- DEPARTAMENTO DE ENGENHARIA AMBIENTAL
- [ ] 4- DEPARTAMENTO DE ENGENHARIA CIVIL
- [ ] 5- DEPARTAMENTO DE ENGENHARIA DE PRODUÇÃO
- [ ] 6- DEPARTAMENTO DE ENGENHARIA ELÉTRICA
- [ ] 7- DEPARTAMENTO DE ENGENHARIA MECÂNICA
- [ ] 8- DEPARTAMENTO DE ENGENHARIA QUÍMICA
- [ ] 9- DEPARTAMENTO DE ESTATÍSTICA E CIÊNCIAS ATUARIAIS
- [ ] 10- DEPARTAMENTO DE FÍSICA
- [ ] 11- DEPARTAMENTO DE GEOLOGIA
- [ ] 12- DEPARTAMENTO DE MATEMÁTICA
- [ ] 13- DEPARTAMENTO DE QUÍMICA
- [ ] 14- DEPARTAMENTO DE TECNOLOGIA DE ALIMENTOS
- [ ] NÚCLEO DE GRADUAÇÃO EM ENGENHARIA DE PETRÓLEO

APENDICE B

Comandos das Análise no R

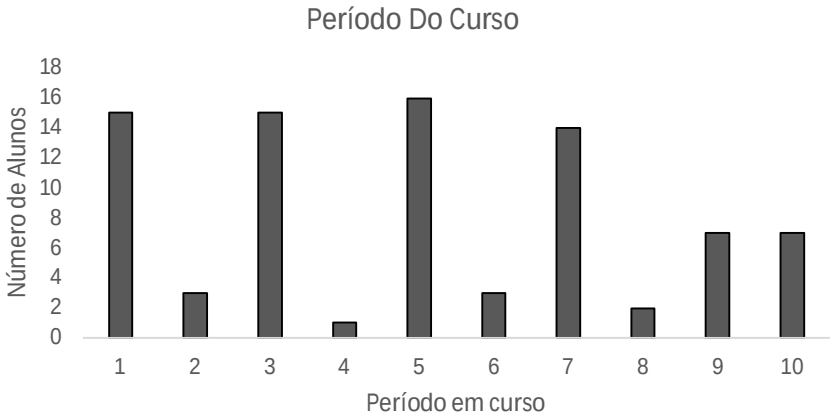
```
#Instalando Componentes
if (!requireNamespace("BiocManager", quietly = TRUE))
 install.packages("BiocManager")
BiocManager::install("Rgraphviz")
library(bnlearn)
library(Rgraphviz)
library(gRain)
library(dplyr)
library(readxl)
library(forecast)
library(lmtest)
library(arules)
library(arulesViz)

#Importando dados
AA<-read_excel(file.choose(), col_names = TRUE)
class(AA)
AA1<-as.data.frame(AA)
attach(AA1)

#Passando para fatores
AA1$IMPORTANCIA<- as.factor(IMPORTANCIA)
AA1$RENDA<- as.factor(RENDA)
AA1$CONHECIMENTO<- as.factor(CONHECIMENTO)
AA1$SEXO<- as.factor(SEXO)
AA1$IDADE<- as.factor(IDADE)
AA1$DEPTO<- as.factor(DEPTO)
AA1$PERIODO<- as.factor(PERIODO)
AA1$TURNO<- as.factor(TURNO)
AA1$COR<- as.factor(COR)
AA1$CNH<- as.factor(CNH)
AA1$TEMPCNH<- as.factor(TEMPCNH)
AA1$PVEICULO<- as.factor(PVEICULO)
AA1$TVEICULO<- as.factor(TVEICULO)
AA1$SVEICULO<- as.factor(SVEICULO)
AA1$USOSERG<- as.factor(USOSERG)
AA1$TSINISTRO<- as.factor(TSINISTRO)
AA1$TEMPSEGURO<- as.factor(TEMPSEGURO)
AA1$ANTSEGURO<- as.factor(ANTSEGURO)
AA1$MOTIVONTS<- as.factor(MOTIVONTS)

#Criação das regras
at = as(AA1, "transactions")
regras <- apriori(at, parameter = list(support = 0.5, confidence = 0.90))
summary(regras)
regras <- sort(regras, by ="lift")
```

```
inspect(regras, n=3)
is.redundant(regras, measure="lift" )
regras[is.redundant(regras)]    #Regras redundantes
regras[!is.redundant(regras)]   #Regras não redundantes
inspect(head(sort(regras, by="lift"), n=3))
regras <- regras[!is.redundant(regras)]  #Remover regras redundantes
inspect(regras)

#Testar significância
is.significant(regras,at, method="fisher", alpha=0.1)
regras[is.significant(regras, at)]      #Regras significaivas
regras[!is.significant(regras, at)]    #Regras não significativas
inspect(regras[is.significant(regras, at)])
regras <- regras[is.significant(regras,at)]

#Gerar Rede
hc <- hc(AA1, start = NULL, whitelist = NULL, blacklist = NULL, score = NULL,
       debug = FALSE, restart = 0, perturb = 0, max.iter = Inf, maxp = Inf,
       optimized = TRUE)
bnlearn::score(hc, data = AA1, type ="bic")
graphviz.plot(hc, main = "HC BIC= -1778.468", shape = "rectangle")
plot(hc)
arc.strength(hc,AA1)
strength<-arc.strength(hc,AA1)
strength.plot(hc,strength, shape="rectangle")

#layout= "dots", "neato", "twopi", "circo", "fdp"
#shape= "circle", "ellipse", "rectangle"

#Inferencia#
TPC<-bn.fit (hc, AA1)
cpquery(TPC, event=(PVEICULO =="1"), evidence = ( CNH=="2" ))
cpquery(TPC, event=(TSINISTRO=="4"), evidence = (USOSERG=="1" & CNH=="3"
       & PVEICULO=="1" & SVEICULO=="1"))
cpquery(TPC ,event=(MOTIVONTS=="1"), evidence=(RENDA=="2" & CNH=="1" &
       PVEICULO=="1" & SVEICULO=="2" & ANTSEGURO=="2"))
barplot(table(RENDA))
barplot(table(PVEICULO))
table(PVEICULO)
table(MOTIVONTS)
table(RENDA)
barplot(table(MOTIVONTS))
barplot(prop.table(table(SEXO, CONHECIMENTO)))
plot(CONHECIMENTO)
prop.table(table(CONHECIMENTO))
```

APENDICE C

Estatísticas Descritivas de algumas Variáveis

Figura 1. Gráfico de Alunos por Períodos



Fonte: Próprio autor

Figura 2. Arrecadação do mercado de seguros regulado.

Arrecadação do mercado de seguros regulado pela Susep em abril de 2019				
	abril 2019 (R$ milhões)	Var. abr. 2019 / mar. 2019	Var. 2019 até abr./ 2018 até abr.	Var. 12 m. até abr.2019 / 12 m. até abr.2018
Ramos Elementares (sem DPVAT)	**6.003,7**	**-2,8%**	**7,0%**	**7,3%**
Automóvel	2.992,1	7,1%	-0,4%	2,5%
Patrimonial	1.349,3	2,6%	13,9%	11,6%
Habitacional	334,8	-1,9%	2,4%	-3,2%
Transportes	312,2	9,3%	10,8%	13,6%
Crédito e Garantia	369,6	-50,8%	38,4%	19,6%
Responsabilidade Civil	145,8	-1,5%	22,6%	18,6%
Rural	408,4	20,5%	6,5%	10,7%
Marítimos e Aeronáuticos	66,5	-21,5%	52,5%	19,1%
Satel./ Nucl./ Petróleo	24,9	-78,5%	1,1%	108,5%
Coberturas de Pessoas	**13.266,0**	**7,0%**	**-14,9%**	**-8,9%**
Planos de Risco de Seguros de Pessoas	**3.728,6**	**14,3%**	**14,8%**	**11,0%**
Vida e Acidentes pessoais	1.874,1	9,9%	8,8%	8,0%
Prestamista	1.165,3	6,0%	26,7%	21,3%
Outros	689,3	51,2%	13,6%	3,9%
Planos de acumulação*	**9.537,3**	**4,4%**	**-24,6%**	**-14,8%**
Capitalização	**2.026,3**	**7,1%**	**10,5%**	**2,3%**
Mercado Segurador (sem DPVAT)	**21.295,9**	**4,0%**	**-6,5%**	**-3,5%**
DPVAT	165,0	-15,2%	-52,5%	-34,3%
Mercado Segurador (Total)	**21.460,9**	**3,9%**	**-7,8%**	**-4,1%**

Fonte: SUSEP. Ramos Elementares

Tabela 1- Descritivas variável TVEICULO

Tipo Sinistro	Freq Absoluta	FreqRelativa	FreqRelativa(%)
Furto	0	0	0%
Roubo	2	0,024096386	2%
Carroceria	7	0,084337349	8%
Outros	5	0,060240964	6%
Ausência	69	0,831325301	83%

Fonte: Próprio autor

Tabela 2- Descritivas variável SVEICULO

SVEICULO	Freq Absoluta	FreqRelativa	FreqRelativa(%)
Sim	38	0,457831325	46%
Não	22	0,265060241	27%
Não Possui	23	0,277108434	28%
Total	83	1	1

Fonte: Próprio autor

Tabela 3- Descritivas variáveis COR

COR OU RAÇA	Freq Absoluta	FreqRelativa	FreqRelativa(%)
Branco	15	0,180722892	18%
Negro	15	0,180722892	18%
Pardo	53	0,638554217	64%
Total	83	1	100%

Fonte: Próprio autor

Tabela 4- Descritivas variável TURNO

TURNO	Freq Absoluta	FreqRelativa	FreqRelativa(%)
Diurno	67	0,807228916	81%
Noturno	16	0,192771084	19%
Total	83	1	100%

Fonte: Próprio autor

Tabela 5- Descritivas variável CNH

CNH	Freq Absoluta	FreqRelativa	FreqRelativa(%)
Sim	63	0,759036145	76%
Não	20	0,240963855	24%
Total	83	1	100%

Fonte: Próprio autor

Tabela 6- Descritivas variável TEMPCNH

TEMPCNH	Freq Absoluta	FreqRelativa	FreqRelativa(%)
Anos			
1	47	0,56626506	57%
2	9	0,108433735	11%
3	6	0,072289157	7%
>3	1	0,012048193	1%
Não Possui CNH	20	0,240963855	24%
Total	83	1	100%

Fonte: Próprio autor

Tabela 7- Descritivas da variável CONHECIMENTO

Notas	Freq Absoluta	XY	Freq Relativa	FreqRelativa(%)
0	6	0	0,072289157	7%
1	5	5	0,060240964	6%
2	7	14	0,084337349	8%
3	5	15	0,060240964	6%
4	7	28	0,084337349	8%
5	14	70	0,168674699	17%
6	10	60	0,120481928	12%
7	15	105	0,180722892	18%
8	10	80	0,120481928	12%
9	3	27	0,036144578	4%
10	1	10	0,012048193	1%
Total	83	414	1	100%
Media	4,988			
Desvio Padrão	2,597			

Fonte: Próprio autor

Tabela 8- Descritivas variável IMPORTANCIA

Notas	Freq Absoluta	XY	Freq Relativa	FreqRelativa(%)
0	0	0	0	0%
1	1	1	0,012048193	1%
2	0	0	0	0%
3	2	6	0,024096386	2%
4	0	0	0	0%
5	3	15	0,036144578	4%
6	2	12	0,024096386	2%
7	8	56	0,096385542	10%
8	20	160	0,240963855	24%
9	14	126	0,168674699	17%
10	33	330	0,397590361	40%
Total	83	706	1	100%
Media	8,506			
Desvio Padrão	1,804			

Fonte: Próprio autor

Tabela 9- Descritivas variável PVEICULO

PVEICULO	Freq Absoluta	FreqRelativa	FreqRelativa(%)
Sim	61	0,734939759	73%
Não	22	0,265060241	27%
Total	83	1	100%

Fonte: Próprio autor

Tabela 10- Descritivas variável USOSERG

USOSERG	Freq Absoluta	FreqRelativa	FreqRelativa(%)
Sim	14	0,168674699	17%
Não	25	0,301204819	30%
Não Possui	44	0,530120482	53%
Total	83	1	100%

Fonte: Próprio autor

Tabela 11- Descritivas variável TSINISTRO

Tipo Sinistro	Freq Absoluta	FreqRelativa	FreqRelativa(%)
Furto	0	0	0%
Roubo	2	0,024096386	2%
Carroceria	7	0,084337349	8%
Outros	5	0,060240964	6%
Ausência	69	0,831325301	83%
Total	83	1	100%

Fonte: Próprio autor

Tabela 12- Descritivas variável MOTIVONTS

MOTIVO	Freq Absoluta	FreqRelativa	FreqRelativa(%)
Não Importante	3	0,036144578	4%
Caro	7	0,084337349	8%
Sem Condições	7	0,084337349	8%
Outros	7	0,084337349	8%
Ausência	59	0,710843373	71%
Total	83	1	100%

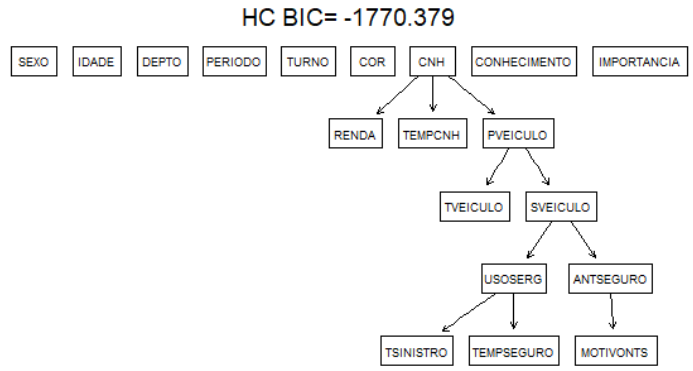
Fonte: Próprio autor

APÊNDICE D

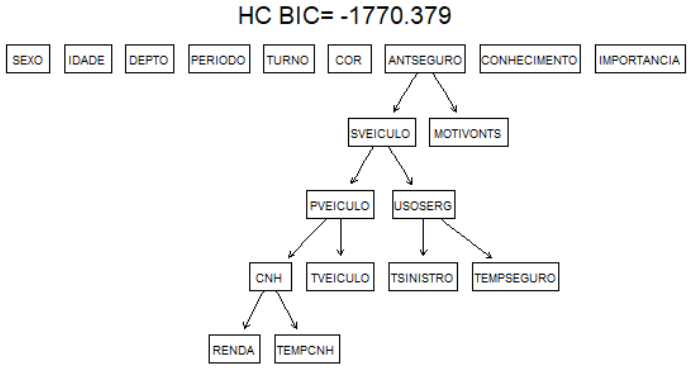
Redes Não Aceitas por BIC ou Variáveis de Forma Adequadas

Figura 1. Redes Geradas e Não Aceitas

A)



B)



C)