UNIVERSIDADE FEDERAL DO ESTADO DO RIO DE JANEIRO
CENTRO DE CIÊNCIAS EXATAS E TECNOLOGIA
PROGRAMA DE PÓS-GRADUAÇÃO EM INFORMÁTICA

Avaliação Econômica de uma Oportunidade Exploratória de Petróleo através de Mineração de Dados e com Apoio de uma Ontologia de Domínio

Marcos Antônio Affonso

Orientadores

Leila Cristina Vasconcelos de Andrade

Kate Cerqueira Revoredo

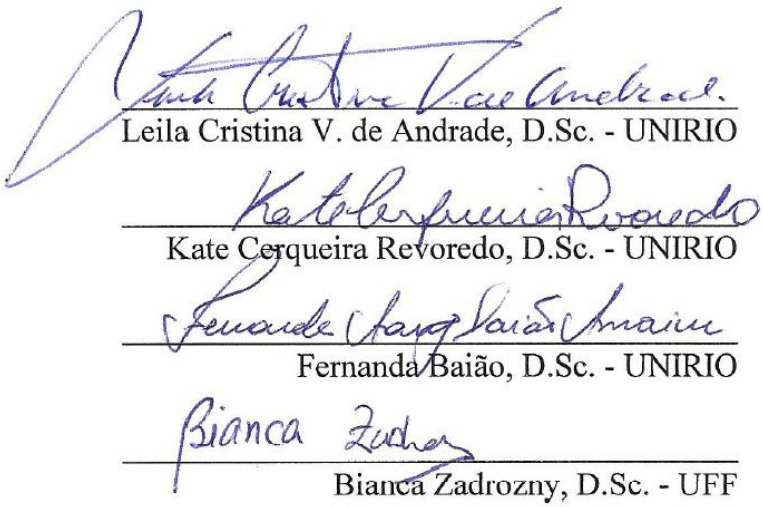
RIO DE JANEIRO, RJ – BRASIL
Agosto de 2012

AVALIAÇÃO ECONÔMICA DE UMA OPORTUNIDADE EXPLORATÓRIA DE PETRÓLEO ATRAVÉS DE MINERAÇÃO DE DADOS E COM APOIO DE UMA ONTOLOGIA DE DOMÍNIO

Marcos Antônio Affonso

DISSERTAÇÃO APRESENTADA COMO REQUISITO PARCIAL PARA OBTENÇÃO DO TÍTULO DE MESTRE PELO PROGRAMA DE PÓS-GRADUAÇÃO EM INFORMÁTICA DA UNIVERSIDADE FEDERAL DO ESTADO DO RIO DE JANEIRO (UNIRIO). APROVADA PELA COMISSÃO EXAMINADORA ABAIXO ASSINADA.

Aprovada por:



Leila Cristina V. de Andrade, D.Sc. - UNIRIO

Kate Cerqueira Revoredo, D.Sc. - UNIRIO

Fernanda Baião, D.Sc. - UNIRIO

Bianca Zadrozny, D.Sc. - UFF

RIO DE JANEIRO, RJ – BRASIL
Agosto de 2012

Affonso, Marcos Antônio.
A257 Avaliação econômica de uma oportunidade exploratória de petróleo através de mineração de dados e com apoio de uma ontologia de domínio / Marcos Antônio Affonso, 2012.
112f. ; 30 cm

Orientador: Leila Cristina Vasconcelos de Andrade.
Coorientador: Kate Cerqueira Revoredo.
Dissertação (Mestrado em Informática) – Universidade Federal do Estado do Rio de Janeiro, Rio de Janeiro, 2012.

1. Tecnologia da informação. 2. Mineração de dados. 3. Processo decisório. 4. Ontologia (Informática). 5. Petróleo - Exploração. I. Andrade, Leila Cristina Vasconcelos. II. Revoredo, Kate Cerqueira. III. Universidade Federal do Estado do Rio de Janeiro. Centro de Ciências Exatas e Tecnologia. Curso de Mestrado em Informática. III. Título.

CDD – 005.5

Agradecimentos

Gostaria de expressar meus agradecimentos a todos que me apoiaram com paciência, recursos e confiança na realização deste trabalho. Ao corpo docente da UNIRIO agradeço pelos conhecimentos repassados a mim, em especial às minhas orientadoras Dr.ªs Leila Andrade e Kate Revoredo pela cooperação e preocupação pelo desenvolvimento deste trabalho.

À Petrobras, especialmente aos colegas do departamento de Exploração e Produção Luciano Arantes, Jalimar, Cleomar, Armando e Regis pelas suas contribuições técnicas e de recursos que ajudaram a viabilizar este trabalho.

Agradeço à compreensão de meus familiares, aos quais privei da minha presença, e, sobretudo, a minha mãe que viu este trabalho iniciar, mas não conseguiu vê-lo concluído. A ela, em memória, dedico este esforço de conhecimento.

Aos meus colegas de mestrado, meu muito obrigado pelo companheirismo que ajudou a mitigar as dificuldades da caminhada.

Affonso, Marcos Antônio**. Avaliação Econômica de uma Oportunidade Exploratória de Petróleo através de Mineração de Dados e com Apoio de uma Ontologia de Domínio.** UNIRIO, 2012. 123 páginas. Dissertação de Mestrado. Departamento de Informática Aplicada, UNIRIO**.**

RESUMO

Uma Oportunidade Exploratória de Petróleo (OE) é definida como uma região geográfica com potencial de possuir uma acumulação de petróleo de valor suficiente para justificar um projeto exploratório. As avaliações econômicas de uma OE são baseadas em técnicas convencionais de Estatística e Economia que não conseguem modelar o comportamento imprevisível presente em algumas variáveis como preço do óleo no mercado e decisões governamentais, ocasionando imprecisões nas avaliações.

Por outro lado, técnicas de mineração de dados são utilizadas para aprender um modelo a partir de um conjunto de dados históricos e têm sido aplicadas na construção de modelos capazes de realizar predições na área de petróleo, como na predição de viscosidade e do valor de ofertas em leilões de concessões exploratórias.

Este trabalho propõe a construção de um modelo preditivo capaz de avaliar economicamente uma OE utilizando técnicas de mineração de dados, especificamente Redes Bayesianas. Uma ontologia de domínio foi construída e utilizada para apoiar a etapa de mineração de dados. Além disto, o conhecimento do domínio, estando representado formalmente em uma rede Bayesiana, abre caminho para realização de outros tipos de inferências além da predição econômica. A nossa proposta foi avaliada através de um estudo de caso que comprovou a sua viabilidade.

Palavras-chave: Mineração de Dados, Sistemas de Suporte à Decisão, Ontologia, Exploração de Petróleo.

ABSTRACT

A petroleum exploration opportunity (EO) is defined as a mapped region with potential for possessing a sufficient petroleum accumulation that may justify an exploration project. Currently, economic evaluations of an EO are based on conventional techniques of Statistics and Economics that are not able to model unpredictable behavior peculiar to some variables as oil price and government decisions, resulting in evaluation inaccuracies.

On the other hand, Data Mining techniques have been applied to build models based on historical data capable of making predictions in the petroleum area such as the viscosity value and the bid value of an exploration concession in an auction.

This work proposes the construction of a predictive model able to economically evaluate an EO using Data Mining techniques, specifically Bayesian Networks. A domain ontology was built and used to support the data Mining step. Moreover, domain knowledge being formally represented by a Bayesian Network, paves the way for achievement of other types of inferences beyond the economic prediction. Our proposal was evaluated through a study case that proved its viability.

Keywords: Data Mining, Decision Making System, Ontology, Petroleum Exploration

Sumário

Lista de Figuras

Lista de Tabelas

Capítulo 1 – Introdução

Este capítulo apresenta a motivação, o problema a ser tratado e a hipótese da pesquisa, bem como a organização deste trabalho.

1.1 Motivação

Apesar de pesquisas na busca por fontes alternativas de energia (LELLIS, 2007) como álcool, biomassa (fonte renovável) e energia eólica, o consumo de derivados de petróleo no Brasil vem aumentando consideravelmente, como demonstrado pelo gráfico de consumo por fonte de energia, publicado pelo Ministério de Minas e Energia (MME) em 2011 e exibido na Figura 1. O eixo da ordenada expressa a quantidade de energia consumida em TEP (tonelada de petróleo equivalente), enquanto a abscissa representa a passagem dos anos a partir de 1970 até 2011.

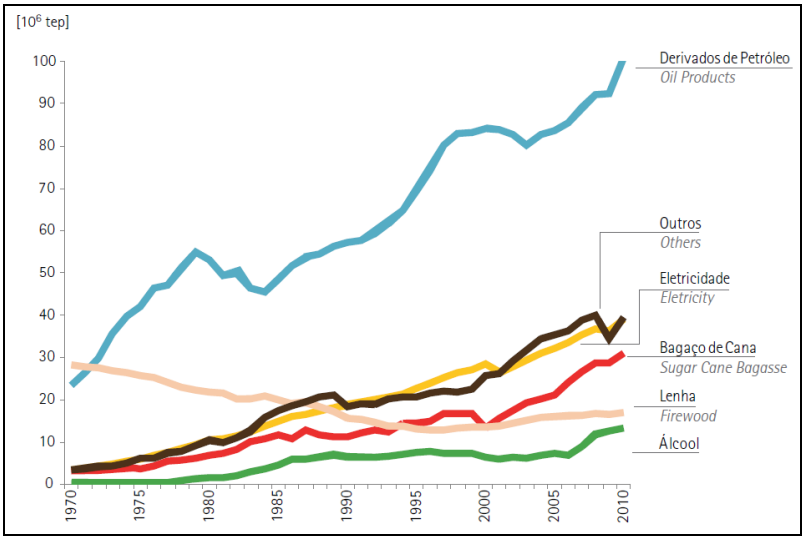


Figura 1: Consumo de energia no Brasil por fonte (Fonte: MME - 2011)

Enquanto estas fontes alternativas não atingirem um estágio de evolução que permita ampla utilização, a indústria petrolífera vem assumindo a responsabilidade de suprir a maior parte da energia necessária ao desenvolvimento. Dessa forma, a crescente demanda por energia exige a descoberta de novas acumulações de petróleo, ou seja, novas jazidas.

Quando determinada área geográfica é submetida a estudos e apresenta indícios de possuir petróleo é denominada Oportunidade Exploratória (OE) e esforços são feitos para se avaliar a possibilidade de realização de um projeto que confirme as expectativas geradas. Os estudos geológicos procuram identificar as condições geológicas necessárias à presença do petróleo e servem de base para os estudos econômicos que avaliarão a viabilidade econômica do projeto.

A modalidade de avaliação econômica de uma OE comumente utilizada é através da análise do Valor Presente Líquido (VPL) (SCHUYLER, 2001). VPL se baseia no fluxo de caixa e na curva de produção e será detalhada na Seção 2.1. A tarefa de avaliação lida com informações incertas como preço do óleo no mercado ou qualidade do óleo e é

realizada por especialistas (engenheiros e geólogos com formação em Economia) que aplicam seus conhecimentos e experiências para realizar estas avaliações.

Dependendo da quantidade e da complexidade da OE, o trabalho de avaliação pode se estender de 2 a 4 dias, comprometendo o cumprimento de prazos para a tomada de decisão. A situação torna-se crítica quando se trata de altos investimentos. Além disso, as avaliações são dependentes dos artefatos fluxo de caixa e curva de produção que, às vezes, não estão disponíveis a tempo. Estes problemas foram relatados pelos especialistas.

Além da possível demora na finalização da avaliação de uma OE, outra questão é a imprecisão dessa avaliação. Corrobora para isto o Índice de Sucesso Exploratório (ISE) que varia conforme a geologia e o critério de cálculo de cada país. Por exemplo, na Austrália o ISE está entre 10 e 40% [1] enquanto que nos EUA o ISE está por volta de 55% [2]. Algumas empresas como Atoka [3] colocam em sua página o valor de seu ISE como forma de mostrar a sua eficiência.

Se considerarmos uma média mundial de 20%, então de cada 100 poços de petróleo perfurados em apenas 20 descobre-se petróleo em quantidade e qualidade suficientes para transformar uma OE em um campo produtor de petróleo. Parte do motivo para este baixo índice é devido a imprecisões na avaliação econômica provocadas pela utilização de métodos convencionais e também pelas incertezas inerentes à atividade exploratória

[1] http://www.dmp.wa.gov.au/documents/Petroleum_Exploration.pdf

[2] http://www.petrostrategies.org/Graphs/drilling_success_rates.htm

[3] http://www.atoka.com

que lida com informações estimadas como a qualidade do óleo a ser encontrado e profundidade em que se encontra.

O valor das avaliações econômicas de uma OE, sendo expresso pelo VPL, é baseado em Estatística convencional e modelos econômicos clássicos que tendem a apresentar baixo desempenho, pois trabalham com pressupostos incapazes de captar padrões imprevisíveis (YU et al., 2007) como preço do óleo, impostos e decisões governamentais.

Por outro lado, técnicas de mineração de dados (WITTEN et al., 2011) tem sido utilizadas para, a partir de uma base de dados, aprender um modelo que descreva o comportamento do domínio. Há na literatura vários exemplos de aplicação de mineração de dados para construção de modelos, como na saúde pública (CANLAS, 2009) e detecção de fraude em cartões de crédito (OGWELEKA, 2011).

Neste trabalho, argumentamos que técnicas de mineração de dados podem ser utilizadas para melhorar a avaliação econômica de uma OE através da descoberta automática de um modelo que descreva dados históricos de avaliações econômicas. Esse modelo será utilizado para auxiliar o especialista na predição do valor econômico de uma OE, dando suporte às decisões de investimento, aumentando a eficiência do trabalho e oferecendo uma ferramenta capaz de realizar inferências sobre diversos cenários. Além disso, uma ontologia de domínio foi construída para auxiliar a tarefa de aprendizado do modelo preditivo, já que sendo uma ontologia uma especificação explícita e formal de uma conceitualização compartilhada (GRUBER, T., 1995),

julgamos que uma ontologia ajudaria na representação e entendimento do domínio em questão, fornecendo informações necessárias à construção do modelo preditivo.

1.2 Hipótese de Pesquisa

O trabalho de avaliação econômica de uma OE apresenta vários problemas. Alguns relatados pelos especialistas e outros identificados por nós após análise e expostos na Seção 1.1, como utilização de métodos tradicionais e dependência de artefatos como curva de produção e fluxo de caixa. A Figura 2 mostra o diagrama BPM representando os processos envolvidos no cálculo de uma avaliação, desde a identificação de uma OE até a sua colocação em uma Carteira de Projetos onde poderá ser escalonada e receber uma prioridade.

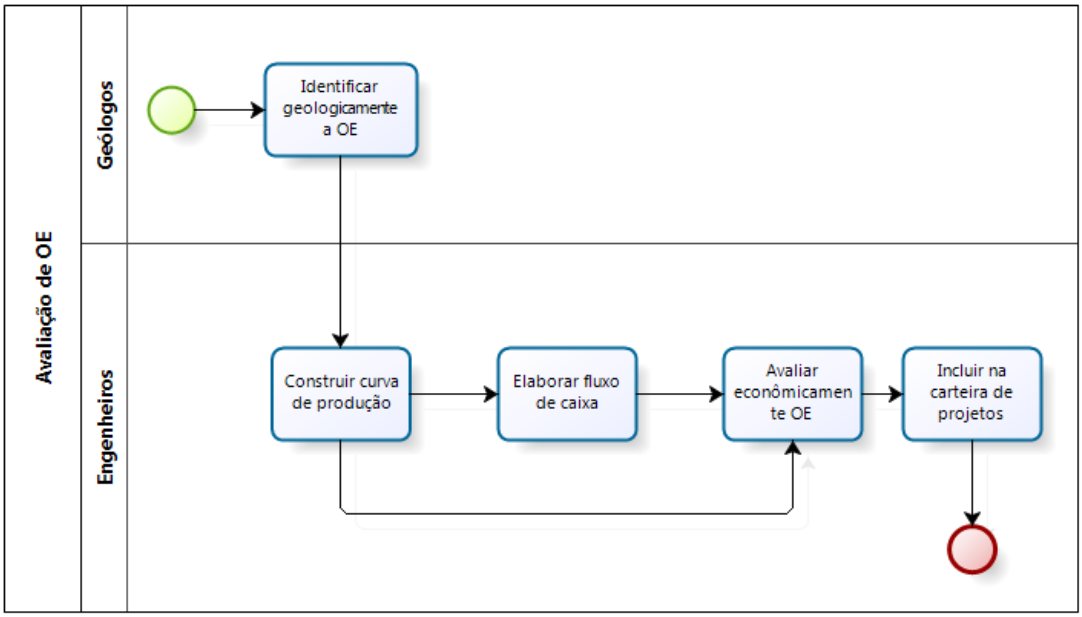


Figura 2: Processos relacionados à avaliação econômica

O tempo necessário para realizar uma avaliação também é considerado um fator crítico e pode comprometer o cumprimento de prazos dependendo do momento em que ocorre. Por ocasião da realização dos Leilões de Concessão de Área, promovido periodicamente pela ANP[4] (Agência Nacional de Petróleo), há um acúmulo de trabalhos de avaliações, pois vários cenários de várias áreas são considerados para se chegar a uma decisão sobre o valor econômico de uma concessão em disputa. Uma avaliação mal executada equivale a uma aposta mal sucedida, ocasionando prejuízos que serão repassados para a sociedade em forma de aumento do preço do combustível nos postos de abastecimento ou em forma de racionamento.

Outro fator a ser considerado é que quanto mais a tarefa de avaliação for eficiente menos dependeremos de importações de combustíveis, pois caso o custo de descoberta de novas acumulações se eleve em demasia será economicamente mais vantajoso importar do que produzir em nosso próprio território, ocasionando uma dependência em relação a produtores externos e, consequentemente, gerando uma dependência energética e, por vezes, política. Por estas razões a exploração de petróleo é considerada elemento estratégico para a soberania de uma nação.

Além disto, existe muita incerteza nesta atividade ocasionando um baixo índice de sucesso em torno de 20%, o que em termos práticos significa que cada descoberta tem que pagar a si e a outras quatro que foram mal sucedidas.

Pelo exposto acima definiremos nossa Hipótese como:

[4] http://www.anp.gov.br/

- **Hipótese**
    - Se for possível aprender um modelo, aplicando técnicas de mineração de dados e com o apoio de uma ontologia de domínio, capaz de fazer a predição das avaliações econômicas, então:
- **Tese**
    - A tarefa de avaliação econômica poderá contar com um sistema de suporte à decisão tornando o processo mais eficiente e, estando o domínio representado por um modelo, será possível realizar inferências sobre as variáveis que influenciam esta tarefa.

Adicionalmente teremos, como uma consequência desejada, uma diminuição no tempo de execução da tarefa de avaliação e, espera-se, mais precisão também. É importante ressaltar que o modelo apenas dará apoio ao trabalho dos especialistas, não tendo a intenção de substituir suas habilidades e conhecimentos, mas de complementá-las com técnicas de mineração de dados.

1.3 Metodologia de Pesquisa

Primeiramente foi necessário levantar os problemas que afetam a trabalho de avaliação econômica como tempo de execução, dependência de artefatos como curva de produção e tratamento de comportamentos imprevisíveis com métodos tradicionais. Foram levantados na literatura os trabalhos relacionados conforme mostra o Capítulo 5.

Para comprovação da Hipótese foi feito um Estudo de Caso. Para isto, foi necessário contatar uma empresa de petróleo e solicitar acesso ao histórico de avaliações a ser utilizado no aprendizado do modelo preditivo.

De posse destas informações iniciou-se o Estudo de Caso com a aplicação da técnica de Descoberta de Conhecimento em Banco de Dados (DCBD) que é constituída por três etapas: **Pré-processamento**, **Mineração de Dados** e **Pós-processamento** (HAN, et al., 2006).

Na etapa de pré-processamento foram identificados e corrigidos problemas na base de dados como o comportamento tipo cauda longa do atributo classe. Na etapa de mineração de dados foram aplicados algoritmos de aprendizado de Rede Bayesiana, considerados por nós como os mais adequados ao domínio em questão, caracterizado como de grande incerteza (NEWENDORP et al., 2009). Durante esta etapa os especialistas do domínio, pertencentes à empresa dona das informações, puderam inserir conhecimento no modelo através de sugestões na disposição dos nós da estrutura da Rede Bayesiana.

Paralelamente foi construída a Ontologia de Exploração de Petróleo que auxiliou no conhecimento do domínio e forneceu opções de processamento para os algoritmos de mineração de dados.

Ainda no Estudo de Caso, o modelo bayesiano aprendido foi avaliado segundo as métricas de mineração de dados e finalmente submetido aos especialistas para que pudessem avaliar a contribuição do modelo aos trabalhos de avaliação econômica.

No decorrer desta pesquisa elaboramos e submetemos quatro artigos em importantes eventos acadêmicos com o objetivo de verificar o interesse da comunidade científica pelo tema e colher contribuições que enriqueceram o nosso trabalho.

O primeiro evento que participamos foi o Workshop de Teses e Dissertações (WTDSI/2011, Salvador). Nesta ocasião nosso trabalho estava na fase inicial e basicamente versava sobre a importância da atividade exploratória no Brasil e como a mineração de dados poderia contribuir para sua melhoria. Neste evento apresentamos alguns resultados preliminares (AFFONSO, M. et al, 2011) . O segundo a aceitar nosso artigo foi o International Association for Development of Information Society (IADIS/2011, Rio de Janeiro) (AFFONSO, M. et al, 2011), nesta oportunidade nossos experimentos apresentavam resultados mais elaborados e já dispúnhamos de um modelo com acurácia satisfatória. O terceiro evento foi o Simpósio Brasileiro de Sistemas de Informação (SBSI/2012, São Paulo) (AFFONSO, M. et al, 2012) quando já possuíamos conclusões mais definitivas sobre a nossa proposta. O quarto evento foi o International Conference on Enterprise Information Systems (ICEIS/2012, Polônia) (AFFONSO, M. et al, 2012) ocasião em que foi aceito, mas não apresentado.

1.4 Organização da Dissertação

A Figura 3 apresenta a Organização desta Dissertação.

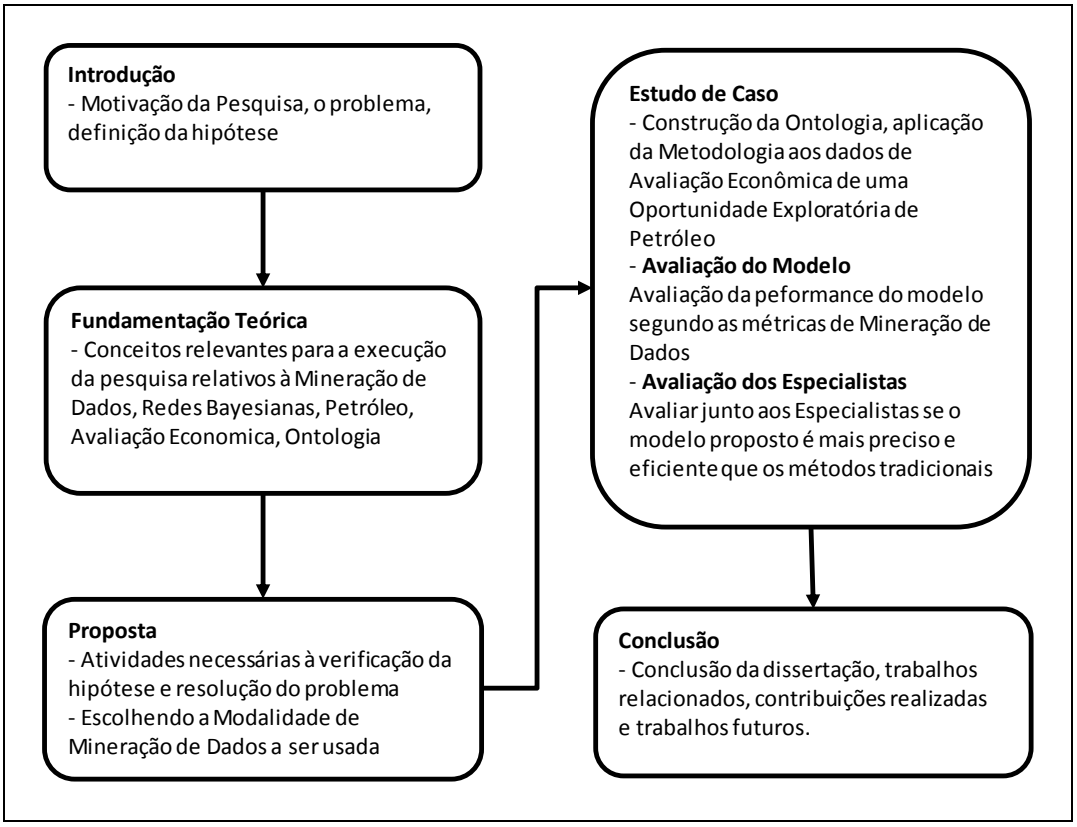


Figura 3: Organização da dissertação

Capítulo 2 – Fundamentação Teórica

Neste capítulo serão apresentados os conceitos considerados relevantes para a compreensão deste trabalho. Em Avaliação Econômica discorremos sobre o atual contexto e os principais métodos usados. O processo de Descoberta do Conhecimento em Banco de Dados é detalhado nas suas principais etapas (pré-processamento, mineração de dados e pós-processamento). Também consideramos necessário explicar os fundamentos de Redes Bayesianas, modalidade escolhida para representação do modelo preditivo. Com relação à Ontologia, foram descritos seus princípios, linguagens de construção e ferramentas.

2.1 Avaliação Econômica

Esta Seção descreve o atual contexto econômico relativo ao petróleo e as modalidades de avaliações econômicas utilizadas.

Contudo, é necessário esclarecer que na indústria do petróleo existem duas atividades principais: Exploração e Produção (conhecidas pela sigla E&P) que geralmente são consideradas uma só para efeito de simplificação, mas que guardam algumas peculiaridades importantes para os envolvidos.

Grosso modo, a Exploração tenta responder à pergunta onde se localiza uma jazida (ou acumulação) de petróleo que valha à pena seguir com um projeto para colocá-la em produção, quer dizer, o recurso ainda não foi descoberto. A atividade Produção não trabalha mais com esta incerteza e se encarrega de retirar desta jazida, já descoberta, o petróleo da maneira mais eficiente. Nesta ocasião, diminuíram-se as incertezas (talvez o volume seja um pouco menor ou maior, talvez a qualidade do óleo não seja tão boa).

Na Produção, a rocha reservatório é então caracterizada com maior precisão e passa a denominar-se reservatório. Um reservatório é composto por zonas de produção e atribui-se a ele um valor de reserva de petróleo que pode ser provada ou provável. Contudo, as atividades Exploração e Produção são estreitamente relacionadas e guardam conceitos em comum.

Não foi encontrada na literatura e nem na Petrobras uma ontologia que descrevesse os conceitos essenciais da avaliação econômica de uma OE como descrita neste trabalho.

2.1.1 Contexto Econômico da Atividade Exploratória de Petróleo

A atividade petrolífera no Brasil vem passando por transformações nestas últimas décadas. Até o ano de 1997 esta atividade era tipicamente monopolista, exercida apenas por uma empresa estatal. No entanto, este quadro viria a mudar com a promulgação da Lei do Petróleo nº 9.478/1997 que possibilitou que empresas estrangeiras e nacionais participassem desta atividade. Como consequência, a atividade petrolífera ficou exposta

à competição do mercado. A Tabela 1 mostra a quantidade de empresas por país que atualmente exercem atividades exploratórias no Brasil.

Tabela 1: Países que atuam no Setor Petróleo no Brasil (Fonte página ANP - Relação de concessionários 2012[5])

País Origem	Quant empresas		País Origem	Quant empresas
Angola	2		Espanha	1
Austrália	3		Estados Unidos	8
Brasil	39		França	1
Canadá	4		Holanda	1
Ilhas Cayman	1		Índia	2
China	2		Japão	1
Cingapura	1		Noruega	2
Colômbia	2		Panamá	1
Coréia	1		Portugal	1
Dinamarca	1		Reino Unido	4
Total Empresas 78 / Total países 20				

O intuito desta lei, abrindo o mercado a quem quiser nele investir, era modernizar os procedimentos e aumentar a produção nacional de petróleo, considerado insumo energético estratégico para se atingir o desenvolvimento. A lei criou a Agência Nacional de Petróleo (ANP) que seria o órgão regulador desta indústria e responsável pela definição de diretrizes para a participação do setor privado na pesquisa, exploração, refino, exportação e importação de petróleo e derivados.

O petróleo é composto por uma complexa mistura de hidrocarbonetos. Hidrocarbonetos são compostos orgânicos que ocorrem na Natureza e compõem-se de hidrogênio e carbono. Podem ser simples como o metano [CH4], porém alguns possuem moléculas complexas. Apresentam-se na forma de gases, líquidos ou sólidos. Suas moléculas podem ter a forma de cadeias, anéis e outras estruturas. Petróleo, gás e carvão são exemplos de hidrocarbonetos.

Com o intuito de propiciar transparência e igualdade de acesso às áreas detentoras de hidrocarbonetos a ANP instituiu os Leilões de Oferta de Concessões Exploratórias tipo Envelope Fechado. Esta decisão estimulou o desenvolvimento da produção de petróleo a partir do ano 2000. A Figura 4 mostra o efeito destas iniciativas na produção de petróleo ao longo dos anos.

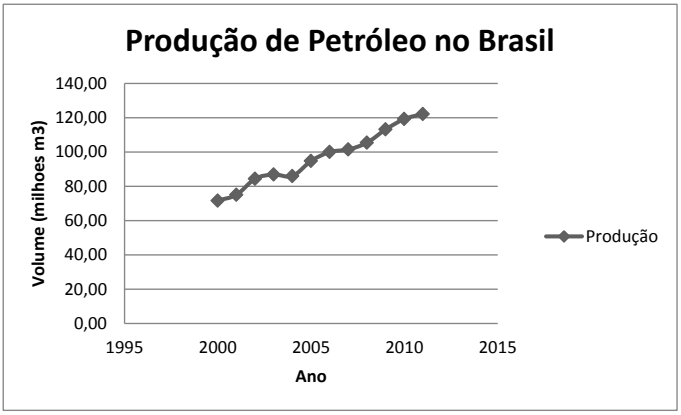


Figura 4: Evolução da produção de petróleo no Brasil (Fonte página ANP – Relatório de produção de derivados de petróleo 2012[6])

Como consequência destas mudanças, o Brasil alcançou a nona posição mundial na produção de petróleo conforme mostra a Tabela 2.

[5] http://www.brasil-rounds.gov.br/portugues/lista_de_concessionarios.asp

[6] http://www.anp.gov.br/?dw=8485

Tabela 2: Posição mundial da produção de petróleo (Fonte: Website U.S. Energy Information Administration – 2011[7])

	País	Produção MilhõesBarris/dia
1	Arábia Saudita	11,15
2	Russia	10,23
3	USA	10,14
4	China	4,30
5	Irã	4,23
6	Canadá	3,60
7	Emirados Árabes	3,10
8	México	2,96
9	**Brasil**	2,69
10	Kuwait	2,68
11	Iraque	2,64
12	Nigéria	2,53
13	Venezuela	2,47
14	Noruega	2,01
15	Angola	1,88

Há outras formas de obtenção de uma concessão exploratória como a modalidade **Cessão Onerosa**, reservada às acumulações do pré-Sal. Nesta modalidade a concessão é efetuada por lei e motivada por interesses nacionais. Na modalidade farm-in uma empresa se associa a outra que já detenha uma concessão, participando com um percentual. A ANP já promoveu 10 rodadas de negociação tipo leilão desde a promulgação da Lei 9.478/1997.

Com a descoberta das acumulações na camada pré-sal, ao longo do litoral brasileiro, a Lei 9.478/1997 foi atualizada pela Lei nº 12.351, de 22 de dezembro de 2010, que garantiu salvaguardas inerentes a grandes descobertas.

[7] http://www.eia.gov/countries/

Estas mudanças provocaram nas empresas a necessidade de buscarem novos métodos de prospecção, avaliação e produção de petróleo em território nacional. Uma empresa para sobreviver em um mercado competitivo deve avaliar seus riscos, participar de consórcios quando conveniente e estar atenta aos movimentos do mercado como preço do óleo, ameaça de conflitos externos, impostos e mudanças de estratégias governamentais. Acrescenta-se a isto, o fato de que o petróleo é uma fonte de energia não renovável, o que torna a sua descoberta mais rara na medida em que é produzido. Por isto, a cada dia é necessário melhorar os procedimentos de investigação se deseja encontrar uma grande acumulação.

Atualmente já não é suficiente encontrar uma jazida e avaliá-la apenas pelo seu potencial geológico. Tem-se de avaliá-la também economicamente para decidir se vale à pena continuar com o projeto ou redirecionar os recursos para outra OE que prometa maiores retornos financeiros, objetivo fim de uma empresa.

As atividades de grande importância trabalham frequentemente sob o risco de algo não sair conforme esperado, podendo ocasionar a inviabilidade do projeto. Exploração de petróleo é um clássico exemplo de decidir sob incerteza (NEWENDORP et al., 2009). Por isto, é importante a realização de uma Análise de Decisão que avalie os riscos de cada escolha e se decida pela opção que ofereça a melhor chance de se obter sucesso.

Mesmo que um poço de petróleo consiga-se descobrir uma grande acumulação, existem outros fatores que podem tornar esta descoberta um grande fiasco financeiro. Por exemplo, dependendo da profundidade em que a acumulação se encontra, da sua

distância da costa ou qualidade do hidrocarboneto, talvez a colocação desta jazida em produção não retorne os investimentos aplicados nela.

É necessária uma avaliação econômica conjugada com uma avaliação geológica para que o projeto seja considerado viável. Uma análise de decisão só será completa se forem levados em consideração também aspectos econômicos. Decisões tomadas apenas baseadas na experiência e intuição podem estar contaminadas de vieses, e não serão perdoadas em um mercado competitivo.

2.1.2 Métricas de Rentabilidade Econômica (NEWENDORP et al., 2009).

Não existe um consenso a respeito de qual método de avaliação econômica é melhor do que os demais. Porém existem características desejáveis como: (i) ter poder de comparação entre investimentos, como ordenação do pior para o melhor, (ii) considerar o valor do investimento no tempo (R$1,00 hoje é diferente de R$1,00 no futuro), (iii) considerar a incerteza do empreendimento e (iv) considerar outros fatores como metas corporativas, disposição ao risco e volume do investimento.

A avaliação mais simples e usada em projetos de baixo investimento é **Lucro = Renda – Custos**. Esta avaliação não leva em conta a passagem do tempo e a influência que este tem sobre o capital empregado. Tampouco considera o montante sendo investido. Outro ponto não levado em conta é a depreciação do bem em questão. No entanto, é simples e útil para uma primeira abordagem.

Payout (PO) é um método de avaliação de investimento que se preocupa em informar quanto tempo um determinado projeto levará para começar a pagar todo o

investimento. O empreendedor credita e debita suas receitas e despesas em uma única conta e periodicamente tira o saldo para saber se a conta está zerada. No início do projeto só haverá despesas, mas assim que começar a creditar os retornos financeiros a conta tenderá a zero (Figura 5). O ponto fraco deste método é considerar o fluxo de caixa apenas até o ponto onde começa a ter retorno.

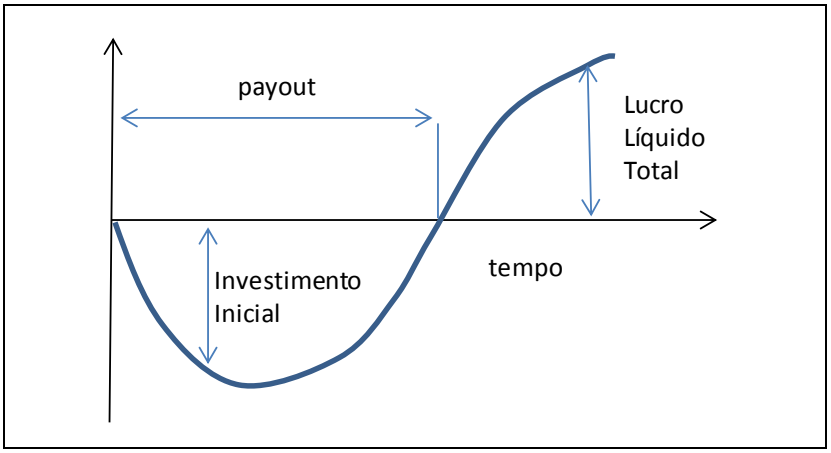


Figura 5: Curva de investimentos acumulados

Retorno de Investimento (ROI, sigla em inglês), diferentemente de Payout, considera o lucro total na avaliação de um investimento. ROI é uma medida sem dimensão e expressa a capacidade do projeto em gerar renda a cada unidade de moeda investida. Sendo **N** o número de períodos, sua fórmula é dada por:

$$ROI = \frac{\sum_{i=1}^{N} FluxoCaixa_i}{Investimento}$$

Os métodos de avaliação econômica vistos até o momento não levam em consideração o valor do tempo sobre os investimentos. No entanto, a moderna economia

financeira analisa o poder de compra de uma quantia em dinheiro sob dois aspectos: (i) valores recebidos no presente valem mais que os mesmos valores recebidos no futuro, isto é expresso através das taxas de juros praticadas nas operações de empréstimo bancário e (ii) no conceito de liquidez, que concede ao detentor do valor monetário o poder de escolher entre investir agora ou aguardar uma melhor oportunidade.

Valor Presente Líquido é um método de avaliação que considera o valor do tempo no seu cálculo. Uma sequência de valores a receber no futuro pode ser projetada para o presente numa operação denominada desconto. Desconto é a operação inversa dos juros compostos que definem um valor a ser pago no futuro por um empréstimo realizado no presente. O Valor Presente (VP) de uma quantia no futuro (VF) é dado por $VP = FV / (1 + i)^t$, sendo **i** a taxa de juros e **t** a quantidade de períodos.

O Valor Presente Líquido, considerado o critério de decisão mais usado, baseia-se no VP e é definido pela fórmula (NEWENDORP et al., 2009) a seguir:

$$VPL = \sum_{i=1}^{N} \frac{E(FC_i)}{(1 + TMA)^{i-j}}$$

Na fórmula acima **N** representa o número de períodos do investimento (ex. anos), **FC** o fluxo de caixa, **E** é a probabilidade de **FC** ocorrer, **TMA** é a Taxa Mínima de Atratividade, isto é, a taxa de juros que tornaria o investimento atrativo. A variável **j** ajusta o expoente **i** para que este reflita se os valores de **FC** incidem no início de cada período (j=1), no meio (j=0,5) ou o final do período (j=0).

O VPL apresenta algumas vantagens sobre os demais métodos por considerar a independência de cada FC e ajustar seus valores no tempo. Utiliza apenas uma taxa (TMA) no cálculo, porém, em casos de instabilidade do mercado, é possível definir uma TMA para cada período. Usa a probabilidade **E,** que pode ser fixa ou variável**,** para considerar os riscos envolvidos na realização do **FC**.

Se o valor de VPL for igual a zero o investimento nem adiciona nem subtrai valores. Se for maior que zero, o investimento gerará retorno dado pelo valor do VPL. Se o VPL for negativo, provavelmente o projeto será descartado a não ser que haja motivos não financeiros para realizá-lo. Outra vantagem na utilização do VPL é que os projetos poderão ser ordenados em uma fila de retorno de investimento, base para a formação de uma carteira de projetos.

Porém, como podemos perceber pela fórmula do VPL, este é um método analítico convencional. Alguns domínios como exploração de petróleo trabalham com conceitos imprevisíveis como preço do óleo no mercado, flutuações econômicas e impostos. Estas aproximações podem levar a erros e acrescentar mais riscos aos investimentos.

Na indústria de petróleo usa-se basicamente o VPL para se avaliar economicamente o quanto uma descoberta é capaz de oferecer como retorno de investimento (JUNIOR, 2003). Quando se identifica uma OE com características de bom investimento, o primeiro passo é elaborar uma curva de produção de óleo, onde a estimativa de produção, ano a ano, é refletida levando-se em conta informações aproximadas sobre a OE. A Figura 6 ilustra uma curva de produção típica que começou a produção em 2011, alcançará produção máxima em 2015 e se esgotará em 2041.

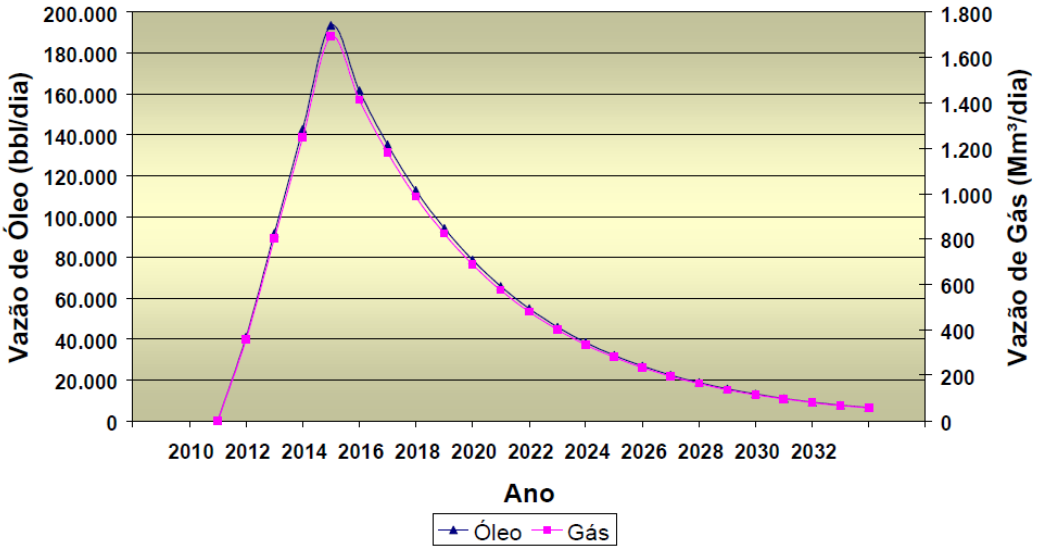


Figura 6: Curva de produção típica (SILVA, B. et al., 2006)

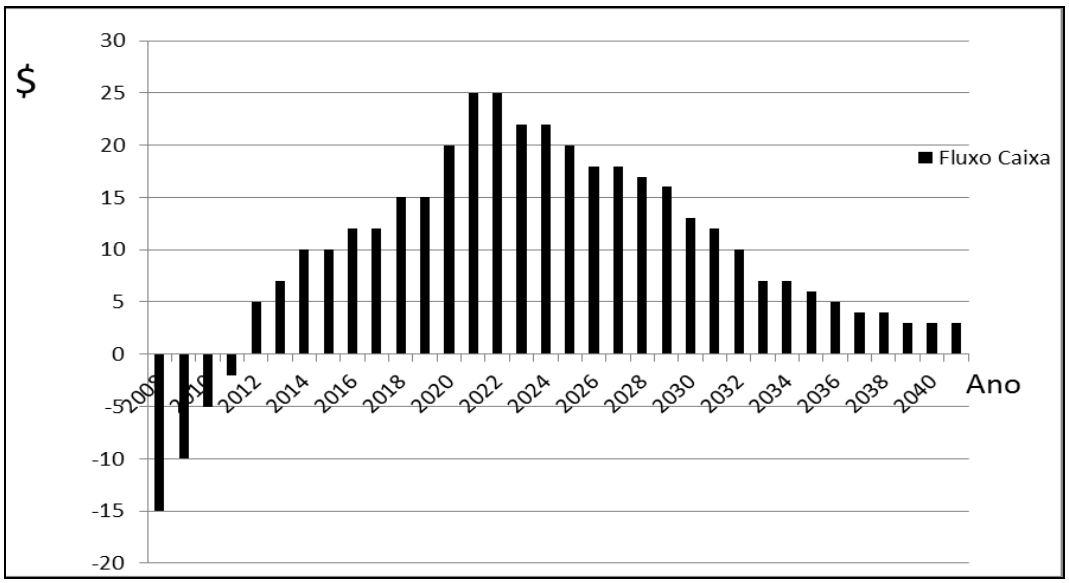


Figura 7: Fluxo de caixa hipotético

O segundo passo é o cálculo do Fluxo de Caixa relativo a esta curva de produção. O fluxo de caixa expressa o retorno líquido (receita menos despesa) ano a ano, por toda vida útil do campo. A Figura 7 representa o fluxo de caixa hipotético correspondente à curva de produção da Figura 6. Note que os investimentos se iniciaram em 2008. O fluxo de caixa se torna positivo à medida que o campo entra em produção em 2011.

De posse destas informações, pode-se calcular o VPL que consistirá no somatório dos fluxos de caixa corrigidos para o valor presente.

2.2 Descoberta do Conhecimento em Banco de Dados (DCBD)

A quantidade de dados armazenados em dispositivos digitais vem crescendo ano a ano, a uma taxa de 23%, de acordo com o site da University of Southern California [8], publicado em 2011. Uma das causas foi o progressivo barateamento dos meios de armazenamento, popularização do uso de computadores e a oferta de drives, inclusive grátis, na Web. A internet, as câmaras fotográficas digitais, logs de aplicações e as ferramentas de coleta automática contribuem para esta inundação de dados. Isto sem contar com os dados antigos que estão sendo digitalizados.

Com todos estes dados à disposição, surgiu a possibilidade de utilizá-los para extração de informações úteis e inéditas. O objetivo do processo de Descoberta do Conhecimento em Banco de Dados (DCBD) é identificar padrões que representem conhecimento interessante e novo, implícito nos dados armazenados em banco de dados, data warehouse, na Web e outros repositórios (HAN, J. et al., 2006). É comum o uso do termo mineração de dados como sinônimo de DCBD, porém mineração de dados refere-se apenas a uma das suas etapas. DCBD é um processo multidisciplinar e faz uso de técnicas de várias áreas como: Aprendizado de Máquina, Banco de Dados, Reconhecimento de Padrões, Estatística e outras.

[8] http://news.usc.edu/#!/article/29360/How-Much-Information-Is-There-in-the-World

DCBD é um processo formado pelos seguintes passos: pré-processamento, mineração de dados e pós-processamento, conforme ilustrado na Figura 8.

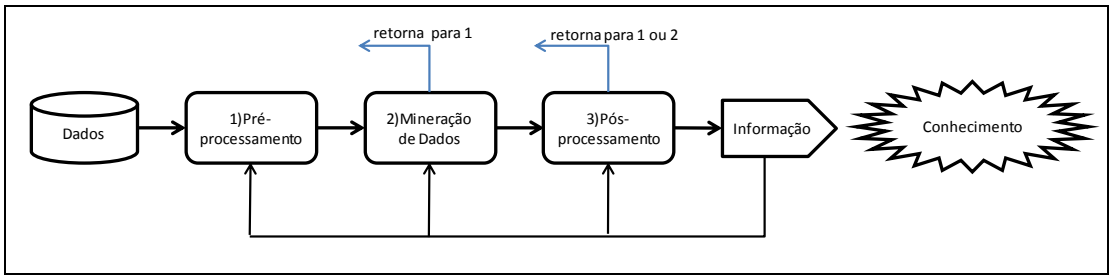


Figura 8: Descoberta de Conhecimento em Bases de Dados

Nas próximas seções iremos detalhar cada uma destas etapas.

2.2.1 Pré-Processamento

O passo de pré-processamento tem a finalidade de preparar os dados para a etapa de mineração de dados. Tarefas como integração de dados, limpeza e remoção de outsiders são realizadas nesta etapa (HAN, J. et al., 2006). O pré-processamento é considerado a etapa que demanda maior esforço de todo processo de DCBB. A seguir as técnicas de pré-processamento consideradas nessa dissertação são descritas.

2.2.1.1 Integração dos Dados

Integração dos dados visa reunir os dados coletados em um repositório consistente. Em geral, os dados estão distribuídos por várias fontes como banco de dados, planilhas eletrônicas e outros. O trabalho de integração consiste em consolidar estes dados em uma fonte única, removendo as redundâncias e resolvendo conflitos de informação. Por exemplo, um mesmo indivíduo pode aparecer com nomes diferentes em fontes diversas.

2.2.1.2 Limpeza dos Dados

A Limpeza dos Dados consiste em: preencher os atributos faltantes com valores legítimos, identificar/remover outliers e tratar as inconsistências (HAN, J. et al., 2006). Estes problemas podem ocorrer por erro humano ou indisponibilidade dos dados no momento da sua entrada no sistema. Geralmente ocorrem quando o dado é armazenado em ambientes como planilhas eletrônicas que possuem poucas verificações automáticas de consistência, diferentemente de um sistema de gerenciamento de banco de dados (SGBD) que possui recursos como integridade referencial e controle de acesso que ajudam a manter a integridade.

Com relação a dados faltantes, o ideal é que se tente recuperar os valores originais consultando outras fontes e fazer o preenchimento manualmente. Caso não seja possível, utilizam-se técnicas para preenchimento dos dados faltantes com valores válidos (HAN, J. et al., 2006). Estas técnicas dependem do tipo do atributo. Existem dois principais tipos: nominais que assumem apenas valores pertencentes a um conjunto finito, por exemplo, {Azul, Vermelho, Verde} e atributos contínuos cujos valores pertencem ao conjunto dos números reais.

Uma maneira simples e rápida de tratar dados faltantes é a remoção das instâncias que contêm valores faltantes, porém, se houver poucas instâncias e muitos valores faltantes, pode haver comprometimento da etapa de mineração de dados. Outra solução seria, tratando-se de atributos nominais, preencher o campo com o valor mais frequente do atributo considerando-se a classe. Caso o atributo seja contínuo, preenche-se com a média aritmética.

Os métodos de imputação citados acima podem inserir um viés no conjunto de dados a serem minerados. Um método que menos compromete é determinar o valor do atributo faltante através dos atributos restantes. Por exemplo, podemos aprender um modelo classificador que tenha o atributo faltante como classe e classificar as tuplas incompletas submetendo-as a este classificador.

Um erro que pode ocorrer, principalmente quando o ambiente não possui recursos de verificações automáticas, é quando um atributo nominal assume um valor fora do conjunto permitido. Por exemplo, pode ocorrer o valor Amarelo quando o conjunto permitido é {Azul, Vermelho, Verde}. Nestes casos pode-se proceder como no caso dos dados faltantes para corrigir o valor.

2.2.1.3 Remoção de Outliers

Outliers são dados contínuos que se destacam dos demais por apresentarem valores fora de uma faixa comum. Em algumas situações é possível identificá-los através da construção de gráficos, porque outliers se destacam visualmente dos demais. No entanto, este método não é preciso, podendo induzir a erros. O salário do presidente de uma companhia se destaca das demais remunerações, mas nem por isto podemos considerar esta situação como outlier. Tem-se que ter o cuidado de analisar cada caso, pois, por exemplo, num sistema de Detecção de Fraude em Cartão de Crédito a presença de um outlier pode indicar uma possível fraude. Geralmente usam-se métodos estatísticos para identificação e a partir daí fazer uma análise caso a caso.

Um método estatístico utilizado na detecção de outliers devido à sua eficiência e rapidez é o Interquartil Range (**IQR**) (MACLAVE et al., 2009). IQR é considerada uma medida de dispersão, semelhante ao desvio padrão, e é definido por **IQR = Q3 – Q1**, sendo **Q3** o terceiro quartil e **Q1** o primeiro. A Figura 9 ilustra o conceito de IQR.

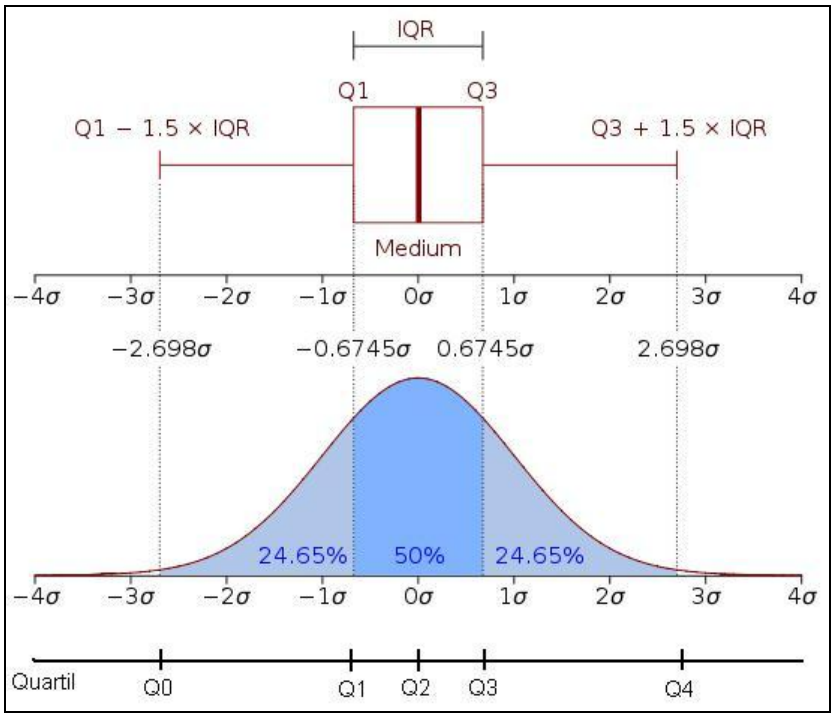


Figura 9: Diagrama do IQR para identificação de outliers

Os quartis dividem o conjunto de dados em quatro partes aproximadamente iguais (duas com 24,65% e outras duas com 25%) conforme ilustra o eixo Quartil da Figura 9. Se tomarmos a parte central (Q3-Q1) estaremos considerando 50% da totalidade dos dados. Os dados se estendem para a esquerda até (Q1-1.5 * IQR). E à direita, se estendem até (Q3 + 1.5 * IQR). Percebe-se que a totalidade dos dados fica compreendida entre estes limites. Os dados que se localizam fora destes limites já poderiam ser considerados outliers. Na prática, costuma-se definir como Outlier os valores de **x** tais que (Q1 – 6 * 1.5 * IQR) < **x** ou **x** > (Q3 + 6 * 1.5 * IQR).

IQR é uma medida de dispersão, assim como o desvio padrão, porém possui cálculos mais simplificados, sendo ideal quando se lida com grande quantidade de dados. Usaremos este método nos nossos experimentos para identificação dos outliers e analisaremos cada caso para decidir pela remoção ou não da instância.

2.2.1.4 Normalização

Uma vez que a limpeza esteja realizada, inicia-se o passo de transformação dos dados. Este passo refere-se à aplicação de funções matemáticas aos atributos da base de dados de modo a prepará-la para a mineração de dados. Por exemplo, às vezes é mais interessante saber a idade de uma pessoa do que sua data de nascimento. Outras vezes é necessário concatenar dois atributos em um atributo único. Ou então, é conveniente gerar um atributo baseado nos demais. Há situações onde é desejável ocultar o nome do atributo, por motivos de segurança da informação, alterando-o para um nome obscuro.

Algumas transformações são muito importantes e usadas constantemente, como normalização e discretização. Normalização consiste em transformar os valores numéricos de uma base de dados para uma escala entre limites pré-definidos como, por exemplo, entre 0 e 1. A fórmula de normalização, dados os limites inferior e superior do intervalo, é dada na Figura 10:

Legenda:
$A_{normalizado}$ = valor normalizado
valor_min = valor mínimo que A pode assumir
valor_max = valor máximo que A pode assumir
Lim_{sup} = Limite superior do intervalo
Lim_{inf} = limite inferior do intervalo

Fórmula:

$$A_{normalizado} = \frac{valor\_min}{valor\_max - valor\_min} * (Lim_{Sup} - Lim_{inf}) + Lim_{inf}$$

Figura 10: Fórmula para normalização de um valor

O objetivo da normalização é não permitir que variáveis que possuam um intervalo grande de valores dominem o processo de aprendizado dos modelos. Pode ser usado, também, quando se pretende ocultar os valores da variável.

2.2.1.5 Discretização

A discretização é aplicada aos dados com o objetivo de transformar dados contínuos em dados discretos. Esta transformação divide os valores do atributo em intervalos, diminuindo a quantidade de valores que um atributo pode assumir. A discretização permite, na etapa de mineração de dados, que sejam aplicados algoritmos que apenas utilizam dados nominais, aumentando assim o número de algoritmos disponíveis para a experimentação, porém com perda de informação.

Existem dois métodos principais: Igual-largura e Igual-frequência. Em Igual-largura dividimos o intervalo total de valores do atributo em intervalos com larguras iguais, sendo o comprimento dos intervalos definidos da seguinte maneira (Figura 11):

Legenda:
Comp_Intervalo = Comprimento do Intervalo
valor_min = valor mínimo do atributo
valor_max = valor máximo do atributo
N = número de Intervalos

Fórmula:

$$\text{Comp\_Intervalo} = \frac{\text{valor\_max} - \text{valor\_min}}{N}$$

Figura 11: Cálculo do intervalo de discretização

A escolha do número de intervalos é feita de maneira empírica e pode ocorrer que alguns intervalos não possuam nenhum elemento na base de dados que o represente, comprometendo o processo de aprendizado do modelo. A abordagem Igual-frequência propõe que cada intervalo possua o mesmo número de elementos, não importando a largura deste. Isto resolve o problema dos intervalos não representados.

Identificamos na literatura trabalhos que versam sobre tipos mais sofisticados de Discretização de classes como o K-means Clusterização (KM) proposto em (REVOREDO et al, 2004). Este método inicialmente aplica Igual-largura e, a seguir, move cada elemento para o intervalo vizinho se este movimento reduz a distância de cada elemento ao centro do cluster, com a restrição de que cada intervalo deva possuir pelo menos um elemento.

2.2.1.6 Sampling

Outro recurso que faz parte do pré-processamento é chamado de Sampling (HAN et al., 2006). O sampling possibilita a geração aleatória de uma nova base de dados a partir da

base original. Pode ser sem reposição, isto é, as instâncias não se repetem. Ou com reposição, quando as instâncias selecionadas podem se repetir. Este método é usado nas situações em que o número de instâncias na base de dados ou é pequeno demais ou grande demais para serem processado com eficiência pelos algoritmos de mineração de dados.

2.2.1.7 Redução de Dimensionalidade

Em algumas situações, ocorrem dúvidas se não estamos utilizando um número de atributos excessivo em nossa base de dados. Questiona-se se um número menor poderia trazer o mesmo benefício com menos custos computacionais. Além disto, atributos considerados irrelevantes podem deteriorar o desempenho do aprendizado de 5 a 10% (WITTEN et al. 2011). Para evitar estes problemas, os atributos que realmente podem influenciar positivamente o aprendizado do modelo podem ser identificados e assim os demais são removidos. Isto se chama redução de dimensionalidade.

O conhecimento do domínio em questão pode auxiliar o especialista na seleção manual dos atributos, mas existem algoritmos que detectam automaticamente os atributos mais influentes. Entre eles se destacam os Algoritmos Genéticos (AG) que serão abordados na seção a seguir.

2.2.1.7.1 Algoritmos Genéticos (AGs)

A Natureza tem inspirado os cientistas na criação de modelos que procuram resolver problemas científicos complexos. Os Algoritmos Genéticos é um representante desta

classe e foram inspirados nos sistemas naturais biológicos e na Teoria da Evolução das Espécies, enunciada pelo naturalista inglês Charles Darwin (GOLDBERG, D., 2006).

Porém, foi a partir do trabalho de John Holland na Universidade de Michigan, em 1960, que foram definidas as bases para o desenvolvimento dos Algoritmos Genéticos.

Os AGs pertencem à classe dos Algoritmos Evolucionários que são modelos computacionais de processos naturais para resolver problemas. AGs trabalham com operadores básicos chamados **Seleção**, **Reprodução** e **Mutação**. Estes operadores são aplicados sobre uma população de indivíduos para aumentar seus desempenhos num dado meio ambiente conforme visto na Figura 12.

AG utiliza o conceito de população que é um grupo de estruturas ou indivíduos em um determinado meio ambiente. Após a aplicação de operadores sobre uma população surge uma nova população e cada nova população é chamada de geração.

O operador seleção seleciona entre os indivíduos de uma população os mais aptos segundo uma função de avaliação. A seguir, o operador reprodução faz com que os indivíduos selecionados se reproduzam gerando novos indivíduos que serão introduzidos na população no lugar dos menos aptos, resultando numa nova população. Eventualmente ocorre uma mutação, isto é, uma alteração aleatória em alguns dos indivíduos. Este processo se repete um número pré-determinado de vezes.

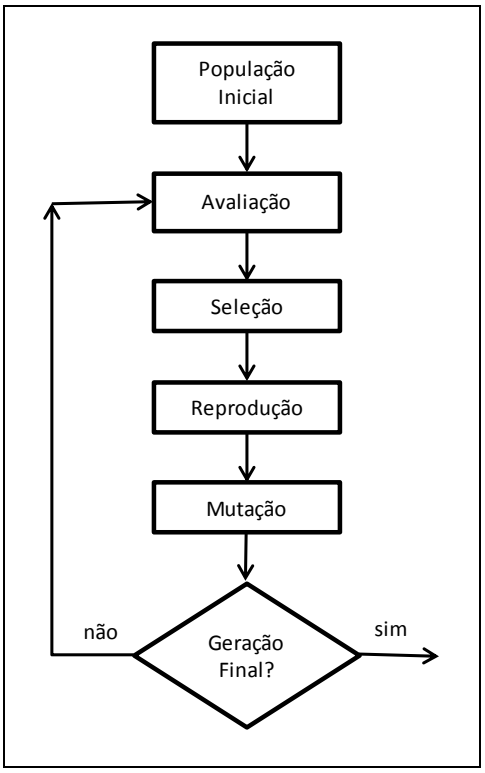


Figura 12: Fluxograma básico de um AG

Algoritmos Evolucionários dependem de fatores estocásticos (probabilísticos) para seu processamento. Por causa disto, são considerados como heurísticas para atingir um resultado ótimo. Entende-se por heurísticas boas práticas que podem levar a um bom resultado, sem oferecer garantias.

Na busca por uma solução ótima, não apenas o tamanho do espaço de amostragem pode dificultar a tarefa. Algumas funções (LINDEN, 2008) apresentam vários picos de valores, ou melhor, são multimodais (Figura 13). Se utilizarmos um algoritmo tipo gradiente (Hill Climbing) este pode ficar preso em um máximo local, dificultando o atingimento do máximo global.

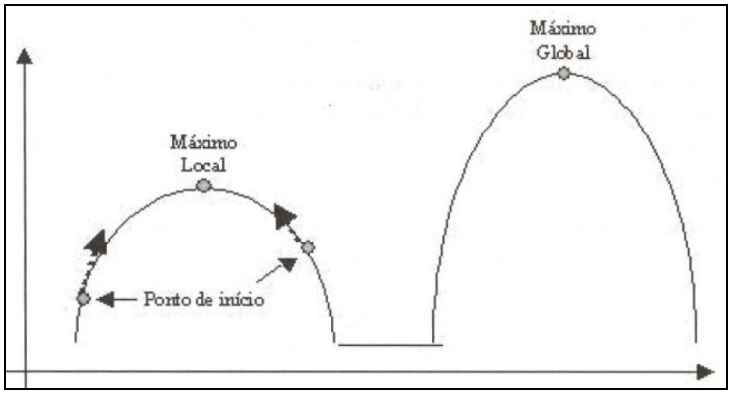


Figura 13: Função multimodal

Com os AGs esta situação é mais difícil de ocorrer. Primeiro porque se trabalha com uma população de indivíduos. Segundo, seus operadores fazem uma varredura mais eficiente do espaço de busca. Por exemplo, o operador de mutação pode criar um indivíduo que, devido a sua avaliação, consiga sair de um mínimo local.

Alguns pesquisadores (LINDEN, 2008) consideram os AGs algoritmos genéricos, portando não são a solução para todos os problemas e nem devem ser usados indiscriminadamente. Existem algoritmos específicos para determinadas ocasiões que apresentam grande eficiência e, portanto, devem ser usados em substituição aos GAs. A Figura 14 compara os desempenhos de algoritmos específicos e genéricos.

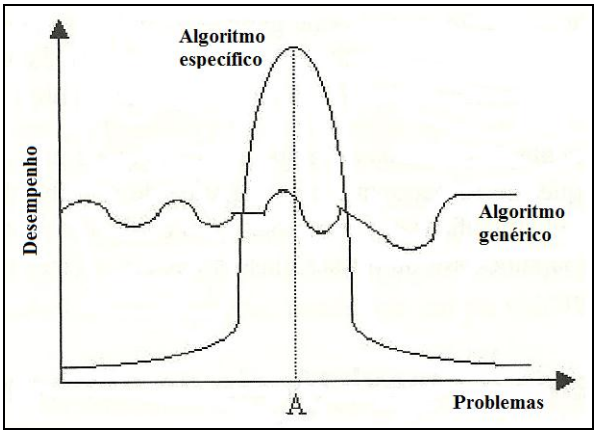


Figura 14: Comparação de desempenho entre algoritmos específicos e genéricos

2.2.1.8 Seleção do Atributo mais Influente

Uma maneira simples de realizar uma classificação é considerar apenas o atributo mais influente do conjunto de atributos. Esta é uma técnica que consome poucos recursos de máquina e que, em algumas circunstâncias, pode apresentar bons resultados. Também é utilizado como uma abordagem inicial para conhecer melhor os dados.

OneR (HOLTE, R., 1993) é um exemplo de algoritmo de aprendizado baseado nesta ideia. Este produz uma regra de classificação que testa apenas o atributo mais influente, conseguindo alcançar bons níveis de acurácia (WITTEN et al., 2011). Seu princípio básico é testar a capacidade de classificação de cada atributo da base e selecionar o melhor. O pseudocódigo de OneR é mostrado na Figura 15.

```
Para cada atributo
    Para cada valor deste atributo construa uma regra conforme a seguir
        Conte a frequencia de cada classe
        Considere a classe com maior frequencia
        Faça uma regra associando a classe a este valor de atributo
    Calcule a taxa de erro das regras
Selecione as regras com a menor taxa de erro
```

Figura 15: Pseudocódigo do algoritmo OneR

2.2.2 Mineração de Dados

Finalizada a etapa de pré-processamento, considera-se que a base de dados está preparada para a etapa de mineração de dados que consiste na aplicação de técnicas de Aprendizado de Máquina (MITCHELL, 1997) que irão extrair o conhecimento interessante, novo e oculto na base. Para tal, será necessário aprender um modelo que será capaz de classificar ou predizer instâncias inéditas.

Define-se como classificação a capacidade de um modelo de classificar dados nominais ou discretos em um dentre vários possíveis valores de classes. Por exemplo, uma financeira poderá tomar a decisão de fornecer empréstimo a uma pessoa após coletar alguns de seus dados como idade, faixa salarial e estado civil. Submetem-se estes dados ao modelo e ele irá classificá-la como apta ou não ao empréstimo (ver Figura 16).

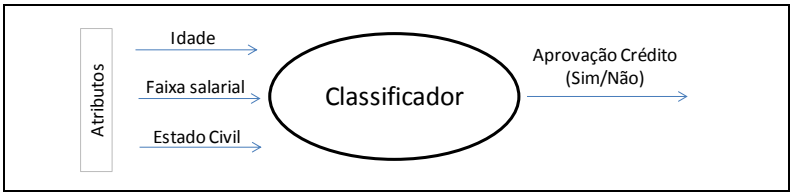


Figura 16: Classificador

A predição funciona de modo similar, porém é aplicada quando o atributo classe é contínuo, não importando o tipo dos demais. Considerando ainda a Figura 16 do exemplo anterior, poderíamos, caso o atributo classe fosse contínuo, predizer não apenas Sim/Não, mas o valor a ser emprestado. Aplicando-se a discretização é possível transformar um problema de predição em classificação quando se avalie que haverá algum ganho de desempenho ou se deseja ter maiores opções de aprendizado que compensem a perda de informação.

O aprendizado de máquina é dito supervisonado quando ocorre a partir de uma base de dados onde a classe das instâncias é conhecida. Nesta situação o modelo aprendido leva em consideração a classe de cada instância. Porém, em alguns casos não dispomos desta informação. Quando isto acontece diz-se que o aprendizado será do tipo não supervisionado. Algoritmos de clusterização e de regras de associação são típicos de aprendizado não supervisionado.

O aprendizado de um modelo é composto por duas fases: treinamento e teste. Na fase de treinamento uma base de dados categorizada é submetida a um algoritmo de aprendizado e este aprende um modelo capaz de prever a categoria de novas instâncias.

Após a fase de treinamento realiza-se a fase de teste que tem o objetivo de avaliar o desempenho do modelo quando classifica dados inéditos. Para isto reserva-se um percentual (por exemplo, 1/3) das instâncias que compõem a base de dados para serem classificados. Esta técnica é conhecida por hold-out. Um dos problemas desta abordagem é que se a base for pequena estaremos privando a fase de treinamento de

preciosas instâncias. Outro problema é definir quais instâncias deveremos reservar para testes. O ideal é que sejam representativos da população.

Uma técnica que visa minimizar estes problemas é chamada validação cruzada (WITTEN et al., 2011). Consiste em, a cada rodada de treinamento, utilizar partições (folds) diferentes para treinamento e teste retiradas da mesma base e, no final, o desempenho do modelo será dado pela média de desempenho das rodadas. Por exemplo, na validação cruzada 10-fold particiona-se a base de dados em 10 partes iguais. A cada rodada utiliza-se nove partições para treinamento e realiza-se o teste com a partição restante. Ao final da décima rodada calcula-se a média de desempenho. A validação cruzada é a técnica padrão utilizada para avaliação de modelos aprendidos.

No entanto, existem outras técnicas que se destacam dependendo da situação que se tem em mãos. Uma delas é chamada leave-one-out. Esta técnica funciona similarmente à validação cruzada, mas a cada rodada separa apenas uma instância para teste e o restante para treinamento. É ideal quando se trata de uma base com poucas instâncias, pois uma grande parte da base será usada para treinamento. Porém, para grandes bases haverá um custo computacional enorme, pois exigirá uma rodada para cada instância. Outra vantagem da técnica leave-one-out é ser considerada um procedimento determinístico porque, mantendo-se a mesma base, sempre se chegará ao mesmo resultado devido ao fato de não utilizar escolhas aleatórias.

Nas seções seguintes serão discutidos em mais detalhes os algoritmos de aprendizado de máquina utilizados nesta dissertação, sendo que o aprendizado de RB, a técnica principal utilizada nessa dissertação, será abordado na Seção 2.3 em destaque.

2.2.2.1 Aprendendo Redes Neurais Artificiais

A Rede Neural Artificial (RNA) é um modelo inspirado na biologia do funcionamento do cérebro humano (HAYKIN, 1999). A sua topologia é composta de camadas e cada camada possui nós que se conectam aos nós da camada seguinte. Os algoritmos utilizados para construir uma RNA requerem um longo tempo de treinamento e, por isto, não são adequados aos casos de aprendizado de máquina em que o tempo é crítico. A topologia é determinada empiricamente (número de camadas ocultas e número de nós em cada camada oculta).

A RNA é adequada quando se trabalha com valores contínuos ou quando se tem pouca informação sobre o domínio estudado tal como no trabalho de Reconhecimento de Caracteres Escritos. Ela possui a qualidade de tolerar bem dados com muitos ruídos. A Figura 17 mostra uma topologia de RNA com seis nós, uma camada de entrada, uma oculta e uma de saída. Cada aresta, ligando um nó a outro, possui um peso **w** associado.

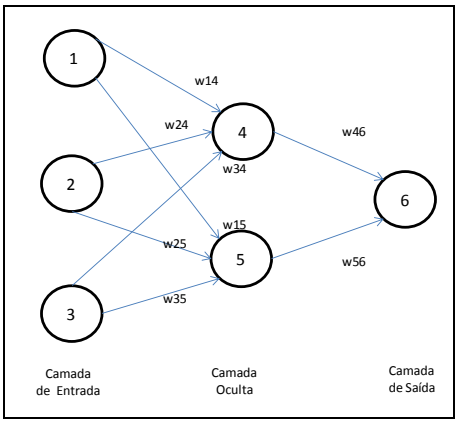


Figura 17: Topologia de uma rede neural artificial

Para classificar uma instância, esta é alimentada na rede pela camada de entrada que possui um nó para cada atributo. Cada valor de atributo é multiplicado por um peso **w** e,

caso atinja um valor limiar mínimo (Threshold), será repassado para a camada oculta seguinte. Após passarem pela última camada oculta, os valores intermediários são propagados para a camada de saída que somará as entradas e classificará a tupla.

Faz parte do treinamento da RNA a escolha de uma topologia eficiente que se dá por tentativa e erro. Não há regras claras para se definir o melhor número de camadas ocultas nem os valores iniciais dos pesos, esta escolha dependerá do problema a ser resolvido. Às vezes é necessário reiniciar o treinamento usando uma nova topologia e novos valores de pesos caso os resultados não sejam satisfatórios.

O algoritmo Backpropagation é o mais utilizado para treinamento de modelos RNA. Consiste em passar uma instância pela rede, calcular o valor predito e compará-lo com o valor real. Caso haja uma diferença, os valores dos pesos são reajustados na ordem inversa, isto é, da camada de saída para a primeira camada oculta (daí o nome Backpropagation). Este é um processo iterativo que só termina quando o erro converge para um valor aceitável. Os pesos geralmente são inicializados com valores aleatórios entre 0 e 1.

Backpropagation utiliza como função de saída (Threshold) a função sigmóide devido a sua facilidade de derivação, usada no cálculo do erro. Outra vantagem é que, por ser não-linear permite a criação de modelos que representem domínios complexos. Outro parâmetro importante de uma RNA é a taxa de aprendizado que tem a finalidade de controlar a velocidade de correção do erro, evitando uma convergência prematura dos valores dos pesos **w**.

2.2.2.2 Métricas de Avaliação de Modelos Aprendidos

O desempenho dos classificadores (dados nominais) é dado pelo percentual de instâncias corretamente classificadas em relação ao total de instâncias de teste. Esta métrica chama-se acurácia (Acc). Alguns autores preferem usar taxa de erro que é dada por taxa_erro = 1-Acc (HAN et al., 2006). Esta é uma medida padrão de desempenho do modelo.

A acurácia pode ser detalhada por valores de classe, dada por uma matriz chamada matriz de confusão. Isto ajuda a identificar onde se localizam os problemas de aprendizado. Através da matriz de confusão podemos verificar qual valor de classe do nosso modelo foi bem aprendida e quais não foram, podendo levar o processo de DCBD de volta ao pré-processamento, como coletar mais dados.

Na Figura 18 mostramos a matriz de confusão e a acurácia correspondente a um modelo aprendido. O número total de instâncias é 14 e existem dois valores de classe (Sim/Não). Das nove instâncias com classe = Sim, seis foram classificadas corretamente e três foram incorretamente A acurácia do modelo para este conjunto de dados é 9/14 = 64% e a Taxa de erro = 36%.

```
=== Matriz de Confusão ===

 a     b  <-- classificado como
 6     3 | a = Sim
 2     3 | b = Não

Exemplos Corretamente Classificados         9      64%
Exemplos Incorretamente Classificados       5      36%
```

Figura 18: Exemplo de matriz de confusão

Porém, existem outras métricas importantes que complementam a avaliação dada pela acurácia. Primeiro precisamos definir os seguintes conceitos:

- TP (True Positive) = Verdadeiro Positivo; totaliza as instâncias positivas e que foram classificadas desta forma. Na Figura 18 temos TP = 6, elemento (1,1) da matriz.
- FP (False Positive) = Falso Positivo; totaliza as instâncias negativas que foram classificadas como positivas. Na Figura 18 temos FP = 2, elemento (2,1) da matriz.
- TN (True Negative) = Verdadeiro Negativo; totaliza as instâncias negativas e classificadas desta forma. Na Figura 18 temos TN = 3, elemento (2,2) da matriz.
- FN (False Negative) = Falso Negativo; totaliza as instâncias positivas que foram classificadas como negativas. Na Figura 18 temos FN = 3, elemento (1,2) na matriz.

Usando os conceitos acima é possível definir as seguintes taxas

- Taxa de Verdadeiros Positivos (TP rate) = TP / (TP + FN). Esta taxa expressa a capacidade do modelo em identificar verdadeiros positivos. Também chamada de Recall
- Taxa de Falso Positivo (FP rate) = FP / (FP + TN). Esta taxa expressa a tendência do modelo em errar ao afirmar que identificou um positivo.
- Precision = TP / (TP + FP). Expressa a precisão do modelo ao identificar os Verdadeiros Positivos (TP).

A métrica estatística Kappa ajusta o valor medido pela acurácia descontando desta os acertos produzidos por um modelo puramente aleatório, quer dizer, produzidos pela sorte. Por exemplo, um modelo aleatório poderia produzir, usando os mesmos dados usados na Figura 18, a matriz de confusão da Figura 19:

```
Classificador Aleatório
=== Matriz de Confusão ===
 a     b   <-- classificado como
 5     4 | a = Sim
 3     2 | b = Não

Exemplos Corretamente Classificados          7      50%
Exemplos Incorretamente Classificados        7      50%
```

Figura 19: Matriz confusão de modelo aleatório

Este classificador aleatório classificou corretamente 5 + 2 = 7 instâncias. Efetuando a diferença entre os valores classificados corretamente de ambos os modelos temos 9 – 7 = 2. Isto é, o modelo não aleatório classificou corretamente mais duas instâncias que o aleatório. Kappa terá o valor 2/(total instâncias – acertos aleatórios), Kappa = 2/(14-7)=

28,6%. O valor Kappa permite distinguir modelos que apresentem valores de acurácias iguais.

Curva ROC (Receiver Operating Characteristic) é uma métrica de avaliação de modelos que utiliza recursos gráficos. Deve-se a sua origem à Teoria de Detecção de Sinais utilizada na Segunda Guerra Mundial, sendo depois adaptada para trabalhar com mineração de dados (WITTEN et al., 2011). A curva ROC expressa visualmente a relação (trade-off) entre a taxa verdadeiro positivo (TP rate) e a taxa de falso positivo (FP rate), isto é, a capacidade do modelo em identificar TP à medida que erradamente identifica FP em diferentes porções da base de dados. Dado um modelo, o aumento da taxa de TP ocorre sempre à custa de um aumento da taxa de FP.

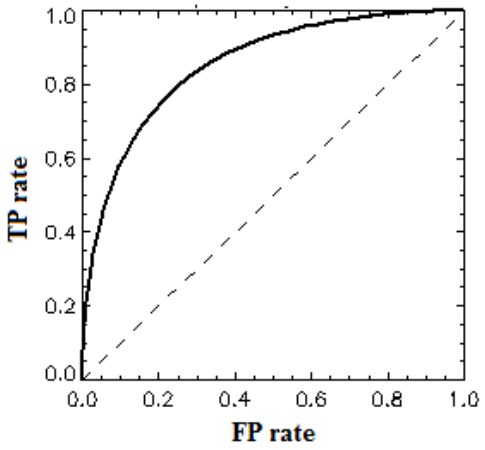


Figura 20: Curva ROC

A linha pontilhada do gráfico da Figura 20 representa um modelo aleatório que servirá como referência. E a curva contínua o modelo em avaliação. Quanto mais a curva representando o modelo se aproximar da linha pontilhada menos precisão terá. A área sob a curva ROC pode variar de 0,5 (precisão mínima) a 1 (precisão máxima).

Em se tratando de variáveis contínuas, a métrica de avaliação mais usada devido à sua simplicidade matemática é a Média Quadrática dos Erros (Mean-Squared Error) que é calculada somando-se as diferenças entre o valor predito e o real, Erro = (predito - real)^2 e dividindo-se o resultado pelo número de casos N. Esta métrica é muito influenciada pelos valores outliers, pois, os erros são elevados ao quadrado.

Dependendo da situação, será mais conveniente usar a Média Absoluta dos Erros (Mean Absolute Error) que leva em conta apenas a magnitude do erro e desprezando o sinal. As fórmulas para estas métricas são mostradas na Figura 21.

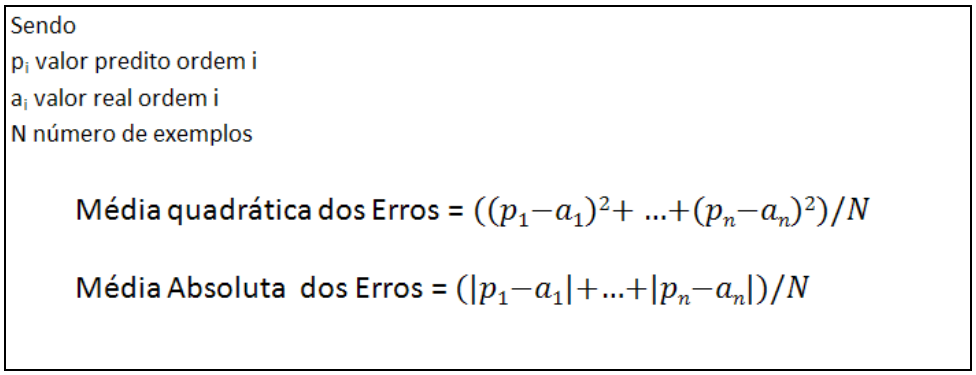


Figura 21: Métricas para valores contínuos

2.2.3 Pós-processamento

Nesta etapa são apresentados e avaliados os padrões extraídos e sumarizados os resultados. Nesta ocasião, os padrões podem ser interpretados. Isto pode ser realizado por meio de documentação e visualização de gráficos. Executa-se, também, a remoção de padrões redundantes. Depois que o conhecimento é extraído poderá ser usado num sistema especialista ou diretamente pelo usuário através de uma interface gráfica.

2.3 Redes Bayesianas (RB)

RB é um modelo que alia Teoria dos Grafos à Teoria da Probabilidade, com forte embasamento teórico, que representa distribuições de probabilidade de modo conciso e utiliza grafos para expressar as dependências entre as variáveis do domínio (BEN-GAL, et al., 2007). Ou seja, RB é um grafo direcionado e acíclico onde os nós representam variáveis do domínio e as arestas representam dependência entre essas variáveis. A cada variável está associada uma tabela de probabilidade conjunta (TPC) que indica o grau de influência das variáveis pais na variável filha.

RBs pertencem à família de Modelos Gráficos Probabilísticos (MGP) (KOLLER, *et al.,* 2009) e ganharam aceitação e notoriedade durante os primórdios da Inteligência Artificial (IA) nos anos 1980. Os estudiosos desta disciplina, baseada nos formalismos lógicos, rejeitavam a abordagem probabilística alegando que o ser humano não manipula números quando raciocina. No entanto, sentiam a necessidade de um método que permitisse a inserção de evidências e fornecesse suporte à tomada de decisão sob incerteza.

Porém, esta postura iria mudar no final dos anos 1980 devido ao influente livro de Judea Pearl (PEARL, 1988) no qual se estabelecem os fundamentos de RB. Outro marco importante foi o artigo de Lauritzen e Spiegelhalter (LAURITZEN et al., 1988) onde são desenvolvidos os fundamentos de como raciocinar e inferir usando modelos gráficos probabilísticos (MGP). Estes trabalhos forneceram a base para construção em larga escala de sistemas especialistas bem sucedidos, como diagnóstico médico e detecção de falhas que ratificaram a eficiência da abordagem proposta.

Em Teoria da Probabilidade os conceitos de probabilidade a priori e probabilidade condicional são a base para o raciocínio probabilístico. Probabilidade a priori expressa o grau de crença em um fato ocorrer sem que seja considerado qualquer outro conhecimento ou evidência. Por exemplo, podemos considerar que a probabilidade a priori de chover amanhã seja **P(C)**. No entanto, se consultarmos a previsão do tempo **T**, a nossa crença se alterará para mais ou para menos. Dizemos então que a probabilidade condicional de chover amanhã, dada a previsão do tempo, é **P(C | T)**. Formalmente a probabilidade condicional é definida por **P(C|T) = P(C∩T)/P(T)**:

O nome Redes Bayesianas foi escolhido em homenagem ao Reverendo Thomas Bayes autor do Teorema de Bayes [9] que relaciona probabilidade a priori e condicional. Por exemplo, se **A** for uma doença e **B** for um sintoma poderemos calcular **P(A|B)** a partir de **P(B|A),** isto é, da probabilidade do sintoma **B** ocorrer sabendo-se que o paciente padece da doença **A**. Este teorema serve de base para realização de inferências sobre RB, apesar de não ser o único recurso utilizado. Na construção de TPCs utiliza-se também a probabilidade frequentista que se baseia na contagem do número de casos. A Figura 22 mostra o Teorema de Bayes, sendo **P(A) e P(B)** as probabilidades a priori de **A** e **B**.

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

Figura 22: Teorema de Bayes

[9] Publicado postumamente por Laplace no livro Théorie Analytique des Probabilités (1812)

A probabilidade de um evento qualquer **a** ocorrer obedece as seguintes propriedades:

- $0 \leq P(\mathbf{a}) \leq 1$
- $P(\mathbf{a}) = 0$ significa que o evento é impossível
- $P(\mathbf{a}) = 1$ significa que o evento é certo

No entanto, existem duas linhas de pensamento a respeito do significado de probabilidade. A interpretação frequentista considera probabilidade como uma frequência de eventos que se repetem. A probabilidade de um evento é expressa como a fração das vezes em que um evento ocorre se o experimento é repetido infinitas vezes. Um exemplo clássico é o jogo de lançar um dado, onde a probabilidade de se obter um número par (2,4 ou 6) tende a 1/2 à medida que repetimos o experimento.

Isto funciona bem quando nos restringimos a eventos tangíveis que possam ser repetidos infinitamente. Porém existem eventos que ocorrem apenas uma vez. Por exemplo, qual a probabilidade de um candidato à presidência conseguir se reeleger?

Uma interpretação alternativa seria considerar probabilidade como um grau subjetivo de crença, uma decisão baseada na experiência pessoal de uma pessoa. Apesar da controvérsia, ambas as interpretações são usadas dependendo do problema sendo tratado.

Variável aleatória é definida como uma variável que a cada valor que possa assumir possui uma probabilidade associada a este valor. Este conceito auxilia o entendimento de distribuição de probabilidade conjunta que será utilizado neste trabalho. Dado um conjunto de variáveis aleatórias $X = \{X_1,X_2,...,X_n\}$, a distribuição de probabilidade conjunta $P(X_1,X_2,...,X_n)$ irá associar valores de probabilidades a cada tupla deste

conjunto. Para efeito de ilustração, considere as variáveis Pessoas com renda acima de R$5.000,00 (**R**) e Pessoas que possuem carro (**C**). Uma distribuição de probabilidade hipotética é mostrada na Tabela abaixo.

Tabela 3: Distribuição de probabilidade de P(R ∧ C)

R=Renda>5.000	C=Possui Carro	P(R ∧ C)
V	F	0,10
V	V	0,20
F	F	0,55
F	V	0,15

$$\sum_{i=1}^{N} \mathbf{P}(\boldsymbol{R} \wedge \boldsymbol{C})_{\boldsymbol{i}} = \mathbf{1}$$

MGP, da qual RB se origina, tem como propriedade principal o poder de representar, de modo conciso, sistemas probabilísticos com muitas variáveis aleatórias. Considere um sistema simples de diagnóstico de Gripe(**G**) ou Resfriado(**R**). O médico leva em consideração para o seu diagnóstico três variáveis: Estação do Ano(**E**), Dor Muscular(**D**) e Congestão(**C**). Para ilustrar este sistema, considere a Figura 23 representando a RB modelada para este domínio.

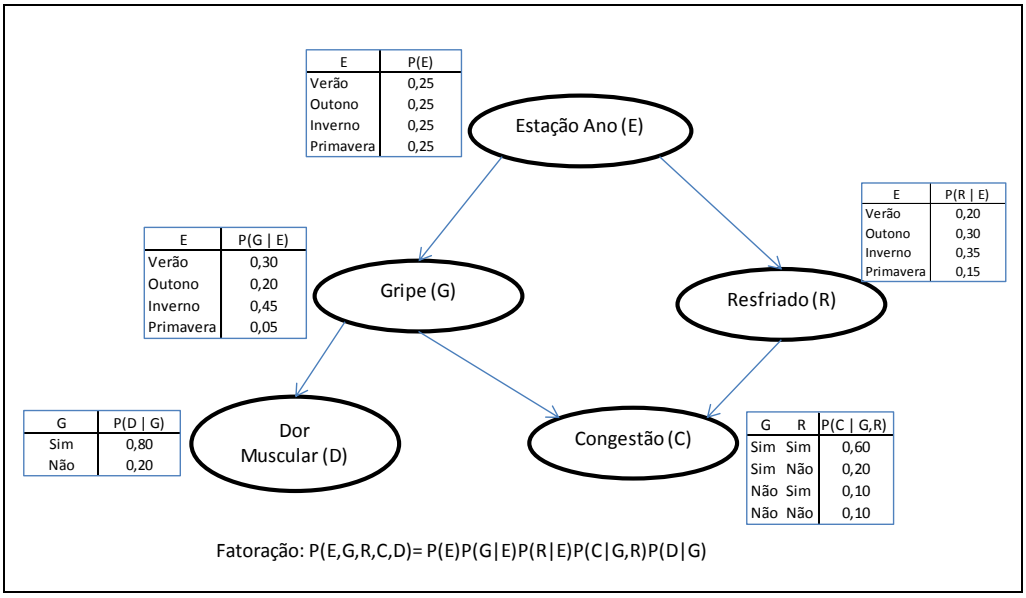


Figura 23: RB Diagnóstico gripe ou resfriado

A RB acima expressa que a estação do ano pode influenciar diretamente o diagnóstico de Gripe ou Resfriado. Tanto Gripe como Resfriado podem ocasionar Congestão, mas apenas Gripe poderá ocasionar Dor Muscular. Considerando que a variável **E** pode assumir os valores {Verão, Outono, Inverno, Primavera} e as demais {Sim, Não}, uma tabela completa da distribuição conjunta de probabilidade deste sistema que cobrisse todas as possibilidades teria 4 x 2 x 2 x 2 x 2 = 64 parâmetros. No entanto, a exploração das independências entre os nós condicionados aos pais, a ser detalhadas mais adiante, resulta na fatoração da distribuição em questão, conforme visto na parte inferior da Figura 23. Sendo assim, será necessário especificar apenas 3 + 4 + 4 + 4 + 2 = 17 parâmetros.

Levando-se em consideração que a maioria dos sistemas probabilísticos utiliza grande número de nós e complexa interação entre eles, como no caso de um diagnóstico médico realístico com dezenas de sintomas e enfermidades, fica evidente que a

representação oferecida por RB tornará o modelo mais conciso e, consequentemente, computacionalmente tratável. Outra questão é que seria enfadonho ou até mesmo impossível para um especialista do domínio preencher manualmente uma tabela de probabilidade conjunta.

A representação em forma de grafo apresenta vantagens como ser amigável ao entendimento humano, podendo ser validada por especialistas do domínio. Modelos que são opacos ao entendimento tendem a levantar dúvidas sobre os resultados apresentados.

Após esta introdução sobre RB poderemos defini-la formalmente como um grafo acíclico dirigido (DAG) com suas respectivas tabelas de probabilidades condicionais. O grafo é dito dirigido porque as arestas que estabelecem as dependências entre os nós possuem direção, e é acíclico porque não existe maneira de começar em um nó e voltar ao mesmo nó seguindo uma sequência de arestas direcionadas, ou seja, não há ciclos. Existem outras abordagens derivada de MGP como Redes de Markov (KOLLER et al., 2009) na qual as arestas não possuem direção, porém estas não serão tratada no presente trabalho.

Em suma, uma RB possui dois componentes principais:

- Um grafo acíclico direcionado representado por nós e arestas, chamado Estrutura ou Grafo da RB (sigla em inglês DAG).
- Um conjunto de distribuições de probabilidade condicionais (uma para cada nó) chamado Parametrização da RB (TPC).

2.3.1 Independências Condicionais entre os Nós de uma RB

O conceito de independência entre os nós de uma RB é de suma importância, pois é na identificação dessas independências representadas na estrutura, que a RB se torna uma abordagem concisa, pois é sabido que uma tabela de probabilidade conjunta cresce exponencialmente em relação ao número de nós (variáveis) nela representados.

Há um tipo de RB chamado Naive Bayes que considera independência total entre as variáveis da base de dados. Apesar da sua simplicidade, apresenta bom desempenho em muitas situações reais. Neste trabalho usaremos Naive Bayes como base de comparação em nossos experimentos.

Voltando a Figura 23, podemos considerar que um quadro de Dores Musculares (**D**) pode influenciar a crença em que estação do ano estamos, mas se é evidente que o paciente está gripado (**G**) então será irrelevante este conhecimento. Obviamente a estação influencia minha crença em **D**, mas esta influência já está expressa em Gripe por meio da sua TPC.

Estes exemplos nos dão uma ideia das independências condicionais expressas em uma estrutura de RB. Para formalizar melhor estas ideias, considere **V** uma variável, **G** um DAG, Pais(V) os pais do nó **V** e:

- Descendentes(**V**) o conjunto de nós descendentes de **V** no DAG **G.**
- Não-Descendentes(**V**) o conjunto de todas as variáveis em **G** que não sejam **V**, Pais(**V**) e Descendentes(**V**).

Então o conjunto **I** de independências condicionais no DAG **G** será de definido como:

I(V, Pais(V),Não-Descendentes(V)), para todas as variáveis V no DAG G

Esta notação estabelece que cada variável **V** no DAG **G** é condicionalmente independente de seus não-descendentes dado seus Pais. A definição acima é conhecida como Suposição de Markov ou Markov(**G**).

Percebe-se que esta definição de independência condicional em um DAG não faz uso de percepções de causa e efeito. Contudo, estas percepções servem para auxiliar o ser humano a construir um DAG, e serão expressas por meio de independências condicionais entre os nós.

Uma RB corresponde a uma distribuição de probabilidade conjunta. Dada uma instanciação qualquer ($\mathbf{X_1}$,$\mathbf{X_2}$,...$\mathbf{X_n}$) é possível calcular a sua probabilidade através da fórmula (Figura 24), sendo P($\mathbf{X_i}$ | $\mathbf{Pa_i}$) a probabilidade da variável **Xi** dado os pais (**Pa**). Se todas as instâncias forem consideradas teremos uma distribuição.

$$P(X_1, X_2, \ldots, X_n) = \prod_{i=1}^{n} P(X_i \mid \boldsymbol{Pa}_i)$$

Figura 24: Fórmula para cálculo da probabilidade da instância (X_1,X_2,...X_n) considerando uma RB

A Suposição de Markov permite identificar as independências mais evidentes, porém existem outros tipos de independências implícitas em um DAG capazes de serem inferidas através de uma técnica chamada *d-separation* (DARWICHE, 2009) descrita a seguir.

Considere X,Y,Z três variáveis aleatórias distintas. Dizemos que X e Y são d-separadas por Z se todos os caminhos que levam de X a Y estão bloqueados por Z,

representado por dsep(X,Z,Y). Para efeito de entendimento vamos considerar este bloqueio como uma válvula que interrompe ou permite o caminho entre duas variáveis.

Há três tipos de válvulas de bloqueio: sequencial, divergente e convergente. A Figura 25 a seguir mostra estas estruturas.

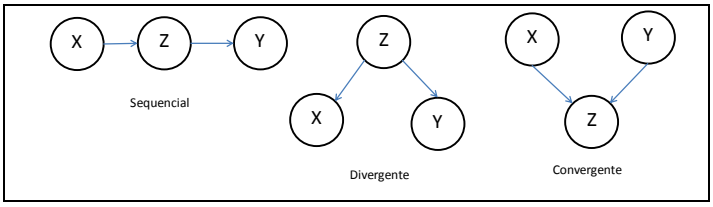


Figura 25: Tipos de d-separação

No tipo d-separação sequencial, (Figura 25, à esquerda) uma vez que sabemos o valor da variável Z o caminho entre X e Y estará bloqueado, isto é, nossa crença em X não mais afetará nossa crença em Y. No tipo Divergente (centro) a evidência da variável Z bloqueará a Dependência entre as variáveis X e Y. E por último, d-separation Convergente (direita) estabelece que não havendo evidências em Z ou qualquer de seus descendentes o caminho entre X e Y estará bloqueado, isto é, a falta de evidência em Z torna os nós X e Y independentes entre si.

2.3.2 Abordagens de Construção de uma RB

Existem três principais abordagens para a construção de uma RB. Na primeira delas um especialista de uma área auxiliado por um analista de representação do conhecimento, constroem uma RB usando seus conhecimentos e intuições de causalidade. Este método insere muita subjetividade e é viável quando se trata de soluções simples. A segunda abordagem ocorre quando o conhecimento do domínio já está representado de alguma

forma. Em (NADKARNI, 2004) é proposta uma metodologia para derivar uma RB de um Mapa Causal ou Mapa Cognitivo que é utilizado para dar suporte à decisão. Já em (ANDEAS et al., 2012) é apresentada uma abordagem para extrair uma RB de uma ontologia.

O terceiro método usado é o do Aprendizado do modelo RB a partir de um conjunto de dados históricos. Tanto a estrutura quanto as TPCs podem ser derivadas de uma base de dados através de um processo de indução utilizado por algoritmos de aprendizado de máquina. Um dos princípios em que se baseia este aprendizado é o conceito de função de avaliação. Máxima verossimilhança (Maximum Likelihood) é um exemplo de função de avaliação da qualidade da RB dada uma base de dados **D**.

Neste trabalho utilizaremos a terceira opção complementada com a segunda. Isto é, aprenderemos uma RB, mas partiremos de um conhecimento inicial expresso em uma ontologia de exploração de petróleo.

2.3.3 Inferências em uma RB

Um sistema geral de inferência e raciocínio proposto por (McCarthy, 1959) é composto por duas partes: Uma base de conhecimentos e uma máquina de inferência conforme mostrado na Figura 26 a seguir.

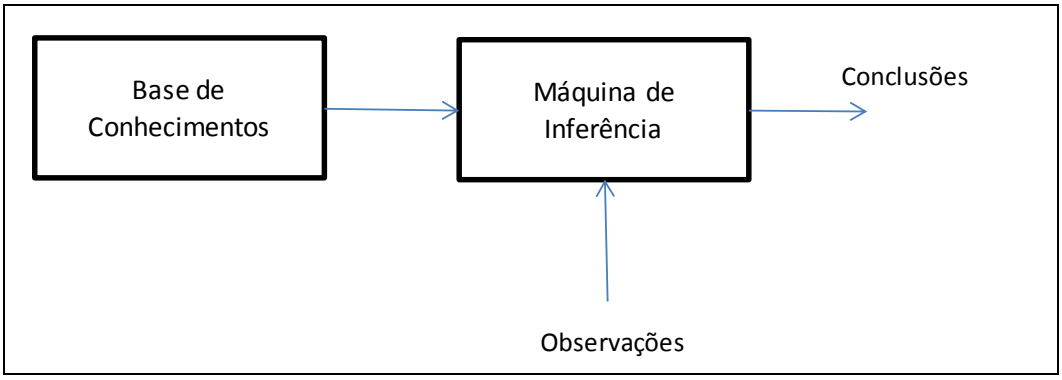


Figura 26: Sistema de raciocínio de McCarthy

Uma RB formada pelo seu grafo e TPCs constitui a representação da base de conhecimentos da Figura 26. Apesar de ser possível, pelo menos teoricamente, calcular manualmente qualquer valor de probabilidade, não expressos explicitamente pelas TPCs usando a Teoria da Probabilidade, esta seria uma tarefa que envolveria grande esforço mesmo para uma pequena rede. Visando tratar esta dificuldade, foi desenvolvido um conjunto de algoritmos que realizam inferências de maneira eficaz e que constituem a máquina de inferência proposta no modelo de MacCarthy. Este conjunto se subdivide em algoritmos exatos e aproximados.

Inferência sobre uma RB significa basicamente inferir a probabilidade de uma ou mais variáveis, dado um conjunto de evidências. Em geral, uma consulta de inferência é constituída por um par (**Q**,**e**), onde **Q** é o conjunto de variáveis que se quer inferir e **e** o conjunto de evidência. Um tipo de consulta simples e muito utilizada é a inferência da probabilidade a priori.

2.3.3.1 Algoritmos de Inferência Exatos

O método mais simples de se calcular inferências exatas em uma RB é chamado eliminação de variáveis (KOLLER et al., 2009). Este método consiste em remover variáveis sucessivamente até que só restem as que participam da consulta. Dada a distribuição da Figura 27, caso quiséssemos saber a probabilidade a priori da variável E ser igual a Verdadeiro, P(E=V), bastaria somar (Summing out) as probabilidades referentes às linhas onde E=V, independente das demais variáveis. Resultando P(E=V) = ¾.

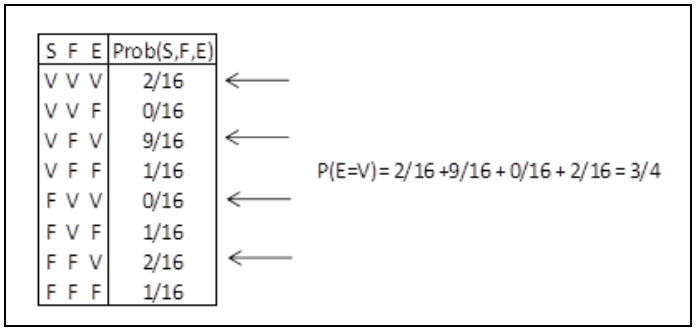


Figura 27: Cálculo de P(E)

Porém este método raramente é viável, pois a distribuição de probabilidade conjunta nem sempre está disponível. Para contornar esta situação, o método eliminação de variáveis utiliza uma estrutura de dados chamada fator que é semelhante a uma TPC.

Inicialmente os fatores correspondem às TPCs. Então, aplicam-se operações de multiplicação e soma entre fatores com o objetivo de eliminar as variáveis que não fazem parte da consulta até se chegar à solução.

A dificuldade de se fazer operações utilizando fatores é que os fatores intermediários de uma operação de eliminação podem alcançar grandes dimensões, dificultando o processamento no tempo e consumindo recursos. Isto pode ser minimizado utilizando-se heurísticas que forneçam uma ordem ótima de eliminação das variáveis que produza os menores fatores intermediários.

Outro método de inferência exata é chamado Inferência por Condicionamento ou Análise por Casos (Case Analysis). Este método se destaca por necessitar de poucos recursos de espaço e tempo. Baseia-se em processar uma consulta, por exemplo P(x), primeiramente considerando um número de casos **C** em que **x** participa. E no final se obtém P(x) pela soma das probabilidades ($x \wedge c_i$). Para isto, é necessário que o conjunto de casos **C** seja exaustivo e mutuamente exclusivo, isto é, cada ($x \wedge c_i$) forma uma partição de x. Assim sendo, expressamos P(x) pela Figura 28:

$$\boldsymbol{P(x) = \sum_{i=1}^{N} \mathbf{P}(x \wedge ci)}$$

Figura 28: Análise por Caso

2.3.3.2 Algoritmos de Inferência Aproximados

Em algumas situações os algoritmos de Inferência Exata, apesar da sua precisão, não são viáveis devido à complexidade do problema proposto ou aos limitados recursos de tempo de execução e de espaço de memória. Dependendo do caso, valerá a pena sacrificar a precisão em troca de uma resposta rápida onde o erro de aproximação é

aceitável (DARWICHE, 2009). Não entraremos em detalhe, pois não os utilizaremos em nossos experimentos.

2.3.4 Aprendizado de uma Estrutura de RB

O objetivo do aprendizado de estrutura é encontrar um DAG que melhor explique os dados. Esta busca pela melhor estrutura é considerada um problema NP-hard e sua complexidade é expressa por 2 elevado a $O(N^2)$, sendo N o número de variáveis. O número de possíveis DAGs contendo N variáveis é superexponencial em **N** (KOLLER et al., 2009).

É necessário, então, efetuar restrições que diminuam esta complexidade. O algoritmo de busca K2 (COOPER, G. et al., 1990), por exemplo, restringe o espaço de busca definindo uma ordem topológica fixa dos nós. Na prática, isto impõe que a variável X_n só poderá ter como pais o conjunto $\{X_1, X_2,... X_{n-1}\}$, isto é, seus predecessores.

Outra restrição é utilizar algoritmos de busca tipo guloso (RUSSEL, S. et al., 2009) que efetuam uma busca aproximada por meio de uma heurística capaz de encontrar um máximo local.

A limitação da quantidade de pais que um nó pode ter também é uma medida restritiva que melhora o desempenho da busca e tem o efeito positivo de evitar o overfitting. Ocorre overfitting quando o modelo aprendido se adequa perfeitamente aos dados de treinamento, mas se comporta com baixo desempenho quando submetido a instâncias inéditas.

2.3.5 Aprendizado dos Parâmetros de uma RB

Para o entendimento do aprendizado dos parâmetros de uma RB será necessário definir o conceito de verossimilhança. Verossimilhança é um tipo de função de avaliação da qualidade da RB expressa pelo produto das probabilidades de se observar cada instância da base de dados na RB. Se considerarmos o Conjunto de TPCs, a base **D** e a instância di = (X_1,X_2,...Xn), então Verossimilhança (**L**) é dada pela fórmula da Figura 29, onde P(di) é calculado aplicando-se a fórmula de cálculo da probabilidade expressa na Figura 24:

$$L(\text{Conj. de TPCs} \mid D) = \prod_{i=1}^{N} P(di)$$

Figura 29: Fórmula da verossimilhança

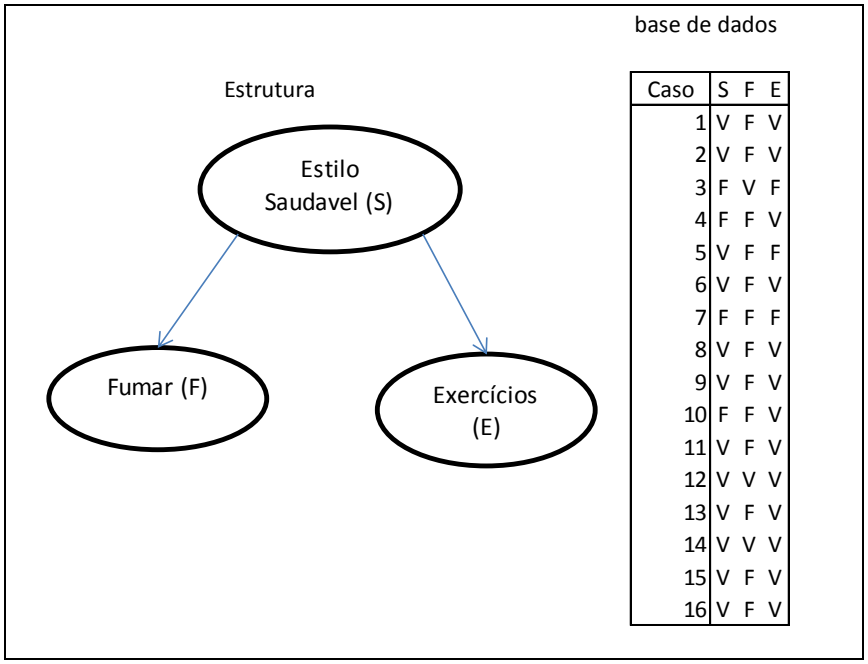
A Verossimilhança por conveniência é representada pelo seu logaritmo (Log-Verossimilhança) devido ao seu baixo valor e a vantagem de se trabalhar com soma ao invés de multiplicação. Resultando em (Figura 30):

$$LL(\text{Conj. de TPCs} \mid D) = \sum_{i=1}^{N} \log P(di)$$

Figura 30: Log-verossimilhança

O aprendizado dos Parâmetros de uma RB (TPCs) pode ocorrer simultaneamente ao da estrutura. Porém por questões didáticas consideraremos o caso em que a estrutura já é conhecida e a partir daí aprenderemos as tabelas de probabilidades condicionais. A Figura 31 representa uma RB com três variáveis aleatórias {EstiloSaudavel (**S**), Fumar

(**F**), Exercícios (**E**)} e os artefatos necessários para a construção das probabilidades condicionais: a estrutura e uma base de dados.



Caso	S	F	E
1	V	F	V
2	V	F	V
3	F	V	F
4	F	F	V
5	V	F	F
6	V	F	V
7	F	F	F
8	V	F	V
9	V	F	V
10	F	F	V
11	V	F	V
12	V	V	V
13	V	F	V
14	V	V	V
15	V	F	V
16	V	F	V

Figura 31: Componente para o aprendizado da RB Estilo Saudável

Com estas informações é possível construir as TPCs por meio da contagem da frequência. Por exemplo, para calcular P(S = V) conta-se na base de dados as instâncias em que S=V (12 instâncias) e divide-se pelo total de instâncias (16), resultando em P(S=V) = ¾. Para calcular P(S=F) faz-se 1- P(S = V) =1/4.

Para calcular P(F=V|S=V), contam-se as instâncias em que F=V (2 instâncias) e divide-se pelo total de instâncias em que S=V (12 instâncias), isto é 2/12 ou 1/6. Procedendo-se desta maneira poderemos calcular os restantes dos parâmetros conforme mostrados na Figura 32:

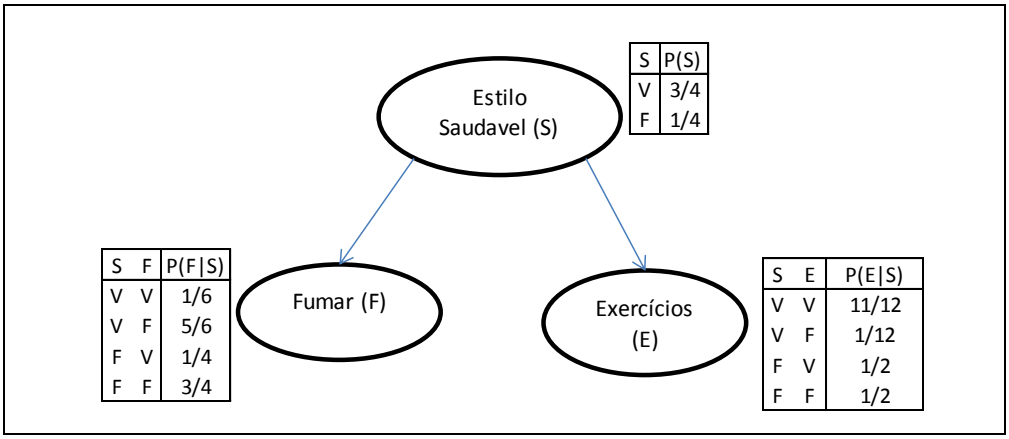


Figura 32: RB Vida Saudável com TPCs

2.4 Ontologia

Ontologia, no campo da ciência da computação, é uma especificação explícita e formal de uma conceitualização compartilhada (GRUBER, T., 1995). Explícita porque seus objetos têm de ser explicitamente definidos. Formal para que não restem dúvidas quando uma máquina resolva interpretar. E compartilhada porque expressa um consenso entre os interessados na ontologia.

Existem várias classes de ontologias (BORST, N., 1997):

- De Fundamentação– não está preocupada com um domínio específico. Procura definir conceitos gerais que regem qualquer domínio. Existem linhas de pesquisa a respeito da ontologia de fundamentação. Uma ontologia de fundamentação específica é denominada UFO (Unified Foundational Ontology) proposta por (GUIZZARDI et al., 2004) e utiliza OntoUML como linguagem.

- De Domínio – Neste caso já existe a preocupação em se estudar certo domínio e extrair relações e objetos inerentes a este.

  Existem exemplos de modelagem de ontologia no domínio petróleo. (CAPPELLI et al., 2007) construíram uma ontologia da atividade reservatórios, derivada de uma modelagem de processos de negócio, composta, na ocasião, por 179 conceitos. Esta ontologia representa os conceitos relevantes na área de reservas e propicia a interoperabilidade semântica, pois uniformiza a nomenclatura. Os autores utilizaram a linguagem OWL, considerada na ferramenta Protege, pois possui um nível de expressividade que não comprometeria o resultado final. Contudo esta ontologia se refere à atividade de produção, o que a difere da nossa proposta que trata exclusivamente da atividade de exploração.

- Tarefas – O domínio já não é o foco e sim o modo de realizar uma tarefa. Pode-se dizer que metodologias são ontologias de tarefas.

- Aplicação – Esta ontologia trata da maneira como se implementa uma determinada solução, dentro de um domínio.

Atualmente as principais linguagens de ontologia são RDF (Resource Description Framework) e OWL (Web Ontology Language) sendo esta última de maior expressividade dentre as linguagens definidas pelo World Wide Web Consortium (W3C), órgão responsável pela padronização.

Uma vez construída a ontologia é possível utilizar um *reasoner* (RUSSEL, S. et al., 2009). *Reasoners* são mecanismos computacionais criados para se realizar inferências

lógicas a partir de um conjunto de fatos e regras. Também chamados de mecanismos de inferência, eles baseiam-se em regras especificadas por uma linguagem de ontologia e uma linguagem descritiva. As ontologias são representações baseadas na lógica, de forma que mecanismos de inferências podem ajudar a construir ontologias, descobrindo inconsistências, dependências ocultas e possíveis redundâncias. Isso permite que novos conhecimentos, além dos contidos explicitamente nas ontologias, sejam agregados.

Sem a utilização de um reasoner seria muito difícil manter grandes ontologias em um estado logicamente correto. Computar e manter múltiplas heranças são o trabalho dos reasoners.

Usaremos a plataforma Protege[10] na construção da ontologia, devido às facilidades que esta oferece, pois é um ambiente gráfico que facilita a inserção de classes, objetos de propriedade, objetos de dados e inclusive de indivíduos. Além disto, é possível utilizar os vários reasoners que podem ser anexados ao Protege em forma de plug-ins como o Fact++ e HermIT que fazem uso da Lógica Descritiva (DL) para realizar inferências. No entanto, quanto mais expressiva for a linguagem maior dificuldade terá o reasoner para fazer inferências. Podendo chegar a não-computabilidade no caso da OWL Full.

Protege é uma ferramenta open-source desenvolvida pela Stanford Medical Informatics. Possui uma comunidade de milhares de usuários. Apesar do desenvolvimento de Protege ter sido historicamente direcionado para aplicações

[10] http://protege.stanford.edu/

médicas, o sistema é independente de domínio e tem sido utilizado em aplicações de outras áreas.

A plataforma Protege suporta duas maneiras de se modelar ontologias:

- Protege-Frames possibilita ao usuário construir e alimentar com instâncias as ontologias baseadas em estruturas, quer dizer, considera uma ontologia como uma estrutura de classes organizadas em hierarquia representando conceitos relevantes do domínio, relacionadas por um conjunto de slots que descrevem suas propriedades e um conjunto de valores específicos de classe chamados instâncias.
- Protege-OWL possibilita construir ontologias para Web Semântica usando o W3C's Web Ontology Language (OWL). Uma ontologia-OWL inclui descrição das classes, propriedades e instâncias.

Capítulo 3 - Proposta

Neste Capítulo iremos descrever como pretendemos desenvolver nossa pesquisa no sentido de encontrar uma solução que minimize os problemas que afetam a avaliação econômica de uma Oportunidade Exploratória de Petróleo com o auxílio da mineração de dados.

3.1 Contatando uma empresa de petróleo

Para iniciarmos o nosso trabalho se fez necessário contatar uma empresa de petróleo de grande porte que se dispusesse a nos ajudar nas pesquisas. Para isto, agendamos reuniões com a equipe da Petrobras envolvida na realização de avaliações econômicas de Oportunidades Exploratórias.

O objetivo destas reuniões era buscar apoio ao nosso projeto e adquirir os conhecimentos necessários à pesquisa, já que se trata de um domínio específico. Desta maneira, seria possível conhecer os problemas que afetam a atividade exploratória e buscar uma solução que minimizasse estas questões. Dentre os problemas relatados pelos especialistas na realização de suas atividades se destacam: demora no tempo de avaliação, uso de técnicas convencionais que podem eventualmente inserir imprecisões no atual processo e adicionar mais incertezas ao processo.

Levantados estes problemas, consideramos que a utilização de técnicas de mineração de dados, aliadas a uma representação formal do domínio fornecida por uma ontologia, poderiam contribuir para a melhoria do processo de avaliação. Sendo assim, julgamos ser possível construir um modelo de mineração de dados que auxiliasse esta atividade a minimizar os problemas relatados, pois os especialistas teriam a sua disposição uma ferramenta que daria suporte às suas tarefas e aceleraria a tomada de decisão. Para testar a nossa hipótese resolvemos realizar um estudo de caso que contou com um histórico de avaliações econômicas.

3.2 Coletando informações sobre avaliação econômica

Identificados os problemas que afetam a tarefa de avaliação econômica, decidimos que a melhor solução seria construir um modelo baseado em mineração de dados que apoiasse esta tarefa. Para isto se faz necessário coletar dados sobre avaliações econômicas. A Petrobras colocou a nossa disposição suas informações sobre avaliações econômicas. Ressalto que por serem estes dados sensíveis, haverá necessidade de se estabelecer sigilo sobre os mesmos, isto é, não torná-los públicos, pois possuem valor estratégico. Representam uma vantagem competitiva em relação a outras empresas. Por exemplo, a taxa mínima de atratividade (TMA) é, em suma, um índice que expressa o quanto uma empresa espera como retorno financeiro dos seus investimentos em uma OE e pode ser considerada um índice de produtividade da empresa.

3.3 Construir a ontologia do domínio exploratório

Consideramos importante realizar a construção de uma ontologia do domínio, pois isto nos ajudaria a entender o processo exploratório como um todo e perceber como a avaliação econômica se ajusta nele. Por ser considerado um domínio complexo, restringiremos a nossa atenção aos principais conceitos e relacionamentos, pois nosso propósito principal é a construção de um modelo que faça a predição da avaliação econômica, o que, por si só, constitui um tema bem abrangente, pois envolve conhecimentos geológicos, econômicos e de DCBD. Procuraremos restringir nosso trabalho ao tempo estabelecido em nosso cronograma.

A ontologia exploratória nos ajudará a identificar as variáveis relevantes referentes à análise econômica de uma OE e manterá uma coerência com o modelo de mineração a ser construído.

3.4 Aplicação do processo de descoberta do conhecimento em banco de dados

Nesta ocasião começaremos a aplicação do processo de DCBD, especificamente o pré-processamento. Utilizaremos os acessos concedidos a nós para obter os dados sobre as avaliações econômicas. Dependendo de como estes dados estão armazenados haverá facilidade na realização desta etapa. Porém, se estiverem espalhados por várias fontes haverá necessidade de uma consolidação mais elaborada.

Para isto, será necessária a construção de ferramentas como macros e consultas a banco de dados. Caso haja instâncias com dados faltantes, serão aplicadas técnicas que insiram valores válidos. A princípio desejamos coletar o maior número possível de

instâncias e de variadas regiões do Brasil, pois isto, respectivamente, facilitará o aprendizado do modelo e o tornará com maior poder de generalização. Construiremos duas bases de dados: uma com a totalidade das variáveis e outra composta apenas das variáveis consideradas mais influentes. Com isto objetivamos testar a redução da dimensionalidade dos dados. Faremos uso do software Office 2010 para nos auxiliar nesta fase, pois este conta com uma planilha eletrônica (MS Excel) e também com um SGBD (MS Access) onde serão armazenados os dados coletados.

3.5 Análise dos dados coletados

Estando os dados consolidados e limpos, já é possível realizar uma análise de como estes se comportam. Construção de histogramas, cálculo das medidas centrais, de extremos (mínimos e máximos) e valores de dispersão (desvio padrão) ajudam a conhecer o comportamento das variáveis e, se detectado algum problema, fazer a correção através da aplicação de filtros de pré-processamento.

Esta análise irá nos ajudar na resolução de eventuais problemas que possam vir a ocorrer na etapa a seguir, de mineração de dados. Um dos recursos a ser utilizado será plotar o gráfico de cada variável em relação a variável de classe com o objetivo de descobrir alguma correlação que se destaque. **Normalização** e **discretização** das variáveis contínuas serão alguns dos procedimentos que aplicaremos em nossos experimentos com RB. E, como a nossa base de dados possui a variável classe, o aprendizado do modelo será do tipo supervisionado.

3.6 Mineração dos dados

Para iniciar esta etapa, já dispomos de uma estrutura inicial de RB derivada da ontologia exploratória que nos fornecerá tanto as variáveis essenciais como uma sugestão de ordem das variáveis já que alguns algoritmos são sensíveis a esta ordem. Também será possível observar se a estrutura inicial faz sentido em termos de noções de causalidade. Caso não faça, esta é a oportunidade de rearranjar alguns nós como forma de inserir conhecimento do especialista.

Os experimentos são realizados aplicando-se diferentes algoritmos de aprendizado de RB com diferentes opções e analisando-se os resultados através das métricas de mineração de dados como acurácia, matriz de confusão, valor Kappa e outros. Para a fase de teste do aprendizado usaremos a validação cruzada 10-fold. Em caso de dificuldades em atingir o desempenho desejado poderemos voltar à etapa de pré-processamento e ajustar a base de dados utilizando-se filtros de pré-processamento. É nosso objetivo também experimentar, em menor escala, outras modalidades de algoritmos de aprendizado como Redes Neurais com o intuito de fazer uma comparação. Esta etapa corresponde à mineração de dados em DCBD.

3.7 Apresentação e análise dos resultados

Realizados os experimentos necessários e já de posse dos resultados, poderemos realizar uma análise através da construção de gráficos, comparação entre os vários modelos preditivos gerados e extraindo conhecimentos por meio de análise. Finalmente,

apresentaremos nosso trabalho aos especialistas para que eles possam discorrer sobre a contribuição que o modelo poderá oferecer à tarefa de avaliação econômica.

3.8 Ontologia para atividade exploratória de petróleo

Após estudo do domínio em questão e com a ajuda dos especialistas demos início à construção da ontologia da atividade exploratória do petróleo usando a plataforma Protege versão 4.1.0 que segue as recomendações para construção de ontologias definidas pelo World Wide Web Consortium (W3C), órgão responsável pela padronização[11].

Tivemos a preocupação de nos certificarmos de que não havia uma ontologia específica sobre o domínio de avaliação econômica disponível. Foram inseridos os principais conceitos e definidas as propriedades de objetos que estabelecem relacionamentos entre estes conceitos. Houve a preocupação de elaborar uma ontologia que fosse comum a todos que atuem na área de exploração de petróleo no Brasil, desconsiderando particularidades inerentes a uma empresa específica.

Após finalizar a ontologia, foi realizada a verificação de consistência usando o Reasoner FaCT++ que a considerou consistente. O objetivo desta ontologia é principalmente identificar os conceitos relevantes e seus respectivos relacionamentos para apoiar a construção da RB. Não sentimos necessidade de expressar todos os detalhes, pois apenas queríamos extrair uma ideia geral deste domínio considerado complexo por alguns autores (CAPPELLI, et al.; 2007). Por isto, consideramos que

expressá-la em OWL-DL, apesar de alguns autores considerarem de pouca expressividade, seria suficiente para nosso propósito. Sendo que, num trabalho futuro, um maior detalhamento poderá ser realizado.

A Figura 33 mostra o diagrama da nossa ontologia de primeiro nível construído usando-se a ferramenta gráfica do aplicativo MS-Excel, mais amigável do que a disponível no Protege.

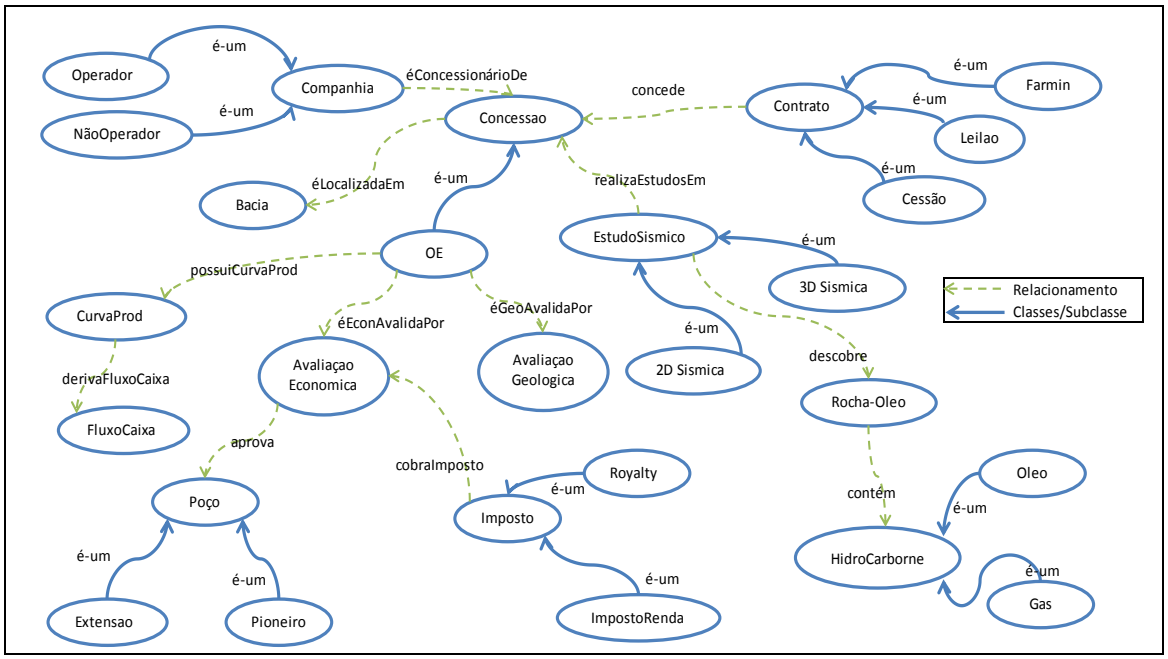


Figura 33: Ontologia da atividade de exploração de petróleo

Baseados na Figura 33, descreveremos o domínio conforme a seguir. Uma **Companhia** para explorar petróleo no Brasil deve primeiramente receber uma **Concessão** do governo outorgada por meio de um **Contrato** de concessão onde ficam estabelecidas as condições como tempo de validade, trabalhos a serem executados,

[11] http://www.w3.org/TR/owl-features/

relatórios sobre o andamento dos projetos e bônus de assinatura (quantia a ser paga pela concessão).

Esta outorga pode ser conquistada por três meios: (i) **Leilão** promovido pela ANP pelo maior lance, (ii) concessão por meio de lei motivada por interesses nacionais (**Cessão**) ou (iii) uma empresa pode comprar uma participação em um bloco exploratório já licitado (**Farm-in**). A concessão outorgada (chamada bloco exploratório) é uma grande área onde a presença de petróleo ainda não foi confirmada e dependerá de **Estudos** para se chegar a uma conclusão. Esta é uma fase de grande incerteza, onde o investimento na aquisição da concessão encontra-se em risco.

A empresa detentora da concessão inicia os estudos na área concedida a fim de localizar acumulações de petróleo. O principal meio de investigação é o **Estudo Sísmico** que consiste em levantamentos sísmicos que podem ser do tipo **2D** ou **3D**. Normalmente inicia-se com levantamento do tipo 2D que são menos detalhados, mas mais abrangentes. Assim que se identifica uma possível acumulação faz-se o levantamento 3D que é mais detalhado e focado em um objetivo de área geográfica menor. Caso ocorra a identificação de estruturas que indiquem uma boa probabilidade da presença de uma rocha que possua hidrocarbonetos (óleo ou gás), esta área é denominada **Oportunidade Exploratória** e iniciam-se estudos mais aprofundados como a construção da C**urva de produção** que vai refletir a provável produção de óleo ano a ano e a seguir é calculado o **Fluxo de Caixa** que explicitará as receitas e despesas necessárias para a realização da produção expressa na curva.

Após coletar todas estas informações é possível realizar a avaliação econômica da OE quando são considerados fatores econômicos como impostos, preço de mercado do hidrocarboneto e taxa mínima de atratividade (TMA). Caso a avaliação econômica tenha um resultado promissor, um colegiado de especialistas avalia as evidências e toma a decisão de perfurar (ou não) um poço exploratório que irá confirmar (ou não) as expectativas geradas. Este poço será denominado **Pioneiro** se for o primeiro de uma área ainda não explorada. Em algumas situações é necessário mais um poço para se fazer uma avaliação segura. Neste caso o poço é chamado poço de **Extensão**.

3.9 RB derivada de uma ontologia

Alguns autores consideram a possibilidade de derivar uma RB de uma ontologia (FENZ et al., 2009) e (DEVITT et al., 2006). A ideia básica, por eles defendida, é aproveitar ontologias disponíveis para extrair informações necessárias à construção de uma RB. Este método, além de facilitar a modelagem da RB, manterá uma coerência entre as duas formas de representação de um mesmo domínio.

A proposta de (DEVITT et al., 2006) exige que a ontologia que servirá de base esteja carregada com informações que servirão para a construção das TPCs. Não sendo este o nosso caso e por dispormos de uma base de dados, optamos para a tarefa de derivação pela abordagem de (FENZ et al., 2009) que destaca os seguintes passos necessários para esta transformação:

1. Construção de uma nova ontologia, baseada na ontologia de exploração de petróleo que, apesar de explicitar os principais conceitos e mostrar onde se

posiciona a avaliação econômica, é muito geral e de alto nível para realização de uma derivação. Esta nova ontologia será construída com a preocupação de relacionar os conceitos intimamente relacionados com a avaliação econômica como qualidade do óleo e respectivo volume. Chamaremos esta nova ontologia de ontologia focada.

2. Conceitos → Nós: Os conceitos contidos na ontologia, relevantes para o problema considerado, devem ser representados como nós na RB.
3. Relacionamentos → Arestas: Os objetos de relacionamentos entre os conceitos serão transformados em dependências (arestas) entre os nós da Rede.
4. Axiomas → Conjunto de Valores: Os Axiomas definirão os valores que os nós da rede poderão assumir, ou seja, os estados possíveis dos nós.
5. Instâncias → Evidências: Instâncias de conceitos serão usadas pra derivar evidências.

Faremos uso da primeira, segunda e terceira diretivas. A quarta será substituída pelas informações contidas na base de dados que será usado no aprendizado. Não utilizaremos a quinta diretiva, pois nossa ontologia não está carregada com uma quantidade suficiente de instâncias.

3.10 Definindo a Modalidade de Mineração de Dados

Antes de iniciarmos a etapa de mineração de dados, tivemos de definir que modalidade de mineração de dados seria mais adequada ao domínio em mãos. Segundo Newendorp (2009, p.1) "Exploração de petróleo é um clássico exemplo de decisão sob incerteza". Outro domínio com mesmas características é diagnóstico médico, sendo RB considerada, nesta área, como uma das primeiras técnicas ser aplicada e obter bons resultados (KOLLER et al., 2009). Um dos primeiros sistemas baseado em RB, no domínio de doenças abdominais agudas, que superou os especialistas é tratado em (DE DOMBAL, T. et al., 1974).

A lógica tradicional trabalha com implicações tipo A → B (sempre que A ocorre, B também ocorre). Em um ambiente de incerteza raramente isto é verdadeiro. Nesta situação é mais certo considerar que sempre que A ocorre, B ocorrerá com uma probabilidade que varia de zero a um [0;1]. Isto é, a lógica probabilística é mais adequada quando se lida com incerteza. RB modela essa incerteza e possibilita a realização de inferências probabilísticas através de algoritmos específicos. (RUSSEL, S. et al., 2009) afirmam que "A presença de incerteza muda radicalmente a maneira que um agente toma decisões". E a utilização de grafos torna RB amigável ao entendimento, facilitando a inserção de conhecimento do domínio baseada na noção de causalidade.

Por todas estas razões, consideramos que RB é a modalidade mais adequada. Ademais, a utilização de Rede Bayesiana apresentou-se como um desafio, já que outras modalidades, segundo a literatura atual, já são bem exploradas em pesquisas como é o caso de Rede Neural, utilizadas em vários trabalhos conforme descrito no Capítulo 5.

Capítulo 4 - Estudo de Caso

Neste estudo de caso mostraremos a viabilidade de se aprender um modelo bayesiano, com apoio de uma ontologia, que dê suporte à avaliação econômica de uma OE. Este modelo será mais eficiente em termos de tempo e permitirá inferências não oferecidas pelos métodos tradicionais.

Para iniciar o estudo de caso, dispomos de uma ontologia do domínio, conforme descrita no Capítulo 3, que será usada para extrair informações importantes para construção de uma estrutura básica de RB que será evoluída e terá suas TPCs aprendidas através da utilização de algoritmos de aprendizado de RB até que seja considerada eficiente pelas métricas de mineração de dados. Foi também realizada uma avaliação pelos especialistas que a avaliaram sob o aspecto da sua contribuição ao trabalho de avaliação econômica.

4.1 Estrutura inicial de RB derivada de uma ontologia

Inicialmente iremos construir uma nova ontologia baseada na ontologia de exploração de petróleo conforme descrito na Seção 3.9, porém focada na tarefa de avaliação econômica, conforme sugere a diretiva 1 (FENZ et al, 2009). O resultado é mostrado na Figura 34.

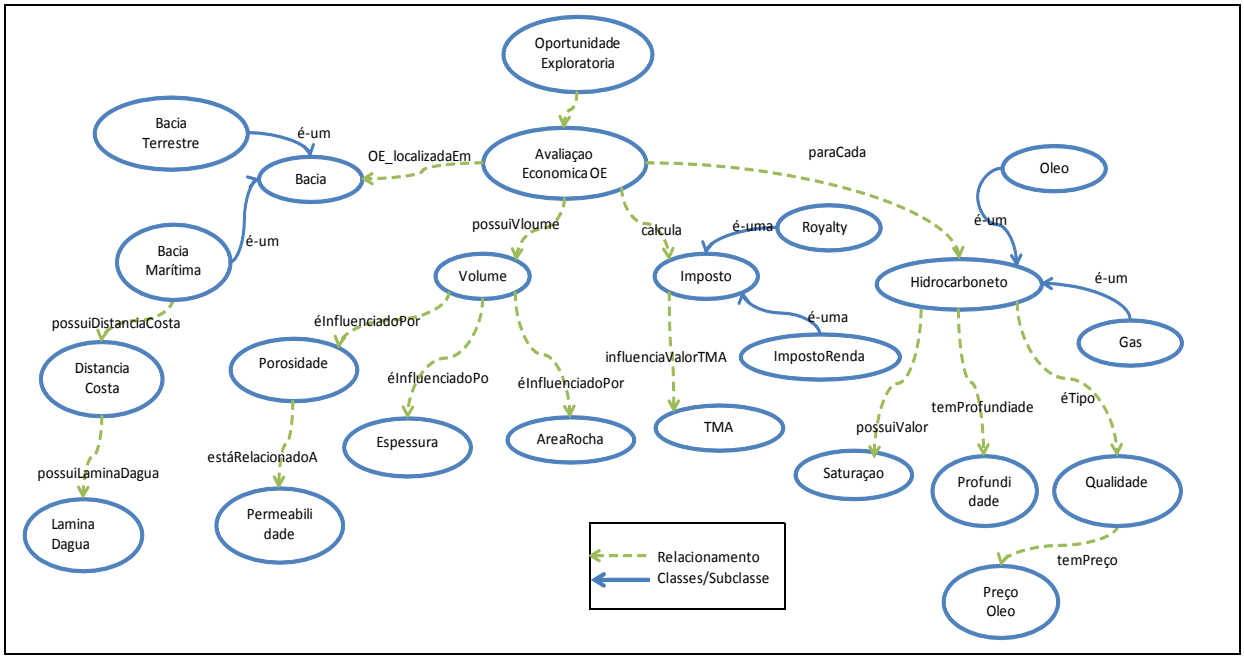


Figura 34: Ontologia focada na avaliação econômica

A Ontologia acima foca em parte do domínio responsável pela avaliação econômica de uma OE e é descrita a seguir.

A avaliação da OE é diretamente influenciada pela **Bacia Sedimentar**, que pode ser terrestre ou marítima, pelo **Volume** do óleo que se espera encontrar e pelo tipo do **hidrocarboneto** (óleo ou gás). A bacia marítima está localizada a uma **Distância da costa** e esta possui uma **Lâmina d'água**. O **Volume** é fortemente influenciado pelos valores de **Porosidade**, **Espessura** e **Área**, sendo que estas características se referem à rocha acumuladora. Os **Impostos** são considerados nos cálculos da avaliação e a **TMA** é definida de forma que seja possível pagar os impostos e obter retorno financeiro. O **Hidrocarboneto** (que poder ser óleo ou gás) está localizado a uma **Profundidade** e possui uma **Qualidade** que define o seu preço de mercado.

E seguindo as diretrizes 2 e 3, derivamos uma estrutura de RB inicial (Figura 35), que poderá ser melhorada com aplicação dos algoritmos de aprendizado como inclusão de arestas para aumentar a sua performance. Estando a estrutura construída, aprenderemos as TPCs.

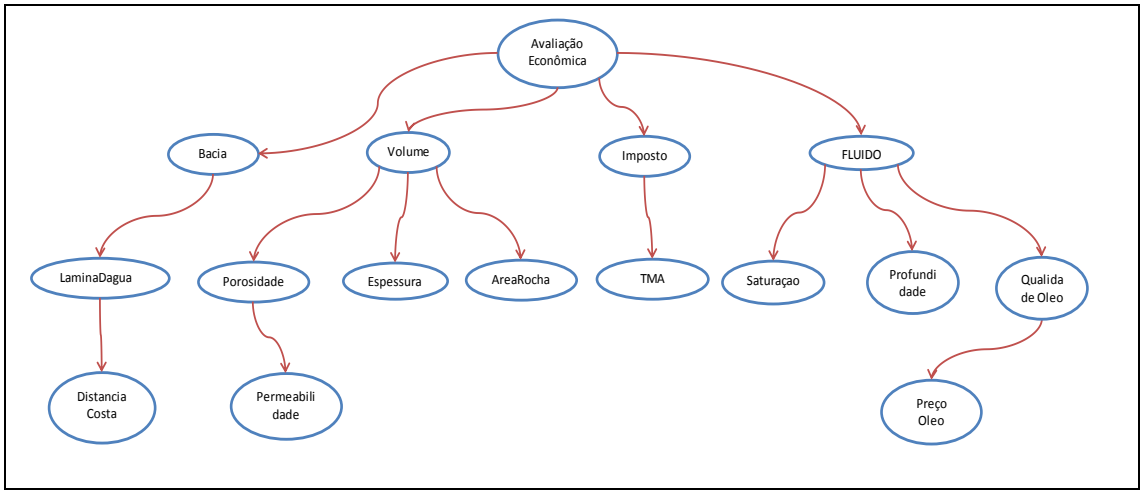


Figura 35: Estrutura de RB derivada de ontologia

Observando a ontologia focada na avaliação econômica, temos uma sugestão de uma possível ordem em que as variáveis essenciais deverão ser introduzidas nos algoritmos de busca pela melhor estrutura, pois nota-se que algumas variáveis são obtidas em primeiro lugar em relação às demais. Por exemplo, a bacia e os impostos são de conhecimento dos especialistas mesmo antes de começarem os cálculos propriamente ditos, e influenciam as demais variáveis. Isto é importante, pois esta ordem influencia o desempenho do algoritmo que iremos aplicar no aperfeiçoamento desta estrutura. No nosso caso, como temos 15 variáveis mais a classe, seria impossível testar todos os 15! arranjos possíveis. A ordem sugerida é apresentada na Tabela 4, a seguir.

Tabela 4: Ordem das variáveis essenciais

Ordem	Variável	Tipo	Descrição
1	Bacia	nominal	Bacia Sedimentar onde se localiza a OE
2	Volume	contínuo	Volume estimado de óleo (m3)
3	Impostos	contínuo	Impostos sobre a produção de petróleo (%)
4	Fluido	nominal	Tipo de Fluido (óleo ou gás)
5	LDA	contínuo	Profundidade do oceano (m)
6	Porosidade	contínuo	Porosidade da rocha (%). A percentagem de volume de poros que pode conter fluidos.
7	Espessura	contínuo	Espessura da rocha (m). Medida da espessura vertical de uma camada de rocha sedimentar
8	Area	contínuo	Área da acumulação de petróleo (Km2)
9	TMA	contínuo	Taxa mínima de atratividade (returno esperado)
10	Saturação	contínuo	Saturação da rocha (%). Quantidade relativa de água, petróleo e gás nos poros de uma rocha, expressa como percentagem do volume.
11	Profundidade	contínuo	Profundidade da OE (m). Profundidade em que se encontra a acumulação de petróleo
12	Qualidade_Oleo	contínuo	Medida de qualidade do óleo (API). Medida relativa a densidade do petróleo
13	Distancia	contínuo	Distância da OE até a costa (m)
14	Permeabilidade	contínuo	Permeabilidade da rocha (mD). Medida da capacidade de uma rocha para transmitir fluidos
15	Preço_oleo	contínuo	Preço do óleo no mercado (USD)
16	VPL	contínuo	Valor Presente Líquido

Após a construção da ontologia focada foi possível identificar as variáveis relevantes (Tabela 4), a ordem em que são adquiridas e definir uma estrutura básica de RB a ser evoluída no processo. Dados históricos a respeito dessas variáveis foram coletados totalizando uma base de dados com 700 instâncias de avaliações de OE, sendo cada uma composta de 15 variáveis mais a variável de classe. Estas informações foram fornecidas pela Petrobras e têm caráter privado. Como esta é uma empresa líder neste mercado e com atuação em todo território nacional, consideramos as informações bem representativas e abrangentes do domínio em questão. De certo, cada empresa tem suas peculiaridades, mas nossa pesquisa se preocupou em evitar vieses e procurou se basear nos procedimentos comuns a todos os envolvidos na atividade.

Para a execução dos experimentos foram utilizadas as ferramentas Weka[12], que possui filtros e algoritmos de mineração, Matlab[13] para geração de gráficos e a SamIam[14] para complementar os recursos da Weka. As etapas de DCBD estão descritas em detalhes a seguir.

4.2. Pré-processamento

A análise dos dados iniciou por identificar na base de dados as instâncias outliers, isto é, os pontos fora da curva. Aplicando um algoritmo baseado nos valores dos quartis conforme descrito na Seção 2.2.1.3, foi possível identificar e extrair quatro instâncias que se mostraram fora dos limites aceitáveis.

Em seguida, aplicamos a normalização dos dados contínuos para valores entre zero e um [0;1] por dois motivos: (i) impedir que variáveis com grande intervalo numérico dominem as de pequeno durante o aprendizado e (ii) ocultar os verdadeiros valores, compromisso assumido junto à empresa detentora da informação. Apenas duas variáveis eram do tipo nominal: **Bacia Sedimentar** e **Fluido**.

O próximo passo foi desenhar os histogramas das variáveis e observar como os valores estavam distribuídos, identificar valores extremos (máximo e mínimo) e a dispersão. Percebeu-se, também, que a variável classe VPL possui o comportamento tipo cauda longa, isto é, possui um grau acentuado de desvio ou afastamento do eixo de simetria de sua distribuição, conforme mostra o respectivo histograma da Figura 36.

[12] http://weka.wikispaces.com/

[13] http://www.mathworks.com/products/matlab/

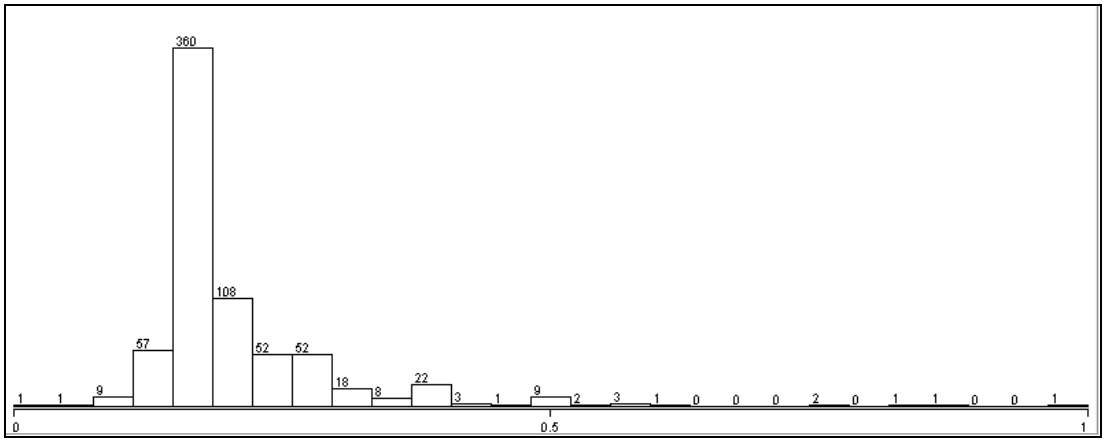


Figura 36: Histograma da classe VPL

Isto foi confirmado ao traçarmos o gráfico Kernel Density Estimates (JANERT, 2010) que é uma técnica similar ao histograma, mas que não sofre influência do tamanho e da localização dos intervalos (bins) necessários para à construção dos histogramas, dependendo apenas da função kernel escolhida.

Comparando o gráfico Kernel Density (Figura 37) com o histograma do VPL ficou evidenciado o comportamento cauda longa da classe VPL e deduzimos ser esta uma característica do problema. Resolvemos, então, aplicar a função logarítmica aos valores de VPL (Figura 38) para atenuar os problemas que este comportamento poderia trazer à etapa de mineração de dados.

[14] http://reasoning.cs.ucla.edu/samiam/

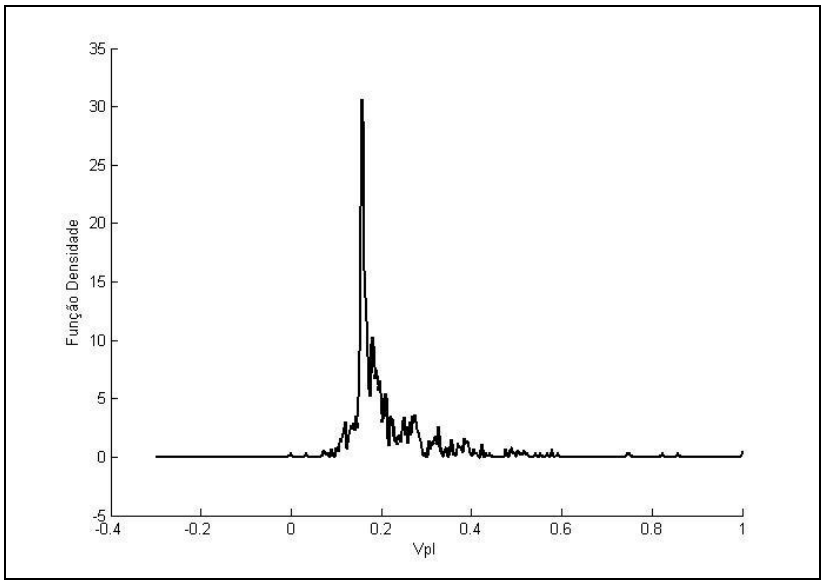


Figura 37: Gráfico Kernel Density Estimate (KDE) do VPL

A transformação logarítmica aplicada à classe VPL resultou no histograma abaixo:

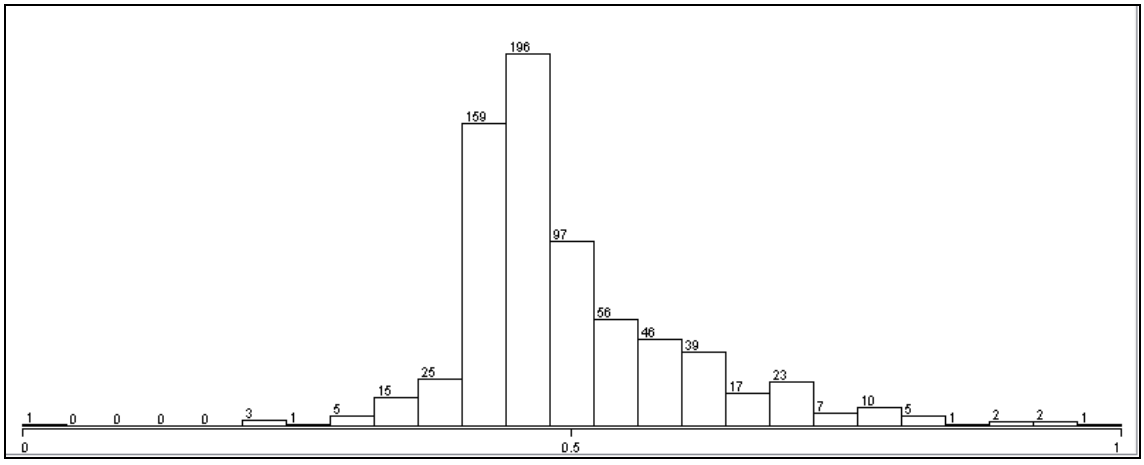


Figura 38: Histograma após transformação logarítmica da classe VPL

Em seguida discretizamos os atributos numéricos. Existem dois principais tipos de Discretização: (i) Intervalos de igual largura e (ii) Intervalos de igual frequência. Foram realizados experimentos preliminares utilizando o algoritmo NaiveBayes (JOHN, G. et al., 1995) para descobrir que método de discretização melhor se adequa aos dados que dispúnhamos. A opção (i), no nosso caso, apresentou melhor desempenho em testes preliminares. O resultado da discretização após normalização é mostrado na Tabela 5.

Tabela 5: Quadro de discretização das variáveis após normalização

Variável	#Bins	Bins	Variável	#Bins	Bins
volume	10	(0-0.1],(0.1-0.2],...,(0.8-0.9](0.9-1]	Qualidade_Oleo	10	(0-0.1],(0.1-0.2],...,(0.8-0.9](0.9-1]
LDA	10	(0-0.1],(0.1-0.2],...,(0.8-0.9](0.9-1]	Saturação	10	(0-0.1],(0.1-0.2],...,(0.8-0.9](0.9-1]
Profundidade	10	(0-0.1],(0.1-0.2],...,(0.8-0.9](0.9-1]	distancia	10	(0-0.1],(0.1-0.2],...,(0.8-0.9](0.9-1]
area	10	(0-0.1],(0.1-0.2],...,(0.8-0.9](0.9-1]	Impostos	10	(0-0.1],(0.1-0.2],...,(0.8-0.9](0.9-1]
Espessura	10	(0-0.1],(0.1-0.2],...,(0.8-0.9](0.9-1]	TMA	10	(0-0.1],(0.1-0.2],...,(0.8-0.9](0.9-1]
Porosidade	10	(0-0.1],(0.1-0.2],...,(0.8-0.9](0.9-1]	Preço_oleo	10	(0-0.1],(0.1-0.2],...,(0.8-0.9](0.9-1]
Permeabilidade	10	(0-0.1],(0.1-0.2],...,(0.8-0.9](0.9-1]	VPL	10	(0-0.1],(0.1-0.2],...,(0.8-0.9](0.9-1]

A decisão de discretizar os dados numéricos foi motivada pelos seguintes fatores. As plataformas que dispúnhamos para nossos experimentos exigiam a discretização dos dados para aplicação dos algoritmos de aprendizado bayesiano. Para provar a nossa hipótese consideramos, a princípio, que esta restrição não comprometeria nossos resultados, pois apesar da perda de um pouco de precisão seriamos recompensados com uma gama de opções que facilitariam o atingimento dos nossos objetivos.

A construção de algoritmos que aceitem dados contínuos exige a determinação do comportamento de cada variável, isto é, a definição de um tipo de distribuição de probabilidade para cada variável da base, seja ela uma distribuição normal, log-normal ou triangular. O complemento BNT[15] da plataforma Matlab só dispõe de recursos para processar variáveis que apresentem comportamento gaussiano, não sendo este o nosso caso.

Ao término da etapa de pré-processamento, a base de dados estava preparada para ser submetida aos algoritmos de mineração de dados propriamente dito. A Figura 39 mostra o diagrama em bloco da etapa de pré-processamento.

[15] http://bnt.googlecode.com/svn/trunk/docs/usage.html#basics

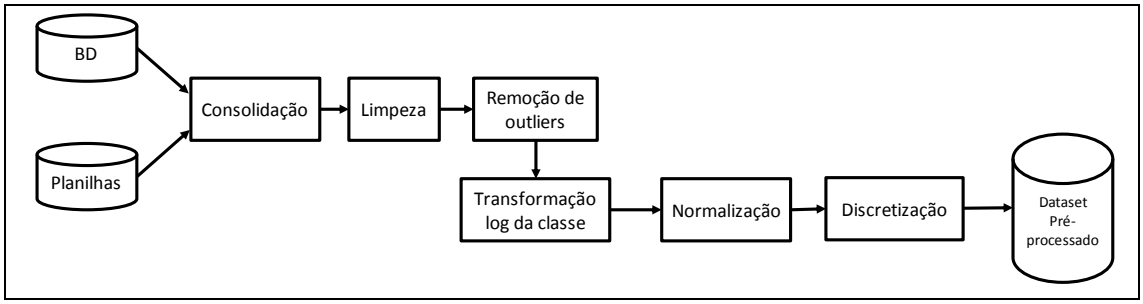


Figura 39: Etapa de pré-processamento realizada durante o estudo de caso

4.3 Mineração de dados

Neste ponto dos experimentos contamos com uma estrutura de RB básica a ser melhorada e a base de dados de avaliações pré-processada. Iniciamos, então, a etapa de mineração de dados. Consideramos como algoritmo de aprendizado de RB o K2 local nos experimentos preliminares. Utilizamos validação cruzada 10-fold nas fases de testes de aprendizado, inclusive nos experimentos mais avançados. Além disso, restringimos que cada nó da rede tivesse no máximo dois pais. O modelo preliminar aprendido apresentou uma acurácia de 72,5%.

Uma questão que primeiramente pretendíamos responder era se a ordem das variáveis deduzida pela ontologia havia acrescentado algum ganho de acurácia em relação a uma ordem puramente aleatória. Geramos 10 ordens aleatórias e comparamos os resultados com a ordem deduzida e observamos um ganho médio de 1,2 ponto percentual em favor da ordem deduzida (Tabela 6). Apesar de não representar muito, serviu para corroborar a ideia de que uma ordem baseada na sequência em que as informações são obtidas pode melhorar o desempenho do aprendizado.

Tabela 6: Comparação entre ordem deduzida e aleatória

Ordem	Acurácia	Média (%)
Deduzida	72,50	
aleatória1	71,10	
aleatória2	71,52	
aleatória3	70,50	
aleatória4	72,30	
aleatória5	72,00	71,34
aleatória6	71,41	
aleatória7	72,33	
aleatória8	71,20	
aleatória9	68,50	
aleatória10	71,34	

Prosseguindo com os experimentos e analisando a matriz de confusão da Figura 40, percebe-se que classes menos representadas apresentam baixas taxas de Verdadeiro Positivo (TP rate).

===Matriz Confusão (teste com validação cruzada 10-fold===

a	b	c	d	e	f	g	h	i	j	<-classificada como	Taxa TP
0	6	1	1	0	0	0	0	0	0	a = [0 - 0,1)	0,00
0	435	39	7	0	0	0	0	0	0	b = [0,1 - 0,2)	0,90
0	72	66	11	1	2	0	0	0	0	c = [0,2 - 0,3)	0,43
0	13	16	13	1	0	0	0	0	0	d = [0,3 - 0,4)	0,30
0	2	3	6	2	1	0	0	0	0	e = [0,4 - 0,5)	0,14
0	1	2	4	2	0	0	0	0	0	f = [0,5 - 0,6)	0,00
0	0	0	0	0	0	0	0	0	0	g = [0,6 - 0,7)	0,00
0	0	0	2	0	0	0	0	0	0	h = [0,7 - 0,8)	0,00
0	0	2	0	0	0	0	0	0	0	i = [0,8 - 0,9)	0,00
0	0	0	1	0	0	0	0	0	0	j = [0,9 - 1]	0,00

Figura 40: Matriz de confusão inicial

Plotando-se o gráfico TP rate x Nº exemplos por classe (Figura 41), percebe-se uma correlação entre essas medidas: quanto menor a representatividade da classe na base de dados, menor a TP rate. Isto ressalta a dificuldade do algoritmo de aprendizado de RB em modelar estas classes minoritárias. Pela Figura 40 percebe-se que classes minoritárias como **a,f,g,h** não foram aprendidas satisfatoriamente.

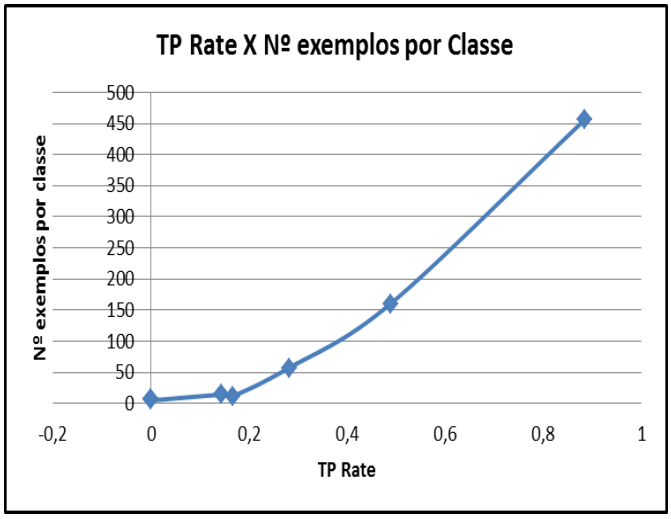


Figura 41: Gráfico TP Rate X Nº exemplos por classe

Para resolver este problema, aplicamos o algoritmo SMOTE (CHAWLA et al., 2002) que tem a finalidade de fazer um Resampling aplicando a técnica Synthetic Minority Oversampling. Esta técnica gera instâncias sintéticas das classes menos representadas, com poucas instâncias (instâncias minoritárias) baseadas nos **k** vizinhos mais próximos conforme a seguir:

- Para cada atributo **a** da instância minoritária
    - Computa os **k** vizinhos mais próximos de **a**
    - Escolhe aleatoriamente **v** entre os **k** vizinhos
    - Calcula a distância **d** entre **a** e **v**
    - Sintético **a** = **a** + **d** * (número aleatório entre 0 e 1)

Após aplicação do SMOTE houve um ganho em média de 0,51 na TP rate. Houve melhora também na acurácia que passou de 72% para 83%. A Figura 42 apresenta a matriz de confusão resultante.

===Matriz Confusão (teste com validação cruzada 10-fold===

a	b	c	d	e	f	g	h	i	j	<-classificada como	Taxa TP
8	4	1	0	0	0	0	0	0	0	a = [0 - 0,1)	0,62
1	468	21	6	0	0	0	0	2	0	b = [0,1 - 0,2)	0,94
0	47	75	12	0	6	0	0	3	0	c = [0,2 - 0,3)	0,52
0	10	10	27	3	2	0	0	4	0	d = [0,3 - 0,4)	0,48
0	1	0	6	11	2	0	0	0	0	e = [0,4 - 0,5)	0,55
0	0	3	1	0	14	0	0	1	0	f = [0,5 - 0,6)	0,74
0	0	0	0	0	0	0	0	0	0	g = [0,6 - 0,7)	0,00
0	0	0	0	0	0	0	16	0	0	h = [0,7 - 0,8)	1,00
0	0	0	0	0	0	0	0	12	0	i = [0,8 - 0,9)	1,00
0	0	0	1	0	0	0	0	0	6	j = [0,9 - 1]	1,00

Figura 42: Matriz de confusão após resampling com SMOTE

Com o objetivo de encontrar um modelo mais preciso, experimentamos outros algoritmos de aprendizado de RB. O algoritmo Hill Climber (RUSSEL, S. et al., 2009), ao contrário do K2, não tem restrição de ordem e efetua adição, deleção e reversão de arcos. A busca Tabu (GLOVER, F. et al., 1985) é similar a Hill Climber, mas estende um pouco a busca mesmo quando encontra um suposto ponto ótimo.

Além disto, testamos algumas opções de processamento que consideramos de grande influência em aprendizado de redes bayesianas:

(i) Escopo da métrica de avaliação: escopo **Local** avalia cada nó da rede individualmente e assume que a rede ótima é dada pela soma das avaliações de cada nó da rede. Escopo **Global** avalia a rede considerando métricas globais como acurácia, não importando o desempenho de cada nó individualmente.

(ii) Número máximo de pais para os nós (2 ou 3). Além destes valores a estrutura da rede fica muito complexa para uma análise, correndo-se o risco de haver overfitting.

Outro ponto considerado nos experimentos foi testar a redução da dimensionalidade da base de dados. Esta Redução tem por objetivo verificar se todo o conjunto das 16 variáveis é necessário para o aprendizado do modelo ou apenas um subconjunto seria suficiente.

Para isso, utilizamos o algoritmo genético simples (SGA) descrito por (GOLDBERG, 2006) com uma população de 20 indivíduos, número de gerações 20, crossover (prob=0,6) e mutação (prob=0,033) para selecionar as variáveis mais influentes. Como resultado, obtivemos que as variáveis relevantes são: Volume, LDA, Profundidade, Área da Acumulação, Espessura, Porosidade, TMA, Preço do Óleo. Acrescentamos as variáveis: Imposto e Qualidade do óleo, pois os especialistas as consideraram importantes. Nosso subconjunto ficou, assim, constituído por dez variáveis mais a classe. Os resultados obtidos com a base de dados completa e reduzida estão na Figura 43 a seguir, onde cada modelo RB foi construído com diferentes opções de algoritmo de busca, escopo da métrica e número máximo de pais por nó conforme mostra a Tabela 7.

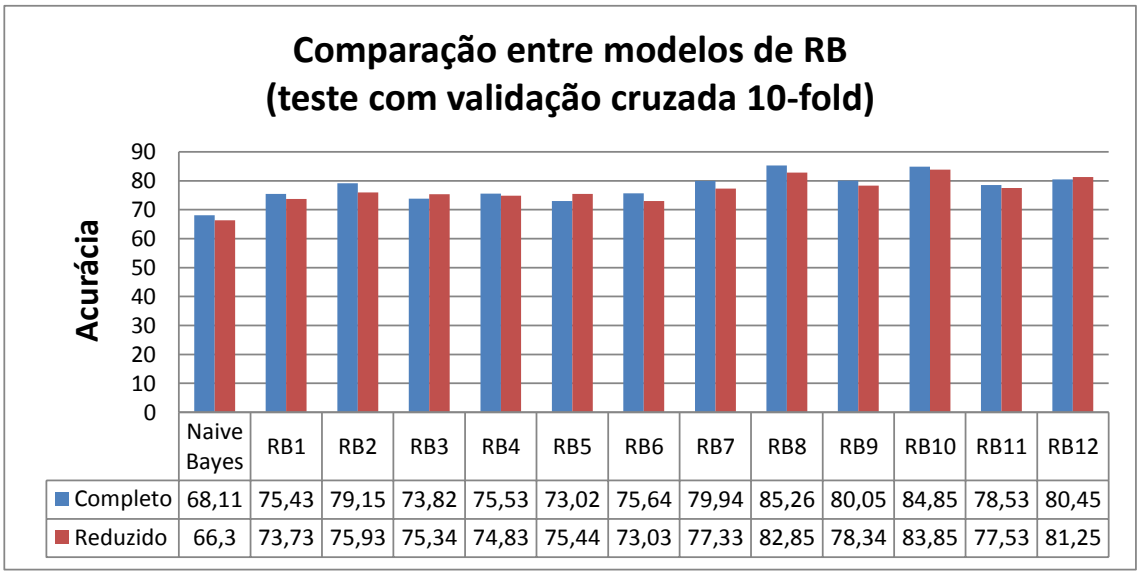


Figura 43: Comparação de desempenho dos modelos RB

Tabela 7: Opções utilizadas por cada Modelo Bayesiano (RB#)

Modelo	Escopo	Algoritmo	# pais
RB1	local	K2	2
RB2	local	K2	3
RB3	local	HillClimber	2
RB4	local	HillClimber	3
RB5	local	Tabu	2
RB6	local	Tabu	3
RB7	global	K2	2
RB8	global	K2	3
RB9	global	HillClimber	2
RB10	global	HillClimber	3
RB11	global	Tabu	2
RB12	global	Tabu	3

Percebe-se que todas as propostas foram melhores do que a base de comparação. Não houve diferença significativa entre os modelos aprendidos com a base completa e a reduzida, sendo o modelo RB8 o de melhor desempenho. Os valores das demais

métricas de avaliação de mineração de dados e a matriz de confusão final são mostrados na Figura 44. A curva ROC de um dos valores de classe é mostrada na Figura 45.

===Avaliação Modelo RB8===

Estatística Kappa 0,8175

Métrica por classe

taxa TP	taxa FP	Precisio	Recall	F-Measure	área ROC	classe
1,00	0,00	1,00	1,00	1,00	1,00	[0 - 0,1)
1,00	0,00	0,99	1,00	1,00	1,00	[0,1 - 0,2)
0,89	0,00	0,92	0,89	0,91	1,00	[0,2 - 0,3)
0,72	0,01	0,80	0,72	0,76	0,98	[0,3 - 0,4)
0,88	0,06	0,89	0,88	0,88	0,97	[0,4 - 0,5)
0,85	0,06	0,75	0,85	0,80	0,97	[0,5 - 0,6)
0,64	0,02	0,73	0,64	0,69	0,96	[0,6 - 0,7)
0,69	0,01	0,86	0,69	0,77	0,96	[0,7 - 0,8)
0,82	0,01	0,90	0,82	0,85	0,97	[0,8 - 0,9)
0,96	0,02	0,77	0,96	0,86	0,99	[0,9 - 1]

===Matriz Confusão===

a	b	c	d	e	f	g	h	i	j	**<-- classificada como**
34	0	0	0	0	0	0	0	0	0	a = [0 - 0,1)
0	108	0	0	0	0	0	0	0	0	b = [0,1 - 0,2)
0	0	24	3	0	0	0	0	0	0	c = [0,2 - 0,3)
0	0	2	48	12	4	1	0	0	0	d = [0,3 - 0,4)
0	1	0	7	317	22	10	2	0	1	e = [0,4 - 0,5)
0	0	0	1	18	139	5	0	0	0	f = [0,5 - 0,6)
0	0	0	0	7	12	58	3	3	7	g = [0,6 - 0,7)
0	0	0	1	2	2	4	31	1	4	h = [0,7 - 0,8)
0	0	0	0	0	7	1	0	44	2	i = [0,8 - 0,9)
0	0	0	0	1	0	0	0	1	47	j = [0,9 - 1]

Figura 44: Medidas de desempenho do modelo RB8

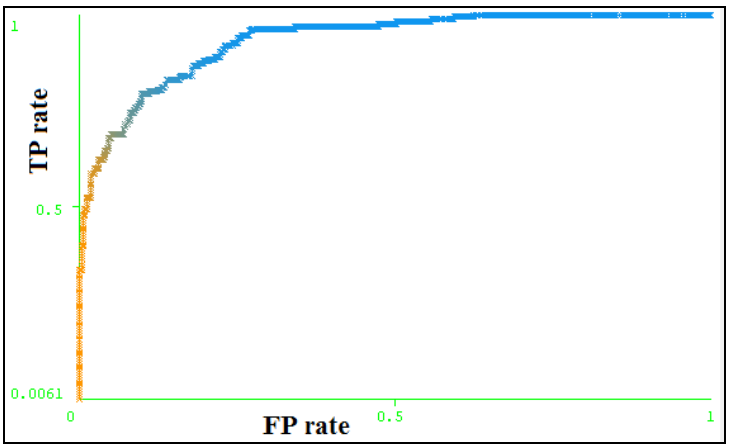


Figura 45: Curva ROC de um dos valores de classe

A Figura 46 mostra a diagrama em bloco da etapa de mineração de dados

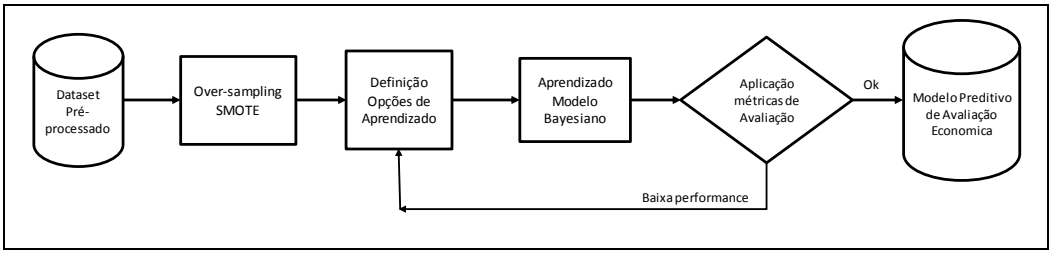


Figura 46: Diagrama em boco da etapa mineração de dados

4.4 Pós-Processamento

Completadas estas etapas, apresentamos o grafo da RB aos especialistas para a análise das dependências entre as variáveis ali representadas. Seguindo suas orientações rearranjamos alguns nós e arestas com o objetivo de inserir o conhecimento do domínio. Por exemplo, eles consideram que as variáveis Área, Espessura e Porosidade (todas referentes à acumulação) influenciam fortemente o valor do Volume de óleo. Sendo assim, nós modificamos a estrutura para expressar este conhecimento. Em seguida, as TPCs tiveram que ser reaprendidas e o modelo foi novamente testado apresentando acréscimo de 1 ponto percentual na acurácia.

Construído o modelo, já é possível começar a fazer inferências sobre o domínio. A plataforma Weka não dispõe de recursos para realização de inferências complexas. Porém, a plataforma SamIam, desenvolvida em Java pela equipe Automated Reasoning da University of California Los Angeles (UCLA)[16], oferece estas facilidades. SamIam possui uma interface gráfica amigável para a construção da RB, permite inferências exatas e aproximadas com várias opções de algoritmos.

[16] http://reasoning.cs.ucla.edu/samiam/

As principais inferências disponíveis nesta plataforma são cálculo das probabilidades a priori, most probable explanation (MPE: dadas as evidências calcula a instanciação da rede que explica tais evidências) e Maximum a Poteriori Hypothesis (MAP: similar a MPE, mas trabalha com um conjunto restrito de variáveis). Ambas as plataformas foram usadas nos nossos experimentos. O formato XML_BIF permite a transferência de informações de uma plataforma para outra.

Uma inferência que resolvemos realizar utilizando a base de dados reduzida e a plataforma SamIam foi qual seria o volume de óleo e o respectivo valor de vpl que justificam as seguintes evidências: área da acumulação = 0,73, espessura = 0,52, impostos = 0,81, lda = 0,33, tma = 0,47 e preço do óleo =0,77. Sendo que nada se sabe sobre as variáveis porosidade, profundidade da OE e qualidade do óleo (Tabela 8).

A RB nos forneceu como resposta que, dadas as evidências, o vpl esperado será de 0,41 e o volume será de 0,20 com uma probabilidade de 0,3%.

Tabela 8: Inferência utilizando RB

	Variável	Valor
Evidências	Area	0,73
	Espessura	0,52
	Impostos	0,81
	LDA	0,33
	TMA	0,47
	Preço_oleo	0,77
Desconhecidos	Porosidade	?
	Profundidade	
	Qualidade_Oleo	
Justificativa (0,3%)	Volume	0,20
	VPL	0,41

4.5 Outras Considerações

Apesar das evidências coletadas sobre o domínio apontarem para utilização de RB, resolvemos construir um modelo baseado na modalidade Rede Neural Artificial, já que, sendo nossas variáveis de maioria contínua, esta modalidade forneceria resultados que serviriam de comparação com nossos experimentos com RB.

De fato, em testes preliminares, foram alcançados índices aceitáveis de desempenho. Porém, o que nos inclinou a abandonar esta modalidade é que a participação do especialista ficaria prejudicada já que Redes Neurais possuem camadas ocultas que não possuem uma interpretação real no domínio em estudo, impedindo a inserção do conhecimento do domínio no modelo.

Contudo, fizemos alguns testes sem aplicação de discretização e variando as principais opções como: (i) número de vezes (500 ou 1000) que as instâncias passam pela rede (Época), (ii) se a taxa de aprendizado decresce (sim/não) e (iii) quantidade de camadas ocultas (uma ou duas). O resultado do experimento é mostrado na Tabela 9.

Tabela 9: Resultado dos experimentos com RNA

Modelo	Nº Épocas	Taxa aprendizado descresce?	Nº camadas ocultas	Acurácia
RNA1	500	não	1	0,81
RNA2	500	não	2	0,83
RNA3	500	sim	1	0,79
RNA4	500	sim	2	0,75
RNA5	1000	não	1	0,84
RNA6	1000	não	2	0,81
RNA7	1000	sim	1	0,78
RNA8	1000	sim	2	0,74

Considerando a acurácia dos modelos, o modelo RNA5 se destaca dos demais, pois apresenta acurácia de 84%.

Outra abordagem considerada, especialmente por se destacar pela rapidez de processamento, foi Máquina de Suporte Vetorial (SVM - Support Vector Machine). Esta se baseia em achar um hiperplano que divide o espaço de instâncias de maneira ótima conforme a classe. Em nossos experimentos esta modalidade apresentou valores semelhantes aos encontrados com Redes Neurais. Mostrou-se rápida com relação ao tempo de execução, mas sofre dos mesmos problemas de caixa-preta das Redes Neurais.

Durante a análise da base de dados realizada na etapa de pré-processamento algumas informações relevantes para os especialistas foram encontradas. Ao aplicar o algoritmo OneR (HOLTE, R., 1993), que seleciona a variável mais influente para classificação, identificamos que a variável Área da Acumulação é a que mais influencia o valor econômico de uma OE. Esta descoberta foi ratificada pelos especialistas, pois a área da acumulação está diretamente relacionada ao volume de óleo.

Para analisar e evidenciar melhor como a Área da Acumulação se relaciona com o VPL plotamos o gráfico VPL x Área com a respectiva curva exponencial de tendência que se ajusta aos dados conforme visto na Figura 47. Percebe-se uma proporcionalidade entre as duas grandezas apesar da influência das demais variáveis.

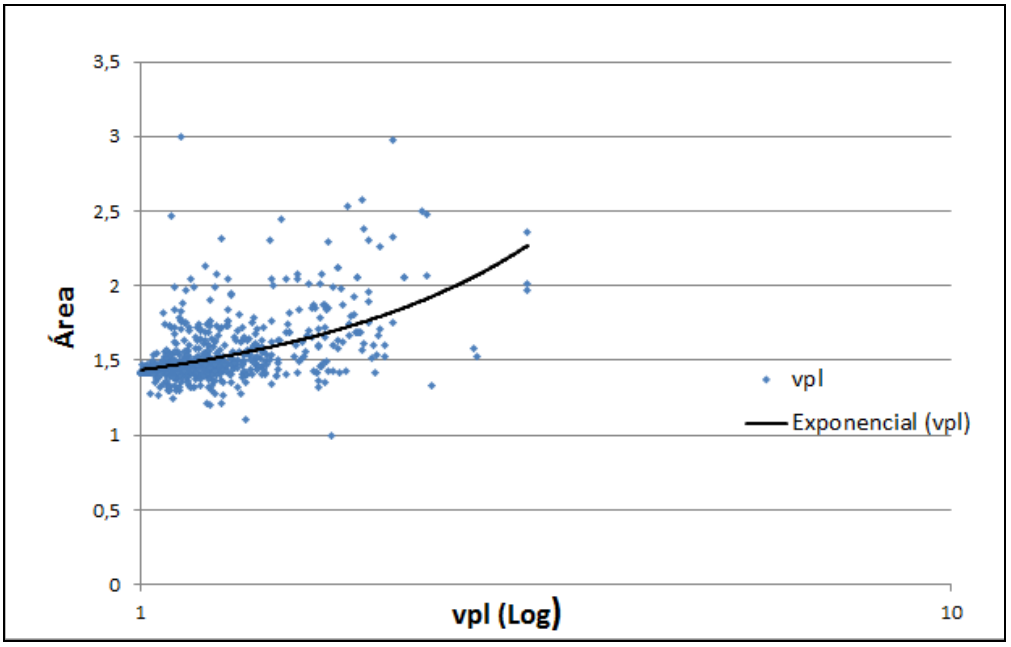


Figura 47: Gráfico VPL x Área da Acumulação

4.6 Avaliação dos Especialistas

Após conclusão do trabalho de pesquisa resolvemos submeter nossos resultados aos especialistas para que pudessem avaliar se o modelo proposto resultaria em ganho de produtividade em relação aos métodos convencionais por eles utilizados.

Foi realizada uma reunião nas dependências da Petrobras com a equipe multidisciplinar e pós-graduada responsável pelas avaliações econômicas de Oportunidade Exploratórias da Petrobras, composta pelos engenheiros Marcus Vinicius Rodrigues, Jalimar Guimarães e Regis Yuzo, pelo geofísico Juliano Binder e liderada pelo geólogo Luciano Arantes. Todos integrantes do Setor de Gestão de Portfólio Exploratório responsável por gerar e administrar a carteira de projetos gerando informações econômicas sobre as oportunidades e avaliando os riscos envolvidos nos

projetos. Após e durante a apresentação, onde foram expostos a metodologia e os resultados tanto parciais quanto finais, as seguintes observações foram emitidas pelos presentes:

- Foi reconhecida a aplicabilidade do trabalho e, devido aos resultados alcançados em nossos experimentos, formou-se a convicção de que a sua utilização melhorará as avaliações econômicas, tanto em termos de tempo quanto de precisão. No estágio atual, o modelo poderá ser usado como suporte à decisão.
- Sugeriram o acréscimo de outras variáveis relevantes ligadas à litologia (estudo das amostras geológicas coletadas durante a perfuração do poço) para aumentar a precisão da rede. Também houve sugestão de se aprender um modelo que prediga o tempo que uma sonda exploratória leva para perfurar um poço de petróleo, considerada por eles uma predição importante porque sondas são recursos escassos e demandam investimentos elevados.
- De um modo geral, a aprovação de nosso trabalho foi unânime e considerado por eles, estudiosos da atividade exploratória, como uma importante ferramenta para auxiliar no trabalho de avaliação econômica.
- Com relação à ontologia construída, os presentes consideraram que os principais conceitos estavam representados e que não foi percebida perda de expressividade significativa.

4.7 Análise do estudo de caso

Neste estudo de caso, com o apoio de uma ontologia de domínio, pudemos conhecer os principais conceitos que influenciam a avaliação econômica de uma OE e extrair informações necessárias à construção da RB do domínio. Na etapa de pré-processamento percebemos o comportamento cauda longa da variável de classe e resolvemos aplicar o seu logaritmo, preparando a base de dados para a etapa seguinte de mineração de dados.

Nossos experimentos na etapa de mineração de dados, especificamente RB, mostraram, através das métricas de avaliação, a viabilidade de se predizer o valor da avaliação econômica de uma OE utilizando dados históricos. O tempo de aprendizado do modelo, considerando uma base de 700 instâncias, foi de 1,3s e mostrou que o modelo é mais eficiente em termos de tempo do que uma avaliação convencional que pode tomar de 2 a 4 dias. Além disto, com a utilização do modelo, o especialista não ficará mais dependente de artefatos como curva de produção e o fluxo de caixa para prosseguir com suas avaliações.

A avaliação dos especialistas considerou o modelo como um dispositivo eficiente que contribuirá para a melhoria da tarefa de avaliação econômica.

Capítulo 5 – Trabalhos Relacionados

Existem algumas pesquisas no sentido de aplicar a mineração de dados para predição e classificação de valores no domínio da atividade petrolífera. Em (SCHOENINGER, 2003) é desenvolvida uma proposta de Representação da Certeza de Sucesso Exploratório através de um sistema especialista baseado em Lógica Fuzzy (RUSSEL, S. et al., 2009) que estima os riscos da exploração petrolífera e calcula a probabilidade de sucesso geológico de uma OE. Neste estudo a autora dividiu os dados em dois tipos: (i) objetivos seriam os dados históricos, enquanto que (ii) dados subjetivos seriam as informações sobre o prospecto fornecidas pelo especialista, usando sua experiência e sentimento.

Schoeninger analisou três modalidades de mineração de dados para tratar os dados Objetivos: Redes Neurais, RB e Lógicas Fuzzy. Tendo considerado Lógica Fuzzy como sendo a mais adequada ao seu trabalho. RB foi considerada como uma opção devido a sua característica de tratar informações incertas. Chegou-se a desenvolver um protótipo na plataforma Netica[17]. Porém, não havia, na ocasião, uma base de dados de atributos do sistema petrolífero, pois a defesa de uma OE era feita de forma descritiva. Desta maneira, as tabelas condicionais das RBs teriam de ser preenchidas de forma subjetiva,

[17] http://www.norsys.com/netica.html

isto é, manualmente. Isto inseriria fatores de intuição aumentando a incerteza do domínio. Além disto, a autora considerou que seria necessário construir uma RB para cada bacia sedimentar devido às diferenças geológicas inerentes. Neste experimento não houve preocupação com o aspecto econômico da OE, como cálculo do VPL. Sua preocupação era calcular as chances geológicas do projeto. Além disto, a autora utilizou dados fornecidos pela intuição conjuntamente com dados observados. Nosso trabalho se baseia totalmente em um histórico de avaliações e fornece resultados gerados por algoritmos que usam lógica probabilística, mais adequada a ambientes de incerteza.

A variável bacia sedimentar não foi considerada pelos nossos experimentos de seleção de atributos como uma variável de grande influência na classificação, por isto, não nos sentimos motivados a dar-lhe um tratamento especial como sugere a autora em seus estudos. Por fim, um modelo RB permitirá outros tipos de inferências além da avaliação econômica, portanto supomos ser mais abrangente.

Em (JUNIOR, 2003) o autor apresenta um estudo do atual quadro econômico da indústria de Exploração e Produção de Petróleo e do processo de decisão de investimentos sob incerteza, num ambiente de competição pela aquisição de áreas. É apresentado um modelo preditivo baseado em Redes Neurais, com auxílio de uma simulação de Monte Carlo, objetivando determinar as ofertas ótimas em um leilão de concessões. Foram obtidos 9.600 casos, divididos em dois grupos de 4.800. Estes dados foram utilizados para treinar a Rede Neural, que permitiram determinar o valor percebido por cada possível competidor, em função de 5 variáveis: volume de reservas, investimentos no desenvolvimento, custos operacionais, taxa de desconto do fluxo de caixa, preço do óleo no mercado. Uma vez otimizada e treinada, a rede está apta a

realizar predições a partir de um conjunto de valores de entrada. A Rede Neural utilizada foi do tipo Multilayer Perceptron. A primeira camada foi composta de 5 neurônios de acordo com o número de parâmetros de entrada do modelo. A última camada possui apenas um único neurônio, que corresponde ao valor presente do fluxo de caixa de um campo de petróleo. O número de camadas ocultas e a quantidade de neurônios em cada uma, foram otimizados geneticamente, assim como as funções de ativação dos neurônios e os seus pesos. Nosso trabalho se aplica a uma fase posterior, onde o leilão já ocorreu, a concessão exploratória já foi concedida e deixou de existir a competição pela concessão. Além disto, estamos utilizando a técnica RB.

Em (JUNIOR, 2010) utiliza-se mineração de dados, especificamente Redes Neurais, para melhorar o gerenciamento da produção de um campo de petróleo por meio da predição da produção período a período. O autor tira proveito da enormidade de dados armazenados pelos chamados poços inteligentes. Os resultados obtidos são comparados com os gerados pela simulação numérica (simulador comercial STARS), método utilizado na prática para a realização de predição da produção de petróleo considerando informações geológicas como os dados de rocha, os dados de fluido e as propriedades rocha-fluido. Foram feitas comparações com a técnica Equação de Balanço de Materiais.

O autor utilizou Redes Neurais Recorrentes (RNR), utilizadas como técnica para resolver problemas de séries temporais. RNR possui conexões de realimentação em sua topologia, ou seja, as saídas dos neurônios podem ser alimentadas de volta para suas entradas ou para os neurônios das camadas anteriores.

O autor justificou o uso de RNR afirmando que estas têm a capacidade de lidar

com sistemas de séries temporais que apresentam dificuldades para a predição, porque são não-lineares e possuem muitos ruídos. Os modelos estatísticos tradicionais são lineares e não são capazes de lidar com a natureza não-linear e a não-estacionaridade de certos sinais.

A grande vantagem observada pelo autor na utilização de RNRs foi o tempo de processamento das informações. Enquanto que no simulador as simulações duravam aproximadamente duas horas e meia, com as redes neurais o tempo gasto ficou em torno de 10 segundos. O conjunto de treinamento usado se refere aos 10 primeiros anos de produção de petróleo.

Foi verificado que tanto a predição da produção acumulada e da vazão não apresentou erro médio quadrático superior a 10%.

Nosso trabalho trata da atividade exploratória que é uma fase anterior a de produção, que se caracteriza por ser mais gerencial, e busca a otimização dos procedimentos.

Na pesquisa de (OMOLE et al., 2009) desenvolveu-se uma Rede Neural, tipo Back Propagation (BPNN), que prediz o valor de viscosidade de uma acumulação de petróleo localizada na Nigéria (delta do Rio Niger) e compara com os valores fornecidos pelas observações empíricas. A viscosidade é considerada uma característica importante, pois afetará o processo de produção assim como o custo de extração. Os autores utilizaram 32 instâncias. 22 para treinamento da rede e 10 para testes de acurácia. As variáveis essenciais foram: Temperatura do Reservatório, Grau API do óleo, Razão gás/óleo, densidade do gás, pressão do ponto bolha e viscosidade (classe). O erro absoluto

relativo da RN foi de 0,06781, enquanto que os métodos convencionais apresentaram erros de 0,45852 a 0,1741, levando a concluir que a solução por RNA teve melhor desempenho. Observa-se aqui que as instâncias utilizadas referem-se a uma região fora do Brasil, Nigéria. Além disto, o objeto da predição neste caso é a viscosidade e não os aspectos econômicos.

(MARTINELLI et al., 2011) apresenta uma proposta de RB que possibilita a análise de um prospecto (similar a um OE, mas melhor identificado) localizado no Mar do Norte. Explorando as semelhanças geológicas de uma área será possível predizer a viabilidade de um prospecto geológico. Neste estudo, tanto o grafo quanto os parâmetros são definidos pelos especialistas. Não há treinamento da RB. Em nosso trabalho, dispomos de uma base de dados que nos permitirá treinar uma RB, o que torna a nossa abordagem diferente.

Em (AVANSI, 2009) é proposto o uso de Modelos Proxy na definição das principais atividades da Engenharia de Reservatórios: Estratégia de Produção e Avaliação Econômica de Campos de Petróleo. Modelos Proxy são baseados na integração de métodos estatísticos como a Teoria Experimental Design (EDT) e a Metodologia Response Surface (RSM). Os processos tradicionais, como Simulação Numérica, demandam um número grande de simulações devido ao número de variáveis e cenários, e, mesmo assim, produzem resultados sub-ótimos. Modelos Proxy podem ser usados quando não se necessita de alta precisão nos resultados, reduzindo o número de simulações. O autor usou as seguintes variáveis nos seus experimentos: número de poços, capacidade de produção, capacidade de injeção, camadas de completação e timeline do poço. Foi aplicada discretrização nas variáveis contínuas. As funções de

avaliação utilizadas foram VPL, Produção Acumulada de Óleo e Produção Acumulada de Água. Os experimentos do autor mostraram uma redução de 80% no número de simulações, comprovando que seu método pode ser considerado uma alternativa eficaz aos métodos tradicionais.

No entanto, (ZUBAREV, 2009) considera que em situações de alta complexidade, grande espaço de soluções e incertezas, a aplicação de Modelos Proxy não é recomendada.

Capítulo 6 – Conclusão e Trabalhos Futuros

Este trabalho expôs os problemas envolvidos na avaliação de uma Oportunidade Exploratória de Petróleo e propôs uma solução para minimizar estes problemas.

Os resultados desta pesquisa apontam na direção de que é possível construir um modelo preditivo de avaliações econômicas. O tempo de aprendizado foi de apenas 1,3s e as métricas de acurácia (Figura 44) comprovam a viabilidade do modelo. A medida de acurácia alcançada pelo modelo RB8 (85%) foi considerada satisfatória.

No entanto, como os resultados reais sobre uma OE não estão disponíveis, pois seu valor econômico real só será conhecido após a conclusão do projeto de produção do campo de petróleo (que pode variar de 20 a 30 anos), não foi possível fazer uma comparação entre o valor real e o predito. Além disto, os dados mais antigos que compõem a base de dados são de 2002. Caso estas informações estivessem disponíveis seria possível extrair um significado mais preciso do erro de 15% apresentado pelo modelo.

De certo, a exploração de outros tipos de discretização, além das experimentadas por nós, poderia contribuir para construções de modelos mais eficientes.

Consideramos que o modelo bayesiano tem a seu favor o fato de basear as suas predições em um histórico de 700 avaliações. O que não ocorre com uma avaliação tradicional, que tende a ser uma experiência única, sujeita a julgamentos subjetivos por parte dos especialistas.

Nossos resultados restringem-se ao nosso território nacional, pois a base de dados possui informações apenas sobre avaliações em áreas localizadas no Brasil e exploradas por uma determinada empresa. Contudo, houve a preocupação de coletar dados referentes a todas as áreas geográficas com potencial de petróleo onde a Petrobras atua.

O tema por nós abordado mostrou ser de interesse da comunidade científica, pois em todos os eventos a que foi submetido, foi aceito (AFFONSO, M. et al., 2011 e 2012). E após as apresentações foi percebido o interesse que havia despertado.

De certo, avaliação econômica envolve outras coisas que não foram abordadas neste trabalho. Uma Análise de Risco completa, na qual o cálculo do VPL tem papel importante, envolve outros conceitos como VME (Valor Monetário Esperado) e Diagrama de Decisão que juntos permitem a análise de um cenário em que várias decisões devem ser tomadas uma após a outra (NEWENDORP et al., 2009). Teoria da Utilidade trata das preferências pessoais que o investidor assume diante de um investimento como o grau de aversão ao risco. Estes conceitos poderiam ser incluídos em um modelo preditivo posterior.

Outro trabalho a ser desenvolvido é gerar a distribuição de probabilidade de VPL por bacia sedimentar. Apesar de nossos experimentos não identificar a variável bacia

como uma das variáveis mais influentes é costume da área exploratória a análise por bacia, pois se sabe que possuem geologias específicas.

Referências

AFFONSO, M.; REVOREDO, K.; ANDRADE, L.; (2011); "Predição do Valor Econômico de uma Oportunidade Exploratória de Petróleo", Workshop de Teses e Dissertações de Sistemas de Informação, WTDSI 2011

AFFONSO, M.; REVOREDO, K.; ANDRADE, L.; (2011); "Prediction of Economic Value of a Petroleum Exploration Opportunity Using Bayesian Network", International Association for Development of the Information Society, IADIS 2011

AFFONSO, M.; REVOREDO, K.; ANDRADE, L.; (2012); "Avaliando uma Oportunidade Exploratória de Petróleo através de Mineração de Dados", Simpósio Brasileiro de Sistemas de Informação, SBSI 2012

AFFONSO, M.; REVOREDO, K.; ANDRADE, L.; (2012); "Evaluating a Petroleum Exploration Opportunity through Data Mining", International Conference on Enterprise Information System – ICEIS 2012

ANDEAS, B.; FRANCO, T.; (2012) "Mining Bayesian networks out of ontologies", Journal of Intelligence Information System (2012) 38:507–532

AVANSI, GUILHERME (2009); "Use of Proxy Models in the Selection of Production Strategy and Economic Evaluation of Petroleum fields", International Student Paper Contest 2009 - Society of Petroleum Engineers (SPE)

BEN-GAL, I.; RUGGERI, F.; FALTIN F.; KENETT R.; (2007) “Bayesian Networks”, Encyclopedia of Statistics in Quality & Reliability, Wiley & Sons

BORST, N.; (1997); “Construction of Engineering Ontologies”, PhD thesis, Centre for Telematica and Information Technology, University of Twenty, Enschede, HL.

CANLAS, RUBEN; (2009) “Data Mining in Healthcare: Current Applications and Issues”, Master of Science in Information Technology, Carnegie Mellon University - Australia

CAPPELLI, C.; BAIÃO, F.; SANTORO, F.; IENDRIKE, H.; LOPES, M.; NUNES, V.; (2007); “Uma Abordagem de Construção de Ontologia de Domínio a partir do Modelo de Processos de Negócio”, II Workshop on Ontologies and Metamodeling in Software and Data Engineering, WOMSDE 2007

CHAWLA, N.; BOWYER, K.; HALL, L.; KEGELMEYER, W. (2002) “SMOTE: Synthetic Minority Over-sampling Technique”, Journal of Artificial Intelligence Research 16 (2002) 321–357, Morgan Kaufmann Publishers.

COOPER, G.; HERSKOVITS, E.; (1990); “A Bayesian method for constructing Bayesian belief networks from databases”, UAI'91 Proceedings of the Seventh Conference on Uncertainty in Artificial Intelligence, Pages 86-94

DARWICHE, ADNAN; (2009) “Modeling and Reasoning with Bayesian Networks”, 1st ed. Cambridge University Press.

DE DOMBAL, T.; LEAPER, J.; HORROCKS, C.; STANILAND, R. (1974); “Human and computer-aided diagnosis of abdominal pain: Further report with emphasis on performance of clinicians”, British Medical Journal, 376-380

DEVITT, A.; DANEV, B.; MATUSIKOVA, K.; (2006) “Constructing Bayesian Networks Automatically using Ontology”, IOS Press Applied Ontology, Network Management Research Centre, Ericsson, Ireland.

FENZ, S.; TJOA, A.; HUDEC, M. (2009); “Ontology-based generation of Bayesian networks”, International Conference on Complex, Intelligent and Software Intensive Systems, IEEE

GLOVER, F.; McMILLAN, C.; (1985); “Interactive Decision Software and Computer Graphics for Architectural and Space Planning”, Annals of Operations Research 1985/6

GOLDBERG, DAVID; (2006) “Genetic Algorithms in Search, Optimization and Machine Learning”, Addison Wesley

GRUBER, T.; (1995); “Toward Principles for the Design of Ontologies Used for Knowledge Sharing”, International Journal Human-Computer Studies Vol. 43

GUIZZARDI, G.; WAGNER, G.; (2005); “Some Applications of a Unified Foundational Ontology in Business Modeling”, Applications of a Unified Foundational Ontology, Idea Group Inc

HAN, J.; KAMBER, M.; (2006) “Data Mining: Concepts and Techniques”, 2nd ed. Morgan Kaufmann. Pag 5-7.

HAYKIN, SIMON; (1999) “Neural Network: A Comprehensive Foundation”, New Jersey: Prentice Hall, 1999.

HOLTE, R.; (1993); “Very Simple Classification Rules Perform Well on Most Commonly Used Datasets”, Machine Learning 11:63-91

JANERT, PHILIPP; (2010) “Data Analysis with Open Source Tools”, Published by O’Reilly Media, Inc., Sebastopol - USA

JOHN, G.; LANGLEY, P.; (1995); “Estimating Continuous Distributions in Bayesian Classifiers”, In Proceedings of the Eleventh Conference on Uncertainty in Artificial Intelligence, Morgan Kaufmann Publishers, San Mateo, 1995.

JUNIOR, ALDAYR; (2010) “Predição não-linear de Curvas de Produção de Petróleo Via Redes Neurais Recursivas”, Dissertação de Mestrado – Universidade Federal do Rio Grande do Norte – UFRN – Engenharia de Petróleo

JUNIOR, REPSOL; (2003) “A Competição e a Cooperação na Exploração e Produção de Petróleo” – COPPE/UFRJ - Dissertação Mestrado. Pag 62-63; 171. Planejamento Energético.

KOLLER, D.; FRIEDMAN, N.; (2009) “Probabilistic Graphical Models”, MIT Press

LAURITZEN, S.; SPIEGELHALTER, D.; (1988) “Local Computations with Probabilities on Graphical Structures and Their Application to Expert Systems”, Journal of the Royal Statistical Society, Vol. 50, No 2, pp. 157-224, Blackwell Publishing

LELLIS, MAURO; (2007) “Fontes Alternativas de Energia Elétrica no Contexto da Matriz Energética Brasileira”, Dissertação de Mestrado em Engenharia da Energia – Universidade Federal do Itajubá – UNIFEI.

LINDEN, RICARDO; (2008) “Algoritmos Genéticos”, Rio de Janeiro: Brasport.

MACLAVE, J.; BENSON, P.; SINCICH, T.; (2009), “Estatística para Administração e Economia”, 10º Edição, Prentice Hall.

MARTINELLI, G.; EIDSVIK, J.; HAUGE, R.; FORLAND, M.; (2011) “Bayesian Networks for Prospect Analysis in the North Sea”, AAPG Bulletin 2011, PP. 1423-1442.

MITCHELL, TOM; (1997) “Machine Learning”, McGraw-Hill.

NADKARNI, SUCHETA; (2004) “A Causal Mapping Approach to Constructing Bayesian Networks”, Decision Support Systems, Vol. 38, Issue 2, 2004, pp. 259--281

NEWENDORP, P.; SCHUYLER, J.; (2009), “Decision Analysis for Petroleum Exploration”, 2nd Edition, Planning Press

OGWELEKA, FRANCISCA; (2011) “Data Mining Application in Credit Card Fraud Detection System”, Journal of Engineering Science and Technology

OMOLE, O.; FALODE, O.; DENG D.; (2009) “Prediction of Nigerian Crude Oil Viscosity Using Artificial Neural Network”, Petroleum & Coal, ISSN 1337-7027

PEARL, JUDEA; (1988) “Probabilistic Reasoning in Intelligent Systems”, Morgan Kaufmann Publisher

REVOREDO, K.; ZAVERUCHA, G.; (2004); “Search-based Class Discretization for Hidden Markov Model for Regression”, SBIA - Brazilian Symposium on Artificial Intelligence 2004, Volume 3171/2004, 317-325

RUSSEL, S.; NORVIG, P.; (2009); “Artificial Intelligence: A Modern Approach”, 3rd Edition, Prentice Hall, Part III and V

RUSSEL, S.; NORVIG, P.; (2009); “Artificial Intelligence: A Modern Approach”, 3rd Edition, Prentice Hall, Part III and V

SCHOENINGER, CINTIA; (2003) “Tratamento de Informações Imperfeitas na Análise de Risco de Prospectos em Exploração Petrolífera” - Universidade Federal de Santa Catarina (UFSC) - Dissertação Mestrado em Ciências da Computação/Inteligência Artificial.

SCHUYLER, JOHN (2001); “Risk and Decision Analysis in Projects”, 2nd Edition, Project Management Institute

SILVA, B.; LEONARDO, G.; MEDEIROS, R.; (2006); “Análise de Risco de Projetos de Produção Marítima de Petróleo”, Brazilian Business Review – BBR/2006, Vol. 3 Nº 2 pp. 229-244

WITTEN, I.; FRANK, E.; (2011) “Data Mining: Practical Machine Learning Tools and Techniques”, 3rd ed. Elsevier. Pag 5; 9; 278-279.

YU, L.; LAI, K.; WANG, S.; HE K.; (2007) “Oil Price Forecasting with an EMD-Based Multiscale Neural Network Learning Paradigm”, Part III, LNCS 4489, pags 1; 925–932, Springer-Verlag

ZUBAREV, D.; (2009); “Pros and Cons of Applying Proxy-models as a Substitute for Full Reservoir Simulations”, Annual Technical Conference and Exhibition 2009 - Society of Petroleum Engineers (SPE)