UNIVERSIDADE DO VALE DO RIO DOS SINOS - UNISINOS
UNIDADE ACADÊMICA DE PESQUISA E PÓS-GRADUAÇÃO
PROGRAMA DE PÓS-GRADUAÇÃO EM CIÊNCIAS CONTÁBEIS
NÍVEL MESTRADO

RAFAEL HERDEN CAMPOS

ANÁLISE DA RELAÇÃO ENTRE RISCO OPERACIONAL E PROCESSOS TECNOLÓGICOS

SÃO LEOPOLDO
2014

Rafael Herden Campos

ANÁLISE DA RELAÇÃO ENTRE RISCO OPERACIONAL E PROCESSOS TECNOLÓGICOS

Dissertação apresentada ao Programa de Pós-Graduação em Ciências Contábeis da Universidade do Vale do Rio dos Sinos – UNISINOS, como requisito parcial para obtenção do titulo de Mestre em Ciências Contábeis

Orientador: Prof. Dr. Adolfo Alberto Vanti

São Leopoldo

2014

C198a Campos, Rafael Herden.
Análise da relação entre risco operacional e processos tecnológicos / Rafael Herden Campos. – 2014.
130 f. : il. ; 30 cm.

Dissertação (mestrado) – Universidade do Vale do Rio dos Sinos, Programa de Pós-Graduação em Ciências Contábeis, 2014.
"Orientador: Prof. Dr. Adolfo Alberto Vanti."

1. Governança de tecnologia da informação. 2. COBIT (Padrão de gestão de tecnologia da informação). 3. Mineração de dados (Computação). 4. Risco operacional. I. Título.

CDU 657

Dados Internacionais de Catalogação na Publicação (CIP)
(Bibliotecário: Flávio Nunes – CRB 10/1298)

RAFAEL HERDEN CAMPOS

ANÁLISE DA RELAÇÃO ENTRE RISCO OPERACIONAL E PROCESSOS TECNOLÓGICOS

Dissertação apresentada ao Programa de Pós-Graduação em Ciências Contábeis da Universidade do Vale do Rio dos Sinos – UNISINOS, como requisito parcial para obtenção do titulo de Mestre em Ciências Contábeis.

Orientador: Dr. Adolfo Alberto Vanti

Aprovado em 22/05/2014.

BANCA EXAMINADORA

_____

Profª. Drª Eliana Rocío Rocha Blanco
Universidad de Cantábria – Santander/Espanha

_____

Prof. Dr. Marcos Antônio de Souza
Universidade do Vale do Rio dos Sinos - UNISINOS

_____

Prof. Dr.Clóvis Antônio Kronbauer
Universidade do Vale do Rio dos Sinos – UNISINOS

_____

Prof. Dr.Adolfo Alberto Vanti - Orientador
Universidade do Vale do Rio dos Sinos - UNISINOS

Dedico este trabalho aos meus pais José Roberto e Elaine
e aos meus irmãos Fernanda e Eduardo.

AGRADECIMENTOS

Agradeço a Deus por me fortalecer e me guiar sempre.

Agradeço aos meus familiares, que acompanharam esta jornada. Obrigado pela compreensão e apoio. Especialmente a minha avó Teresinha e tia Helena.

Agradeço a grandes amigos que colaboraram para a realização deste trabalho e que com sua amizade e companheirismo fizeram com que eu acreditasse em meu potencial. Três pessoas em especial, Ana Lúcia, Marineiva e Daniela.

Agradeço ao professor Dr. Adolfo A. Vanti, que me orientou nesta construção, sua dedicação, ensinamentos, paciência e seu esforço para que me superasse a cada dia.

Agradeço a todos os demais professores que souberam transmitir seus conhecimentos e foram fundamentais para o aprendizado que levarei para minha vida.

Agradeço à Universidade do Vale do Rio dos Sinos e a AES Sul por esta oportunidade de crescimento humano e profissional.

Agradeço aos colegas de trabalho, colegas do mestrado e colegas do grupo de pesquisa que estiveram presentes durante o curso.

Agradeço a todos que me incentivaram e torceram para que esta meta fosse alcançada.

A imaginação é mais importante que o conhecimento, o conhecimento é limitado, a imaginação infinita. (Albert Einstein)

RESUMO

O objetivo deste estudo é analisar as relações entre risco operacional e processos tecnológicos. São analisados os 34 processos pertencentes ao modelo de governança Cobit versão 4.1 com o auxilio de recurso computacional ligado à aprendizagem de máquina para extração de conhecimento - Data Mining. O método utilizado para desenvolvimento deste estudo foi definido como Design Research. Foram coletadas 341 respostas sobre os processos de governança de TI de 140 empresas localizadas principalmente no estado do Rio Grande do Sul. Para aplicação da mineração de dados foi utilizado o software de código aberto *Waikato Environment for Knowledge Analysis* (Weka), com o uso de algoritmos de geração de agrupamentos, seleção de atributos, classificação e associação. Os dados obtidos evidenciam que o processo com maior nível de maturidade nas 140 empresas pesquisadas é o ES5 – Garante a segurança dos sistemas, enquanto que o de menor nível é o PO7 – Gerencia os recursos humanos de TI. Os resultados também indicam que entre os 34 processos, PO7 - Gerencia os recursos humanos de TI, PO10 – Gerencia os projetos, AI1 – Identifica soluções de automação, AI2 – Adquiri e mantém software aplicativo, PO8 – Gerencia a qualidade são os processos de governança que possuem maior relação com a avaliação e gerenciamento de riscos (PO9).

Palavras-chave: Governança de TI. Cobit. Riscos Operacionais. *Data Mining*.

ABSTRACT

The objective of this study is to analyze the relation between operational risk and technological processes. 34 processes were analyzed belonging to the governance model COBIT version 4.1, with the aid of a computational resource associated with machine learning for knowledge extraction, known as Data Mining. The method used to conduct this study is defined as Design Research. It was collected 341 replies about the IT governance processes of 140 companies, located, mainly in the state of Rio Grande do Sul. For the application of Data Mining was used a software open source called Waikato Environment for Knowledge Analysis (Weka), using algorithms to generate clusters, attribute selection, classification and association. The obtained data show that the process with the highest level of maturity in the surveyed 140 companies is the ES5 - it ensures the security of the systems, while the lowest level is the PO7 - which Manage Human resources in IT. The results also show that between 34 processes, those which have greater relations with P09 - evaluating and managing risks, are the governance processes PO7 - Managing Human Resources in IT, PO10 - managing the projects, Al1 - Identifies automation solutions, Al2 - acquire and maintain application software, PO8 - Management the quality.

Keywords: IT governance. Cobit. Operational Risks. Data Mining.

LISTA DE FIGURAS

LISTA DE GRÁFICOS

LISTA DE QUADROS

LISTA DE TABELAS

LISTA DE ABREVIATURAS

BIS	*Basel Committee on Banking Supervision*
COBIT	*Control Objectives for Information and Related Technology*
COSO	*Committee of Sponsoring Organizations of the Treadway Commission*
DCBD	Descoberta de Conhecimento em Base de Dados
DM	*Data Mining*
DR	*Design Research*
DSS	*Decision Support System*
DW	*Data Warehouse*
ERM	*Enterprise Risk Management*
ISACA	*Information Systems Audit and Control Association*
ITGI	*The Information Technology Governance Institute*
KDD	*Knowledge Discovery in Database*
SAD	Sistema de Apoio à Decisão
TI	Tecnologia da Informação
WEKA	*Waikato Environment for Knowledge Analysis Software*

SUMÁRIO

1 INTRODUÇÃO

O uso de tecnologia da informação e seus respectivos sistemas são regulados por uma Governança de TI (WEILL; ROSS, 2006), necessária para garantir principalmente o princípio da transparência empresarial sustentado por sistemas de informação.

As empresas que realizam governança em informação conseguem decidir melhor, sobressaindo-se perante os concorrentes (WEILL; ROSS, 2006). Desta forma, revisam processos, métodos e estratégias e também consideram os riscos operacionais (COSO, 2010) de suas atividades, principalmente de fatores internos como falhas ou inadequações junto a processos, pessoas e sistemas, tendo a participação do usuário (SPEARS; BARKI, 2010).

Vaughan (1997) conceitua o risco como uma condição em que existe a possibilidade de ocorrer um resultado adverso ao esperado, e essa possibilidade não precisa ser mensurável, pois apenas a sua existência já o caracteriza. Portanto, para o autor, o risco só existe quando há mais de uma possibilidade de resultado futuro e pode se apresentar em diferentes maneiras, como disponibilizado em falhas na base de dados de *Data Warehouse* (DW), o qual sustenta o Sistema de Informação Contábil (BAI; NUNEZ; KALAGNAM, 2012).

Conforme Witten e Frank (2005), a garimpagem de dados compreende a avaliação automática ou semiautomática de grandes volumes de dados possibilitando a previsão de conhecimento e comportamentos, identificando tendências e descobrindo relações entre diferentes variáveis, entre outras vantagens para a tomada de decisões.

Gartner, Zwicker e Rödder (2009) estabeleceram relação entre investimentos em tecnologia da informação e produtividade. Sendo assim, a análise de um sistema de apoio à decisão aliado a uma rigorosa análise dos riscos envolvidos nas atividades, podem representar significativamente bem a realidade dos processos, permitindo a análise crítica do estado atual da maturidade, bem como possibilitando a criação de metas objetivas.

O presente estudo é continuidade de pesquisas realizadas em conjunto pelo grupo de pesquisa GTI Unisinos/CNPQ com pesquisadores da Universidad de

Cantabria / Santander – Espanha. Os dados coletados também são oriundos das atividades do grupo de pesquisa.

1.1 Problema de pesquisa

Considerando os processos tecnológicos e os riscos operacionais envolvidos nas mais diferentes organizações, bem como suas complexidades, o presente estudo tem definido como questão de pesquisa: Quais as relações entre risco operacional e processos tecnológicos?

1.2 Objetivos

1.2.1 Objetivo Geral

Para responder a questão de pesquisa foi desenvolvido o seguinte objetivo: Analisar a relação entre os riscos operacionais e os processos tecnológicos.

1.2.2 Objetivos Específicos

- Analisar as relações entre os processos de governança de TI.
- Analisar os processos que possuem maior influência sobre o processo de avaliação e gestão de riscos.
- Definir e processar diferentes regras de garimpagem de dados sobre o processo de gestão de riscos.

1.3 Justificativa

A presente pesquisa se justifica diante de duas perspectivas. A primeira diz respeito às lacunas existentes na literatura sobre o tema que está sendo estudado. Já a segunda trata da forma como os processos tecnológicos são tratados no contexto prático das empresas, tratados de maneira mais subjetiva, menos

automatizada e com isso impactam nos riscos operacionais e no planejamento do desenvolvimento de processos tecnológicos das organizações.

Riscos operacionais são os “riscos de perdas resultantes da inadequação ou falha de processos internos, pessoas e sistemas, ou de eventos externos, envolvendo risco legal, porém excluindo o risco estratégico e reputacional” (BIS, 2006, p. 144). Goldstein, Chernobai e Benaroch (2011) argumentam que mesmo em empresas não-financeiras ocorrem perdas operacionais significativas, como no caso da United Airlines que sofreu um desligamento de um sistema crítico de dados no ano de 2007, ocasionando mais de 20 vôos cancelados e o atraso de outros 250, resultando em uma perda superior a US$ 10 milhões.

Riscos operacionais e as grandes perdas acabam por tornarem-se importantes e de relevância neste assunto, pois envolvem questões sobre identificação e mitigação dos mesmos.

O processo de *Data Mining* é viável para identificar novas informações, válidas e potencialmente úteis, embora não possa ser considerado um processo trivial tendo como objetivo gerar resultados de forma eficiente, que possam ser visualizados e interpretados pelos gestores (FAYYAD; PIATETSKY-SHAPIRO; SMYTH, 1996).

Segundo Markus, Majchrzak e Gasser (2002), as informações resultantes da análise de processos de TI devem estender as fronteiras da resolução dos problemas humanos e suas capacidades organizacionais, fornecendo capacidades, bem como ferramentas computacionais e teorias sobre sua aplicação e o impacto que irá acompanhar o seu desenvolvimento e utilização. No entendimento que esta dissertação foi conduzida, busca-se trazer maior conhecimento sobre possíveis oportunidades que técnicas automatizadas de extração de conhecimento podem trazer às organizações.

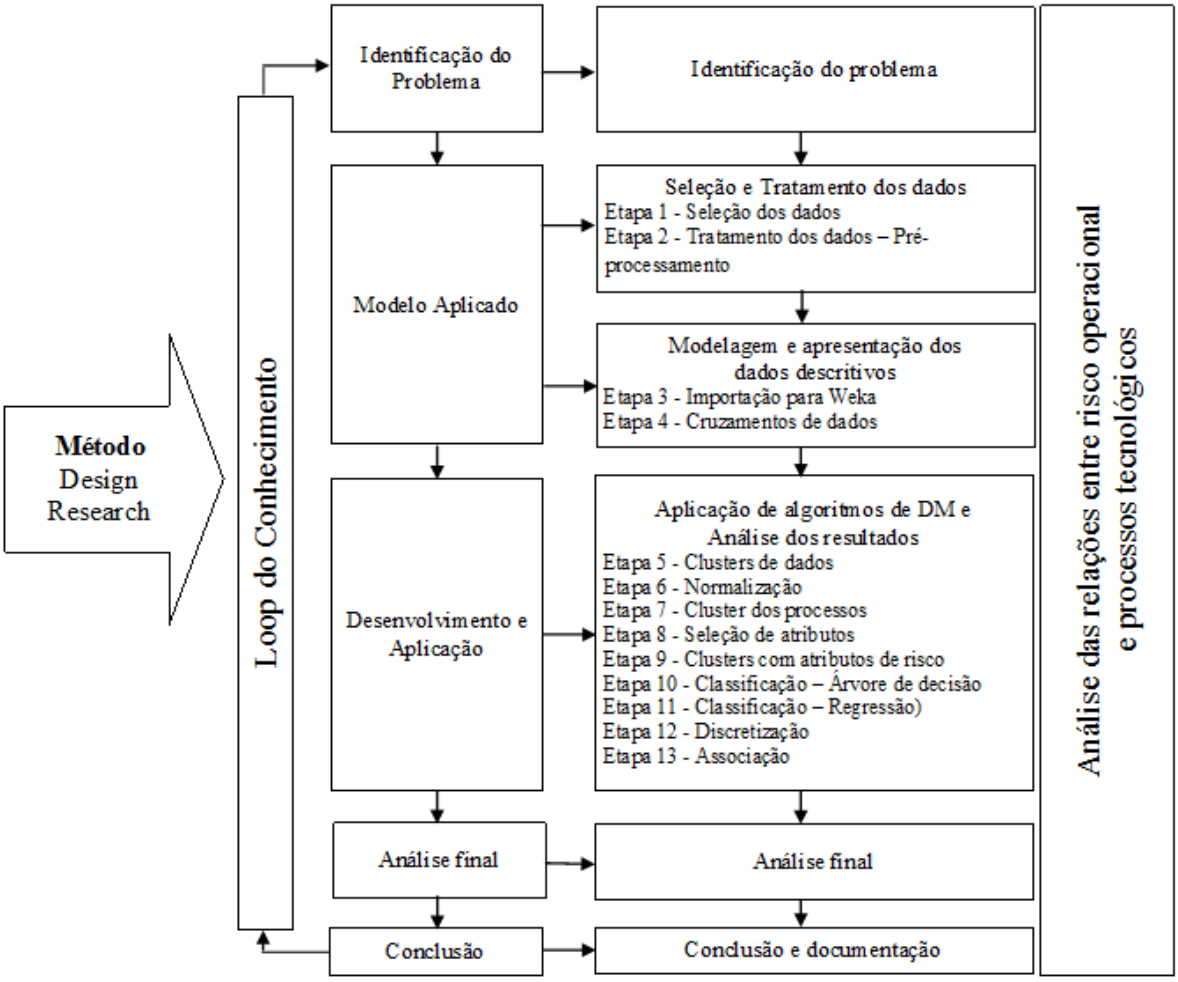
No âmbito teórico, pesquisas que utilizam-se de técnicas de *Data Mining* têm sido conduzidas com diversos objetivos, entre eles, detecção de fraude (ANGELOS *et al*., 2011) e (GONZÁLEZ; VELAZQUEZ, 2013), e de riscos operacionais (MURAYAMA *et al*., 2011) e (CHENG *et al*., 2012). O último autor explorou os riscos operacionais, suas causas e sua distribuição em acidentes de trabalho na indústria da construção em Taiwan através de técnicas de mineração de dados, buscando estabelecer padrões de causa-e-efeito sobre acidentes de trabalho.

A pesquisa de Shiri, Amini e Raftar (2012) aborda a detecção de risco de falências de empresas através de ferramentas de mineração de dados, gerando relatórios de previsão de dificuldades financeiras, de possibilidade de fraude, de gestão de estimativa de risco de crédito, e de previsão de desempenho corporativo, demonstrando diversas aplicações possíveis para a mineração de dados.

Quanto ao âmbito prático da pesquisa, esta se fundamenta na dificuldade de identificação das relações entre riscos operacionais e processos tecnológicos nas organizações, visto que existe uma ampla gama de variáveis que interferem nos processos de governança e que muitas vezes são analisadas de forma subjetiva, baseadas nas experiências dos componentes da equipe que está gerando tal análise, e ainda, ficam atreladas ao fato de envolver enorme quantidade de possibilidades, podendo passar despercebido algum fator relevante. Assim, esta abordagem, com o uso de técnicas de mineração de dados, pode auxiliar os gestores na tomada de decisões mais acertadas por basearem-se em informações com maior acurácia.

Com o objetivo de apresentar as etapas metodológicas realizadas no desenvolvimento da pesquisa, através do método de *Design Research* e *Data Mining* que são pouco utilizados na área das ciências sociais aplicadas, foi gerado um framework, presente na Figura 1.

Figura 1: Framework da pesquisa



Fonte: Elaborado pelo autor a partir de Hevner et al. (2004) e Shiri, Amini e Raftar (2012).

A Figura 1 apresenta o framework metodológico desenvolvido na pesquisa que é composto em duas partes. A primeira, à esquerda, contempla o método de Design Research e a segunda considera a aplicação das técnicas de *Data Mining* ligadas ao método de DR.

1.4 Estrutura do trabalho

O presente projeto está organizado em cinco capítulos. O primeiro trata da introdução, apresentando o tema que está sendo estudado, o objetivo geral e os

objetivos específicos, a justificativa e a estrutura do trabalho. No segundo capítulo tem-se o referencial teórico, relacionado ao tema de riscos operacionais, governança de TI e mineração de dados. No capítulo três são desenvolvidos os procedimentos metodológicos utilizados na pesquisa, apresentando o *Design Research* e as etapas do método. No quarto são apresentados os resultados gerados pelos algoritmos através das análises realizadas e no quinto, é apresentada a conclusão do estudo. Por fim são apresentadas as referências bibliográficas utilizadas, apêndices e anexos.

1.5 Delimitação

Este estudo delimita-se considerando os temas de riscos operacionais (BIS, 2003) e (HAIJA; HAYEK, 2012), e governança de tecnologia da informação, mais especificamente referente ao processo de avaliação e gerenciamento de riscos em relação aos demais processos tecnológicos referente ao modelo de maturidade do Cobit versão 4.1. O estudo refere-se a uma análise de técnicas legadas do sistema *Data Mining* (DM), que quando aplicado na análise e gestão de riscos operacionais, auxilia na racionalização do processo de mensuração e avaliação de níveis de maturidade das organizações, aumentando ainda mais a transparência empresarial.

O presente estudo se propõe ao desenvolvimento de um artefato, conhecido com método, através da aplicação da análise das relações entre riscos operacionais e processos tecnológicos, onde os riscos operacionais são tratados pelo processo de governança de avaliação e gerenciamento de riscos e os processos tecnológicos tratam dos demais processos de governança do Cobit 4.1.

Os resultados dos classificadores foram testados com a opção *Use Training Set*, do software Weka, na qual ocorre o treinamento com os dados disponíveis e na sequência é realizado os testes para verificação das regras geradas.

2. FUNDAMENTAÇÃO TEÓRICA

2.1 Gestão de riscos operacionais

Guimarães *et al.* (2009) entendem que gestão de risco é um tema de importantes discussões e ponderam que as incertezas devem ser alvo de monitoramento. Dessa forma, a gestão de risco dá suporte a decisões, atuando de forma a minimizar riscos ou fatores de exposição aos riscos.

Conforme Oliveira e Pacheco (2011, p.250) “risco é a probabilidade de que ocorra algo não esperado quanto ao retorno do investimento ou a probabilidade de ocorrer algo diferente do esperado”. Bernstein (1996, p.197), compreende que “quando corremos um risco, apostamos em um resultado que será consequência da decisão tomada, embora não se saiba ao certo qual será o resultado”. Desta forma, se compreende que o risco é a possibilidade de ocorrer algo inesperado ou incerto.

Os riscos podem resultar em perdas para a organização e para uma correta gestão de riscos se faz necessária uma avaliação prévia dos objetivos da organização em relação aos ambientes interno e externo, bem como a probabilidade de eventos que afetarão o desempenho da empresa (ZONATTO; BEUREN, 2010).

Segundo Varella (2005), a gestão de riscos é um processo sistemático de definição, análise e resposta aos riscos, que tem como meta maximizar os eventos positivos e minimizar os negativos. O autor ainda traz que risco é qualquer condição ou evento em potencial que possa afetar os objetivos propostos e Woon, Azizan e Samad (2011) mencionam que a natureza do risco é consequência do desempenho incerto que compromete recursos e resultados.

Segundo Silva, Fernandes e Grande (2008), um marco na gestão de riscos foi o documento Gerenciamento de Risco Empresarial - Estrutura Integrada (*Enterprise Risk Management – Integrated Framework*) em 2004, que acabou sendo conhecido como ERM. Esse documento foi desenvolvido por seis associações do *Committee of Sponsoring Organizations of the Treadway Commission* (COSO) e a consultoria *Pricewaterhouse-Coopers* como um instrumento que efetivamente permite identificar, avaliar e gerenciar riscos e surgiu para incorporar uma filosofia diferenciada de gestão de riscos e não substituir o modelo de controles internos proposto pelo COSO.

De acordo com Oliveira (2004), o gerenciamento do risco operacional é a identificação das falhas nas diferentes tarefas, que acabam resultando em perdas financeiras. De acordo com a relevância das falhas e o capital envolvido, irá possibilitar a identificação de quais indicadores de risco devem ser monitorados e com que frequência isto será feito.

O comitê de Basiléia define o risco operacional como "o risco de perdas resultantes da inadequação ou deficiência de processos internos, pessoas e sistemas, ou de eventos externos" (BIS, 2003, p. 2).

Amaral *et al.* (2009), atribuem o risco operacional a fatores internos e externos. Exemplificando, internamente ligados a fatores de falha de produção, de projetos, de processos e controles, ao fator humano e recursos materiais inadequados ou insuficientes inseridos no contexto. Os fatores externos são independentes da ação da organização cabendo apenas a assimilação ou ainda uma forma de construir uma proteção através de meios de segurança. Corrobora para esse entendimento Haija e Hayek (2012) quando afirmam que o risco operacional é o risco de perdas “diretas ou indiretas resultantes de processos inadequados ou deficientes, pessoas e sistemas ou de eventos externos. Risco operacional pode se manifestar de forma tangível nas interrupções dos negócios, as falhas de controle, erros ou evento externo”.

De uma maneira ampla, o risco operacional é o risco ligado às atividades cotidianas de uma organização e faz necessário o gerenciamento do mesmo através da gestão de desempenho, processos, colaboradores e sistemas. Riscos operacionais incluem tudo que pode levar a perdas financeiras, como falhas nos controles internos e governança corporativa, sejam através de fraudes ou pela não realização das operações em tempo hábil. Existem ainda outros eventos que podem ser incluídos como risco operacional, dentre eles estão as falhas significativas nos sistemas de TI ou desastres como um grande incêndio (SAVIC, 2008).

Relacionado aos riscos operacionais, as organizações devem compreender quais fatores são suas causas bem como suas prováveis consequências quando expostas aos mesmo e, a partir disto, criar programas de administração e controle de riscos. Também é importante que o entendimento sobre os riscos operacionais seja completa e compreenda todos os aspectos que possam causar perdas operacionais significativas (BIS, 2003).

Nesse sentido, o Comitê de Basiléia desenvolveu classificações por "tipo de evento" de riscos operacionais. Estes “tipos de eventos” são agrupados em componentes distintos de acordo com a natureza do evento de risco operacional e são definidos pela fonte de risco que a empresa enfrenta, não aplicados a uma atividade específica (DEMPSER, 2008).

Com o intuito de identificar e classificar as possíveis fontes de perdas operacionais foi elencado pelo *Risk Management Group of the Basel Committee on Banking Supervision* os principais tipos de riscos operacionais, conforme Quadro 1 (BIS, 2003).

Quadro 1: Classificação de riscos operacionais segundo acordo de Basiléia II.

Tipo	Principais Ocorrências
Fraude interna	Erros intencionais sobre as posições, roubos por parte dos empregados, utilização de informação confidencial em beneficio pessoal.
Fraude externa	Roubo, falsificação, cheques sem fundos, invasão dos sistemas de informação.
Relações trabalhistas e segurança do trabalho	Solicitações de indenizações por parte dos empregados, infração às normas trabalhistas de segurança e saúde, acusações de discriminação e obrigações de maneira geral. Assedio.
Práticas com os clientes, produtos e negócios	Abusos de confiança, Uso indevido de informação confidencial sobre os clientes, negociação fraudulenta, bloqueio de capitais, venda de produtos não autorizados.
Danos a ativos materiais	Terrorismo, vandalismo, terremotos, incêndios, inundações.
Interrupção das atividades e Falha nos sistemas	Falhas de hardware e software, problema de telecomunicação, interrupção de serviços públicos.
Execução, entrega e processamento	Erros na entrada de dados, documentação jurídica incompleta, concessão de acesso não autorizado à conta de clientes, litígios com fornecedores e clientes.

Fonte: Adaptado de BIS (2003).

É relevante notar que a classificação dos dados por tipos de eventos que geram perda, tal como foi proposto pelo Comitê de Basiléia, foi desenvolvida pelos reguladores do setor bancário em conjunto com as próprias instituições financeiras, portanto, em algum grau, é específico para este setor. No entanto, como os riscos operacionais são muito relevantes para as empresas, estas categorias podem ser

usadas como ponto de partida para a criação de uma taxonomia de risco específico para outros setores da economia (DEMPSER, 2008)

Dempser (2008) também entende que utilizando-se de uma estrutura já existente, é possível realizar comparações entre empresas. A importância destas comparações é a possibilidade de incentivar os investidores a procurarem em todos os setores da economia, aquelas empresas que propiciem o risco/retorno apropriado em seus investimentos. Exemplificando, um investidor que queira incluir em seu portfólio um investimento com alto risco/alto retorno pode considerar investir em qualquer setor com segurança, desde que os empresários destes setores desenvolvam uma taxonomia de risco comparável (DEMPSER, 2008).

Outra classificação que visa complementar o entendimento sobre a classificação de riscos operacionais e pode ser mais amplamente utilizado por empresas não financeiras é a trazida por Savic (2008), quando apresenta as cinco principais categorias de Risco Operacional: 1. Organização; 2. Processos e políticas; 3. Sistemas e tecnologia; 4. Pessoas; 5. Eventos externos. Para o entendimento, cada uma das categorias é apresentada no Quadro 2.

Quadro 2: Categorias de Risco Operacional.

Tipo	Principais Ocorrências
Organização	Riscos decorrentes de questões como a gestão da mudança, gerenciamento de projetos, cultura corporativa e comunicação, responsabilidades, alocação e planejamento de continuidade de negócios.
Processo e Política	Riscos decorrentes de deficiências em processos como a liquidação e pagamento, não-conformidade com as políticas internas ou a regulação externa, ou falhas em produtos ou negócios do cliente.
Sistemas e Tecnologia	Riscos decorrentes de defeito falhas de hardware ou software, em outras tecnologias, tais como redes ou telecomunicações, bem como violações de segurança de TI.
Pessoas	Riscos decorrentes de falha de empregados, empregador, conflito de interesses ou de comportamento interno ou de fraudes.
Externa	Riscos decorrentes de fraudes ou litígio por partes externas à empresa, bem como a falta de segurança física para a instituição e seus representantes.

Fonte: Adaptado de Haija e Hayek (2012).

Haija e Hayek (2012), entendem que estas cinco principais categorias, presentes no Quadro 2, apresentam uma base válida para a identificação de problemas de gestão de riscos operacionais. No entanto os autores vão além, pois compreendem que para resolução de novos desafios e riscos, é importante a disciplina e o desenvolvimento de um *framework* claro para que se possa resolver os problemas.

Savic (2008) entende que dependendo das características da empresa, alguma categoria se sobrepõe parcialmente a outra. Isto pode ser visto em uma empresa financeira, onde os sistemas e a tecnologia interagem para produzir um processo bem sucedido, e quando um não funciona, o outro acaba sendo afetado.

Savic (2008, p.89) também apresenta alguns dos principais fatores que representam direcionadores para a identificação do risco operacional quando afirma:

> Novos produtos, a sofisticação do produto, canais de distribuição, novos mercados, novas tecnologias, a complexidade da tecnologia, e-commerce, o volume de negócios, a nova legislação, a globalização, pressões de regulamentação, fusões e aquisições, reorganizações, rotatividade de pessoal, a diversidade cultural de funcionários e clientes.

Sendo assim, não existe um modelo definido para gestão de riscos operacionais e que as organizações os monitoram conforme os impactos que podem causar em suas atividades bem como em seus resultados e deve ser realizado pela direção da organização, gerentes e funcionários (SILVA; FERNANDES; GRANDE, 2008). Os autores também afirmam que monitorar os riscos e garantir que eles estejam em conformidade com a propensão ao risco aceitável, permite prover, com razoável segurança, o alcance dos objetivos

Assim, para uma gestão dos riscos mais eficiente e eficaz é necessário considerar dois aspectos: probabilidade de ocorrência do risco e seu provável impacto na organização. Depois de identificados os riscos potenciais, os mesmos devem ser analisados e mensurados, e para isso, a construção de uma matriz de risco torna-se importante. Após, a direção da organização deve dar respostas aos riscos, estes podendo ser evitados, aceitados ou transferidos, de forma a mitigá-los (SILVA; FERNANDES; GRANDE, 2008).

2.2 Governança de TI

Considerando que na atualidade as empresas necessitam dos sistemas de TI para desenvolverem suas atividades, os controles sobre os sistemas de TI são críticos. Controles da TI se relacionam com a regulação dos ambientes de TI e incluem o acesso aos sistemas, programas, operações e gestão (KUHN JR.; AHUJA; MUELLER, 2013).

A evolução dos sistemas de informação e sua infraestrutura interna, bem como nos intercâmbios de informações entre as organizações, resulta em uma crescente necessidade de gerir estrategicamente toda esta infraestrutura. O alinhamento da governança de TI com as estratégias e objetivos da organização se tornam um pré-requisito fundamental para a sustentabilidade a longo prazo (SCHOLL; PATIN, 2014).

De acordo com Herath e Herath (2014), os controles de TI têm um efeito importante sobre a realização de muitos objetivos, uma vez que os dados que são utilizados nos relatórios financeiros são extraídos, armazenados e processados em sistemas de TI e estes devem atingir um nível condizente de controles internos a fim de se manterem seguros e disponíveis para as organizações.

Segundo Kuhn Jr., Ahuja e Mueller (2013), o processo de conformidade desenvolvido a partir da lei Sarbanes-Oxley (SOX), se propôs a mitigar riscos, evitar a ocorrência de fraudes, permitindo maior transparência nos negócios. A governança de TI ajuda a atender as necessidades de informação financeira, garantir a integridade dos dados financeiros, bem como de documentos e garantir a segurança dos sistemas de TI para que os funcionários tenham acesso somente aos dados designados, mantendo registros e informações confiáveis preservadas.

O propósito da lei SOX segundo Kuhn Jr., Ahuja e Mueller (2013) é proteger os investidores, melhorando a precisão e confiabilidade das divulgações corporativas e, restaurando a confiança do investidor na integridade dos relatórios financeiros das empresas. Também trouxe para executivos e investidores a noção que a governança de TI é essencial para o bom desenvolvimento das atividades das empresas garantindo segurança e confiabilidade. Confirmando este entendimento, Herath e Herath (2014) concordam que as normas contábeis tiveram um impacto visível sobre as práticas de segurança da informação nas organizações.

Abordagens sobre a padronização de segurança em TI têm evoluído rapidamente e atualmente existem numerosas normas de segurança internacionais fornecendo diretrizes, promovendo as melhores práticas, e servindo como base para certificações. Os padrões de segurança mais amplamente aceitos e aplicados incluem ISO 27001, que trata da segurança da informação operacionalmente e a versão 4.1 do Control Objectives for Information and Related Technology – COBIT, que se trata de um modelo de governança de TI (LAMBRINOUDAKIS, 2013).

O modelo de governança de TI Cobit 4.1 foi desenvolvido pela *Information Systems Audit and Control Association* (ISACA), e construído de forma estruturada, gerenciável e lógica, com objetivo de atender as necessidades da governança corporativa e os requisitos do negócio (ISACA, 2014). O Modelo é composto por processos de TI e estes possuem três características inter-relacionadas que são: (a) A capacidade de TI é composta por um portfólio de processos de TI individuais. (b) Individualmente, os processos de TI tem diferentes impactos sobre a capacidade da empresa para responder ao ambiente competitivo. (c) A padronização de processos individuais é uma questão fundamental para atingir sucesso do portfólio da capacidade de TI (DEBRECENY; GRAY, 2013).

Debreceny e Gray (2013) entendem que o framework do Cobit 4.1 trata da gestão e controle de TI e que é relevante ao longo do ciclo dos sistemas de informação, no planejamento através do desenvolvimento e aquisição de soluções para posterior implantação, juntamente com um conjunto de controles de monitoramento e feedback. Segundo Kerr e Murthy (2013), o Cobit está organizado em quatro domínios, que correspondem as principais áreas de responsabilidade dentro da TI: Planejar e Organizar (PO), Adquirir e Implementar (AI), Entregar e Suportar (ES), e Monitorar e Avaliar (MA), dentre os quais, os domínios PO e MA tratam dos principais aspectos de governança, enquanto que AI e ES estão mais focados nos aspectos de gestão.

Conforme Debreceny e Gray (2013), o COBIT é uma abordagem abrangente para a governança de TI e gestão. Os principais componentes do Cobit são divididos em 34 processos de TI que estão alocados nos quatro domínios apresentados. Os 34 processos, na visão de Kerr e Murthy (2013), são fatores-chave para o controle de TI e devem ser gerenciados cuidadosamente para que haja uma governança de TI eficaz.

Os 34 processos do Cobit 4.1 podem ser visualizados no Quadro 3.

Quadro 3: Processos do Cobit 4.1

Domínio	Processo	Descrição
Planejar e Organizar	PO1	Define o planejamento estratégico de TI.
	PO2	Define a arquitetura da informação.
	PO3	Determina as Diretrizes da Tecnologia.
	PO4	Define a organização de TI e seus relacionamentos.
	PO5	Gerencia o Investimento de TI.
	PO6	Comunica as metas e diretrizes gerenciais.
	PO7	Gerencia os recursos humanos de TI.
	PO8	Gerencia a qualidade.
	PO9	Avalia e gerencia os riscos.
	PO10	Gerencia os projetos.
Adquirir e Implementar	AI1	Identifica soluções de automação.
	AI2	Adquiri e Mantém Software Aplicativo.
	AI3	Adquire e mantém a arquitetura tecnológica.
	AI4	Desenvolve e mantém procedimentos de TI.
	AI5	Obtém recursos de TI.
	AI6	Gerenciar mudanças
	AI7	Instala e certifica soluções e mudanças.
Entregar e Suportar	ES1	Define níveis e mantém os acordos de níveis de serviços.
	ES2	Gerencia os serviços de terceiros.
	ES3	Gerenciar desempenho e capacidade da TI.
	ES4	Garante a continuidade dos serviços.
	ES5	Garante a segurança dos sistemas.
	ES6	Identifica e Aloca Custos
	ES7	Educa e treina os usuários.
	ES8	Gerencia a central de serviços e incidentes.
	ES9	Gerencia a configuração.
	ES10	Gerencia os problemas.
	ES11	Gerencia os dados.
	ES12	Gerencia a infra-estrutura.
	ES13	Gerenciar operações.
Monitorar e Avaliar	MA1	Monitora e avalia a desempenho de TI.
	MA2	Monitora e avalia o controle interno.
	MA3	Assegura a conformidade aos requisitos externos.
	MA4	Fornecer governança de TI.

Fonte: ISACA (2014)

A Governança de TI é um processo pelo qual os objetivos do impacto da tecnologia da informação são determinados, gerenciados e controlados. A Governança de TI inclui o estabelecimento de direitos de decisão, definição de objetivos, bem como a capacidade organizacional para atender esses objetivos, que

serão monitorados através de *loops* de *feedback*, que empregam métricas para sua mensuração (DEBRECENY; GRAY, 2013).

Um tema importante na análise das capacidades organizacionais, segundo Debreceny e Gray (2013), tem sido o desenvolvimento de métricas para verificar o nível de maturidade dos processos.

O enfoque desenvolvido pela ISACA é derivado do modelo de maturidade do *Software Engeneering Institute* (SEI), porém difere do mesmo devido ao Cobit promover uma definição em formato de escala "interpretada de acordo com a natureza dos processos de gerenciamento de TI" pelo qual entende-se que "não há intenção de medir os níveis de maneira precisa ou tentar certificar que aquele nível foi exatamente atingido. A avaliação de maturidade do Cobit espera resultar em um "perfil" pelo qual, atendendo as condições pré-estabelecidas atingirá um determinado nível de maturidade (ISACA, 2014, p.21). No Quadro 4, pode ser visualizada os níveis de maturidade do Cobit 4.1.

Quadro 4: Níveis de maturidade do Cobit 4.1

	Nível de Maturidade	**Descrição**
0	Inexistente	Completa falta de um processo reconhecido. A empresa nem mesmo reconheceu que existe uma questão a ser trabalhada.
1	Inicial / *Ad hoc*	Existem evidências de que a empresa reconheceu que existem questões e que precisam ser trabalhadas. No entanto, não existe processo padronizado; ao contrário, existem enfoques *Ad Hoc* que tendem a ser aplicados individualmente ou caso a caso. O enfoque geral de gerenciamento é desorganizado.
2	Repetível	Os processos evoluíram para um estágio onde procedimentos similares são seguidos por diferentes pessoas fazendo a mesma tarefa. Não existe um treinamento formal ou uma comunicação dos procedimentos padronizados e a responsabilidade é deixada com o indivíduo. Há um alto grau de confiança no conhecimento dos indivíduos e consequentemente erros podem ocorrer.
3	Definido	Procedimentos foram padronizados, documentados e comunicados através de treinamento. É mandatório que esses processos sejam seguidos, no entanto possivelmente desvios não serão detectados. Os procedimentos não são sofisticados, mas existe a formalização das práticas existentes.

4	Gerenciado e Mensurável	A gerência monitora e mede a aderência aos procedimentos e adota ações onde os processos parecem não estar funcionando muito bem. Os processos estão fundamentados em um constante aprimoramento e fornecem boas práticas. Automação e ferramentas são utilizadas de uma maneira limitada ou fragmentada.
5	Otimizado	Os processos foram refinados a um nível de boas práticas, baseado no resultado de um contínuo aprimoramento e modelagem da maturidade como outras organizações. TI é utilizada como um caminho integrado para automatizar o fluxo de trabalho, provendo ferramentas para aprimorar a qualidade e efetividade, tornando a organização rápida em adaptar-se.

Fonte: ISACA (2014)

Segundo Debreceny e Gray (2013), a padronização de processos resulta em qualidade, no qual pode-se melhorar a confiabilidade, a previsibilidade, gerar custos mais baixos, e aumentar a flexibilidade e agilidade. Esta melhoria na padronização de processos também é associada com o conceito de maturidade do processo.

De forma genérica, o Cobit trata de duas grandes categorias de controles internos de TI. A primeira categoria se refere aos controles de aplicação, que atuam sobre os sistemas de processos de negócio e relatórios financeiros e incluem medidas preventivas, bem como medidas de detecção de transações não autorizadas. A segunda categoria, de controles gerais, atua na infraestrutura de TI e visa garantir operações seguras e ininterruptas. Com esta forma de organização, busca-se manter os níveis de Governança de TI desejados pelos gestores (KUHN JR.; AHUJA; MUELLER, 2013).

2.3 Mineração de dados

Segundo Turban *et al.* (2009, p.153), o processo de mineração de dados é a descoberta de conhecimento em banco de dados, o qual se utiliza de “técnicas estatísticas, matemáticas, de inteligência artificial e de aprendizagem automática para extrair e identificar informações úteis de banco de dados”.

Tan, Steinbach e Kumar (2009) afirmam que a vasta quantidade de informações em bases de dados, aliada à importância da extração de conhecimento útil como suporte à decisão, têm exigido investimentos consideráveis das empresas

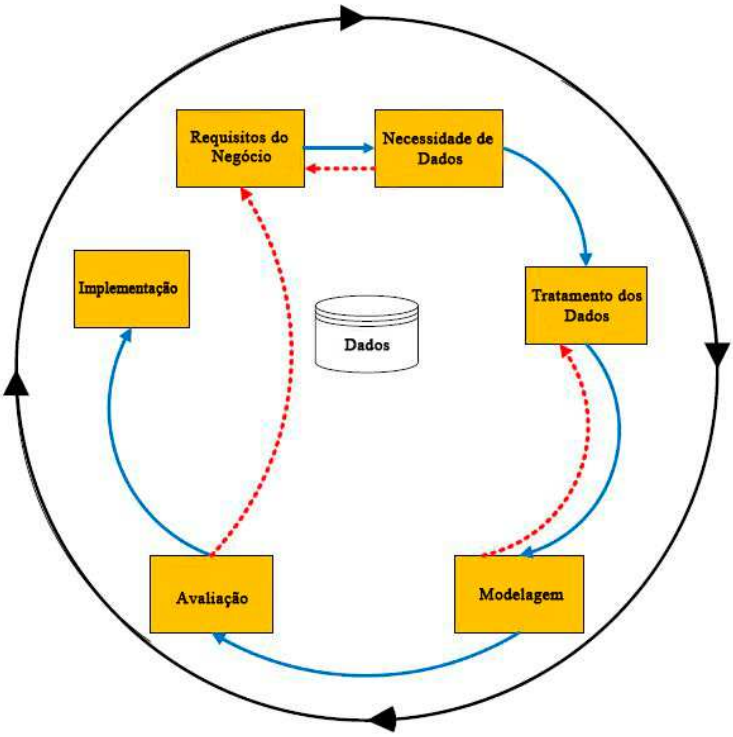
e da comunidade científica, por ser reconhecidamente uma tarefa desafiadora. No mesmo entendimento, Witten e Frank (2005) argumentam que para lidar com volumes de dados significativos existem técnicas que tem se mostrado de grande valia na obtenção de informações.

Na indústria da informação e na sociedade como um todo, as técnicas de DM tem atraído atenção devido a ampla variedade aplicações, em áreas como previsão de padrões comportamentais, identificação de tendências e descoberta de relações entre dados, com o objetivo de otimizar a tomada de decisão (COBO *et al.*, 2012).

As técnicas de Mineração de Dados pertencem a um campo da ciência conhecido como Descoberta de Conhecimento em Base de Dados (DCBD) - *Knowledge Discovery in Database* (KDD), no qual se aplicam técnicas para análise inteligente e automática de imensos repositórios de dados com o objetivo de encontrar informações relevantes (WITTEN; FRANK, 2005).

Segundo Sharma, Osei-Bryson e Kasper (2012), a Descoberta de Conhecimento através de Mineração de Dados é um processo de diversas fases que incluem o entendimento sobre o negócio (*Business Understanding*), a necessidade de dados, o tratamento destes dados, a modelagem, a avaliação e a implantação. O processo KDD, que pode ser visualizado na Figura 2, é interativo e complexo, uma vez que cada fase envolve múltiplas tarefas e existem numerosas dependências intra e interfases entre as várias tarefas do processo.

Figura 2: Processo de KDD



Fonte: Sharma, Osei-Bryson e Kasper (2012)

A descoberta de conhecimento em bancos de dados é um processo não trivial de identificar novas informações, válidas e potencialmente úteis com o objetivo de gerar, de forma eficiente, resultados que possam ser visualizados e interpretados através da interação homem-máquina (FAYYAD; PIATETSKY-SHAPIRO; SMYTH, 1996). Ainda, Seng e Chen (2010) corroboram afirmando que o KDD busca conhecimentos úteis, regularidades e informações de alto nível que possam ser extraídas a partir dos conjuntos de dados relevantes de repositórios de dados, bem como a visualização a partir de ângulos diferentes, possibilitando uma melhor visão do negócio.

Data Mining é a análise inteligente e automática de imensos repositórios de dados com objetivo de encontrar informações relevantes (WITTEN; FRANK, 2005). Na mesma linha de entendimento Tan; Steinbach; Kumar (2009) afirmam que a mineração de dados é o processo de descoberta de informações úteis de forma automática em grandes bancos de dados.

O *Data Mining* integra técnicas e algoritmos distintos. Na área das ciências, a Estatística, que é construída sua tecnologia, para avaliar e validar resultados; Banco

de Dados, refere-se aos recursos para se manipular grandes repositórios de dados, através de armazenamento, indexação e processamento de consultas; e Inteligência Artificial, através do reconhecimento de padrões e aprendizagem de máquina (*machine learning*), utiliza-se da heurística, que é a ciência que busca entender como o homem pensa a resolução de problemas (TAN; STEINBACH; KUMAR, 2009).

Segundo Williams (2012), atualmente a técnica de DM tem sido muito utilizada como uma ferramenta crítica para a coleta de informações relevantes para as empresas e para a avaliação de riscos operacionais. As agências reguladoras também estão utilizando essa técnica como forma de avaliar os grandes volumes de dados gerados pelo *mandatory compliance* e pelo *disclosure* na busca de agentes de mercado, cujas práticas são suspeitas e merecedoras de uma análise mais aprofundada, que é por vezes referido como análise de tendências de risco. Segundo González e Velásquez (2013), nos últimos anos, as técnicas de DM e inteligência artificial tem sido incorporadas nos planos de atividades das auditorias.

As técnicas de *Data Mining* contribuem com métodos para melhorar os controle dos negócios, reduzindo o risco ao prever situações indesejáveis e fornecendo recomendações com base em experiências anteriores (BAJO *et al*., 2012). Ainda, a mineração de dados provê técnicas relevantes, facilitando a interpretação dos dados e ajudando a melhorar a compreensão dos processos por trás dos dados, com o auxilio de algoritmos rápidos e eficientes (GONZÁLEZ; VELÁSQUEZ, 2013).

Conforme Shmueli e Koppius (2011), o poder preditivo (ou a precisão da previsão) se refere à capacidade de um modelo gerar previsões precisas de novas observações, considerando um período futuro ou observações que não foram incluídas na amostra original. Segundo Campos *et al*. (2014), essa previsibilidade é um recurso robusto do DM e isso permite aos gestores uma melhoria nos processos organizacionais e de tomada de decisão bem como no desenvolvimento de estratégias.

No âmbito de gestão, o *Data Mining* tem se mostrado bastante útil por analisar e verificar situações a fim de predizer ou descrever os fatos. Estudos na área com o intuito de mitigar riscos têm sido conduzidos por recentes pesquisas

(Valle, Varas e Ruz, 2012; Shiri, Amini e Raftar, 2012; Murayama *et al*., 2011;Cheng *et al*., 2012; González e Velásquez, 2013).

Conforme Tan, Steinbach e Kumar (2009), as tarefas de *Data Mining* podem ser divididas em duas categorias, as Preditivas e as Descritivas. As Preditivas tratam de predizer algum dado baseado em atributos de outros dados, que se apresentam como variáveis dependentes e independentes. As Descritivas buscam derivar padrões através de tendências, grupos e correlações a fim de sumarizar as relações entre os dados.

Deve-se notar que existe diferença entre tarefas e técnicas de *Data Mining*. As tarefas referem-se ao tipo de informação que se quer buscar nos dados e seus padrões, enquanto que a técnica engloba os métodos que nos permitem encontrar os padrões e informações buscadas (SILVA; RALHA, 2011). No próximo capítulo será tratado sobre as principais técnicas e algoritmos de Mineração de Dados.

O processo de Mineração em banco de dados consome muito tempo na parte de pré-processamento, devido que parte dos dados são ruidosos (com erros, ou valores fora do padrão), incompletos (ausentes), inconsistentes (discrepâncias semânticas), tudo isso especialmente pelo volume significativo de dados a serem trabalhados. Com esta etapa concluída, que visa uma melhor qualidade e riqueza semântica, inicia-se a etapa de mineração de dados, ou seja, é realizada a seleção das técnicas e ajustes dos parâmetros dos algoritmos que melhor realizarão a busca por padrões representativos, bem como sua efetiva busca (SILVA, 2004).

Quanto à aplicação dos algoritmos de *Data Mining*, estes podem ser divididos em dois tipos: Aprendizagem Supervisionada e Aprendizagem Não-Supervisionada.

Aprendizagem Supervisionada

Nesta categoria, é fornecido um atributo no qual cada amostra da mineração pertence. São classificados como algoritmos preditivos, ou seja, apresentam inferência nos dados a fim de se obter tendências. Busca informações não disponíveis a partir de dados disponíveis (SILVA, 2004).

Os grupos de algoritmos desta categoria podem ser verificados no Quadro 5.

Quadro 5: Grupos de algoritmos de Aprendizagem Supervisionada

Algoritmo	Utilização	Exemplo
Classificação	Determinar o valor de um atributo através dos valores de um subconjunto dos demais atributos da base de dados	Inferir (prever) que “clientes do sexo feminino, com renda superior a R$ 1.500,00 e com idade acima de 30 anos compram cosméticos importados”.
Seleção de atributos	Para bases de dados que se encontram atributos que têm um peso maior ou até determinante nas tarefas de mineração de dados	No caso de cliente, a sua renda com certeza é um atributo determinante nos seus hábitos de consumo, no entanto seu nome não influência hábitos de consumo.

Fonte: Adaptado de Silva (2004).

Aprendizagem Não-Supervisionada

Nesta categoria, segundo Silva (2004), não é fornecido um atributo, nem o número de atributos no qual a amostra pode pertencer. São classificados como algoritmos descritivos, ou seja, descrevem de forma clara as características das propriedades dos dados encontrados.

Os tipos de algoritmos desta categoria podem ser verificados no Quadro 6

Quadro 6: Grupo de algoritmos de Aprendizagem Não-supervisionada

Algoritmo	Utilização	Exemplo
Associação	Quando a classe de uma tarefa de mineração não é determinada como no caso da classificação. O próprio algoritmo elege os atributos determinantes.	Clientes do sexo masculino, casados, com renda superior a R$ 1.800,00 têm o seguinte hábito de consumo: roupas de grife, perfumes nacionais, relógios importados.
Clustering	Definição de uma métrica de similaridade para cada atributo e uma função de combinação destas métricas em uma métrica global, os objetos são agrupados com base no princípio da maximização da similaridade intraclasse e da minimização da similaridade interclasse.	Identificar subgrupos homogêneos de clientes de uma determinada loja.

Fonte: Adaptado de Silva (2004).

Diante destas categorias apresentadas nos Quadros 5 e 6, existem algoritmos que fazem parte de cada uma destas. No quadro 7, são apresentados algoritmos utilizados no estudo, conforme sua categoria.

Quadro 7: Algoritmos de *Data Mining* usados no estudo

Métodos de classificação	Método de Seleção de Atributos	Métodos de Agrupamento	Método de Associacão
•REPTree •M5	•CfsSubsetEval	•EM •SimpleKMeans	•Apriori

Fonte: Elaborado pelo autor.

A utilização de *Data Mining* nas organizações pode ser tratada como uma inovação tecnológica, por possuir a função de extrair conhecimento de uma grande base de dados, de forma automatizada e agiliza o processo de transferência de conhecimento (TAN; STEINBACH; KUMAR, 2009). Para aplicar técnicas de mineração de dados, se faz o uso de algoritmos, presentes na próxima sub-seção.

2.3.1 Algoritmos de *Data Mining*

Nesta subseção são apresentados algoritmos de *Data Mining* empregados na pesquisa. Todos os algoritmos foram utilizados a partir do software Weka. Foram aplicadas técnicas de clustering, classificação, seleção de atributos e de associação.

2.3.1.1 Algoritmos de Cluster

Clustering é o processo de agrupar os itens de banco de dados sob determinados clusters ou agrupamentos. Todos os itens do mesmo cluster tem características semelhantes e itens pertencentes a diferentes clusters tem características diferentes (REVATHI; SUMATHI, 2013).

Na pesquisa são apresentados dois algoritmos de agrupamento usados o EM (*expectation maximisation*), que gera os agrupamentos baseado em probabilidade e o KMeans, que gera os agrupamentos a partir distância total de cada um dos pontos do cluster ao seu centro.

2.3.1.1.1 Algoritmo EM - *Expectation Maximisation*

O algoritmo de *clustering* EM atribui uma distribuição de probabilidade para cada instância e também indica a probabilidade de que pertencem a cada um dos agrupamentos. O EM pode decidir quantos agrupamentos é possível criar por validação cruzada, ou pode ser especificado a priori o número de grupos que se deseja criar (WITTEN; FRANK, 2005).

Para gerar a expectativa de maximização, são realizados dois passos, o primeiro, é gerado o cálculo das probabilidades do cluster (que são os valores esperados), a expectativa; o segundo, é realizado o cálculo dos parâmetros de distribuição, que é a maximização da probabilidade das distribuições dos dados.

No algoritmo EM o algoritmo converge para um ponto fixo dos agrupamentos, mas sua pretensão não é o atingir de maneira simples, mas a partir de uma medida de qualidade.

A probabilidade global é obtida multiplicando-se as probabilidades das instâncias individuais i, onde estas são dadas aos agrupamentos A e B sendo determinadas a partir da função de distribuição normal.

De maneira prática, a probabilidade global é a medida da qualidade dos agrupamentos e aumenta a cada iteração do algoritmo EM, ou seja, o logaritmo global é calculado a partir da soma dos logaritmos dos componentes individuais, eatravés da realização de sucessivas iterações até que a melhoria torna-se insignificante (WITTEN; FRANK, 2005).

2.3.1.1.2 Algoritmo KMeans

KMeans é um método de agrupamento iterativo, no qual os itens são movidos pelo algoritmo entre os conjuntos de cluster até o conjunto desejado a ser atingido (REVATHI; SUMATHI, 2013). É um método de agrupamento simples e eficaz . Witten e Frank (2005)

O método Kmeans é uma técnica de agrupamento amplamente utilizada que procura minimizar a distância ao quadrado médio entre pontos no mesmo cluster. Embora ele não ofereça garantia de exatidão, a sua simplicidade e velocidade são muito atraentes na prática (ARTHUR; VASSILVITSKII, 2007).

Na técnica de agrupamento KMeans, o algoritmo começa com um conjunto arbitrário de centros ou são especificados antecipadamente quantos grupos estão

sendo procurados, representado pelo parâmetro k. Então k pontos são escolhidos aleatoriamente como centros de cluster. Todas as instâncias são atribuídas a seu centro de cluster mais próximo de acordo com a métrica da distância euclidiana.

A escolha do centro do agrupamento ou centróide é baseada no cálculo pelo mínimo do quadrado da distância total de cada um dos pontos do cluster ao seu centro. Este processo é realizado iterativamente até que se estabilize e não se verifique melhora na redução ao mínimo quadrado da distância total entre os centros e os demais pontos. Desta forma, cada ponto é atribuído ao seu agrupamento central mais próximo (WITTEN; FRANK, 2005).

Complementando, Larose (2005) apresenta o processo do algoritmo KMeans em cinco passos:

Passo 1: Peça ao usuário em quantos agrupamentos k o conjunto de dados deve ser dividido.

Passo 2: Aleatoriamente são atribuídos k registros para serem os locais iniciais do centro cluster.

Passo 3: Para cada registro, se encontra o centro mais próximo cluster. Assim, cada centro de cluster é "dono" de um subconjunto dos registros, representando, uma divisão do conjunto de dados, encontrando k clusters (C1, C2,..., Ck).

Passo 4: Para cada um dos agrupamentos k, se encontra o baricentro do cluster e se atualiza a localização de cada centro de cluster para o novo valor do centróide.

Passo 5: É repetido os passos 3 a 5 até que a convergência seja completa e as distâncias entre os registros e os centros não encontre signifivativa redução na soma dos erros quadrados.

O critério de "mais próximo" no passo 3 é normalmente a distância Euclidiana, embora outros critérios possam ser aplicados também.

O algoritmo termina quando os centróides não mudam, ou seja, quando algum critério de convergência é cumprido, com nenhuma redução significativa na soma dos erros quadrados.

2.3.1.2 Algoritmo de Seleção de Atributos

2.3.1.2.1 Algoritmo CfsSubsetEval

O CfsSubsetEval é um algoritmo de seleção de atributos baseado em correlação, aplicados a problemas de aprendizagem supervisionada, incluindo aqueles em que a variável é numérica. O algoritmo classifica subconjuntos de atributos de acordo com uma função de avaliação heurística baseada em correlação (HALL, 1998).

A função de avaliação busca nos subconjuntos de dados os atributos que são altamente correlacionados com a classe e não correlacionadas entre si. Atributos irrelevantes são ignoradas porque possuem baixa correlação com a classe. Subconjuntos redundantes também são eliminados devida alta intercorrelação. A aceitação de um atributo irá depender do grau com que ele prevê as classes a partir das instâncias (HALL, 1998).

O algoritmo gera um ranking de subconjuntos de atributos a partir da base de dados e busca todos os possíveis subconjuntos de atributos, permitindo realizar a função de avaliação. No entanto, essa enumeração exaustiva de todos os possíveis subconjuntos de atributos requer alto custo computacional e é proibitiva na maioria dos casos. Para permitir o funcionamento otimizado do algoritmo, existem e estratégias heurísticas de pesquisa, que são os buscadores do algoritmo e estão integrados ao mesmo (HALL, 1998)..

As estratégias de pesquisa, utilizadas no presente estudo conjuntamente com o avaliador CfsSubsetEval foram quatro: Linear Foward Selection, Greedy Stepwise, Best First e Evolutionary Search, cujos principais aspectos podem ser vistos na sequência.

Linear Foward Selection: Inicia a pesquisa com um número restrito de atributos, e vai aumentando o número de atributos em cada etapa, até atingir um determinado tamanho. A busca utiliza tanto a ordenação inicial para selecionar os melhores atributos, ou realiza um ranking entre os atributos (GUETLEIN *et al.*, 2009).

Greedy Stepwise: Executa uma busca sem atributos, ou com o total dos atributos do banco de dados e vai adicionando ou eliminando atributos até a adição ou exclusão dos atributos restantes resulta em uma piora na avaliação (HALL, 1998).

Best First: Busca nos subconjuntos de atributos por adição ou eliminação, iniciando por um conjunto vazio de atributos e adicionando novos atributos em cada etapa ou começa com o conjunto completo de atributos e vai eliminando atributos a cada etapa até que nenhuma melhora se verifique na avaliação de cinco subconjuntos consecutivos em relação ao atual melhor subconjunto (HALL, 1998).

Evolutionary Search: Realiza a busca na base de dados usando um algoritmo evolutivo segundo os critérios de inicialização aleatória e substituição geracional com elitismo , isto é, o melhor indivíduo é mantido sempre.

2.3.1.3 Algoritmo de Classificação/Regressão

Classificação é o processo localizar um conjunto de regras para determinar a classe de um objeto a partir de seus atributos. (REVATHI; SUMATHI, 2013). Os algoritmos de classificação utilizados na pesquisa foram RepTree, M5 e de Regressão.

2.3.1.3.1 Algoritmo REPTree

O REPTree (*Reduced Error Pruning Tree*) é um algoritmo de aprendizagem de árvores de decisão que utiliza a informação de ganho/variância e poda, sendo que a poda é usada para redução de erro (ZHAO; ZHANG, 2008).

O algoritmo está otimizado para gerar resultados de forma otimizada, com baixo custo computacional e por isso classifica atributos numéricos uma única vez. Entre as opções, podem ser definidas como o número mínimo de instâncias por folha, profundidade máxima da árvore e proporção mínima de conjunto de treinamento de variância, bem como o número de dobras para a poda (WITTEN; FRANK, 2005).

Em relação aos resultados obtidos pelo algoritmo, após cálculo dos valores contidos nos parênteses e colchetes, chega-se a cobertura no conjunto de treinamento / erros no conjunto de treinamento e cobertura após a poda / erros no conjunto poda. Pode haver casos de números fracionários (instâncias com peso <1) (WITTEN; FRANK, 2005).

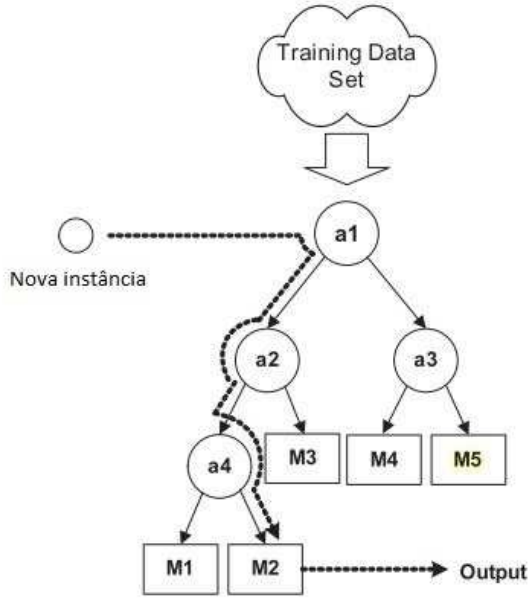
2.3.1.3.2 Algoritmo M5P

M5P é um modelo de aprendizagem de árvore de decisão que constrói árvores com o uso de regressão logística. As variáveis aceitas são do tipo binárias e multiclasse, atributos numéricos e nominais. Ao montar as funções de regressão logística em um nó, ele usa a validação cruzada para determinar quantas iterações são necessárias para sua execução em vez de fazer a validação cruzada em cada nó. Esta heurística, a qual pode se optar por não utilizar, melhora o tempo de execução consideravelmente, requerendo menor custo computacional, com pouco efeito sobre a precisão (WITTEN; FRANK, 2005).

Algoritmos M5 utilizam-se da seguinte idéia: dividir o espaço de parâmetros em áreas (subespaços) e construir em cada um deles um modelo de regressão linear especializada local. A divisão segue a lógica utilizada na construção de uma árvore de decisão sendo que a qualidade do modelo é geralmente medida pela precisão com a qual se prevê para os atributos-alvo.

Um exemplo de árvore de decisão pode ser visualizado na Figura 3.

Figura 3: Exemplo de árvore de decisão



Fonte: Bhattacharya e Solomatine (2005).

Onde, ai são os nós de divisão e Mj são os modelos.

Os modelos de aprendizagem de árvore de decisão com M5 possuem aprendizagem de forma eficiente e podem enfrentar tarefas de alta dimensionalidade com centenas de atributos, .funcionando da seguinte maneira:

Tendo como objetivo construir um modelo que relaciona um atributo alvo em relação aos valores de outros atributos de entrada, modelos baseados em árvore são construídos por um método de dividir para conquistar. Para divisão, o conjunto de dados é associado a uma folha ou utiliza-se um teste que divide o mesmo em subconjuntos correspondentes aos resultados do teste e assim aplica-se de forma recursiva para os demais dados. O critério de divisão para o algoritmo de árvore modelo M5 é baseado na medida de erro do desvio padrão e calcula a redução esperada neste erro como resultado de teste para cada atributo do nó.

Depois de examinar todos os desdobramentos possíveis, isto é, os atributos e os prováveis valores de divisão, o M5 escolhe aquele que minimiza a probabilidade de erros estimados. A divisão termina quando os valores de todos os registros de um determinado nó variam ligeiramente. Muitas vezes a divisão gera estruturas que devem ser podadas, como uma sub-árvore de uma única folha.

Por fim, um processo de suavização é executado para compensar as descontinuidades que ocorrem entre os modelos lineares adjacentes nas folhas da árvore podada, especialmente nos modelos construídos a partir de um pequeno número de itens. Na suavização, os dados das equações lineares adjacentes são atualizados de modo que os resultados previstos para os vetores de entrada vizinhos possuam valores próximos para as diferentes equações (BHATTACHARYA; SOLOMATINE, 2005).

2.3.1.3.3 Algoritmo de Regressão

Entre os algoritmos de classificação do Weka estão os que foram projetados para treinamento e previsão de um único atributo (classe), alvo para previsão. Alguns classificadores podem utilizar somente classes nominais e outros só podem utilizar classes numéricas (para problemas de regressão). Ainda existem os que permitem a utilização de ambos os tipos de classes.

Para classificação baseada em uma única estrutura, o Weka apresenta um desempenho muito bom, porém, podendo ser de difícil interpretação quando há

muitas opções de nós, porque dificulta a visualização de uma grande quantidade de informações. No entanto, verifica-se que o também pode ser usado para construir árvores de decisão muito eficazes que não incluem todos os aspectos, utilizando-se da poda, ou seja, a redução das respostas baseadas em certos parâmetros a fim de que não se perca a acurácia da informação.

O modelo de regressão linear trabalha de forma a minimizar o erro quadrado das previsões como um todo e utiliza a métrica de Akaike, que é uma estatística utilizada para a escolha da especificação ótima de uma equação de regressão no caso de alternativas não aninhadas. O processo de geração dos resultados é realizado de forma iterativa no conjunto dos dados e o número de iterações necessárias é determinado por meio de validação cruzada, na qual o processo se encerra quando não se obtêm maior êxito na redução de erros.

Havendo um grande número de classes o resultado será amplo e, neste caso, pode ser usado um dos métodos de poda. Os autores afirmam que a operação de poda é muito importante para melhorar a visualização dos resultados produzindo árvores pequenas, porém muito precisas (WITTEN; FRANK, 2005).

2.3.1.4 Algoritmo de Associação

Os algoritmos de associação encontram regras, espacialmente relacionadas no banco de dados. Uma regra de associação tem o seguinte formato: A -> B (s%; c%), onde s é o suporte (a probabilidade, que A e B se mantenham unidos em relação a todos os casos possíveis) e c é a confiança (a probabilidade condicional de que B seja verdadeiro sob a condição de A) (REVATHI; SUMATHI, 2013).

A mineração de regras de associação encontra todas as regras existentes na base de dados que são satisfatórias ao suporte mínimo e também as restrições mínimas de confiança, enquanto que a mineração através de classificação tem como objetivo descobrir um pequeno conjunto de regras na base de dados que formem um classificador preciso. Outra diferença entre elas é que para a mineração de regras de associação, a meta da descoberta não é pré-determinada, enquanto que para mineração de regras de classificação existe um alvo pré-estabelecido(LIU; HSU; MA. 1998).

2.3.1.4.1 Algoritmo Apriori

O algoritmo de associação Apriori busca gerar todas as regras de associação entre os dados, com um suporte e confiança maior do que o mínimo especificado pelo usuário, chamados minsup e minconf, respectivamente (AGRAWAL; SRIKANT, 1994). O Apriori é um dos algoritmos de mineração de regras de associação mais influentes e sua regra é expressa pela frequência de itens (ZHONG, 2013).

Segundo He e Qi (2013), a idéia básica do algoritmo Apriori é: (a) Buscar, em primeiro lugar, todas as regras presentes no conjunto de dados, com um suporte mínimo pré-definido;(b) Geração de regras fortes de associação a partir do conjunto de dados, atendendo a um suporte e confiança mínimos.

Algoritmo Apriori é um tipo de algoritmo que procura conjuntos de itens frequentes e gera regras a partir de um suporte e confiança mínimos. O princípio é usar um método de iteração chamado pesquisa passo a passo que usa (k) - item a explorar e (k+1) - conjuntos de dados .

Algoritmo Apriori tem alto custo computacional, uma vez que necessita de múltiplas varreduras no banco de dados para que se consiga produzir as regras.

Para a geração das regras, após a verificação dos itens frequentes k em k+1, se chega aqueles que são candidatos as regras tipo C, que são comparadas com o banco de dados, gerando o subconjunto L, que é um subconjunto de C. Cada elemento do subconjunto C é verificado para ver se pode ou não pertencer ao subconjunto L. Por fim, cada regra gerada a partir da interação C-L tem seu suporte e confiança contado e verificado em relação ao que foi estabelecido previamente e, caso seja confirmado, a regra é disponibilizada (HE; QI, 2013).

3 METODOLOGIA

No presente capitulo, procura-se descrever a metodologia utilizada na pesquisa bem como os métodos empregados para o entendimento sistemático da construção de um conjunto inter-relacionado de variáveis e de como suas relações ajudam a explicar (ou prever) os fenômenos existentes (CRESWELL, 2007).

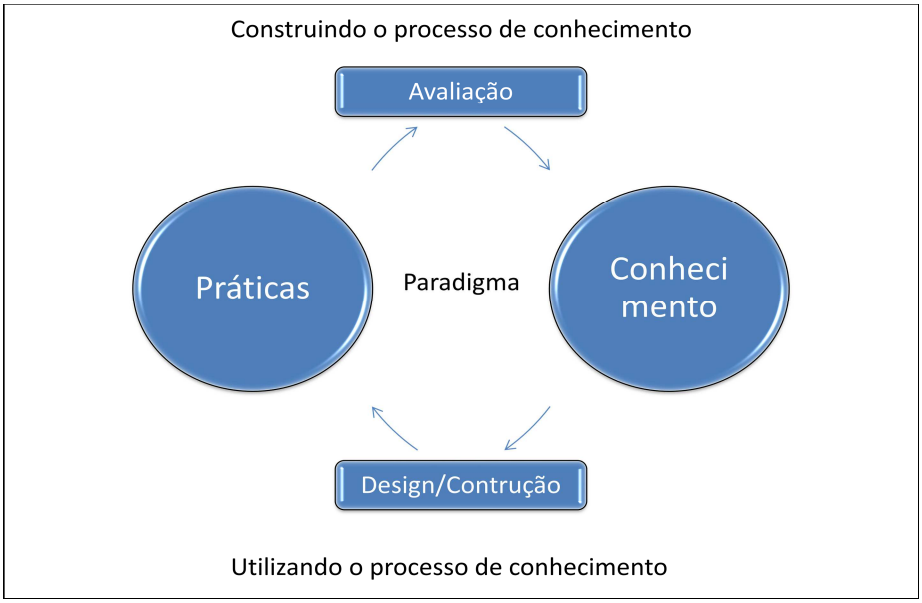
Segundo Hevner *et al*. (2004), existem dois paradigmas que caracterizam grande parte das pesquisas em Sistemas de Informação: Ciência comportamental e o *Design Science.* A primeira trata de desenvolver e verificar teorias que explicam ou predizem o comportamento humano e organizacional, enquanto que a segunda procura estender os limites das capacidades humanas e organizacionais, criando artefatos novos e inovadores. Os autores ainda afirmam que os sistemas de informação, desenvolvidos a partir do *Design Science*, são implementados em uma organização com o propósito de melhorar a eficácia e eficiência da mesma. O conhecimento e a compreensão do problema e a sua solução são obtidos na construção e aplicação do artefato projetado e o nível de implementação é atingido na medida em que os efeitos são alcançados.

Chakrabarti (2010) igualmente compreende a importância da eficiência e eficácia, mas complementa que este tipo de pesquisa permite na prática desenvolver produtos melhores sucedidos, dentro de uma visão que o conhecimento científico deve estar ligado à utilidade em relação ao propósito, ou seja, o trabalho deve ser acadêmico (possuir um problema a ser resolvido) e ser útil (relacionado com sua utilidade como finalidade).

No entendimento de Hevner *et al.* (2004), artefatos não são necessariamente algo materialmente concreto, mas podem ser constructos, símbolos, algoritmos, práticas ou metodologias que permitam aos pesquisadores resolverem um problema inerente ao desenvolvimento e execução de um sistema de informação com sucesso. No presente estudo, pretende-se desenvolver uma análise das relações de avaliação e gestão de riscos com processos tecnológicos, através de técnicas de garimpagem de dados, e desta forma gerar um artefato metodológico para análise das relações entre processos tecnológicos.

Na Figura 4, é apresentado um modelo geral para a geração e acumulação de conhecimento.

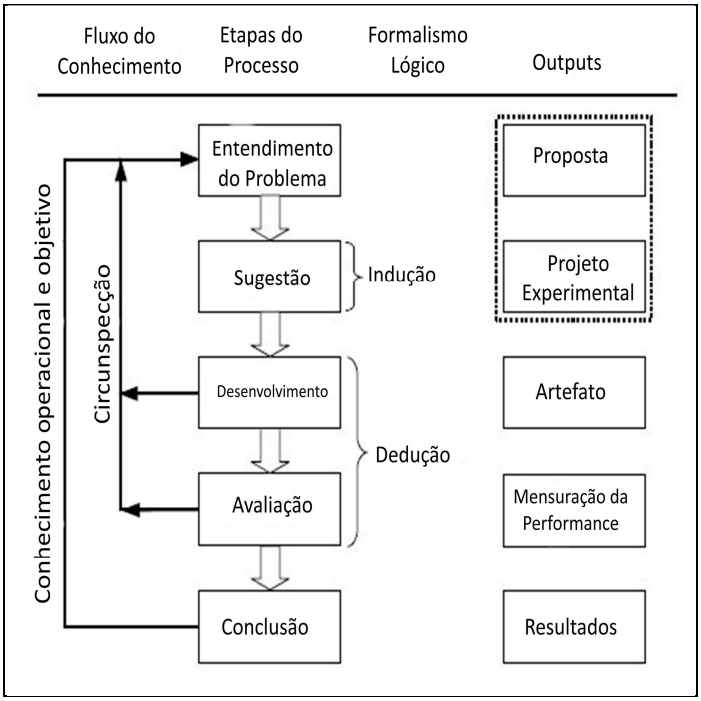
Figura 4: Processo de acúmulo de conhecimento



Fonte: Manson (2006)

A metodologia do *Design Research* (DR), foi proposta por Takeda *et al*. (1990) e aprimorada por Manson (2006) e contempla 5 etapas: Entendimento do Problema; Sugestão; Desenvolvimento; Avaliação e Conclusão, como que pode ser observado na Figura 5.

Figura 5: Modelo geral de DR



Fonte: Manson (2006, p.4)

Para o entendimento de cada uma destas etapas do DR (Figura 5), desenvolvidas a partir da visão Takeda *et al.* (1990) e Manson (2006), pode ser verificada na sequência:

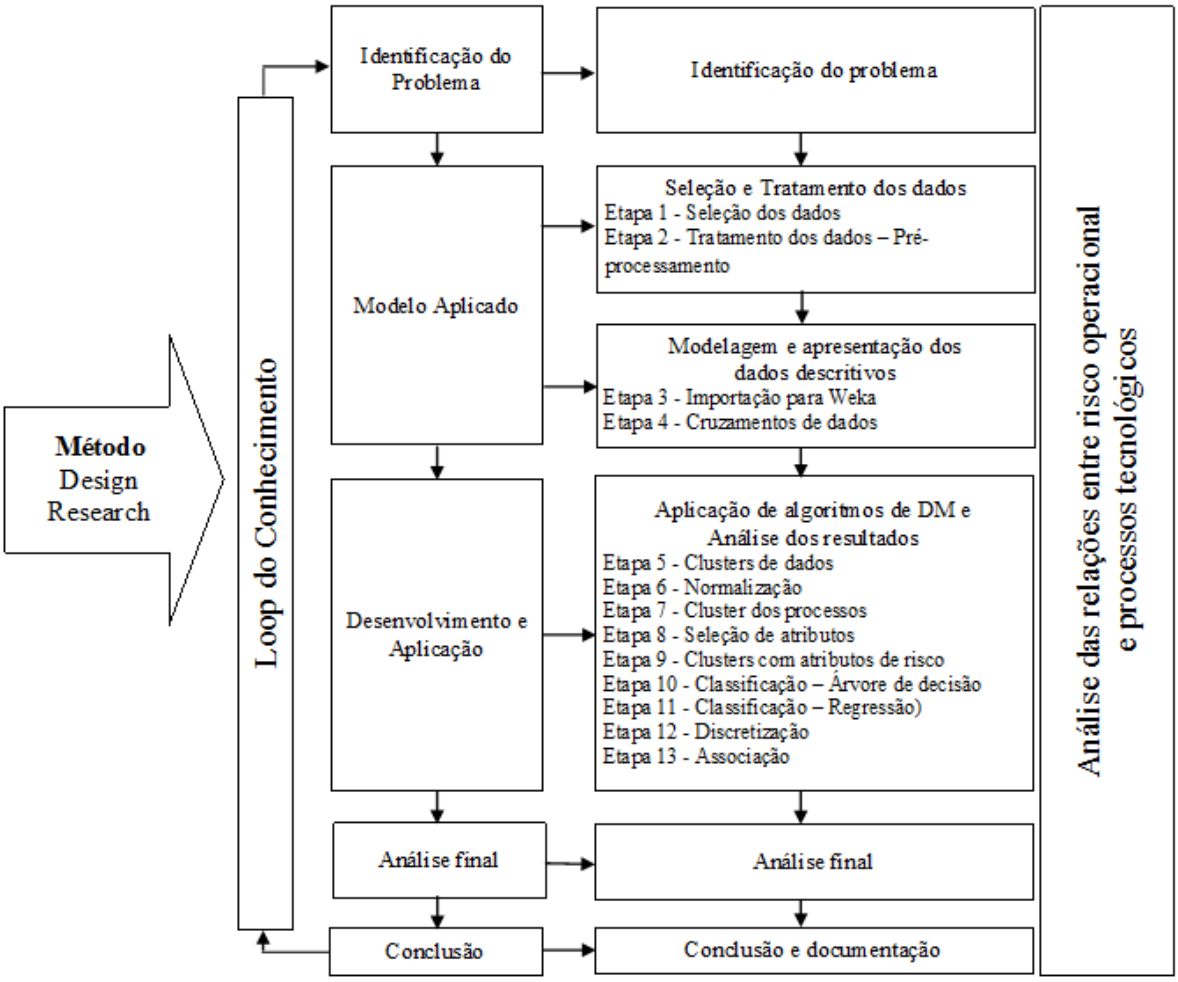
- **Entendimento do problema:** É a primeira etapa do processo da pesquisa e se inicia quando o pesquisador torna-se ciente do problema, que pode surgir de diversas fontes, dentre elas, necessidades da indústria, do governo ou de pesquisas relacionadas. Como resultado desta etapa, o pesquisador deve gerar uma proposta formal ou informal para iniciar um estudo.
- **Sugestão:** Nesta etapa, o pesquisador faz experimentos ou “tentativas” de design, como proposta preliminar. Esta etapa é essencialmente criativa, e é aonde diferentes pesquisadores chegam em diferentes projetos experimentais.
- **Desenvolvimento:** Durante esta etapa, o pesquisador irá desenvolver um ou mais modelos; as técnicas utilizadas irão variar muito, dependendo dos artefatos a serem construídos. Alguns exemplos de artefatos são algoritmos como uma prova formal, um software ou sistemas . A construção em si não requer nenhuma novidade, pois sua concepção já é a própria inovação.
- **Avaliação:** Uma vez desenvolvido o projeto, ele deve ser avaliado em função dos critérios que estão implícita ou explicitamente contidos na proposta. Quaisquer desvios das expectativas, antes e durante o desenvolvimento, devem ser minuciosamente detalhados, buscando formular explicações sobre estas disparidades.
- **Conclusão:** Em algum momento, após os *loops* de circunscrição e feedback de melhorias, a análise dos dados pode ser considerado "bom o suficiente”, assim se conclui o processo e os resultados são escritos e consolidados (loop do conhecimento operacional e objetivo). Os conhecimentos produzidos são classificados em dois tipos, os sólidos e os *loose-ends*. Os primeiros são fatos que foram aprendidos e que podem ser aplicados repetidamente, já o outro trata de anomalias que não podem ser explicadas e, frequentemente, tornam-se objetos de outras investigações.

Hevnet et at. (2004) ainda argumentam que para o desenvolvimento de pesquisas no paradigma do DR, devem ser observados os aspectos da relevância e do rigor cientifico. A relevância trata propriamente da utilidade que seja gerada como inovação para a ciência e para a sociedade . Já o rigor é atingido de forma adequada na aplicação da fundamentação e das metodologias existentes. Na ciência comportamental, as metodologias são tipicamente enraizadas na coleta de dados e nas técnicas de análise empírica. No DR, métodos computacionais e matemáticos são utilizados para avaliar a qualidade e a eficácia das análises, porém, técnicas empíricas também podem ser empregadas.

Pode-se dizer que o resultado buscado pelo presente estudo é uma análise das relações entre avaliação e gerenciamento de risco operacional e processos tecnológicos com o uso de regras automáticas de garimpagem de dados bem como a geração do artefato, desenvolvido a partir da metodologia de DR, permitindo uma melhor verificação das relações entre os processos e informações para possibilitar melhores decisões por parte dos gestores. Desta forma, mitigando riscos e otimizando o planejamento de investimentos, os resultados acabarão refletindo nos custos. uma vez que quando se investe em processos que possuem maior relação, maiores são as possibilidades de se atingirem e manterem os objetivos relacionados ao nível de maturidade desejados, reduzindo a necessidade de outros tipos de investimentos desnecessários.

Para o presente estudo foi desenvolvido um framework, apresentado na Figura 6, com o intuído de identificar as relações entre de risco operacional e processos tecnológicos através de processo de mineração de dados que se pretende trabalhar à luz da metodologia de *Design Research*.

Figura 6: Framework da pesquisa



Fonte: Elaborado pelo autor

Para o desenvolvimento deste estudo, foi definido o *Design Research* como método de pesquisa a ser seguida. Esse é desenvolvido através de passos metodológicos com o intuito de atingir o objetivo de pesquisa e possuiu como base a metodologia geral do DR proposta por Takeda (1990), Hevner *et al.* (2004) e Manson (2006). Porém, para evidenciar mais especificamente o processo de Data Mining relacionado a esta pesquisa, contemplou-se as cinco etapas de DR da seguinte maneira:

1. Identificação do problema;

2. Modelo aplicado envolvendo: (a) Seleção de dados e tratamento de dados, (b) Modelagem e apresentação dos dados descritivos.

3. Desenvolvimento e aplicação: (a) Aplicação de algoritmos de DM, (b) Análise dos resultados

4. Análise final: Análise final

5. Conclusão: Conclusão e documentação.

Seguindo a especificidade do DM dentro do DR, as próximas etapas se operacionalizam da maneira apresentada na continuação.

3.1 Identificação do problema

O presente estudo trata da análise das relações entre a avaliação e o gerenciamento de riscos operacionais e processos tecnológicos, que remetem ao problema de pesquisa anteriormente apresentado. Objetiva-se desenvolver uma análise com o intuito de analisar as relações entre gestão de riscos operacionais e processos tecnológicos, através de mineração de dados.

O uso de técnicas sofisticadas de mineração de dados permite encontrar padrões nas relações entre os processos estudados, cujo principal benefício é auxiliar na tomada de decisões a fim de mitigar riscos bem como ser fonte de informação para o melhor planejamento do desenvolvimento e dos investimentos. Portanto, isto reflete em uma tomada de decisão mais precisa e eficaz, reduzindo a subjetividade na melhoria dos processos tecnológicos.

3.2 Seleção e tratamento dos dados

A seleção ou coleta de dados ocorreu por meio de questionário elaborado envolvendo a metodologia de governança Cobit 4.1, que contempla 34 processos através da avaliação dos níveis de maturidade de cada um destes processos, e aplicado a funcionários de 140 empresas localizadas principalmente no estado do Rio Grande do Sul. Os dados utilizados neste estudo foram coletados e tratados nos anos de 2012 e 2013.

Os dados coletados são compostos por 42 duas variáveis (atributos) e destas, 34 referem-se a processos tecnológicos do modelo de governança Cobit 4.1 e 8 fazem referência a dados descritivos dos respondentes e das empresas onde atuam. O questionário foi disponibilizado através de formulário eletrônico na plataforma

Google Doc’s, que é uma ferramenta conhecida dos respondentes, permitindo agilidade e confiabilidade na coleta de dados.

O tratamento dos dados também conta com as etapas de limpeza e transformação dos dados. A limpeza visa eliminar as respostas inválidas para que se produzam corretas análises e a transformação deles tem por finalidade deixá-los no formato necessário para a aplicação dos algoritmos de mineração de dados, bem como para a geração de novas classes a partir de metadados. Na pesquisa foi desenvolvido o atributo região a partir do atributo cidade.

Nos atributos que contém os 34 processos do Cobit 4.1 de forma nominal, com respostas entre inexistente e otimizado, e estas foram transformadas para valores numéricos, conforme previsto pela metodologia e representadas entre 0 e 5, sendo 0 inexistente e 5 otimizado, padronizando os valores na aplicação de regras de mineração das informações colhidas., Além disso, uma parte significativa dos algoritmos presentes no estudo exige variáveis numéricas para sua operacionalização.

Nesta etapa é apresentada a forma como foi realizada a coleta e seleção dos dados do estudo, bem como seu tratamento, através da limpeza e transformação das informações para o desenvolvimento da aplicação dos algoritmos de mineração de dados.

3.2.1 Etapa 1: Seleção dos dados

Os dados da pesquisa foram selecionados a partir da metodologia de governança Cobit 4.1, a qual se encontra consolidada nos meios acadêmicos e profissionais, com validação do ITGI e ISACA.

A coleta dos dados deu-se ao longo dos anos de 2012 e 2013 e os respondentes são colaboradores das empresas pesquisadas, possuindo acesso e familiarizados com os processos de suas organizações. As respostas foram inseridas pelos próprios respondentes em um formulário de pesquisa eletrônico Google Doc’s disponível na internet de forma gratuita e previamente disponibilizado a eles.

Os tipos de dados coletados compreendem em sua totalidade o Quadro 8 com 42 variáveis, no entanto nem todas são úteis para a análise de dados ou, foram

removidas na etapa de limpeza dos dados para que não seja possível a identificação dos respondentes. Estas atividades de transformação e limpeza foram desenvolvidas na fase de pré-processamento dos dados.

Quadro 8: Variáveis coletadas

	Variáveis
1	Indicação de data e hora
2	Empresa
3	Setor de atividade
4	Porte da empresa
5	Departamento/Setor
6	Cidade
7	Nome do respondente
8	Cargo do respondente
9	PO1 - Define o planejamento estratégico de TI.
10	PO2 - Define a arquitetura da informação.
11	PO3 - Determina as Diretrizes da Tecnologia.
12	PO4 - Define a organização de TI e seus relacionamentos.
13	PO5 - Gerencia o Investimento de TI.
14	PO6 - Comunica as metas e diretrizes gerenciais.
15	PO7 - Gerencia os recursos humanos de TI.
16	PO8 - Gerencia a qualidade.
17	PO9 - Avalia e gerencia os riscos.
18	PO10 - Gerencia os projetos.
19	AI1 - Identifica soluções de automação.
20	AI2 - Adquiri e Mantém Software Aplicativo.
21	AI3 - Adquire e mantém a arquitetura tecnológica.
22	AI4 - Desenvolve e mantém procedimentos de TI.
23	AI5 - Obtém recursos de TI.
24	AI6 - Gerenciar mudanças
25	AI7 - Instala e certifica soluções e mudanças.
26	DS1 - Define níveis e mantém os acordos de níveis de serviços.
27	DS2 - Gerencia os serviços de terceiros.
28	DS3 - Gerenciar desempenho e capacidade da TI.
29	DS4 - Garante a continuidade dos serviços.
30	DS5 - Garante a segurança dos sistemas.
31	DS6 - Identifica e Aloca Custos
32	DS7 - Educa e treina os usuários.
33	DS8 - Gerencia a central de serviços e incidentes.
34	DS9 - Gerencia a configuração.
35	DS10 - Gerencia os problemas.
36	DS11 - Gerencia os dados.
37	DS12 - Gerencia a infra-estrutura.
38	DS13 - Gerenciar operações.
39	MO1 - Monitora e avalia a desempenho de TI.
40	MO2 - Monitora e avalia o controle interno.
41	MO3 - Assegura a conformidade aos requisitos externos.
42	MO4 - Fornecer governança de TI.

Fonte: Dados da pesquisa

O modelo de governança Cobit 4.1 é constituído por 4 domínios divididos em 34 processos, e cada uma das respostas das variáveis relativas aos processos do Cobit (variáveis 9 a 42 do Quadro 8), constituem de uma única resposta compreendida em um dos níveis de maturidade dos processos. As respostas possíveis para cada uma dessas 34 variáveis e seu significado podem ser vistos no Quadro 9.

Quadro 9: Níveis de maturidade em processo de TI

	Nível de Maturidade	**Descrição**
0	Inexistente	Completa falta de um processo reconhecido. A empresa nem mesmo reconheceu que existe uma questão a ser trabalhada.
1	Inicial / *Ad hoc*	Existem evidências que a empresa reconheceu que existem questões e que precisam ser trabalhadas. No entanto, não existe processo padronizado; ao contrário, existem enfoques *Ad Hoc* que tendem a ser aplicados individualmente ou caso-a-caso. O enfoque geral de gerenciamento é desorganizado.
2	Repetível	Os processos evoluíram para um estágio onde procedimentos similares são seguidos por diferentes pessoas fazendo a mesma tarefa. Não existe um treinamento formal ou uma comunicação dos procedimentos padronizados e a responsabilidade é deixada com o indivíduo. Há um alto grau de confiança no conhecimento dos indivíduos e consequentemente erros podem ocorrer.
3	Processo Definido	Procedimentos foram padronizados, documentados e comunicados através de treinamento. É imprescindível que esses processos sejam seguidos; no entanto, possivelmente desvios não serão detectados. Os procedimentos não são sofisticados, mas existe a formalização das práticas existentes.
4	Gerenciado e Mensurável	A gerência monitora e mede a aderência aos procedimentos e adota ações onde os processos parecem não estar funcionando muito bem. Os processos estão debaixo de um constante aprimoramento e fornecem boas práticas. Automação e ferramentas são utilizadas de uma maneira limitada ou fragmentada.
5	Otimizado	Os processos foram refinados a um nível de boas práticas, baseado no resultado de um contínuo aprimoramento e modelagem da maturidade como outras organizações. TI é utilizada como um caminho integrado para automatizar o fluxo de trabalho, provendo ferramentas para aprimorar a qualidade e efetividade, tornando a organização rápida em adaptar-se.

Fonte: Adaptado de ISACA (2014)

Os dados da pesquisa possuem um bom nível de respostas válidas, ou seja, em sua maioria foram respondidas de forma completa, permitindo um maior aproveitamento. No entanto houveram respostas que necessitaram ser eliminadas e este procedimento está explicado na etapa 2, no item limpeza.

3.2.2 Etapa 2: Tratamento dos dados – Pré-processamento

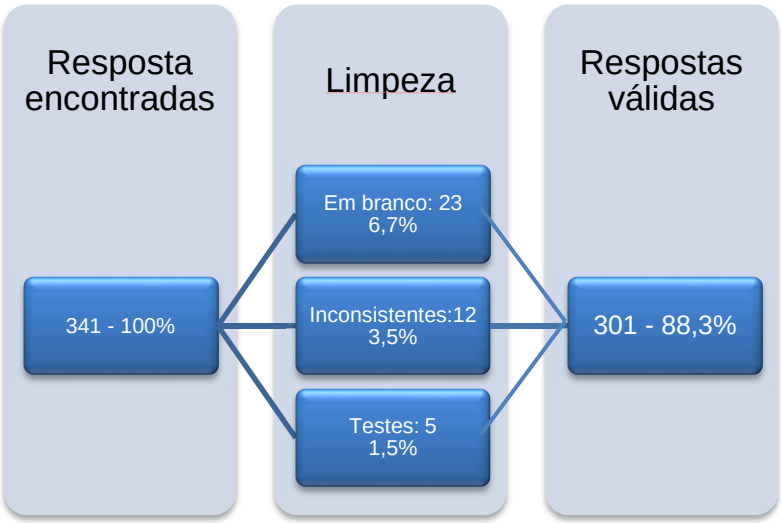
Nesta etapa foi realizado todo o tratamento dos dados, sua limpeza e transformação, para que o processo de *Knowledge Discovery in Database* (KDD) tenha o maior aproveitamento possível dos dados disponíveis. Também são apresentados os gráficos de aspectos descritivos do banco de dados.

3.2.2.1 Limpeza de instâncias

Foram coletadas 341 respostas, também chamadas de instâncias no processo de mineração de dados, porém nem todas são válidas, o que demandou uma limpeza sobre os aspectos não válidos, como respostas que eram testes de consistência do banco de dados e, que foram desenvolvidas inicialmente para verificar a funcionalidade do banco de dados. Após o processo de limpeza das respostas, restaram 301 (88,3%) válidas e estas, multiplicadas pelas 42 variáveis/classes, totalizaram 12.642 dados.

As respostas que apresentavam em todos os processos o mesmo nível de maturidade também foram excluídas. Essa constância ficou evidenciada nos casos de empresas que do primeiro ao último processo atribuíram o mesmo nível de maturidade, não existindo variabilidade entre os processos, o que tende a ser um indicativo de que o respondente não verificou com cuidado as questões. Devido ao risco de distorção nas respostas, optou-se preventivamente em excluir os registros que possuíam todos os processos com o mesmo nível de maturidade. Os registros que estavam completamente em branco também foram excluídos do banco de dados. Na Figura 8, são apresentadas as quantidades de respostas coletadas para a pesquisa e os totais da limpeza de dados.

Figura 7: Quantidades de respostas coletadas e válidas



Fonte: Dados da pesquisa

Após a transformação dos dados, foi realizado outro processo de limpeza, porém diferente deste, devido que se refere aos atributos/classes.

O próximo passo para a realização das tarefas de mineração de dados e de transformação dos mesmos, foi deixá-los padronizados com o objetivo de permitir a aplicação das atividades desejadas.

3.2.2.2 Transformação

A transformação de dados preparou as informações no formato aceito pelo programa de análise e permitiu a geração de novos dados a partir dos existentes com o objetivo de proporcionar melhores análises.

Segundo Goldschimidt e Passos (2005), em sua maioria, os métodos de mineração de dados pressupõem a organização dos dados em uma tabela, normalmente grande e bidimensional, o que ocorre com o software Weka utilizado neste estudo. Para extração das informações do banco de dados original para a tabela, o autor afirma que existem duas formas. A primeira através de junção direta, onde se colocam os registros na tabela sem uma análise crítica sobre a contribuição da variável para o processo. A segunda, por junção indireta, seleciona os atributos que possuem maior potencial de contribuição para o processo. A junção indireta foi a forma utilizada no presente estudo e por esta razão, a transformação e limpeza dos dados se tornaram importantes para a geração de melhores resultados.

Devido à necessidade de se responder a questão problema, optou-se por trabalhar as respostas por empresas e não por respondente. Para isso, foram geradas médias das respostas, uma vez que havia mais de um respondente por empresa, nos processos das empresas entre os respondentes e esta média resultou no nível de maturidade da empresa

Para possibilitar a geração de médias, é necessário que as variáveis estejam em formato numérico e não de forma nominal e para tal, a primeira tarefa de transformação dos dados foi a substituição dos níveis de maturidade (nominal) por seu correspondente numérico, sendo o nível com nota mais baixa o inexistente (0) e o mais alto otimizado (5) conforme apresentado no Quadro 10.

Quadro 10: Transformação níveis de maturidade para variável numérica

Nominal	Numérico
Inexistente	0
Inicial	1
Repetitivo	2
Definido	3
Gerenciado	4
Otimizado	5

Fonte: Isaca (2014)

Depois de substituídas por sua representação numérica, as respostas foram agrupadas por empresa e gerada a média para cada um dos 34 processos em uma nova instância. Nesta nova instância foram colocados os demais aspectos da empresa.

Nos títulos de cada atributo foi substituído o nome completo por sua respectiva sigla, para facilitar a visualização dos resultados obtidos. Exemplo: “PO1 - Define o planejamento estratégico de TI” por somente “PO1”. Isto ocorreu para cada um dos 34 processos do Cobit.

Quanto ao atributo setor, foi padronizado através do uso do critério de setores da Bovespa, com exceção de militar e público por não estarem contemplados na mesma.

No atributo cidade houve necessidade de padronização do termo usado, porque alguns respondentes usaram siglas de cidades e outros usaram o nome completo da cidade. Foram transformadas as cidades para siglas, sendo que as maiores respostas foram Porto Alegre (POA), Novo Hamburgo (NH) e São Leopoldo (SL).

O atributo empresa, a qual originalmente continha o nome da empresa foi substituído através de uma codificação por uma ordem sequencial, totalizando 140 empresas a fim de não serem identificadas individualmente e desta forma foi mantido o sigilo estabelecido para o estudo.

Os cargos dos respondentes também foram padronizados, porém como o objetivo da pesquisa está vinculado às médias das empresas, a característica da resposta não permitia gerar uma média, bem como também não se poderia assumir algum cargo predominante entre os respondentes, esta informação foi descartada na segunda limpeza deste atributo.

Ocorreu a geração de novos atributos através de metadados, a partir do atributo cidade, que gerou o atributo região com o objetivo de verificar as principais características dos níveis de maturidade em empresas de cada região. Foram criadas três categorias que eram as com maior frequência entre as respostas: Capital (Porto Alegre), Vale do Rio dos Sinos composto pelas cidades do Vale do Rio dos Sinos e Interior, com as demais cidades do interior do estado do Rio Grande do Sul. Existem somente sete empresas de fora do estado, que representam 5% das respostas, e por sua baixa frequência optou-se por não criar uma nova categoria de região para a análise dos dados.

Nos registros que não possuíam um valor atribuído, foi colocado o símbolo de interrogação, “?”, que é o formato que o software entende que neste registro não existe informação disponível. De outra forma, ocorre erro na importação ou atribuição de valor do próximo registro neste, o que iria desconfigurar a correta importação dos dados e poderia gerar resultados equivocados.

A partir dos dados transformados e organizados, se realizou uma nova limpeza de dados que não foram úteis para processamento, apresentado na sequencia.

3.2.2.3 Limpeza de atributos/classes e informações incompatíveis

Após a transformação dos dados e geração de médias dos processos por empresa, foram excluídas todas as instâncias (linhas) que continham dados individuais das respostas e mantidas as instâncias com as médias.

Os atributos excluídos por não auxiliarem a trazer melhores resultados para o processo ou por sigilo nas informações foram: Indicação de data e hora da resposta (irrelevante), nome da empresa (sigilo e irrelevante), nome do respondente (sigilo e irrelevante), departamento do respondente (não foi possível gerar uma padrão para a empresa, pois havia casos de mais de um respondente por empresa) e cargo do respondente (não foi possível gerar um padrão para a empresa, pois havia casos de mais de um respondente por empresa). Após a limpeza, transformação e nova limpeza dos dados, foram contabilizados 39 atributos e 140 instâncias, gerando um total de 5.460 dados.

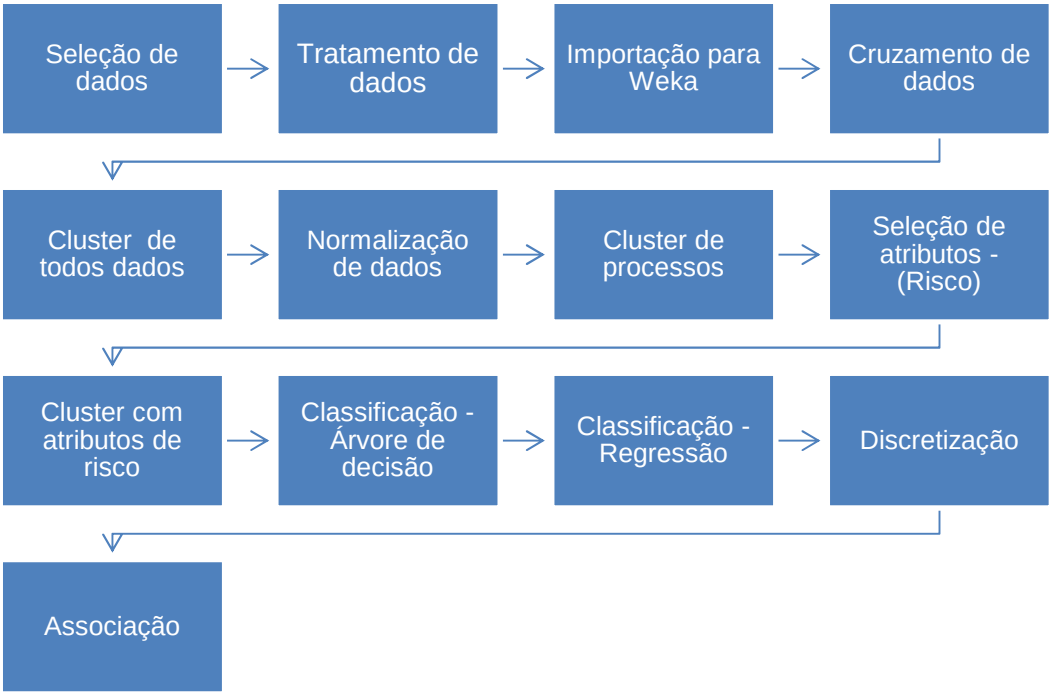
Foram encontrados registros sem informações, ocorridos por erro na entrada de dados ou pela real inexistência de informação. Para tratar destas incompatibilidades, Larose (2005) afirma que é possível ignorar os valores ausentes, filtrar e eliminar as instâncias com valores ausentes ou prever os valores baseados em outros atributos avaliando especificamente cada caso. Devido a baixa quantidade de registros sem valor encontrados, 21 em 5.460, optou-se por ignorar os valores ausentes, pois os mesmos não iriam afetar nas análises a serem desenvolvidas.

A próxima etapa trata da modelagem, onde foi realizado o planejamento dos demais processos e a apresentação dos dados descritivos.

3.3 Modelagem e apresentação de dados descritivos

O processo de modelagem compreende o planejamento das demais etapas para análise dos dados a partir do objetivo proposto. Para tanto foi desenvolvido um modelo de aplicação para mineração dos dados coletados, que pode ser visto na Figura 7.

Figura 8: Modelagem de pesquisa



Fonte: Elaborado pelo autor

Conforme visualizado na Figura 7, após a seleção e tratamento dos dados, são aplicadas técnicas de mineração de dados - geração de clusters, classificação, seleção de atributos e associação. Primeiramente, para todas as classes e em seguida direcionando para as classes que possuem relação com o processo PO9 (Avalia e gerencia riscos) para que finalmente seja possível analisar os dados.

Quanto às variáveis, o Quadro 11 diz respeito as 39 variáveis utilizadas na pesquisa. Foi gerado um quadro onde pode se observar o nome da variável, sua descrição, o tipo de variável e os possíveis valores a serem encontrados. As variáveis ligadas aos processos de TI (Cobit 4.1), desde PO1 a MA4 possuem um escala entre 0 e 5, sendo 0 - Inexistente e 5 – Otimizado, no entanto na coluna de valores foi atribuído somente para o processo PO1 e nos demais “Idem” pois se referem igualmente aos valores presentes em PO1.

Quadro 11: Variáveis coletadas

Váriavel	Descrição	Tipo de variável	Valores
Empresa	Número da empresa.	Numérico sequencial	140 diferentes empresas
Setor	Setor no qual a empresa esta inserida.	Categorico	24 diferentes setores
Porte da empresa	Setor no qual a empresa esta inserida.	Categórico	Grande; Médio; Pequeno; Micro
Cidade	Cidade da empresa.	Categórico	30 diferentes cidades
Região	Região onde esta localizada a empresa.	Categórico	Capital; Vale do Rio dos Sinos; Interior
PO1	Define o planejamento estratégico de TI.	Numérico escalar	0 - Inexistente 1 - inicial 2 - Repetitivo 3 - Definido 4 - Gerenciado 5 - Otimizado
PO2	Define a arquitetura da informação.	Numérico escalar	Idem
PO3	Determina as Diretrizes da Tecnologia.	Numérico escalar	Idem
PO4	Define a organização de TI e seus relacionamentos.	Numérico escalar	Idem
PO5	Gerencia o Investimento de TI.	Numérico escalar	Idem
PO6	Comunica as metas e diretrizes gerenciais.	Numérico escalar	Idem
PO7	Gerencia os recursos humanos de TI.	Numérico escalar	Idem
PO8	Gerencia a qualidade.	Numérico escalar	Idem
PO9	Avalia e gerencia os riscos.	Numérico escalar	Idem
PO10	Gerencia os projetos.	Numérico escalar	Idem
AI1	Identifica soluções de automação.	Numérico escalar	Idem
AI2	Adquiri e Mantém Software Aplicativo.	Numérico escalar	Idem
AI3	Adquire e mantém a arquitetura tecnológica.	Numérico escalar	Idem
AI4	Desenvolve e mantém procedimentos de TI.	Numérico escalar	Idem
AI5	Obtém recursos de TI.	Numérico escalar	Idem
AI6	Gerenciar mudanças	Numérico escalar	Idem
AI7	Instala e certifica soluções e mudanças.	Numérico escalar	Idem
ES1	Define níveis e mantém os acordos de níveis de serviços.	Numérico escalar	Idem
ES2	Gerencia os serviços de terceiros.	Numérico escalar	Idem
ES3	Gerenciar desempenho e capacidade da TI.	Numérico escalar	Idem
ES4	Garante a continuidade dos serviços.	Numérico escalar	Idem
ES5	Garante a segurança dos sistemas.	Numérico escalar	Idem
ES6	Identifica e Aloca Custos	Numérico escalar	Idem
ES7	Educa e treina os usuários.	Numérico escalar	Idem
ES8	Gerencia a central de serviços e incidentes.	Numérico escalar	Idem
ES9	Gerencia a configuração.	Numérico escalar	Idem
ES10	Gerencia os problemas.	Numérico escalar	Idem
ES11	Gerencia os dados.	Numérico escalar	Idem
ES12	Gerencia a infra-estrutura.	Numérico escalar	Idem
ES13	Gerenciar operações.	Numérico escalar	Idem
MA1	MAnitora e avalia a desempenho de TI.	Numérico escalar	Idem
MA2	MAnitora e avalia o controle interno.	Numérico escalar	Idem
MA3	Assegura a conformidade aos requisitos externos.	Numérico escalar	Idem
MA4	Fornecer governança de TI.	Numérico escalar	Idem

Fonte: Dados da pesquisa

Após a geração do quadro sobre as variáveis foram gerados gráficos descritivos sobre a composição dos dados, que serão cruzados com as demais variáveis, presente no capitulo 4 desta pesquisa.

3.4 Aplicação de algoritmos de mineração de dados e análise de resultados

Para realização da análise dos dados após o tratamento que consiste principalmente na seleção, limpeza transformação e importação dos dados no software Weka de mineração de dados, foram aplicados algoritmos de geração de clusters que permitem buscar agrupamentos maximizando a procura por similaridades entre os elementos e diferenças entre os grupos. Também foram utilizados algoritmos de associação e classificação para encontrar padrões nas relações entre processos de gestão de risco operacional e processos tecnológicos originando regras a partir destas.

A pesquisa utilizou-se de sofisticadas técnicas de mineração de dados, cuja validade comprovada pretende auxiliar na busca automática de relações entre risco operacional e processos tecnológicos, minimizando a subjetividade normalmente presente na avaliação e gerenciamento de riscos. Busca-se a identificação de padrões de agrupamentos e associações entre os processos tecnológicos envolvidos na avaliação e gestão de riscos operacionais, a fim de permitir um melhor planejamento estratégico e operacional, que poderá ser traduzido num melhor nível de governança de tecnologia da informação.

A pesquisa foi realizada com o uso do software *Waikato Environment for Knowledge Analysis (*Weka), desenvolvido a partir de 1993, através de um projeto de pesquisa da Universidade de Waikato, Nova Zelândia e financiada pela *New Zealand Government's Foundation for Research, Science and Technology*, órgão governamental com o intuito de agregar diversos algoritmos com abordagens distintas na área da inteligência artificial conhecida como Aprendizagem de Máquina, através do uso de técnicas de *Data Mining*. O software foi desenvolvido na linguagem de programação JAVA e está licenciado sob a *General Public License*, permitindo aos usuários e pesquisadores utilizar, estudar e inclusive alterar o código fonte (TAN; STEINBACH; KUMAR, 2009).

Após a aplicação de algoritmos de mineração de dados e análise de resultados, se retoma os aspectos com maior relevância no capitulo de análise final.

3.5 Análise final e documentação

Fase na qual se retoma os principais aspectos encontrados na pesquisa. A documentação trata do documento final, que no presente caso, foi a entrega da pesquisa, realizada no formato de dissertação para o Programa de Pós-Graduação em Ciências Contábeis da Universidade do Vale do Rio dos Sinos.

3.6 Conclusão

Na conclusão, utilizando-se de informações provenientes da base de dados e das análises, demonstra-se a relevância do estudo, ou seja, o novo conhecimento gerado a partir das verificações e do processo de garimpagem de dados.

4. RELAÇÃO DE RISCOS OPERACIONAIS E PROCESSOS TECNOLÓGICOS

A análise da relação entre risco operacional e processos tecnológicos ocorreu através de um ciclo de desenvolvimento de *Design Research* e para tal envolveu as etapas apresentadas na metodologia e as demais na sequência.

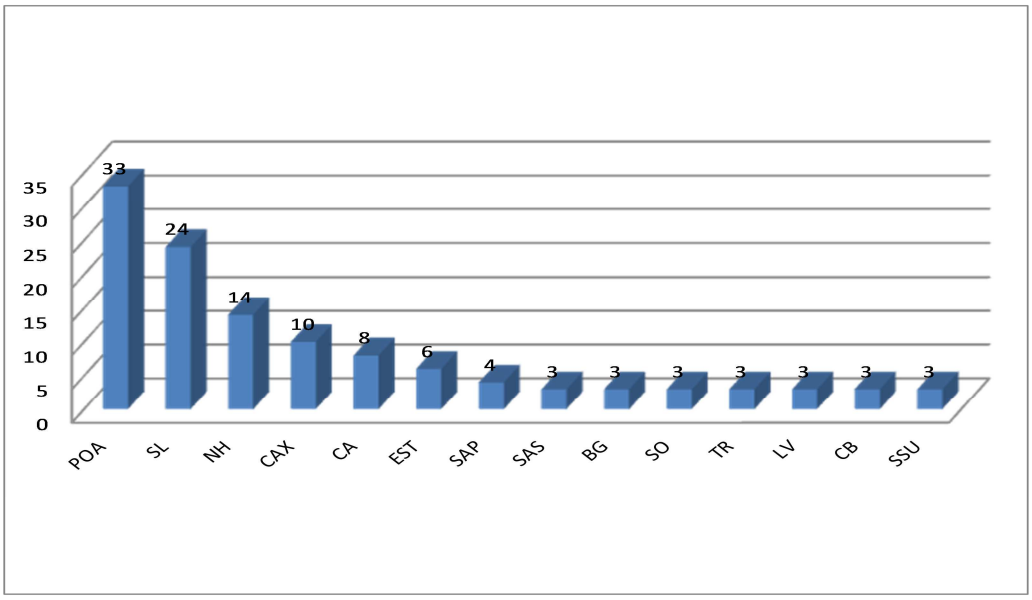
4.1 Modelagem e apresentação de dados descritivos

Neste capítulo são apresentados os dados descritivos das empresas pesquisadas bem como as etapas 3 - Importação de dados para o Weka e 4 – Cruzamento de dados.

Quanto à modelagem, foi apresentado na metodologia da pesquisa, e se refere às etapas do processo de mineração dos dados para se responder aos objetivos. A contagem dos atributos dos dados gerou gráficos descritivos para uma maior aproximação com o tema tratado e entendimento do contexto das empresas pesquisadas bem como permitiu fazer as primeiras inferências, de forma mais descritiva.

No Gráfico 1, é possível observar as cidades maior número de empresas presentes nas respostas.

Gráfico 1: Cidades das empresas



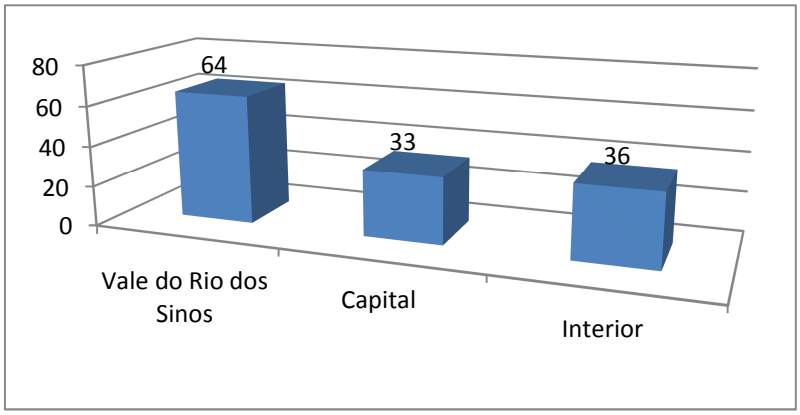
Fonte: Dados da pesquisa

A partir do Gráfico 1, pode-se verificar que as cidades de Porto Alegre (POA), São Leopoldo (SL) e Novo Hamburgo (NH) representam em conjunto (50,7%) a maior parte das localidades das empresas.

As cidades com 6 empresas ou menos representam 36,4% do total de respostas. As cidades que não estão presentes no gráfico, são as que possuíam apenas uma ou duas empresas. compreendendo 4 cidades com 2 empresas e 12 cidades com uma única empresa.

No Gráfico 2 é apresentada a frequência das empresas distribuídas em três regiões: Capital, Vale do Rio dos Sinos e Interior.

Gráfico 2: Empresas por região



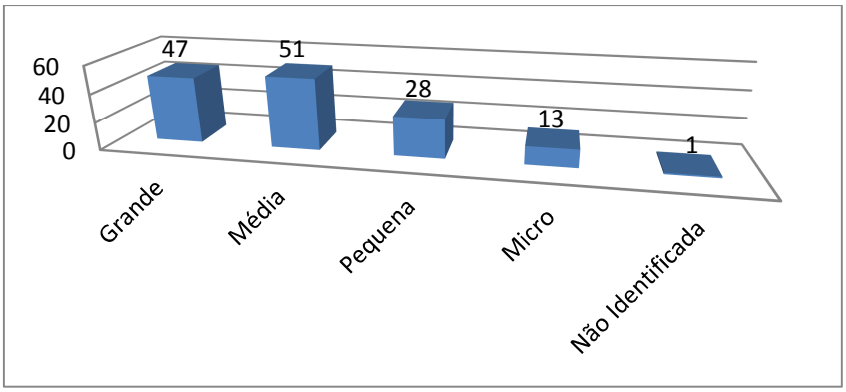
Fonte: Dados da pesquisa

As empresas do Vale do Rio dos Sinos, conforme apresentado no Gráfico 2, representam, aproximadamente metade (45,7%) das empresas presentes no estudo.

Embora não esteja apresentado no gráfico de forma descritiva, as empresas de fora do estado do RS representam apenas 5% dos dados, que devido a baixa frequência não justifica a criação de uma categoria específica para análise de dados. No software de mineração de dados essa categoria não foi classificada em nenhum dos itens quanto à região.

Quanto ao porte, as empresas são divididas em 4 categorias e foram classificadas pelos respondentes conforme apresentado no Gráfico 3

Gráfico 3: Porte das empresas



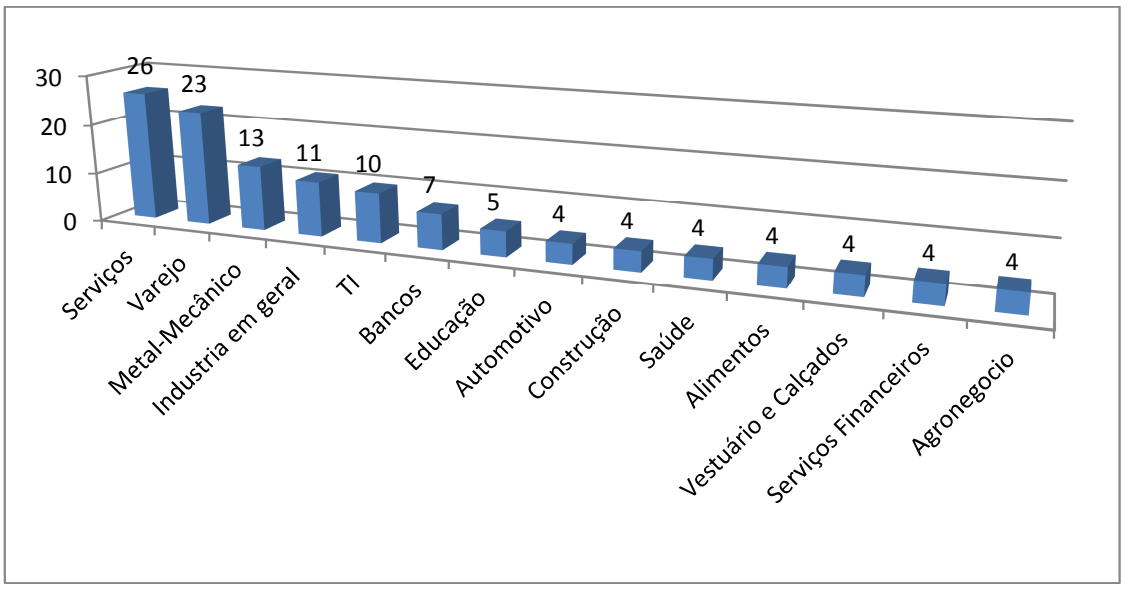
Fonte: Dados da pesquisa

Conforme o Gráfico 3, a categoria que apresentou maior frequência de respostas foram as empresas de médio porte (36,4%), seguidas pelas de grande porte (33,6%). A diferença entre elas é de apenas 4 respostas e em conjunto respondem por 70% das respostas.

Houve uma única empresa na qual não foi possível identificar o porte nem o nome, pois o respondente não assinalou estas opções, ficando fora da análise.

No que se refere ao setor, as empresas foram divididas conforme o critério de classificação da Bovespa, com exceção de duas organizações, uma militar e outra pública, em 25 diferentes setores. No Gráfico 4 pode-se visualizar os setores com mais de quatro empresas presentes no estudo.

Gráfico 4: Setores das empresas pesquisadas



Fonte: Dados da pesquisa

Conforme apresentado no Gráfico 4, os setores de serviços e varejo são os mais representativos e em conjunto respondem por 35% das respostas. Embora exista uma maior frequência de respostas presentes em dois setores específicos, nota-se uma diversificação das respostas nos mais diversos setores.

Embora não estejam representadas no Gráfico 4 devido a sua menor frequência, outras 11 empresas estão presentes no estudo, das quais 2 setores possuem três empresas, outros 2 setores possuem duas empresas e 7 setores possuem uma única empresa.

A próxima etapa trata da importação dos dados para o software Weka, no qual será possível realizar comparações e mineração de dados.

4.1.1 Etapa 3: Importação de dados para o Weka

Para poder trabalhar com os dados no Weka, foi necessário a importação dos dados para o programa. Embora o Weka seja um programa que importe os dados de tabelas bidimencionais, essas tabelas devem estar em um formato específico para que sejam carregadas corretamente.

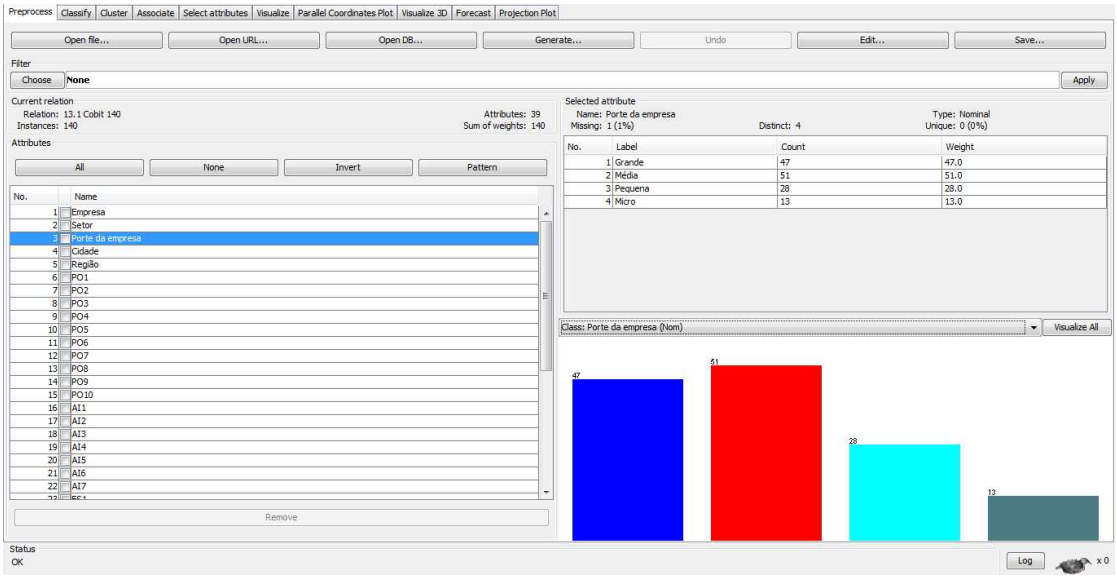
Um dos formatos aceitos pelo Weka é o “.csv” que pode ser gerado a partir do Excel, no entanto os dados do arquivo Excel (.xls) precisam ser transformados para que possam estar adequados aos critérios do Weka.

Para a organização do arquivo foram colocados os atributos na primeira linha do arquivo e cada uma das instâncias em uma nova linha, estando cada resposta na coluna de seu respectivo atributo. Após, foi salvo um novo arquivo com o formato “CSV separado por vírgulas”, que salva um “;” entre os registros e permite uma única pasta de trabalho.

Depois de salvo em “.csv” o arquivo foi aberto através do bloco de notas para serem realizadas as substituições necessárias. Primeiro se busca e substitui todas as vírgulas dos números por ponto, em função da notação decimal que o Weka entende ser diferente da aplicada no Brasil. Depois de realizada esta substituição, uma segunda também foi necessária, onde todos os “;” são substituídos por vírgula, forma na qual o software distingue os diferentes registros. Importante que seja realizada nesta ordem para que não ocorram problemas quanto à diferenciação entre a notação e a separação dos registros.

O próximo passo foi importar os dados através da opção “*open file...*” na qual selecionou-se o arquivo “.csv” preparado. Todos os formatos estavam de acordo com as especificações do Weka, e após um período de importação, os dados apareceram na tela principal do software, que pode ser vista na Figura 9.

Figura 9: Interface gráfica do Weka na importação de dados



Fonte: Dados da pesquisa

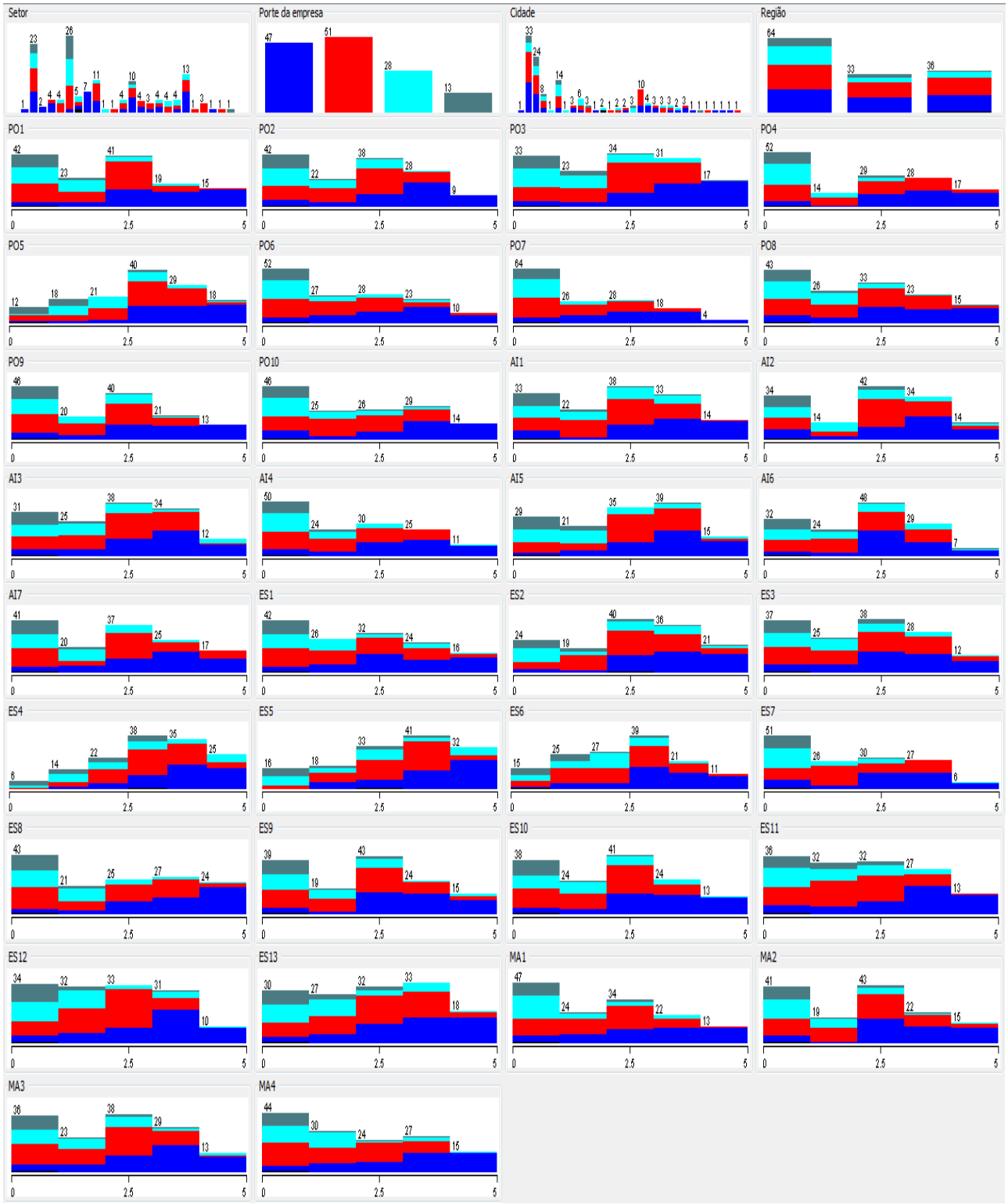
Na sequência foi gravado um arquivo no formato do software Weka (.arff), através da opção “*Save*...” presente na tela principal. A partir deste momento as tarefas de mineração de dados foram realizadas através do arquivo “.arff” com exceção daquelas específicas cuja necessidade de um pré-processamento distinto se fez necessário.

Na próxima etapa são apresentados os cruzamentos de dados entre os atributos em relação ao porte, região e setor.

4.1.2 Etapa 4: Cruzamentos de dados

O primeiro cruzamento se trata de um histograma mostrando todos os atributos em relação ao porte (Gráfico 5). As cores azul representam grande porte; vermelho médio; verde claro pequeno e verde escuro às micro-empresas.

Gráfico 5: Histogramas por porte



Fonte: Dados da pesquisa

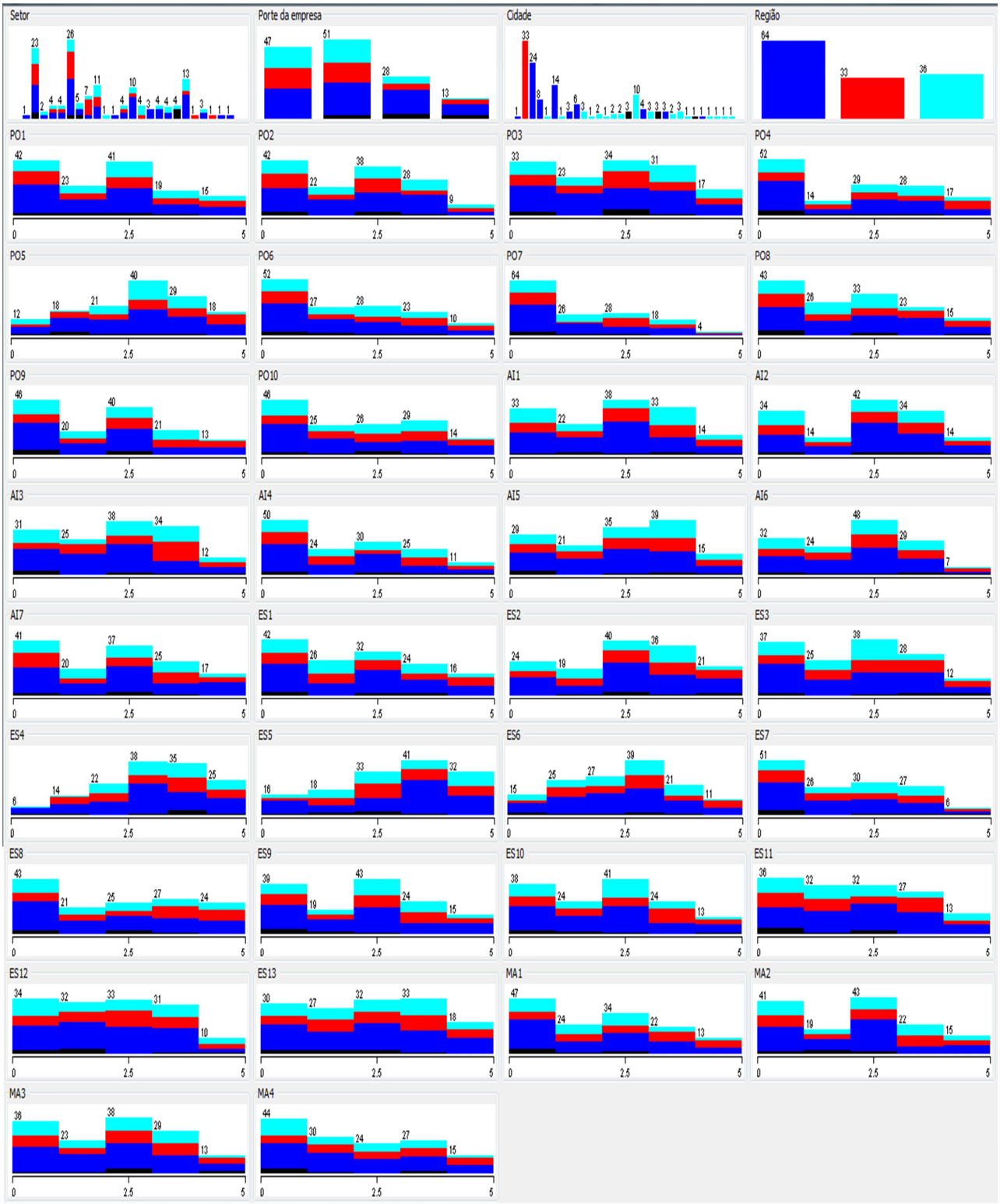
A partir do Gráfico 5, se verifica que em 31 dos 34 processos, as empresas de grande porte (azul) possuem níveis de maturidade mais altos em relação aos demais

portes, o que comprova uma certa obviedade empírica e experimental. As empresas de médio porte (vermelho) apresentam maiores níveis intermediários de maturidade em seus processos, no entanto bastante presente em níveis iniciais em determinados processos como pode ser visto em PO6, PO7, ES11, MA4.

Verifica-se ainda que as micro empresas (verde escuro) estão mais ligadas aos processos com nível de maturidade inicial ou mesmo inexistente.

O próximo cruzamento realizado foi por região (Gráfico 6), onde as cores azul representam as cidades do Vale do Rio dos Sinos, vermelho a capital Porto Alegre e azul claro as cidades do interior do estado.

Gráfico 6: Histogramas por Região

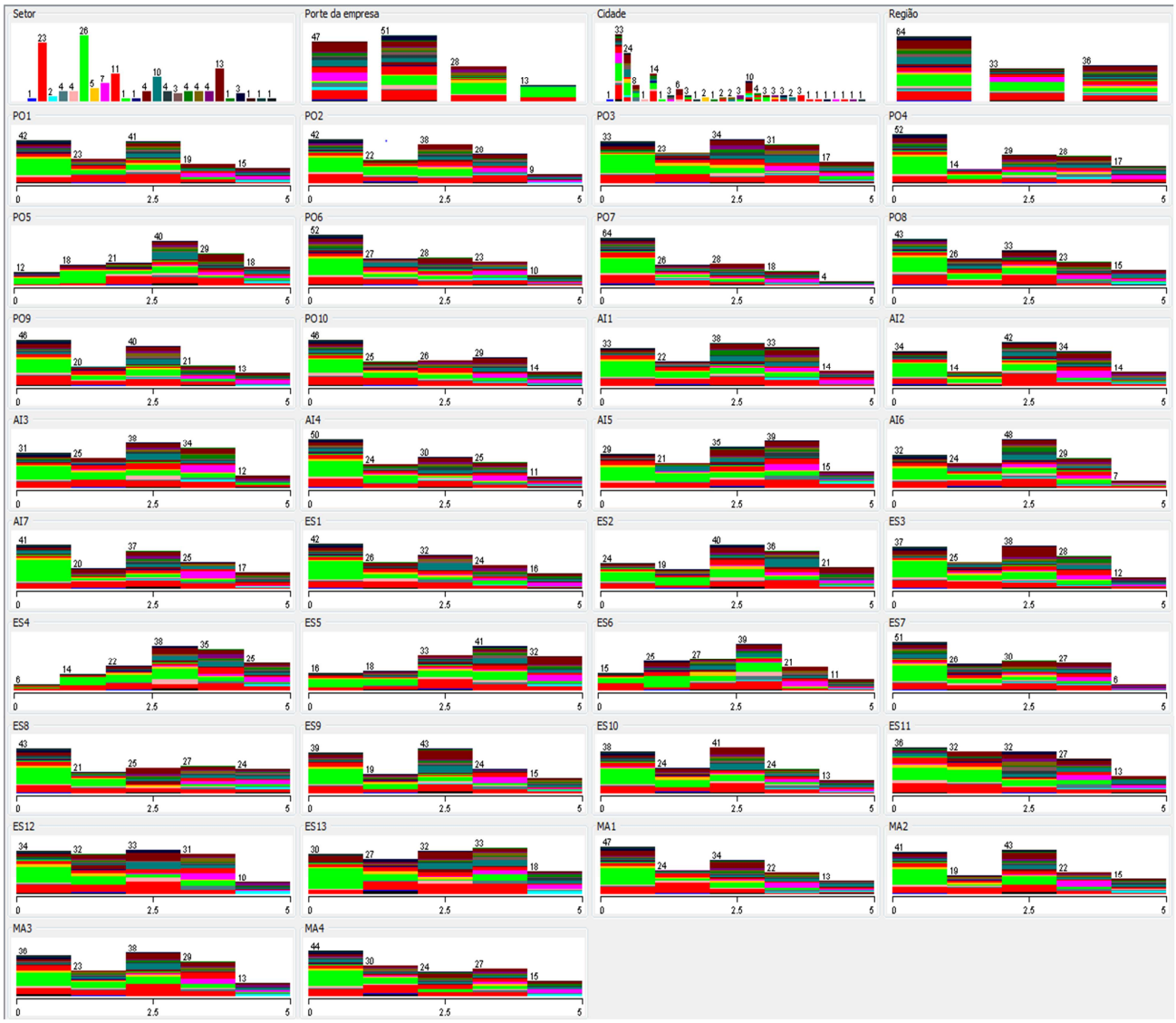


Fonte: Dados da pesquisa

Conforme pode se verificar no Gráfico 6, as empresa das cidades do Vale do Rio dos Sinos (em azul) estão mais presentes nos níveis iniciais e intermediários com exceção dos processos ES4 – Garante a continuidade dos serviços e ES5 – Garante a segurança dos sistemas, no qual apresentam uma maior frequência nos níveis de maturidade mais altos. As empresas da capital possuem uma distribuição mais equilibrada entre os níveis. Já as empresas do interior possuem predominantemente a maturidade nos níveis iniciais e intermediários.

Por fim, visualiza-se no histograma por setor (Gráfico 7), cada um dos setores e como sua frequência se distribui entre os processos de governança.

Gráfico 7: Histograma por setor



Fonte: Dados da pesquisa

Os setores mais representativos, verificados a partir do Gráfico 7, são os de serviços (verde claro) com 26 empresas, seguido pelo de varejo (vermelho) com 23 empresas, sendo que as empresas de serviços estão mais presentes em níveis de maturidade mais baixos e o varejo em níveis intermediários. Nem um dos dois setores se destaca nos níveis com maior maturidade.

Tendo finalizado a importação de dados e apresentação dos principais dados descritivos, inicia-se a aplicação de algoritmos de mineração de dados.

4.2 Aplicação de algoritmos de mineração de dados e análise de resultados

4.2.1 Etapa 5 - Clusters dos dados

Neste capítulo são apresentados os resultados dos clusters (agrupamentos) encontrados por porte das empresas e por sua região.

O algoritmo selecionado para esta tarefa foi o EM (Expectation Maximisation), ou maximização de expectativa. O algoritmo atribui uma distribuição de probabilidade para cada instância e indica a probabilidade de pertencerem a cada um dos agrupamentos. No algoritmo EM, a quantidade de clusters criadas é decidida de forma automática através de inteligência artificial a partir da validação cruzada ou ainda, é especificada a priori pelo pesquisador que indicará o número de grupos que se deseja gerar (WITTEN; FRANK, 2005).

Foram gerados cluster por região, porte e setor, sem as instâncias de empresa e cidade, no entanto o atributo setor gerou informações imprecisas para análise, as quais não foram incorporadas na presente pesquisa.

No Quadro 12 são apresentados os clusters gerados a partir da classe Porte e Região, gerando um cluster para cada um ,as características (nível de maturidade) de cada processo presente nas empresas do referido cluster e a quantidade percentual de empresas presentes em cada um dos agrupamentos.

Quadro 12: Cluster por porte das empresas

Cluster				
	0	1	2	3
	Pequena	**Micro**	**Média**	**Grande**
	26%	20%	34%	19%
Região				
Vale do Rio dos Sinos	22,69%	25,31%	34,70%	17,29%
Capital	32,55%	10,81%	32,31%	24,32%
Interior	27,50%	20,00%	33,20%	19,30%
Processos				
PO1	1,6532	0,9762	2,8178	3,9367
PO2	1,4044	1,0160	2,8265	3,9067
PO3	1,8096	1,1374	3,1784	4,2803
PO4	1,3741	0,6933	2,9216	4,1930
PO5	2,5292	1,0714	3,1602	4,1570
PO6	1,2390	0,4878	2,4348	4,0495
PO7	0,8503	0,2169	2,1702	3,4650
PO8	1,4320	1,1549	2,9615	3,7747
PO9	1,5993	0,6040	2,7392	4,1311
PO10	1,4288	0,6223	2,8685	4,1700
AI1	1,8965	1,1846	3,0948	4,1937
AI2	2,1118	1,2441	3,1367	3,8871
AI3	2,0528	0,8447	3,2086	3,9472
AI4	1,4070	0,4883	2,6963	3,8687
AI5	2,4290	0,7767	3,1512	4,2601
AI6	2,0127	0,7970	2,9824	3,6663
AI7	2,0112	0,5747	2,9430	4,0498
ES1	1,6796	0,7528	2,8877	3,9619
ES2	2,2935	1,5036	3,1525	4,2287
ES3	1,8163	0,8650	2,9397	3,8057
ES4	2,7293	1,6084	3,4649	4,4391
ES5	2,8651	1,6054	3,5965	4,5604
ES6	2,0409	0,8005	2,8637	3,8830
ES7	1,8011	0,3424	2,4475	3,6197
ES8	1,4756	0,6656	3,2397	4,4539
ES9	1,9206	0,5797	2,8664	4,0134
ES10	2,0494	0,4402	2,8077	4,0280
ES11	1,7965	0,9737	2,7258	4,0814
ES12	1,8827	0,6780	2,8347	4,0197
ES13	2,1473	0,9076	3,1475	4,3915
MA1	1,4698	0,2854	2,7472	4,0798
MA2	1,6478	0,7648	2,9286	4,2347
MA3	1,9544	0,7802	2,9472	3,9918
MA4	1,6390	0,4824	2,7323	4,1007

Fonte: Dados da pesquisa

No Quadro 12 se verifica que as empresas de grande porte possuem seus processos com maior nível de maturidade, estando classificados entre definido e gerenciado e esta característica é mais presente em empresas da capital. As empresas de médio porte se encontram com os processos com menor nível de maturidade e nota-se que sua distribuição por região foi mais equilibrada, entre capital, interior e vale dos sinos, possuindo aproximadamente 1/3 para cada uma das três regiões.

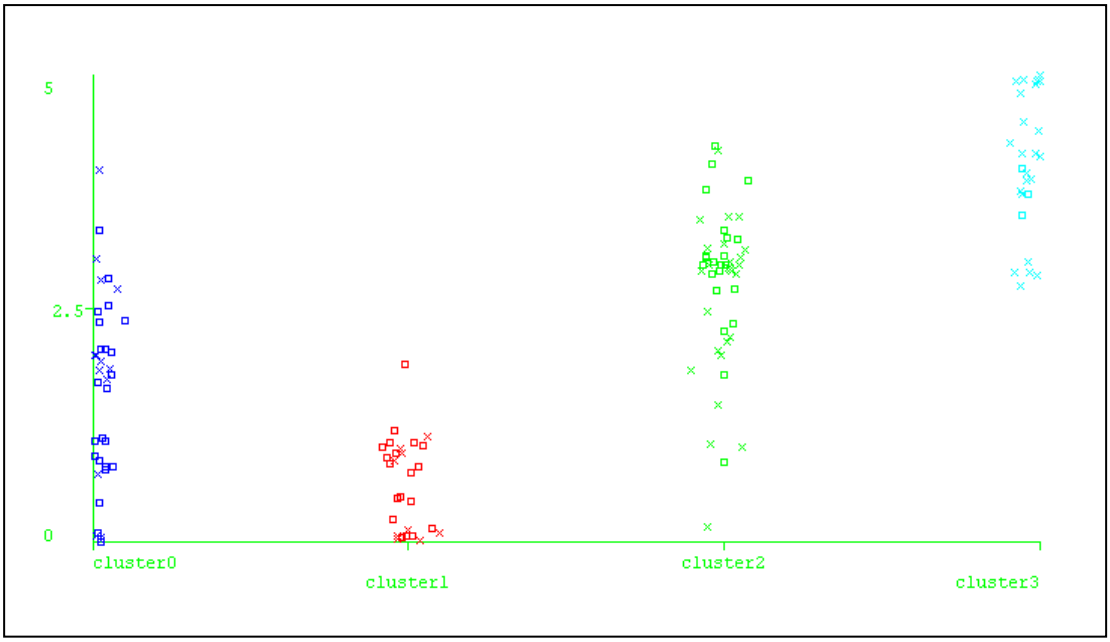
Verifica-se que PO7 (Gerencia os recursos humanos de TI) foi o processo com menor nível de maturidade em relação aos demais processos do mesmo agrupamento, inclusive foi o único com média abaixo de 1 nas pequenas empresas. Isto é um indicativo de baixa preocupação por parte das empresas em relação a este processo, gerando riscos nas operações que envolvam recursos humanos.

O processo ES5 (Garante a segurança dos sistemas) apresentou o mais alto nível em cada um dos clusters nas empresas de porte pequeno, médio e grande, e nas microempresas este processo apresentou o mesmo nível que ES4 (Garante a continuidade dos serviços), diferenciando-se apenas na terceira casa decimal.

Os indicativos de baixo nível de maturidade em processos ligados aos recursos humanos e alto nível nos processos ligados à segurança dos sistemas e hardware não propicia um nível de segurança adequado, que segundo o estudo de SPEARS; BARKI (2010) geram riscos operacionais, pois falhas ou inadequações junto a processos internos podem ter a participação do usuário, mesmo que não intencionalmente. Isto devido a fatores ligados aos recursos humanos, entre eles, baixo nível de treinamento.

No Gráfico 8, fica evidenciado cada um dos 4 clusters por porte em relação ao nível de maturidade.

Gráfico 8: Cluster por Porte



Fonte: Dados da pesquisa

No Gráfico 8, se verifica o cluster 1 (microempresas) nos níveis mais baixos e o cluster 3 (grandes empresas) nos níveis mais altos. Existem empresas que são consideradas *outliers*, ou seja, fora do padrão de normalidade para aquelas características. No entanto entre as empresas dos clusters 1, 2 e 3 nota-se uma maior concentração no nível de maturidade, e nas empresas do cluster 0 (pequenas) está mais pulverizado não apresentando um padrão claramente estabelecido, o que indica que entre as pequenas empresas existem algumas com processos mais iniciais e outras mais organizadas, podendo ser feito um estudo futuro para verificar se estas empresas menores com níveis mais altos de maturidade estariam melhor preparadas para o crescimento.

Após a análise de clusters, de forma mais descritiva, a pesquisa foi direcionada para a análise dos 34 processos do Cobit 4.1, e foram excluídas as classes que não eram relativas a processos, entre elas cidade, região e setor.

Depois de excluídas as classes descritivas e mantidas somente as classes referentes aos processos, realizou-se um procedimento de normalização dos dados que consistiu em mantê-los sob o mesmo intervalo, porem sem perder suas características estatísticas (DODGE, 2003).

4.2.2 Etapa 6 - Normalização

A normalização é útil para transformar os dados, deixando-os sobre um mesmo intervalo sem perder suas características estatísticas. No Weka, somente é possível normalizar atributos numéricos e existindo atributos nominais inseridos, os mesmos são ignorados. Nesta pesquisa o intervalo selecionado foi definido entre 0 e 1, sendo 0 o nível mais baixo de maturidade e 1 o mais alto.

Para realizar a função de normalizar, o software Weka possui uma função de pré-processamento chamada Normalize, que realiza a normalização dos dados conforme os critérios estabelecidos. A função pode ser acessada nos seguintes passos: *filters/unsupervisioned/instance/normalize*.

Depois de selecionada o filtro, nas opções de função, se escolhe adicionar o número “1” no campo *norm* indicando o intervalo máximo estabelecido e se ativa o software para realizar a função, que disponibilizou na tela principal os dados já normalizados.

Para normalização dos dados, se utiliza a fórmula da Figura 10.

Figura 10: Fórmula da normalização

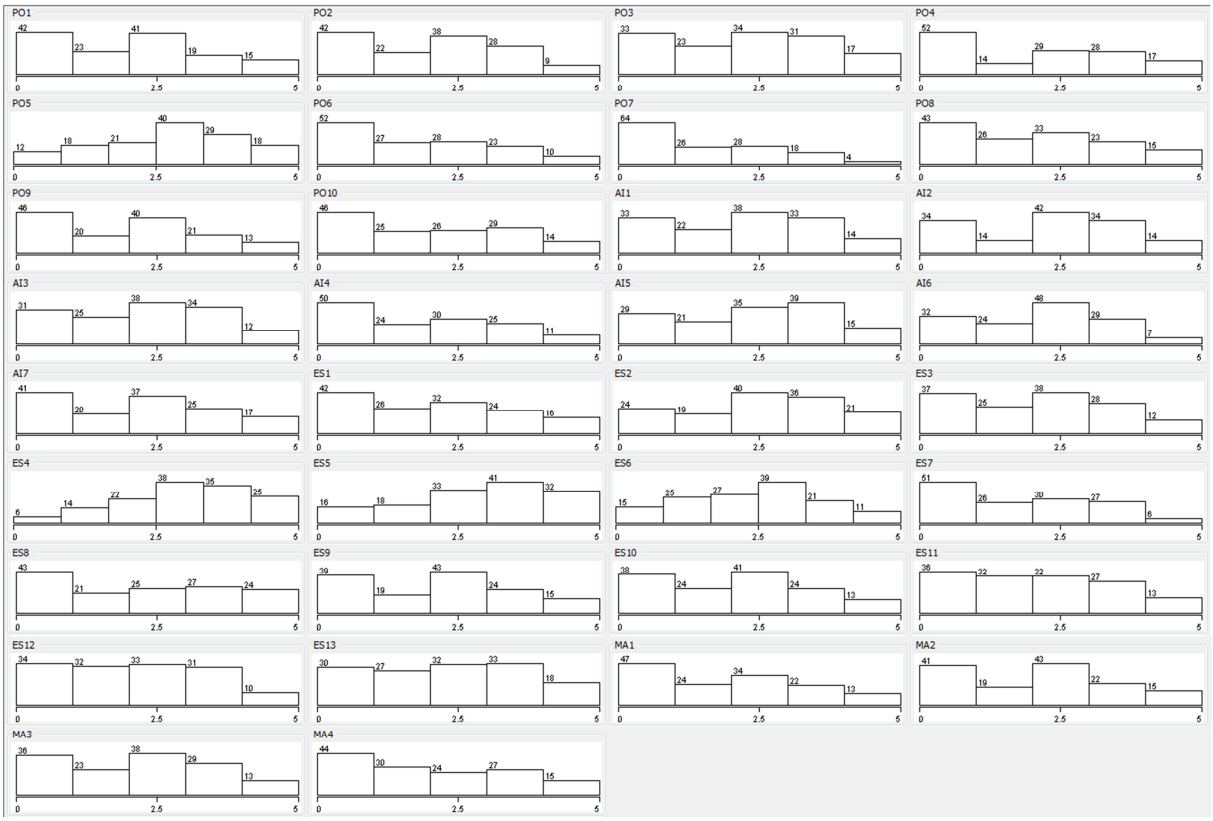
$$Z = \frac{X - \mu}{\sigma}$$

Fonte: Dodge (2003)

Onde Z é o valor que se deseja normalizar, X é o valor atual μ é a média dos valores e σ é o desvio padrão.

Na sequência são apresentados os gráficos de cada uma das classes, aparecendo no primeiro os dados recebidos não normalizados (Gráfico 9) e no segundo já com os dados normalizados (Gráfico 10)

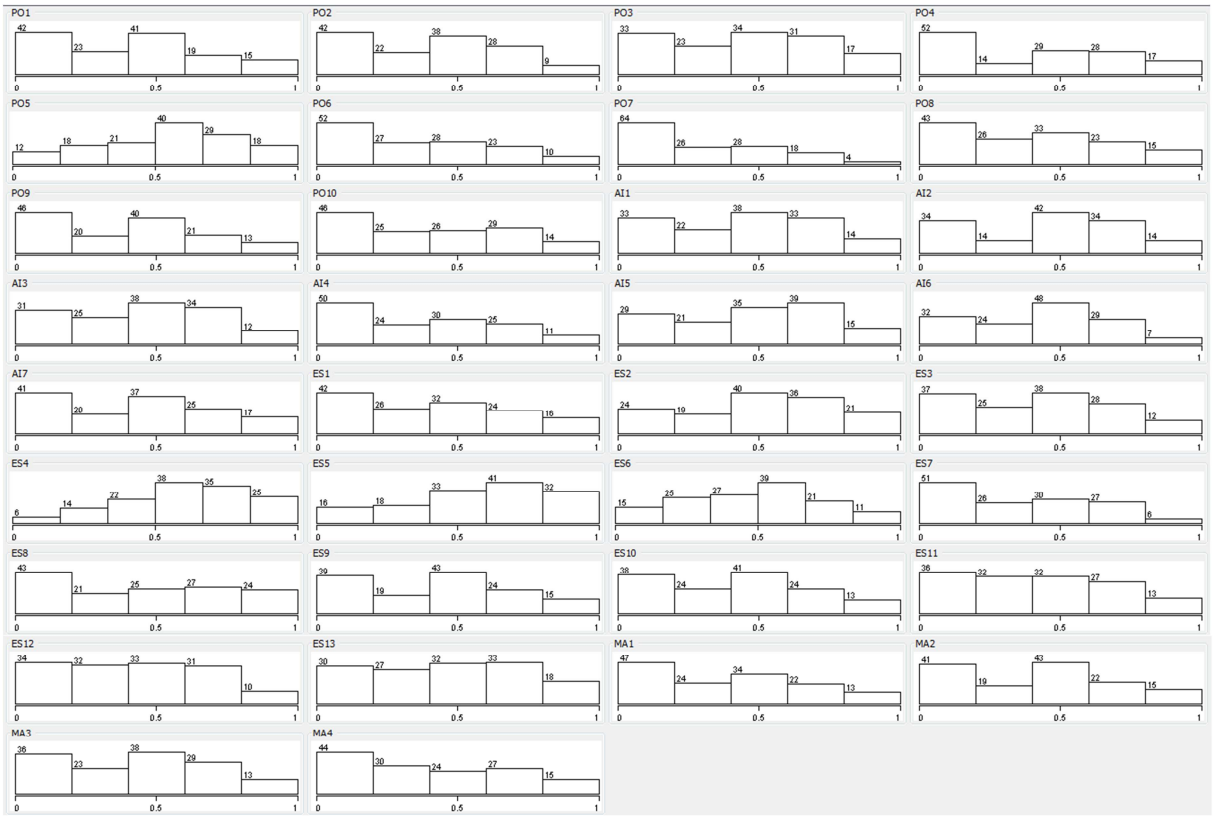
Gráfico 9: Histogramas dos processos - Não Normalizados



Fonte: Dados da pesquisa

Pode se visualizar que o intervalo dos dados do gráfico 9 (não normalizado) está entre 0 e 5, que corresponde desde o nível inexistente (0) até o nível otimizado(5), diferente do apresentado no Gráfico 10, normalizado.

Gráfico 10: Histogramas dos processos - Normalizados



Fonte: Dados da pesquisa

No Gráfico 10, fica evidenciado que o intervalo encontra-se entre 0 e 1. Como já exposto pelo autor, a normalização permite colocar os dados em um intervalo que neste caso ficou entre 0 e 1, sem que se perdessem suas características estatísticas, o que fica comprovado quando comparados os dois gráficos. Na Tabela 1 pode-se verificar a equivalência entre o nível de maturidade do Cobit e o valor do processo após a normalização.

Tabela 1: Equivalência Cobit - Atributo normalizado

COBIT	De	Até
0	0,000000	0,166667
1	0,166667	0,333333
2	0,333333	0,500000
3	0,500000	0,666667
4	0,666667	0,833333
5	0,833333	1,000000

Fonte: Elaborado pelo autor.

Na Tabela 1 verifica-se que os valores mínimo e máximo são respectivamente 0 e 1, devido a normalização dos dados, onde quanto mais próximo a zero, o processo possui menor nível de maturidade e quanto mais próximo a 1 possui um maior nível de maturidade.

A próxima etapa compreende a geração de clusters por processos com o objetivo de gerar agrupamentos de características semelhantes entre si e diferentes entre os grupos.

4.2.3 Etapa 7 - Cluster dos processos

Nesta etapa foram selecionados dois algoritmos de geração de clusters, o EM e o Kmeans. Os aspectos descritivos não foram analisados nesta etapa , sendoutilizados somente os processos do Cobit 4.1.

Os resultados gerados a partir de dois algoritmos, EM e Kmeans, trouxeram resultados muito semelhantes, que podem ser visualizados no Quadro 13. A tabela contém cada um dos 34 processos do Cobit 4.1, os 4 clusters encontrados, sua participação em relação ao total de empresas e o nível de cada um.

Quadro 13: Clusters gerados pelos algoritmos EM e Kmeans

Cluster - EM				
	0	1	2	3
	32%	30%	21%	17%
PO1	0,3203	0,5895	0,7822	0,1694
PO2	0,3064	0,5715	0,7793	0,1597
PO3	0,3769	0,6229	0,8690	0,2070
PO4	0,2529	0,6050	0,8334	0,1535
PO5	0,4876	0,6573	0,8166	0,1750
PO6	0,2613	0,4817	0,8098	0,0639
PO7	0,1599	0,4446	0,6919	0,0424
PO8	0,3337	0,5540	0,7810	0,2028
PO9	0,3265	0,5357	0,8116	0,1243
PO10	0,3051	0,5481	0,8335	0,1285
AI1	0,4103	0,6055	0,8278	0,2097
AI2	0,4656	0,6200	0,7545	0,1984
AI3	0,4322	0,6183	0,7870	0,1556
AI4	0,2851	0,5224	0,7903	0,0889
AI5	0,4707	0,6286	0,8331	0,1563
AI6	0,4380	0,5858	0,7293	0,0945
AI7	0,3647	0,6168	0,7865	0,1258
ES1	0,3752	0,5671	0,7776	0,1007
ES2	0,4978	0,6245	0,8261	0,2422
ES3	0,4072	0,5599	0,7685	0,1188
ES4	0,5547	0,6963	0,8738	0,2840
ES5	0,5868	0,7261	0,9024	0,2581
ES6	0,3875	0,5817	0,7660	0,1618
ES7	0,3518	0,4760	0,7189	0,0632
ES8	0,3205	0,6400	0,8781	0,1221
ES9	0,3740	0,5777	0,7870	0,1101
ES10	0,3853	0,5597	0,7976	0,0945
ES11	0,3496	0,5500	0,7989	0,2019
ES12	0,3718	0,5765	0,7661	0,1415
ES13	0,4186	0,6567	0,8452	0,1617
MA1	0,2743	0,5612	0,8160	0,0499
MA2	0,3319	0,5920	0,8302	0,1451
MA3	0,4327	0,5625	0,7728	0,1284
MA4	0,3148	0,5669	0,7802	0,1042

Cluster - KMeans					
	Geral	0	1	2	3
	100%	21%	34%	18%	26%
PO1	0,4709	0,7809	0,5473	0,1947	0,3072
PO2	0,4588	0,7777	0,5339	0,1853	0,2876
PO3	0,5236	0,8617	0,5980	0,2307	0,3509
PO4	0,4618	0,8289	0,5680	0,1473	0,2387
PO5	0,5531	0,8107	0,6264	0,1840	0,4986
PO6	0,4074	0,8024	0,4585	0,0933	0,2329
PO7	0,3354	0,6896	0,4050	0,0407	0,1572
PO8	0,4701	0,7788	0,5416	0,2187	0,2968
PO9	0,4552	0,8124	0,5140	0,1193	0,3162
PO10	0,4573	0,8292	0,5464	0,1233	0,2658
AI1	0,5210	0,8249	0,6069	0,2173	0,3685
AI2	0,5260	0,7520	0,6312	0,2147	0,4167
AI3	0,5142	0,7823	0,6208	0,1493	0,4050
AI4	0,4274	0,7857	0,5080	0,0933	0,2581
AI5	0,5393	0,8272	0,6297	0,1500	0,4518
AI6	0,4838	0,7251	0,5757	0,1307	0,4077
AI7	0,4868	0,7838	0,5942	0,1207	0,3541
ES1	0,4692	0,7746	0,5670	0,0967	0,3464
ES2	0,5601	0,8216	0,6234	0,2727	0,4604
ES3	0,4785	0,7608	0,5671	0,1540	0,3541
ES4	0,6170	0,8716	0,7016	0,2727	0,5333
ES5	0,6377	0,9023	0,7169	0,2480	0,5838
ES6	0,4856	0,7652	0,5661	0,1553	0,3775
ES7	0,4158	0,7176	0,4758	0,0687	0,3279
ES8	0,4979	0,8766	0,6092	0,1173	0,3036
ES9	0,4755	0,7836	0,5611	0,1060	0,3644
ES10	0,4733	0,7919	0,5542	0,0907	0,3685
ES11	0,4776	0,7983	0,5339	0,2020	0,3306
ES12	0,4755	0,7651	0,5687	0,1360	0,3491
ES13	0,5344	0,8428	0,6396	0,1633	0,3986
MA1	0,4342	0,8137	0,5342	0,0480	0,2577
MA2	0,4812	0,8283	0,5802	0,1553	0,2914
MA3	0,4901	0,7706	0,5903	0,1473	0,3641
MA4	0,4508	0,7793	0,5663	0,1000	0,2716

Fonte: Dados da pesquisa

No Quadro 13 fica evidenciado que embora o número do cluster de um algoritmo não corresponda diretamente ao de outro, ocorreram resultados muito semelhantes. Para realizar uma verificação disto, deve-se analisar individualmente como cada empresa foi classificada, de acordo com o quadro do apêndice A. No entanto, para poder realizar a comparação fez-se necessário criar uma tabela de equivalência, em função da ordem dos agrupamentos gerados pelo sistema serem diferentes entre si.

No Quadro 14, é apresentada a equivalência do cluster EM em relação ao Kmeans para os agrupamentos de todos os processos.

Quadro 14: Equivalência de clusters

Cluster EM	Cluster Kmeans
0	3
1	1
2	0
3	2

Fonte: Dados da pesquisa

No Quadro 14, pode ser visualizado que as características do cluster 0 gerado pelo algoritmo EM equivale ao cluster 3 no algoritmo Kmeans.

Verifica-se, a partir do comparativo presente no apêndice A, que entre as empresas que foram atribuídas a distintos agrupamentos pelo algoritmo EM e Kmeans, o cluster indicado apresentou diferenças significativas entre os níveis. Ou seja, a empresa foi agrupada sob um cluster próximo, mas com algumas características semelhantes entre si, o que pode ser visualizado na empresa 22, onde foi atribuído pelo algoritmo EM que ela pertencia ao cluster 0, caracterizado por processos de nível repetível e pelo KMeans equivalente ao cluster 1 caracterizado por processos de nível definido. Isso significa afirmar que as empresas cujas atribuições de cada um dos algoritmos foram diferentes, apresentam características de maturidade de seus processos na fronteira entre os clusters e- não se observou nenhuma atribuição de empresa cujo agrupamento foi classificado como de baixo nível de maturidade por um algoritmo e alto por outro.

Fica evidenciado, a partir do comparativo, que a grande maioria das empresas foram classificadas de igual forma- indicando uma convergência na forma de qualificar dos algoritmos e uma maior confiabilidade nos agrupamentos encontrados. Para maior clareza dos resultados, foi gerado a Tabela 2 mostrando a frequência e o percentual de equivalência entre os algoritmos.

Tabela 2: Resumo das equivalências entre os clusters

	Frequência	%
Equivalente	131	93,6%
Diferente	9	6,4%
	140	100,0%

Fonte: Dados da pesquisa

Verifica-se a partir da Tabela 2 que 93,6% das respostas foram classificadas de igual forma, ou seja, a ampla maioria das 140 empresas foi atribuída ao mesmo cluster pelos dois diferentes algoritmos (EM e Kmeans), mostrando consistência nos agrupamentos obtidos. Também se verifica que entre as 9 classificações diferentes, em sete delas o cluster EM atribuiu como 0 (nível repetível) e o Kmeans como 1(nível definido), e considerando que a sequencia do nível repetível é o nível definido, as empresas atribuídas possuem características que atendem aos dois agrupamentos, ou seja, possuem processos que estão muito próximos, alguns de nível repetível, e outros definido, permitindo ao algoritmo atribuí-la a diferentes agrupamentos.

Para um entendimento mais completo acerca das principais características de cada cluster, foi gerado a Tabela 3.

Tabela 3: Características dos clusters de processos

Cluster EM				
Cluster	0	1	2	3
% Participação	32%	30%	21%	17%
Valor mínimo	0,1599	0,4446	0,6919	0,0424
Valor máximo	0,5868	0,7261	0,9024	0,2840
Média do cluster	0,3745	0,5848	0,8007	0,1454

Cluster Kmeans				
Cluster	0	1	2	3
% Participação	21%	34%	18%	26%
Valor mínimo	0,6896	0,4050	0,0407	0,1572
Valor máximo	0,9023	0,7169	0,2727	0,5838
Média do cluster	0,7976	0,5724	0,1514	0,3499

Fonte: Dados da pesquisa

.

Está considerada na Tabela 3, a sequência de cada cluster encontrado- e suas principais características segundo os níveis de maturidade do Cobit 4.1.

Cluster 0: Processos com certo gerenciamento, onde alguns procedimentos são seguidos por diferentes pessoas porém a responsabilidade sobre os processos

é deixada com o indivíduo. Há um alto grau de confiança no conhecimento das pessoas e consequentemente erros podem ocorrer. Está ligado ao nível de maturidade Repetível.

Cluster 1: Os processos possuem procedimentos padronizados e foram documentados e comunicados através de treinamento. Possíveis desvios não são detectados e os procedimentos não são sofisticados, embora haja a formalização das práticas existentes. Está ligado ao nível de maturidade Definido.

Cluster 2: Os processos estão bem gerenciados e quase otimizados. É monitorada e medida a aderência aos procedimentos e são adotadas ações quando não estão funcionando muito bem. Os processos estão debaixo de um constante aprimoramento e fornecem boas práticas Está ligado ao nível de maturidade Gerenciado.

Cluster 3: Processos com pouco ou nenhum gerenciamento porém é possível verificar que existem questões a serem trabalhadas. Não existe processo padronizado e os processos não estão bem gerenciados. Está ligado ao nível de maturidade Inexistente/Inicial.

Ainda, pode-se observar a partir do Quadro 17, que as médias dos clusters do algoritmo EM não apresentam uma variação significativa aos seus equivalentes no algoritmo Kmeans. Ainda, se verifica que o cluster que apresentou maior variação entre o processo com menor e maior valor foi o cluster 0, enquanto que o cluster 2 apresentou a menor variação. Isto significa dizer que embora nos processos do cluster 0 exista certo gerenciamento, este é focado a alguns especificamente em detrimento de outros. Já nos do cluster 2, evidencia-se um nível de maturidade mais equilibrado, ou seja, o gerenciamento abrange todos os processos.

A próxima etapa contempla um aprofundamento sobre o processo PO9 – Avalia e Gerencia os Riscos, relacionado com o objetivo do estudo que busca analisar a relação entre risco operacional e processos tecnológicos.

4.2.4 Etapa 8 - Seleção de atributos

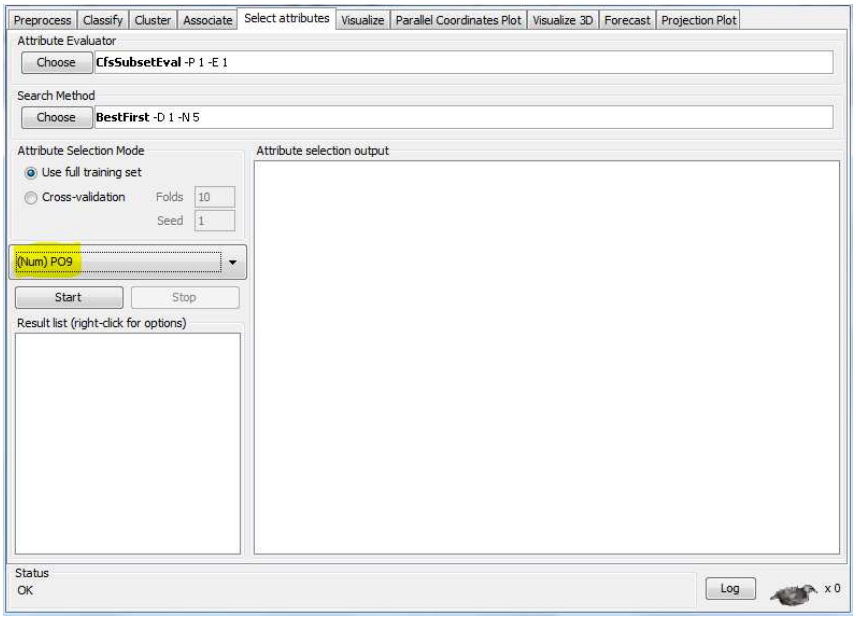
Nesta etapa objetiva-se identificar os atributos que são discriminantes sobre nível de maturidade do processo PO9 – Avalia e gerencia riscos.

O nome deste processo de busca de atributos descriminantes denomina-se seleção de atributos (*Select Attributes*).

A Seleção de atributos no Weka funciona a partir de dois elementos, um algoritmo de avaliação e outro de busca. O avaliador verifica os atributos que são altamente correlacionados com a classe e não correlacionados entre si. O buscador utiliza de estratégias para otimizar a pesquisa na base de dados a fim de encontrar subconjuntos a um menor custo computacional.

Como apresentado anteriormente, o processo PO9 foi o alvo de busca, ou seja, a partir desta classe serão realizadas as buscas dos atributos discriminantes. Para seleção do atributo alvo, após selecionar a aba *Select attributes* escolhe-se na caixa de seleção destacada em amarelo na Figura 11 o atributo desejado.

Figura 11: Seleção do atributo PO9



Fonte: Dados da pesquisa

Na Figura 11, pode ser visualizada a tela do Weka onde foi realizada a seleção de atributos. O algoritmo principal e avaliador selecionado foi o CfsSubsetEval e o método de busca foi o BestFirst. Foram realizadas buscas com mais outros três métodos com a finalidade de encontrar os processos com maior presença nos diversos métodos.

Foram utilizados 4 métodos de busca, que encontraram 12 processos que são descriminantes entre os 34 do Cobit em relação ao processo de avaliação e

gerenciamento de risco (PO9). No Quadro 15 podem ser visualizados os resultados encontrados para cada um dos métodos de busca e ao final do quadro os totais de atributos encontrados.

Quadro 15: Resultado comparativo da seleção de atributos

CfsSubsetEval			
Método de busca			
Best First	**Greedy Stepwise**	**Linear Forward Selection**	**Evolutionary Search**
PO3	PO3	PO3	PO3
PO7	PO7	PO7	PO7
PO8	PO8	PO8	PO8
PO10	PO10	PO10	PO10
AI1	AI1	AI1	AI1
AI2	AI2	AI2	AI2
AI3	AI3	AI3	AI3
ES11	ES11	ES11	ES6
MA3	MA3	MA3	ES8
MA4	MA4	MA4	MA3
			MA4
Numero de atributos encontrados			
10	10	10	11

Fonte: Dados da pesquisa

No Quadro 15 ficou constatado que dentre os 12 processos encontrados, 9 estão presentes em todos os métodos de busca. Como critério de seleção de atributos e refinamento da busca, optou-se por manter os atributos que possuíam maior poder discriminante em relação aos demais e para isso foram selecionados aqueles que estavam presentes em todos os métodos, ou seja, 9 atributos, que podem ser verificados no Quadro 16.

Embora os processos ES6, ES8 e ES11 tenham aparecido na busca, eles não estavam presentes em todos os métodos, indicando que os demais métodos não os consideraram suficientemente descriminantes e, em função disto, foram excluídos.

Quadro 16: Atributos presentes em todos os métodos de busca

PO3	**PO10**	**AI2**	**MA3**
PO7	**AI1**	**AI3**	**MA4**
PO8			

Fonte: Dados da pesquisa

A partir destes atributos, presentes no Quadro 16, incluindo o PO9, foi gerado um novo arquivo onde permanecem estes atributos e os demais são excluídos com o objetivo de estudar o comportamento destes sobre o atributo de riscos (PO9).

Para exclusão das classes que não serão mais utilizadas e geração do novo arquivo de trabalho, é preciso selecionar as que se deseja eliminar usando a opção *Remove.* Após esta operação, foi usada a opção *Save* para criar um novo arquivo ".arff", com nome distinto do primeiro, para utilização posterior. A partir deste momento, todas as atividades foram geradas a partir do novo arquivo com 10 classes distintas.

Na próxima etapa são gerados novos agrupamentos baseados nas similaridades entre os atributos.

4.2.5 Etapa 9 - Clusters com atributos de risco

Nesta etapa são gerados clusters a partir de 10 processos, que foram selecionados através da triagem dos atributos e em função de maior relação com os processos de avaliação e gerenciamento de riscos (PO9). Foram aplicados os algoritmos anteriormente utilizados no total dos processos, o EM e o Kmeans. Nas aplicações realizadas com o total dos 34 processos foi incluída em sua denominação a palavra "total" e naquelas realizadas com os 10 processos de maior influência sobre riscos, foi incluída a palavra "reduzido", bem como gerado uma tabela de equivalência entre os clusters reduzido e total, com o objetivo de comparar segundo as mesmas características.

Foram realizados testes com os algoritmos EM e Kmeans, e realizadas comparações entre os resultados das atribuições dos algoritmos no total dos processos e com os 10 processos encontrados na etapa de seleção de atributos. O algoritmo que apresentou melhor resultado foi o Kmeans. O resultado do comparativo das atribuições do Kmeans reduzido com o cluster de todos os

processos foi gerado a partir dos dados presentes no apêndice B e pode ser visualizado resumidamente na Tabela 4.

Tabela 4: Resumo comparativo entre cluster Kmeans reduzido e cluster de todos processos

	Kmeans Red. X EM Total		Kmeans Red. X Kmeans Total	
	Frequência	%	Frequência	%
Equivalente	106	75,71%	110	78,57%
Diferente	34	24,29%	30	21,43%
	140	100,00%	140	100,00%

Fonte: Dados da pesquisa

A Tabela 4 apresenta como o algoritmo Kmeans reduzido classificou as empresas de igual forma que o EM total e o Kmeans total em 75,71% e 78,57%, respectivamente. Isso indica que nos clusters do algoritmo Kmeans aproximadamente 3/4 dos centróides permanecem iguais, significando que as características dos agrupamentos encontrados no total dos processos e nos reduzidos é idêntico em aproximadamente 3/4 das empresas, e que o nível de maturidade dos processos reduzidos tende a apresentar os mesmos níveis de maturidade quando avaliados os demais processos, não apresentando variação acentuada entre os mesmos.

No Quadro 17 podem ser visualizados os resultados para cada um dos clusters encontrados pelo algoritmo Kmeans.

Quadro 17: Clusters Kmeans com 10 atributos

Cluster Kmeans					
	Geral	0	1	2	3
	100%	27%	28%	21%	24%
PO3	0,5236	0,8311	0,5660	0,4466	0,1971
PO7	0,3354	0,6558	0,3822	0,1603	0,0730
PO8	0,4701	0,7745	0,4966	0,3626	0,1912
PO9	0,4552	0,7818	0,5274	0,3011	0,1387
PO10	0,4573	0,7924	0,5503	0,2644	0,1407
AI1	0,5210	0,7916	0,6179	0,4218	0,1922
AI2	0,5260	0,7235	0,6461	0,4879	0,2000
AI3	0,5142	0,7624	0,6170	0,4224	0,1971
MA3	0,4901	0,7592	0,5804	0,3362	0,2169
MA4	0,4508	0,7767	0,5226	0,3408	0,0980

Fonte: Dados da pesquisa.

Verifica-se a partir do Quadro 17, que o cluster com maior nível de maturidade é o zero e o com menor é o 3. Para um entendimento em relação às principais características de cada cluster do Quadro 17, foi gerada a Tabela 5.

Tabela 5: Característica dos clusters com 10 atributos

Cluster Kmeans				
Cluster	0	1	2	3
% Participação	27%	28%	21%	24%
Valor mínimo	0,6558	0,3822	0,1603	0,0730
Valor máximo	0,8311	0,6461	0,4879	0,2169
Média do cluster	0,7649	0,5507	0,3544	0,1645

Fonte: Dados da pesquisa.

Considerando a Tabela 5, é apresentado na sequência cada cluster encontrado e suas principais características segundo os níveis de maturidade.

Cluster 0: Nível Gerenciado. Os processos estão bem gerenciados e quase otimizados. É monitorada e medida a aderência aos procedimentos e se adota ações quando não estão funcionando muito bem. Os processos estão debaixo de um constante aprimoramento e fornecem boas práticas.

Cluster 1: Nível Definido. Os processos possuem procedimentos padronizados e foram documentados e comunicados através de treinamento. Possíveis desvios não são detectados e os procedimentos não são sofisticados embora exista a formalização das práticas existentes.

Cluster 2: Nível Repetível. Processos com certo gerenciamento, onde alguns procedimentos são seguidos por diferentes pessoas, porém a responsabilidade sobre os processos é deixada com o indivíduo. Há um alto grau de confiança no conhecimento dos indivíduos e consequentemente erros podem ocorrer.

Cluster 3: Nível Inexistente/Inicial. Processos com pouco ou nenhum gerenciamento e verifica-se que existem questões que precisam ser trabalhadas. Não existe processo padronizado e os processos não estão bem gerenciados.

Após a análise dos agrupamentos, foi realizada a tarefa de classificação, gerando uma árvore de decisão que pode ser vista na próxima etapa.

4.2.6 Etapa 10 - Classificação – Árvore de decisão

A tarefa de classificação visa determinar o valor de um atributo através dos valores de um subconjunto dos demais atributos da base de dados e, com isso, ter uma visão clara do comportamento do nível de maturidade do atributo alvo diante dos níveis dos demais atributos.

A classificação permite realizar duas análises. A primeira ao prever qual será o nível de maturidade do atributo alvo, baseado nos processos dos demais, e a segunda, ao se estimar o valor desejado para o alvo, permitindo analisar quais os processos que possuem relação com o mesmo, e as possíveis ações para se atingir o nível de maturidade desejado para o processo alvo, focando os recursos naqueles que trarão melhores resultados.

Para realizar a geração de uma árvore de decisão, foram utilizados dois algoritmos de classificação, chamados Reptree e M5P. O resultado do classificador foi testado com a opção *Use Training Set*, na qual primeiro ocorre o treinamento com os dados disponíveis e na sequência os testa. Os resultados do coeficiente de correlação e os erros, médio absoluto e quadrático, podem ser vistos na Tabela 6.

Tabela 6: Resultados dos algoritmos RepTree e M5P

Algoritmo de Arvore de Decisão		
Indicador	**RepTree**	**M5P**
Coeficiente de Correlação	0,8437	0,8441
Erro médio absoluto	0,1107	0,1324
Erro médio quadrático	0,1544	0,1699

Fonte: Dados da pesquisa

O resultado dos dois algoritmos, segundo a Tabela 6, são semelhantes porém o algoritmo RepTree apresentou melhores resultados em relação aos erros. Assim, este foi selecionado para expor os resultados encontrados.

O REPTree é um algoritmo de aprendizagem que gera árvores de decisão usando informação de ganho/variância e poda, sendo que a poda é usada para redução de erro.

Os resultados encontrados são apresentados de duas formas, a primeira em formato texto que pode ser visualizado na Figura 13 e a segunda em formato de

árvore de decisão, representada graficamente na Figura 14, que permite uma maior compreensão. No formato texto (Figura 12) a coluna inicial representa o item do nó conjuntamente com um teste de verificação do valor do atributo; este, se confirmado, leva a coluna seguinte, com um nova verificação, até atingir a última coluna que, se estiverem satisfeitas as condições anteriores, apresenta o resultado esperado para o atributo alvo. Na presente pesquisa, o atributo alvo é o processo de risco operacional.

Figura 12: Resultado do algoritmo RepTree em formato texto

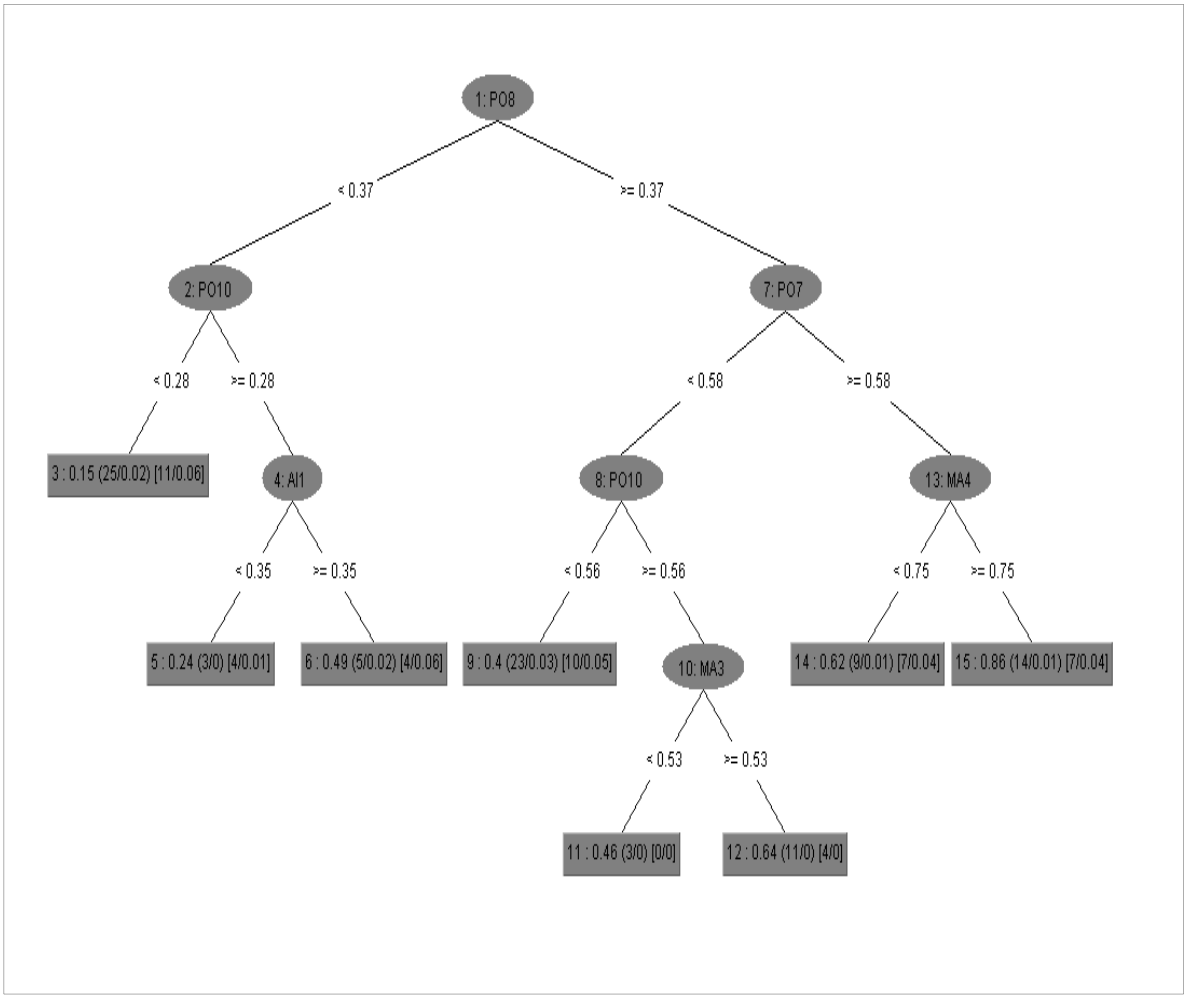
```
PO8 < 0.37
|   PO10 < 0.28 : 0.15 (25/0.02) [11/0.06]
|   PO10 >= 0.28
|   |   AI1 < 0.35 : 0.24 (3/0) [4/0.01]
|   |   AI1 >= 0.35 : 0.49 (5/0.02) [4/0.06]
PO8 >= 0.37
|   PO7 < 0.58
|   |   PO10 < 0.56 : 0.4 (23/0.03) [10/0.05]
|   |   PO10 >= 0.56
|   |   |   MA3 < 0.53 : 0.46 (3/0) [0/0]
|   |   |   MA3 >= 0.53 : 0.64 (11/0) [4/0]
|   PO7 >= 0.58
|   |   MA4 < 0.75 : 0.62 (9/0.01) [7/0.04]
|   |   MA4 >= 0.75 : 0.86 (14/0.01) [7/0.04]
```

Fonte: Dados da pesquisa

Na Figura 12 são geradas regras para o PO9 (Avalia e gerencia riscos) e como cada um dos elementos se comporta. A leitura da última regra gerada diz que o processo PO9 terá valor de 0,86, ou seja, terá um alto nível de maturidade, quando o processo MA4 (Fornecer governança de TI) for maior ou igual a 0,75; PO7 (Gerencia os recursos humanos de TI) maior ou igual a 0,58 e PO8 (Gerencia a qualidade) maior ou igual a 0,37, com suporte de 14 instancias antes da poda, ou seja, com o conjunto completo da base de dados. Isso significa que essa regra se confirmou em 14 vezes em 140 instâncias.

Já na Figura 13, fica mais evidente a compreensão da regra uma vez que ela foi construída em forma gráfica de árvore de decisão. As regras geradas ficam na parte inferior da árvore e se referem ao PO9, alvo da busca das regras de associação.

Figura 13: Árvore de decisão do processo PO9



Fonte: Dados da pesquisa

A árvore de decisão gerada na Figura 13 apresenta “caminhos” a serem percorridos para que se atinja uma determinada regra e para que possa auxiliar a gestão do nível de maturidade do processo PO9. Os pontos em comum foram: O processo PO8 (Gerencia a Qualidade), está presente em todas as regras geradas. O processo MA4 (Fornecer Governança de TI) participa das regras com maior nível de maturidade para o processo de Avaliação e gerenciamento de riscos (PO9) e o

processo PO10 (Gerencia os projetos) participa de seis das dez regras geradas. Isso indica que quanto maior a governança de TI, melhor monitorados estarão os riscos .

Como os dados foram normalizados no intervalo entre 0 e 1, um quadro de equivalência entre o valor obtido pela regra e os níveis originais do Cobit, que pode ser visualizado na Tabela 7, auxilia um melhor entendimento das informações geradas pelos resultados das aplicações de mineração de dados.

.

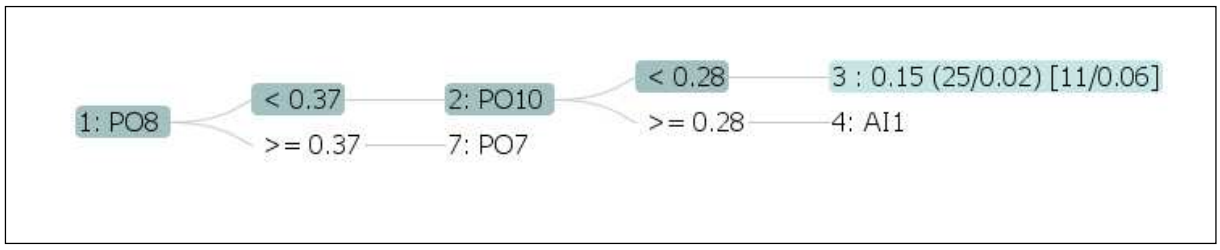
Tabela 7: Equivalência COBIT – Atributos normalizados

COBIT	De	Até
0	0,000000	0,166667
1	0,166667	0,333333
2	0,333333	0,500000
3	0,500000	0,666667
4	0,666667	0,833333
5	0,833333	1,000000

Fonte: Dados da pesquisa

Para uma melhor visualização e interpretação dos dados se utiliza da ferramenta *Profuse,* presente no Weka, que gera uma visualização individual para cada uma das oito regras encontradas. A informação entre parênteses gerada na regra são o suporte e o erro. As visualizações podem ser verificadas da Figura 14 até a 18.

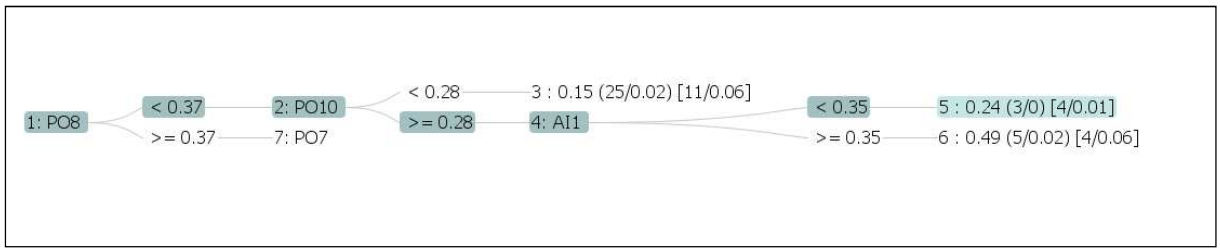
Figura 14: Regra de associação nº 1



Fonte: Dados da pesquisa

A Figura 14 apresenta a regra na qual afirma que o processo PO9 vai ter valor de 0,15 (inexistente) quando PO10 (Gerencia os projetos) for menor que 0,28 e PO8 (Gerencia a qualidade) menor que 0,37.

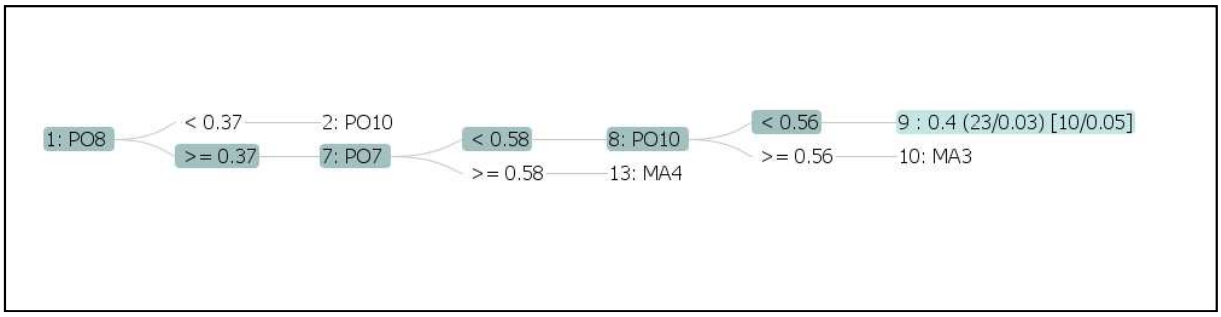
Figura 15: Regra de associação nº 2 e 3



Fonte: Dados da pesquisa

As regras nº 2 e 3 cujo valor é respectivamente 0,24 e 0,49 possuem algumas características em comum, confirmando-se quando PO8 for menor que 0,37 e PO10 maior ou igual a 0,28. Porém, se o processo AI1 for menor que 0,35, o PO9 será 0,24 e quando AI1 for maior ou igual a 0,35 o valor de PO9 é 0,49, com suporte de 4.

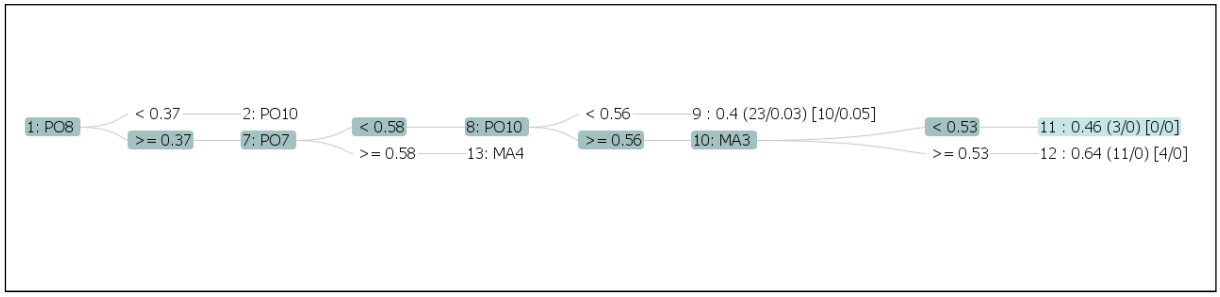
Figura 16: Regra de associação nº 4



Fonte: Dados da pesquisa

A regra nº 4, apresentada na Figura 16 é uma das mais fortes, tendo um suporte de 23. A regra diz que PO9 será 0,40 quando PO8 for maior ou igual a 0,37, PO7 menor que 0,58 e PO10 menor que 0,56.

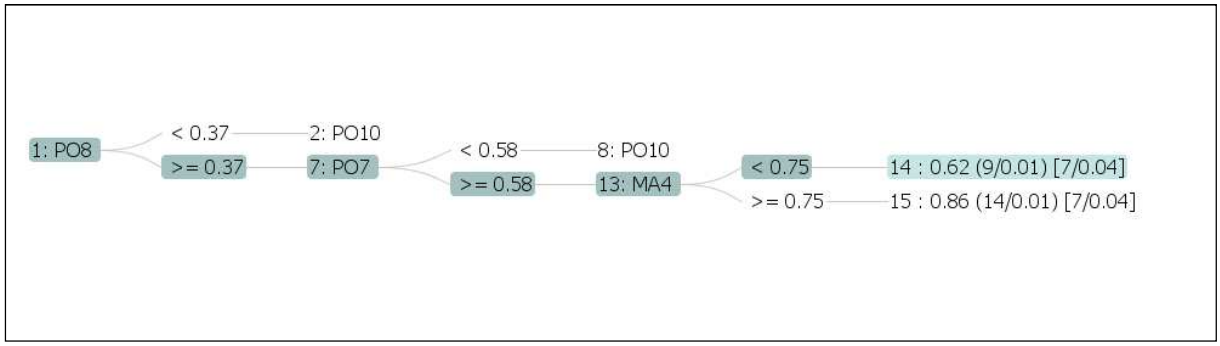
Figura 17: Regra de associação nº 5 e 6



Fonte: Dados da pesquisa

As regras nº 5 e 6 tratam de níveis repetitivos (0,46) e definidos (0,64), respectivamente. As duas regras possuem em comum ocorrerem quando PO8 for maior ou igual a 0,37, PO7 menor que 0,58 e PO10 maior ou igual a 0,56, e apenas diferindo em relação a MA4 (Fornecer governança de TI), que quando for menor de 0,53, PO9 se define como 0,46 e quando for maior ou igual a 0,53 resulta em 0,64.

Figura 18: Regra de associação nº 7 e 8



Fonte: Dados da pesquisa

As regras 7 e 8 apenas diferem no item MA4, que sendo menor de 0,75, reflete um nível de maturidade de PO9 como gerenciado (0,62) e estando maior ou igual a 0,75 é considerado otimizado (0,86). Os demais processos são comuns às duas regras. São eles: PO8 maior ou igual a 0,37 e PO7 maior ou igual a 0,58 e devem estar atendidos da mesma forma para que a regra se confirme.

Nota-se que nem todas as regras tem o mesmo suporte, estando as regras de número 1(25), 4(23), 6(11) e 8(15) com maior suporte.

Verifica-se a presença constante dos processos PO7, PO8 e PO10 na influência sobre o nível do PO9 em detrimento aos demais processos, que embora possuam influência, não geram regras segundo os critérios estabelecidos.

Na próxima etapa foi gerada uma regra de predição numérica, que buscou prever o nível da maturidade do processo PO9 a partir dos demais.

4.2.7 Etapa 11 - Classificação – Regressão

Os algoritmos de predição numérica possuem a função de prever um determinado atributo baseado nos demais, sendo que neste estudo se busca prever o valor (nível de maturidade) do processo PO9 baseado nos outros atributos.

Foi utilizado o arquivo do Weka normalizado, que vem sendo utilizado nas últimas etapas, para a geração desta regra. Para poder aplicar o algoritmo, se seleciona as Funções de Classificação em *Classify* > *Functions* e se seleciona *LinearRegression*.

Nas opções da função, em *Attribute Selection Method* existem as opções de seleção dos atributos M5, Greedy e nenhuma Seleção. Utilizando-se o método de busca M5, se percorre os atributos buscando e removendo os com menor coeficiente que não promovam melhorias no modelo gerado, observando-se a estimativa do erro dada pelo critério de informação de Akaike. No método Greedy utilizam-se as informações geradas pela métrica de Akaike. No método sem nenhuma seleção, são utilizados todos os atributos disponíveis para gerar a regra.

Na Tabela 8 é possível observar os resultados de regressão linear com os três métodos de busca.

Tabela 8: Resultado algoritmo de regressão

Indicador	Greedy	M5P	Sem método de busca
Coeficiente de Correlação	0,8642	0,8642	0,8650
Erro médio absoluto	0,1111	0,1111	0,1109
Erro médio quadrático	0,1447	0,1447	0,1443

Fonte: Dados da pesquisa

Na Tabela 8 se verifica que os métodos de busca M5 e Greedy obtiveram o mesmo desempenho quanto aos itens de correlação e de erros médios. Já na opção sem um método de busca houve uma pequena melhora nos resultados, a partir da terceira casa decimal. Os resultados do algoritmo de regressão podem ser visualizados no Quadro 18.

Quadro 18: Resultado da técnica de regressão a partir do processo PO9

	Nível do Processo PO9 é igual a					
Método de Busca	Greedy		M5		Sem método de busca	
Resultados da regressão	0,1955 * PO7 +	19,13%	0,1955 * PO7	19,13%	0,0008 * PO3 +	0,08%
	0,1176 * PO8 +	11,51%	+ 0,1176 * PO8	11,51%	0,1784 * PO7 +	17,48%
	0,1900 * PO10 +	18,59%	+ 0,19 * PO10	18,59%	0,1145 * PO8 +	11,22%
	0,1730 * AI1 +	16,93%	+ 0,173 * AI1	16,93%	0,1826 * PO10 +	17,89%
	0,1614 * AI2 +	15,79%	+ 0,1614 * AI2	15,79%	0,1591 * AI1 +	15,59%
	0,1097 * AI3 +	10,73%	+ 0,1097 * AI3	10,73%	0,1568 * AI2 +	15,36%
	0,1151 * MA3 +	11,26%	+ 0,1151 * MA3	11,26%	0,1060 * AI3 +	10,38%
	-0,0404	-3,95%	-0,0404	-3,95%	0,1031 * MA3 +	10,10%
					0,0587 * MA4 +	5,75%
					-0,0392	-3,84%

Fonte: Dados da pesquisa

Pode-se verificar, a partir do Quadro 18, que os resultados com o método de busca Greedy e M5 foram idênticos, principalmente devido à métrica de Akeike, utilizada em ambos os casos. No caso da regressão sem um método de busca, os resultados foram semelhantes, porém foram utilizados todos os atributos, encontrando o PO3 e MA4, que nos demais não estavam presentes por sua baixa contribuição para o modelo.

A leitura dos resultados presentes no Quadro 18 pode ser feita da seguinte maneira: no Greedy, o valor do nível de maturidade do processo PO9 é igual a 0,1955 x PO7 mais 0,1176 x PO8 mais 0,1900 x PO10 e assim por diante até chegar ao último valor (constante) na qual diminui 0,0404.

Todos os itens de processo apresentaram influência positiva, ou seja, na medida em que o nível de maturidade do referido processo aumenta, o nível do PO9 também aumenta.

Os processos com maior peso, maior para o menor, no nível da maturidade do processo PO9 encontrados através dos métodos Greedy e M5 foram, PO7, PO10, AI1, AI2, PO8, MA3 e AI3, enquanto que no outro método, PO10, PO7, AI1, AI2, PO8, AI3, MA3, MA4 e PO3. Nota-se que houve inversão na ordem de 4 processos (PO7 e PO10; MA3 e AI3), porém a diferença do peso de influência de cada processo foi muito baixa, inclusive ocorrendo isto nos processos que estão presentes na regressão sem um método de busca que foram classificados como de menor peso.

A próxima tarefa foi o uso de algoritmos de associação, onde se busca prever o comportamento de um determinado atributo baseado nos demais, no entanto antes é necessária a realização da discretização dos dados.

4.2.8 Etapa 12 - Discretização

Uma etapa importante ao processo de regras de associaçãoque trata do agrupamento em categorias, uma vez que há muitos dados distintos, recomenda-se agrupar em categorias, e para isso se utiliza da discretização (TANG; MACLENNAN (2005).

A Discretização é um método no qual os atributos contínuos(no presente caso, numéricos) são transformados para unidades individuais- ou categorias, com o objetivo de gerar cálculos com menor complexidade (WITTEN E FRANK, 2005)

Para realizar a função de discretizar, o software Weka possui uma função de pré-processamento chamada *Discretize*, a qual realiza a discretização dos dados conforme os critérios estabelecidos. A função pode ser selecionada da seguinte forma: *filters/unsupervised/attribute/discretize*

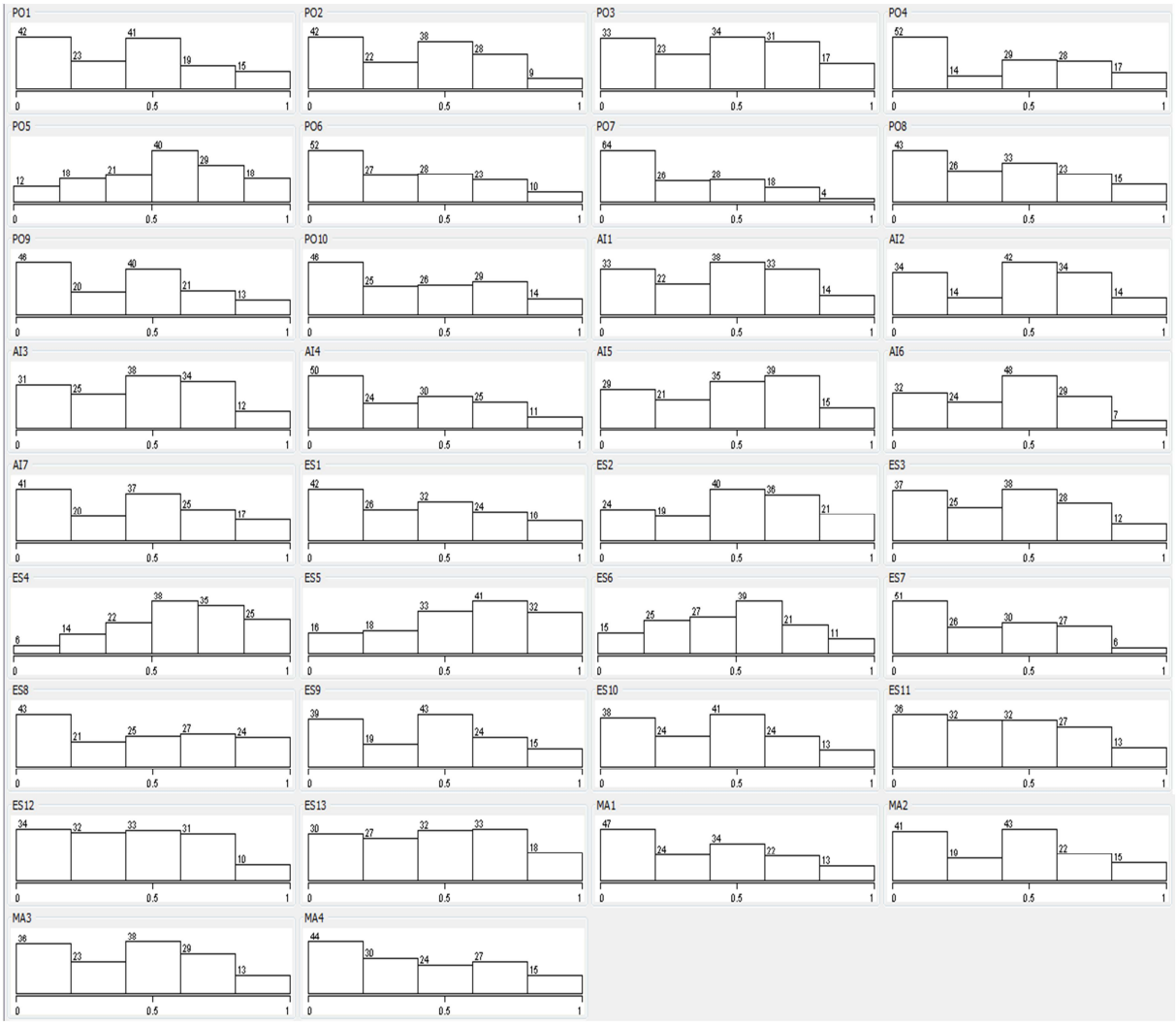
A função de discretização irá gerar categorias a partir dos atributos numéricos. As categorias definidas para o presente estudo são baixo, médio e alto nível da maturidade Para selecioná-las, deve-se ir nas opções da função e, em *bins*, se escolhe o número de categorias que no caso foram três.

A opção *IgnoreClass* deve ser selecionada *True*, para que todas as classes sejam discretizadas porque, do contrário, a última classe que o software considera como indexadora não será discretizada.

A opção *UseEqualFrequency* gera as classes nominais de forma que tenham igual frequência de distribuição. Deve estar *false* para preservar as características de distribuição dos níveis de maturidade.

No Gráfico 11 podem ser vistos os dados antes da discretização e no Gráfico 12 após este processo

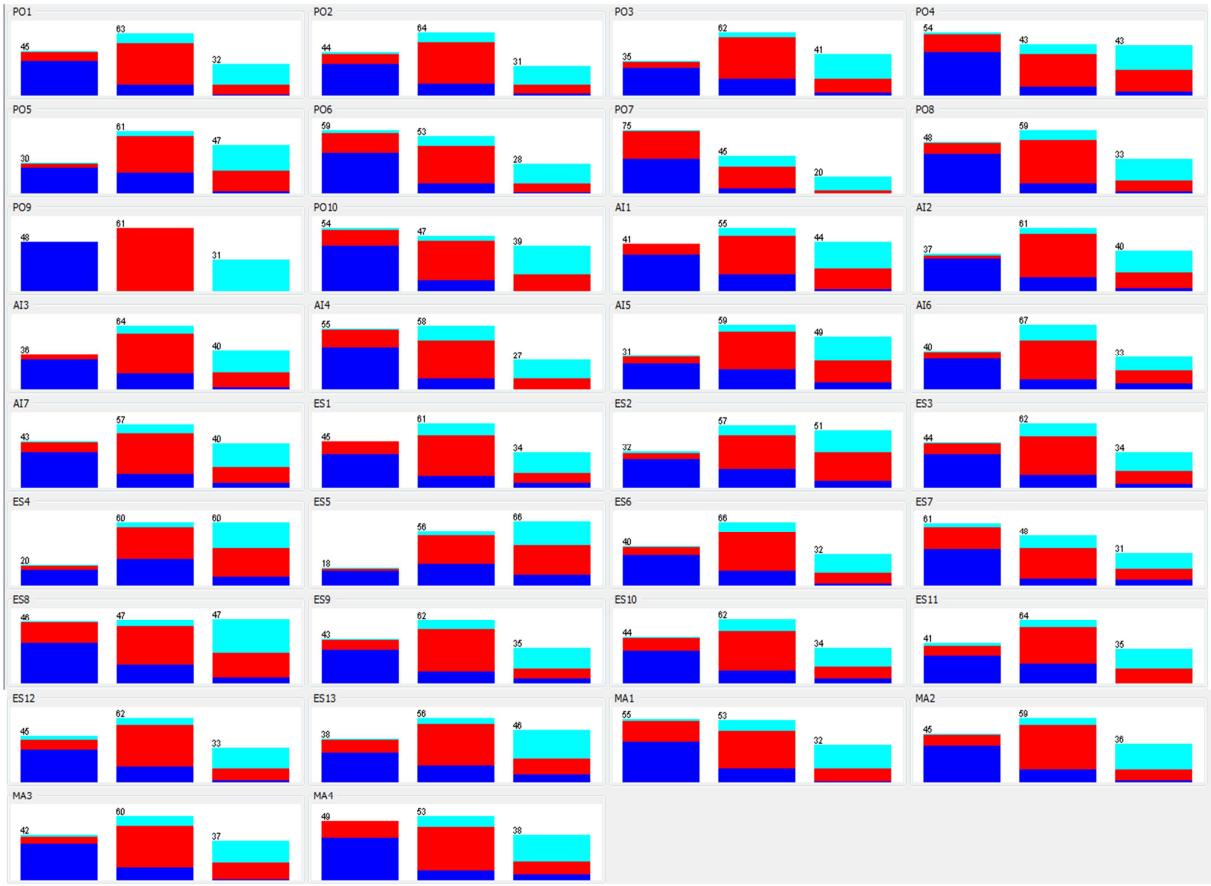
Gráfico 11: Histograma sem discretização



Fonte: Dados da pesquisa

No Gráfico 11 pode ser visualizado cada um dos atributos normalizados, porém não discretizados. Após a discretização, os atributos podem ser visualizados no Grafico 11.

Gráfico 12: Histograma com discretização - PO9



Fonte: Dados da pesquisa

O Gráfico 12 apresenta o resultado dos atributos discretizados em 3 categorias, sendo estas -classificados conforme o critério de maturidade, ou seja, baixo, médio ou alto nível de maturidade no processo.

Na Figura 19 pode ser visto como o software apresenta cada um dos níveis.

Figura 19: Extração do resultado da discretização para o processo PO9

No.	Label	Count
1	'(-inf-0.333333]'	49
2	'(0.333333-0.666667]'	59
3	'(0.666667-inf)'	32

Fonte: Dados da pesquisa

Conforme apresentado na Figura 19, o nível nº 1 (destacado em azul) compreende os níveis mais baixos de maturidade dos processos, e vão desde 0 a 0,333333, encontrando 49 ocorrências no processo PO9.

Com os resultados obtidos da discretização, os dados numéricos foram transformados para variáveis nominais e a partir destes foi gerada uma tabela para entendimento dos níveis de maturidade, que pode ser visualizada na Tabela 9.

Tabela 9: Equivalência entre variáveis numéricas e nominais após discretização

Numérico	Nominal
0 a 0,333333	Baixo
0,333334 a 0.666667	Médio
0,666668 a 1	Alto

Fonte: Dados da pesquisa

A partir da Tabela 9, se verifica que a distribuição dos níveis de maturidade é distribuída de forma equivalente sem, no entanto, comprometer a qualidade da informação quantitativa presente em cada nível, ou seja, conforme o valor numérico do processo ele foi enquadrado no respectivo nível categórico específico não havendo igual frequência de distribuição entre as categorias, permitindo manter a característica principal da informação.

Na sequência são geradas regras de associação entre os dados, as quais foram extraídas a partir do arquivo discretizado.

4.2.9 Etapa 13 - Associação

Os algoritmos de associação buscam prever o comportamento de um determinado atributo a partir de outros.

Nas associações, Tang e Maclennan (2005), afirmam que havendo muitas dados distintos é recomendável agrupar em categorias. Além disto, o algoritmo de associação Apriori, exige para seu funcionamento que os dados não sejam numéricos a fim de possibilitar a geração de regras. Devido a isso, fez-se necessária a etapa anterior de discretização dos dados.

O algoritmo Apriori funciona basicamente a partir de dois elementos, o suporte e a confiança. Por suporte entende-se a quantidade de vezes que um

determinado conjunto de dados ocorre simultaneamente na base de dados. Já a confiança é usada para demonstrar a confiabilidade da regra gerada, ou seja, o percentual de vezes que a regra gerada é confirmada nos dados explorados, e quanto maior, mais confiáveis são os resultados.

A implementação do algoritmo se dá no processo de busca por regras, primeiramente por aquelas de maior suporte e vai reduzindo iterativamente até encontrar o suporte mínimo e o número necessário de regras a uma dada confiança mínima, ou seja, buscará todas as regras que tenham um mínimo de suporte e confiança preestabelecidos.

Para a geração de regras de associação somente foram utilizados aqueles atributos considerados discriminantes para o processo PO9 – Avalia e gerencia riscos, foram: PO3, PO7, PO8, PO10, AI1, AI2, AI3, MA3 e MA4, incluindo o PO9.

A utilização do algoritmo Apriori no Weka ocorre através da aba *Associate*, e opção Apriori. Para a configuração das opções, é dado um duplo clique no nome do algoritmo e surge o menu de itens que podem ser ajustados. Para melhor entendimento, abaixo está presente uma breve explicação dos principais itens a serem configurados.

LowerBoundMinSupport: Item que trata do suporte mínimo desejado. Foi configurado inicialmente para 0,20 (20% - 28 instâncias).

MetricType: Tipo de métrica. A opção *Confidence* é a que trata da confiança.

MinMetric: Segundo o tipo de métrica estabelecido, qual o valor mínimo para que se gere a regra. Foi estipulado 0,90 (90% de confiança)

NumRules: Este item trata do número máximo de regras possíveis de serem geradas. Para encontrar todas as regras possíveis, deve ser indicado um número elevado. No estudo, foram designadas 1000 possíveis regras.

Após a ativação do algoritmo e o respectivo processamento, são apresentados os resultados, encontrando 50 regras válidas segundo os critérios estabelecidos (Suporte de 20% e confiança de 90%). Destas, 22 foram eliminadas por redundância. As demais estão expostas nos Quadros 20, 21 e 22.

As 22 eliminadas apresentavam em sua formação regras mais simples que individualmente já confirmavam a mesma coisa, como pode ser visto no exemplo do Quadro 19.

Quadro 19: Exemplo de regra com redundância

Se		Então	Suporte	Confiança
PO3 =Baixo	PO10 =Baixo	PO7=Baixo	30	0,97

Fonte: Dados da pesquisa

No Quadro 19 é apresentado uma regra redundante, (se PO3 – Determina as diretrizes de tecnologia for baixo e PO10 - Gerencia os projetos for baixo, PO7 - Gerencia os recursos humanos de TI - também será baixo) que é explicado individualmente pelas regras nº 1 (caso o PO3 - for baixo, PO7 também será baixo) e nº 4 (PO10 baixo, PO7 também será baixo). Assim foram eliminadas as regras que não traziam novas informações ao conjunto de regras.

Os resultados encontrados foram organizados em quadros separados pela quantidade de itens de verificação e organizados por ordem de confiança decrescente da maior para a menor.

Embora as regras geradas estejam dispostas continuamente nos resultados do Weka, se optou por separá-las em diversos quadros, conforme o número de atributos de entrada para permitir uma melhor visualização e análise. No Quadro 20 são apresentadas as regras mais simples, com apenas um atributo de entrada.

Quadro 20: Regras simples com um atributo de entrada

Nº	Se	Então	Suporte	Confiança
1	PO3 =Baixo	PO7 =Baixo	33	0,94
2	MA3 =Baixo	PO7 =Baixo	40	0,93
3	AI3 =Baixo	PO7 =Baixo	33	0,92
4	PO10 =Baixo	PO7 =Baixo	50	0,91
5	AI1 =Baixo	PO7 =Baixo	37	0,90

Fonte: Dados da pesquisa

Todas as regras com uma classe de entrada geraram regras para o PO7 baixo. Dentre os 10 processos, somente cinco (PO3, MA3 – Assegura a conformidade aos requisitos externos, AI3 - Adquiri e mantêm a arquitetura tecnológica, PO10 e AI1 – Identifica soluções de automação) possuíram suporte e confiança mínimos, encontrando relações entre as variáveis.

Verifica-se, a partir do Quadro 20, que as regras dos processos MA3 e PO10, respectivamente, representam 28,6% (40) e 35,7% (50) de suporte entre as

instâncias, ou ainda, as regras mais fortes e com maior presença estão presentes em aproximadamente em um terço das instâncias.

Regras com duas classes de entrada são um pouco mais complexas, por envolverem mais de um atributo na sua geração. Isto pode ser visto no Quadro 21.

Quadro 21: Regras mais complexas com dois atributos de entrada

Nº	Se		Então		Suporte	Confiança
6	PO3 =Baixo	MA4 =Baixo	PO10 =Baixo		28	0,97
7	PO7 =Baixo	AI2 =Baixo	PO9 =Baixo		31	0,97
8	PO7 =Baixo	AI2 =Baixo	PO10 =Baixo		30	0,94
9	PO7 =Baixo	AI3 =Baixo	PO10 =Baixo		31	0,94
10	PO9 =Baixo	AI2 =Baixo	PO7 =Baixo		31	0,94
11	PO10 =Baixo	AI2 =Baixo	PO9 =Baixo		30	0,94
12	PO9 =Baixo	AI3 =Baixo	PO10 =Baixo		28	0,93
13	PO9 =Baixo	MA4 =Baixo	PO10 =Baixo		33	0,92
14	PO9 =Baixo	MA4 =Baixo	PO7 =Baixo		33	0,92
15	PO8 =Baixo	MA4 =Baixo	PO7 =Baixo		33	0,92
16	PO8 =Baixo	PO10 =Baixo	PO7 =Baixo		36	0,92
17	PO9 =Baixo	AI2 =Baixo	PO10 =Baixo		30	0,91
18	PO3 =Baixo	PO7 =Baixo	PO10 =Baixo		30	0,91
19	MA3 =Baixo	MA4 =Baixo	PO9 =Baixo		29	0,91
20	PO10 =Baixo	AI2 =Baixo	PO7 =Baixo	PO9 =Baixo	29	0,91
21	PO7 =Baixo	AI2 =Baixo	PO9 =Baixo	PO10 =Baixo	29	0,91
22	PO3 =Baixo	PO10 =Baixo	MA4 =Baixo		28	0,90
23	AI1 =Baixo	MA4 =Baixo	PO10 =Baixo		28	0,90
24	PO8 =Baixo	AI1 =Baixo	PO9 =Baixo		28	0,90

Fonte: Dados da pesquisa

A leitura do Quadro 21 pode ser feita da seguinte maneira: Se na regra nº 6 informar que se os processos PO3 e MA4 - Fornecer governança de TI possuem baixos níveis de maturidade, o PO10 também será baixo, com suporte de 28 e confiança de 97%.

Observa-se, a partir do Quadro 21, que todas as regras geradas são de baixo nível de maturidade. Foram encontradas 19 das 27 regras válidas (70,4%) com dois atributos de entrada. A maioria das regras encontradas se referem aos processos - PO10 (8) seguidos por PO7(5), PO9(5) - Avalia e gerencia riscos - e uma para MA4.

Apenas as regras nº 20 e 21 possuem dois atributos de saída e podem ser lidas. Por exemplo na regra nº 20, se PO10 e AI2 - Adquiri e mantêm software aplicativo forem baixos, PO7 e PO9 também serão baixos.

No Quadro 22 podem ser vistas regras com três atributos de entrada.

Quadro 22: Regras mais complexas com três atributos de entrada

Nº	Se			Então	Suporte	Confiança
25	PO7 =Baixo	PO8 =Baixo	PO9 =Baixo	PO10 =Baixo	31	0,94
26	PO7 =Baixo	PO8 =Baixo	MA4 =Baixo	PO10 =Baixo	30	0,91
27	PO8 =Baixo	PO10 =Baixo	MA4 =Baixo	PO9 =Baixo	29	0,91

Fonte: Dados da pesquisa

A leitura do Quadro 22, na regra nº 25, afirma que se PO7, PO8 (Gerencia a qualidade) e PO9 forem baixos, PO10 também o será, com suporte de 31 e confiança de 94%.

Verifica-se também a frequência recorrente das classes PO7, PO8, PO9 e PO10, que mostram bastante proximidade, como comprovado nas regras encontradas.

De maneira mais abrangente, as regras de 1 a 27 mostram que as classes previstas com maior frequência são PO7 – Gerencia os recursos humanos de TI (10 - 34,5%), PO9 – Avalia e gerencia riscos (7 - 24,1%), PO10 – Gerencia os projetos (9 - 31%) e somente uma regra gera previsão do processo MA4 – Fornecer governança de TI (3 - 10,3%).

Na aplicação inicial, não foram encontradas regras com nível médio ou alto de maturidade entre os processos. Optou-se pela geração de novas simulações iterativamente com o objetivo de buscar regras que estivessem neste intervalo de nível médio a alto. As novas aplicações obtiveram resultado positivo, porém com um suporte menor embora tenham mantido o nível de confiança (90%).

Foram encontradas 547 regras, possuindo suporte mínimo de 15% (21 instâncias) e confiança de 90%. Foram desconsideradas nesta busca as regras de baixo nível de maturidade e devido as de nº 1 a 27 possuírem maior suporte e de igual forma, as redundantes também não foram consideradas.

Nos Quadros 23, 24 e 25 podem ser visualizados os resultados da busca de regras (nº 28 a 43) com níveis médio e alto de maturidade.

Quadro 23: Regras com nível de maturidade médio com três atributos de entrada

Nº	Se			Então	Suporte	Confiança
28	PO9=Médio	PO10=Médio	MA4=Médio	MA3=Médio	22	0,96
29	PO8=Médio	PO10=Médio	MA3=Médio	PO9=Médio	24	0,96
30	PO3=Médio	AI1=Médio	MA4=Médio	PO8=Médio	21	0,95
31	PO8=Médio	PO9=Médio	PO10=Médio	MA3=Médio	24	0,92
32	PO3=Médio	PO8=Médio	PO10=Médio	MA3=Médio	22	0,92
33	PO9=Médio	PO10=Médio	AI2=Médio	PO3=Médio	22	0,92
34	PO9=Médio	PO10=Médio	MA3=Médio	PO8=Médio	24	0,92
35	PO3=Médio	PO8=Médio	PO10=Médio	PO9=Médio	22	0,92
36	PO8=Médio	PO10=Médio	AI2=Médio	PO3=Médio	21	0,91
37	PO10=Médio	AI2=Médio	MA3=Médio	PO3=Médio	21	0,91
38	PO9=Médio	PO10=Médio	MA4=Médio	PO8=Médio	21	0,91
39	AI1=Médio	AI3=Médio	MA4=Médio	PO8=Médio	21	0,91
40	PO8=Médio	PO10=Médio	AI2=Médio	PO9=Médio	21	0,91

Fonte: Dados da pesquisa

O Quadro 23 apresenta 13 regras com 3 atributos de entrada cada uma e fica evidenciado que não foram encontradas regras que misturassem níveis de maturidade, ou seja, o conjunto dos atributos de nível médio prevê em seu resultado atributos de nível médio.

Nas regras geradas, os processos PO (Planejar e Organizar) possuem maior frequência, no entanto foram encontradas 3 regras que previam processos de MA (Monitorar e Avaliar) com nível médio de maturidade.

No Quadro 24 podem ser observadas as regras de nível médio com 4 processos de entrada.

Quadro 24: Regras com nível de maturidade médio com quatro atributos de entrada

Nº	Se				Então	Suporte	Confiança
41	PO3=Médio	PO9=Médio	AI2=Médio	MA3=Médio	PO8=Médio	21	0,95
42	PO3=Médio	PO8=Médio	AI2=Médio	MA3=Médio	PO9=Médio	21	0,95
43	PO8=Médio	PO9=Médio	AI2=Médio	MA3=Médio	PO3=Médio	21	0,91

Fonte: Dados da pesquisa

Pode ser visualizado no Quadro 24, que as regras geradas possuem o suporte mínimo de 15% (21 instâncias) e prevêem somente processos PO. Nas três

regras geradas com 4 atributos de entrada, todas possuem dois processos, sendo um PO, um AI (Adquirir e Implementar) e um MA (Monitorar e Avaliar).

De igual forma que apresentado no Quadro 24, pois neste não ocorre mistura de níveis, estando todos no nível médio de maturidade, e isto se repete no Quadro 25 com alto nível de maturidade.

Entre os níveis intermediários de maturidade o único processo que não está presente em nenhuma das regras geradas é o PO7, lembrando que este era bastante recorrente entre os níveis baixos de maturidade.

Em relação a regras de nível médio de maturidade envolvendo o gerenciamento e avaliação de riscos (PO9), foram encontradas 4 que o previa, nas regras 29, 35, 40 e 42, ou seja, 25% das 16 regras previram o atributo.

Se verifica que 10 entre 16 regras de nível médio de maturidade possuem o atributo de PO9 na sua composição, confirmando a significativa relação com os demais atributos, encontrados na etapa de seleção de atributos.

No Quadro 25 pode ser observadas regras de nível alto de maturidade.

Quadro 25: Regras com nível alto de maturidade

Nº	Se		Então	Suporte	Confiança
44	PO10=Alto	MA4=Alto	AI1=Alto	22	0,92
45	PO9=Alto	MA4=Alto	AI1=Alto	21	0,91

Fonte: Dados da pesquisa

Segundo pode verificar-se no Quadro 25, para níveis altos de maturidade somente duas regras foram encontradas, e previram processos do AI (Adquirir e Implementar), diferente das demais regras, previram processos do PO (Planejar e Organizar) e MA (Monitorar e Avaliar).

Pode-se verificar, que entre as regras encontradas, as que apresentaram maior suporte e maior confiança, são aquelas que possuem baixos níveis de maturidade em seus processos e que no conjunto das regras geradas, todos os processos estiveram envolvidos em pelo menos um regra.

4.3 Análise final e documentação

O estudo desenvolvido teve como objetivo analisar as relações entre risco operacional e processos tecnológicos e foi utilizado o método de DR em uma aplicação de DM. Foram coletadas 341 respostas e, após o processo de limpeza, restaram 301 consideradas válidas, representando 88,3% do total. Estas, multiplicadas pelas 42 variáveis/classes, totalizaram 12.642 dados. Após uma nova limpeza, mais específica, além da transformação dos dados, foram contabilizados 39 atributos e 140 instâncias, somando 5.460 dados.

As localidades que representaram, em conjunto, a maior parte das cidades onde as empresas presentes no estudo estavam situadas foram Porto Alegre (POA), São Leopoldo (SL) e Novo Hamburgo (NH) com 50,7% das empresas.

As empresas do Vale do Rio dos Sinos representam aproximadamente a metade (45,7%) das empresas presentes estudadas.

Quanto ao porte, a categoria que apresentou maior frequência foi a de empresas de médio porte (36,4%) seguidas pelas de grande porte (33,6%). A diferença entre elas é de apenas 4 respostas e em conjunto elas correspondem a 70% das respostas.

Verificou-se que as empresas de grande porte possuem seus processos com maior nível de maturidade, estando classificados entre definido e gerenciado e estas empresas estão localizadas principalmente na capital. As empresas de médio porte encontram-se com os processos em menor nível de maturidade e sua distribuição por região foi mais equilibrada, entre capital, interior e vale dos sinos, possuindo aproximadamente 1/3 para cada uma das três regiões.

Verifica-se que PO7 (Gerencia os recursos humanos de TI) foi o processo com menor nível de maturidade em relação aos demais processos do mesmo agrupamento, inclusive sendo o único com média abaixo de 1 nas pequenas empresas. Isto é um indicativo de baixa preocupação por parte das empresas em relação a este processo, o que as tornem mais suscetíveis a riscos nas operações que envolvam recursos humanos.

O processo ES5 (Garante a segurança dos sistemas) apresentou o mais alto nível em cada um dos clusters nas empresas de porte pequeno, médio e grande e

nas micro empresas este processo apresentou o mesmo nível que ES4 (Garante a continuidade dos serviços).

Os indicativos de baixo nível de maturidade em processos ligados aos recursos humanos e alto nível nos processos ligados à segurança dos sistemas e hardware não propicia um nível de segurança adequadopois, segundo o estudo de SPEARS; BARKI (2010), geram riscos operacionais em função de falhas ou inadequações junto a processos internos que podem ter a participação do usuário Estas inconformidades, mesmo que não intencionais, estão relacionadas a fatores ligados aos recursos humanos, entre eles, baixo nível de treinamento.

Na geração de cluster por região, os processos de mais alto nível de maturidade coincidem com o fato da maioria das empresas de grande porte (59,76%) presentes no estudo são oriundas da capital. As empresas do interior mostraram que possuem um nível de maturidade intermediário (repetível – definido) com presença mais marcante de empresa de médio (59,68%) e pequeno porte (42,93%). Já as do Vale do Rio dos Sinos, que representam 35% das empresas pesquisadas, detêm os menores índices de maturidade e são constituídas por 81,25% das micros empresas da pesquisa e cerca de metade das pequenas empresas (53,84%).

O processo PO5 (Gerencia o investimento de TI) na região do Vale dos Sinos e Interior se sobressai em relação a média, no entanto na capital o processo se encontra dentro da média.

As empresas das cidades do Vale do Rio dos Sinos possuem seu nível de maturidade mais disperso, enquanto que as empresas da capital e interior possuem seus níveis de maturidade mais concentrados.

Ficou evidenciado que a grande maioria das empresas foram classificadas de maneira igual independentemente do algoritmo utilizado, indicando uma convergência na forma de classificar entre eles e uma maior confiabilidade nos agrupamentos encontrados. Onde 93,6% das respostas foram classificadas igualmente, ou seja, a ampla maioria das 140 empresas foram qualificadas no mesmo cluster nos dois diferentes algoritmos, o EM e o Kmeans, mostrando consistência nos agrupamentos obtidos.

Não se observou qualquer indicação de empresa cujo agrupamento foi classificado como de baixo nível de maturidade por um algoritmo e por outro de alto nível de maturidade.

A partir dos 34 processos do Cobit, os algoritmos encontraram 12 processos que estariam relacionados com o processo PO9 (Gerencia e avalia riscos), e somente 9 que estavam presentes em todos em todos os métodos de busca. Como critério de seleção de atributos e refinamento da busca, optou-se por manter os atributos que possuíam maior poder discriminante em relação aos demais e, para isso, foram selecionados aqueles que estavam contidos em todos os métodos de busca, ou seja, 9 atributos.

Embora os processos ES6, ES8 e ES11 tenham aparecido na busca, eles não estavam presentes em todos os métodos, indicando que os demais métodos não os consideraram suficientemente discriminantes e em função disto foram excluídos.

O algoritmo Kmeans reduzido classificou as empresas de igual forma que o EM total e o Kmeans total em 75,71% e 78,57% respectivamente. Isso indica que os clusters do algoritmo Kmeans permanecem com aproximadamente 3/4 dos centroides iguais, isto significa dizer que as características dos agrupamentos encontrados no total dos processos e nos reduzidos é idêntico em aproximadamente 3/4 das empresas, e que o nível de maturidade dos processos reduzidos tende a apresentar os mesmos níveis de maturidade quando avaliado os demais processos, não apresentando variação acentuada entre os mesmos.

O resultado dos algoritmos REPTree e M5P mostraram-se semelhantes, porém o algoritmo RepTree apresentou melhores resultados em relação aos erros, e por isso foi selecionado para mostrar as conclusões encontrados através de ferramenta gráfica *Profuse,* que facilitaram a compreensão das regras geradas para o atributo alvo de riscos operacionais.Foram geradas 8 regras, observando-se que apresentam suportes diferentes, estando as regras de número 1 com suporte de 25, 4(23), 6(11) e 8(15), sendo estas as de maior suporte.

Foi verificado que os processos PO7, PO8 e PO10 possuem maior influência sobre o nível do PO9 em detrimento aos demais processos, que embora possuam influência, não geram regras segundo os critérios estabelecidos.

Na regressão, verificou-se que os métodos de busca M5 e Greedy obtiveram o mesmo desempenho quanto aos itens de correlação e erros médios. Já na opção sem um método de busca houve uma pequena melhora nos resultados, a partir da terceira casa decimal

Os resultados com o método de busca Greedy e M5 foram idênticos, principalmente devido à métrica de Akeike, utilizada em ambos os casos. No caso da regressão sem um método de busca, os resultados foram semelhantes porém foram utilizados todos os atributos, encontrando o PO3 e MA4, que nos demais não estavam presentes por sua baixa contribuição para o modelo.

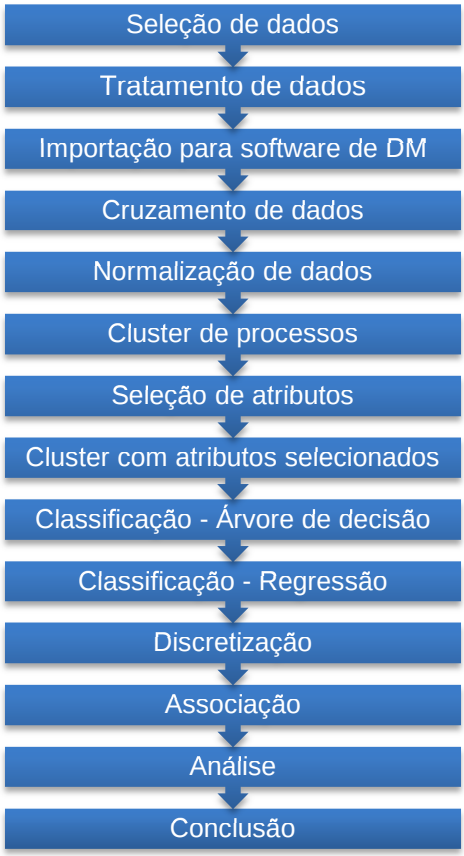
Todos os itens de processo apresentaram influência positiva, ou seja, na medida em que o nível de maturidade do referido processo aumenta, o nível do PO9 também sobe.

Os processos com maior peso, do maior para o menor, no nível da maturidade do processo PO9 encontrados, através dos métodos Greedy e M5 foram: PO7, PO10, AI1, AI2, PO8, MA3 e AI3, enquanto que sem um método de busca no foram: PO10, PO7, AI1, AI2, PO8, AI3, MA3, MA4 e PO3.

Em relação a ordem do peso dos processos, houve inversão em 4 deles(PO7 e PO10; MA3 e AI3), porém a diferença do peso de influência de cada um foi muito baixa, inclusive ocorrendo que processos que estavam presentes na regressão sem um método de busca, acabaram sendo classificados como de menor peso.

Assim, após a aplicação da metodologia do DR e DM, bem como a análise, evidencia-se o artefato gerado como um método para análise de relações entre processos tecnológicos, que pode ser visualizado na Figura 20.

Figura 20: Artefato gerado



Fonte: Elaborado pelo autor

O artefato gerado (Figura 20), é um método que pode ser aplicado para verificação das relações entre os níveis de maturidade de cada um dos 34 processos do Cobit 4.1, bem como pode ser utilizado no Cobit 5 (versão mais recente), uma vez que se aplicam os níveis de maturidade aos processos.

O método proposto apresenta um resumo da abordagem e aplicação desta pesquisa, demonstrando que a interação entre o método de *Design Research* e *Data Mining* proposto permite encontrar uma grande quantidade de informações uteis e de qualidade significativa para dar suporte a decisões com maior acurácia e menor nível de subjetividade, permitindo a organização uma maior transparência e objetividade nas decisões relacionadas aos processos tecnológicos.

5. CONCLUSÃO

O objetivo geral deste estudo consistiu na análise da relação entre risco operacional e processos tecnológicos, considerando a análise das relações entre os processos de governança de TI, a análise dos processos que possuem influência sobre a avaliação e gestão de riscos, bem como a definição e o processamento de diferentes regras de garimpagem de dados relacionadas a avaliação e gestão de risco.

Tendo sido considerados principalmente os temas teóricos de governança de TI, risco operacional e *Data Mining*, foram analisados 34 processos pertencentes ao modelo de governança de TI Cobit versão 4.1. O método utilizado para o desenvolvimento deste estudo foi definido como *Design Research* integrado às etapas do *Data Mining,* para extração e geração de novos conhecimentos sumariamente descritos na continuação.

A coleta de dados resultou em 341 respostas em 140 empresas sobre os processos de governança de TI, gerando um total de 5.460 dados, nos quais foram aplicados algoritmos de geração de agrupamentos (EM e Kmeans), seleção de atributos (CfsSubsetEval), classificação (REPTree, M5P e LinearRegression) e associação (Apriori) com a utilização do software de código aberto *Waikato Environment for Knowledge Analysis* (Weka).

Mediante a seleção de atributos verificou-se entre os 34 processos, os que possuíam maior poder discriminante em relação ao atributo de risco (PO9). O algoritmo CfsSubsetEval verificou os atributos correlacionados com a classe e não correlacionados entre si e foram encontrados 9 atributos (processos do Cobit) que possuem relação com o atributo de risco. Estes são: PO3 (Determina as diretrizes de tecnologia), PO7 (Gerencia os recursos humanos de TI), PO8 (Gerencia a qualidade), PO10 (Gerencia os projetos), AI1 (Identifica soluções de automação), AI2 (Adquire e mantêm software aplicativo), AI3 (Adquire e mantêm a arquitetura tecnológica), MA3 (Assegura a conformidade aos requisitos externos) e MA4 (Fornece governança de TI).

Foi evidenciado que o algoritmo Kmeans dos processos selecionados atribuiu os agrupamentos às empresas de igual forma que o algoritmo EM e Kmeans com

todos os 34 processos, em 75,71% e 78,57%, respectivamente. Isto significa dizer que as características dos agrupamentos encontrados no total dos processos e nos reduzidos são idênticas em aproximadamente 3/4 das empresas e que o nível de maturidade dos processos reduzidos tende a apresentar os mesmos níveis de maturidade quando avaliados os demais processos, não apresentando variação acentuada entre os mesmos.

Foram encontrados também quatro clusters de empresas distribuídos da seguinte maneira:

- O primeiro possui características de nível Gerenciado e quase Otimizado;
- O segundo, as empresas possuem nível Definido.
- O terceiro cluster, as empresas possuem nível Repetível
- O quarto e o último cluster, as empresas possuem nível entre Inexistente e Inicial.

Ficou evidenciado que não existe cluster para nível Otimizado, o que significa afirmar que na amostra destas empresas ainda é necessário uma melhoria dos processos tecnológicos.

Quanto aos algoritmos de Classificação (REPTree), foi gerada uma árvore de decisão com 8 regras que prevê o nível de maturidade do atributo de risco (PO9) a partir dos demais. Nas regras geradas, verificou-se que os processos PO7, PO10 e PO8 foram os com maior presença naquelas onde o nível do PO9 estava relacionado. Esta constatação também foi confirmada pelo algoritmo LinearRegression, quando os mesmos se encontraram entre os cinco atributos de maior peso (PO7, PO10, AI1, AI2, PO8) na regra de previsão numérica.

Todos os 9 processos selecionados apresentaram influência positiva, ou seja, na medida em que cada nível de maturidade de cada um dos processos aumenta, o nível do PO9 também é aumentado.

O uso do algoritmo Apriori (Associação) gerou 27 regras válidas do total de 50 regras encontradas com suporte mínimo de 28 instâncias (20%) e confiança mínima de 90%. Todas as regras geradas anteviram níveis baixos de maturidade nos processos envolvidos, sendo que as previstas com maior frequência foram PO7 (10 - 34,5%), PO9 (7 - 24,1%), PO10 (9 - 31%) e somente uma regra do processo MA4 (1 – 3,4%).

Através do algoritmo Apriori também foram geradas regras que predisseram níveis médios e altos de maturidade, no entanto com suporte mínimo de 21 instâncias (15%) e mantendo-se a confiança em 90%. Foram encontradas 547 regras, das quais somente 15 previram níveis médios e 2 níveis alto de maturidade.

No mesmo uso de regras geradas pela associação, foi possível verificar que aquelas que apresentaram maior suporte e maior confiança, foram as que possuíam baixos níveis de maturidade em seus processos. Já no conjunto das regras geradas, todos os processos que possuem relação com o atributo de risco (PO9) estiveram envolvidos ao menos em uma regra, exceto o processo PO7.

Por fim, o texto final da dissertação também seguiu pelo método *Design Research* fechando o ciclo de geração de conhecimento da análise das relações entre risco operacional e processos tecnológicos bem como gerando um artefato que se constituiu em um método para análise das relações entre processos tecnológicos, utilizado na análise de governança de TI.

Recomendações para estudos futuros

Dadas as conclusões encontradas neste estudo, recomenda-se como estímulo aos desafios da pesquisa em governança de TI e riscos:

1. Expansão da pesquisa para os diversos estados do país com o objetivo de gerar comparativos entre as realidades de governança de TI presentes nos diferentes locais.

2. Gerar e aplicar a técnica de mineração de dados a partir de informações de empresas que utilizam a norma ISO 27001, buscando a identificação de padrões de risco a nível operacional.

REFERÊNCIAS

AMARAL, I. C.; NEVES, M. C.; FREITAS, A. F.; BRAGA, M. Gerenciamento dos riscos operacionais: os métodos utilizados por cooperativa de crédito. **RCO – Revista de Contabilidade e Organizações – FEARP/USP**, v. 3, n. 7, p. 93-108, 2009.

AGRAWAL, R.; SRIKANT, R. Fast Algorithms for Mining Association Rules in Large Databases. In: 20th International Conference on Very Large Data Bases, 478-499, 1994.

ANGELOS, E. W. S. D.; SAAVEDRA, O. R.; CORTÉS, O. A. C.; SOUZA, A. N. D. Detection and identification of abnormalities in customer consumptions in power distribution systems. **IEEE Transactions on Power Delivery**, v.26, n.4, p. 2436-2442, 2011.

ARTHUR, D.; VASSILVITSKII, S. k-means ++ : The Advantages of Careful Seeding Eighteenth annual ACM-SIAM symposium on Discrete algorithms. Anais..., 2007.

BAI, X; NUNEZ, M.; KALAGNANAM, J. Managing data quality risk in accounting information systems. **Information Systems Research**, v. 23, n. 2, p. 453–473, 2012.

BAJO, J.; BORRAJO, M.L.; PAZ, J. F.D.; CORCHADO, J. M.; PELLICER, M. A multi-agent system for web-based risk management in small and medium business. **Expert Systems with Applications**, v. 39, n. 8, p. 6921–6931, jun. 2012.

BHATTACHARYA, B.; SOLOMATINE, D. P. Neural networks and M5 model trees in modelling water level–discharge relationship. **Neurocomputing**, v. 63, p. 381–396, 2005.

BANK FOR INTERNATIONAL SETTLEMENTS - BIS. **Basel Committee on Banking Supervision** – International Convergence of Capital Measurement and Capital Standards. 2006. Disponível em: <http://www.bis.org/publ/bcbs128.pdf>. Acesso em: 13 nov. 2012.

_____. **Sound practices for the management and supervision of operational risk.** Bank for International Settlements. Fev, 2003. Disponível em: <www.bis.org>. Acesso em: 16 set. 2012.

BERNSTEIN, P. L. **Against the gods:** the remarkable story of risk. New York: John Wiley & Sons, 1996.

CAMPOS, R.H.; ROCHA, R.; COBO, A.; VANTI, A.A. Measuring technological maturity levels with data mining. In: 11º CONTECSI - The 11th International

Conference on Information Systems and Technology Management. Anais... São Paulo:FEA-USP, 2014.

CHAKRABARTI, A. A course for teaching design research methodology. **Artificial Intelligence for engineering design, analysis and manufacturing,** v. 24, p.317-334, 2010.

CHENG, C.; LEU, S.; CHENG, Y.; WU, T.; LIN, C. Applying data mining techniques to explore factors contributing to occupational injuries in Taiwan's construction industry. **Accident Analysis and Prevention,** v. 48, p.214–222, 2012.

COBO, A.; ROCHA, R.; VANTI, A.A.; SCHNEIDER, G. Fuzzy clustering: application on organizational metaphors in brazilian companies. **Journal of Information Systems and Technology Management**. v. 9, n. 2, p.197-212, 2012.

COMMITTEE OF SPONSORING ORGANIZATIONS OF A TREADWAY COMMISSION - COSO. **Gerenciamento de Riscos Corporativos** – Estrutura Integrada, 2009. Disponível em: <http://www.coso.org/Publications/ERM/COSO_ERM_ExecutiveSummary.pdf>. Acesso em: 14 abr. 2012.

CRESWELL, J. W. **Projeto de Pesquisa:** métodos qualitativo, quantitativo e misto. 2ª ed. Porto Alegre: Artmed, 2007.

DEBRECENY, R. S.; GRAY, G. L. IT Governance and Process Maturity: A Multinational Field Study. **Journal of Information Systems**, v. 27, n. 1, p. 157–188, jun. 2013.

DEMPSTER, A.M. An operational risk framework for the performing arts and creative industries. **Creative Industries Journal**, v.1, n. 2, p.151-170, 2008.

DODGE, Y. **The Oxford Dictionary of Statistical Terms.** Oxford: Oxford University Press, 2003.

FAYYAD, U. M.; PIATESKY-SHAPIRO, G.; SMYTH, P. **From data mining to knowledge discovery: an overview**. Advances in Knowledge Discovery and Data Mining, AAAI Press, 1996.

GARTNER, I.; ZWICKER, R.; RÖDDER, W. Investimentos em Tecnologia da Informação e impactos na produtividade empresarial: uma análise empírica à luz do paradoxo da produtividade. **RAC,** v.13, n.3, p.391-409, 2009.

GOLDSCHMIDT, R.; PASSOS, E **Data Mining**: um guia prático. Rio de Janeiro: Elsevier, 2005

GOLDSTEIN, J.; CHERNOBAI, A.; BENAROCH, M. An event study analysis of the economic impact of IT operational risk and its subcategories. **Journal of the Association for Information Systems,** v.12, n. 9, p. 606-631, 2011.

GONZÁLEZ, P. C..; VELÁSQUEZ, J. D. Characterization and detection of taxpayers with false invoices using data mining techniques. **Expert Systems with Applications**, v. 40, n. 5, p. 1427–1436, 2013.

GUIMARAES, I. C.; PARISIS, C.; PEREIRA, A. C.; WEFFORT, E. F. J. A importância da controladoria na gestão de riscos das empresas não-financeiras: um estudo da percepção de gestores de riscos e controllers. **Revista Brasileira de Gestão de Negócios**, v. 11, n. 32, 2009.

GUETLEIN, M.; FRANK, E.; HALL, M.; KARWATH, A. Large Scale Attribute Selection Using Wrappers. In: Proc IEEE Symposium on Computational Intelligence and Data Mining. Anais… United States: IEEE Computational Intelligence Society, p. 332-339, 2009.

HAIJA, M. F. A. E.; HAYEK, A. F. A. operational risk disclosures in Jordanian commercials banks: it's enough. **International Research Journal of Finance and Economics**, n. 83, p. 49-61, 2012.

HALL, M. A. **Correlation-based Feature Subset Selection for Machine Learning**. New Zealand: Hamilton, 1998.

HE, L.; QI, J. Enterprise Human Resources Information Mining Based on Improved Apriori Algorithm. **Journal of Networks**, v. 8, n. 5, p. 1138–1146, 2013.

HERATH, H. S. B.; HERATH, T. C. IT security auditing: A performance evaluation decision model. **Decision Support Systems**, v. 57, p. 54–63, 2014.

HEVNER, A. R.; MARCH, S. T.; PARK, J.; RAM, S. Design science in information systems research. **Management Information Systems Quarterly**, v. 28, n. 1, p. 75–105, 2004.

INFORMATION SYSTEMS AUDIT AND CONTROL ASSOCIATION – ISACA. **COBIT** - Control Objectives for Information and related Technology. 4.1. Disponível em:<http://www.isaca.org/AMTemplate.cfm?Section=Downloads&Template=/MembersOnly.cfm&ContentFileID=14002>. Acesso em: 05 abr. 2014.

KERR, D. S.; MURTHY, U. S. The importance of the CobiT framework IT processes for effective internal control over financial reporting in organizations: An international survey. **Information & Management**, v. 50, n. 7, p. 590–597, 2013.

KUHN JR., J. R.; AHUJA, M.; MUELLER, J. An examination of the relationship of IT control weakness to company financial performance and health. **International Journal of Accounting and Information Management**, v. 21, n. 3, p. 227–240, 2013.

LAMBRINOUDAKIS, C. Evaluating and enriching information and communication technologies compliance frameworks with regard to privacy. **Information Management & Computer Security**, v. 21, n. 3, p. 177–190, 2013.

LAROSE, D. T. **Discovering knowledge in data**: an introduction to data mining. Haboken: Wiley-Interscience, 2005.

LIU, B; HSU, W.; MA, Y.: Integrating Classification and Association Rule Mining. In: Fourth International Conference on Knowledge Discovery and Data Mining, 80-86, 1998.

MANSON, N. J. Is operations research really research? **Operations Research Society of South Africa - ORSSA**, v. 22, n. 2, p. 155-180, 2006.

MARKUS, M. L.; MAJCHRZAK, A.; GASSER, L. A design theory for systems that support emergent knowledge processes. **MIS Quarterly**. v. 26, n. 3, p.179-212, 2002.

MURAYAMA, S.; OKUHARA, K.; SHIBATA, J.; ISHII, H. data mining for hazard elimination through text information in accident report. **Asia Pacific Management Review**, v.16, n.1, p. 65–81, 2011.

OLIVEIRA, A. J. F. Método para avaliação de riscos operacionais em bancos. 2004. 143 f. **Dissertação (Mestrado em Engenharia de Produção) – Universidade Federal de Santa Catarina**, Florianópolis, 2004.

OLIVEIRA, G.; PACHECO, M. **Mercado Financeiro.** 2ªed. São Paulo: Fundamento Educacional, 2011.

REVATHI, T.; SUMATHI, P. A Survey on Data Mining using Clustering Techniques. **International Journal of Scientific & Engineering Research**, v. 4, n. 1, p. 1–5, 2013.

SAVIC, A. Managing IT-related operational risks. **Ekonomski Anali,** v. 53 n.176 p.88-109, 2008.

SCHOLL, H. J.; PATIN, B. J. Resilient information infrastructures: Criticality and role in responding to catastrophic incidents. **Transforming Government: People, Process and Policy**, v. 8, n. 1, p. 28–48, 2014.

SENG, J. L.; CHEN, T. C. An analytic approach to select data mining for business decision. **Expert Systems with Applications**, v. 37, n. 12, p. 8042-8057, 2010.

SHARMA, S.; OSEI-BRYSON, K.M.; KASPER, G.M. Evaluation of an integrated Knowledge Discovery and Data Mining process model. **Expert Systems with Applications** v. 39 p.11335–11348, 2012.

SHMUELI, G.; KOPPIUS, O. Predictive analytics in information systems research. **MIS Quarterly**, v. 35, n. 3, p. 553–572, 2011.

SHIRI, M. M.; AMINI, M. T.; RAFTAR, M. B. Data Mining techniques and predicting corporate. **Interdisciplinary Journal of Contemporary Research in Business**. v.3, n. 2, p. 61–69, 2012.

SILVA, A. J.; FERNANDES, F. C.; GRANDE, J. F. Controles Internos como Elementos de Mitigação e Gestão de Riscos: Um Estudo nas Maiores Empresas do Sul do Brasil. In: XI SEMEAD – SEMINÁRIOS EM ADMINISTRAÇÃO. Anais... São Paulo:FEA-USP, 2008.

SILVA, C. S.; RALHA, C. G.. Detecção de Cartéis em Licitações Públicas com Agentes de Mineração de Dados. **Revista Eletrônica de Sistemas de Informação**, v. 10, n. 1, p.1-19, 2011.

SILVA, M. S. **Mineração de Dados** - Conceitos, Aplicações e Experimentos com Weka. 2004. Disponível em:<http://bibliotecadigital.sbc.org.br/?module= Public&action=Publication Object&subject=154&publicationobjectid=9>. Acesso em: 15 set. 2012.

SPEARS, J. L.; BARKI, H. User participation inf information systems security risk management. Management Information Systems. **MIS Quarterly Executive**, v. 34, n. 3, p. 503-522, 2010.

TAKEDA, H.; VEERKAMP, P.; TOMIYAMA, T.; YOSHIKAWA, H. Modeling design processes. **Artificial Intelligence Magazine**, v. 11, n. 4, p. 37-48, 1990.

TAN, P.; STEINBACH, M.; KUMAR, V. **Introduction to data mining**. United States: Addison-Wesley Longman Publishing, 2009.

TANG, Z.; MACLENNAN, J. **Data mining with SQL server 2005**. Indianopolis: Wiley, 2005.

TURBAN, E.; SHARDA, R.; ARONSON, J.; KING, D. **Business Intelligence**: um enfoque gerencial para a inteligência do negócio. Porto Alegre: Bookman, 2009.

VALLE, M.; VARAS, S.; RUZ, G. Job performance prediction in a call center using a naive Bayes classifier. **Expert Systems with Applications**, v. 39, n. 11, p. 9939–9945, 2012.

VARELLA, L. DINSMORE, P. C.; CAVALIERI, A. **Gerenciamento de Riscos em gerenciamento de projetos**. 2° ed. Rio de Janeiro: Qualitymark, 2005. Cap. 9 p. 191-214.

VAUGHAN, E.J. **Risk management.** New Baskerville: John Wiley & Sons, 1997.

WEILL, P.; ROSS, J. **Governança de Tecnologia da Informação.** São Paulo: M. Books do Brasil, 2006.

WILLIAMS, J. Regulatory technologies, risky subjects, and financial boundaries: Governing “fraud” in the financial markets. **Accounting, Organizations and Society**, v. 38, n. 7, p. 544-588, 2012.

WITTEN, I; FRANK, E. **Data mining**: practical machine learning tools and techniques. New Zealand: Morgan Kaufmann Publishers, 2005.

WOON, L. F.; AZIZAN, A. N.; SAMAD, A. F. M. A Strategic framework for value enhancing enterprise risk management. **Journal od Global Business and Economics,** v. 2. n. 1, 2011.

ZHAO, Y.; ZHANG, Y. Comparison of decision tree methods for finding active objects. **Advances in Space Research**, v. 41, n. 12, p. 1955–1959, jan. 2008.

ZHONG, X. The application of Apriori algorithm for network forensics analysis. **Journal of Theoretical and Applied Information Technology**, v. 50, n. 2, p. 430–435, 2013.

ZONATTO, V. D. S.; BEUREN, I. Categorias de riscos evidenciadas nos relatórios da administração de empresas brasileiras com ADRs. **Revista Brasileira De Gestão De Negócios**, v. 12, n. 35, p.141-155, 2010.

APÊNDICE A - COMPARATIVO DE CLASSIFICAÇÃO DO ALGORITMO EM E KMEANS

EM		Kmeans			Equivalência
Empresa	Cluster	Empresa	Cluster	Cluster Equivalente	
0	0	0	3	0	Equivalente
1	0	1	3	0	Equivalente
2	2	2	0	2	Equivalente
3	2	3	0	2	Equivalente
4	0	4	3	0	Equivalente
5	0	5	3	0	Equivalente
6	3	6	2	3	Equivalente
7	1	7	1	1	Equivalente
8	3	8	2	3	Equivalente
9	0	9	3	0	Equivalente
10	2	10	0	2	Equivalente
11	2	11	0	2	Equivalente
12	0	12	3	0	Equivalente
13	2	13	0	2	Equivalente
14	0	14	3	0	Equivalente
15	0	15	3	0	Equivalente
16	2	16	0	2	Equivalente
17	3	17	2	3	Equivalente
18	0	18	3	0	Equivalente
19	1	19	1	1	Equivalente
20	0	20	3	0	Equivalente
21	0	21	3	0	Equivalente
22	0	22	1	1	Diferente
23	0	23	3	0	Equivalente
24	1	24	1	1	Equivalente
25	2	25	0	2	Equivalente
26	3	26	2	3	Equivalente
27	0	27	3	0	Equivalente
28	0	28	3	0	Equivalente
29	0	29	3	0	Equivalente
30	0	30	3	0	Equivalente
31	1	31	1	1	Equivalente
32	2	32	0	2	Equivalente
33	1	33	1	1	Equivalente
34	1	34	1	1	Equivalente
35	1	35	1	1	Equivalente
36	1	36	1	1	Equivalente
37	1	37	1	1	Equivalente
38	1	38	1	1	Equivalente
39	2	39	0	2	Equivalente
40	0	40	3	0	Equivalente
41	1	41	1	1	Equivalente
42	0	42	3	0	Equivalente
43	1	43	1	1	Equivalente
44	1	44	1	1	Equivalente
45	1	45	1	1	Equivalente
46	0	46	3	0	Equivalente
47	2	47	0	2	Equivalente
48	0	48	3	0	Equivalente
49	0	49	3	0	Equivalente
50	3	50	2	3	Equivalente

EM		Kmeans			Equivalência
Empresa	Cluster	Empresa	Cluster	Cluster Equivalente	
51	3	51	2	3	Equivalente
52	1	52	1	1	Equivalente
53	1	53	1	1	Equivalente
54	3	54	2	3	Equivalente
55	0	55	1	1	Diferente
56	1	56	1	1	Equivalente
57	0	57	3	0	Equivalente
58	3	58	2	3	Equivalente
59	1	59	1	1	Equivalente
60	0	60	3	0	Equivalente
61	0	61	3	0	Equivalente
62	0	62	3	0	Equivalente
63	2	63	0	2	Equivalente
64	2	64	0	2	Equivalente
65	0	65	1	1	Diferente
66	3	66	2	3	Equivalente
67	2	67	0	2	Equivalente
68	3	68	2	3	Equivalente
69	3	69	2	3	Equivalente
70	1	70	1	1	Equivalente
71	2	71	0	2	Equivalente
72	1	72	1	1	Equivalente
73	0	73	3	0	Equivalente
74	1	74	1	1	Equivalente
75	1	75	1	1	Equivalente
76	3	76	2	3	Equivalente
77	0	77	1	1	Diferente
78	1	78	1	1	Equivalente
79	3	79	2	3	Equivalente
80	1	80	1	1	Equivalente
81	3	81	2	3	Equivalente
82	1	82	1	1	Equivalente
83	0	83	3	0	Equivalente
84	0	84	3	0	Equivalente
85	0	85	3	0	Equivalente
86	2	86	0	2	Equivalente
87	0	87	3	0	Equivalente
88	2	88	0	2	Equivalente
89	0	89	1	1	Diferente
90	2	90	0	2	Equivalente
91	3	91	2	3	Equivalente
92	1	92	1	1	Equivalente
93	1	93	1	1	Equivalente
94	0	94	1	1	Diferente
95	1	95	1	1	Equivalente
96	3	96	2	3	Equivalente
97	3	97	2	3	Equivalente
98	0	98	3	0	Equivalente
99	2	99	0	2	Equivalente
100	0	100	3	0	Equivalente

EM			Kmeans			
Empresa	Cluster		Empresa	Cluster	Cluster Equivalente	**Equivalência**
101	2		101	0	2	Equivalente
102	0		102	3	0	Equivalente
103	1		103	1	1	Equivalente
104	3		104	2	3	Equivalente
105	1		105	1	1	Equivalente
106	0		106	3	0	Equivalente
107	1		107	0	2	Diferente
108	2		108	0	2	Equivalente
109	2		109	0	2	Equivalente
110	3		110	2	3	Equivalente
111	1		111	1	1	Equivalente
112	1		112	1	1	Equivalente
113	2		113	0	2	Equivalente
114	1		114	1	1	Equivalente
115	1		115	1	1	Equivalente
116	3		116	2	3	Equivalente
117	0		117	2	3	Diferente
118	2		118	0	2	Equivalente
119	1		119	1	1	Equivalente
120	1		120	1	1	Equivalente
121	2		121	0	2	Equivalente
122	1		122	1	1	Equivalente
123	0		123	3	0	Equivalente
124	2		124	0	2	Equivalente
125	0		125	1	1	Diferente
126	1		126	1	1	Equivalente
127	2		127	0	2	Equivalente
128	2		128	0	2	Equivalente
129	0		129	3	0	Equivalente
130	3		130	2	3	Equivalente
131	3		131	2	3	Equivalente
132	2		132	0	2	Equivalente
133	3		133	2	3	Equivalente
134	2		134	0	2	Equivalente
135	3		135	2	3	Equivalente
136	0		136	3	0	Equivalente
137	1		137	1	1	Equivalente
138	1		138	1	1	Equivalente
139	1		139	1	1	Equivalente

APÊNDICE B – COMPARATIVO ENTRE A CLASSIFICAÇÃO DOS AGRUPAMENTOS REDUZIDOS E TOTAIS

Kmeans Red.		Equivalentes ao EM Red.		Comparação	
Empresa	Cluster	EM Total	Kmeans Total	Kmeans Red. X EM Total	Kmeans Red. X Kmeans Total
0	3	3	3	Equivalente	Equivalente
1	0	3	3	Diferente	Diferente
2	1	1	1	Equivalente	Equivalente
3	1	1	1	Equivalente	Equivalente
4	3	3	3	Equivalente	Equivalente
5	0	3	3	Diferente	Diferente
6	3	0	0	Diferente	Diferente
7	2	2	2	Equivalente	Equivalente
8	0	0	0	Equivalente	Equivalente
9	2	3	3	Diferente	Diferente
10	1	1	1	Equivalente	Equivalente
11	1	1	1	Equivalente	Equivalente
12	3	3	3	Equivalente	Equivalente
13	1	1	1	Equivalente	Equivalente
14	0	3	3	Diferente	Diferente
15	3	3	3	Equivalente	Equivalente
16	1	1	1	Equivalente	Equivalente
17	0	0	0	Equivalente	Equivalente
18	2	3	3	Diferente	Diferente
19	3	2	2	Diferente	Diferente
20	3	3	3	Equivalente	Equivalente
21	0	3	3	Diferente	Diferente
22	2	3	2	Diferente	Equivalente
23	3	3	3	Equivalente	Equivalente
24	2	2	2	Equivalente	Equivalente
25	1	1	1	Equivalente	Equivalente
26	0	0	0	Equivalente	Equivalente
27	3	3	3	Equivalente	Equivalente
28	0	3	3	Diferente	Diferente
29	3	3	3	Equivalente	Equivalente
30	2	3	3	Diferente	Diferente
31	2	2	2	Equivalente	Equivalente
32	1	1	1	Equivalente	Equivalente
33	2	2	2	Equivalente	Equivalente
34	1	2	2	Diferente	Diferente
35	2	2	2	Equivalente	Equivalente
36	2	2	2	Equivalente	Equivalente
37	2	2	2	Equivalente	Equivalente
38	2	2	2	Equivalente	Equivalente
39	1	1	1	Equivalente	Equivalente
40	3	3	3	Equivalente	Equivalente
41	2	2	2	Equivalente	Equivalente
42	3	3	3	Equivalente	Equivalente
43	1	2	2	Diferente	Diferente
44	1	2	2	Diferente	Diferente
45	2	2	2	Equivalente	Equivalente
46	3	3	3	Equivalente	Equivalente
47	1	1	1	Equivalente	Equivalente
48	0	3	3	Diferente	Diferente
49	3	3	3	Equivalente	Equivalente
50	0	0	0	Equivalente	Equivalente

Continuação

Kmeans Red.		Equivalentes ao EM Red.		Comparação	
Empresa	Cluster	EM Total	Kmeans Total	Kmeans Red. X EM Total	Kmeans Red. X Kmeans Total
51	0	0	0	Equivalente	Equivalente
52	1	2	2	Diferente	Diferente
53	2	2	2	Equivalente	Equivalente
54	3	0	0	Diferente	Diferente
55	2	3	2	Diferente	Equivalente
56	1	2	2	Diferente	Diferente
57	3	3	3	Equivalente	Equivalente
58	0	0	0	Equivalente	Equivalente
59	2	2	2	Equivalente	Equivalente
60	0	3	3	Diferente	Diferente
61	0	3	3	Diferente	Diferente
62	0	3	3	Diferente	Diferente
63	1	1	1	Equivalente	Equivalente
64	1	1	1	Equivalente	Equivalente
65	2	3	2	Diferente	Equivalente
66	0	0	0	Equivalente	Equivalente
67	1	1	1	Equivalente	Equivalente
68	0	0	0	Equivalente	Equivalente
69	0	0	0	Equivalente	Equivalente
70	2	2	2	Equivalente	Equivalente
71	1	1	1	Equivalente	Equivalente
72	2	2	2	Equivalente	Equivalente
73	3	3	3	Equivalente	Equivalente
74	2	2	2	Equivalente	Equivalente
75	2	2	2	Equivalente	Equivalente
76	0	0	0	Equivalente	Equivalente
77	2	3	2	Diferente	Equivalente
78	2	2	2	Equivalente	Equivalente
79	0	0	0	Equivalente	Equivalente
80	2	2	2	Equivalente	Equivalente
81	0	0	0	Equivalente	Equivalente
82	2	2	2	Equivalente	Equivalente
83	3	3	3	Equivalente	Equivalente
84	3	3	3	Equivalente	Equivalente
85	3	3	3	Equivalente	Equivalente
86	1	1	1	Equivalente	Equivalente
87	3	3	3	Equivalente	Equivalente
88	1	1	1	Equivalente	Equivalente
89	1	3	2	Diferente	Diferente
90	1	1	1	Equivalente	Equivalente
91	0	0	0	Equivalente	Equivalente
92	3	2	2	Diferente	Diferente
93	1	2	2	Diferente	Diferente
94	3	3	2	Equivalente	Diferente
95	2	2	2	Equivalente	Equivalente
96	0	0	0	Equivalente	Equivalente
97	0	0	0	Equivalente	Equivalente
98	3	3	3	Equivalente	Equivalente
99	1	1	1	Equivalente	Equivalente
100	0	3	3	Diferente	Diferente

Continuação

Kmeans Red.		Equivalentes ao EM Red.		Comparação	
Empresa	Cluster	EM Total	Kmeans Total	Kmeans Red. X EM Total	Kmeans Red. X Kmeans Total
101	1	1	1	Equivalente	Equivalente
102	3	3	3	Equivalente	Equivalente
103	2	2	2	Equivalente	Equivalente
104	0	0	0	Equivalente	Equivalente
105	2	2	2	Equivalente	Equivalente
106	0	3	3	Diferente	Diferente
107	1	2	1	Diferente	Equivalente
108	1	1	1	Equivalente	Equivalente
109	1	1	1	Equivalente	Equivalente
110	0	0	0	Equivalente	Equivalente
111	1	2	2	Diferente	Diferente
112	2	2	2	Equivalente	Equivalente
113	1	1	1	Equivalente	Equivalente
114	2	2	2	Equivalente	Equivalente
115	2	2	2	Equivalente	Equivalente
116	0	0	0	Equivalente	Equivalente
117	3	3	0	Equivalente	Diferente
118	1	1	1	Equivalente	Equivalente
119	3	2	2	Diferente	Diferente
120	2	2	2	Equivalente	Equivalente
121	1	1	1	Equivalente	Equivalente
122	2	2	2	Equivalente	Equivalente
123	3	3	3	Equivalente	Equivalente
124	1	1	1	Equivalente	Equivalente
125	2	3	2	Diferente	Equivalente
126	2	2	2	Equivalente	Equivalente
127	1	1	1	Equivalente	Equivalente
128	1	1	1	Equivalente	Equivalente
129	0	3	3	Diferente	Diferente
130	0	0	0	Equivalente	Equivalente
131	0	0	0	Equivalente	Equivalente
132	1	1	1	Equivalente	Equivalente
133	0	0	0	Equivalente	Equivalente
134	1	1	1	Equivalente	Equivalente
135	0	0	0	Equivalente	Equivalente
136	3	3	3	Equivalente	Equivalente
137	2	2	2	Equivalente	Equivalente
138	2	2	2	Equivalente	Equivalente
139	2	2	2	Equivalente	Equivalente