UNIVERSIDADE DE SÃO PAULO
FACULDADE DE ECONOMIA, ADMINISTRAÇÃO E CONTABILIDADE DE
RIBEIRÃO PRETO
DEPARTAMENTO DE ADMINISTRAÇÃO
PROGRAMA DE PÓS-GRADUAÇÃO EM ADMINISTRAÇÃO DE ORGANIZAÇÕES

LUÍS FELIPE BARBOSA FERNANDES

Aplicação de Redes Bayesianas em modelos de classificação de risco de crédito

Orientador: Prof. Dr. Evandro Marcos
Saidel Ribeiro

RIBEIRÃO PRETO
2019

Prof. Dr. Vahan Agopyan
Reitor da Universidade de São Paulo

Prof. Dr. André Lucirton Costa
Diretor da Faculdade de Economia, Administração e Contabilidade de Ribeirão Preto

Prof. Dr. Jorge Henrique Caldeira de Oliveira
Chefe do Departamento de Administração

LUÍS FELIPE BARBOSA FERNANDES

Aplicação de Redes Bayesianas em modelos de classificação de risco de crédito

Dissertação apresentada ao Programa de Pós-Graduação em Administração de Organizações da Faculdade de Economia, Administração e Contabilidade de Ribeirão Preto da Universidade de São Paulo, para obtenção do título de Mestre em Ciências. *Versão Corrigida. A original encontra-se disponível na FEA-RP/USP.*

Orientador: Prof. Dr. Evandro Marcos
Saidel Ribeiro

RIBEIRÃO PRETO

2019

Autorizo a reprodução e divulgação total ou parcial deste trabalho, por qualquer meio convencional ou eletrônico, para fins de estudo e pesquisa, desde que citada a fonte.

Fernandes, Luís Felipe Barbosa

Aplicação de Redes Bayesianas em modelos de classificação de risco de crédito / Luís Felipe Barbosa Fernandes – Ribeirão Preto, 2019.

88f.: il.; 30 cm

Dissertação apresentada ao Programa de Pós-Graduação em Administração de Organizações da Faculdade de Economia, Administração e Contabilidade de Ribeirão Preto da Universidade de São Paulo, para obtenção do título de Mestre em Ciências. *Versão Corrigida. A original encontra-se disponível na FEA-RP/USP.* – Universidade de São Paulo

Orientador: Ribeiro, Evandro Marcos Saidel

1. Pontuação de Crédito. 2. Redes Bayesianas. 3. Modelagem de crédito. 4. Risco de Crédito.

LUÍS FELIPE BARBOSA FERNANDES

Aplicação de Redes Bayesianas em modelos de classificação de risco de crédito

Dissertação apresentada ao Programa de Pós-Graduação em Administração de Organizações da Faculdade de Economia, Administração e Contabilidade de Ribeirão Preto da Universidade de São Paulo, para obtenção do título de Mestre em Ciências. *Versão Corrigida. A original encontra-se disponível na FEA-RP/USP.*

Área de Concentração: Administração de Organizações

Data de Aprovação:

____/____/____

Banca Examinadora:

Prof. Dr. Evandro Marcos Saidel Ribeiro
Orientador

Professor
Avaliador 1

Professor
Avaliador 2

Professor
Avaliador 3

Aos meus pais, Luís Augusto e Marisa, que foram
essenciais durante mais essa jornada acadêmica.

Agradecimentos

Agradeço a todos que, direta ou indiretamente, contribuíram para a realização desse trabalho. Aos meus pais, Luís Augusto Fernandes e Marisa Barbosa Fernandes, agradeço pelo apoio na decisão de realizar mais essa etapa acadêmica, em todo suporte e compreensão durante esses dois anos.

Ao meu orientador, Evandro Marcos Saidel Ribeiro, pelos conhecimentos compartilhados, pelas oportunidades criadas durante esse período e pela colaboração durante as atividades de pesquisa.

Ao professor Gilberto Aparecido Prata, por ter me levado a conhecer a ferramenta Latex.

A instituição SERASA S.A. inscrita no CNPJ sob o nº 62.173.620-0001/80, com sede na Alameda dos Quinimuras, 187 – Planalto Paulista – São Paulo – SP e a FUNDAÇÃO PARA PESQUISA E DESENVOLVIMENTO DA ADMINISTRAÇÃO, CONTABILIDADE E ECONOMIA, designada FUNDACE, com sede à Rua Bernardino de Campos, 1001 – sala 401- Higienópolis – Ribeirão Preto – SP, inscrita no CNPJ/MF sob o nº 00.934.542/0001-31, pelo apoio financeiro recebido.

Ao colega analista de sistemas, André Pignata, por compartilhar seus conhecimentos em programação e contribuir para o desenvolvimento do aplicativo de cálculo do risco de crédito.

Aos professores da Universidade de São Paulo que contribuíram para o meu desenvolvimento profissional e pessoal durante esses anos, ressaltando a importância da dedicação e persistência em todos os campos da vida.

Agradeço também a Deus, por ter me concedido mais uma oportunidade de crescimento profissional e proporcionado condições para que eu chegasse até aqui.

“In God we trust, all others bring data”

William Edwards Deming

RESUMO

FERNANDES, Luís Felipe Barbosa. *Aplicação de Redes Bayesianas em modelos de classificação de risco de crédito*. 2019. 88f. Mestrado em Administração de Organizações – Faculdade de Economia, Administração e Contabilidade de Ribeirão Preto, Universidade de São Paulo, Ribeirão Preto, 2019.

A demanda pelo estudo e aprimoramento de modelos de crédito que auxiliem na tomada de decisões, relativas a concessão creditícia, cresce de forma acelerada. Frente às dificuldades de ordem financeira que atingem os mais diversos países, incluindo o Brasil, verifica-se uma crescente preocupação dos órgãos reguladores do mercado financeiro, bem como, das próprias instituições credoras que atuam no mercado por modelos de crédito. A dificuldade para a obtenção de informações que reflitam a saúde financeira das empresas – assimetria informacional – aliada à carência de informações no mercado financeiro contribui para o aumento dos casos de *default* e empresas que decretam concordata. Em face dos problemas e dificuldades apresentados, a pesquisa empregou o método probabilístico de Redes Bayesianas com o objetivo de desenvolver um modelo de crédito que calcule o risco de crédito de uma empresa baseado apenas em um conjunto de indicadores financeiros, obtidos a partir das demonstrações financeiras dessas empresas. Para isso, foi usado um conjunto de demonstrações financeiras, referentes a um total de 852 empresas com faturamento superior à 200 milhões, cedidas pela instituição Serasa Experian. A partir dessas demonstrações foram implementadas as fórmulas usadas pela Serasa Experian para o cálculo de indicadores financeiros, a partir dos quais a Rede Bayesiana inicial foi formada. A técnica de Redes foi implementada através do algoritmo denominado Algoritmo Pc, que combina elementos de grafos probabilísticos e definições de probabilidades condicionais, para a selecionar as variáveis, representadas pelos indicadores financeiros, mais significantes para o cálculo do risco de crédito. Além disso, foi realizada uma comparação da técnica probabilística de Redes Bayesianas com a técnica de Regressão Logística, para verificar qual dos modelos melhor se adequava ao conjunto de dados. Após implementar a técnica, foi desenvolvido também um aplicativo, que calcula o risco de crédito de uma empresa, a partir de um conjunto de 17 indicadores financeiros e exibe ao usuário final, a classe de risco, dentre cada uma das 13 classes possíveis, a que uma empresa possui maior probabilidade de pertencer. Para validar a técnica de Redes Bayesianas foram empregadas duas métricas: a RMSE(Raiz Quadrada do Erro Médio) e o MAE(Erro Absoluto Médio). As métricas mostraram que o modelo de Redes Bayesianas foi pouco preditivo, com resultados aquém do esperado. Os resultados da técnica de Regressão Logística porém, mostram um percentual de acertos muito superior, classificando um percentual de 82% das empresas classificadas como de risco de crédito baixo, de forma correta.

Palavras-chave: Pontuação de Crédito, Redes Bayesianas, Modelagem de Crédito, Risco de Crédito.

ABSTRACT

FERNANDES, Luís Felipe Barbosa. *Application of Bayesian Networks in models of classification of credit risk*. 2019. 88f. Mestrado em Administração de Organizações – Faculdade de Economia, Administração e Contabilidade de Ribeirão Preto, Universidade de São Paulo, Ribeirão Preto, 2019.

The demand for studies and enhancement of credit models that helps at the decision making, associated with the granting credit, grows in a high speed. In the face of the recent troubles of financial order that accomplish innumerous countries nowadays, including Brasil, financial authorities have shown an increasing concern, as well as, the financial institutions that plays at the market for credit models. The challenge of search for informations that shows the financial health's companies – information asymmetry – together with the lack of data at the financial market contribute to increase the number of default cases and number of companies that fails. Due to the issues and difficulties described, this research used the probabilistic approach of Bayesian Networks to develop a credit model capable of calculate the credit risk of a company based on a set of financial indexes, obtained by the financial statements of these companies. For that, it was used a set of financial statements, regarding a set of 852 companies with revenues higher than 200 hundred billion reais, obtained through an agreement with the institution Serasa Experian. These financial statements were used to calculate the financial indexes through the formulae adopted by Serasa Experian, which gave the inicial set of the Bayesian Network. The Network technique was used through an algorithm called Pc Algorithm, that mix elements of probabilistic graphs with conditional's probability definitions, to select variables, represented by financials indexes, that are more significant to the calculation of credit's risk. Besides that, it was made a comparison between Bayesian Network and Logistic Regression technique, with the purpose of verify which one was the best to this set of variables. After the technique was implemented, it was also developed an application, capable of calculate the credit risk of a corporation, using a dataset of seventeen financial indexes. As a result, the app shows to the final user which of the thirteen risk's classes, has the biggest chance of being associated with the enterprise. To validate the technique it were employed two measurements, the RMSE(root mean square error) and the MAE(mean absolute error). The measurements showed that the Bayesian Networks model was not very predictive to the sample of companies which it was trained, since the outcomes fell short of expectations. On the other hand, the Logistic Regression technique showed better results when compared with the Bayesian Network technique. The percentage of right risk's class classifications were much higher, resulting at a percentage of 82% of companies classified as "low risk" , in the right way.

Keywords: Credit Score, Bayesian Networks, Credit Modeling, Credit Risk.

LISTA DE ILUSTRAÇÕES

Figura 1 – Modelo KMV de probabilidade de <i>default</i>	26
Figura 2 – Representação de um Grafo	32
Figura 3 – Tipos de Arestas	33
Figura 4 – Representação de um DAG e um PDAG	34
Figura 5 – Representação do modelo Naïve-Bayes	35
Figura 6 – Elementos de um Grafo probabilístico	36
Figura 7 – Rede Bayesiana do Risco Operacional	39
Figura 8 – Rede Bayesiana do Risco Operacional com CPDs	40
Figura 9 – Atribuição de Evidências na Rede do Risco Operacional	41
Figura 10 – Diagrama de caixa para os 20 indicadores	49
Figura 11 – Gráfico de dispersão da correlação entre 19 indicadores e a variável risco	51
Figura 12 – Correlação entre os 19 indicadores e a variável risco	51
Figura 13 – Rede Bayesiana do Exame Clínico	52
Figura 14 – Rede Bayesiana Inicial	56
Figura 15 – Rede obtida através do algoritmo Greedy Search	58
Figura 16 – Rede obtida através do algoritmo Hybrid	59
Figura 17 – Esqueleto da Rede Bayesiana	60
Figura 18 – Esqueleto da Rede Bayesiana com direções	61
Figura 19 – Curva ROC	65
Figura 20 – Cross-validation pelo método k-fold	68
Figura 21 – Rede após a implementação da cobertura de Markov	70
Figura 22 – Print de Tela da guia Estrutura de Capitais	77
Figura 23 – Print de Tela da guia Indicadores de Liquidez	78
Figura 24 – Print de Tela da guia Risco Calculado	79
Figura 25 – Print de Tela com valores atribuídos	80
Figura 26 – Resultado após atribuição de valores na rede	81

LISTA DE TABELAS

Tabela 1 – Classes de Risco e probabilidade de inadimplência associada às classes de risco.	24
Tabela 2 – Valores e Abreviações para a Rede Bayesiana do Risco Operacional . .	38
Tabela 3 – Atribuição de Evidências	38
Tabela 4 – Distribuição de empresas por segmentos de indústrias	43
Tabela 5 – Classificação das variáveis	44
Tabela 6 – Indicadores financeiros: abreviações e significados.	46
Tabela 7 – Resumo das estatísticas descritivas dos 20 indicadores financeiros para o ano de 2013	49
Tabela 8 – Base de dados chestSim500	54
Tabela 9 – Exemplo de atribuição de evidências	54
Tabela 10 – Atribuição de probabilidades pela Rede Bayesiana	55
Tabela 11 – Classes de Risco adotadas para as variáveis	55
Tabela 12 – Resultados da Regressão Logística	64
Tabela 13 – Definições para classificação de modelos que produzem resultados dicotômicos	65
Tabela 14 – Matriz de confusão	66
Tabela 15 – Teste de Hosmer e Lemeshow	66
Tabela 16 – Métricas para o processo de validação cruzada	68
Tabela 17 – Matriz de confusão para a técnica de Redes Bayesianas	69
Tabela 18 – Matriz de confusão após implementação da cobertura de Markov . . .	71
Tabela 19 – Siglas e significados	73
Tabela 20 – Classes de Valores dos Indicadores	73
Tabela 21 – Valores dos indicadores	75

SUMÁRIO

	1 INTRODUÇÃO	14
1.1	Trabalhos relacionados	15
1.2	Problema de Pesquisa	19
1.3	Objetivos	19
1.3.1	Geral	19
1.3.2	Específicos	19
1.4	Justificativa	19
1.5	Estrutura do trabalho	20
	2 REVISÃO BIBLIOGRÁFICA	21
2.1	Risco de Crédito	21
2.2	<i>Credit Scoring</i>	22
2.3	Modelos de Crédito	24
	3 CONCEITOS PROBABILÍSTICOS	28
3.1	Espaço Amostral e Eventos	29
3.2	Variáveis Aleatórias	29
3.3	Probabilidade Condicional	30
3.4	Teorema Bayesiano	31
	4 GRAFOS PROBABILÍSTICOS	32
4.1	Vértices e arestas	32
4.2	Classificação das arestas	33
4.3	Grafos Acíclicos Dirigidos (DAGs)	33
4.4	Grafos Acíclicos parcialmente dirigidos (PDAG)	33
4.5	Redes Bayesianas	34
4.5.1	Modelo Naïve-Bayes	34
4.5.2	Elementos Básicos	35
4.5.3	Definição de Redes Bayesianas	36
4.5.4	Exemplo de Rede Bayesiana	37
	5 BASE DE DADOS	42
5.1	Classificação das Variáveis	44
5.2	Indicadores financeiros selecionados	45
	6 ANÁLISE DOS DADOS	48
6.1	Padronização e categorização dos dados	48
6.2	Correlação	50
6.3	Correlação entre 20 indicadores	50

6.4	Categorização	52
6.4.1	Exemplo do Exame Clínico	52
6.4.2	Criação de categorias de dados	55
6.4.3	Esboço da Rede Bayesiana inicial	56
	7 MODELAGEM	57
7.1	Greedy Search	57
7.2	Algoritmo Híbrido	58
7.3	Algoritmo PC	59
7.3.1	Primeira etapa	59
7.3.2	Segunda etapa	61
	8 REGRESSÃO LOGÍSTICA	62
8.1	Modelo de regressão logística múltipla	62
8.2	Aplicação dos dados	62
8.3	Coeficientes da regressão	63
8.4	Receiver Operating Characteristic	64
8.5	Teste de Hosmer e Lemeshow	66
	9 VALIDAÇÃO DO MODELO	67
9.1	Resultados	68
9.2	Implementação da cobertura de Markov (Markov Blanket)	69
	10 O PACOTE SHINY	72
10.1	Dados	72
10.2	Tratamento dos dados	73
10.3	Descrição do aplicativo	74
10.4	Simulação com o aplicativo Shiny	75
	11 CONSIDERAÇÕES FINAIS	82
	REFERÊNCIAS	84

Capítulo 1

Introdução

O crédito, que hoje compõe uma das atividades centrais de bancos e outras instituições financeiras, tem suas origens muito remotas. Os primeiros registros de um empréstimo remontam à 2000 a.C. e são atribuídos ao povo babilônico. Naquela época, o dinheiro foi emprestado sob a condição de que o mesmo fosse devolvido quando o devedor obtivesse seu próximo rendimento, neste caso, representado pela próxima colheita (THOMAS; EDELMAN; CROOK, 2017).

Embora a atividade de crédito tenha se iniciado há tanto tempo, as técnicas para avaliar o risco intrínseco a essa atividade surgiram apenas na década de 30. A primeira iniciativa nesse sentido é atribuída à Fisher, no campo da estatística, em 1936. Fisher definiu uma forma de diferenciar variedades de uma mesma espécie de plantas usando as suas características físicas. Poucos anos depois, em 1941, David Durand notou que a técnica proposta por Fisher poderia ser aplicada à atividade de crédito. A técnica mostrou-se útil para classificar o conjunto de empréstimos cedidos por um credor em dois tipos: os empréstimos bons e os empréstimos ruins (THOMAS, 2000).

As primeiras tentativas de estabelecer um modelo que pontuasse os consumidores segundo a sua qualidade de crédito são atribuídas à Bill Fair e Earl Isaac. Na década de 50, esses profissionais foram os responsáveis pela expansão dos modelos de pontuação de crédito ao desenvolver a FICO, uma das empresas pioneiras no desenvolvimento de sistemas de (ABRAMOWICZ; MAREK; SZTYKIEL, 2003). O *credit scoring* pode ser definido como uma técnica que auxilia organizações na decisão de conceder o crédito aos clientes que o solicitam (THOMAS, 2000). O *credit scoring* assim como o *credit rating* compõe o conjunto de técnicas mais importantes para os processos de decisão de crédito dos bancos. Esse processo é composto por três etapas: a fase de coleta de informações, a etapa de análise e a fase de classificação, que analisa quais variáveis que influenciam a decisão do crédito.

A avaliação do crédito é um dos processos vitais para a sobrevivência de bancos e outras instituições que atuam como credores. Isso porque, a qualidade dos empréstimos que essas organizações cedem são fatores-chaves para a sua competitividade, sobrevivência e lucratividade perante os concorrentes (ABDOU; POINTON, 2011). A análise do risco de crédito é

indispensável para instituições financeiras que concedem crédito seja para fomentar negócios ou atender demandas de indivíduos. A importância de realizar a análise de risco de crédito se dá à medida que constrói-se um retrato dos clientes, ajudando a mitigar o risco de *default* e não-pagamento, além de promover o sucesso de longo-prazo de qualquer organização bancária (QASEM; NEMER, 2018).

Observa-se nos últimos anos uma evolução no processo de gerenciamento de risco de crédito. Os métodos de tomada de decisões que se pautavam apenas nos chamados critérios julgamentais perderam espaço nas atividades dos bancos, que exigem instrumentos mais eficientes para o gerenciamento da sua exposição ao risco de crédito (BRITO; ASSAF NETO; CORRAR, 2009). As instituições financeiras sempre utilizaram regras ou princípios desenvolvidos por analistas a fim de decidir quanto aqueles que receberiam o crédito. Porém, o aumento gradual no número de solicitantes de crédito tornou impossível sustentar esse modelo de concessão. Assim, a automatização do processo de decisão para aprovar o crédito tornou-se uma necessidade, de acordo com (LEE et al., 2002).

Nesse sentido, diversas técnicas unindo conhecimentos estatísticos e computacionais destacam-se como ferramentas para essa tarefa. Dentre elas, podemos citar as seguintes: Regressão Logística (NIKOLIC et al., 2013), aplicação de Redes Neurais (HUANG et al., 2004; ABDOL; POINTON, 2011), Naïve-Bayes classifier (PATIL; SHEREKAR, 2013; WU, 2011), Árvore de decisão (PANDEY et al., 2017), Extreme Learning Machine (QASEM; NEMER, 2018), Support Vector Machine (HUANG; CHEN; WANG, 2007), Artificial Neural Network (LEE et al., 2002) e as Redes Bayesianas (ASHCROFT, 2012).

Esse trabalho concentra-se na aplicabilidade da técnica de Redes Bayesianas para o desenvolvimento de um sistema de avaliação que classifique a situação financeira de empresas a partir de uma quantidade definida de índices financeiros obtidos através das demonstrações contábeis de cada uma delas. Para isso, será usada uma base de dados formada por empresas do setor de alimentos e bebidas com balanços contábeis para três anos consecutivos: 2013, 2014 e 2015.

1.1 Trabalhos relacionados

Pavlenko e Chernyak (2010) desenvolveram um estudo focado na gestão de risco de concentração de crédito. Para isso os autores selecionaram a modelagem bayesiana que foi implementada através de duas técnicas: as redes K-BN, um caso particular das redes Naïve-Bayes e também pela técnica de árvore Bayesiana aumentada (TAN). Para comprovar a aplicabilidade das técnicas, os pesquisadores utilizaram a base de dados de um banco de médio porte ucraniano. O intuito da pesquisa foi selecionar técnicas que possibilitassem gerir o risco de concentração envolvido em transações bancárias. O risco de concentração é definido pelos autores como o risco de crédito de um grupo de empresas reunidas por uma característica em comum. Essas características podem ser: as empresas do grupo possuem os mesmos sócios, ou as empresas compartilham muitos parceiros financeiros em comum, por exemplo. Pavlenko

e Chernyak (2010) comprovam que as técnicas de rede k-BN e TAN são técnicas apropriadas para esse objetivo já que elas possuem duas características essenciais para a análise do risco de concentração: as técnicas permitem visualizar de forma prática a maneira como as empresas do grupo estão relacionadas, em outras palavras, as técnicas permitem compreender essas relações através de grafos. Além disso, outra vantagem dessas técnicas é que elas permitem incorporar a informação de analistas de crédito e outros especialistas na área de risco num primeiro momento e depois atualizar essas probabilidades conforme novas evidências são obtidas por eles.

Abdou e Pointon (2011) trazem importantes contribuições para a modelagem de crédito. Os pesquisadores conduziram um estudo focado na revisão de literatura de modelos de avaliação de crédito. Foram 214 trabalhos analisados que incluíam: artigos, teses e livros com enfoque no segmento de finanças e também no segmento bancário. Ao término da pesquisa, os autores puderam concluir que há um consenso sobre o uso de modelos de avaliação de crédito. A revisão dos trabalhos indica que esses modelos são muito usados porque garantem uma tomada de decisão desprovida de julgamento pessoal. Ao contrário do julgamento pessoal, esse método decide a favor ou contra o devedor com base na sua pontuação. É, portanto, uma medida numérica. Outra grande contribuição da pesquisa foi revelar a inexistência de um método de avaliação de crédito ideal. De acordo com os seus resultados, não há um método capaz de indicar quais as variáveis mais importantes, qual o tamanho da amostra ideal e qual o ponto de *cut-off* perfeito para cada uma das situações que se quer modelar. A última contribuição que merece destaque é o resultado da comparação realizada entre algumas das técnicas de avaliação de crédito quanto a sua capacidade preditiva. Abdou e Pointon (2011) concluíram que dentre os métodos estatísticos aplicados nos modelos de avaliação de crédito, as técnicas mais avançadas como as redes neurais e a programação genética, produzem melhores resultados quando comparadas com técnicas mais tradicionais como, por exemplo, a análise discriminante e a regressão logística.

Wu (2011) propõe um estudo visando selecionar as melhores técnicas para uma análise de crédito consistente, de maneira a melhorar a acurácia da classificação de proponentes de créditos e obter conhecimentos mais aprofundados sobre reconhecimentos de padrões dos devedores, que resultem na tomada de decisões de crédito mais assertivas. Após uma revisão dos principais trabalhos na área de modelagem de crédito Wu (2011) testa a hipótese de que um modelo híbrido (modelo que envolve mais de uma técnica) de aprendizagem de máquina é mais eficiente para a modelagem de crédito do que o uso de uma única técnica. As técnicas selecionadas foram: o classificador bayesiano e o algoritmo de busca TAN (FRIEDMAN; GEIGER; GOLDSZMIDT, 1997). A pesquisa realizada permite concluir que o SMO (*Sequential Minimal Optimization*) ou MLP (*Multilayer Perceptron*) são técnicas eficientes para o pré-processamento de dados. Além disso, o classificador bayesiano quando aplicado junto ao algoritmo de busca TAN são técnicas complementares eficazes que aumentam a confiabilidade na tomada de decisões de crédito.

Patil e Sherekar (2013) desenvolveram um trabalho comparando o desempenho de duas técnicas de classificação de dados sob três aspectos: acurácia, sensibilidade e especificidade.

As técnicas escolhidas foram o classificador Naïve-Bayes e algoritmo J48 (SHARMA; SAHNI, 2011). O método Naïve-Bayes baseia-se na teoria de probabilidades enquanto que, o algoritmo J48 usa a técnica de árvore de decisão. Para comparar os métodos descritos, os autores selecionaram um banco de dados de um banco formado por uma série de atributos sobre os seus clientes, como por exemplo: gênero, idade, estado civil, se possui filhos entre outros. Os autores focaram em dois atributos: hipoteca e gênero. As duas técnicas foram testadas sobre essa base de dados e para comparar qual delas classificava o maior percentual de casos de forma correta utilizou-se uma matriz de confusão. Uma vez que o percentual de falso-positivos foi inferior na técnica J48 e o percentual de positivos verdadeiros superior à técnica bayesiana, os autores concluíram que a técnica J48 é mais eficiente em termos de acurácia e eficiência que o método Naïve-Bayes.

Pandey et al. (2017) desenvolveram um trabalho de comparação das principais técnicas para análise de risco de crédito. As nove técnicas selecionadas foram: *Bayesian classifier*, *Naïve-Bayes classifier*, *Decision tree*, KNN (*K-Nearest Neighbor*), K-means clustering, MLP (*Multilayer perceptron*), ELM (*Extreme Learning Machine*), SVM (*Support Vector Machine*) e ANN (*Artificial Neural Network*). Para comparar essas técnicas e selecionar a que possuía a maior acurácia dentre elas, os autores recorreram a duas bases de dados: a German Credit e a Australian credit, ambas disponíveis no repositório UCI. A base German possuía catorze atributos e a Australian contava com vinte atributos que podem ser compreendidos como variáveis explicativas, que juntas determinavam se o crédito seria aprovado ou não para cada um dos casos da base de dados. Os pesquisadores submeteram cada uma das técnicas às duas bases de dados. Pandey et al. (2017) concluíram que a técnica ELM era a de maior acurácia entre as técnicas testadas, já que ela foi a que revelou maior percentual de acertos na classificação das bases.

Dentre as pesquisas brasileiras na área de risco de crédito e modelagem destacam-se os trabalhos apresentados abaixo.

Louzada, Ferreira-Silva e Diniz (2012) desenvolveram um estudo comparando duas técnicas de regressão logística: a técnica naïve e a técnica com estado dependente da seleção da amostra (*Logistic regression with state-dependent sample selection*). Para comparar a capacidade preditiva das técnicas, os pesquisadores criaram uma amostra de dados hipotética, composta por 1.000.000 de clientes classificados como bons pagadores e 100.000 clientes classificados como maus pagadores. As métricas empregadas para analisar a capacidade preditiva dos modelos são: sensibilidade, especificidade e acurácia. A pesquisa mostrou que, embora as duas técnicas apresentem valores similares de sensibilidade, especificidade e acurácia, a técnica de regressão logística naïve, subestima as probabilidades de *default*. Outra importante conclusão é que adotar amostras balanceadas, com a mesma porcentagem de clientes classificados como *default* e não-*default* contribui para obtenção de modelos de previsão com capacidade preditiva elevada. As simulações com amostras de porcentagens distintas para essas duas classes de clientes resultaram em modelos menos preditivos.

Guimarães e Resende (2018) desenvolveram um trabalho aplicando a técnica de re-

gressão logística para previsão de falência empresarial no Brasil. O objetivo da pesquisa era desenvolver um modelo capaz de prever o estado de inadimplência de empresas não-financeiras brasileiras com um ano de antecedência. Para isso, os pesquisadores recorreram à base de dados do software Economática, no qual selecionaram um total de 401 empresas não-financeiras como base de dados para a pesquisa. Essa base era composta por um total de 43 empresas insolventes e 358 empresas solventes. As variáveis explicativas do modelo eram formadas por um conjunto de 34 indicadores contábeis que compreendiam indicadores como a Liquidez Seca, Ebitda sobre despesas financeiras e também Patrimônio Líquido sobre Ativo Total, por exemplo. Os resultados dos testes realizados com o modelo mostraram que as variáveis de rentabilidade possuem grande significância estatística, e o modelo mostrou grande capacidade preditiva, principalmente na classificação das empresas solventes. Para a amostra adotada na pesquisa o modelo foi capaz de prever de forma correta maior percentual de empresas solventes do que de empresas insolventes. Porém, conforme apontam os autores, essa característica pode ser específica da amostra.

Sá et al. (2018) desenvolveram um trabalho na área de análise de crédito usando a técnica de mineração de dados. O objetivo da pesquisa era testar a aplicação da técnica de mineração denominada CRISP-DM (Processo Padrão Inter-Indústrias para Mineração de Dados) através de métodos de classificação para a aprovação de concessão de crédito. A base de dados usada para a pesquisa é composta por 500.000 instâncias e 176 atributos que desempenham o papel de classificar os clientes em dois tipos: aptos ou não aptos para receber o crédito. Os atributos dos clientes incluem características do tipo: cadastrais, financeiras, geográficas e até mesmo partidárias. Quatro técnicas de modelagem foram adotadas na pesquisa: o método Naïve Bayes, o método k vizinhos mais próximos (KNN), a regressão logística e a técnica de redes neurais artificiais. A pesquisa revelou, através da comparação de medidas de performance das técnicas que nenhuma delas apresentou grau de acurácia satisfatório. Dentre os motivos que justificam esse resultado os autores apontam a baixa correlação entre as variáveis usadas para classificar os clientes como uma das principais causas. Os autores sugerem o uso de técnicas de Deep learning como medida para obter resultados mais consistentes.

Os resultados dos testes realizados com o modelo mostraram que as variáveis de rentabilidade possuem grande significância estatística, e o modelo mostrou grande capacidade preditiva, principalmente na classificação das empresas solventes. Para a amostra adotada na pesquisa o modelo foi capaz de prever de forma correta maior percentual de empresas solventes do que de empresas insolventes. Porém, conforme apontam os autores, essa característica pode ser específica da amostra.

Esse trabalho dedica-se a desenvolver um modelo de crédito com a técnica de Redes Bayesianas com o objetivo de validar a técnica de Redes Bayesianas como uma alternativa mais preditiva na classificação do risco de crédito associado a um cliente.

1.2 Problema de Pesquisa

O problema de pesquisa desse trabalho consiste em avaliar a capacidade da técnica de Redes Bayesianas como uma técnica para previsão do risco de crédito de empresas do segmento de alimentos e bebidas.

1.3 Objetivos

1.3.1 Geral

Desenvolver um modelo de crédito, a partir da técnica probabilística de Redes Bayesianas, que atribua a classe de risco de crédito em que uma empresa se enquadra, a partir de um conjunto de 17 indicadores financeiros calculados através das suas demonstrações financeiras.

1.3.2 Específicos

- (a) Desenvolver um modelo de crédito a partir da técnica de Redes Bayesianas que classifique empresas em classes de risco que variam segundo os valores dos seus indicadores financeiros;
- (b) Comparar a acurácia do modelo bayesiano com o modelo de Regressão Logística;
- (c) Desenvolver um aplicativo que implemente a Rede Bayesiana, possibilitando assim a utilização da mesma nos processos de decisão de crédito que envolvem empresas.

1.4 Justificativa

A revisão dos trabalhos e artigos que discutem a temática de modelagem de crédito demonstram uma carência no emprego de técnicas estatísticas e probabilísticas para o cálculo do risco de crédito. Isso abre a oportunidade para que novas técnicas sejam implementadas e testadas como possíveis alternativas ao processo tradicional de cálculo de risco de crédito. Nesse sentido, o emprego da técnica probabilística de Redes Bayesianas na modelagem de crédito contribui na prospecção de novas técnicas.

Conforme apontam Hand, McConway e Stanghellini (1997) a grande contribuição da técnica de Redes Bayesianas para a modelagem de crédito reside em dois aspectos. O primeiro, é quanto a compreensão das relações que existem entre as variáveis que explicam o comportamento dos devedores. Neste trabalho, os devedores são representados pelas empresas e as variáveis são o conjunto de indicadores financeiros calculados a partir delas. A técnica de Redes Bayesianas automatiza esse processo, já que, através do cálculo de correlações e covariâncias entre as variáveis, a própria técnica encarrega-se de identificar possíveis correlações entre os indicadores e assim selecionar apenas os mais relevantes. O segundo aspecto é a possibilidade de modelar um conjunto de variáveis, sem definir previamente, qual dessas variáveis é

a variável resposta. Essa característica da técnica de grafos probabilísticos traz resultados mais próximos a realidade, já que dispensa a necessidade de que o gestor de risco configure ou defina as variáveis que mais impactam a variável risco, tornando assim o modelo de classificação de risco de crédito menos subjetivo.

Dentre as técnicas presentes nos artigos que abordam a modelagem de crédito, destacam-se as técnicas de: Support Vector Machine (HUANG; CHEN; WANG, 2007), Redes Neurais (HUANG et al., 2004), análise discriminante (LEE et al., 2002), Regressão Logística (HOSMER; LEMESHOW, 2000) e a técnica de cluster (LIM; SOHN, 2007). Justifica-se portanto a relevância dessa pesquisa na medida que ela acrescenta uma abordagem pouco explorada na modelagem de crédito que pretende contribuir para tornar o processo de concessão de crédito mais objetivo e menos dependente da avaliação julgamental de analistas de crédito, que por vezes, prejudicam a obtenção de notas de risco mais condizentes com a saúde financeira das empresas avaliadas.

Para viabilizar esse objetivo, essa pesquisa dedica-se ao desenvolvimento de um modelo de crédito que atribui uma classe de risco para uma empresa, a partir de um conjunto de indicadores financeiros calculados a partir da mesma. Os indicadores financeiros são as variáveis que compõe a Rede Bayesiana e as relações entre eles são representadas através de grafos probabilísticos. Essas relações definem a configuração da Rede Bayesiana, de maneira que, é possível identificar através do modelo as variáveis com maior e menor influência sobre a variável resposta risco.

1.5 Estrutura do trabalho

Esta dissertação está organizada da seguinte maneira: o capítulo 2 apresenta a revisão bibliográfica dos principais conceitos envolvendo o crédito. O capítulo 3 apresenta os conceitos probabilísticos que fundamentam os grafos probabilísticos e o capítulo 4 apresenta o conceito e os principais tipos de grafos probabilísticos.

O capítulo 5 apresenta as características da base de dados usada para a modelagem, enquanto o capítulo 6 analisa esses dados através de técnicas estatísticas de análise de dados.

O capítulo 7 trata do processo de modelagem e apresenta os principais tipos de modelos para representar um conjunto de dados. O capítulo 8 apresenta a validação do modelo através da técnica de Regressão logística. O capítulo 9 apresenta a validação do modelo bayesiano e o capítulo 10 apresenta a interface do aplicativo desenvolvido para implementar a modelagem bayesiana. O capítulo 11 apresenta as conclusões.

Capítulo 2

Revisão Bibliográfica

2.1 Risco de Crédito

O risco de crédito é considerado a forma mais antiga de risco no mercado financeiro. Se o conceito de crédito pode ser compreendido como a expectativa de recebimento de uma quantia em um período definido, então a definição de risco de crédito pode ser entendida como a expectativa de que isso não se concretize (CAOINETTE; ALTMAN; NARAYANAN, 1998). A gestão desse risco é essencial para a atividade de diversas empresas, principalmente para aquelas pertencentes a indústria bancária. Isso porque, em cenários de crises financeiras, os bancos e outras instituições financeiras estão sujeitos à grandes perdas financeiras (GARCÍA; GIMÉNEZ; GUIJARRO, 2013).

Iscoc et al. (2012) define o risco de crédito como sendo o resultado das perdas financeiras em que uma empresa incorre ao não cumprir uma obrigação, evento definido como *default*. O *default* é um termo associado a eventos financeiros que implicam em perdas para ambas ou apenas uma das partes envolvidas em uma transação. Ainda não há um consenso sobre a definição desse conceito, porém, em geral, diz respeito a atrasos em pagamentos por períodos superiores à 60 ou 90 dias. O *default* ainda é considerado como um dos três componentes que formam o risco de crédito, juntamente com o risco de exposição e de recuperação (BRITO; ASSAF NETO, 2008).

No âmbito empresarial, a administração desse risco engloba duas categorias de perdas diferentes: as perdas esperadas e as perdas inesperadas. As perdas esperadas podem ser calculadas ou previstas e por isso mesmo, não constituem um risco de crédito propriamente dito. Por isso, uma definição mais apropriada é assumir que o risco de crédito é a probabilidade de uma deterioração inesperada na qualidade de crédito de um devedor (SAITA, 2007).

Historicamente, nota-se que até mesmo os bons clientes estão sujeitos ao não cumprimento das suas obrigações financeiras, o que tem levado bancos a adotarem seguros contra todos os empréstimos que eles concedem. Esse mesmo pensamento é válido quando consideramos transações entre instituições financeiras e empresas. Neste caso, o risco de não-recebimento dos clientes é embutido no valor dos empréstimos cobrados das instituições, que criam um fundo

de provisão para potenciais clientes inadimplentes (BLUHM; OVERBECK; WAGNER, 2002).

Para mensurar o risco de crédito, bancos classificam seus clientes corporativos quanto a probabilidade de *default* associada a cada um deles, o que permite classificá-los e torna a tarefa de conceder ou não o crédito, mais segura. Para isso, adota-se uma função linear que utiliza um conjunto de variáveis de caráter econômico e financeiro dessas empresas. Isso possibilita aos credores uma avaliação completa sobre os devedores, já que essas variáveis englobam diferentes aspectos sobre os clientes corporativos, por exemplo, tamanho da empresa, liquidez, solvência e lucratividade (GARCÍA; GIMÉNEZ; GUIJARRO, 2013).

2.2 Credit Scoring

Thomas, Edelman e Crook (2017) definem o processo de atribuir uma pontuação de crédito como um conjunto de modelos de decisão que auxilia os credores na decisão de conceder o crédito. Esses modelos são utilizados para as seguintes decisões: definir quem irá receber o crédito, quanto crédito será ofertado para esses clientes, qual valor será cobrado por esse empréstimo e quais estratégias podem ser adotadas em prol da lucratividade dos credores nessas transações.

De acordo com Sánchez e Lechuga (2016) o processo que determina a pontuação de crédito de uma pessoa ou empresa, é um dos mais importantes nas decisões de gestão de crédito de um banco. O processo inclui a coleta, análise e classificação de diversas variáveis que influenciam nas decisões relativas a concessão de crédito. Dentro desse processo, uma das ferramentas mais importantes a disposição dos bancos na tarefa de reduzir o risco de conceder crédito a maus pagadores é a pontuação de crédito.

A primeira utilização da pontuação de crédito de que se tem registro remete à década de 60. Naquela época a técnica era usada para avaliar se os solicitantes de crédito seriam capazes ou tinham condições de pagar as dívidas assumidas ou honrar as obrigações contraídas. No século 21, no entanto, a definição de pontuação de crédito adquiriu maior amplitude. Até então o *credit scoring* era associado apenas a decisões do tipo: aceitar ou rejeitar. Nesse novo contexto, o *credit scoring* passou a representar o uso de modelos estatísticos para a gestão do crédito (ANDERSON, 2007).

Sullivan (1981) e Bailey (2004) argumentam a existência de dois métodos para mensurar o risco de crédito: o método julgamental e o método estatístico. O método julgamental é baseado na experiência de profissionais considerados especialistas na área de risco, baseado também na experiência com negócios desses mesmos profissionais e em um comitê de crédito, que constitui um grupo de profissionais especialistas no assunto que tomam a decisão de conceder ou não o crédito com base em 5 critérios, conhecidos como os 5 Cs de crédito. Os 5Cs são definidos como:

- a) Caráter: mede as características do devedor como integridade, reputação e honestidade;
- b) Capital: mede as diferenças entre os ativos (bens como veículo e casa própria, por

exemplo) do solicitante de crédito e os seus passivos (despesas com aluguel e outras dívidas);

c) Colateral: mensura as garantias como carro e casa que o devedor possui caso não consiga arcar com as dívidas contraídas;

d) Capacidade: mede a capacidade do devedor em arcar com as despesas assumidas. Essa capacidade é medida através do cargo que o cliente ocupa e do seu salário, por exemplo;

e) Condição: mensura as condições econômicas que os tomadores de crédito estão sujeitos, como comportamentos sazonais e condições de mercado que podem impactar na capacidade de pagamento dos devedores.

Esses cinco critérios fornecem conteúdo para que os analistas e gestores de crédito conduzam uma análise de crédito, seja ela subjetiva ou quantitativa, segundo Baesens, Rösch e Scheule (2016). O método julgamental caracteriza-se pela sua subjetividade. As aplicações de crédito são avaliadas individualmente por um gestor transferindo assim o sucesso da solicitação para sua experiência como profissional na área de crédito e para o senso comum dos analistas de crédito (SULLIVAN, 1981; BAILEY, 2004). Por isso mesmo, as críticas em torno da análise julgamental apontam a subjetividade, a inconsistência e a influência de preferências individuais para a tomada de decisões como os principais defeitos dessa técnica. Brito e Assaf Neto (2008), sustentam essa afirmação à medida que relatam maior ênfase dos modelos quantitativos no processo de revisão de métodos para avaliação do risco de crédito. Segundo os autores, um comportamento recente, notado em instituições financeiras e bancos é o abandono dos métodos mais tradicionais de tomada de decisão, baseados em critérios julgamentais, em prol de instrumentos mais eficientes para medir o risco de tomadores de crédito e também de carteiras de crédito.

O método estatístico para cálculo do *credit scoring* destaca-se pela sua consistência, acurácia e eficiência. Por se tratar basicamente de uma fórmula matemática, a pontuação de crédito pode ser programada e calculada de forma rápida e eficiente. Essa característica é interessante porque permite a tomada de decisões mais rápidas. A consistência do método é explicada pela ausência de subjetividade na tomada de decisão, já que a experiência e intuição dos especialistas não interfere na avaliação do método. Finalmente, a eficiência do método é resultado da sua performance superior aos métodos julgamentais, que reduz a perda com maus pagadores e custos operacionais que a empresa incorreria se concedesse crédito para clientes com essas características, segundo Baesens, Rösch e Scheule (2017).

O resultado das técnicas e modelos estatísticos usados no *credit scoring* geram pontuações que permitem avaliar o risco de um tomador ou da operação de crédito. Nessa pesquisa, as pontuações serão usadas para classificar empresas em classes de risco que variam conforme a probabilidade de inadimplência de cada uma delas. A Tabela 1 abaixo representa a classificação que será usada nessa dissertação.

Tabela 1 – Classes de Risco e probabilidade de inadimplência associada às classes de risco.

Classes de Risco	Faixas Probabilidade de Inadimplência (%)	Probabilidade Média de Inadimplência (%)
1	0,00 a 0,10	0,05
2	0,10 a 0,20	0,15
3	0,20 a 0,30	0,25
4	0,30 a 0,40	0,35
5	0,40 a 0,50	0,45
6	0,50 a 0,75	0,62
7	0,75 a 1,00	0,87
8	1,00 a 1,25	1,12
9	1,25 a 1,50	1,37
10	1,50 a 2,00	1,75
11	2,00 a 3,00	2,5
12	3,00 a 4,00	3,5
13	4,00 a 5,00	4,5
14	5,00 a 8,00	6,5
15	8,00 a 10,00	9
16	10,00 a 15,00	12,5
17	15,00 a 30,00	22,5
18	30,00 a 50,00	40
19	50,00 a 99,99	75
20	Default - Cesta de Eventos/Refin/CCF	
21	Default - Recuperação Judicial e Extrajudicial	
22	Default - Falência	

Fonte: Experian (2018)

A Tabela 1 de classificação é usada pela Serasa Experian para classificar o risco de crédito das empresas que compõe seu portfólio de clientes. O risco de inadimplência é representado pelas Classes de Risco, ordenadas de forma crescente: a classe de risco 1, compreende as empresas com probabilidade de inadimplência pequena e consequente risco de crédito baixo. Já a classe de risco 19, compreende empresas com probabilidade de inadimplência elevada e consequente risco de crédito alto.

2.3 Modelos de Crédito

Os modelos de risco de crédito são ferramentas e aplicações para medir o risco transações individuais e o risco existente numa carteira de crédito como um todo (BRITO; ASSAF NETO, 2008). Segundo Andrade (2003), esses modelos podem ser classificados em três grupos: modelos de classificação de risco, modelos estocásticos de risco e modelos de risco de portfólio. Esses modelos são definidos da seguinte forma:

- (a) **modelos de classificação de risco:** tem como objetivo avaliar o risco de um proponente ao crédito ou de uma operação, atribuindo dessa forma uma medida que traduza a expectativa de risco de *default*, usualmente expressa na forma de uma pontuação ou ainda

como uma classificação de risco, também conhecido como pontuação de crédito. Esses modelos são muito usados pelas instituições financeiras em seus processos de concessão de crédito;

- (b) **modelos estocásticos de risco:** objetivam analisar o comportamento estocástico do risco de crédito, bem como, o comportamento das variáveis dele determinantes. Esses modelos possuem como principal objetivo a precificação de títulos e derivativos de crédito;
- (c) **modelos de risco de portfólio:** buscam estimar a distribuição estatística das perdas ou ainda do valor de uma carteira de crédito, por meio da qual, são obtidas medidas que permitem quantificar o risco de crédito. Esses modelos são muito importantes para a gestão do risco dentro das instituições, uma vez que a partir deles é possível avaliar o risco de crédito considerando a diversificação resultante das correlações existentes entre os ativos de uma mesma carteira.

Dentre esses três tipos de modelos, Fayyad (2008) destaca que os mais difundidos no mercado para processos de análise de risco de crédito são: o CreditMetrics, o modelo KMV, CreditRisk+ e o CreditPortfolioView, descritos a seguir.

- **Modelo CreditMetrics**

Desenvolvido pelo banco JP Morgan em 1997, o modelo tornou-se um dos principais instrumentos para a gestão de risco de instituições financeiras. O CreditMetrics é um modelo para análise e gestão do risco de crédito, que se difere dos outros modelos por ser baseado na análise do risco de crédito atrelado aos instrumentos de investimento em uma carteira e não apenas em títulos individuais, como ocorre em outros modelos (BORIS; IVANAA; ANNAA, 2015).

Esse modelo é muito conhecido entre estudiosos de econometria como um modelo do tipo probit ordenado. Isso porque no CreditMetrics, os eventos de crédito são atribuídos à oscilações em variáveis latentes não-observáveis, cujo comportamento está atrelado a variáveis macroeconômicas, segundo Gordy (2000).

A distribuição dessas variáveis, segundo Saunders e Allen (2002), é significativa para a estimação do modelo, e por isso mesmo, é fundamental destacar os pressupostos assumidos quanto a distribuição e calibração dos parâmetros do modelo. Os autores argumentam ainda que, ao assumir que as variáveis macroeconômicas seguem uma distribuição normal, como ocorre no CreditMetrics, a comparação em termos de resultado com um modelo que adota outra distribuição para as mesmas variáveis, pode acarretar em resultados diferentes.

A importância dessas distribuições ficam mais evidentes em Gordy (2000), que realiza uma comparação entre dois modelos, o CreditMetrics e o CreditRiskPlus. O modelo CreditMetrics, quando comparado ao CreditRiskPlus, que adota a distribuição Gamma para a probabilidade média de *default* divergem de forma significativa quanto as estimativas de perdas não-esperadas, quando o desvio-padrão do fator de risco sistêmico apresenta alta variabilidade.

• Modelo KMV

O método KMV é um método para o cálculo de risco de crédito baseado no modelo de precificação de ações desenvolvido por Merton (1974). Também definido como modelo de precificação de ações ou ainda sob o nome de método estrutural para mensuração de risco de crédito, o modelo KMV modela o processo de *default* através da estimação da probabilidade de que esse evento aconteça (PAPALAMPROU; ANTONIOU, 2019).

A frequência esperada de *default* (EDF) ou estimação da probabilidade de *default* é uma métrica usada para calcular o quão distante uma empresa está de incorrer em um *default*. Para calcular o EDF, o modelo KMV utiliza-se de um processo estocástico, que acompanha as flutuações de valores do ativo de uma instituição em um horizonte de tempo. Assume-se que o logaritmo do retorno dessas ações segue uma distribuição empírica e o desvio padrão dos valores permanece constante nesse horizonte de tempo. Os *ratings* são dados a partir da EDF, de forma que quanto maior a distância do *default* de uma empresa, melhor a classificação que ela recebe (FAYYAD, 2008).

Figura 1 – Modelo KMV de probabilidade de *default*

Fonte: Adaptado de Kealhofer (2003, p.31)

A Figura 1 representada acima ilustra a ideia por trás do modelo KMV. A região de *default* é representada na Figura 1 pela área definida abaixo do ponto de *default*, na cauda direita da curva normal, em negrito. Essa região em destaque representa a probabilidade de que o valor de mercado dos ativos de uma empresa, no período de um ano, seja inferior aos ganhos dessa empresa. Dessa forma, se o valor de mercado dos ativos da empresa decrescem, ou o montante de dívidas apresentam um aumento substancial, e ainda, se a volatilidade do valor de mercado dos ativos apresentarem uma elevação, a probabilidade de *default* aumenta (SCHOMMER, 2007).

- **Modelo CreditRisk+**

É um modelo para o cálculo de risco de crédito considerado o maior representante dos modelos atuariais. Desenvolvido pelo *Credit Suisse Financial Products*(CSFP), o CreditRisk+ é um método estatístico baseado no risco de *default*. As taxas de default são consideradas como variáveis aleatórias contínuas e as variações dessas taxas são incorporadas à incerteza associada à essas variáveis. Alguns fatores externos como, por exemplo, o estado da economia, podem contribuir para a elevação ou redução das taxas de *default*. Porém, como não existe uma relação causal entre eles, o efeito desses fatores são incorporados na variação das taxas (SANFINS; CLARK, 2010).

O *default* no CreditRisk+ é a união da frequência com que os eventos de *default* ocorrem e a severidade de perdas que eles acarretam. A estimação de perdas na carteira está indissociada da distribuição de eventos de *default*. O modelo CreditRisk+ assume que a distribuição desses eventos aproxima-se de uma distribuição de Poisson. O modelo inicial considera que essas variáveis são independentes, ou seja fixas, porém como esse pressuposto é questionável, já que estudos anteriores provam a variação dessas taxas, desenvolveu-se uma forma mais sofisticada para o modelo. O método propõe que as taxas de *default* assumam uma distribuição Gama, com a mesma média de distribuição, porém com caudas mais grossas. Isso impacta diretamente no aumento da quantidade de eventos extremos, consequência da correlação implícita da volatilidade quando incorporada à taxa de *default* (SANFINS; CLARK, 2010).

- **Modelo CreditPortfolioView**

Crouhy, Galai e Mark (2000) definem o CreditPortfolioView como um modelo multifator usado para simular distribuições condicionais conjuntas de *default* e também probabilidade de migração de rating dentro de diversos grupos. Esses grupos variam conforme o segmento de indústria, o país em que a indústria está inserida, fatores macroeconômicos como a taxa de desemprego, o crescimento do PIB, taxa de câmbio e também gastos governamentais.

O CreditPortfolioView baseia-se na premissa de que a probabilidade de *default* está atrelada às condições econômicas. Quando as condições econômicas evidenciam uma piora, a probabilidade de ocorrer um *default* aumenta. “Em outras palavras, os ciclos de crédito seguem o ciclo de negócios de perto” Crouhy, Galai e Mark (2000, p. 113). Uma vez que a economia é fortemente afetada por fatores macroeconômicos, o CreditPortfolioView propõe um método que relaciona os eventos de *default* e migração de risco aos fatores macroeconômicos.

Após essa revisão de literatura a respeito dos principais conceitos sobre o risco de crédito, o próximo capítulo apresenta os conceitos que fundamentam a técnica probabilística de Redes Bayesianas empregada neste trabalho.

Capítulo 3

Conceitos probabilísticos

Para adentrar na teoria de probabilidade iniciamos com a seguinte citação do Marquês de Laplace:

“Vemos que a teoria da probabilidade é no fundo somente o senso comum reduzido ao cálculo; ela nos faz apreciar com exatidão o que mentes pensantes percebem como que por instinto, muitas vezes sem se dar conta disso. As mais importantes questões da vida são, em sua grande maioria, apenas problemas de probabilidade.” (LAPLACE, 1814)

A frase do matemático Laplace resume de maneira objetiva a principal função da probabilidade: traduzir em números aquilo que percebemos ao nosso redor.

Thomas Bayes foi um pastor britânico e matemático vindo de uma família abastada. Os trabalhos produzidos durante a vida de Bayes mesclaram entre publicações de cunho religioso e publicações de caráter matemático. Na época, seus trabalhos tiveram boa aceitação por parte dos leitores, embora nenhum desses artigos mostrasse preocupação com a teoria de probabilidade, segundo Tabak (2004).

A notoriedade do autor deve-se a um de seus trabalhos que ele nunca publicou. Após a sua morte, o colega e pastor Richard Price (1723 – 91), dedicou à reunir alguns de seus trabalhos. Dentre eles, um artigo intitulado : “An essay towards Solving a problem in the doctrine of chances”, chamou atenção por introduzir conceitos de probabilidade. Nesse artigo encontra-se a primeira definição do que conhecemos hoje como teorema de Bayes (TABAK, 2004).

Para compreender a relevância e aplicabilidade desse teorema apresentamos um exemplo simples e corriqueiro onde o teorema pode ser usado. Em um jogo de baralho, quando um participante tira uma carta ao acaso, é intuitivo que a chance da carta retirada ser de ouros é de 1 em 4. Porém, se alguém diz que a carta retirada é vermelha, a probabilidade de que a carta seja de ouros aumenta para 50%. Isso acontece porque, segundo o teorema Bayesiano, a probabilidade de que a carta seja de ouros é dada pela seguinte fórmula:

$$P(A|B) = \frac{P(B|A).P(A)}{P(B)} \quad (3.1)$$

A= representa a chance de que as cartas sejam de ouros.

B= representa a chance de que as cartas sejam vermelhas.

Dessa forma, temos que:

$$P(ouros|vermelho) = \frac{P(vermelho|ouros).P(ouros)}{P(vermelho)} \quad (3.2)$$

- $P(vermelho|ouros) : 1$
- $P(ouros) : \frac{1}{2}$
- $P(vermelho) : \frac{1}{4}$

O cálculo resulta no seguinte valor: $(1 \times 0.25)/0.5 = 0.5$ ou 50%. Nas próximas seções detalharemos os principais conceitos probabilísticos para compreensão do teorema de Bayes, assim como, o teorema Bayesiano.

3.1 Espaço Amostral e Eventos

Para introduzir o conceito de espaço amostral considere um experimento do qual não é possível prever o resultado final com certeza. Suponha ainda que, embora o resultado não seja conhecido, o conjunto dos resultados possíveis para o experimento é conhecido. Esse conjunto é o *espaço amostral* de um experimento e é representado pela letra "S" (ROSS, 2010).

Exemplo 1.0

Imagine um experimento que consiste em jogar duas moedas. O espaço amostral desse experimento é formado pelo conjunto:

$$S = \{(Ca, Ca), (Ca, Co), (Co, Ca), (Co, Co)\}$$

Observação: *Ca* simboliza cara e *Co* simboliza coroa.

Se as duas moedas derem cara, o resultado será (Ca,Ca). Se a primeira der cara e a segunda der coroa, o resultado será (Ca,Co). Mas, se a primeira der coroa e a segunda der cara, o resultado será (Co,Ca) e se as duas derem coroa, o resultado será (Co,Co).

Um *evento* é um qualquer subconjunto *E* do espaço amostral. Em outras palavras, um evento é qualquer conjunto formado pelos possíveis resultados de um experimento (ROSS, 2010).

No exemplo 1.0, $E = \{(Ca, Ca), (Ca, Co)\}$ é um evento em que a primeira moeda resulta em cara.

3.2 Variáveis Aleatórias

Os objetivos de um experimento podem ser os mais variados possíveis. No lançamento de duas moedas, por exemplo, alguns podem estar interessados no evento em que a segunda moeda resulta em coroa, dado por $E_1 = \{(Ca, Co), (Co, Co)\}$ e outros no evento em que as duas

moedas dão o mesmo resultado, dado por $E_2 = \{(Co, Co), (Ca, Ca)\}$. Esses eventos de interesse são definidos como *variáveis aleatórias*. De maneira mais formal, podemos definir uma variável aleatória como: uma função real definida no espaço amostral de um experimento (ROSS, 2010).

Exemplo 2.0

Suponha que estejamos interessados no seguinte experimento: lançamento de três moedas honestas. Vamos adotar que Ca representa os lançamentos que resultam em cara e Co os que resultam em coroa. Y é uma variável aleatória que assume o número de caras que aparecem nesses lançamentos. Quais os possíveis valores que Y assume?

Solução

A variável Y pode assumir quatro valores: 0, 1, 2 ou 3. "0" representa o evento em que todas as moedas resultaram em coroa. "1" representa o evento em que apenas uma moeda resultou em cara e assim sucessivamente. Os valores que variável assume e as respectivas probabilidades são representados abaixo:

$$\begin{aligned} P(Y = 0) &= P(Co, Co, Co) = \frac{1}{8} \\ P(Y = 1) &= P\{(Ca, Co, Co), (Co, Ca, Co), (Co, Co, Ca)\} = \frac{3}{8} \\ P(Y = 2) &= P\{(Ca, Ca, Co), (Co, Ca, Ca), (Ca, Co, Ca)\} = \frac{3}{8} \\ P(Y = 3) &= P(Ca, Ca, Ca) = \frac{1}{8} \end{aligned}$$

3.3 Probabilidade Condicional

Para dois eventos quaisquer A e B , com $P(B) > 0$, a probabilidade condicional de A dado B é definida pela fórmula 3.3 a seguir:

$$P(A|B) = \frac{P(B|A) \times P(A)}{P(B)} \quad (3.3)$$

Exemplo 3.0

Considere o Exemplo 2.0. Qual a probabilidade de sair cara em um dos três lançamentos das moedas? Nesse caso, a probabilidade é dada por $P(Y = 1)$ e conforme já mostramos anteriormente essa probabilidade resulta em:

$$P(Y = 1) = P\{(Ca, Co, Co), (Co, Ca, Co), (Co, Co, Ca)\} = \frac{3}{8}$$

Para compreender como a probabilidade condicional altera a nossa resposta vamos pensar em uma situação mais complexa. Qual a probabilidade de sair cara, dado que uma moeda deu coroa?

Quando inserimos a evidência de que uma moeda deu coroa e queremos saber as chances de obter uma única cara, restringimos nosso espaço amostral. Neste caso, dos três lançamentos, temos informações de apenas dois deles. Porém e quanto ao outro lançamento? Como os

lançamentos são independentes, a probabilidade de obter uma cara ou uma coroa no próximo lançamento é a mesma :50%.

Considere A = probabilidade de sair cara e B = probabilidade de sair coroa. Aplicando a fórmula 3.3 temos que:

$$P(A|B) = \frac{3/8 \times 1/2}{1/2} = \frac{3}{8} \quad (3.4)$$

Assim, a probabilidade de sair cara, dado que já saiu um coroa é dada por $P(A|B) = 3/8$. Da relação expressa em 3.3 obtemos ainda a *regra do produto*:

Regra 1 (Regra do Produto).

$$P(A \cap B) = P(B|A).P(A) \quad (3.5)$$

O que a fórmula 3.5 diz é que: a probabilidade de que A e B ocorram é igual a probabilidade de que B ocorra, dado que A já tenha ocorrido, multiplicado pela probabilidade de A ocorrer. Essa relação é muito usada para resolver problemas que envolvem intersecção de eventos.

3.4 Teorema Bayesiano

O teorema de Bayes é definido como :

Teorema 1 (Teorema de Bayes). *A probabilidade de um evento C_i ocorrer, supondo que um outro evento A tenha ocorrido, é dado por:*

$$P(C_i|A) = \frac{P(C_i)P(A|C_i)}{\sum_{j=1}^n P(C_j)P(A|C_j)} \quad (3.6)$$

para todo $i = 1, 2, \dots, n$.

Assim, o teorema bayesiano mostra que a probabilidade à posteriori de um evento (C_i) condicionado à uma evidência (A) é alterada através do seguinte quociente: a probabilidade da intersecção entre esses dois eventos, representado por $P(A|C_i)$, que corresponde ao numerador da fórmula 3.6, dividido pela probabilidade de que A ocorra, dado na fórmula 3.6 pelo denominador $\sum_{j=1}^n P(C_j)P(A|C_j)$. O próximo capítulo apresenta esses conceitos implementados na estrutura de grafos probabilísticos.

Capítulo 4

Grafos Probabilísticos

4.1 Vértices e arestas

Definição 1. Um grafo é a união de um conjunto de vértices e arestas. Pode-se representá-lo por $G=(V,E)$, onde V representa o conjunto de vértices e E o conjunto de arestas. Na Figura 2 tem-se o exemplo de um grafo.

Figura 2 – Representação de um Grafo

Fonte: Elaboração Própria. Os círculos em cinza são os vértices e as linhas em preto são as arestas do grafo.

Na Figura 2, os **vértices** ou nós são identificados pelos círculos simbolizados pelas letras A, B, C, D, E e F . E as **arestas**, são representadas pelas linhas que conectam essas letras, cujas extremidades podem conter setas ou flechas orientadas. A linha que conecta os vértices A e B , A e C são exemplos de arestas.

Para representar a aresta que existe entre dois “nós” do grafo presente na Figura 2, por exemplo A e C , a seguinte notação é usada: $A \rightarrow C$. Porém, caso os vértices estejam ligados por arestas não dirigidas, como ocorre na ligação entre C e D , representa-se a conexão entre os vértices por: $C-D$

4.2 Classificação das arestas

As arestas de um grafo são classificadas em dois tipos: dirigidas ou não-dirigidas. Essa classificação varia conforme a orientação da ligação entre os “nós” do grafo (KOLLER; FRIEDMAN, 2009). A Figura 3 mostra exemplos de grafos com diferentes tipos de arestas.

Figura 3 – Tipos de Arestas

Fonte: Elaboração Própria. A figura (a) e (b) representam as arestas dirigidas. A aresta presente na figura (c) é exemplo de uma aresta não-dirigida.

Na Figura 3, as letras a e b representam **arestas dirigidas**. Arestas dirigidas conectam dois “nós” através de uma linha com sentido definido. O sentido é dado pela direção apontada pela seta na extremidade do vértice. Representa-se esse tipo de aresta da seguinte forma: $X_i \Rightarrow X_j$, onde i e j representam dois vértices distintos.

Ainda na Figura 3, a letra c representa uma **aresta não-dirigida**. Arestas não-dirigidas conectam dois “nós” através de uma linha sem direção definida. A notação para representar esse tipo de ligação é: $X_i - X_j$.

4.3 Grafos Acíclicos Dirigidos (DAGs)

Caracteriza-se pela ausência de ciclos na sua estrutura. Ciclo é um conceito que designa um tipo especial de ligação entre “nós”. Matematicamente, define-se o conceito de ciclo a partir da definição de vértices (KOLLER; FRIEDMAN, 2009).

Definição 2. Dado um conjunto V e um grafo G , onde $V = V_1, V_2, \dots, V_n$ e n representa o número de vértices que compõe o grafo G , diz-se que um ciclo num grafo G é um caminho que parte de V_1 e chega a V_n , sendo $V_n = V_1$. Na Figura 4 a seguir, a letra a, é um exemplo de DAG.

4.4 Grafos Acíclicos parcialmente dirigidos (PDAG)

Os DAGs e PDAGs compartilham uma característica em comum: ambos são acíclicos. A diferença entre essas estruturas reside na classificação das suas arestas. Enquanto que os

DAGs são compostos, exclusivamente, por arestas direcionadas, os PDAG possuem uma composição mesclada.

Esses grafos são formados por duas partes bem definidas: uma composta apenas por arestas dirigidas e outra apenas por arestas não-dirigidas, segundo Koller e Friedman (2009). A Figura abaixo mostra a diferença entre esses tipos de grafos.

Figura 4 – Representação de um DAG e um PDAG

Fonte: Elaborado pelo autor. A letra (a) representa um DAG e (b) representa um PDAG.

A letra a, da Figura 4 exibe um exemplo de grafo acíclico dirigido (DAG). Não há formação de ciclos nesse grafo e existe apenas um caminho que conecta os quatro vértices da estrutura: $A \rightarrow B \rightarrow C \rightarrow D$.

A letra b, da Figura 4 mostra um exemplo de grafo acíclico parcialmente direcionado (PDAG). Note que há dois caminhos no grafo que justificam a definição "parcialmente dirigido" atribuída para esse tipo de grafo. O caminho: $A \rightarrow B \rightarrow D$ é um exemplo de caminho composto apenas por arestas dirigidas. Já o caminho: $A - C - E - D$ é formado exclusivamente por arestas não-dirigidas.

4.5 Redes Bayesianas

O primeiro passo para definir o que são Redes Bayesianas é apresentar o modelo Naïve-Bayes. O modelo é considerado como dos mais simples onde a parametrização condicional é combinada com pressupostos de independência condicional para produzir uma representação simples de uma distribuição de probabilidade de grande dimensão (KOLLER; FRIEDMAN, 2009).

4.5.1 Modelo Naïve-Bayes

O modelo Naïve-Bayes considera que as variáveis do modelo assumem um número definido de valores mutuamente exclusivos. Suponha que exista uma variável denominada risco e que o modelo seja formado apenas por essa variável. O conjunto de valores que essa variável

pode assumir(o espaço amostral) é dado por dois eventos: risco elevado e risco pequeno. Os "valores"risco elevado e risco pequeno são mutuamente exclusivos, porque eles não podem ocorrer simultaneamente, segundo Koller e Friedman (2009).

O modelo inclui também atributos dessas variáveis, representados simbolicamente por $X_1, \dots, X_n$ e assume que esses atributos são condicionalmente independentes dado o valor da classe da variável (KOLLER; FRIEDMAN, 2009). A figura 5 abaixo representa um exemplo de modelo Naïve-Bayes.

Figura 5 – Representação do modelo Naïve-Bayes

Fonte: Adaptado de Koller e Friedman (2009, p.50)

Para compreender o modelo Naïve-Bayes de maneira mais clara suponha o seguinte exemplo. Suponha que estejamos interessados em avaliar o risco de uma empresa. A variável "Class" é a variável risco e o conjunto $X_1, \dots, X_n$ são os atributos dessa variável. Suponha ainda que X_1 seja crédito e X_2 seja poder de barganha com fornecedores.

Nesse exemplo, assumimos que quanto maior o risco de uma empresa menor a probabilidade de que ela consiga crédito em instituições financeiras e ao mesmo tempo, menor o seu poder de barganha com fornecedores. O que o modelo Naïve-Bayes nos diz é que esses dois atributos são independentes, dado a variável risco. Ou seja, não há qualquer relação entre os atributos.

4.5.2 Elementos Básicos

Para definir o que são Redes Bayesianas é preciso definir alguns elementos básicos que podem aparecer na sua estrutura. Segundo Koller e Friedman (2009):

Definição 3. (nó filho) Dado um grafo $G = (V, E)$, onde V representa o conjunto de nós do grafo e E representa o conjunto de arestas desse grafo, sempre que tivermos $X_i \rightarrow X_j$ dizemos que X_j é filho de X_i e que X_i é pai de X_j .

Definição 4. (vizinhos) Dado um grafo $G = (V, E)$, se existe uma aresta entre X_i e X_j , de forma que pode-se representar essa ligação através da seguinte notação $X_i - X_j$, dizemos que

X_i e X_j são vizinhos.

A Figura 6 a seguir descreve os principais elementos de um grafo probabilístico. Nessa estrutura composta por dez nós, o conjunto formado pelos nós G, F, C e B são os predecessores de A. Já os nós B e C são os pais de A, porque exercem influência direta sobre esse nó. Os nós descendentes de A são definidos como filhos de A e nesse caso compreendem os nós: D e E. O nó H é pai de D, assim como A.

Figura 6 – Elementos de um Grafo probabilístico

Fonte: Adaptado de Nagarajan, Scutari e Lèbre (2013, p.3)

4.5.3 Definição de Redes Bayesianas

As Redes Bayesianas são definidas como modelos gráficos, analíticos e também como modelos de decisão. A característica mais importante das Redes Bayesianas é a sua capacidade de absorver tanto conhecimentos quantitativos quanto qualitativos. Essa característica permite que possamos confiar ainda mais no modelo, já que ele engloba não apenas dados estatísticos, mas também a experiência de especialistas (ABRAMOWICZ; MAREK; SZTYKIEL, 2003).

Segundo Pavlenko e Chernyak (2010), podemos definir o conceito de Rede Bayesiana como:

Definição 5. Uma Rede Bayesiana é um grafo acíclico dirigido que codifica a distribuição de probabilidade conjunta sobre um conjunto de variáveis aleatórias $\mathcal{X} = \{X_1 \dots X_d\}$. Formalmente, pode-se definir uma BN para $\mathcal{X}$ pelo par $\langle G, P \rangle$.

G representa um grafo acíclico dirigido cujos nós são dados pelo conjunto $\mathcal{X} = \{X_1 \dots X_d\}$ e P , representado pela notação: $P = \{P(X_1 | \prod_{[i]}), \dots, P(X_d | \prod_{[d]})\}$ representa o conjunto de distribuições de probabilidade condicionais de d , dado o conjunto de nós definidos como "pai" desses vértices, representados como: $\prod_{[i]}$, para cada $X_i, i = 1, \dots, d$.

Assim, podemos dizer que se existe uma aresta dirigida entre dois vértices, por exemplo, $X_1 \rightarrow X_j$, há uma influência ou a crença de que exista uma relação de influência, entre essas duas variáveis (PAVLENKO; CHERNYAK, 2010).

4.5.4 Exemplo de Rede Bayesiana

Para compreender as atribuições de probabilidades em uma Rede Bayesiana é apresentado o seguinte exemplo adaptado de Yoon (2003). O exemplo descreve o caso hipotético de uma empresa de seguros, que recentemente implementou uma página de serviços on-line para venda de seguros básicos, como, seguro para viagens e seguro para acidentes pessoais. Para garantir a continuidade das suas operações e sua competitividade no mercado, a empresa está implementando ações para se proteger do que ela define como riscos operacionais.

Esses riscos compreendem problemas como: ataque de vírus, quedas no sistema, ação de hackers e outras ameaças que implicam em falhas na operação dos seus serviços on-line. Yoon (2003) selecionou dezessete variáveis para compor os nós da Rede Bayesiana. A relação de variáveis está detalhada na Tabela 2. A fim de medir as consequências econômicas e de mercado que esses riscos podem ocasionar à empresa Yoon (2003) definiu a variável *Cost* como variável resposta.

A variável *Cost* representa o montante financeiro a ser investido para cobrir as despesas com os riscos operacionais evitando assim que novos negócios sejam prejudicados e que a reputação da empresa seja acometida. Essa variável é influenciada por todas as outras dezesseis variáveis da rede e por isso, é a *variável resposta*. Cada um desses dezessete eventos são os nós da Rede Bayesiana, representados pelas suas respectivas abreviações através dos retângulos amarelos na Figura 7 abaixo.

Os nós foram posicionados conforme a percepção da autora quanto a influência de cada um sob a variável *Cost*.

Cada variável é associada a uma *distribuição de probabilidade condicional (CPD)*, que estabelece uma distribuição de valores de X para cada conjunto de valores atribuídos aos seus pais no modelo. Assim, para os nós que não possuem pais no modelo, a CPD é dada pela distribuição marginal de cada variável (KOLLER; FRIEDMAN, 2009). As CPDs são representadas na Rede Bayesiana da Figura 8 abaixo através de retângulos e contém o nome da variável e a probabilidade de que cada variável assuma cada um dos valores descritos na Tabela 2, na coluna *Values*.

Conforme evidências são fornecidas essas probabilidades, representadas nas CPDs, são atualizadas, de maneira que a probabilidade de que cada nó assuma um determinado valor se altera. Suponha, por exemplo, que as seguintes evidências a respeito de seis nós dessa rede sejam dadas, conforme a Tabela 3 abaixo.

A Tabela 3 atribui a cada uma das seis variáveis um valor. Esses valores são então absorvidos pela rede Bayesiana, conforme a Figura 9 a seguir. Comparando as Figuras 8 e 9 é possível notar que após a absorção das evidências na rede, a probabilidade de que a variável

Tabela 2 – Valores e Abreviações para a Rede Bayesiana do Risco Operacional

No.	Description	Values	Abbreviation
1.	Application Failure	Application corruption, Lockup, OK(No failure)	AF
2.	Cost of losses from network risk (Cost)	0.0m, 0.5m, 1.0m,	Cost
3.	Data Loss	0%, 50%, 100%	DL
4.	End User Modification	Yes, No	EUM
5.	Firewall	Application Proxy, Packet	F
6.	File Access Control	High, Low	FAC
7.	High Availability Network	Yes, No	HAN
8.	High Availability Server	Yes, No	HAS
9.	Hacker Attack	Yes, No	Hack
10.	Network Failure	Yes, No	NF
11.	Power Surge	Yes, No	PS
12.	Server Downtime	0 day, 0.5 day, 1 day	SD
13.	Server Failure	Yes, No	SF
14.	Server Hardware Quality	High, Low	SQ
15.	Transaction Downtime	0 day, 0.5 day, 1 day	TD
16.	Uninterrupted Power Sup	Yes, No	UPS
17.	Virus Attack	Yes, No	V

Fonte: Yoon (2003, p.41)

Cost(Custo) assumo o valor 0 aumenta. O valor que na Figura 8 era de 69,21% passou para 91,27%, conforme exhibe a Figura 9. Em outras palavras, podemos interpretar que a atribuição dessas evidências teve um impacto positivo sobre o custo da empresa, uma vez que, a probabilidade de que essa empresa tenha poucas despesas com riscos operacionais aumenta.

Tabela 3 – Atribuição de Evidências

Variable	Evidence
F	Application Proxy
FAC	High
HAN	Yes
HAS	Yes
SQ	High
UPS	Yes

Fonte: Adpatado de Yoon (2003, p.51)

Figura 7 – Rede Bayesiana do Risco Operacional

Fonte: Adaptado de Yoon (2003, p.42)

Figura 8 – Rede Bayesiana do Risco Operacional com CPDs

Fonte: Yoon (2003, p.50)

Figura 9 – Atribuição de Evidências na Rede do Risco Operacional

Fonte: Yoon (2003, p.52)

Capítulo 5

Base de dados

A base de dados foi obtida através de um convênio de pesquisa firmado entre a Universidade de São Paulo, a instituição SERASA S.A. inscrita no CNPJ sob o nº 62.173.620-0001/80, com sede na Alameda dos Quinimuras, 187 – Planalto Paulista – São Paulo – SP e a FUNDAÇÃO PARA PESQUISA E DESENVOLVIMENTO DA ADMINISTRAÇÃO, CONTABILIDADE E ECONOMIA, designada FUNDACE, com sede à Rua Bernardino de Campos, 1001 – sala 401- Higienópolis – Ribeirão Preto – SP, inscrita no CNPJ/MF sob o nº 00.934.542/0001-31.

Os dados estão organizados em três planilhas, que guardam informações sobre três períodos distintos: 2013-2015. Os períodos escolhidos estão de acordo com a necessidade de dados da própria instituição para a sua tomada de decisões. O último exercício é usado para avaliar a situação atual da empresa, enquanto os dois anteriores são usados para comparar e calcular as variações nos índices financeiros e contas patrimoniais que houveram para cada uma delas durante esse período. Esses três períodos são suficientes para decidir quanto à *rating* ou decréscimos no *rating* de cada empresa.

Para cada ano são disponibilizadas 240 contas patrimoniais, divididas entre contas de ativo, passivo e algumas contas de resultado. As observações são identificadas por um conjunto de quatro variáveis: um código confidencial atribuído pela Serasa, o ano a que pertence os valores associados aquela observação, o ramo de atuação em que a instituição enquadra-se e o *rating* atribuído para cada empresa. No total, são 852 empresas distribuídas em 155 ramos de atuação. A Tabela 4 abaixo mostra a quantidade de empresas distribuídas em cada um desses ramos.

Tabela 4 – Distribuição de empresas por segmentos de indústrias

EMPRESAS	NÚMERO DE EMPRESAS
USINA DE AÇÚCAR E	
ÁLCOOL	47
IND DE PRODUTOS INTERMEDIÁRIOS QUÍMICOS	32
IND DE PRODUTOS FARMACÊUTICOS	30
ABATE E PREPARAÇÃO DE BOVINOS	28
IND DE ADUBOS E FERTILIZANTES	25
IND DE AUTOPEÇAS E ACESSÓRIOS	22
IND DE LATICÍNIOS	18
IND DE MECÂNICA PESADA	18
ABATE E PREPARAÇÃO DE AVES	16
IND DE CANOS E TUBOS DE FERRO E AÇO	15
IND DE CALÇADOS	14
IND DO PAPEL	14
IND DE CIMENTO E CLINQUER	13
IND DE MECÂNICA LEVE	13
BENEFICIAMENTO DE ARROZ	12
IND DE EMBALAGENS PLÁSTICAS	12
IND DE SOJA	12
IND DE CONFECCÕES EM GERAL	11
IND DE DERIVADOS DE PETRÓLEO	11
IND DE MÁQUINAS E EQUIPAMENTOS AGRÍCOLAS	11
IND DE PISOS E AZULEJOS	11
METALURGIA – METAIS NÃO FERROSOS FORMA PRIMÁRIA	11
EDITORIA DE LIVROS	10
IND DE ADITIVOS INDLS E ÓLEOS VEGETAIS E ANIMAIS EM BRUTO	10
IND DE RAÇÕES	10
IND DE RESINAS	10
IND DE TRIGO E DERIVADOS	10
IND DE CABOS	8
IND DE GERADORES	8
IND DE LAMINADOS DE AÇO	8
IND DE PRODUTOS VETERINÁRIOS	8
IND DE REFRIGERANTES	8
IND DE TECIDOS SINTÉTICOS E ESPECIAIS	8
IND DE ARTEFATOS DE PAPEL	7
IND DE CAFÉ	7
IND DE EQUIPAMENTOS DE COMUNICAÇÃO	7
IND DE EQUIPAMENTOS DE INFORMÁTICA	7
IND DE LIGAS DE METAIS	7

Fonte: Elaboração própria

5.1 Classificação das Variáveis

Para possibilitar o cálculo dos indicadores financeiros os dados presentes na amostra foram classificados segundo quatro parâmetros: a classe numérica, o tipo de conta, abreviação, uma breve descrição. Esses parâmetros auxiliam a organizar as informações sobre os dados, ao mesmo tempo que, facilitam a aplicação de ferramentas estatísticas sobre a base de dados.

Tabela 5 – Classificação das variáveis

Tipo	Classificação	Código	Descrição
Categórica	-	UF	Estado
Categórica	-	ANO	Ano do balanço
Categórica	-	RISCO	Risco de crédito
Número	Passivo	PASC	Passivo Circulante
Número	Passivo	EXLP	Exigível a Longo Prazo
Número	Passivo	PL	Patrimônio Líquido
Número	Passivo	DEPAC	Depreciação Acumulada
Número	Ativo	ATVPER	Ativo Permanente
Número	Ativo	DISP	Disponível
Número	Ativo	APLFIN	Aplicações Financeiras
Número	Ativo	CLIENTES	Clientes
Número	Passivo	DIVINST	Dívidas com instituições financeiras
Número	Passivo	IMPOSTOS	Impostos
Número	Ativo?	FATLIQ	Faturamento Líquido
Número	DRE	CMV	Custo da mercadoria vendida
Número	Passivo	FORN	Fornecedores
Número	Passivo	DIVFIN	Dívidas com instituições financeiras
Número	Ativo	ATVTOT	Ativo Total
Percentual	Índice	ENDON	Endividamento
Percentual	Índice	CDEXD	Composicao das exigibilidades
Percentual	Índice	ODAI	Obsolescência do Ativo Imobilizado
Percentual	Índice	IDPL	Imobilização do Patrimônio Líquido
Número	Índice	LQG	Liquidez Geral
Número	Índice	LQC	Liquidez Corrente
Número	Índice	LQS	Liquidez Seca
Percentual	Índice	PFNEND	Participação de Financiamento no Endividamento
Percentual	Índice	PFNATCR	Participação de Financiamento no Ativo Circulante
Número	Índice	PMRV	Prazo médio de recebimento de Vendas
Número	Índice	GDATV	Giro do Ativo

Tipo	Classificação	Código	Descrição
Percentual	Índice	RTBV	Rentabilidade de Vendas
Número	Índice	PMRE	Prazo médio de Renovação de Estoques
Número	Índice	PMPC	Prazo médio de pagamento de compras
Número	Índice	CICLFN	Ciclo Financeiro
Percentual	Índice	RENTPL	Rentabilidade do Patrimônio Líquido
Percentual	Índice	EVRV	Evolução Real de Vendas
Percentual	Índice	ROI	Retorno sobre investimento

Fonte: Elaboração própria.

A variável Tipo classifica as variáveis em três tipos, que variam de acordo com a forma que as variáveis são medidas. Contas patrimoniais medidas em percentuais são do tipo percentual, já contas medidas em valores nominais pertencem a categoria número, por exemplo. As variáveis do Tipo categóricas são aquelas que assumem uma quantidade definida de valores em um intervalo. Ano, por exemplo, assume apenas três valores: 2013, 2014 ou 2015. A variável Risco, assume uma quantidade definida pelo conjunto: $\{1, 20\}$, sendo 1 a classe de risco com menor probabilidade de inadimplência e 20, a classe com maior risco de inadimplência.

5.2 Indicadores financeiros selecionados

A partir das contas patrimoniais disponíveis foram escolhidos um conjunto de indicadores financeiros que poderiam ser calculados a partir desse conjunto. Esse grupo formado por 21 indicadores mais a variável risco correspondem aos mesmos parâmetros usados pela Serasa Experian nas suas decisões de crédito. Os indicadores financeiros medem quatro diferentes aspectos de uma mesma empresa: o nível de estrutura de capitais, o grau de relacionamento com instituições de crédito, o grau de liquidez, além de acompanhar a evolução dos resultados.

Esses indicadores juntos a análise de outros aspectos da empresa como: formação acadêmica dos proprietários e gestores do alto escalão, tradição ou anos de atuação do mercado, por exemplo, ajudam a compor nota de risco que será atribuída a cada empresa. Além disso, esses indicadores também são importantes para rever as notas de risco atribuídas. Anualmente essas notas são revisadas porque as empresas sofrem variações ao longo de cada ano que podem prejudicar sua capacidade em arcar com novas dívidas ou despesas, em virtude de oscilações econômicas e outras variáveis macroeconômicas que podem afetar sua saúde financeira. Esses indicadores auxiliam o gestor de crédito a avaliar a nota de risco atribuída a essas empresa, na medida que esses índices detectam de forma rápida e menos trabalhosa essas oscilações.

A Tabela 6 abaixo apresenta os índices financeiros e as respectivas siglas adotadas.

Tabela 6 – Indicadores financeiros: abreviações e significados.

INDICES FINANCEIROS	SIGLAS
Endividamento	EDV
Composição das exigibilidades	CDEXD
Imobilização do patrimônio líquido	IDPL
Obsolescência do ativo imobilizado	ODAI
Participação de Financiamentos no Ativo	PFNATV
Endividamento Oneroso	ENDON
Participação de Financiamentos no Endividamento	PFNEND
Participação de Financiamento no Ativo circulante	PFNATCR
Nível de Desconto de Duplicatas	NDD
Liquidez Geral	LQG
Liquidez Corrente	LQC
Liquidez Seca	LQS
Prazo médio de recebimento de Vendas	PMRV
Prazo médio de renovação de Estoques	PMRE
Prazo médio de pagamento de compras	PMPC
Ciclo Financeiro	CICLFN
Giro do Ativo	GDATV
Rentabilidade do Ativo Médio	ROI
Rentabilidade do Patrimônio Líquido médio	RTPL
Rentabilidade das vendas	RTBV
Evolução real de vendas	EVRV

Fonte: Elaborado pelo autor.

As fórmulas adotadas pela Serasa para o cálculo desses 21 indicadores financeiros encontram-se descritas abaixo.

$$EDV = (\text{Passivo Circulante} + \text{Exigível a Longo Prazo} / \text{Patrimônio Líquido}) * 100$$

$$CDEXD = (\text{Passivo Circulante} / (\text{Passivo Circulante} + \text{Exigível a Longo Prazo})) * 100$$

$$ODAI = (\text{Depreciação Acumulada} / \text{Imobilizado}) * 100$$

$$IDPL = (\text{Ativo Permanente} / \text{Patrimônio Líquido}) * 100$$

$$LQG = (\text{Ativo Circulante} + \text{Realizável a Longo Prazo}) /$$

(Passivo Circulante + Exigível a Longo Prazo)

$$LQC = (\text{Ativo Circulante} / \text{Passivo Circulante})$$

$$LQS = (\text{Disponível} + \text{Aplicações Financeiras} + \text{Clientes} / \text{Passivo Circulante})$$

$$PFNEND = (\text{Dívidas com instituições financeiras} / \text{Passivo Circulante} + \text{Exigível a Longo Prazo}) * 100$$

$$PFNATCR = (\text{Dívidas de curto prazo com instituições financeiras} / \text{Ativo Circulante}) * 100$$

$$PFNA = (\text{Dívidas com instituições financeiras} / \text{Ativo Total}) * 100$$

$$PMRV = [\text{Saldo médio de duplicatas a receber} / ((\text{Faturamento Líquido} + \text{Impostos}) / (\text{PDR} * 12))] * 360$$

$$GDATV = (\text{Faturamento Líquido} / \text{PDR} * 12) / \text{Ativo Total}$$

$$RTBV = (\text{Resultado do Exercício} / \text{Faturamento Líquido}) * 100$$

$$PMRE = [\text{Estoque Médio} / (\text{Custo dos produtos ou mercadorias vendidas} / (\text{PDR} * 12))] * 360$$

$$PMPC = [\text{Saldo Médio de Fornecedores} / (\text{Compras} / (\text{PDR} * 12))] * 360$$

$$CICLFN = (PMRV + PMRE) - PMPC$$

$$RENTPL = (\text{Resultado do exercício} / \text{PDR} * 12) / \text{Patrimônio Líquido Médio}$$

$$EVRV = [\text{Faturamento Líquido Atual} / (\text{Faturamento Líquido Anterior} / \text{Somatório do IGP-DI médio})] * 100$$

$$ROI = [(\text{Resultado do exercício} / \text{PDR} * 12) / \text{Ativo Total Médio}]$$

$$ENDON = [\text{Dívidas com instituições financeiras} / \text{Patrimônio Líquido}] * 100$$

Capítulo 6

Análise dos dados

Para obter o primeiro esboço da Rede Bayesiana foi necessário adotar a padronização das variáveis explicativas da rede. O procedimento de padronização pela mediana justifica-se pela quantidade e amplitude de valores. Se multiplicarmos o número de empresas pelo total de indicadores contáveis temos 5.454 dados. Essa quantidade de dados implica em um número elevado de *outliers* que poderiam ser tratados como erros ou descartados da análise. A fim de garantir que todos os dados sejam computados e facilitar a compreensão da posição desses valores na rede adota-se a padronização descrita na Seção 6.1.

6.1 Padronização e categorização dos dados

Para obter o primeiro desenho da Rede Bayesiana que modela o risco de crédito das empresas adotamos duas técnicas: a padronização e a categorização dos dados. A padronização consiste na escolha de uma função científica matemática que padroniza os dados da amostra, de maneira que todas as observações tenham como referência um valor central e sejam fornecidas em termos de uma unidade padrão. Usualmente, define-se “0” como sendo o valor dentro dessa escala que representa a média ou então a mediana e “1” como a distância de um desvio padrão, ou então a distância interquartil. A fórmula adotada para a padronização é descrita como:

$$Z_i = \frac{v_i - \tilde{v}}{IQR} \quad (6.1)$$

Na qual:

Z_i : valor do indicador padronizado

v_i : valor do indicador financeiro antes da padronização

$\tilde{v}$: é a mediana dos valores de cada indicador

IQR : Distância Interquartil

Os resultados da padronização efetuada foram sintetizados através de uma Análise Exploratória de dados. Essa análise compreende um conjunto de medidas descritivas, medidas de centro e de posição que auxiliam a obter o sentido prático dos resultados estatísticos. As medidas de centro são a mediana e a média do conjunto de dados, enquanto que as medidas

de posição são os quartis e percentis. Essas medidas são apresentadas para seis dos vinte e um indicadores selecionados, na Tabela 7 abaixo.

Tabela 7 – **Resumo das estatísticas descritivas dos 20 indicadores financeiros para o ano de 2013**

Medidas Estatísticas	EDV	CDEXD	ODAI	IDPL	LQG	LQC
Mínimo	-23,4493	-2,4109	-0,3565	-8,3912	-2,1948	-1,827
1st Quartil	-0,3556	-0,5341	-0,3565	-0,4295	-0,4516	-0,4236
Mediana	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000
Média	1,5653	-0,1024	0,2556	1,3537	0,4402	0,4102
3rd Quartil	0,6444	0,4659	0,6435	0,5705	0,5484	0,5764
Máximo	148,4437	1,2816	4,953	288,2347	19,8022	14,2731

Fonte: Elaborado pelo autor. As estatísticas descritivas foram obtidas através da padronização pelo software RStudio.

A Figura 10 mostra os resultados após a padronização dos dados.

Figura 10 – **Diagrama de caixa para os 20 indicadores**

Fonte: Elaboração própria através do software RStudio.

Note que, após aplicar a técnica de padronização, analisar a variabilidade dos indicadores e a proporção de valores em cada percentil torna-se mais intuitivo. A padronização pela mediana garante ainda que, as observações com valores iguais a mediana sejam 0 e as demais sejam medidas pelo número de distâncias inter-quartis em relação à mediana.

As próximas seções dedicam-se a aplicação de técnicas de Análise Exploratória de dados (EDA) para inferir características do comportamento dos indicadores financeiros das empresas selecionadas.

6.2 Correlação

A correlação existe entre duas variáveis quando elas estão relacionadas de alguma maneira. Quando essa relação segue ou aparenta possuir um padrão linear, de forma que os dados distribuem-se em torno de uma reta, há uma grande probabilidade de que exista uma correlação linear entre esses dados, que pode ser medida através do coeficiente de correlação linear (TRIOLA, 2009).

O coeficiente de correlação linear é definido como:

$$r = \frac{n(\sum xy) - (\sum x)(\sum y)}{\sqrt{n(\sum x^2) - (\sum x)^2} \sqrt{n(\sum y^2) - (\sum y)^2}} \quad (6.2)$$

Notações usadas:

r : coeficiente de correlação linear

n : quantidade de pares(x,y) da amostra

$\sum x$: somatório dos valores de x

$\sum x^2$: somatório dos valores de x ao quadrado

$\sum x^2$: o quadrado da soma dos valores de x

$\sum xy$: soma dos valores resultantes da multiplicação $x \times y$

Essas definições dão sustentação para implementar o coeficiente de correlação linear sobre a nossa amostra de empresas com o objetivo de identificar de maneira preliminar a posição dessas variáveis na rede.

6.3 Correlação entre 20 indicadores

Conforme vimos na seção 6.2, para afirmar que há correlação linear entre quaisquer variáveis é preciso garantir que três condições sejam satisfeitas: a amostra de dados deve ser aleatória, o diagrama de dispersão dos dados deve mostrar um comportamento linear e os valores extremos da amostra, quando se tratarem de erros, devem ser retirados da amostra. A primeira condição é facilmente demonstrada. A única característica em comum entre os dados das empresas selecionadas pela Serasa Experian é o faturamento anual, superior à R\$200.000.000. A segunda condição é demonstrada pela Figura 11 abaixo. O gráfico de dispersão foi elaborado a partir do coeficiente de correlação r entre os indicadores financeiros selecionados e variável risco.

A Figura 11 representa no eixo x os 19 indicadores financeiros e no eixo y , os valores do coeficiente de correlação entre esses indicadores e o risco. Adicionando a linha de tendência ao gráfico, representada pela cor azul, é possível notar que existe uma relação linear entre os indicadores e o risco. Note que os pontos se ajustam de maneira satisfatória à reta, distribuindo-se de maneira uniforme em torno dela.

Os vinte índices observados podem ainda ser agrupados em dois conjuntos: o grupo dos índices que mantém relação linear direta com a variável risco e aqueles que estabelecem uma relação linear inversa. Esses dois grupos são muito visíveis na Figura 12, na qual, os índices

Figura 11 – Gráfico de dispersão da correlação entre 19 indicadores e a variável risco

Fonte: Elaboração própria. O gráfico foi elaborado através do pacote *ggplot2* do software *R*.

situados a esquerda do gráfico, com valores predominantemente negativos representam o grupo das variáveis inversamente correlacionadas ao risco e os índices situados a direita, com valores predominantemente positivos, representam o grupo das variáveis diretamente correlacionadas ao risco.

Figura 12 – Correlação entre os 19 indicadores e a variável risco

Fonte: Elaboração própria. O gráfico foi elaborado através do pacote *ggplot2* do software *R*.

Os dois grupos formados são compostos pelos seguintes índices financeiros:

- **Indicadores financeiros inversamente relacionados ao risco:** IDPL, ENDON, CDEXD, RTBV, EDV, PFNA, PFNEND, GDATV, PMRV.

- **Indicadores financeiros diretamente relacionados ao risco:** CICLFN, LQG, ROI, ODAI, LQS, LQC, PFNATCR, EVRV, PMPC, PMRE.

É importante notar que os indicadores com maior valor absoluto da correlação são os que tem maior impacto no modelo, pois espera-se que a maior contribuição ao modelo esteja em relações lineares.

6.4 Categorização

Essa seção apresenta o processo de categorização adotado no trabalho, assim como demonstra a necessidade desse processo para o cálculo de probabilidades efetuado na Rede Bayesiana. Antes disso porém, é apresentado um exemplo clássico envolvendo grafos e probabilidades condicionais: o exemplo do Exame Clínico. Esse exemplo auxilia a compreender a utilidade de aplicar a categorização em Redes Bayesianas e também ilustra a forma como ocorrem o cálculo de probabilidades na rede.

6.4.1 Exemplo do Exame Clínico

A atribuição de probabilidades condicionais na Rede Bayesiana ocorre através de classes ou categorias. Cada variável que compõe a rede assume uma quantidade discreta de valores, que podem ser numéricos ou alfabéticos. A Figura 13 ilustra o processo de atribuição de valores na rede.

Figura 13 – Rede Bayesiana do Exame Clínico

Fonte: Højsgaard, Edwards e Lauritzen (2012)

O exemplo retrata um conjunto de variáveis que influenciam e podem causar a dificuldade respiratória de um indivíduo: asia, smoke, tub, lung, bronc, either e dysp. A variável "asia" indica se um indivíduo esteve no continente Asiático ou não.

A variável "smoke" indica se a pessoa é fumante, já a variável "tub" representa a doença tuberculose e indica se a pessoa possui ou não essa enfermidade. A variável "either" indica se um indivíduo se encaixa em uma das três situações: o indivíduo possui tuberculose e câncer pulmonar, ou o indivíduo possui apenas tuberculose ou o indivíduo possui apenas câncer pulmonar. Nessas três situações a variável "either" recebe valor positivo. A última variável presente na rede é a variável "dysp". Essa variável indica se o indivíduo possui ou não dificuldade respiratória (dyspnoea) e é a variável resposta nessa Rede Bayesiana.

A dificuldade respiratória pode ser causada por algumas doenças como: tuberculose, câncer pulmonar ou bronquite. Porém, a dificuldade respiratória pode ser causada não por apenas uma dessas doenças, mas por mais de uma ou também por nenhuma das doenças citadas.

Portanto, para obter as probabilidades de que o indivíduo tenha dificuldade respiratória cada uma dessas sete variáveis (a variável "dysp" não é apontada porque é a variável resposta) são categorizadas em duas classes: "yes" ou "no". Quando o indivíduo afirma possuir a enfermidade citada, ela recebe o valor positivo "yes". Porém, quando ele não a possui, a mesma recebe o valor negativo "no". Quando não há evidências sobre a variável, nenhum valor é assumido por ela. Os valores das variáveis são atualizados conforme as evidências são absorvidas dentro da rede.

Para ilustrar a atribuição de probabilidades pela Rede Bayesiana, vamos supor que tenhamos as seguintes evidências: um indivíduo esteve na Ásia recentemente e tem dificuldade respiratória. A partir dessas evidências deseja-se calcular a probabilidade de que esse indivíduo tenha câncer pulmonar ou bronquite.

O primeiro passo para obter essa probabilidade é alimentar a Rede Bayesiana da Figura 13 com dados. Esses dados são um conjunto de observações que resultam em valores positivos ("yes") ou negativos ("no") para a presença de dificuldade respiratória. Højsgaard, Edwards e Lauritzen (2012) adotam um conjunto de dados definido como **chestSim500**. Esse conjunto é ilustrado na Tabela 8 abaixo.

Tabela 8 – Base de dados chestSim500

	Asia	Tub	Somke	Lung	Bronc	Either	Xray	Dysp
1	no	no	no	no	yes	no	no	no
2	no	no	no	no	yes	no	no	yes
3	no	no	yes	no	yes	no	no	yes
4	no	no	yes	no	yes	no	yes	no
5	no	no	yes	no	yes	no	no	yes
6	no	no	no	no	no	no	no	yes
7	no	no	yes	no	yes	no	no	yes
8	no	no	no	no	no	no	no	no
9	no	no	no	no	no	no	no	no
10	no	no	no	no	yes	no	no	yes
11	no	no	no	no	no	no	no	no
12	no	no	no	no	no	no	no	no
13	no	no	yes	yes	yes	yes	yes	yes

Fonte: Adaptado de Højsgaard, Edwards e Lauritzen (2012). A figura apresenta apenas uma parte das 500 observações que compõe a base de dados.

Os números na primeira coluna da figura representam o número da observação. As colunas seguintes são as oito variáveis que formam a Rede Bayesiana e os respectivos valores que elas assumem em cada uma das observações.

A partir desse conjunto de observações podemos retomar nossa pergunta inicial: qual a probabilidade de que um indivíduo tenha câncer pulmonar ou bronquite, tendo como evidências sua visita à Ásia e a presença de dificuldade respiratória? Essa situação pode ser resumida conforme a Tabela 9 abaixo.

Tabela 9 – Exemplo de atribuição de evidências

	asia	smoke	tub	lung	either	bronc	dysp
1	yes	?	?	?	?	?	yes

Fonte: Elaboração de autoria própria

Alimentando a Rede Bayesiana representada na Figura 13 com os dados de **chestSim500** e atribuindo valores positivos para as variáveis "asia" e "dysp", conforme a Tabela 9, chegamos aos seguintes valores descritos na Tabela 10 abaixo: Na Tabela 10, a probabilidade de que o indivíduo com as características citadas na Tabela 9 possua câncer pulmonar, representado pela variável "lung", é de 0.055, enquanto a de que não possua é de 0.945. As chances de que esse mesmo indivíduo possua bronquite, representado pela variável "bronc", é de 0.45 e de que não possua é de 0.55.

Tabela 10 – Atribuição de probabilidades pela Rede Bayesiana

lung		bronc	
yes	no	yes	no
0.055	0.945	0.45	0.55

Fonte: Adaptado de Højsgaard, Edwards e Lauritzen (2012)

A Rede Bayesiana proposta nesse trabalho será categorizada de forma semelhante à Rede Bayesiana do exame clínico. Cada variável da rede poderá assumir um número n de classes que serão representadas por letras. Essas letras indicam a distância de cada variável em relação a mediana (neste caso igual à 0).

Essa atribuição de classes auxiliará na atribuição de probabilidades condicionais, representadas por CPTs, que determinam a classe de risco de uma empresa, a partir das classes atribuídas à cada um dos indicadores de que se tem evidências dentro da rede.

6.4.2 Criação de categorias de dados

A primeira tentativa de categorizar os indicadores foi através de criação de quatro classes de dados. Essas classes segregam os valores dos indicadores em dois tipos: aqueles cujos valores estão acima da mediana e aqueles cujos valores estão abaixo dela. Devido à padronização adotado nesse trabalho, o valor da mediana é o mesmo para todos os indicadores: 0. Dessa forma, os valores acima da mediana pertencem às classes A ou B, enquanto os valores abaixo pertencem à C ou D. As classes propostas para categorizar as variáveis seguem a seguinte ordenação descrita na Tabela 11:

Tabela 11 – Classes de Risco adotadas para as variáveis

<i>Classes</i>	<i>Descrição das Classes</i>	<i>Porcentagem de dados</i>
A	$X > 0.5$	25,88%
B	$0 < X < 0.5$	18,59%
C	$-0.5 < X < 0$	27,74%
D	$-\infty < X < -0.5$	27,77%

Fonte: Elaboração de autoria própria

As classes da Tabela 11 classificam cada um dos 2 indicadores em 4 intervalos. A classe **A** é formada pelos indicadores cujos valores estão acima de 0.5. A Classe **B** é constituída pelos indicadores cujos valores são maiores do que 0, mas menores do que 0.5. A classe **C** é composta pelos valores acima de -0.5 e menores do que 0. A classe **D** é formada pelos indicadores menores que -0.5.

A última coluna da Tabela 11 mostra a porcentagem de indicadores das empresas em nosso banco de dados que se enquadram em cada uma dessas classes. Buscamos segregar as classes de maneira uniforme, de forma que garantíssemos poucas classes, com concentração proporcional de dados.

6.4.3 Esboço da Rede Bayesiana inicial

As técnicas de Análise de Dados Exploratória aplicadas aos indicadores selecionados inicialmente permitiram a elaboração do primeiro esboço da Rede Bayesiana, representado pela Figura 14 abaixo.

Figura 14 – Rede Bayesiana Inicial

Fonte: Elaboração própria através do software *Gephi* 0.9.2.

A Figura 14 representa a primeira ilustração da Rede Bayesiana composta por 21 indicadores financeiros. Os indicadores estão classificados de acordo com a classe de modularidade a que pertencem. Assim, os indicadores de cor vermelha representam os indicadores pertencentes à uma classe de modularidade, os de azul à uma segunda e os de verde à uma terceira e os de cinza à uma quarta classe. A modularidade é uma medida de teoria de redes que segrega os nós da rede segundo o grau de interação entre eles. Assim, as classes de modularidades são comunidades que agrupam os nós com maior nível de interação. O nível de interação constitui o número de ligações, representada por arestas, que os indicadores estabelecem entre si. Assim, as quatro classes de modularidade obtidas indicam possíveis grupos de indicadores, que formarão comunidades na futura rede bayesiana.

Capítulo 7

Modelagem

O primeiro passo para desenvolver uma Rede Bayesiana capaz de atribuir o risco para uma empresa a partir de um conjunto de indicadores financeiros obtidos dela é selecionar um modelo, processo descrito por Højsgaard, Edwards e Lauritzen (2012) como "Model Selection".

O uso de grafos para a representação de modelos muda o foco da estimação de parâmetros para a estimação da estrutura do modelo. Essa é uma tarefa desafiadora porque a quantidade de modelos possíveis é muito grande. Uma estrutura composta por 30 nós, por exemplo, possui 2^{435} possibilidades, número superior à quantidade de átomos observada no universo (10^{80}), segundo Højsgaard, Edwards e Lauritzen (2012).

Dentre os modelos existentes para representar a Rede Bayesiana dos indicadores selecionamos apenas os modelos do tipo Grafos Gaussianos Dirigidos (DGGM), uma vez que, a Rede Bayesiana proposta pressupõe a existência de arestas dirigidas entre os nós da sua estrutura. Os três principais modelos desse tipo são obtidos pelos algoritmos Greedy Search, Algoritmo Pc e o Algoritmo Híbrido. Nesse capítulo será testada a implementação dos três algoritmos e a estrutura de rede que melhor representar a relação entre os indicadores será a rede bayesiana adotada neste trabalho.

7.1 Greedy Search

O algoritmo Greedy Search é um método para modelagem de redes que busca encontrar a estrutura por trás de um conjunto de dados. O algoritmo é implementado através da função `hc(hill-climbing)` do pacote `bnlearn` (Bayesian Network Structure Learning), pertencente ao software R. O processo de busca ocorre através da maximização do valor do critério de informação Bayesiano (BIC).

A primeira etapa da modelagem é definir os nós que irão formar a rede. O usuário tem duas opções: adotar um grafo “vazio” ou usar uma estrutura prévia. O grafo vazio é formado pelos dados inputados ao algoritmo, sem arestas e sem a presença de qualquer tipo de estrutura. A estrutura prévia porém, é obtida a partir de um DAG (Grafo acíclico dirigido) resultante da implementação de um algoritmo que já foi implementado sobre esses mesmos dados,

por exemplo, o Algoritmo PC.

A segunda etapa do algoritmo é a maximização. Após a definição da estrutura da rede o algoritmo testa cada uma das arestas da rede até obter o maior valor para um critério de eliminação, que pode ser o critério de informação Bayesiano (BIC) ou o critério de Akaike (AIC). Por padrão, o algoritmo Greedy Search usa o critério BIC.

O processo de maximização percorre cada um nós da estrutura e testa a remoção, adição ou mudança na orientação de cada uma das arestas da rede até obter a maior pontuação possível pelo critério BIC. A rede final é a aquela que maximizou esse critério. O nome Greedy Search (busca gananciosa) é devido à essa etapa do processo, uma vez que, o algoritmo percorre todos os nós da rede até que a maior pontuação seja obtida. A aplicação desse algoritmo na base de dados resultou na rede representada pela Figura 15 abaixo.

Figura 15 – Rede obtida através do algoritmo Greedy Search

Fonte: Elaboração própria do autor. *Layout Circular aplicado através do pacote igraph, RStudio.*

A rede 15 apresenta os 22 nós e as respectivas arestas dirigidas que formam essa estrutura. As relações representadas na estrutura estão de acordo com as percepções sobre o relacionamento entre os indicadores, exceto pela variável Risco. Assume-se que a variável Risco é a variável resposta da Rede, portanto, essa variável não exerce influência sobre nenhuma outra variável da Rede. Na Figura 15, a variável Risco influencia duas variáveis: LQG e LQC. Uma vez que isso vai contra os pressupostos assumidos, essa estrutura não foi adotada como representação da rede bayesiana.

7.2 Algoritmo Híbrido

É um algoritmo para encontrar a estrutura por trás de um conjunto de dados que busca, assim como o algoritmo Greedy Search, maximizar o critério BIC. A diferença entre esse algoritmo e o Greedy Search é a função. O algoritmo híbrido combina um algoritmo do tipo

constraint-based (modelo em que os resultados são limitados a um intervalo) usado para determinar o esqueleto da rede à um algoritmo do tipo greedy hill-climbing, que determina as direções das arestas na rede. O algoritmo é atribuído à Tsamardinos, Aliferis e Statnikov (2003) e é implementado pela função mmhc (max-min hill-climbing) do pacote bnlearn.

O algoritmo é implementado em duas etapas: o primeiro passo é a definição do esqueleto da rede obtido através da aplicação de um teste de significância, com alfa de 5%. A segunda etapa é a aplicação do critério de informação BIC. A rede obtida através do algoritmo Hybrid é representada na Figura 16 a seguir.

Figura 16 – Rede obtida através do algoritmo Hybrid

Fonte: Elaboração própria do autor

A estrutura de redes obtida pelo algoritmo está de acordo com os pressupostos assumidos para a rede bayesiana. Porém a decisão quanto ao melhor modelo para os dados depende da implementação do algoritmo PC.

7.3 Algoritmo PC

O Algoritmo Pc desenvolvido por Spirtes e Glaymour (1991) é implementado computacionalmente através do pacote pcalg do R. O algoritmo é implementado em duas etapas: a primeira produz um grafo completo não dirigido (Undirected Graph) definido como "esqueleto" da rede. A segunda etapa atribui as direções às arestas do grafo, de acordo com as relações de independência condicionais identificadas.

7.3.1 Primeira etapa

O algoritmo pressupõe que todos os vértices estão conectados por uma aresta. Então, para testar a significância dessas ligações e verificar quais arestas permanecem no grafo e quais

são retiradas, o algoritmo adota o teorema de independência condicional enunciado abaixo.

Teorema 2. *Para todo u e v vértices de um grafo, dizemos que u e v não são adjacentes (ou vizinhos) em um grafo, se, e somente se, existe um subconjunto S , contido em V de maneira que u é condicionalmente independente de v dado o subconjunto S .*

Para obter o esqueleto da rede, são calculadas nessa etapa: a matriz de covariância, a matriz de correlação e o teste de independência condicional via Z de Fisher. A matriz de covariância testa a independência dos nós presentes na rede. De maneira simplificada, quando a covariância entre duas variáveis é 0, podemos dizer que essas variáveis são independentes.

A matriz de correlação, testa se existe correlação entre as variáveis. De maneira geral, quanto maior a correlação entre duas variáveis, maior a probabilidade de que esses nós estejam conectados ou que estabeleçam um tipo de ligação na Rede.

O teste de independência condicional via Z de Fisher é um método para testar a independência entre as variáveis que compõe a Rede Bayesiana. A estatística de teste é expressa pela fórmula abaixo:

$$Z = \sqrt{n - |S| - 3} * \log((1 + r)/(1 - r))/2 \quad (7.1)$$

que possui distribuição normal sob a hipótese nula de correlação 0. Assim, para um dado r e um dado nível de significância α , se a estatística de teste Z está na região de rejeição a função *condIndFisher* retorna um valor lógico que indica se as variáveis analisadas são condicionalmente independentes dado S ou não.

Para um nível de significância $\alpha = 0.05$ e vinte um indicadores, o resultado é a estrutura descrita na Figura 17.

Figura 17 – Esqueleto da Rede Bayesiana

Fonte: Figura obtida através do Gephi 0.9.2. O layout foi obtido através de uma combinação de distribuições do Gephi.

Na Figura 17 os vértices EDV, IDPL e ENDON formam um grafo acíclico a parte da estrutura, e portanto não exercem influência sobre a variável Risco. Os vértices NDD e RTPL não estabelecem ligações com nenhum outro nó do grafo. Esses vértices são retirados da Rede Bayesiana e portanto, não aparecem na próxima etapa do algoritmo.

7.3.2 Segunda etapa

A estrutura da rede é transformada em um objeto do tipo grafo acíclico completo parcialmente dirigido (CPDAG). Essa estrutura é caracterizada por conter um conjunto de nós dirigidos e outro conjunto não-dirigido. Para obter essa estrutura o algoritmo realiza dois passos: identifica as v-structures presentes na rede e atribui direções para as arestas restantes não-dirigidas. Esse processo resulta na seguinte estrutura descrita na Figura 18 a seguir.

Figura 18 – Esqueleto da Rede Bayesiana com direções

Fonte: Figura obtida através do *Gephi 0.9.2*. O layout foi obtido através de uma combinação de distribuições do *Gephi*.

Para selecionar o modelo que representará os dados, foram comparados os algoritmos Híbrido e o algoritmo PC. O critério para decidir qual a rede mais adequada para representar a relação entre os dados foi baseado em dois parâmetros: menor número de arestas e maior número de variáveis diretamente relacionadas a variável risco. Uma vez que ambas possuíam 21 arestas na sua estrutura, o critério que decidiu quanto a melhor estrutura foi o número de indicadores que influenciavam diretamente a variável risco. Assim, uma vez que a rede resultante do algoritmo PC possuía quatro indicadores influenciando o risco, frente os três do algoritmo Híbrido, a rede Bayesiana selecionada é a rede resultante do algoritmo PC, presente na Figura 18. O próximo capítulo apresenta a aplicação do método de Regressão Logística sobre a Rede Bayesiana presente na Figura 18.

Capítulo 8

Regressão Logística

Os métodos de regressão tornaram-se técnicas essenciais para análise de dados interessantes em descrever a relação entre modelos compostos por uma variável resposta e uma, ou mais, variáveis explicativas. Dentre esses métodos, a regressão logística consolidou-se como método padrão para a análise de dados em diversos campos de estudo (HOSMER; LEMESHOW, 2000).

A regressão logística é a técnica mais usada para analisar o desempenho de um modelo, mesmo com a presença de técnicas como: regressão linear, programação matemática, árvores de classificação e redes neurais. Outras técnicas como algoritmos genéticos, support vector machines e nearest neighbours tem sido testados, mas ainda não se consolidaram na indústria, segundo Crook, Edelman e Thomas (2007).

8.1 Modelo de regressão logística múltipla

O modelo de regressão logística múltipla é definido por Hosmer e Lemeshow (2000) conforme a definição 6 abaixo.

Definição 6. Considere um conjunto de p variáveis independentes. p é representado pelo vetor: $x' = (x_1, x_2, \dots, x_p)$. Assuma que a probabilidade condicional da variável resposta (Y) é denotada por $P(Y = 1|x) = \pi(x)$. O modelo logit de regressão múltipla é dado pela equação:

$$g(x) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_p x_p \quad (8.1)$$

quando o modelo de regressão logística é definido como:

$$\pi(x) = \frac{e^{g(x)}}{1 + e^{g(x)}} \quad (8.2)$$

8.2 Aplicação dos dados

Para aplicar a técnica de regressão logística sobre a base de dados, definimos a variável resposta como a variável Risco e as variáveis explicativas como o conjunto de 21 indicadores fi-

nanceiros. Os indicadores financeiros, conforme já apresentado, podem ser divididos em quatro grupos: os Indicadores de Estrutura de Capitais, Indicadores de Relacionamento, Indicadores de Liquidez e os Indicadores de Resultado. A equação do modelo logístico pode ser escrita como:

$$g(x) = \beta_0 + \beta_1.PFNA + \beta_2.PFNEND + \beta_3.CDEXD + \beta_4.PFNATCR + \beta_5.GDATV + \beta_6.LQC + \beta_7.LQS + \beta_8.ODAI + \beta_9.ROI + \beta_{10}.RTBV + \beta_{11}.EVRV + \beta_{12}.PMRE + \beta_{13}.PMRV + \beta_{14}.CICLFN + \beta_{15}.PMPC + \beta_{16}.LQG \quad (8.3)$$

, onde $g(x)$ é a variável Risco.

A variável risco pode assumir 13 valores, que correspondem às classe de risco definidas pela metodologia Serasa, conforme exhibe a Tabela 1, na página 24. Para relizar a Regressão Logística, os valores que a variável categórica Risco assume foram reclassificados em duas classes: risco baixo e risco alto. A classe “risco baixo”, representa as empresas cuja classe de risco varia de 6 até 12. Para essas classes foi atribuído o valor “0”. As classes de risco de 12 até 18 foram classificadas como “risco elevado”. Para essas classes foi atribuído o valor “1”. Risco baixo recebe o valor 1 e risco elevado recebe o valor 0. A regressão foi aplicada através do software R e as métricas para avaliar os resultados são apresentadas abaixo. O primeiro passo para avaliar os resultados do modelo foi a análise dos seus coeficientes.

8.3 Coeficientes da regressão

Os coeficientes da regressão indicam a influência de cada variável preditora sobre a variável resposta. Quanto mais positivo esse valor, maior a influência da variável preditora sobre a variável. Quanto mais negativo, maior a influência negativa desses coeficientes sobre a variável resposta. O erro padrão mede a qualidade da estimativa dos coeficientes da regressão. Quanto menor o erro, maior a qualidade das estimativas. O teste de Wald é uma métrica que mede a significância dos coeficientes estimados para a regressão. O teste é dado pela seguinte fórmula:

$$W_i = \frac{\hat{\beta}_i}{\sigma \hat{\beta}_i} \quad (8.4)$$

$\hat{\beta}_i$ representa o coeficiente estimado e $\sigma \hat{\beta}_i$ representa o erro padrão na estimativa desse coeficiente. Para avaliar a significância dos coeficientes o valor W_i , representado na coluna Wald Z, da Tabela 12 é comparado à coluna $\Pr(>|Z|)$. Essa coluna mostra os p-valores calculados para cada um dos indicadores e permite apontar quais indicadores tem significância estatística à um dado valor α . Para um $\alpha = 0.05$, por exemplo, os únicos indicadores estatisticamente significantes são: a Obsolescência do Ativo Imobilizado(ODAI) e o Retorno sobre o investimento(ROI). O valor-p de 0.0015 da obsolescência é inferior ao α de 0.05, o que oferece evidência estatística

suficiente para rejeitar a hipótese nula de que o coeficiente é igual a zero. O mesmo ocorre com o valor-p da retorno sobre o investimento, inferior à 0.0001.

Tabela 12 – Resultados da Regressão Logística

	Coeficientes	Erro padrão	Wald Z	Pr(> Z)
Intercepto	0.2096	0.1205	1.74	0.0819
CDEXD	0.2597	0.2293	1.13	0.2574
ODAI	0.3204	0.1011	3.17	0.0015
PFNA	-0.1146	0.2418	-0.47	0.6355
PFNEND	0.0009	0.2252	0.00	0.9968
PFNATCR	-0.3133	0.1987	-1.58	0.1149
LQS	0.1150	0.0994	1.16	0.2473
PMRV	0.0104	0.3023	0.03	0.9727
LQG	0.1159	0.1570	0.74	0.4601
LQC	0.2205	0.1597	1.38	0.1672
GDTV	-0.1219	0.0672	-1.81	0.0697
RTBV	-0.0085	0.0327	-0.26	0.7940
PMRE	-0.0916	0.3538	-0.26	0.7957
PMPC	-0.0802	0.3817	-0.21	0.8337
CICLFN	0.0621	0.5179	0.12	0.9045
ROI	0.8730	0.1279	6.83	<0.0001
EVRV	0.0075	0.0209	0.36	0.7206

8.4 Receiver Operating Characteristic

A curva ROC (*Receiver Operating Characteristic*) é uma ferramenta gráfica para avaliar o poder discriminante de um modelo, segundo Tasche (2006). A técnica baseia-se nos conceitos de sensibilidade e especificidade. A sensibilidade é a proporção de acertos na previsão de um evento que realmente ocorreu, definida como a proporção de verdadeiros-positivos (TP). Em outras palavras, a sensibilidade calcula o número de empresas que tiveram sua classe de risco corretamente calculadas. A especificidade por sua vez, representa a proporção de acertos na previsão de não ocorrência de um evento que realmente não ocorreu. Neste caso, a especificidade representa a proporção de acertos do modelo ao descartar determinadas classes de risco para uma empresa às quais ela realmente não se encaixa (BRITO; ASSAF NETO, 2008).

Existem ainda outros dois conceitos importantes ao analisar os resultados de um modelo dicotômico: o conceito de Falso-negativo (FN) e Verdadeiro-negativo (TN). O conceito de Falso-negativo representa a proporção de resultados classificados como negativos, quando os mesmos eram positivos. Neste caso, os falso-positivos representam as classes de risco que o modelo descarta para uma determinada empresa, mas que correspondiam à classe em que ela se encaixava. Já os verdadeiro-negativos representam o número de resultados classificados como negativos e que realmente eram negativos. Em outras palavras, corresponde ao número de classes de risco descartadas pelo modelo, para uma determinada empresa, que realmente deveriam ser descartadas. A Tabela 13 abaixo resume os conceitos apresentados de forma esquemática.

Tabela 13 – Definições para classificação de modelos que produzem resultados dicotômicos

Resultado	Real	
Classificação do modelo	Positivo	Negativo
Positivo	Verdadeiro-positivo (TP)	Falso-positivo (FP)
Negativo	Falso-negativo (FN)	Verdadeiro-negativo (TN)

Fonte: Adaptado de Louzada, Ferreira-Silva e Diniz (2012, p.8073)

De acordo com Louzada, Ferreira-Silva e Diniz (2012), podemos expressar ainda esses conceitos através das seguintes relações:

$$SEN = \frac{TP}{TP + FN} \quad (8.5)$$

, onde SEN representa a sensibilidade.

$$SPE = \frac{TN}{TN + FP} \quad (8.6)$$

, onde SPE representa a especificidade.

Após introduzir os conceitos que englobam a curva ROC, representamos a curva resultante dessa regressão através da Figura 19 abaixo.

Figura 19 – Curva ROC

Fonte: Elaboração própria

A Curva ROC representa as medidas de especificidade e sensibilidade para diversos pontos de corte. Na Figura 19, o eixo x, False positive rate, é 1-especificidade e o eixo y é a sensibilidade Brito e Assaf Neto (2008). A área sob a curva Roc, definida como AUC(area under curve), também considerada como medida do poder de discriminação do modelo, foi de 0.82122766, o que de acordo com Hosmer e Lemeshow (2000), constitui um nível de discriminação excelente.

Tabela 14 – Matriz de confusão

Classificação do modelo	Real	
	0	1
0	386	132
1	84	240

Fonte: Elaboração própria

A matriz de confusão representada na Tabela 14 acima, mostra o percentual de empresas classificadas de maneira correta e as classificadas incorretamente. Assim, de um total de 470 empresas com risco baixo(0), 82% foram classificadas de forma correta. Enquanto que, do total de 372 empresas com risco considerado alto(1), o percentual de acertos foi de 64%.

8.5 Teste de Hosmer e Lemeshow

Outra métrica muito utilizada para avaliar a aderência de modelos de regressão logística é o teste de aderência de Hosmer e Lemeshow (2000). O teste avalia a diferença entre os valores observados e os previstos pelo modelo, de maneira que, quanto menor essa diferença, maior a acurácia do modelo. O teste foi empregado para analisar a aderência do modelo de regressão logística descrito pela Equação 8.3 e a análise dos resultados é apresentada na Tabela 15 abaixo.

Tabela 15 – Teste de Hosmer e Lemeshow

	Valores
χ^2	13.913
DF (graus de liberdade)	8
p- valor	0.08406

Fonte: Elaboração própria

Ao nível de significância de 5%, o p-valor desse teste indica que não há evidência estatística suficiente para rejeitar a hipótese de que não há diferenças significativas entre os valores previstos e observados, o que nos garante que o modelo se ajusta aos dados. O próximo capítulo dedica-se à validação da técnica de Redes Bayesianas.

Capítulo 9

Validação do modelo

Para validar o modelo bayesiano desenvolvido adotamos a técnica de validação *Cross-Validation* (validação cruzada). A técnica de *Cross-validation* é um dos métodos mais simples e também um dos mais usados para validação de modelos. A validação cruzada tem sido aplicada nas mais diversas classes de modelos, desde modelos de regressão até modelos de classificação (NAGARAJAN; SCUTARI; LÈBRE, 2013).

A validação cruzada é uma alternativa para medir a qualidade de um modelo de previsão, já que, dificilmente é possível separar duas amostras, nas quais, uma seria selecionada para validar o modelo e a outra para treiná-lo, segundo Hastie, Tibshirani e Friedman (2009).

Dentre os métodos existentes para validação do modelo quatro métodos se destacam, segundo Yadav e Shukla (2016):

- Validação por resubstituição
- Validação Hold-out
- K-fold Cross-validation
- Leave One Out Cross-validation
- K-fold Cross-validation repetida

Neste trabalho, selecionamos a técnica de Cross-validation pelo método K-fold. Aplicamos a técnica sobre o conjunto de dados relativos ao ano de 2014, formado por 21 indicadores de 852 empresas. O valor de K selecionado para aplicação da técnica foi 10. O processo de validação pela técnica K-fold Cross-validation segue as seguintes etapas:

- (a) Primeiro seleciona-se um valor K, que irá dividir a base de dados em K partes iguais;
- (b) o treino do modelo é feito com K-1 partes e uma é deixada de fora para validação;
- (c) o processo é repetido K vezes, até que todas as K partes sejam usadas uma vez como amostra de validação.

A Figura 20 abaixo exemplifica esse processo.

Figura 20 – Cross-validation pelo método k-fold

Validation	Train	Train	Train	Train	Train	Train	Train	Train	Train	Train
------------	-------	-------	-------	-------	-------	-------	-------	-------	-------	-------

Fonte: Elaboração própria. Adaptado de Hastie, Tibshirani e Friedman (2009, p.242)

A amostra de 852 empresas é dividida em 10 partes. Cada parte é formada por aproximadamente 85 elementos. O processo é realizado em 10 iterações. Para cada iteração uma amostra ou subconjunto é definida como amostra de validação enquanto as K-1 são usadas para treinar o modelo. A Figura 20 representa a primeira iteração do processo. O processo de validação foi realizado através do software RStudio.

9.1 Resultados

Para o processo de validação os dados foram separados da seguinte forma: as empresas de 2014 foram usadas para treinar o modelo e as empresas de 2015 para validar a técnica. As métricas empregadas para validar o modelo foram : RMSE, o R^2 e o MAE.

O processo de validação cruzada pelo método k-fold, com k=10, resultou nos seguintes valores:

Tabela 16 – Métricas para o processo de validação cruzada

Métricas empregadas	RMSE	MAE
Valores	2.324	1.850765

O RMSE (Root Mean Squared Error) é uma métrica que avalia a diferença entre os valores previstos pelo modelo e os valores reais. Esse erro é elevado ao quadrado, então tira-se a média dos valores calculados e finalmente, é calculada a raiz quadrada desse valor. O valor de RMSE obtido para o modelo bayesiano indica que o modelo erra um grande percentual das suas previsões.

A medida MAE(Mean Absolut Error) mostra o erro absoluto médio no processo de previsão. Calcula-se a diferença entre o valor previsto e o valor real para cada uma das observações e então divide-se a soma desses valores pela quantidade de observações. O valor de 1.85 reforça o que já foi constatado pela medida RMSE.

Outra técnica empregada para analisar as classificações realizadas pelo modelo é desenvolver uma matriz de confusão. Nessa matriz, as linhas representam a classe de risco das variáveis e as colunas, a classe de risco prevista pelo modelo para as mesmas variáveis. A Tabela 17 representa a matriz de confusão do modelo bayesiano.

A primeira linha, na qual Real = 6, por exemplo, representa as empresas classificadas com classe de risco 6. As colunas dessa mesma linha, que variam de 6 a 18, representam as

Tabela 17 – Matriz de confusão para a técnica de Redes Bayesianas

	Previsto												
Real	6	7	8	9	10	11	12	13	14	15	16	17	18
6	0	0	4	5	6	2	0	0	2	1	1	0	0
7	0	2	3	1	6	7	3	5	0	1	1	0	0
8	0	1	4	3	10	10	6	10	2	6	1	0	0
9	3	1	5	8	12	9	10	6	8	7	3	1	0
10	2	3	5	8	22	14	15	19	7	10	8	0	0
11	2	0	6	8	15	16	20	10	10	9	2	0	0
12	4	0	2	10	16	10	12	19	7	9	4	0	0
13	1	2	6	9	11	13	12	9	9	12	2	0	1
14	2	0	6	3	12	10	5	11	7	11	7	1	0
15	1	1	7	7	10	10	7	16	9	18	7	2	0
16	1	0	4	3	14	6	7	12	7	6	4	0	0
17	0	0	0	2	4	3	2	2	5	5	2	0	0
18	0	0	0	1	3	1	4	5	3	6	1	1	0

Fonte: Elaboração própria. Matriz obtida através do software R

classes de risco em que essas empresas foram classificadas pelo modelo bayesiano. Os campos da matriz nos quais o número da linha é igual ao da coluna, representam o número de empresas classificadas corretamente pelo modelo. A soma do número de empresas de cada uma das linhas representa o número de empresas que pertence a cada classe de risco. Para o cálculo da acurácia do modelo foram observados apenas os valores da diagonal principal da matriz de confusão. A fórmula da acurácia é dada por:

$$ACC = \frac{TP + TN}{TP + FP + TN + FN} \quad (9.1)$$

, onde ACC representa a acurácia do modelo. O cálculo dessa métrica resultou em um valor de acurácia para o modelo de 12%.

9.2 Implementação da cobertura de Markov (Markov Blanket)

Para obter uma Rede Bayesiana com menor número de nós e que contenha apenas os nós mais determinantes para o cálculo do risco adotamos o conceito de Markov Blanket. A ideia desse conceito é que para uma dada variável de interesse A, pertencente a um $DAG(G)$, existe um conjunto mínimo de variáveis condicionadas as outras variáveis que formam o DAG, que são independentes de A. O conjunto dessas variáveis, que pode ser definido com $MB(A)$, onde MB representa a abreviatura de Markov Blanket, são suficientes para calcular a distribuição de probabilidades de A, de maneira que as outras variáveis do DAG podem ser deixadas de lado, Tsamardinos, Aliferis e Statnikov (2003). A definição desse conceito é detalhada conforme Margaritis e Thrun (2000) na definição 7 abaixo.

Definição 7. Dado V um conjunto de variáveis, suponha X e Y duas variáveis condicionalmente dependentes dado um conjunto S . O Markov Blanket de X , $MB(X)$, com $X \in V$, é o conjunto de variáveis tal que para qualquer Y pertencente ao conjunto formado por $V - MB(X) - X$, a variável X é independente de Y .

Neste caso, X representa a variável RISCO e Y é o conjunto de todas as variáveis que a influenciam na rede bayesiana. Aplicando essa definição através do software R a Figura adquiriu o seguinte formato, descrito na Figura 21 abaixo.

Figura 21 – Rede após a implementação da cobertura de Markov

Fonte: Elaboração própria

Note que a nova estrutura é composta por cinco nós, sendo quatro variáveis preditoras e a variável resposta Risco. A implementação dessa nova estrutura trouxe ganhos significativos para o modelo bayesiano, conforme pode-se interpretar pela Tabela 18 a seguir.

A acurácia da técnica bayesiana saltou de 12% para um total de 30,7%. Embora o valor continue abaixo do esperado, houve um ganho significativo na qualidade do modelo. O próximo capítulo descreve o aplicativo que calcula o risco de crédito de empresas, através da técnica de Redes Bayesianas implementada nesse trabalho e exemplifica o seu funcionamento através de uma implementação detalhada.

Tabela 18 – Matriz de confusão após implementação da cobertura de Markov

Real	Previsto												
	6	7	8	9	10	11	12	13	14	15	16	17	18
6	5	3	6	3	2	1	1	0	0	0	0	0	0
7	4	9	7	3	2	3	1	0	0	0	0	0	0
8	4	5	14	12	5	9	4	0	0	0	0	0	0
9	7	3	9	19	3	14	5	7	1	3	2	0	0
10	10	3	10	18	20	22	10	10	2	6	1	1	0
11	5	2	4	3	9	43	7	9	5	9	2	0	0
12	7	1	2	7	6	14	38	3	4	6	4	0	1
13	3	5	3	4	6	9	6	26	9	8	9	0	0
14	2	1	2	5	6	6	6	5	20	14	6	1	1
15	3	2	2	5	10	9	4	7	5	38	9	0	1
16	4	1	2	0	6	7	2	5	6	9	21	0	1
17	1	1	0	0	2	2	1	1	3	7	4	3	0
18	0	0	1	0	1	4	2	2	0	3	6	0	6

Fonte: Elaboração própria. *Matriz obtida através do software R*

Capítulo 10

O Pacote Shiny

O Shiny é um pacote do software R que possibilita desenvolver aplicações na web a partir de códigos construídos no R. O desenvolvimento de interfaces dinâmicas, capazes de absorver os inputs fornecidos pelos usuários e gerar outputs a partir deles, além das inúmeras opções de customização que o pacote oferece, permitem aos desenvolvedores construir aplicações dinâmicas e robustas, RStudio (2017).

10.1 Dados

Os dados que alimentam esse aplicativo são um conjunto de indicadores de risco e indicadores macroeconômicos. Os indicadores de risco correspondem ao conjunto de 17 indicadores financeiros obtidos após a segunda etapa da implementação do algoritmo Pc, na seção 7.3.2, cuja rede bayesiana resultante é representada pela Figura 18. Dessa forma, podemos dividir esses indicadores em quatro grupos:

- **Indicadores de Estrutura de Capitais:** Composição das exigibilidades e Obsolescência do Ativo Imobilizado.
- **Indicadores de Relacionamento com Instituições de Crédito:** Participação de financiamentos no ativo, Participação de financiamentos no endividamento e Participação de financiamentos no ativo circulante.
- **Indicadores de Liquidez:** Liquidez Geral, Liquidez Corrente e Liquidez Seca, Prazo médio de recebimento de vendas, Prazo médio de renovação de estoque, Prazo médio de pagamento de compras e Ciclo Financeiro.
- **Indicadores de Resultado:** Giro do Ativo, Rentabilidade do Ativo médio, Rentabilidade de vendas, Evolução real de vendas.

Esses indicadores são representados por siglas que seguem o seguinte esquema descrito na Tabela 19.

Tabela 19 – Siglas e significados

Código	Descrição
CDEXD	Composição das exigibilidades
ODAI	Obsolescência do ativo imobilizado
PFNA	Participação de financiamentos no ativo
PFNEND	Participação de financiamentos no Endividamento
PFNATCR	Participação de financiamentos no ativo circulante
LQS	Liquidez Seca
LQC	Liquidez Corrente
LQG	Liquidez geral
PMRV	Prazo médio de recebimento de vendas
PMRE	Prazo médio de renovação de estoques
PMPC	Prazo médio de pagamento de compras
CICLFN	Ciclo Financeiro
GDTV	Giro do Ativo
ROI	Rentabilidade do Ativo médio
RTBV	Rentabilidade de Vendas
EVRV	Evolução real de vendas

Fonte: Elaboração própria.

10.2 Tratamento dos dados

Os quatro grupos de indicadores foram padronizados pela mediana e categorizados em 4 classes. A variável Risco manteve sua classificação original, com valores variando entre 1 e 19, já que as empresas cujo risco fosse entre 20 e 22 foram retiradas dessa pesquisa. As classes são organizadas conforme a Tabela 20.

Tabela 20 – Classes de Valores dos Indicadores

Classes	Descrição das classes	Faixas
A	Valores maiores que 0.5	$0.5 < X$
B	Valores entre 0 e 0.5	$0 < X < 0.5$
C	Valores entre -0.5 e 0	$-0.5 < X < 0$
D	Valores inferiores à 0.5	$X < -0.5$

Fonte: Elaboração própria

10.3 Descrição do aplicativo

O aplicativo é dividido em 5 guias. Cada guia pode ser vista como uma página que concentra diferentes informações para o cálculo do risco final, principal objetivo do aplicativo. As duas primeiras guias concentram os indicadores de Estrutura de Capitais, indicadores de Relacionamento com instituições de crédito e os Indicadores de Liquidez, conforme pode ser observado na Figura 22.

Para cada indicador existem 4 campos em branco. Esses campos recebem quaisquer valores digitados pelo usuário. Após digitar esses valores e pressionar o botão “SUBMIT” os valores digitados são padronizados e categorizados. O resultado desse processo é exibido para o usuário na tabela ao lado da seleção dos indicadores.

Essa tabela é composta por quatro colunas, conforme podemos ver na Figura 22. A primeira coluna, denominada **Índices** representa os indicadores, a segunda coluna, chamada **Valores Brutos** representa os valores digitados pelo usuário. A coluna **Valores Padronizados** representa os valores digitados pelo usuário após a sua padronização pela mediana. A última coluna, **Categorias**, representa os valores padronizados categorizados segundo as faixas descritas na Tabela 20.

A guia Indicadores de Liquidez é composta por seis indicadores: Liquidez Geral, Liquidez Corrente, Liquidez Seca, Prazo médio de recebimento de vendas, Prazo médio de pagamento de compras e também pelo Ciclo Financeiro. O princípio é o mesmo usado para os indicadores de Estrutura de Capitais: após digitar os valores nos campos especificados, os valores são padronizados e categorizados. Os resultados são mostrados na tabela ao lado da caixa de seleção dos indicadores, conforme a Figura 23.

A guia seguinte, Risco Calculado, apresenta o cálculo do risco. Esse cálculo é feito com base nos valores fornecidos para os indicadores de Estrutura de Capitais, Relacionamento com instituições de crédito, Indicadores de Liquidez e os indicadores de resultado que alimentam a Rede Bayesiana.

O resultado é exibido numa “régua” formada por 13 classes de risco, que variam de 6 à 18 e representam as classes de risco que uma empresa pode assumir. A classe de risco da empresa é aquela que tiver a maior probabilidade associada, conforme exibe a Figura 24.

A Figura 24 é dividida em três partes: na parte superior, abaixo do título de cada uma das 5 guias do aplicativo há o conjunto de classes de risco, que cada uma das empresas pode assumir, que variam no intervalo 6,18. Abaixo há duas figuras, a do lado esquerdo, representa a rede bayesiana do risco e a do lado direito, representa um histograma. Esse histograma auxilia a visualizar as classes de risco com maior probabilidade associadas a cada uma das observações inseridas no aplicativo. Assim, conforme os indicadores associados a cada observação são absorvidos o aplicativo indica as probabilidades de que essa empresa esteja em cada uma das 13 classes de risco, o que é exibido na primeira linha após o título das guias do aplicativo e ao mesmo tempo, o aplicativo evidencia através do histograma as classes para as quais foram atribuídas as maiores probabilidades.

10.4 Simulação com o aplicativo Shiny

Com o intuito de mostrar como o aplicativo funciona foi selecionada uma observação do banco de dados como exemplo. Essa observação corresponde a uma empresa da indústria de soja e o risco atribuído a ela foi de 14, o que corresponde a uma média de 6,5% de inadimplência. Os valores brutos dos 16 indicadores dessa empresa são descritos na Tabela 21.

Tabela 21 – Valores dos indicadores

Indicadores	Valores
CDEXD	77,13
ODAI	6,17
PFNA	64,77
PFNEND	53,36
PFNATCR	96,05
LQG	0,8154
LQC	1,04
LQS	0,45
PMRV	25
PMRE	26
PMPC	39
CICLFN	12
GDATV	1,69
ROI	0,00234
RTPL	0,02493
RTBV	0,2417
EVRV	-16

Fonte: Elaboração própria.

Esses valores são digitados em cada uma das guias do aplicativo a fim de obter a classe de risco mais apropriada às características da empresa. Dessa forma, para o primeiro conjunto de indicadores o aplicativo adquire a seguinte configuração, descrita na Figura 25. O processo é repetido e as probabilidades associadas a cada classe de risco são mostradas na aba RISCO CALCULADO, conforme a Figura 26.

Assim, dado o conjunto de indicadores calculados dessa empresa o resultado obtido pela Rede Bayesiana nos diz que a classe de risco mais provável que essa empresa pertença é a classe 15, que obteve 25.1% de probabilidade. Comparando a classe de risco obtida pela Rede Bayesiana com a classe de risco atribuída pela Serasa, a diferença entre os valores é bem pequena. O risco atribuído pela Rede Bayesiana foi 15, enquanto o valor atribuído pela Serasa

foi 14. Ainda segundo as probabilidades da Rede Bayesiana a classe 13 obteve 20.99% de probabilidade e a classe 14 obteve 12.76%.

Figura 22 – Print de Tela da guia Estrutura de Capitais

Rede Bayesiana

INDICADORES DE ESTRUTURA E RELACIONAMENTO

INDICADORES DE LIQUIDEZ

INDICADORES DE RESULTADO

RISCO CALCULADO

Selecione os valores de cada indicador

Composição das exigibilidades

Obsolescência do ativo imobilizado

Participação de financiamentos no ativo

Participação de financiamentos no Endividamento

Participação de financiamentos no Ativo Circulante

SUBMIT

Indices	Valores_Brutos	Valores_Padronizados	Categorias
CDEXD	4	-1.42	D
ODAI	4	-0.23	C
PFNA	4	-0.37	C
PFNEND	4	-0.36	C
PFNATCR	4	-0.49	C

Fonte: Elaboração própria. Figura obtida através do Shiny App.

Figura 23 – Print de Tela da guia Indicadores de Liquidez

Fonte: Elaboração própria. Figura obtida através do Shiny App.

Figura 25 – Print de Tela com valores atribuídos

Fonte: Elaboração própria. Figura obtida através do Shiny App.

Capítulo 11

Considerações Finais

O trabalho propôs o desenvolvimento de um modelo de crédito baseado na técnica probabilística de Redes Bayesianas, que atribuisse a pontuação de crédito de um cliente institucional com base em conjunto de 17 indicadores financeiros calculados a partir das demonstrações financeiras dessas mesmas empresas.

A Rede Bayesiana obtida porém, mostrou-se pouco capaz de prever com exatidão a classe de risco de cada um dessas empresas. Dentre as causas que justificam esses erros uma merece destaque: o número de classes da variável resposta do modelo. A técnica de Redes Bayesianas calcula a probabilidade associada a uma variável condicionada a um conjunto de outras evidências, advindas das variáveis explicativas do modelo, de que se tem alguma evidência. Esse cálculo é regido pela fórmula de probabilidades condicionais, de maneira que, quanto maior o número de evidências maior o cálculo de probabilidades e consequentemente, maior a complexidade e tempo de processamento do modelo.

Ao assumir um modelo composto por 17 indicadores financeiros, a probabilidade de que uma empresa pertença a cada uma das treze classes de risco possíveis, exige o cálculo de muitas probabilidades condicionais, aumentando a complexidade do modelo e tornando-o mais suscetível a erros. A comparação desse modelo, baseado na técnica de Redes Bayesianas, com o mesmo modelo, porém, desenvolvido com a técnica de Regressão Logística, apenas confirma que o número de classes da variável resposta, pode ser um dos principais responsáveis pelo baixo poder preditivo da técnica.

Ao aplicar a técnica de Regressão Logística, as classes de risco do modelo foram reduzidas de treze para duas, uma vez que utilizamos o modelo de regressão logística binomial. Nessa nova categorização da variável resposta porém, houve um aumento significativo da proporção de acertos do modelo, que saltou de 12% para 66%. Assim, pode-se concluir que para esse conjunto de variáveis, e para os procedimentos de padronização e categorização adotados nesse trabalho, a técnica probabilística de Redes Bayesianas pode não ser a mais adequada. Abdou e Pointon (2011) sustentam essa constatação e enfatizam que embora não seja possível estabelecer um modelo de pontuação de crédito ideal, podemos comparar e verificar quais são as técnicas mais eficientes para obter a pontuação que reflita a capacidade financeira de uma

empresa, o que foi realizado neste trabalho.

Assim, além de promover a comparação entre duas técnicas diferentes, uma probabilística (Redes Bayesianas) e outra de regressão (Regressão Logística), como técnicas para classificação de risco de crédito, o trabalho também contribuiu para o estado da arte dos modelos de classificação, na medida em que ele traz um produto : o aplicativo de crédito. Esse aplicativo operacionaliza todo o processo de classificação de crédito e possibilita que profissionais da área utilizem-o como ferramenta de apoio nas decisões relativas a concessão de crédito.

Referências

- ABDOU, Hussein A.; POINTON, John. Credit scoring, statistical techniques and evaluation criteria: a review of the literature. *Intelligent Systems in Accounting, Finance and Management*, v. 18, p. 59–88, 2011. Acesso em: 14 mar. 2018. Disponível em: <<https://doi.org/10.1002/isaf.325>>.
- ABRAMOWICZ, Witold; MAREK, Nowak; SZTYKIEL, Joanna. Bayesian networks as a decision support tool in credit scoring domain. In: *Managing data mining technologies in organizations: Techniques and Applications*. Hershey: IGI Global, 2003. p. 302. Acesso em: 2 out. 2018. Disponível em: <<https://www.igi-global.com/gateway/chapter/25757>>.
- ANDERSON, Raymond. *The Credit Scoring Toolkit: theory and practice for retail credit risk management and decision automation*. New York: Oxford University Press Inc, 2007. 731 p. ISBN 978-0199226405.
- ANDRADE, F. W. M. Modelos de risco de crédito. *Tecnologia de Crédito*, p. 23–53, 2003.
- ASHCROFT, MICHAEL. Bayesian networks in business analytics. In: *Proceedings of the Federated Conference on Computer Science and Information Systems FedCSIS 2012*. [S.l.: s.n.], 2012.
- BAESENS, Bart; RÖSCH, Daniel; SCHEULE, Harald. *Credit Risk Analytics: Measurement Techniques, Applications, and Examples in SAS*. New Jersey: John Wiley & Sons, Inc., 2016. ISBN 978-1-119-14398-7.
- BAILEY, Murray. *Consumer Credit Quality: Underwriting, Scoring, Fraud Prevention and Collections*. 1. ed. Bristol: White Box Publishing, 2004. 234 p. ISBN 9780954005337.
- BLUHM, Christian; OVERBECK, Ludger; WAGNER, Christoph. *An Introduction to Credit Risk Modeling*. [S.l.]: Chapman & Hall/CRC, 2002. ISBN 1-58488-326-X.
- BORIS, Kollar; IVANAA, Weissova; ANNAA, Siekelova. Quantification of credit risk with the use of creditmetrics. *Procedia Economics and Finance* 26, n. 26, p. 311–316, 2015. Acesso em: 11 jul. 2019. Disponível em: <<https://www.sciencedirect.com/science/article/pii/S2212567115008588>>.
- BRITO, Giovanni A. S.; ASSAF NETO, Alexandre. Modelo de classificação de risco de crédito de empresas. *Revista Contabilidade & Finanças*, v. 19, n. 46, p. 18–29, 2008. Acesso em: 9 mar. 2018. Disponível em: <<https://doi.org/10.1590/s1519-70772008000100003>>.

BRITO, Giovani A. S.; ASSAF NETO, Alexandre N.; CORRAR, Luiz J. Sistema de classificação de risco de crédito: uma aplicação a companhias abertas no brasil. *Revista Contabilidade e Finanças*, v. 20, n. 51, p. 28–43, 2009. Acesso em: 3 abr. 2018. Disponível em: <http://dx.doi.org/10.1590/S1519-70772009000300003>.

CAOUELETTE, John B.; ALTMAN, Edward I.; NARAYANAN, Paul. *Managing Credit Risk: the next great financial challenge*. New York: John Wiley & Sons, Inc., 1998. 452 p. ISBN 978-0471111894.

CROOK, Jonathan N.; EDELMAN, David B.; THOMAS, Lyn C. Recent developments in consumer credit risk assessment. *European Journal of Operational Research*, v. 183, p. 1447–1465, 2007. Acesso em: 17 set. 2018. Disponível em: <https://www.sciencedirect.com/science/article/pii/S0377221706011866>.

CROUHY, Michel; GALAI, Dan; MARK, Robert. A comparative analysis of current credit risk models. *Journal of Banking & Finance*, v. 24, p. 59–117, 2000. Acesso em: 11 mar. 2018. Disponível em: <https://www.ime.usp.br/~rvicente/risco/crouhy.pdf>.

EXPERIAN, Serasa. *Credit Rating Serasa Experian: manual do produto*. São Paulo, 2018.

FAYYAD, Wael H. Credit risk modeling challenges. *SSRN Eletronic Journal*, p. 1–13, 2008. Acesso em: 20 nov. 2018. Disponível em: <http://dx.doi.org/10.2139/ssrn.1240022>.

FRIEDMAN, Nir; GEIGER, Dan; GOLDSZMIDT, Moises. Bayesian networks classifiers. *Machine Learning*, v. 29, n. 2, p. 131–163, 1997. Acesso em: 20 maio 2019. Disponível em: <https://link.springer.com/article/10.1023/A:1007465528199>.

GARCÍA, Fernando; GIMÉNEZ, Vicente; GUIJARRO, Francisco. Credit risk management: A multicriteria approach to assess creditworthiness. *Mathematical and Computer Modelling*, p. 2009–2015, 2013. Acesso em: 5 jul. 2019. Disponível em: <https://www.sciencedirect.com/science/article/pii/S0895717712000623?via%3Dihub>.

GORDY, Michael B. A comparative anatomy of credit risk models. *Journal of Banking & Finance*, v. 24, n. 1-2, p. 119–149, 2000. Acesso em: 11 jul. 2019. Disponível em: <https://www.sciencedirect.com/science/article/pii/S0378426699000540>.

GUIMARÃES, Paulo R. F.; RESENDE, Moisés F. Uma aplicação do modelo de regressão logística na previsão de falência empresarial no brasil. *Revista brasileira de economia de empresas*, v. 18, n. 2, p. 21–42, 2018. Acesso em: 5 jan. 2019. Disponível em: <https://portalrevistas.ucb.br/index.php/rbee/article/view/7963>.

HAND, D. J.; MCCONWAY, K. J.; STANGHELLINI, E. Graphical models of applicants for credit. *IMA Journal of Mathematics Applied in Business & Industry*, v. 8, p. 143–155, 1997. Acesso em : 6 jun. 2018. Disponível em: <https://academic.oup.com/imaman/article/8/2/143/672927>.

HASTIE, Trevor; TIBSHIRANI, Robert; FRIEDMAN, Jerome. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. 2. ed. California: Springer, 2009. Acesso em: 12 set. 2018. ISBN 0387848576. Disponível em: <https://web.stanford.edu/~hastie/Papers/ESLII.pdf>.

HØJSGAARD, Søren; EDWARDS, David; LAURITZEN, Steffen. *Graphical Models with R*. New York: Springer, 2012. ISBN 978-1-4614-2298-3.

HOSMER, David W.; LEMESHOW, Stanley. *Applied logistic regression*. 2. ed. United States: Wiley & Sons, 2000. Acesso em: 2 nov. 2018. Disponível em: <<https://onlinelibrary.wiley.com/doi/book/10.1002/0471722146>>.

HUANG, Cheng L.; CHEN, Mu C.; WANG, Chieh J. Credit scoring with a data mining approach based on support vector machines. *Expert Systems with Applications*, v. 33, p. 847–856, 2007. Acesso em: 22 out. 2018. Disponível em: <<https://www.sciencedirect.com/science/article/pii/S095741740600217X?via%3Dihub>>.

HUANG, Zan et al. Credit rating analysis with support vector machines and neural networks: a market comparative study. *Decision Support Systems*, v. 37, n. 4, p. 543–558, 2004. Acesso em 27 de novembro de 2018. Disponível em: <<https://www.sciencedirect.com/science/article/pii/S0167923603000861>>.

ISCOE, Ian et al. Portfolio credit-risk optimization. *Journal of Banking & Finance*, v. 36, p. 1604–1615, 2012. Acesso em: 5 Jul. 2019. Disponível em: <<https://www.sciencedirect.com/science/article/pii/S0378426612000283?via%3Dihub>>.

KEALHOFER, Stephen. Quantifying credit risk : Default prediction. *Financial Analysts Journal*, v. 59, n. 1, p. 30–44, 2003. Acesso em: 14 jan. 2018. Disponível em: <<https://www.cfapubs.org/doi/pdf/10.2469/faj.v59.n1.2501>>.

KOLLER, Daphne; FRIEDMAN, Nir. *Probabilistic Graphical Models: principles and techniques*. 1. ed. Massachusetts: The MIT Press, 2009. 1270 p. ISBN 0262013193.

LAPLACE, Pierre S. *Théorie analytique des probabilités*. 2. ed. Paris: Courcier, 1814. Acesso em: 10 nov. 2018. Disponível em: <<https://play.google.com/store/books/details?id=6MRLAAAAMAAJ&rdid=book-6MRLAAAAMAAJ&rdot=1>>.

LEE, Tian S. et al. Credit scoring using the hybrid neural discriminant technique. *Expert Systems with Applications*, v. 23, n. 3, p. 245–254, 2002. Acesso em: 4 ago. 2018. Disponível em: <<https://www.sciencedirect.com/science/article/pii/S0957417402000441?via%3Dihub>>.

LIM, Michael K.; SOHN, So Y. Cluster-based dynamic scoring model. *Expert Systems with Applications*, v. 32, n. 2, p. 427–431, 2007. Acesso em: 12 jun. 2018. Disponível em: <<https://www.sciencedirect.com/science/article/pii/S0957417405003532?via%3Dihub>>.

LOUZADA, Francisco; FERREIRA-SILVA, Paulo H.; DINIZ, Carlos A. R. On the impact of disproportional samples in credit scoring models: An application to a brazilian bank data. *Expert Systems with Applications*, v. 39, p. 8071–8078, 2012. Acesso em: 10 set. 2018. Disponível em: <<https://www.sciencedirect.com/science/article/pii/S0957417412001522>>.

MARGARITIS, Dimitris; THRUN, Sebastian. Bayesian network induction via local neighborhoods. In: PRESS, MIT (Ed.). *Advances in Neural Information Processing Systems*. [S.l.: s.n.], 2000. v. 12, p. 505–511.

MERTON, Robert C. On the pricing of corporate debt: the risk structure of interest rates. *The Journal of Finance*, v. 29, p. 449–470, 1974. Acesso em: 10 jul. 2019. Disponível em: <<https://onlinelibrary.wiley.com/doi/full/10.1111/j.1540-6261.1974.tb03058.x>>.

NAGARAJAN, Radhakrishnan; SCUTARI, Marco; LÈBRE, Sophie. *Bayesian Networks In R With Applications In Systems Biology*. New York: Springer, 2013. ISBN 978-1-4614-6446-4.

NIKOLIC, Nebojsa et al. The application of brute force logistic regression to corporate credit scoring models: evidence from serbian financial statements. *Expert Systems with Applications*, v. 40, n. 15, p. 5932–5944, 2013. Acesso em: 14 jun. 2018. Disponível em: <<https://www.sciencedirect.com/science/article/pii/S0957417413003084>>.

PANDEY, Trilok et al. Credit risk analysis using machine learning classifiers. In: INTERNATIONAL CONFERENCE ON ENERGY, COMMUNICATION, DATA ANALYTICS AND SOFT COMPUTING (ICECDS). Chennai, 2017. p. 1850–1854. Acesso em: 11 abr. 2018. Disponível em: <<https://ieeexplore.ieee.org/document/8389769>>.

PAPALAMPROU, Konstantinos; ANTONIOU, Paschalis. Estimation of capital requirements in downturn conditions via the cbv model: Evidence from the greek banking sector. *Operations Research Perspectives*, v. 6, p. 100102, 2019. Acesso em: 10 jul. 2019. Disponível em: <<http://www.sciencedirect.com/science/article/pii/S2214716017301847>>.

PATIL, Tina R.; SHEREKAR, S. S. Performance analysis of naive bayes and j48 classification algorithm for data classification. *International Journal Of Computer Science And Applications*, v. 6, n. 2, p. 256–261, 2013. Acesso em: 3 mar. 2018. Disponível em: <<http://www.researchpublications.org/IJCSA/NCAICN-13/189.pdf>>.

PAVLENKO, Tatjana; CHERNYAK, Oleksandr. Credit risk modeling using bayesian networks. *International Journal of Intelligent Systems*, v. 25, p. 326–344, 2010. Acesso em: 22 nov. 2018. Disponível em: <<https://doi.org/10.1002/int.20410>>.

QASEM, Mais H.; NEMER, Loai. Extreme learning machine for credit risk analysis. *Journal of Intelligent Systems*, p. 1–13, 2018. Disponível em: <<https://doi.org/10.1515/jisys-2018-0058>>.

ROSS, Sheldon. *Probabilidade: um curso moderno com aplicações*. 8. ed. Porto Alegre: Bookman, 2010. 628 p. ISBN 978-85-7780-621-8.

RSTUDIO. 2017. Acesso em: 20 nov. 2018. Disponível em: <<https://shiny.rstudio.com/>>.

Sá, Joyce M. C. et al. Análise de Crédito Utilizando uma Abordagem de Mineração de Dados. *Revista de Engenharia e Pesquisa Aplicada*, v. 3, p. 146–157, 2018. Acesso em: 18 dez. 2018. Disponível em: <<http://revistas.poli.br/index.php/rep/article/view/967>>.

SAITA, Francesco. *Value at Risk and Bank Capital Management*. [S.l.]: Elsevier. Inc, 2007. 280 p. ISBN 978-0123694669.

SÁNCHEZ, José F.; LECHUGA, Gilberto P. Assessment of a credit scoring system for popularbank savings and credit. *Cuntadoría y Administración*, v. 61, n. 2, p. 391–417, 2016. Acesso em: 4 ago. 2018. Disponível em: <<https://www.sciencedirect.com/science/article/pii/S0186104215001138>>.

SANFINS, Marcos A. S.; CLARK, Thiago M. Creditrisk+: Implementação da modelagem estatística de risco de crédito e cálculos alternativos através da transformada rápida de fourier no r. In: 19º SINAPE. São Pedro, 2010. Disponível em: <http://www2.ime.unicamp.br/sinape/sites/default/files/artigo2_0.pdf>.

SAUNDERS, Anthony; ALLEN, Linda. *Credit Risk Measurement: New Approaches to Value at Risk and Other Paradigms*. 2. ed. [S.l.]: John Wiley & Sons, 2002. ISBN 978-0471219101.

SCHOMMER, Susan. Risco de crédito. Material didático para aulas. 2007. Disponível em: <http://w3.impa.br/~zubelli/RISK/risco_de_cred.pdf>.

SHARMA, Aman K.; SAHNI, Suruchi. A comparative study of classification algorithms for spam email data analysis. *International Journal on Computer Science and Engineering (IJCSE)*, v. 3, n. 5, p. 1890–1895, 2011. Acesso em: 21 maio 2019. Disponível em: <http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.301.7309&rep=rep1&type=pdf>.

SULLIVAN, A. Charlene. Consumer finance. In: *Financial Handbook*, Altman EI. New York: John Wiley and Sons, 1981.

TABAK, John. *Probability and Statistics: The science of Uncertainty*. New York: Facts on File, 2004. ISBN 0816049564.

TASCHE, Dirk. Validation of internal rating systems and PD estimates. *The Analytics of Risk Model Validation*, 2006. Acesso em: 19 dez. 2018. Disponível em: <https://www.sciencedirect.com/science/article/pii/B9780750681582500147>.

THOMAS, Lyn C. A survey of credit and behavioural scoring: forecasting financial risk of lending to consumers. *International Journal of Forecasting*, 2000. Acesso em: 18 set. 2018. Disponível em: <https://www.sciencedirect.com/science/article/pii/S0169207000000340?via%3Dihub>.

THOMAS, Lyn C.; EDELMAN, David B.; CROOK, Jonathan N. *Credit Scoring and its Applications*. 2. ed. Philadelphia: Society for Industrial and Applied Mathematics, 2017. ISBN 9781611974560.

TRIOLA, Mário F. *Estatística*. 10. ed. México: Pearson Education, 2009. ISBN 9780321460929.

TSAMARDINOS, Ioannis; ALIFERIS, Constantin; STATNIKOV, Alexander. Algorithms for large scale markov blanket discovery. In: *Proceedings of the sixteenth international Florida artificial intelligence research society conference*. [S.l.]: Department of Biomedical Informatics, Vanderbilt University, 2003. p. 376–381. Acesso em: 14 ago. 2018.

WU, W. Improving Classification Accuracy and Casual Knowledge for Better Credit decisions. *International Journal of Neural Systems*, v. 21, n. 4, p. 297–309, 2011. Acesso em: 10 jun. 2018. Disponível em: <https://www.worldscientific.com/doi/abs/10.1142/S0129065711002845>.

YADAV, S.; SHUKLA, S. IEEE 6th International Conference on Advanced Computing. In: *Analysis of k-fold cross-validation over hold-out validation on colossal datasets for quality classification*. [s.n.], 2016. Acesso em: 11 nov. 2018. Disponível em: <https://ieeexplore.ieee.org/document/7544814>.

YOON, Y. K. *Modelling Operational Risk In Financial Institutions Using Bayesian Networks*. 2003. 83 p. Dissertação (Master in em Actuarial Management) — Cass Business School, Londres, 2003. Acesso em : 15 fev. 2018.