Universidade Federal de Minas Gerais
Escola de Engenharia
Programa de Pós-Graduação em Engenharia Elétrica

Aplicação do Processo de KDD a um Ambiente Industrial

Lucas Costa Oliveira Santos

Dissertação submetida à banca examinadora designada pelo Colegiado do Programa de Pós-Graduação em Engenharia Elétrica da Universidade Federal de Minas Gerais, como parte dos requisitos necessários à obtenção do grau de Mestre em Engenharia Elétrica.

Orientador: Luiz Themystokliz Sanctos Mendes

Belo Horizonte, junho de 2007

“A excelência não é um feito, mas a seguida repetição de bons hábitos.”
(Aristóteles 384-322 a.C.)

“Não há investimento mais rentável do que o do conhecimento.”
(Benjamin Franklin 1706-1790)

Agradecimentos

Agradeço a todos aqueles que de uma forma ou de outra me ajudaram a passar por essa etapa.

A minha amada e companheira Vitória Régia, pela dedicação, amor e tolerância.

A minha mãe, por ser para mim mais do que um exemplo na vida pessoal, profissional e acadêmica.

Aos meus pais Artur e Otto, que junto com o Pedro, Débora, Luiz Felipe, Ana Luisa e Angela formam a minha família.

Ao Dr. Milton, por me ajudar a ser uma pessoa menos ansiosa.

A ATAN, que na pessoa do Engenheiro Marcelo Szuster me ajudou a viabilizar esse projeto.

A UFMG, pelo fato de a quase 10 anos ser a minha casa acadêmica.

E ao professor Luiz Themystokliz que mais do que um orientador, foi o grande motivador dessa empreitada.

Resumo

O processo de extração de conhecimento de grandes volumes de dados é complexo e pode ter seu custo muito elevado, dependendo das características do problema e do que se quer obter. A quantidade de dados gerada pelos sistemas das organizações atuais supera a capacidade humana de interpretar manualmente e compreender tanta informação. Para abordar este problema surgiu, dentro da Computação, a área de pesquisa conhecida como "Extração de Conhecimento em Bases de Dados" (*Knowledge Discovery in Databases*, ou KDD).

No presente trabalho, motivado pela ainda incipiente exploração das técnicas de KDD em ambientes industriais, apresenta-se um processo completo de aplicação dessa metodologia, utilizando dados reais do processo de Laminação de Tiras a Quente de uma grande Usina Siderúrgica do cenário nacional. Além da apresentação do processo de KDD, com a definição de cada uma das suas etapas, é feita uma revisão do estado-da-arte da aplicação dessa metodologia e das técnicas de *Data Mining* na Siderurgia e, mais especificamente, na área de Laminação a Quente.

Dentre um conjunto de problemas a serem tratados, foi definido como objetivo do trabalho a identificação de variáveis que, de alguma forma, estivessem relacionadas com o "Erro de Força" das cadeiras do processo de Laminação a Quente. O algoritmo CART é empregado como principal ferramenta de *Data Mining* e sua utilização resultou em descobertas válidas e potencialmente úteis para a Usina Siderúrgica, como a correlação entre a atuação do operador da planta e aumento do "Erro de Força de Laminação" e também a influência da "Força de Flexão dos Cilindros" no "Erro de Força". Além da análise dos resultados obtidos, são apresentadas também as dificuldades encontradas, bem como as perspectivas futuras sobre o assunto dentro do ambiente industrial.

Abstract

Knowledge extraction from large databases is a complex process which can imply in very high costs, depending on the problem and on what one wants to get. Nowadays, the amount of data stored in many organizations systems goes far beyond human ability to manually interpret and understand that information. In order to deal with this problem, the research area known as *Knowledge Discovery in Databases*, or KDD, has been created in the Computer Science field.

This project, which was motivated by the short exploration of KDD in the Process Industry environment, shows a complete application of this methodology with real data of a Hot Rolling Mill plant in a large Brazilian Steel Industry. Beyond the KDD process presentation, with the definition of every step, this work also reviews the state-of-the-art of this methodology application and of the Data Mining techniques in the Steel Industry, and more specifically in the Hot Rolling Mill.

From a group of potential problems, the project main target was defined as the identification of variables that could be somehow related to the Hot Rolling Mill process "Force Error". The CART algorithm comes as the main tool for Data Mining, and its usage resulted in valid and potentially useful discoveries to that Steel Industry, as the correlation between the plant operator actions and the increase of the "Roll Force Error", as well as the influence of the "Bending Force". Besides the project results analysis, the difficulties found and the near future perspectives of this subject in the Process Industry are presented.

Sumário

Lista de Figuras

Lista de Tabelas

Lista de Siglas

CART - *Classification And Regression Tree*

CLP - *Controller Logic Programmable*

DB2 - *Database* 2, família de bancos de dados relacionais da IBM

ERP - *Enterprise Resource Planning*

HD - *Hard Disk*

iDA - *intelligent Data Analyzer*

IEEE - *Institute of Electrical and Electronic Engineers*

KDD - *Knowledge Discovery in Databases*

LIMS - *Laboratory Information Management System*

MES - *Manufacturing Execution System*

MIT - *Massachusetts Institute of Technology*

OLAP - *On-Line Analytical Processing*

PCA - *Principal Component Analysis*

PIMS - *Plant Information Management System*

PROCOM - *Process Computer*

RNA - Redes Neurais Artificiais

SCADA - *Supervisory Control And Data Acquisition*

SGDB - Sistema Gerenciador de Banco de Dados

SOM - *Self Organizing Maps*

SQL - *Structured Query Language*

Capítulo 1

Introdução

1.1 Contexto

Nos últimos anos, tem-se observado um aumento explosivo do número de informações armazenadas em Bancos de Dados, e também o crescimento exponencial do tamanho destes Bancos. Isto se tem dado principalmente pelos progressos nas tecnologias de aquisição e armazenamento digital de informações. Mais recentemente, tem crescido o interesse por uma melhor análise destas massas de dados, na tentativa de se extrair informações ocultas, não-triviais e previamente desconhecidas que possam ter algum valor estratégico para os responsáveis pelas mesmas.

O processo de extração de conhecimento de grandes volumes de dados é complexo e pode ter seu custo muito elevado, dependendo das características do problema e do que se quer obter. A quantidade de dados gerada pelos sistemas de informação das organizações atuais supera a capacidade humana de interpretar e compreender tanta informação. Para abordar este problema surgiu, dentro da Computação, a área de pesquisa conhecida como "Extração de Conhecimento em Bases de Dados" (*Knowledge Discovery in Databases*, ou KDD) (Fayyad *et al.*, 1996). KDD é uma tecnologia essencialmente multi-disciplinar envolvendo principalmente as áreas de Bancos de Dados, Inteligência Artificial (como por ex. Redes Neurais e Lógica *Fuzzy*) e Estatística, entre outras.

Dentro desta área, existe uma atividade conhecida como *Data Mining* que corresponde à aplicação de técnicas computacionais para se extrair informações e conhecimento, por meios automáticos ou semi-automáticos, de uma grande massa de dados. É importante ressaltar que não são consideradas como *Data Mining* a execução de consultas diretamente a bancos de dados, ou ainda a simples aplicação de técnicas estatísticas para

se obter informações sobre a distribuição dos dados, por exemplo. Nesses casos, já se sabe *a priori* o que se está buscando dentro da base de dados. Uma definição interessante sobre *Data Mining* é a apresentada por Fayyad *et al.* (1996), a qual reproduzimos abaixo:

> *"Processo não-trivial de identificação de dados que são válidos, novos, potencialmente úteis e com padrões reconhecíveis".*

Data Mining, e de forma mais geral KDD, oferecem uma alternativa promissora de ajudar as diversas áreas do conhecimento a encontrarem informações relevantes e ocultas em bases de dados (Two Crows Corporation, 1999). Uma ressalva importante que deve ser feita é o fato de que as ferramentas de *Data Mining* não fazem nenhuma "mágica", encontrando resultados de forma automática. É preciso que este processo seja conduzido com a participação de alguém que conheça bem os dados com os quais está trabalhando. Encontrar bons modelos é apenas uma das fases entre todas que necessitam ser executadas. Se os dados não forem bem trabalhados e os objetivos claramente identificados, as informações encontradas podem não ter relevância alguma.

1.2 Motivação e Justificativa

A principal motivação deste trabalho decorre do fato de que a tecnologia de *Data Mining*, uma das mais promissoras desta década segundo o MIT (Malone, 2005), ainda é praticamente inexplorada em Sistemas de Automação Industrial. Em pesquisa realizada em novembro de 2006 nos periódicos do IEEE, pôde-se verificar que cerca de apenas 1% dos artigos sobre *Data Mining* está relacionado a aplicações na indústria de processos. Nesta mesma pesquisa, encontrou-se um número muito maior de registros em áreas como Finanças, Medicina, Científica e *Marketing*. Alguns exemplos de aplicações nestas áreas são:

- Aplicação do processo de KDD pelas empresas de cartões de crédito na análise de Bancos de Dados de clientes para identificar seus diferentes grupos e predizer seu comportamento, de forma a direcionar as atividades de *marketing* (exemplo: mala direta);

- Análise dos dados sobre o que os consumidores compram para obter conhecimento sobre quais produtos são comprados juntos e quais são bons candidatos para promoções - *Marketing*;

- Na área de finanças, a identificação de padrões nos hábitos dos correntistas bancários é utilizada como suporte para concessão de crédito. A utilização de técnicas de *Data Mining* também está presente na detecção de fraudes financeiras e evasão fiscal;

- Na Medicina, bancos de dados dos sistemas de saúde têm sido analisados para se buscar padrões entre os resultados de diversos exames realizados pelos pacientes e doenças de elevado impacto sócio-econômico (ex.: Diabetes). Na genética, *Data Mining* é utilizado na análise de genomas quanto à presença ou ausência de seqüências-chaves em regiões específicas dos mesmos, além da detecção de correlações entre estrutura e função de genes.

Já na indústria de processos uma importante tendência, atualmente, é a utilização de sistemas de gestão de informações como PIMS (*Plant Information Management System*), MES (*Manufacturing Execution System*), LIMS (*Laboratory Information Management System*) e outros como ferramentas estratégicas de suporte a decisões. Tais aplicações estão se tornando cada vez mais freqüentes, especialmente nas indústrias de médio a grande porte, e tipicamente operam vinculadas a Bancos de Dados históricos e/ou relacionais que armazenam um grande volume de dados de processo (PIMS) ou de produção (MES), organizados de forma sistemática ao longo do tempo. Os Bancos de Dados destas aplicações são, assim, um excelente laboratório para a aplicação e validação de técnicas de *Data Mining* em aplicações da indústria de processos. Exemplos de aplicações nessa área são apresentados na seção 2.3.

A idéia de aplicar a metodologia de KDD em bases de dados industriais do tipo PIMS, MES, LIMS, etc., é bastante interessante pela potencialidade de se encontrar informações relevantes sobre os processos industriais que geram estes dados. O banco de dados de um sistema PIMS, por exemplo, coleta informações temporais sobre uma ampla gama de variáveis de processo. Espera-se que técnicas de *Data Mining* sejam capazes de identificar padrões novos, com informações úteis sobre esse processo para os técnicos responsáveis pelo mesmo.

1.3 Objetivos do trabalho

Dentre os objetivos deste trabalho, podem-se citar como os mais relevantes os seguintes:

- Explorar a viabilidade da aplicação das técnicas de KDD em bancos de dados industriais;
- Identificar objetivos específicos a serem alcançados junto a estas Bases de Dados. Esses objetivos serão apresentados mais a frente na seção 3.2;
- Aplicar o processo de KDD nessas Bases de Dados, tendo em vista o alcance destes objetivos específicos;
- Atrair o interesse da indústria nacional para esta tecnologia, a partir de um caso concreto de aplicação da mesma.

Além das citadas, uma meta importante é a de conseguir documentar de forma simples e objetiva este processo de extração de conhecimento em Bases de Dados Industriais, de modo que este trabalho possa, de alguma forma, contribuir com iniciativas futuras nesta área.

1.4 Ambiente de elaboração do trabalho

Para viabilizar a realização deste projeto, o primeiro desafio foi a busca por uma base de dados de um processo industrial. Através de contatos com uma grande Indústria Siderúrgica Nacional foi iniciado um projeto de pesquisa. Por questões de sigilo industrial, o nome da usina em questão não será revelado. O objetivo desse projeto é extrair conhecimento dos dados de processo para auxiliar a resolução de problemas associados ao ambiente de produção da Siderúrgica.

Dentro do processo produtivo desta Usina, a área escolhida para aplicação de KDD foi a da Laminação de Tiras a Quente. Esta área possui um elevado grau de automação e uma grande taxa de aquisição de dados referentes ao processo. Dentre os vários sistemas disponíveis, optou-se pela utilização de um sistema PIMS para aquisição de dados históricos.

Integrantes da área de Automação da Siderúrgica acompanharam o trabalho desde o início, tendo participação essencial na definição da base de dados, seleção de objetivos e análise dos resultados obtidos. Várias reuniões foram necessárias para realização deste trabalho e as mesmas ocorreram na própria Usina. Além do suporte técnico fornecido, a realização do projeto só foi possível devido à autorização gerencial da Siderúrgica para liberação dos dados para a UFMG, para fins exclusivos deste trabalho.

1.5 Estrutura da dissertação

Este capítulo introdutório apresenta o contexto no qual este trabalho está inserido, bem como a justificativa para realização do mesmo. Os objetivos a serem alcançados são discriminados juntamente com o ambiente de elaboração do trabalho. O restante da dissertação está estruturado em mais 5 capítulos.

O Capítulo 2 apresenta uma revisão bibliográfica sobre o processo de KDD como um todo e o estado-da-arte desta técnica aplicada à indústria de processos, mais especificamente à indústria siderúrgica. O Capítulo 3 faz uma breve introdução ao processo de Laminação de Tiras a Quente, apresenta o levantamento de problemas que poderiam ser tratados neste trabalho e as características da base de dados utilizada. Em seguida, o Capítulo 4 descreve de forma objetiva o detalhamento técnico das etapas do processo de KDD realizadas, incluindo os resultados obtidos. O Capitulo 5 faz uma análise dos resultados, e o Capítulo 6 finaliza a dissertação com a conclusão do trabalho e apresentação de perspectivas futuras.

Capítulo 2

Revisão Bibliográfica

A evolução dos sistemas de computação ocorrida nos últimos anos proporcionou um aumento expressivo na capacidade de armazenamento de dados. O que começou na década de 60 com os *Mainframes*, passou pelos mini-computadores dos anos 70, pelos PCs na década de 80 e evoluiu até a arquitetura distribuída de cliente-servidor dos anos 90 serviu de base para o modelo difuso da internet que é vivenciado atualmente (Ma, 1998). Com essa evolução, que ocorreu tanto no *hardware* quanto no *software*, a facilidade atual que uma aplicação, seja científica ou comercial, possui para gerar *gigabytes* ou *terabytes* de dados em poucas horas excede em muito a capacidade do ser humano de analisar os mesmos de forma manual.

O objetivo de se manter um banco de dados é, na maioria dos casos, realizar consultas. Estas são feitas periodicamente pelos analistas responsáveis pelos dados, com o intuito de se recuperar informações. No passado, essas consultas eram relativamente simples, e na maioria das vezes podiam ser obtidas através de comandos diretos sobre o banco de dados. Com o passar do tempo, a necessidade de se correlacionar informações e armazená-las de forma mais estruturada levou a uma evolução desses bancos de dados, que passaram a ser mais do que simples repositórios. Ademais, a qualidade e quantidade de informações requeridas também segue uma curva crescente.

É interessante notar que a explosão na capacidade de armazenamento de dados está ocorrendo nas mais diversas áreas do conhecimento, desde situações do cotidiano (como transações em supermercados, registros de utilização de cartões de crédito, detalhes de ligações telefônicas e estatísticas governamentais) até questões mais exóticas e específicas (como imagens de astros captadas por novos telescópios, bases de dados de códigos genéticos e registros médicos em geral) (Hand *et al.*, 2001). A atração do homem por

uma busca cada vez maior de conhecimento faz com que, em todas essas áreas, pesquisas sejam realizadas tendo como ponto de partida os dados armazenados. O problema atual é o fato de ser praticamente impossível realizar uma análise manual dessas enormes bases de dados.

Na área industrial, as aplicações PIMS vêm se consolidando como uma das soluções de aquisição de dados. Esse sistema nada mais é do que um repositório no qual são concentradas todas as informações relevantes das células de produção, diretamente ligadas aos sistemas de supervisão e controle. O PIMS coleta informações dos sistemas de chão-de-fábrica e as armazena em uma base de dados histórica de tempo real. Tal base tem características não encontradas nos bancos de dados convencionais, como grande capacidade de compactação (tipicamente de 10:1) e alta velocidade de resposta a consultas. Devido a isto, é capaz de armazenar um grande volume de dados com recursos mínimos, se comparado às soluções convencionais (Aspen PIMS System, 2006). Apesar de possuir mecanismos de consulta e de possibilitar o desenvolvimento de telas de acompanhamento e relatórios, é consenso entre especialistas da área que as bases de dados dos sistemas PIMS são sub-utilizadas. Em grande parte, isso ocorre devido ao enorme tamanho e características temporais desses repositórios.

Problema semelhante ocorre na medicina. Imagine-se um médico que tem acesso a todos os exames que seus pacientes de meia-idade já fizeram na vida. Uma análise dos resultados de todos esses exames pode trazer informações relevantes e indicar, por exemplo, uma tendência de uma pessoa para certo tipo de doença. O problema é que se existirem registros de um número muito elevado de exames já realizados, o trabalho torna-se inviável de ser feito manualmente. Hoje nos Estados Unidos e em alguns países da Europa, já existem bancos de dados de planos de saúde que registram todos os resultados de exames, consultas e procedimentos que cada paciente realiza desde o nascimento (Breault *et al.*, 2002). A questão é: como extrair conhecimento dessa imensa massa de dados?

O conceito de *Data Warehouse* (Gardner, 1998) é considerado como um dos primeiros passos para se tornar viável a análise de grandes massas de dados no apoio à tomada de decisão. *Data Warehouse* é um repositório central de informações construído para busca e análise eficientes de dados. As grandes corporações de todo o mundo nas mais diversas áreas estão investindo cada vez mais nesse tipo de tecnologia, também conhecida

como BI (*Bussiness Inteligence*). A tecnologia de OLAP (*On-Line Analytical Processing*) (Ma, 1998) surge como principal ferramenta de análise dos dados de repositórios centrais. Essa ferramenta permite que consultas relativamente complexas possam ser realizadas e pré-processadas. O usuário final possui, então, um modo flexível e ágil para realizar buscas dentro das bases de dados.

2.1 Processo de KDD

A grande questão com relação a OLAP e outras ferramentas de análise de grandes bases de dados, está no fato de que as mesmas são orientadas a consultas, ou seja, são dirigidas pelos usuários, os quais possuem hipóteses que gostariam de comprovar ou que simplesmente executam consultas aleatórias (Oliveira, 2000). Essa abordagem pode limitar em muito a descoberta de informações sobre os dados, principalmente tratando-se de massas de dados enormes. A possibilidade de existirem padrões ou características relevantes acerca das informações armazenadas cresce junto com o aumento do armazenamento e da qualidade dos dados. A área de KDD surge, então, para tentar suprir essa lacuna deixada pela técnicas tradicionais de exploração de bancos de dados.

Vale ressaltar que muitos autores utilizam os termos KDD ou *Data Mining* como sinônimos. Apesar de não existir um consenso sobre a questão, a abordagem utilizada neste trabalho segue a mais difundida na área acadêmica, segundo a qual *Data Mining* é apenas uma fase do processo de KDD como um todo. A diferenciação será observada de forma mais clara no detalhamento de cada etapa do processo de KDD (ver seção 2.2).

Apesar de ser uma área de conhecimento relativamente recente, existe hoje na literatura um grande número de publicações sobre KDD e a aplicação de técnicas de *Data Mining*. O artigo publicado por Fayyad *et al.* (1996), é talvez um dos principais marcos da unificação de várias tecnologias para a criação de um modelo para "Descoberta de Conhecimento em Banco de Dados". Tal artigo, além de definir de forma bem sólida o conceito de KDD, já traz a definição do termo *Data Mining* como o processo de aplicação de técnicas computacionais para se encontrar informações novas e úteis dentro de um repositório de dados, sendo muito citado em textos sobre o assunto até os dias de hoje.

O estudo sobre o estado-da-arte das técnicas de aprendizado por computador e busca de conhecimento em bases de dados realizado por Thrun *et al.* (1998) mostra a grande potencialidade da aplicação das técnicas de *Data Mining* e traz também uma re-

visão dos principais resultados obtidos em várias áreas do conhecimento até então. No ano seguinte, uma publicação bem mais completa foi realizada pela Two Crows Corporation (1999). Este texto, além de introduzir bem o assunto e apresentar todo o processo de forma mais detalhada, traz uma seção sobre critérios para seleção de ferramentas para se trabalhar com *Data Mining*.

Hoje, a aplicação das técnicas de KDD já apresenta ganhos em diversas áreas de atividade. A grande maioria das empresas de cartão de crédito, por exemplo, já utiliza técnicas de *Data Mining* em suas enormes bases de dados para monitorar o padrão de compras de seus clientes e avisá-los ou até mesmo cancelar automaticamente o cartão, caso uma compra seja detectada fora do perfil do usuário. Além disso, as mesmas bases de dados são exploradas para direcionar melhor campanhas publicitárias (*Marketing* direto). Vários bancos e outras instituições financeiras também têm aplicado KDD para identificar padrões de fraudes, características relevantes sobre os correntistas e realizar previsões do comportamento dos mercados.

As redes de supermercados oferecem um exemplo clássico de sucesso na aplicação de KDD: a otimização de disposição de produtos nas gôndolas de vendas obtidas pela Wal-Mart, cadeia de supermercados norte-americana, através da análise das transações de compra dos seus clientes. Padrões encontrados foram utilizados para impulsionar as vendas (ex: foi observado que homens casados, na faixa etária de 25 a 30 anos sempre compravam cerveja e fraldas - colocando os dois produtos juntos, a rede conseguiu um aumento de 30% das vendas) (Berry and Linoff, 1997). Outro exemplo de aplicação nessa área ocorreu no Brasil, onde as lojas Brasileiras aplicaram cerca de 1 milhão de dólares em técnicas de *Data Mining* (Dias, 2002). O resultado obtido foi a redução de 51 mil para 14 mil do número de produtos oferecidos em suas lojas. Alguns exemplos de anomalias detectadas foram:

- Roupas de inverno e guarda chuvas encalhadas no nordeste;
- Batedeiras 110v a venda em SC onde a tensão elétrica é padronizada em 220v.

2.2 Definição das Etapas do Processo de KDD

O processo de KDD corresponde à execução de uma série de etapas que podem levar à descoberta de informações relevantes sobre as mais distintas massas de dados.

A definição exata da quantidade de etapas e do conteúdo específico de cada uma varia de autor para autor. Fayyad *et al.* (1996) descrevem um processo baseado em 5 fases, conforme apresentado na Figura 2.1.

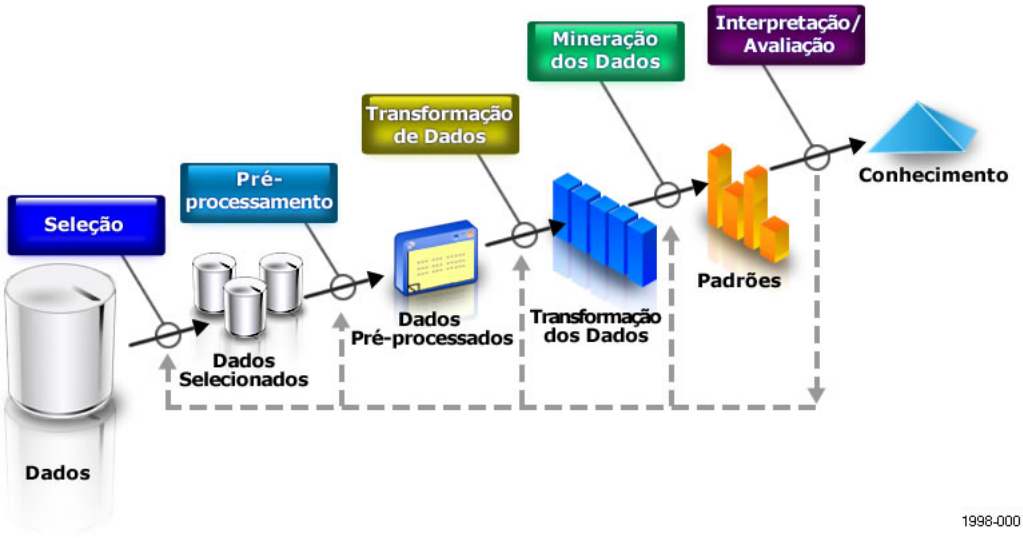

Figura 2.1: Etapas do processo de KDD, figura adaptada de Fayyad *et al.* (1996)

Um outro processo, mais detalhado, é proposto por Berry and Linoff (1997) e composto por 10 passos altamente correlacionados. A Figura 2.2 apresenta este processo, destacando as correlações entre o resultado de um passo e o início de outro. O resultado de cada etapa pode gerar a necessidade de revisão dos passos anteriores.

Nesse processo, grande ênfase é dada à etapa "*Get to know the data*", ou Conhecimento dos dados. Outra diferença em relação à proposta de Fayyad *et al.* (1996) é a etapa inicial, na qual fica explícita a necessidade de se traduzir os problemas de negócio em tarefas de *Data Mining*.

Mais recentemente, o modelo chamado de **CRISP-DM** - *Cross Industry Standard Process for Data Mining* (Shearer, 2000) - é apresentado como um processo mais refinado, destinado a grandes empresas, incluindo aqui ambientes industriais. Esse processo é referenciado por Roiger and Geatz (2002) como um modelo completo, que passa desde o entendimento do negócio da organização até a implantação de aplicações ou modelos gerados a partir de técnicas de *Data Mining* para prover alguma melhoria no processo gerador dos dados.

Na prática, esse último modelo apresentado na Figura 2.3 é muito semelhante ao apresentado por (Berry and Linoff, 1997). Apesar de possuir menos etapas explícitas (são

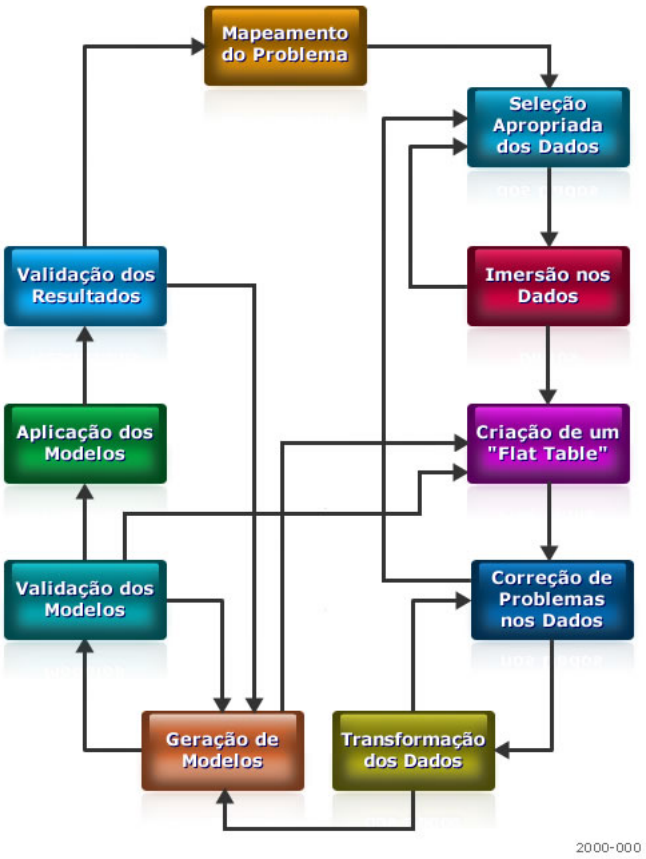

Figura 2.2: Etapas do processo de KDD, figura adaptada de Berry and Linoff (1997)

apenas 6 contra as 10 do modelo descrito na Figura 2.2), ao analisar o detalhamento das atividades vemos que as tarefas previstas são praticamente as mesmas, apenas estruturadas de forma um pouco diferente. Por exemplo, em um modelo existe uma etapa de Preparação dos dados. No outro, essa mesma atividade é explodida em 3 fases (Criação de um *dataset*; Correção de problemas nos dados; e Transformação dos dados) que nada mais são do que as atividades de Preparação.

Apesar de diferirem quanto ao número de fases, é consenso entre os vários autores que se trata de um processo iterativo e que algumas etapas mínimas devem ser cumpridas. No presente trabalho são consideradas as seguintes fases:

- Definição do domínio do problema;
- Seleção do dados;
- Conhecimento dos dados;

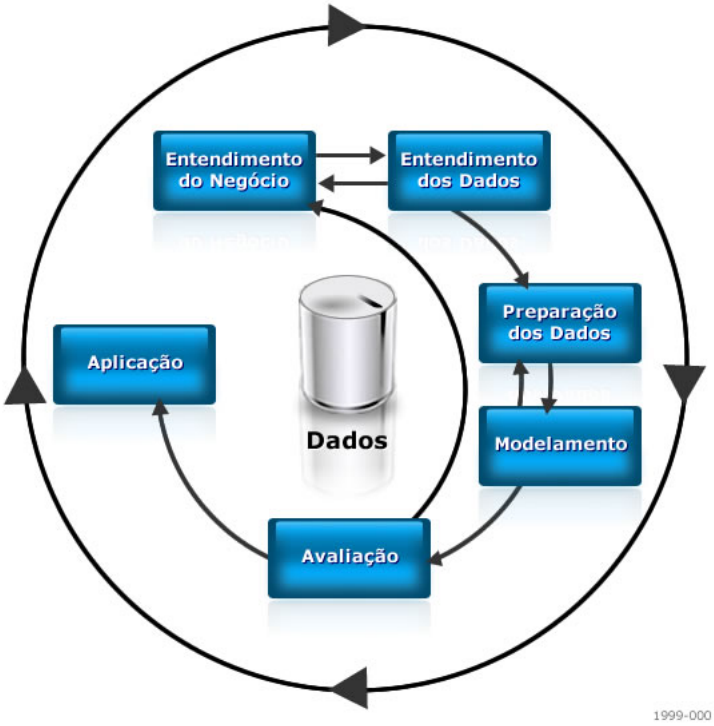

Figura 2.3: Etapas do processo de KDD - Modelo CRISP-DM, figura adaptada de Shearer (2000)

- Preparação e Limpeza dos dados;
- *Data Mining*;
- Interpretação e Avaliação do conhecimento;
- Aplicação do conhecimento obtido.

Por se tratar de um processo complexo, as etapas sugeridas podem não cobrir na sua plenitude todos os passos existentes; porém essa seqüência pode ser considerada como os requisitos mínimos para a extração de conhecimentos de uma base de dados, e a aplicação dos mesmos ao processo do qual esses dados são originados. Nas próximas seções serão detalhadas cada uma das fases, uma vez que o entendimento claro das mesmas é essencial para compreensão do processo como um todo.

2.2.1 Definição do domínio do problema

No conceito de KDD, a primeira etapa de definição do domínio do problema é importante para guiar o processo como um todo. Sem uma definição de que rumo seguir,

as inúmeras possibilidades irão dificultar a convergência e o alcance de algum resultado significativo.

Uma das dificuldades da etapa inicial do processo é o fato de não existir padronização para a mesma (Oliveira, 2000). Apesar disso, algumas considerações são relevantes e podem, de certa forma, servir de guia nessa fase. Dentre elas, pode-se destacar:

- Definição dos objetivos do projeto de KDD;
- Levantamento das possíveis bases de dados com as quais se pretende trabalhar;
- Pesquisa sobre o conhecimento prévio existente.

Com relação aos objetivos levantados, deve-se ter em mente que os mesmos irão guiar o processo de KDD como um todo e, portanto, devem estar claramente definidos. O levantamento de um grande número de objetivos não é recomendável, uma vez que pode dificultar a convergência do trabalho.

No levantamento das bases de dados, além da correlação com os objetivos, deve-se buscar fontes de dados potencialmente úteis para o processo de KDD. Como essa definição é de certa forma subjetiva, é importante o envolvimento de especialistas na área de negócio, ou domínio do problema em que o processo de KDD está sendo aplicado (ex.: Médicos, Engenheiros, Gerentes de *Marketing*, Astrônomos, etc).

Quanto ao conhecimento prévio existente, deve-se buscar informações que facilitem o trabalho de conhecimento dos dados como, por exemplo, outros trabalhos já realizados com a massa de dados, ou até mesmo impressões dos especialistas nos processos que geram esses dados quanto a possíveis problemas existentes. Qualquer informação adicional sobre os dados pode ajudar a guiar o processo de KDD.

2.2.2 Seleção do dados

A etapa de seleção dos dados consiste no trabalho de separação, dentre as bases de dados disponíveis, daquelas que serão utilizadas para atender os objetivos levantados. Tal etapa pode ser muito simples ou muito complexa, dependendo de alguns fatores relacionados aos dados:

- Quantas bases de dados relacionadas aos objetivos estão disponíveis?

- Essas bases são internas (pertencem a diretamente ao domínio da aplicação) ou são externas (de domínio público, como a Internet, por exemplo)?
- Qual é a estrutura de cada base de dados?

As respostas para tais perguntas irão guiar o processo de seleção dos dados. Quando estão disponíveis mais de uma base de dados, pode-se optar por trabalhar inicialmente com a que se julga mais relevante, ou prever uma integração entre as mesmas. O processo de integração por si só pode até requerer um projeto específico, dependendo do tamanho e estrutura das bases de dados envolvidas.

A utilização de bases de dados externas deve sempre ser avaliada com cuidado. Normalmente, esses dados necessitam de um maior esforço de formatação e preparação inicial do que as bases de dados internas. Além disso, a pouca ou nenhuma informação sobre o processo que gera tais dados pode dificultar ainda mais a etapa seguinte do processo de KDD, ou seja, o entendimento desses dados.

A estrutura de armazenamento de cada base de dados com que se vai trabalhar deve ser analisada. Questões de infra-estrutura como espaço em disco, plataformas e ambientes de gerenciamento de dados devem aflorar nessa análise. Deve-se levar em conta que cada base de dados possui uma estrutura interna própria, podendo ser uma simples tabela, passando por um banco de dados relacional ou até um desenho mais complexo de um cubo multi-dimensional.

Uma análise crítica de cada questão acima é necessária para definir qual(is) base(s) de dados serão utilizadas. É importante ressaltar que a seleção não é definitiva e, dependendo do resultado de etapas futuras, uma nova seleção pode ser necessária.

2.2.3 Entendimento dos dados

No conceito de KDD, o entendimento dos dados é o passo mais importante e talvez a tarefa mais penosa de todo o processo. Segundo Berry and Linoff (1997), um bom trabalho de descoberta de conhecimento em bases de dados passa pela intuição do profissional que o está realizando (ex.: decidir qual variável irá guiar um aprendizado supervisionado). Porém, a única maneira de se desenvolver intuição sobre um repositório de dados desconhecido é imergir por completo no mesmo. Ou seja, obter um bom entendimento dos dados com os quais se vai trabalhar.

O entendimento do processo que gerou os dados é um bom ponto de partida. A realização de uma análise estrutural da forma de armazenamento dos dados também é recomendada. No caso de um banco de dados relacional, o entendimento das tabelas envolvidas, bem como o relacionamento das mesmas, é importante. A identificação das principais tabelas e do significado dos atributos (colunas) de cada uma também devem ser feitos.

A análise estatística (ex.: médias de atributos numéricos[1], distribuições de variáveis categóricas[2], entre outras) é uma excelente forma de se conhecer os dados (Hand *et al.*, 2001). Essa análise pode revelar informações relevantes e, na maioria dos casos, contribui para uma visão mais clara do repositório. Durante essa análise informações importantes podem surgir, como por exemplo a observação de que, dentro de uma variável categórica, a maioria dos registros assume um determinado valor; ou ainda, a percepção que duas variáveis apresentam sempre o mesmo valor em todos os registros, etc.

Dependendo da natureza dos dados, a utilização de recursos gráficos (tendências, histogramas, etc.) também pode ser de grande valia. Por exemplo, para análise de variáveis que correspondem a séries temporais, gráficos de tendência podem indicar presença de picos e valores muito fora da média. A análise de histogramas indica a distribuição de variáveis categóricas (ex.: a proporção entre homens e mulheres em uma base de dados de clientes de um banco).

A maioria das ferramentas comerciais ou acadêmicas de *Data Mining* possui módulos específicos que dão suporte em maior ou menor grau para a etapa de Conhecimento dos dados com a disponibilização de funções estatísticas e recursos de visualização dos dados. Algumas, como o *Intelligent Miner* da IBM, possuem módulos dedicados apenas para facilitar a análise inicial dos dados. Outras, como o ROSETTA: *A Rough Set Toolkit for analysis of Data* (1998), já calculam automaticamente os principais indicadores estatísticos de todas as variáveis numéricas. Independente da ferramenta ou técnica aplicada, o importante é não avançar no processo de KDD sem se sentir familiarizado com a base de dados. A etapa seguinte, de preparação e limpeza, só terá valia se forem identificados quais ações devem ser realizadas sobre a massa de dados.

[1]Atributos numéricos são aqueles que assumem valores reais, seja de ponto flutuante - ex.: 3,78 - ou inteiro - ex.: 2. Também são freqüentemente referenciados como atributos ou variáveis contínuas.

[2]Variáveis categóricas são aquelas que assumem valores específicos, normalmente de texto - ex.: A variável categórica **sexo** pode assumir os valores **Masculino** ou **Feminino** - Também são freqüentemente citadas como variáveis ou atributos discretos de uma base de dados.

2.2.4 Preparação e Limpeza dos dados

Uma vez obtido um conhecimento inicial sobre a base de dados, é hora de começar a preparar a mesma para a etapa de *Data Mining*. A maioria das ferramentas disponíveis no mercado trabalha com o conceito de *flat table* como estrutura de dados de entrada para essa etapa (Ye, 2003). Uma *flat table* nada mais é do que uma tabela única com todos os dados que se pretende analisar.

Para gerar a *flat table* inicial, pode ser necessária a realização de junções entre tabelas de uma determinada base de dados ou até de bases de dados distintas. Precisa-se também determinar qual será o tamanho (número de registros) da tabela. É possível realizar a opção de trabalhar com todos os dados disponíveis, ou somente com uma amostra dos mesmos. A análise estatística da distribuição dos dados pode contribuir na tomada de decisão.

Um outro ponto relevante é com relação à limpeza dos dados. Os algoritmos de *Data Mining* normalmente não trabalham bem com dados incompletos; além disso, os dados podem conter valores fora das faixas normais para determinados atributos. As ferramentas mais completas de *Data Mining* também suportam a execução da limpeza dos dados. Algumas, como o *Statistica: Data Mining Software* (2004), possuem módulos com funções especificas para ajudar no trabalho. Entre estas funções podem se destacar:

- Eliminar automaticamente registros com atributos em branco ou fora de faixa;
- Completar atributos em branco com valores pré-definidos (médias, por exemplo);
- Aplicar filtros sobre os dados;
- Agrupar variáveis categóricas com distribuição elevada (vários valores distintos).

Algumas técnicas específicas de *Data Mining* podem exigir ainda mais trabalho na etapa de preparação, como por exemplo a utilização de redes neurais, que exige uma normalização dos dados. A categorização de variáveis também pode ser necessária quando se quiser trabalhar com técnicas como os *Rough Sets* (Cios *et al.*, 1998; Pawlak, 1991) ou Regras de Associação.

2.2.5 Data Mining

A etapa de *Data Mining* consiste basicamente no processo de aplicação de algoritmos das áreas de inteligência artificial e estatística sobre a *flat table* elaborada na etapa anterior. Antes de se definir qual algoritmo se irá aplicar, é necessário a escolha de uma estratégia específica. A Figura 2.4 apresenta a hierarquia na qual estas estratégias estão estruturadas (Roiger and Geatz, 2002).

Figura 2.4: Hierarquia de Estratégias de *Data Mining*, figura adaptada de (Roiger and Geatz, 2002)

A estratégia de agrupamento não-supervisionado tem como objetivo encontrar grupos distintos dentro da massa de dados, e definir quais são as características que distinguem esses grupos. Tal estratégia é recomendada quando não se possui muita informação acerca dos dados, ou quando não se possui uma variável que guie um aprendizado supervisionado. Vários algoritmos podem ser utilizados nesse caso, como por exemplo o *K-Means* (MacQueen, 1967) ou as Redes Neurais de Kohonen (SOM - *Self Organizing Maps*) (Haykin, 1999).

O propósito das Análises de Cestas de Supermercado (do inglês *Market Basket Analysis*) é encontrar relações interessantes entre atributos dentro de uma massa de dados (Roiger and Geatz, 2002). A técnica surgiu da análise dos registros de compras de uma grande cadeia de supermercados nos Estados Unidos e por isso leva este nome. Algoritmos de Regra de Associação (Menzies and Hu, 2003) são normalmente utilizados para aplicação da estratégia. As regras geradas pelo algoritmo descrevem relações entre os atributos, e

podem servir tanto para identificar padrões desconhecidos quanto para validar hipóteses acerca do processo que gera os dados.

Já o aprendizado supervisionado parte do princípio que, dentro da *flat table* selecionada, existem uma ou mais variáveis que possam guiar o processo. Existem basicamente 3 tipos de estratégias com tal premissa (Roiger and Geatz, 2002):

1. Classificação: Construir modelos para classificar os dados de acordo com uma variável-guia (dependente) categórica. Essa estratégia está mais relacionada ao comportamento corrente dos dados e não com uma previsão sobre o futuro. Exemplo: construir um modelo de classificação para determinar se um candidato a empréstimo em um banco apresenta o risco de não saldar a dívida - ou seja, tomar essa decisão baseado nos fatos existentes hoje.

2. Estimação: Construir modelos para estimar o valor de uma variável desconhecida. Diferentemente do modelo de classificação, nesse caso a variável que guia o processo é numérica. Também está relacionada com o comportamento corrente dos dados. Exemplo: Com base nas informações de um banco de dados de clientes de cartões de crédito, estimar a possibilidade de um cartão ter sido clonado.

3. Predição: Essa estratégia consiste na tentativa de se prever o comportamento futuro de uma determinada variável, baseada no histórico existente nos dados. A variável dependente aqui pode ser tanto numérica quanto categórica. Exemplo: Prever qual a probabilidade de chuva com base nos dados históricos das variáveis climáticas de uma determinada região.

De maneira geral, as 3 estratégias apresentadas são muito semelhantes e existem algoritmos de aprendizado supervisionado que suportam todas elas. O que realmente define a melhor estratégia (classificação, estimação ou predição) é a natureza dos dados (Roiger and Geatz, 2002).

Existem vários algoritmos que suportam o aprendizado supervisionado. Podemos classificar esses algoritmos dentre os seguintes grupos, dependendo do paradigma de cada um:

- Aprendizados que constroem representações simbólicas, como por exemplo as árvores de decisão e as regras de associação. Um exemplo de aplicação desse tipo de

algoritmo é o CART (*Classification And Regression Tree*, ou Árvore de Regressão e Classificação) (Breiman *et al.*, 1984). Este algoritmo especifico será detalhado mais a frente na seção 4.4.3;

- Algoritmos que utilizem técnicas estatísticas, como por exemplo inferência bayesiana. Os modelos estatísticos são muito utilizados para tratar dados numéricos (Akaike, 1974);

- Redes Neurais Artificiais (RNA), que implementam um paradigma conexionista baseado em uma abstração do funcionamento das células nervosas que constituem o cérebro. As RNAs são construções matemáticas simples que possuem capacidade de aprender por exemplos e fazer interpolações e extrapolações do que aprenderam (Braga *et al.*, 2000).

A escolha adequada da estratégia de *Data Mining* pode, de certa forma, ser guiada pela característica do problema que se quer resolver. Porém, como as opções são muitas, outros fatores como disponibilidade, usabilidade e performance das ferramentas e algoritmos também são relevantes.

2.2.6 Interpretação e Avaliação do conhecimento

Uma vez vencida a etapa de *Data Mining*, é preciso interpretar e avaliar os resultados obtidos pelas ferramentas utilizadas. A análise deve ser realizada por alguém que conheça bem o processo que gerou os dados, para validar se, dentro dos padrões e informações encontradas, existe algum conhecimento novo e útil.

As saídas dos algoritmos de *Data Mining* ocorrem de várias formas distintas. Os que implementam árvores de decisão normalmente geram uma estrutura, de modo a permitir a visualização da hierarquia entre os nós da árvore. Através de uma análise dessa estrutura, é possível identificar as variáveis mais importantes para o modelo de classificação.

Já os algoritmos de regras de associação, como o próprio nome indica, geram regras sobre os dados com uma determinada "cobertura" (percentual de registros para qual a regra vale) e "confiança" (precisão da regra, dentro da faixa de dados à qual ela se aplica). Normalmente, o número de regras geradas é muito grande e alguma técnica de filtragem deve ser utilizada antes da análise das mesmas.

As redes neurais, por sua vez, por serem algoritmos do tipo "caixa-preta", normalmente não geram nenhuma informação que não o próprio modelo de classificação ou definição de classes no caso do aprendizado não-supervisionado.

Caso os resultados obtidos não alcancem os objetivos propostos e não sejam validados pelos analistas responsáveis pelos dados, é necessário retornar a alguma das etapas anteriores e refazer parte do trabalho. Tal iteração pode ocorrer várias vezes, até que se encontre um resultado satisfatório ou se conclua não ser possível extrair algum conhecimento dos dados através do processo de KDD.

2.2.7 Aplicação do conhecimento obtido

Uma vez identificado um conhecimento novo e potencialmente útil, o mesmo pode ser utilizado para realimentar o processo no qual o dados foram gerados. O desenvolvimento de um sistema baseado em um modelo obtido, ou o ajuste de alguma aplicação existente, baseada em regras de negócio identificadas, são exemplos de aplicações do conhecimento extraído de um processo de KDD. O resultado final pode gerar também uma simples orientação para pessoas que trabalham na geração dos dados.

De qualquer forma, a comunicação do conhecimento adquirido deve ser feita para todas as partes envolvidas, e a melhor aplicação do mesmo passa por uma análise mais sistêmica do domínio do problema. Novos ciclos completos do processo de KDD podem ser demandados para se aprofundar questões levantadas e buscar ainda mais conhecimento na massa de dados.

2.3 KDD Aplicado às Indústrias de Processos

É relativamente recente a aplicação do processo de KDD à Indústria de Processos. Por exemplo, o artigo mais antigo encontrado no IEEE sobre esse assunto data de 1997, com o título *Data Mining: An Industrial Research Perspective* (Apté, 1997). O texto faz apenas uma introdução sobre o assunto e não apresenta nenhum exemplo de aplicação. Um ano depois, Mastrangelo (1998) apresenta resultados reais na indústria química com a utilização do algoritmo CART para identificar as variáveis mais relevantes no processo de fabricação de fibras de nylon. Essa mesma técnica é utilizada por Arieh and Chopra (1998) num processo de seqüenciamento de produção. O classificador é utilizado dentro do sistema de controle distribuído da planta, para indicar quais máquinas são mais adequadas

para produzir cada tipo de material.

O uso de técnicas de *Data Mining* com o objetivo de se obter modelos de uma planta termo-elétrica é descrito por Ogilvie *et al.* (1998). Regras de Associação são geradas, através da base de dados de processo, e confrontadas com o Sistema Especialista existente na planta para atualização das mesmas. O conceito de extração de regras do processo industrial também é aplicado por Irizuki *et al.* (1999), porém numa abordagem diferente, com a utilização de um sistema *neuro-fuzzy* em cascata com o sistema de controle de uma refinaria.

Roed (1999) trata da utilização do processo de KDD para tratar séries temporais em uma base de dados de um processo industrial. A grande contribuição desse exemplo é a utilização da teoria de *Rough Sets* para identificar padrões. Como ferramenta de *Data Mining*, é introduzido o *software ROSETTA* que possui um *framework* completo e específico para aplicação dessa teoria.

Wang (1999) apresenta de forma bem detalhada várias técnicas de *Data Mining*, além de fazer uma série de considerações sobre o processo de KDD nas indústrias de processos. Entre elas destacam-se:

- Análise Estatística Multi-variada: Utilização de técnicas como *PCA - Principal Component Analysis* para a identificação de variáveis mais importantes em processos industriais;
- Aprendizado supervisionado com utilização de redes neurais para definição de modelos de identificação de falhas em processos;
- Aprendizado não-supervisionado para identificar comportamentos semelhantes entre variáveis com a implementação de um classificador Bayesiano.

Kusiak (2000) descreve várias formas de decomposição dos dados de um processo industrial (fabricação de semi-condutores) para melhorar a qualidade dos modelos extraídos por ferramentas de *Data Mining*. As relações obtidas, em forma de regras, foram utilizadas para melhorar a qualidade do produto final, mostrando um exemplo claro onde o processo de KDD foi aplicado com sucesso. Um trabalho semelhante é apresentado por Yoshida and Touzaki (2000), porém com mais ênfase na seleção das regras geradas.

Técnicas de seleção e filtragem de dados são apresentadas por Morello *et al.* (2001) dentro da etapa de preparação no processo de KDD. O objetivo aqui era a extração

do conhecimento do processo de funcionamento de quatro máquinas num processo de batelada, onde uma questão muito importante é o seqüenciamento destes equipamentos. O algoritmo C4.5 (Quinlan, 1993) é utilizado para gerar uma árvore de decisão. O autor ressalta a importância da filtragem dos dados para obtenção de bons resultados, porém não entra em detalhes sobre os conhecimentos adquiridos a partir da base de dados.

Outra aplicação de KDD na indústria de semi-condutores é descrita por Braha and Shmilovici (2002) com a utilização de técnicas de *Data Mining* para melhorar o processo de limpeza das impurezas dos componentes. O trabalho destaca como o processo de KDD pode ser útil mesmo quando o número de dados não é muito grande. Outro ponto ressaltado é a facilidade apresentada por algumas técnicas para tratar, de forma quase transparente, as grandes não-linearidades e correlações das variáveis do processo.

Gao *et al.* (2003) apresentam uma abordagem mais atual do tratamento do acúmulo de informações por um complexo processo da indústria química para otimizar os sistemas de controle automáticos. Um modelo de extração de conhecimento com técnicas de *Data Mining* e realimentação de algumas malhas de controle do processo são apresentadas. Os autores levantam questões interessantes sobre o futuro da aplicação do processo de KDD em bases de dados industriais, como a possível interação dos sistemas de *Data Mining* com outras técnicas inteligentes, que são normalmente utilizadas pelos algoritmos de controle, para com isso melhorar ainda mais o desempenho das malhas[3] industriais.

A extração de conhecimento de um sistema integrado de execução da manufatura é descrito por Zhong and Wang (2003). Apesar de não utilizar a nomenclatura MES, o sistema descrito possui características que o enquadra na mesma camada dentro de uma pirâmide de automação, ou seja entre o nível de controle e supervisão (CLP e SCADA) e dos sistemas ERP (*Enterprise Resource System*). A estrutura do processo de KDD para obter informações relevantes acerca dos dados gerados pelo sistema é apresentada, porém com poucos detalhes e sem exemplos de resultados obtidos.

De maneira geral, observa-se uma certa variação de técnicas sendo utilizadas na aplicação de KDD na indústria de processos. Não podemos afirmar que existe uma forte tendência para maior ou menor utilização de uma ou outra estratégia. É importante mencionar que nenhum dos trabalhos citados faz menção direta ao uso de dados de aplicações

[3]O termo **malha** aqui se refere as malhas de controle que nada mais são do que os sistemas compostos pelos sensores (ex.: pirômetros), atuadores (ex.: válvulas) e algoritmos de controle, além dos equipamentos de processo propriamente ditos.

PIMS para a extração de conhecimento. Porém, várias aplicações tratam da utilização de grandes massas de dados temporais, característica principal de uma aplicação PIMS. Na maioria dos casos descritos, o processo de KDD tem seu alvo em medidas de variáveis de processo. Além disso, observam-se alguns exemplos de utilização da estratégia de aprendizado supervisionado com aplicação do algoritmo CART para identificação de variáveis mais importantes para um determinado processo. Essa técnica será melhor descrita na seção 4.4.3, junto à sua aplicação em uma planta de Laminação a Quente.

2.4 KDD na Siderurgia

Não são muitas as publicações sobre extração do conhecimento de bases de dados na indústria siderúrgica. A maioria dos registros encontrados se resume à área de Laminação, talvez pelo fato da automação destas áreas ser mais recente e permitir uma maior aquisição de dados.

No final da década de 90, técnicas de redes neurais começaram a ser aplicadas, já dentro do conceito de KDD, ao processo de produção de aço. Ge (1999) mostra como pode ser eficaz a modelagem da temperatura do ferro-gusa através do treinamento de uma rede com dados do processo. Cser *et al.* (1999) e Himberg *et al.* (2001) apresentam a utilização de redes neurais do tipo SOM para buscar correlações e dependências "escondidas" nas bases de dados de produção em plantas de Laminação a Quente. Já Elsila and Röning (2002) retrata a aplicação completa do processo de KDD, porém com o objetivo específico de identificar as condições nas quais ocorrem a desclassificação[4] de bobinas. Outra publicação, mais recente, também identifica um objetivo claro: mapear as principais causas de ocorrência de carepas [5] nas placas de aço (Haapamäki *et al.*, 2005). Essas publicações são detalhadas a seguir, ainda nessa seção. Já maiores informações sobre o processo siderúrgico, e mais especificamente sobre o processo de Laminação de Tiras a Quente, são encontradas na seção 3.

Uma revisão sucinta dos sistema de produção de uma planta de Laminação de

[4] O conceito de desclassificação de um material no processo siderúrgico corresponde ao fato do mesmo não atender às especificações mínimas de qualidade exigidas. Estas especificações variam desde características físicas, como dimensão e curvatura, até características químicas como composição de carbono, enxofre e outro elementos presentes nos diversos tipos de aço.

[5] *Carepa* é a denominação dada na siderurgia à película formada na superfície das placas de aço, devido principalmente à forte reação de oxidação do aço com o ar ambiente assim que a placa se forma a temperaturas elevadas - cerca de 900°C - na etapa de Lingotamento Contínuo (processo de conformação e solidificação do aço).

Aço é apresentada por Cser *et al.* (1999). Além de considerações sobre a amostragem e armazenamento, são destacados os pontos de coleta dos dados de processo. As análises foram feitas em cerca de 16.000 registros de produção de bobinas, cada um com mais de 70 variáveis, ou medidas de processo. O resultado da aplicação de KDD trouxe algumas informações interessantes sobre parâmetros técnicos, possibilitando melhoras na qualidade do produto final, porém as mesmas não são divulgadas.

Já Himberg *et al.* (2001) não descrevem as etapas iniciais do processo de KDD, e partem do pressuposto que 47 variáveis já haviam sido previamente selecionadas e agrupadas dentro de uma *flat table* com registros de produção de bobinas. Os resultados apresentados mostram grupos de variáveis com características semelhantes, o que permitiu que especialistas identificassem algumas regras de relevância para o processo, entre as quais destacam-se as seguintes:

- A variação de espessura das tiras tende a diminuir à medida que se a força de flexão dos cilindros das cadeiras de laminação é maior - comportamento mais destacado para tiras de menor espessura;

- Quanto maior a temperatura da tira que se está sendo laminada, maior também a variação de espessura.

Como a variação de espessura é uma variável-chave para o controle de qualidade na produção das bobinas, estas regras foram consideradas úteis para a realimentação do processo produtivo. O autor não entra em detalhes de como o conhecimento adquirido foi aplicado.

Ambos os trabalhos optaram pelo aprendizado não-supervisionado como estratégia de *Data Mining*, por não terem um problema específico ou mesmo um objetivo claro para o processo de KDD. Observa-se que, nesses dois últimos exemplos, o sucesso da aplicação da metodologia estava altamente correlacionado à obtenção de informações novas e úteis relativas ao processo de laminação, devido principalmente à grande quantidade e variabilidade dos dados armazenados nos repositórios analisados.

Ao atacar o problema de desclassificação de bobinas através da análise dos dados do processo de Laminação a Quente de uma usina Siderúrgica na Finlândia, Elsila and Röning (2002) mostra como o processo de KDD pode ser aplicado para resolver problemas na indústria. Quando uma bobina é desclassificada, isto significa que a mesma não poderá ser

repassada diretamente para o cliente e que será sucateada ou reaproveitada para atender outras especificações menos rigorosas. Em ambos os casos, há perda de produtividade. O objetivo de tal trabalho é analisar as centenas de variáveis coletadas em alta freqüência pelos sistemas de automação da planta, de forma a construir um modelo que consiga prever se uma determinada bobina será desclassificada. Para obter correlações entre as variáveis, várias técnicas foram utilizadas nas etapas do processo:

- Análise Estatística na etapa de conhecimento dos dados;
- Correlações Lineares, para retirar da massa de dados variáveis redundantes (preparação dos dados);
- *Self Organizing Maps* (SOM), Coordenadas Paralelas (Inselberg, 1998) e Agrupamento (aprendizado não-supervisionado) com a utilização do algoritmo *k-means* (MacQueen, 1967) foram as técnicas empregadas na etapa de *Data Mining*.

O principal resultado foi a identificação de dependências entre os códigos de desclassificação de bobinas e grupos de variáveis de processo. Segundo Elsila and Röning (2002), esta metodologia ajudou os especialistas de processo a reduzir o número de desclassificações de bobinas.

Outra questão de grande relevância no processo siderúrgico que motivou a aplicação de técnicas de extração do conhecimento de bases de dados é a origem das carepas. Essas películas são consideradas defeitos de fabricação do aço, e as causas de ocorrência destes efeitos não são diretas. Existem vários tipos de carepas, e apesar de quase sempre estarem também associadas ao processo de oxidação do aço, outros fatores podem contribuir para a maior ou menor formação dessas camadas. Haapamäki *et al.* (2005) apresentam a construção de um modelo de predição de ocorrência de carepas, bem como um mecanismo de visualização das mesmas. Além de utilizarem uma rede neural *perceptron multi-camada* (Braga *et al.*, 2000) para construir o modelo preditivo a partir dos dados de processo, as redes SOM são utilizadas para facilitar a visualização das carepas. Como resultado tangível do processo de KDD são apresentadas correlações interessantes (não previstas anteriormente) entre a temperatura e a formação de certos tipos de carepa em aços específicos. Foi levantada ainda, por tais autores, a hipótese de revisão no processo de resfriamento.

De maneira geral, observamos uma forte utilização das redes neurais SOM como estratégia de *Data Mining* para atacar problemas relacionados a aprendizado não-supervisionado na área de Laminação a Quente. Outra característica marcante desse processo é a abundância de dados disponíveis. Os sistemas atuais de automação das plantas de laminação são capazes de armazenar milhares de informações sobre o processo de fabricação de uma única bobina. Acredita-se, portanto, que esse cenário é definitivamente favorável à aplicação de KDD para suportar os especialistas dessas indústrias a extraírem conhecimento das bases de dados existentes.

Capítulo 3

Estruturação do Processo de KDD na Laminação de Tiras a Quente da Usina Siderúrgica

3.1 O processo de Laminação de Tiras a Quente

O processo de Laminação de Tiras a Quente corresponde à transformação física de placas de aço, produzidas em Aciarias, em bobinas de aço através de desbastes realizados por cilindros de laminação. As placas são primeiramente aquecidas a temperaturas superiores a 1000 °C no forno de reaquecimento, visando garantir material que possa ser laminado ao longo da linha. No Laminador Desbastador é feito o primeiro processo de laminação, tanto na largura como na espessura da placa. Nessa etapa, a placa tem sua espessura reduzida a cerca de um quinto da espessura original, através de vários passes sucessivos, sendo a partir daí conhecida como "esboço". A Tesoura de Pontas corta as aparas nas extremidades do esboço, que vai finalmente para o Trem Acabador no qual várias cadeiras fazem a redução final da espessura até as especificações do cliente. A última etapa do processo, antes do bobinamento final, é o resfriamento da tira através de jatos de água fria. A Figura 3.1 mostra as principais etapas desse processo.

O produto laminado a quente atende diversas normas nacionais e internacionais, possuindo uniformidade de propriedades mecânicas e tolerância restrita de espessura, largura e forma. Dentre as variadas aplicações do produto, podem ser citadas:

- Rodas;
- Longarinas;
- Embreagens;

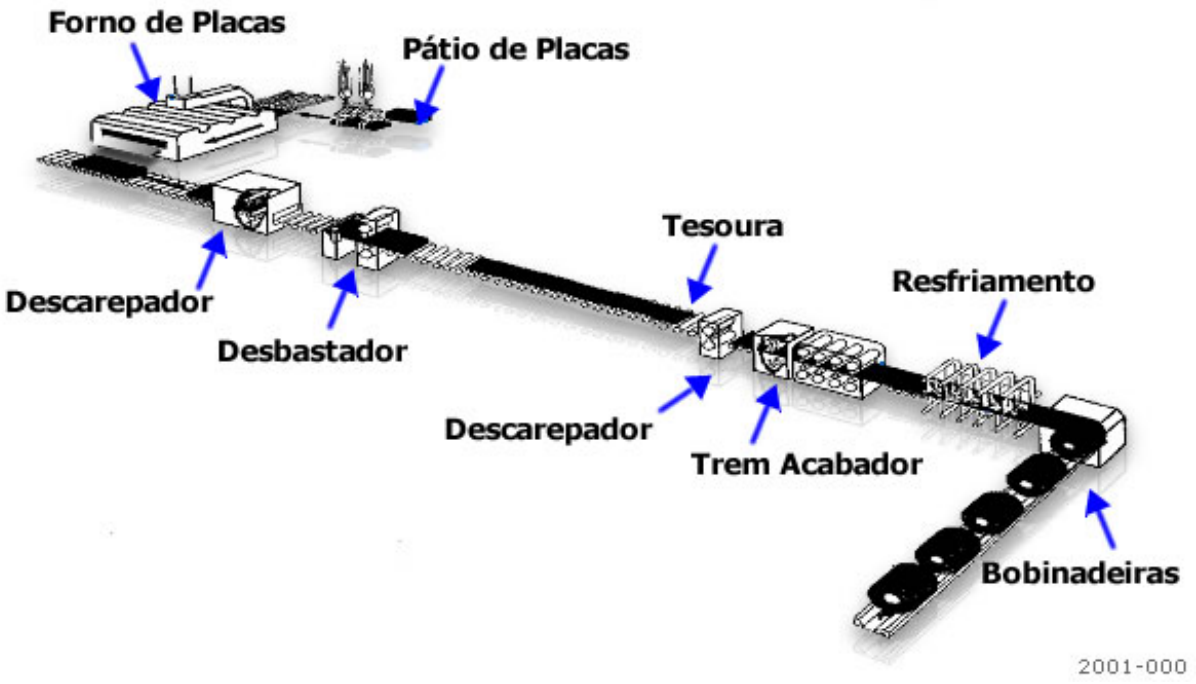

Figura 3.1: Etapas do processo de Laminação de Tiras a Quente

- Botijões;
- Compressores;
- Construção civil;
- Tubos;
- Perfis estruturais;
- Construção naval;
- Implementos agrícolas;
- Matéria-prima para laminação a frio, folha metálica e produtos galvanizados.

A planta de Laminação de Tiras a Quente da Siderúrgica, onde se realizou esse trabalho, possui um elevado grau de automação. Vários sensores e medidores são utilizados para aquisição das informações necessárias ao controle do processo. Observa-se um elevado número de sensores de temperatura (pirômetros) que estão espalhados por toda a linha. Isso se deve ao fato desta grandeza ter que ser muito bem controlada para garantir as propriedades mecânicas do aço. Outros medidores como de largura, espessura

e velocidade, bem como medidores especiais, também são encontrados, conforme figura 3.2.

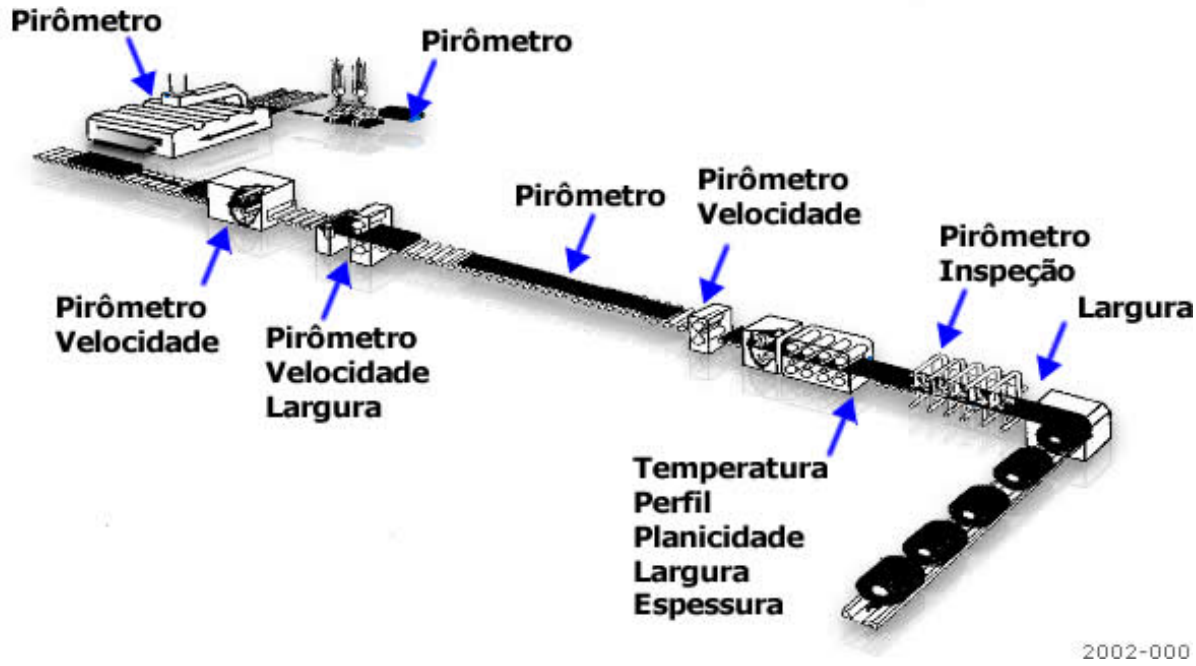

Figura 3.2: Principais medidas ao longo da linha de Laminação

A maioria das medições realizadas pelos sensores, além de utilizadas pelos sistemas de controle e de gerenciamento da produção, também são armazenadas em bancos de dados históricos. A estrutura desse armazenamento será descrita na seção 3.3.

3.2 Levantamento de problemas e respectivas possibilidades de aplicação de KDD

A etapa de levantamento de problemas relativos ao processo de Laminação de Tiras a Quente começou com uma apresentação dos conceitos de KDD para a equipe de Automação da Usina Siderúrgica. Após essa introdução, os analistas vislumbraram de imediato três possibilidades de aplicação, buscando a solução de problemas através da obtenção de conhecimento sobre dados de processo.

A primeira possibilidade foi relativa a um sistema de indicadores de qualidade do processo de produção de bobinas. O sistema classifica algumas variáveis dentro de faixas de qualidade (ex.: OK ou não OK) de acordo com valores pré-determinados. Essa classificação é atribuída às principais características do produto final, como largura, espessura,

planicidade, etc.

Para fins de análise de qualidade, a bobina é divida em três partes: cabeça, corpo e cauda, conforme diagrama da figura 3.3.

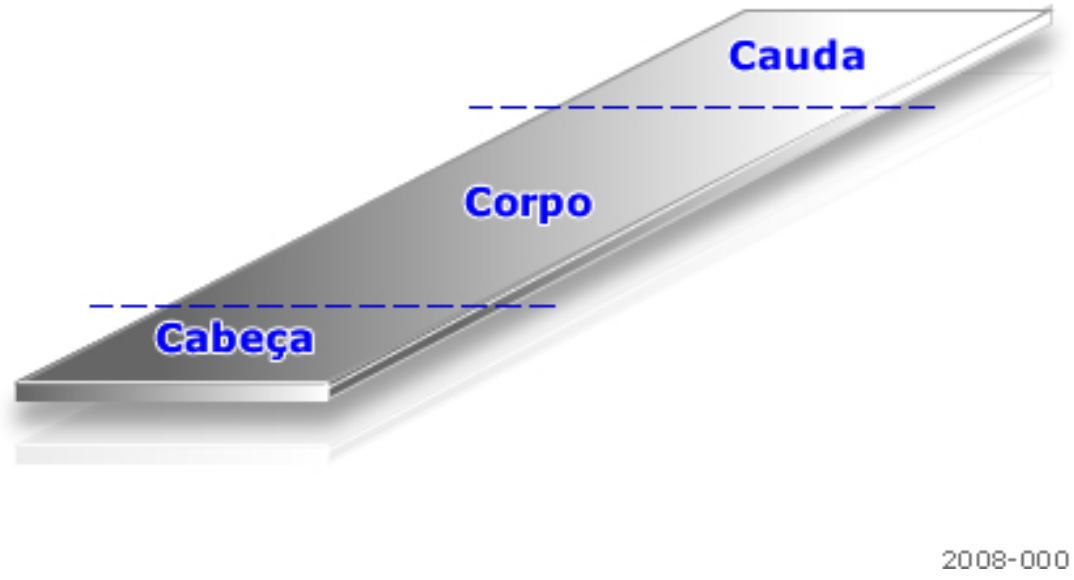

Figura 3.3: Diagrama esquemático de uma bobina

A questão a ser tratada, seria a análise dos dados de processo utilizados pelo sistema de indicadores de qualidade, buscando padrões que poderiam explicar a origem de problemas como, por exemplo, a reprovação de bobinas.

Além dessa possibilidade, foram identificadas mais duas questões que poderiam ser tratadas dentro do projeto de aplicação de KDD:

1. Erro de Força de Laminação: o erro entre o *setup* da força de laminação na cabeça da bobina e a força real aplicada deve ser menor do que um percentual determinado, segundo padrões de qualidade da Usina. Atualmente, o número de bobinas cujos erros de força superam esse percentual tem sido maior do que a meta de qualidade da Siderúrgica.

2. Erro de Largura: o erro de largura de uma bobina deve ser menor do que um determinado percentual, quando se considera a largura de *setup* e a largura real da bobina. O número de bobinas produzidas pelo Laminador de Tiras a Quente com erros acima desse percentual também tem sido maior que o esperado.

Após algumas reflexões optou-se pelo foco inicial do trabalho no problema de erro

de força de laminação, uma vez que esse foi considerado de mais difícil solução do que a questão da largura e também, por ser um objetivo mais específico do que a análise dos dados relacionados ao sistema de indicadores de qualidade. Portanto, definiu-se como objetivo do projeto de KDD a busca por conhecimentos que auxiliem na determinação das razões pelas quais um número significativo de bobinas supera o percentual definido como aceitável para o erro de força.

3.2.1 Definição do Erro de Força na Laminação

No Trem Acabador, em cada uma das cadeiras de laminação, uma força é aplicada sobre a cabeça da tira que está sendo laminada através dos chamados "cilindros de laminação". Essas forças estão entre as principais variáveis a serem controladas dentro do processo de laminação a quente, e seus cálculos são realizados a partir de um modelo matemático que leva em consideração vários fatores como temperatura e composição química do material. Para cada cadeira, o valor definido pelo modelo é denominado de "Força de *Setup*" e sua diferença em relação ao valor medido no processo ("Força Real") é denominado "Erro de Força" conforme descrito na equação 3.1:

$$Erro = \frac{F_{setup} - F_{real}}{F_{setup}} \tag{3.1}$$

O controle adequado do erro de força é muito importante para garantir que o produto final atenda às especificações técnicas e de qualidade. Para fins de análise, essa medida é classificada em 3 faixas, a partir do seu valor percentual absoluto, sendo A e B valores percentuais positivos e B sendo um número maior que A:

1. Erro de Força Aceitável - Valores entre 0
2. Erro de Força Indefinido - Valores entre A% e B%;
3. Erro de Força Ruim ou Inadequado - Valores maiores do que B%.

Essa categorização do erro de força de laminação será a principal orientação para o processo de busca do conhecimento dentro das bases de dados do processo de Laminação de Tiras a Quente. Maiores informações sobre o erro de força e a categorização acima serão apresentadas na seção 4.3.

3.3 Características das Bases de Dados da área de Laminação a Quente

A Usina Siderúrgica possui hoje, na área de Laminação a Quente, um sistema de aquisição de dados que reúne informações dos vários sistemas de controle da planta. Esse sistema é basicamente um PIMS; porém, além de armazenar os dados de forma temporal (*tag* vs. valor vs. *timestamp*), também armazena dados relacionais com informações sobre as bobinas produzidas, ou seja, apresenta a relação bobina vs. variável vs. valor.

O sistema busca informações de vários computadores de processo (*Process Computer*, ou PROCOM) e as consolida dentro de dois grandes bancos de dados, sendo um temporal e outro relacional. Os sistemas que executam em cada um desses PROCOMs controlam várias etapas e características do processo como:

1. Controle de largura do material nos laminadores verticais e redução de espessura;
2. Controle do forno de reaquecimento de placas;
3. Controle de temperatura no leito de resfriamento, através da vazão de chuveiros espalhados ao longo da linha;
4. Classificação de defeitos através de inspeção das superfícies da tira laminada;
5. Gerenciamento da oficina de cilindros, onde é realizada a manutenção dos cilindros desgastados durante a laminação;
6. Aquisição dos dados temporais de Nível 1 (CLPs e Sistemas Supervisórios).

Os analistas da Usina responsáveis pelos sistemas de automação indicaram a base de dados relacional do sistema de consolidação dos dados como a mais adequada para tratar a questão do erro de força. Essa base de dados é atrativa para aplicação de KDD pelo fato de agregar as informações dos vários sistemas que atuam durante cada etapa do processo.

A base de dados relacional do sistema de consolidação dos dados permite o armazenamento de informações de processo correspondentes a um período de até 6 meses de produção de bobinas. O tamanho desse repositório é de aproximadamente 240 *gigabytes*. Um procedimento de *backup* realiza constantemente uma cópia desses dados e a

armazena de forma compactada. Esse *backup* foi disponibilizado para análise no formato do banco de dados *Oracle*. Maiores informações acerca dos dados do Laminador de Tiras a Quente serão apresentadas nas seções 4.2 e 4.3.

Capítulo 4

Detalhamento Técnico do Processo de KDD Aplicado

4.1 Ferramentas de Software Utilizadas

Uma vez definidos tanto o objetivo quanto a base de dados, fez-se necessária uma etapa de preparação da infra-estrutura do projeto para aplicação do processo de KDD. Além da seleção e configuração das ferramentas de *Data Mining*, realizou-se também a instalação de Sistemas de Gerenciamento de Banco de Dados, ou SGBDs. Na seção 4.1.1 apresentam-se os *softwares* de *Data Mining* considerados e os critérios de seleção da ferramenta utilizada. Na seção 4.1.2 são descritos em linhas gerais dois SGBDs, além do *software* Microsoft Excel©, ferramenta utilizada nas etapas de preparação e análise dos dados.

4.1.1 Softwares de *Data Mining*

Existem diversas ferramentas de *Data Mining* disponíveis no mercado, sendo algumas gratuitas e outras existentes dentro de pacotes comerciais completos para análise e tratamento de dados. Algumas em especial foram analisadas dentro do contexto desse trabalho:

- IBM DB2 *Intelligent Miner* (1996);
- *ROSETTA*;
- iDA: *intelligent Data Analyzer* (2002);
- *Statistica*;

- Weka: *Data Mining Software in Java* (2005).

O *Intelligent Miner* é a solução da IBM para área de KDD. Essa robusta e completa ferramenta foi desenhada para trabalhar com grandes bancos de dados e pode ser executada em diversos sistemas operacionais. Apesar de possuir uma interface gráfica, a maneira mais prática de se trabalhar com o *Intelligent Miner* é através de comandos executados diretamente em um console. Contudo, uma das grandes limitações dessa solução é trabalhar apenas com o Banco de Dados proprietário da IBM, o DB2[1]. Para minimizar essa questão, o fabricante disponibiliza um aplicativo para facilitar a migração de Bancos de Dados de diversas plataformas para o formato DB2. O *Intelligent Miner*, além de possuir algoritmos para executar as mais diversas tarefas de *Data Mining* tais como associação, classificação, agrupamento e predição, possui também um módulo dedicado para visualização e análise gráfica dos dados. A instalação e configuração da ferramenta não é trivial, porém existe documentação disponível na internet no *site* da IBM[2].

O *software* ROSETTA, conforme descrito na seção 2.3 é voltado para a utilização da técnica dos *Rough Sets* para análise de dados. De maneira geral, os dados são analisados em buscas de regras que descrevam o seu comportamento. Um dos principais mecanismos da teoria dos *Rough Sets* é a eliminação de atributos da base de dados, sem reduzir o grau de consistência da mesma. Esse grau de consistência é inicialmente calculado e, se algum atributo for retirado sem afetá-lo de maneira significativa, o mesmo é desconsiderado na análise dos dados. Além de possuir uma série de ferramentas específicas para discretização dos dados[3], o ROSETTA possui diversos algoritmos para geração de regras. O *software* pode ser considerado uma ferramenta bem flexível (comunica-se com vários tipos de Bancos de Dados, exporta resultados obtidos para diversos formatos de arquivo) e de interface amigável (fácil utilização). A sua grande limitação está no fato de não trabalhar de forma direta com dados contínuos, o que pode gerar um trabalho enorme na etapa de preparação, dependendo do tamanho do repositório em questão.

Outra ferramenta analisada foi o iDA ou *intelligent Data Analyzer*. Esse *software* nada mais é do que um *plug-in* do Excel© que utiliza o aplicativo da Microsoft com repositório de dados e interface gráfica. O iDA permite a utilização de técnicas de apren-

[1]Na documentação do *Intelligent Miner* está descrito que alguns módulos do sistema são compatíveis também com a base de dados *Oracle*©.

[2]`http://www.ibm.com/br`.

[3]Uma das premissas para aplicação da teoria dos *Rough Sets* é que os dados estejam discretizados, ou seja, os algoritmos existentes não trabalham com dados contínuos.

dizado supervisionado e não supervisionado e, além de um mecanismo para geração de regras, possui dois algoritmos de redes neurais (um para classificação e outro para agrupamento - *clustering*). De maneira geral, o iDA é um aplicativo muito simples e segue as limitações do Excel com relação ao tamanho da base de dados. Além disso, em relação a outras ferramentas, é muito limitado em opções de algoritmos. Porém, para pequenas aplicações, o iDA pode se demonstrar muito útil, pois acima de tudo possui uma instalação simples e é de fácil utilização. Maiores informações sobre o *intelligent Data Analyzer* podem ser obtidas em (Roiger and Geatz, 2002).

O *Statistica* é um pacote completo para análise e tratamento de dados. Além de implementar as mais diversas técnicas estatísticas (desde uma simples análise descritiva até complexos cálculos de distribuição e probabilidade), o aplicativo da STATSOFT possui um módulo específico para *Data Mining*. Esse módulo disponibiliza uma área de trabalho orientada para as etapas do processo de extração de conhecimento, e possui uma extensa gama de algoritmos. A grande maioria das técnicas consolidadas dentro da área de KDD estão implementadas no *Statistica*. Além de várias opções para redes neurais, estão disponíveis algoritmos para classificação e regressão. Existem também diversas opções para aprendizado não-supervisionado e geração de regras de associação. A ferramenta também se destaca nas opções para tratamento dos dados e na flexibilidade para comunicação com bases de dados nos mais diversos formatos (incluindo planilhas do Excel© e vários tipos de Bancos de Dados). Um ponto no qual o *Statistica* não é muito avançado é, porém, a geração de relatórios. Apesar de possuir mecanismos para geração de gráficos, a edição dos mesmos não é trivial. Maiores informações sobre o *Statistica* serão apresentadas nas seção 4.4.2.

A última ferramenta avaliada dentro do contexto desse projeto foi o *Weka*. Esse *software* nada mais é do que uma coleção de algoritmos de aprendizado por computador orientados para execução de tarefas de *Data Mining*. Os algoritmos podem ser aplicados diretamente a uma base de dados, ou chamados por códigos implementados na linguagem de programação Java. Além de funções para classificação, regressão, agrupamento e regras de associação, o *Weka* também possui um módulo para pré-processamento e visualização dos dados.

A escolha da ferramenta de *Data Mining* mais adequada é uma tarefa difícil, porém, alguns critérios podem auxiliar a mesma:

- Diversidade de algoritmos e técnicas de *Data Mining* disponíveis;
- Funcionalidades para tratamento dos dados;
- Mecanismos de integração com distintas bases de dados;
- Ferramentas para visualização e análise dos dados;
- Custo de aquisição e manutenção da ferramenta.

Nesse projeto, optou-se inicialmente por trabalhar com o *Intelligent Miner* e com o *Statistica*, pelo fato dessas duas ferramentas serem mais completas (maior gama de algoritmos e funcionalidades para tratamento e visualização dos dados) e também pelo fato de terem sido conseguidas licenças temporárias para utilização de ambas - uma vez que não se tratam de ferramentas gratuitas como as demais apresentadas[4]. Maiores considerações sobre as ferramentas de *Data Mining* serão apresentadas nas seções 4.2 e 4.4.2.

4.1.2 Outras Ferramentas Utilizadas

Além das ferramentas de *Data Mining* selecionadas, foi necessária a instalação e configuração de dois SGDBs: o *Oracle 9.2i* para restauração do *backup* da base de dados relacional do sistema de consolidação de dados do processo da Usina Siderúrgica, e o DB2 *Enterprise Edition* para viabilizar a utilização do *Intelligent Miner*. Foi também necessária a instalação do IBM *Migration Toolkit*, aplicativo de conversão de bases de dados para o formato DB2.

A instalação do *Oracle 9.2i* demostrou-se mais complexa do que a do DB2, sendo necessário recorrer-se a vários FAQs disponíveis na internet e também à ajuda de profissionais especialistas nessa ferramenta. Uma vez instalado o SGDB da *Oracle*, começou-se o processo de restauração da base de dados. Por se tratar de um repositório de tamanho considerável (240 *gigabytes*), dois desafios precisaram ser vencidos nessa etapa:

1. Transporte do *backup* da Usina para o computador no qual se iria trabalhar: Os arquivos de *backup*, mesmo compactados, ocupavam cerca de 30 *gigabytes* (6 arquivos de 5 *gigabytes* cada) e não era possível gravar cada arquivo em um DVD, pela limitação do sistema operacional do computador de processo de trabalhar com

[4] O iDA possui uma versão gratuita com algumas limitações operacionais. O *Weka* e o ROSETTA são gratuitos e possuem suas versões completas disponíveis na internet.

arquivos maiores que 2 *gigabytes*. Para solucionar esse problema foi utilizado um HD removível com capacidade máxima de 40 *gigabytes*.

2. Comandos para restauração do *backup*: Foram fornecidos pela Siderúrgica os comandos básicos para restauração do Banco de Dados. Porém, o micro disponível possuía um HD com apenas 160 *gigabytes*. Essa questão foi contornada com uma revisão dos comandos fornecidos para permitir que o banco fosse restaurado sem índices e, com isso, ocupasse apenas 42 *gigabytes*. Apesar dessa opção ter resolvido *a priori* o problema de espaço e viabilizar a restauração com sucesso da base de dados, a mesma resultou no problema da lentidão com a qual o *Oracle* passou a responder, uma vez que os índices [5] não foram gerados. Maiores detalhes sobre esse assunto serão apresentados na próxima seção.

O IBM *Migration Toolkit*, que está disponível no *site* do fabricante para *download* na sua versão 1.40, é uma ferramenta de fácil instalação. Sua utilização, entretanto, não é tão simples e o processo de migração da base de dados *Oracle* para o formato DB2 foi bastante penoso. Foram necessários vários ajustes nos parâmetros de configuração desse *software* para executar com sucesso a replicação para a nova plataforma. Problemas de falta de memória e conversão equivocada de tipos de dados numéricos só foram resolvidos através de contato direto realizado via e-mail com a IBM. Consultores responsáveis pelo suporte a essa ferramenta indicaram caminhos alternativos e mecanismos não triviais para solução dos problema encontrados. Apesar de tudo, foi possível deixar a base de dados operacional no ambiente DB2.

Outra ferramenta de *software* muito utilizada durante o processo de KDD foi o Microsoft Excel©. Várias funcionalidades desse *software* tiveram aplicação direta no trabalho para resolver problemas ou ajudar na análises dos dados. Entre essas pode-se destacar:

- Utilização de macros e mecanismos de manipulação de *strings* para geração de *scripts* na etapa de preparação dos dados;
- Criação de repositório de dados a partir da base de dados *Oracle* para facilitar a integração com o *Statistica* - ver seção 4.4.2;

[5] Os índices, em linguagem de Bancos de Dados, nada mais são do que estruturas de acesso auxiliares utilizadas para acelerar a recuperação de registros em resposta a determinadas condições de pesquisa (Navathe and Elmasri, 2000).

- Análise Estatística e Correlações iniciais dos dados: a maioria absoluta dos resultados da etapa de Análises Preliminares (seção 4.3) foi obtida através de recursos do Excel© como *Tabela Dinâmica* e *Estatística Descritiva*;

- Geração de gráficos - o Excel© foi utilizado para geração de todos os gráficos do trabalho devido à grande flexibilidade e facilidade de ajustes de formatação. Mesmo os gráficos gerados pelo *Statistica* foram refeitos nesta ferramenta da Microsoft para melhorar a sua apresentação.

Como se pode observar, uma série de ferramentas de *software* foram necessárias. Para cada uma delas, além do esforço inicial de instalação, foi necessária uma etapa de familiarização (em maior ou menor grau), o que gerou um esforço adicional considerável. Em alguns casos, como o *Statistica* e o Microsoft Excel©, a documentação existente no próprio *software* foi suficiente. Em outros, como por exemplo o SGDB da *Oracle*, a internet foi grande fonte de informações, além da ajuda direta de profissionais especialistas.

4.2 Preparação e Limpeza da Base de Dados

A base de dados restaurada continha informações referentes a aproximadamente seis meses de dados de produção da área de Laminação de Tiras a Quente. Com o objetivo de se reduzir o número de informações a serem tratadas inicialmente, foi acordado com os analistas da Usina Siderúrgica que apenas os dados relativos ao mês de outubro de 2005 seriam utilizados. Esse mês foi especialmente escolhido pelo fato da coleta de dados do PIMS não ter apresentado nenhuma interrupção no período. A partir dessa definição, a etapa inicial de limpeza do banco consistiu na remoção das informações que não eram relativas ao período de 01/10/2005 a 31/10/2005. Mais precisamente, esta restrição foi aplicada ao atributo *Data_inicial* da tabela **TAB_ARQUIVOS**. A figura 4.1 apresenta o modelo de entidades e relacionamentos (MER) da base de dados.

A tabela **TAB_RESULTADOS** é a principal tabela da base de dados e armazena os valores medidos para cada variável armazenada pelo sistema. A **TAB_ARQUIVOS** faz o agrupamento temporal dessas medições, por etapa do processo. São registradas nessa última tabela os instantes de tempo de início e fim de cada etapa. A tabela **TAB_BOBINAS**, por sua vez, faz um agrupamento dos registros de cada bobina produzida, além de conter o identificador único de produção (**BobinaID**). As demais tabelas

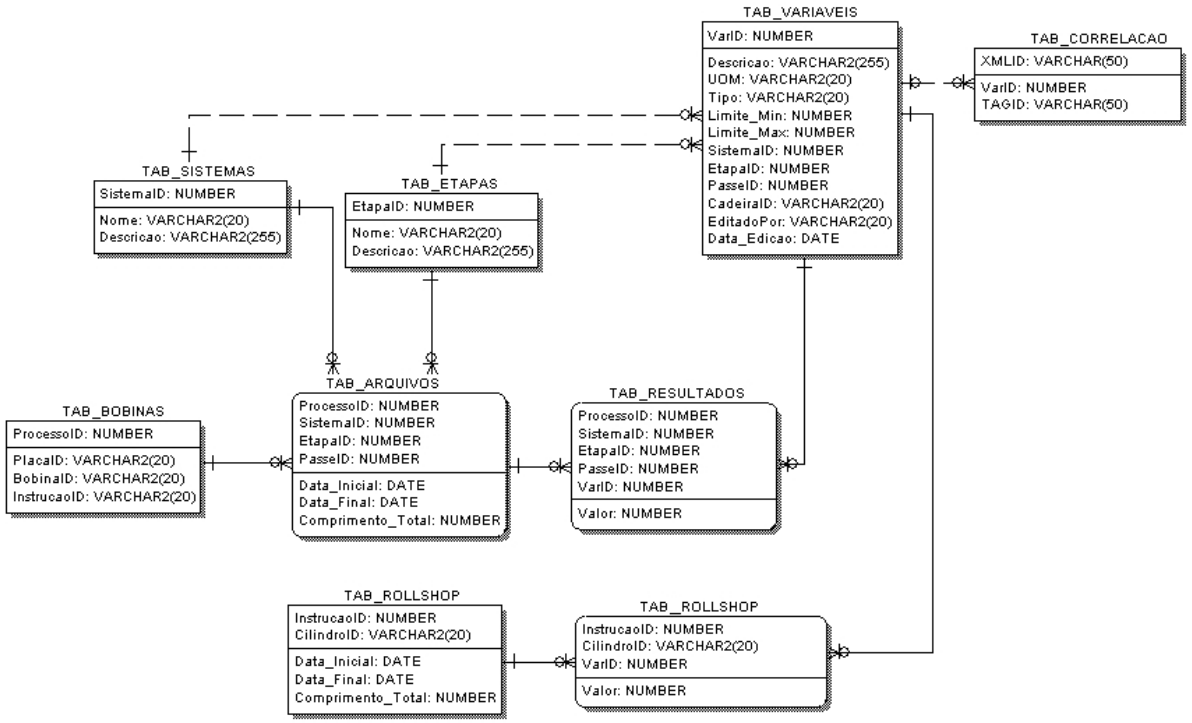

Figura 4.1: MER da base de dados do Sistema de Consolidação de Dados do Processo

do sistema servem apenas de cadastro para as principais variáveis do sistema (ex.: etapas, variáveis, etc).

Como a tabela **TAB_RESULTADOS** possuía mais de 700 milhões de registros e o banco foi restaurado sem índices, a execução de *scripts* diretos para remover registros do Banco de Dados não mostrou-se uma alternativa viável. Uma tentativa nesse sentido foi realizada, porém o *script* chegou a rodar por mais de 24 horas sem um resultado final e foi abortado. Um estudo sobre otimização de comandos SQL [6] foi realizado e gerou a seguinte estratégia para realizar a limpeza, manipulando as principais tabelas do Banco de Dados:

- Criação de uma tabela **TAB_RESULTADOS_OUT** a partir de uma junção com a tabela **TAB_ARQUIVOS** em um novo *tablespace* [7] contendo apenas os dados de outubro de 2005. Este procedimento demorou cerca 12 horas para ser completado;

[6] SQL (do Inglês *Structured Query Language*) é um padrão de linguagem utilizado para se criar e manter Bancos de Dados relacionais através de comandos de texto. A realização de consultas é uma das principais aplicações do SQL.

[7] *tablespaces* nada mais são do que unidades lógicas de armazenamento nas quais um Banco de Dados é dividido. No SGDB da *Oracle* um *tablespace* pode estar fisicamente dividido ainda em um ou mais arquivos de dados.

- Eliminação da tabela **TAB_RESULTADOS** e seu respectivo *tablespace*;
- Exclusão dos registros **TAB_ARQUIVOS** cujos atributos *Data_inicial* não continham valores entre 01/10/2005 a 31/10/2005;
- Criação de uma tabela **TAB_BOBINAS_OUT** contendo apenas os registros da tabela **TAB_BOBINAS** que possuíam correspondência com os registros remanescentes da tabela **TAB_ARQUIVOS**;
- Eliminação da tabela **TAB_BOBINAS**.

As demais tabelas da base de dados não sofreram alterações e, ao final deste procedimento, a base de dados reduziu-se de 42 para aproximadamente 8 *gigabytes*. Apesar disso, devido ao elevado número de registros remanescentes na tabela principal do Banco de Dados (cerca de 115 milhões de linhas), uma lentidão considerável ainda era notada na execução de consultas. Para tentar contornar esse problema foram criados então os índices originais, porém sobre a base de dados reduzida. Após esta etapa, com a execução dos *scripts* de criação dos índices durando mais de 4 horas, o tamanho total da base passou para cerca de 39 *gigabytes*, porém o acesso ao Banco de Dados ficou bem mais rápido (uma consulta ao número total de registros da tabela **TAB_RESULTADOS**, que durava cerca de 10 minutos, passou a ser executada em cerca de 15 segundos).

O objetivo principal da etapa de preparação e limpeza é gerar uma estrutura de dados com a qual as ferramentas de *Data Mining* possam lidar. No caso das ferramentas selecionadas dentro desse projeto, ambas trabalham a partir de uma *flat table* única. Para gerar essa tabela foi necessária a realização de várias manipulações dentro do Banco de Dados relacional e, para permitir esse trabalho, algumas informações relevantes foram levantadas sobre os dados:

- A base de dados contém informações sobre bobinas produzidas na área de Laminação de Tiras a Quente. No mês de outubro de 2005 o sistema possui registros de 8629 bobinas produzidas;
- Para cada bobina, além do seu identificador, são armazenadas as informações de identificação da placa e da instrução de produção que originou aquela bobina;
- Para cada bobina, por fase de processo, é registrada a data de início e fim da fase, bem como o comprimento da bobina nesta etapa. Uma fase é definida por uma

etapa, uma cadeira e um passe, sendo que a definição da cadeira só tem sentido para a etapa correspondente ao Trem Acabador (sendo igual a zero para as demais) e a definição de passe apenas para a etapa correspondente ao Desbastador;

- Para cada fase, várias variáveis de processo têm seu valor armazenado. Existem no sistema um total de 6995 variáveis cadastradas, porém apenas 6541 possuem medidas para o mês de outubro de 2005.

- Cada variável possue um identificador único que defini o sistema, a etapa, o passe e a cadeira correspondente.

A tabelas de 4.1 a 4.3 mostram os totais de variáveis medidas para as etapas, sistemas e cadeiras existentes no banco como um todo e apenas na base final com os dados de outubro de 2005.

Descrição da Etapa	**Código**	**# Variáveis**	
		Total	Out 05
Desbastador	1	3092	2854
Trem Acabador	2	3131	3013
Bobinadeiras	3	42	35
Oficina de Cilindros	5	17	0
Resfriamento	6	84	51
Forno de Reaquecimento	7	81	59
Inspeção	8	10	10
Análise Química	9	165	165
Outras Etapas	25	373	354

Tabela 4.1: Número de Variáveis por Etapas Cadastradas no Sistema de Consolidação de Dados do Processo.

Para tratar o problema do Erro de Força de Laminação, foi construída uma *flat table* com informações sobre cada uma das 8629 bobinas produzidas que constam na base de dados. A primeira opção seria montar essa tabela considerando as 6541 variáveis por bobina como colunas, porém as ferramentas de *Data Mining* muito provavelmente não seriam capazes de lidar com um número tão grande de atributos. Mesmo que as ferramentas suportassem tal estrutura de dados, que teria assim mais de 48 milhões de registros, provavelmente questões como falta de memória e elevado tempo de processamento iriam dificultar muito o trabalho. Além disso, a análise dos resultados seria bem mais complexa pela quantidade de dados a serem tratados. Portanto, foi realizada uma reunião com os

Descrição do Sistema	Código	# Variáveis	
		Total	Out 05
Controle de Laminação	1	6443	6067
Controle do Forno	2	81	59
Controle de Resfriamento	3	84	51
Classificação de Defeitos	4	10	10
Gerenciamento da oficina de cilindros	5	17	0
Concentrador de Dados de Controle (Nível 1)	7	0	0
Sistemas Auxiliares	6	360	354

Tabela 4.2: Número de Variáveis por Sistemas Cadastrados no Sistema de Consolidação de Dados do Processo.

Cadeira	# Variáveis	
	Total	Out 05
1	525	505
2	524	505
3	524	505
4	524	505
5	516	497
6	518	496
Outras	3864	3528

Tabela 4.3: Número de Variáveis por Cadeiras Cadastradas no Sistema de Consolidação de Dados do Processo.

especialistas responsáveis pelos sistemas geradores dos dados da área de Laminação de Tiras a Quente para a seleção manual de variáveis. A partir de um trabalho interno previamente realizado sobre o assunto, os analistas da Usina Siderúrgica selecionaram cerca 200 variáveis que, de alguma forma, poderiam estar relacionadas com o Erro de Força de Laminação.

A partir dessa seleção, foi gerado uma *flat table* com a estrutura descrita na tabela 4.4. Os *scripts* de criação dessa estrutura no Banco de Dados foram escritos manualmente, porém os comandos para popular as colunas com os valores medidos para cada variável do processo foram gerados no Excel© através de macros e funções de manipulação de *strings*. O processo de execução desses *scripts* foi manual e levou bastante tempo para ser executado. Mesmo com os índices tendo sido gerados, para operações de inserção de dados o tempo de resposta do Banco de Dados estava muito ruim.

Além das variáveis selecionadas manualmente, várias outras foram criadas a partir dos dados existentes. Entre elas pode-se citar:

- Erro de Força (Real e Calculado) para cada uma das 6 primeiras cadeiras de lami-

Código Bobina	Variável 1	Variável 2	...	Variável N
bobina bq0001	valor var1 bq0001	valor var2 bq0001	...	valor varN bq0001
bobina bq0002	valor var1 bq0002	valor var2 bq0002	...	valor varN bq0002
bobina bq0003	valor var1 bq0003	valor var2 bq0003	...	valor varN bq0003
.	.	.	.	.
.	.	.	.	.
.	.	.	.	.
bobina bqM-1	valor var1 bqM-1	valor var2 bqM-1	...	valor varN bqM-1
bobina bqM	valor var1 bqM	valor var2 bqM	...	valor varN bqM

Tabela 4.4: Estrutura da *flat table* gerada, na qual *bqXXXX* indica o código das bobinas.

nação - a base de dados não possuía diretamente uma variável com o erro, e o mesmo teve de ser calculado de acordo com a equação descrita na seção 3.3. Além do erro real, foram também geradas variáveis para o erro calculado, uma vez que na base de dados original estava disponível a variável "Força de Laminação Calculada", que corresponde ao valor simulado pelo modelo;

- Erro de Resistência à Deformação (Real e Calculado) - a resistência à deformação da placa que está sendo laminada é utilizada pelo modelo de força como um cálculo intermediário, e sua análise também foi indicada pelos analistas da Usina. Essas novas variáveis foram geradas através da subtração da medida de "*Setup* de Resistência à Deformação" pela "Resistência à Deformação Real" (valor medido do processo) e pela "Resistência à Deformação Calculada" (valor estimado por um modelo matemático);

- Atuação do Operador - através da comparação entre as medidas de "Distribuição de Força Calculada" e "Distribuição de Força Medida", foi gerada uma variável categórica indicando se o operador atuou no processo de laminação ou não. Essa variável pode assumir apenas os valores "sim" ou "não";

- Categorização das variáveis de erro geradas - para cada uma dessas variáveis foi gerada uma nova, categórica, assumindo os valores "OK", "Indefinido" ou "Ruim", de acordo com as faixas estabelecidas pelos analistas da Usina e descritas na seção 3.2.

Após essas inserções, a *flat table* gerada resultou em 8629 linhas por 240 colunas. Para finalizar a etapa de preparação e limpeza dos dados, foi realizada uma varredura em busca de registros em branco e também de variáveis redundantes. Essas operações

foram aceleradas através de recursos específicos da ferramenta *Statistica*. Os registros que estavam com valor zero para variáveis relevantes como força de laminação (valor medido e *setup*) foram excluídos. Também foram eliminadas as variáveis cuja maioria absoluta dos registros estavam nulos. Algumas variáveis foram automaticamente identificadas como redundantes (mesmos valores para todos os registros) e também foram eliminadas. Após todas essas operações, a *flat table* final ficou reduzida a 8421 linhas por 205 colunas.

Uma observação importante com relação à etapa de limpeza e preparação dos dados é que todas as operações realizadas foram validadas pelos especialistas nos sistemas geradores dos dados da área de Laminação de Tiras a Quente. Com isso acredita-se que nenhum dos registros ou variáveis eliminados possuíam qualquer relevância para a análise do erro de força, e poderiam até mesmo dificultar a execução dos algoritmos de *Data Mining*.

4.3 Análises Preliminares ("Get to know the data")

Conforme descrito na seção 2.2.3, é importante imergir nos dados antes da aplicação de algoritmos de *Data Mining* com o objetivo de buscar padrões e informações novas e potencialmente úteis dentro da *flat table* gerada.

Na seção anterior (4.2) foram descritos os números de variáveis de processo medidas para cada etapa, sistema e cadeira de laminação. Observando-se a tabela 4.1, nota-se que a maioria das variáveis armazenadas pelo sistema de consolidação de dados do processo no mês de outubro de 2006 está concentrada nas etapas correspondentes ao Desbastador (43,6%) e ao Trem Acabador (46,1%). A tabela 4.2 mostra, por sua vez, que a maioria absoluta das variáveis é gerada pelo Sistema de Controle de Laminação, além de confirmar a afirmação de que a massa de dados disponibilizada continha apenas informações do Banco de Dados relacional, uma vez que nenhuma variável do Sistema Concentrador de Dados de Controle foi identificada. Analisando-se a *flat table* final, observa-se nas tabelas de 4.5 a 4.7 a mesma descrição feita pelas tabelas de 4.1 a 4.3, porém apenas para as 150 variáveis selecionadas pelos analistas da Usina levando em consideração o foco no problema de força de Laminação.

A tabela 4.5 mostra que a maioria das variáveis selecionadas é relativa à etapa do Trem Acabador. Além dessa etapa, foram consideradas importantes apenas variáveis correspondentes à Análise Química do Material e outras de etapas de menor importância.

Descrição da Etapa	Código	# Variáveis
Trem Acabador	2	107
Análise Química	9	25
Outras Etapas	11	18

Tabela 4.5: Número de Variáveis por Etapas - *flat table* final

Descrição da Sistema	Código	# Variáveis
Controle de Laminação	1	132
Sistemas Auxiliares	6	18

Tabela 4.6: Número de Variáveis por Sistemas - *flat table* final

Una análise direta da tabela 4.7 indica que as cadeiras 1 e 6 tiveram mais variáveis selecionadas que as cadeiras de 2 a 5. Porém, deve-se levar em consideração que, no processo de limpeza dos dados, algumas medidas foram detectadas como redundantes (mesma medida para todas as 6 primeiras cadeiras) e, com isso, optou-se por deixar na base de dados apenas uma variável, sendo ora a da cadeira 1, ora a da cadeira 6.

Dentre as variáveis criadas a partir das 150 originais da *flat table*, talvez as mais importantes sejam as que descrevem o erro de força. Como o erro de força foi calculado para cada uma das cadeiras e depois categorizado dentro das faixas descritas na seção 3.2, foram geradas 12 variáveis relacionadas a essa grandeza: 6 numéricas e 6 categóricas. A tabela 4.8 mostra a análise estatística para as variáveis numéricas, e a figura 4.2 mostra a distribuição do erro nas faixas **OK**, **Indefinido** e **Ruim** para cada cadeira, correspondendo às variáveis categóricas.

A tabela 4.8 mostra que as cadeiras 4 e 5 apresentam as maiores médias para o valor absoluto do erro de força. Além disso, observa-se também que essas cadeiras possuem uma maior variância (mais do que o dobro das cadeiras de 1 a 3) e valores máximos, além de apresentarem também maiores desvios padrões e medianas. Os gráficos da figura 4.2

Cadeira	# Variáveis
1	24
2	16
3	16
4	16
5	16
6	19
Outras	43

Tabela 4.7: Número de Variáveis por Cadeira - *flat table* final

Análise - Cadeira	1	2	3	4	5	6
Média	5,33	4,52	5,90	9,53	10,29	7,28
Erro padrão	0,06	0,05	0,06	0,09	0,09	0,08
Mediana	3,60	3,44	4,39	7,35	8,79	5,20
Modo	2,02	2,34	3,22	9,76	0,44	4,05
Desvio padrão	5,64	4,22	5,53	8,24	8,20	7,22
Variância	31,81	17,78	30,58	67,97	67,29	52,16
Curtose	13,18	32,01	12,61	3,67	3,62	8,00
Assimetria	2,71	3,18	2,31	1,47	1,42	2,31
Intervalo	66,58	92,00	92,37	104,26	93,05	80,56
Mínimo	0,00	0,00	0,00	0,00	0,00	0,00
Máximo	66,58	92,01	92,37	104,27	93,05	80,56
Soma	44815,76	37987,75	49621,22	80191,03	86586,54	61278,39
Total de Registros	8412	8412	8412	8412	8412	8412

Tabela 4.8: Análise Estatística Descritiva do Erro de Força

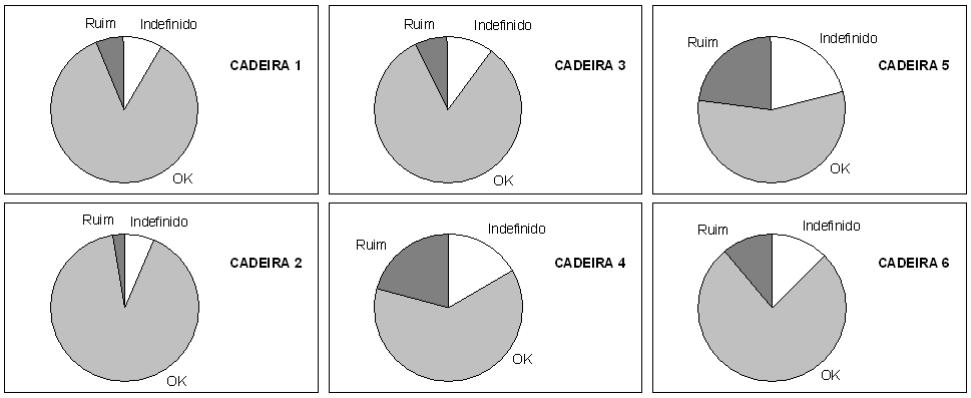

Figura 4.2: Distribuição do Erro de Força por cadeira de Laminação

mostram claramente esse comportamento do erro de força para as cadeiras 4 e 5 em relação às demais. Tais resultados confirmam a percepção dos analistas responsáveis pelos dados de que uma maior ênfase do trabalho deve ser dada a essas cadeiras, uma vez que as mesmas são as mais problemáticas.

Outra variável relevante contida na *flat table* final é a que indica a família do aço (SGF, do inglês *Steel Grade Family*) que foi laminado. A figura 4.3 apresenta os principais tipos de aço existentes e a sua distribuição pelas bobinas produzidas no mês de outubro de 2005. Observa-se que o SGF 3 é responsável por mais da metade dos registros (51%). A segunda família mais representativa é a 9 com 1498 bobinas, ou 18% do total. Além dessas, nota-se também quantidades significativas de registros das familias 2, 5, 6 e 10. As demais juntas correspondem apenas a 6% dos dados.

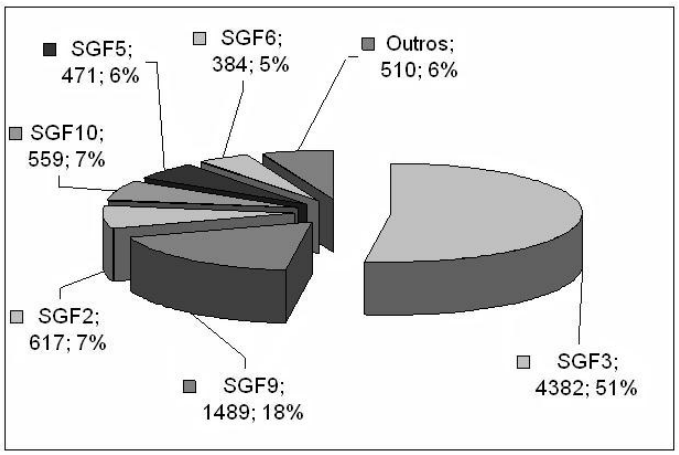

Figura 4.3: Distribuição das bobinas produzidas por Família de Aço, Número de Registros, Valor Percentual

Uma análise interessante a ser feita é a da distribuição do erro de força por cadeira e ao mesmo tempo por tipo de aço, uma vez que as características químicas do material podem ter alguma relação com essa grandeza. As figuras 4.4 e 4.5 mostram essa análise para as cadeiras 4 e 5, levando em consideração as famílias de aço mais significativas da base de dados. A análise completa para todas as cadeiras pode ser vista no Apêndice A.

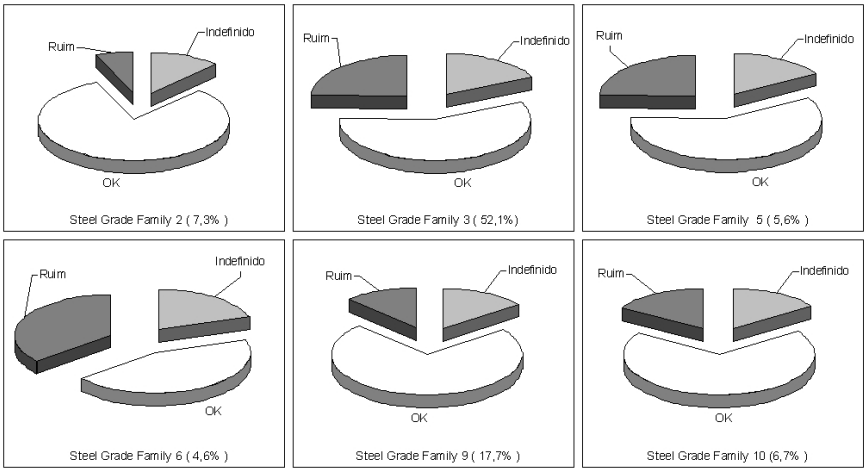

Figura 4.4: Distribuição do Erro de Força por Família de Aço - Cadeira 4

Observa-se, assim, que a família de aço 6 é a que apresenta maior faixa de erro de força considerada **Ruim**. Isso ocorre na cadeira 4, sendo ainda mais acentuado na cadeira 5. A família 2, por sua vez, apresenta um comportamento inverso com a faixa de erro **OK**, cobrindo a maioria absoluta dos registros para as duas cadeiras consideradas nesses gráficos. As demais famílias apresentam um comportamento de certa forma similar dentro de cada cadeira, com uma leve diminuição da faixa **OK** na cadeira 5.

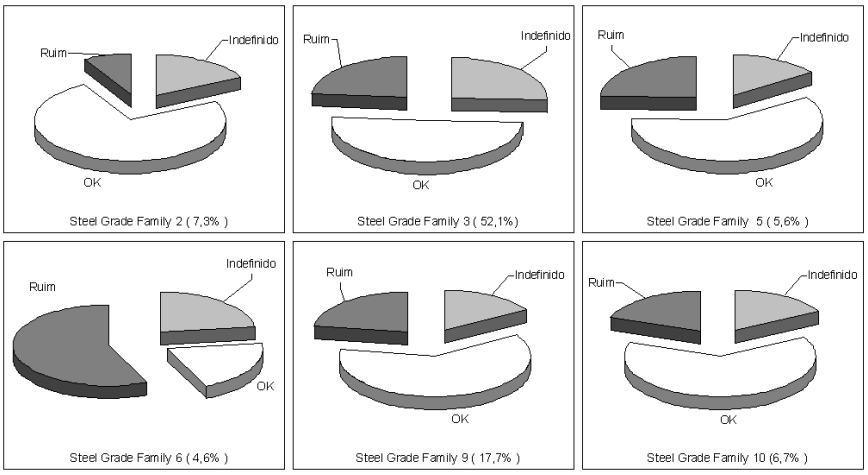

Figura 4.5: Distribuição do Erro de Força por Família de Aço - Cadeira 5

Ainda dentro das análises preliminares do processo de KDD, a figura 4.6 mostra a distribuição de uma medida que despertou muita curiosidade desde o início do projeto: a atuação do operador. Observa-se que a intervenção do operador no processo, alterando manualmente a distribuição de forças dentro de uma determinada cadeira de laminação, ocorre de certa forma regularmente entre as cadeiras (atuação em cerca de um terço dos registros). Exceções ficam apenas pela cadeira 1, com o operador atuando em apenas 24% das bobinas produzidas, e pela cadeira 4, com 41% de registros.

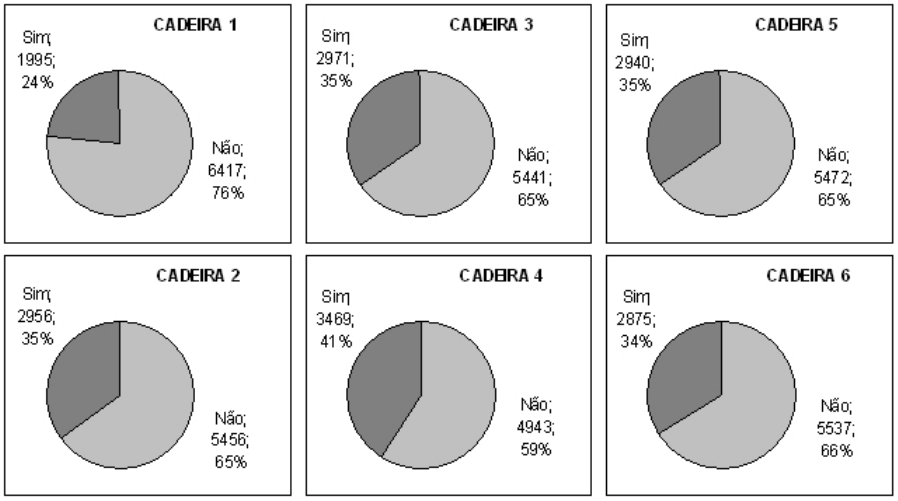

Figura 4.6: Atuação do Operador por Cadeira de Laminação

Através de agrupamentos realizados entre as variáveis de "erro de força" e "atuação do operador"[8], observou-se que a média do erro de força aumentava em cerca de 100%

[8]utilização do recurso de *Pivot Table* do Microsoft Excel© que permite de forma bem simples a

quando o operador atuava no processo nas cadeiras 4 e 5. Essa mesma análise mostrou, entretanto, que isso não ocorria para as demais cadeiras, sendo a intervenção do operador irrelevante no valor médio do erro de força. Para avaliar melhor essa questão, foram traçados gráficos do erro de força para a família de aço 3, para as cadeiras 3 e 4 (figuras 4.7 e 4.8).

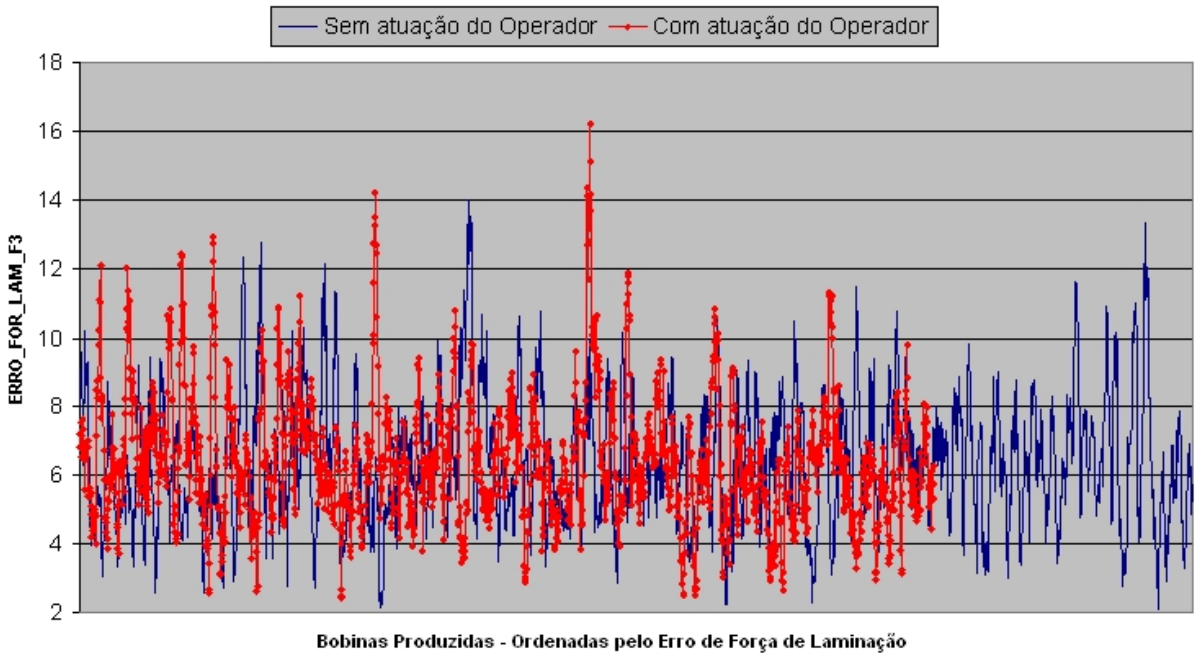

Figura 4.7: Influência da atuação do Operador no Erro de Força - Cadeira 3, Família de Aço 3

Analisando-se esses gráficos, fica clara a influência negativa do operador na cadeira 4, com a média do erro de força aumentando de forma considerável.

É importante ressaltar que a fase de análises preliminares foi realizada mais de uma vez, comprovando a característica iterativa do processo de KDD. No início da aplicação das técnicas de *Data Mining*, foi necessária a construção de novos gráficos, além de realização de mais análises estatísticas, buscando assim obter-se um melhor direcionamento para trabalho. Outra situação relevante ocorreu na apresentação dos resultados iniciais para os analistas responsáveis pelos dados, na qual identificou-se que a variável selecionada para indicar a familia do aço não era a mais adequada. Isso levou a uma repetição de todo o processo, a partir da etapa de análises preliminares, que teve de ser novamente executada com o foco na nova variável para classificação do tipo de aço.

correlação entre variáveis com o cálculo de médias, somatórios e outras operações para consolidação de dados.

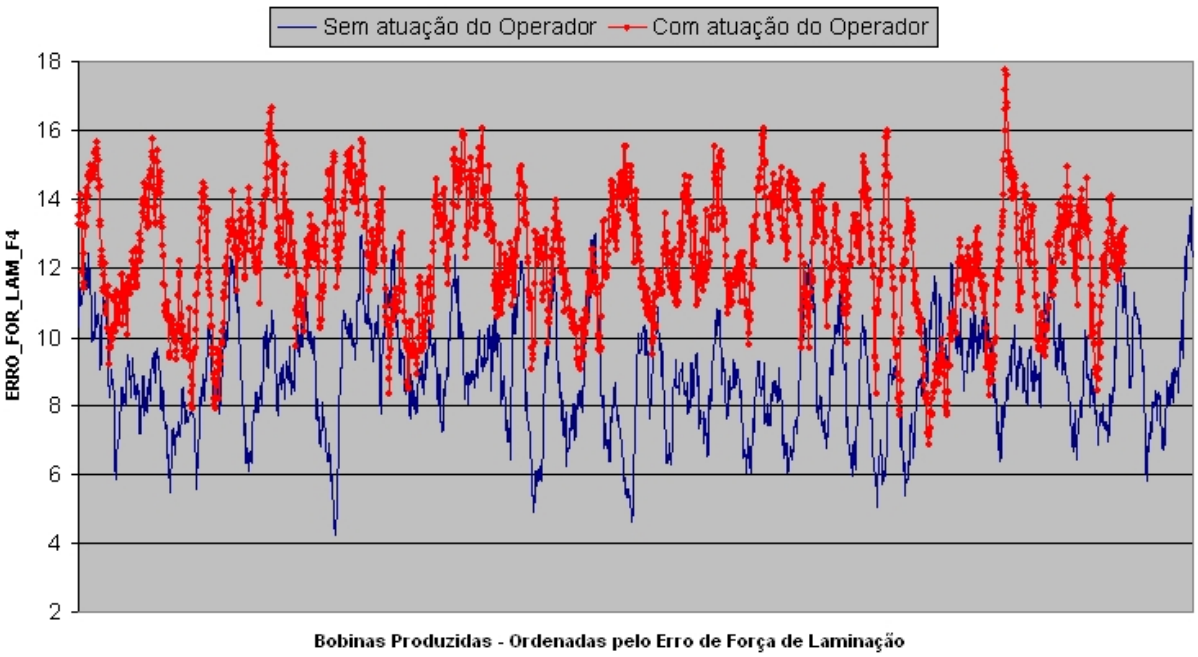

Figura 4.8: Influência da atuação do Operador no Erro de Força - Cadeira 4, Família de Aço 3

4.4 Data Mining

4.4.1 Seleção da Estratégia de *Data Mining*

A etapa de *Data Mining* deve sempre começar com a seleção de uma estratégia, conforme descrito na seção 2.2.5. Para abordar o problema do erro de força de laminação, optou-se por uma estratégia de aprendizado supervisionado: a classificação. Os critérios que suportaram essa escolha foram os seguintes:

- Existência, dentro da *flat table* gerada, de uma variável para "guiar" o processo, no caso o erro de força de laminação;
- O fato da análise a ser feita sobre o erro de força estar diretamente relacionada à categorização realizada sobre essa variável[9].

Dentre os vários algoritmos disponíveis para aplicação da estratégia de classificação, o CART (Breiman *et al.*, 1984) foi o selecionado. A opção por esse algoritmo está associada ao fato do mesmo ser um dos mais consolidados dentro da comunidade

[9]Se a análise fosse focada na variável contínua do erro de força, a estratégia seria de predição e não classificação.

de KDD para esse tipo de tarefa[10], incluindo aplicações na área específica das indústrias de processos (Mastrangelo, 1998). Outro fator determinante na escolha do CART foi a disponibilidade do mesmo dentro da ferramenta *Statistica* e, além disto, pelo fato do algoritmo gerar como saída uma lista de importância dos atributos na classificação, além da árvore de decisão propriamente dita. Maiores detalhes sobre o CART são apresentados na seção 4.4.3.

4.4.2 Detalhamento da Ferramenta Utilizada

A esta altura, apesar do *software* da IBM (o *Intelligent Miner*) também possuir uma implementação de árvore de decisões, optou-se pela utilização exclusiva do *Statistica* como ferramenta para suportar a etapa de *Data Mining*. Além de todas as vantagens já descritas na seção 4.1.1, a aplicação da Statsoft proporcionou um maior grau de aprendizado e domínio das funcionalidades, devido à sua fácil utilização e aplicação.

Uma das características mais interessantes do *Statistica* é a área de trabalho orientada ao processo de KDD. A figura 4.9 mostra um exemplo de aplicação desenhada dentro desse ambiente. Observa-se a divisão em 4 áreas distintas, a saber:

- *Data Acquisition* - Funções de interface com Bancos de Dados;
- *Data Preparation, Cleaning, Transformation* - Funções de manipulação dos dados, como por exemplo, remoção de variáveis, transposição ou divisão dos dados, ordenação, etc;
- *Data Analysis, Modeling, Classification, Forecasting* - Algoritmos de *Data Mining*;
- *Reports* - Exibição de gráficos e relatórios com resultados.

Para desenhar uma aplicação no *Statistica*, basta selecionar as funções a serem utilizadas em cada etapa, configurá-las, através de duplo clique em cima da caixa correspondente, e conectá-las através de comandos diretos executados pelo próprio *mouse*. A configuração original de praticamente todos os parâmetros das funções disponibilizadas já atende à maioria dos casos de utilização (ex.: o número máximo de nós que uma ár-

[10] O CART juntamente com o C4.5 (Quinlan, 1993) são reconhecidos, dentro da comunidade de KDD, como os algoritmos de melhor desempenho para executar tarefas de classificação, sendo os mais utilizados dentre as estratégias de aprendizado supervisionado.

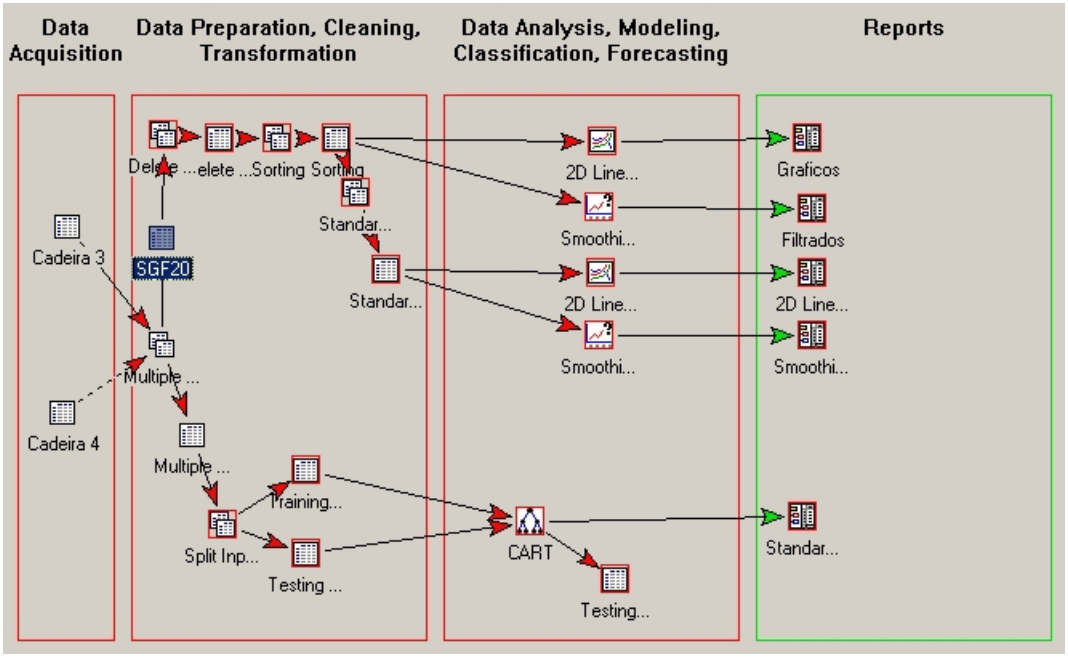

Figura 4.9: Ambiente de Trabalho do Módulo de *Data Mining* do *Statistica*

vore de decisão vai possuir é um parâmetro do algoritmo CART). Porém, para melhor desempenho e maior flexibilidade, ajustes podem ser realizados quando necessário.

Uma vez configurado, a execução do fluxo de trabalho pode ser feita de forma total ou parcial (a partir de um ponto qualquer dentro do fluxo desenhado), utilizando um resultado obtido anteriormente. Além de gerar gráficos, é possível exportar os resultados obtidos para arquivos de diversos formatos, inclusive o do Excel©.

4.4.3 Algoritmo CART

O algoritmo CART é uma das mais conhecidas implementações das Árvores de Decisão. Tais estruturas de dados, por sua vez, podem ser definidas como modelos estatísticos que utilizam um treinamento supervisionado para classificação e previsão de dados.

As Árvores de Decisão podem ser divididas basicamente em 3 componentes:

- Nós-principais ou atributos: representam um mapeamento das variáveis (ou colunas) da *flat table* utilizada como conjunto de treinamento. A cada nó-principal está associado um subconjunto de dados, sendo que ao primeiro nó da árvore corresponde toda a massa de dados do conjunto de treinamento;

- Arcos: provenientes dos nós-principais, esses componentes recebem os valores pos-

síveis para um determinado atributo;

- Nós-folha: representam as diferentes classes do conjunto de treinamento.

Na figura 4.10 observa-se um exemplo de uma árvore para suporte à tomada de decisão de conceder ou não um empréstimo por parte de uma instituição bancária. A estrutura leva em consideração os atributos montante, salário e o saldo existente na conta-corrente do candidato a empréstimo. Cada um desses atributos pode assumir apenas os valores "alto" ou "baixo". Os nós-folhas da árvore indicam a existência de apenas duas classes: "sim" ou "não", para liberação do empréstimo.

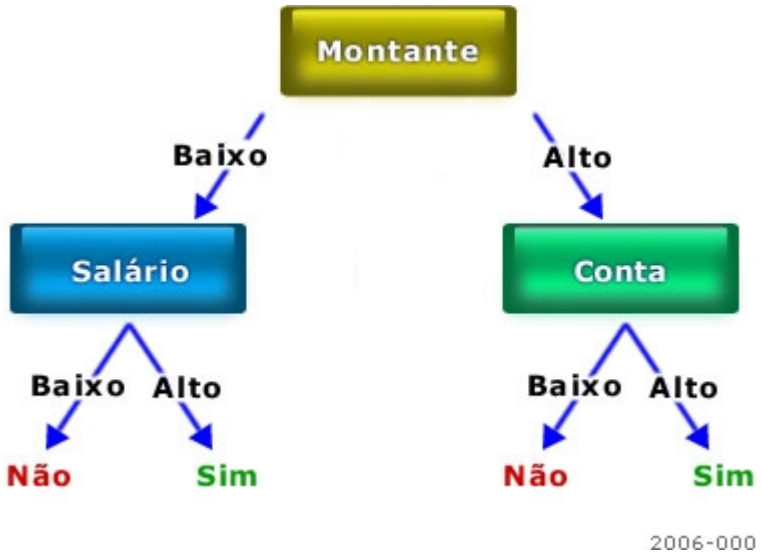

Figura 4.10: Exemplo de Árvore de Decisão - Adaptado de (Garcia, 2000)

A metodologia do modelo CART é tecnicamente conhecida como partição recursiva binária. O processo é binário porque os nós-pais são sempre divididos em apenas dois nós-filhos, e recursivo porque pode ser repetido tratando cada nó-filho como um nó-pai.

O algoritmo CART pode ser resumido nos seguintes passos (Yohannes and Webb, 1999):

1. Dado um determinado nó, o algoritmo aplica todas as possíveis regras para se dividir o conjunto de dados associados ao mesmo. Cada valor que cada variável assume dentro da massa de dados é uma possível regra. Por exemplo, se uma massa de dados possui duas variáveis, uma contínua X assumindo os valores 0.1, 0.7 e 3.4, e outra categórica Y assumindo os valores *baixo*, *médio* e *alto*, as possíveis regras para dividir um nó seriam:

- $X \geq 0.1$?
- $X \geq 0.7$?
- $X \geq 3.4$?
- *Y é baixo?*
- *Y é médio?*
- *Y é alto?*

Para cada possível regra, a amostra de dados é divida em duas, gerando dois nós-filhos. Os casos que respondam "sim" para uma regra vão para o nó-filho da esquerda, e os casos que respondam "não" vão para o nó da direita.

2. O CART aplica então um critério de partição para cada nó-filho gerado por cada uma das possíveis regras. O critério de partição utilizado pelo CART é índice Gini[11]. O grau Gini de impureza de um determinado nó t é definido como 1 - FI, onde FI (função de impureza) é calculado por

$$FI = -\sum p^2(j|t) \; para \; j = 1, 2.., k \tag{4.1}$$

onde p é a probabilidade de ocorrência de cada classe j do modelo de classificação no subconjunto de dados associado ao nó t em questão. Tão melhor será uma regra de divisão quanto maior for a redução de impureza associada à ela. Dado um nó t, o critério de partição gerado por uma regra s é dado por

$$\Delta(s,t) = i(t) - p_E[i(t_E)] - p_D[i(t_D)], \tag{4.2}$$

onde p_E é a proporção de casos associados ao nó t que vão para o nó-filho à esquerda, p_E é a proporção de casos que vão para o nó-filho à direita, i(t_E) é a impureza associada ao nó-filho à esquerda, e i(t_D) a impureza associada ao nó-filho à direita.

3. O algoritmo seleciona então a regra que gerou a maior redução na impureza da árvore.

[11] O índice Gini, desenvolvido por Conrado Gini em 1912, mede o grau de heterogeneidade dos dados. Logo, pode ser utilizado para medir a impureza de um nó de uma árvore de decisão. Quando este índice é igual a zero, o nó é puro. Por outro lado, quando ele se aproxima do valor um, o nó é impuro (aumenta o número de classes uniformemente distribuídas neste nó)(Rätsch *et al.*, 2001).

4. O próximo passo é então dividir o conjunto de dados em dois, a partir da regra selecionada.

5. Cada nó filho é então classificado dentro de uma das possíveis classes do conjunto de treinamento. Essa classificação é feita pela simples análise de distribuição dos registros que foram separados para um determinado nó-filho. Por exemplo, supondo que o conjunto de dados possa ser classificado nas classes A, B e C, se uma divisão CART gerou um nó-filho esquerdo com uma maior quantidade de registros da classe A, esse nó será atribuído como classe A. Caso também existam registros das classes B e C, esse nó ainda não será 100% puro.

6. O CART continua então dividindo a árvore, aplicando os passos acima de forma recursiva aos nós-filhos gerados até que só existam nós-filhos 100% puros, ou com um grau de pureza considerado aceitável. Outro critério de parada que pode ser difinido é o do número máximo de nós que a árvore pode ter.

Durante ou após a geração da árvore, técnicas de "podagem" (*pruning*) podem ser aplicadas com o intuito de estancar o crescimento da árvore ou diminuir o seu tamanho final. Existem várias formas de se realizar a podagem, porém a mais simples e eficaz é a que verifica se o erro de classificação de um determinado nó é menor do que a soma dos erros dos nós-filhos. Quando isso ocorre, os nós-filhos são descartados e o nó em questão se torna uma folha da árvore.

A particularidade das árvores CART serem estruturas binárias permite um tratamento mais simplificado em relação a outras estruturas de dados mais complexas. Além disso, podemos apresentar outras vantagens desse algoritmo:

- Possui, junto com o C4.5, os melhores tempos de resposta médios para geração de árvores de decisão (Garcia, 2004);

- É flexível para trabalhar tanto com atributos numéricos, quanto com atributos categóricos, podendo devido a isso tratar de problemas de regressão, além dos de classificação (Flores, 2005);

- O algoritmo lida bem com pontos muito fora da distribuição padrão dos dados (*outliers*), normalmente os separando em nós isolados (Lewis, 2000).

4.4.4 Metodologia Aplicada

Apesar da *flat table* gerada possuir uma variável categórica para o erro de força de cada uma das 6 primeiras cadeiras da área de Laminação de Tiras a Quente, optou-se por analisar inicialmente apenas o comportamento das cadeiras 3, 4 e 5. Essa decisão foi tomada em conjunto com os analistas da Siderúrgica, que declararam maior interesse pela cadeira 4; com isso, as cadeiras 3 e 5 foram escolhidas por serem respectivamente a anterior e a posterior. Além disso, dada a diferença de comportamento do erro de força por família de aço, foi identificada a necessidade de análise sob a esfera dessa variável. Porém, para focar o trabalho nos registros mais significativos, definiu-se selecionar apenas as famílias 2, 3, 9 e 10 uma vez que essas representam cerca de 86% das bobinas produzidas.

O algoritmo CART foi executado para cada uma das combinações possíveis dentre as cadeiras e famílias de aço selecionadas. A figura 4.11 mostra um exemplo de árvore de decisão gerada pelo *Statistica*. Essa árvore foi gerada com o algoritmo CART configurado para um número máximo de nós igual a 13.

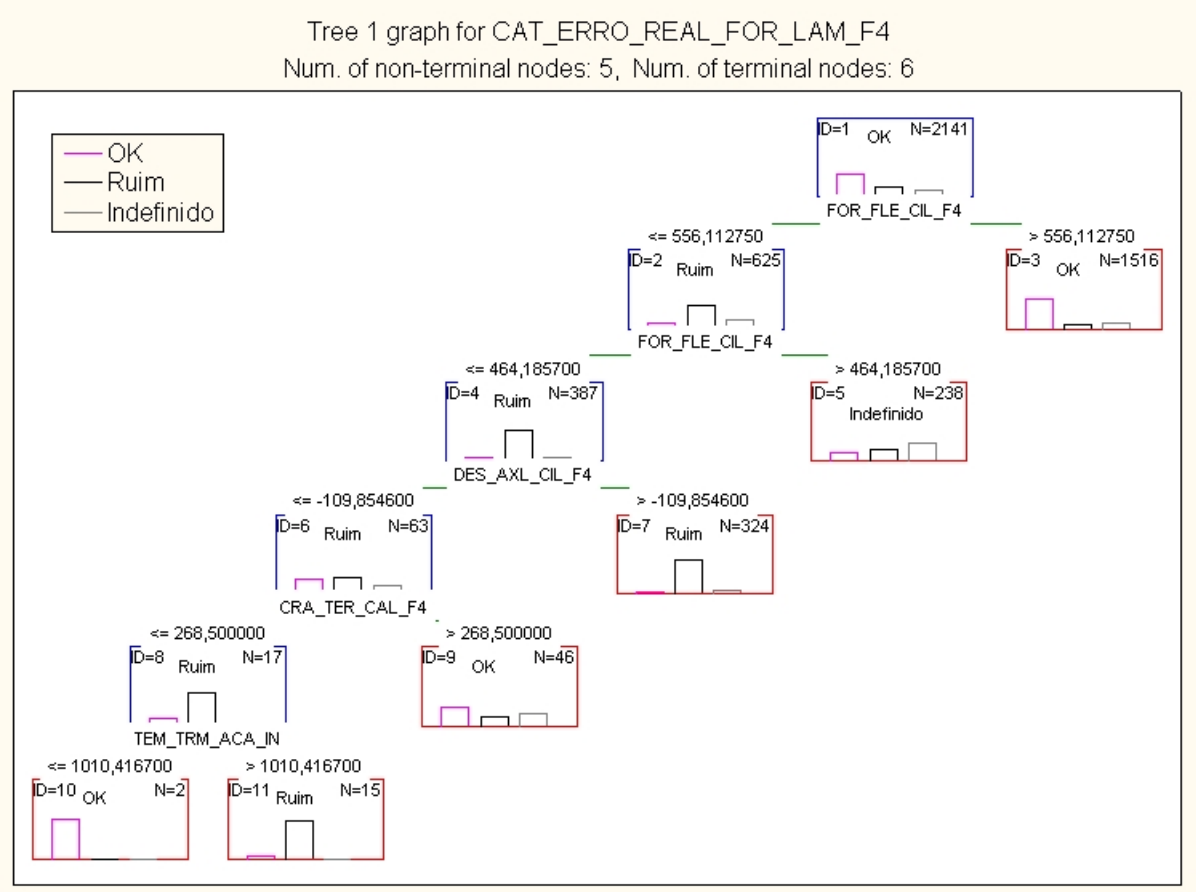

Figura 4.11: Árvore de Decisão para Cadeira 4, família de Aço 3

Pela figura 4.11 podemos observar que a árvore de decisão gerada é binária, pois cada nó se ramifica em no máximo dois nós-filhos. A legenda da árvore mostra as três classes existentes para variável guia do processo de classificação (OK, Indefinido, Ruim). Essa representação gráfica de árvore de decisão do *Statistica* é interessante, pois além de representar os atributos e valores de corte de cada nó, também apresenta um gráfico com a distribuição dos dados classificados pelo CART até o nó em questão. A figura 4.12 mostra a mesma árvore, porém de forma mais clara, contendo apenas a descrição dos nós e as regras de divisão de cada um.

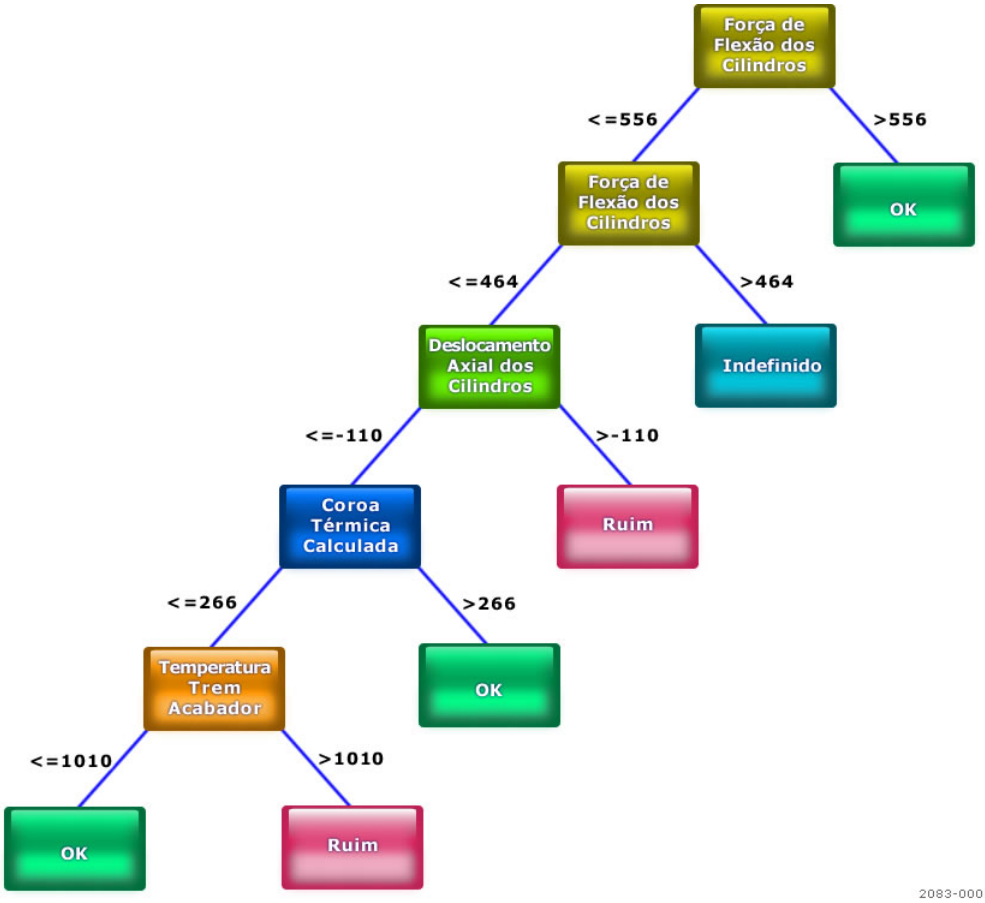

Figura 4.12: Árvore de Decisão para Cadeira 4, família de Aço 3 - Apenas descrição dos nós e regras.

Para identificar as principais variáveis associadas ao erro de força de laminação, foram geradas tabelas para cada cadeira com o *ranking* de importância resultante de cada execução do algoritmo CART[12]. Esse *ranking* mostra de forma direta quais variáveis têm maior influência no modelo gerado para classificar o erro de força dentro das

[12]As tabelas geradas podem ser observadas no Apêndice B

faixas definidas. A partir da análise dessas tabelas, identificou-se que as variáveis mais importantes para as cadeiras 3, 4 e 5 são praticamente as mesmas.

A execução inicial do CART foi realizada com os parâmetros *default* disponibilizados pela ferramenta de *Data Mining*. A variação manual de alguns deles, como por exemplo o número máximo de nós da árvore, mostrou maior impacto diretamente no índice de acerto dos modelos de classificação gerados, porém não influenciou na indicação de variáveis mais relevantes.

4.5 Resultados obtidos

Uma vez definidas quais variáveis foram consideradas pelo CART como as mais significativas para a classificação do erro de força de laminação, as mesmas foram apresentadas para os analistas da Siderúrgica que as separaram em basicamente 3 categorias:

- Variáveis que fazem parte da entrada do modelo para cálculo da força de laminação;
- Variáveis de "aprendizado", utilizadas para auto-corrigir tal modelo;
- Variáveis que não são consideradas diretamente pelo modelo de força.

Para cada uma das cadeiras e para cada família de aço para as quais o algoritmo foi executado, foram gerados gráficos das variáveis mais importantes. Como o objetivo do projeto é avaliar o comportamento das variáveis em função do erro de força, as mesmas foram ordenadas em função dessa medida.

Uma vez que os dados de processo variam bastante, um filtro de média móvel foi aplicado às variáveis, com uma janela de 25 amostras[13], para permitir uma análise mais clara das tendências. Além disso, os dados das variáveis foram normalizados entre 0 e 1 assim como o erro de força, para permitir a visualização de ambos em um mesmo gráfico.

Em cada gráfico foram também acrescentadas duas linhas verticais para indicar a transição entre as faixas **Ruim**, **Indefinido** e **OK** do erro de força.

As figuras de 4.13 a 4.15 mostram exemplos de algumas variáveis consideradas representativas na questão do erro de força, em análise conjunta realizada com os especialistas da Usina:

[13]O filtro de média móvel não altera a tendência das curvas, e apenas aproxima os pontos do gráfico para o valor médio da janela considerada. No caso em questão, como foi utilizado uma janela de 25 amostras, o valor plotado é sempre a média dos últimos 25 registros.

- Força de Flexão dos Cilindros;
- Velocidade de Laminação;
- Erro de Resistência à Deformação.

Além dessas três, também foram consideradas significativas as seguintes variáveis, após as análises dos gráficos:

- Coroa Térmica Calculada;
- Deslocamento Axial dos Cilindros.

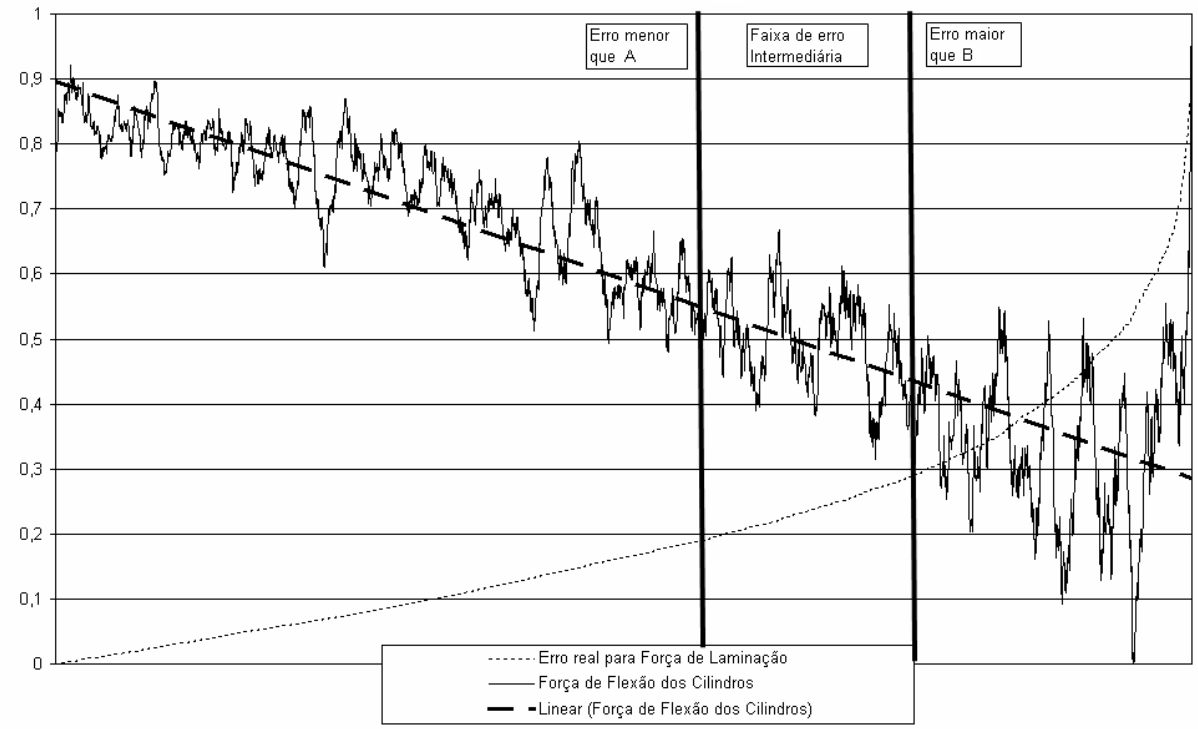

Figura 4.13: Gráfico da Força de Flexão do Cilindros - Cadeira 4, Família de Aço 3

Após análise inicial dos especialistas da Siderúrgica, foi levantada a necessidade de um maior detalhamento dessas 5 variáveis, juntamente com o erro de força de Laminação, para as cadeiras 3, 4 e 5. Como a análise por tipo de aço pareceu mais dificultar do que auxiliar a interpretação dos gráficos, foi decidido em conjunto com os analistas que apenas os dados correspondentes às famílias 3 e 9 seriam considerados, devido à grande representatividade dos mesmos na massa de dados e também devido à semelhança dos mesmos em

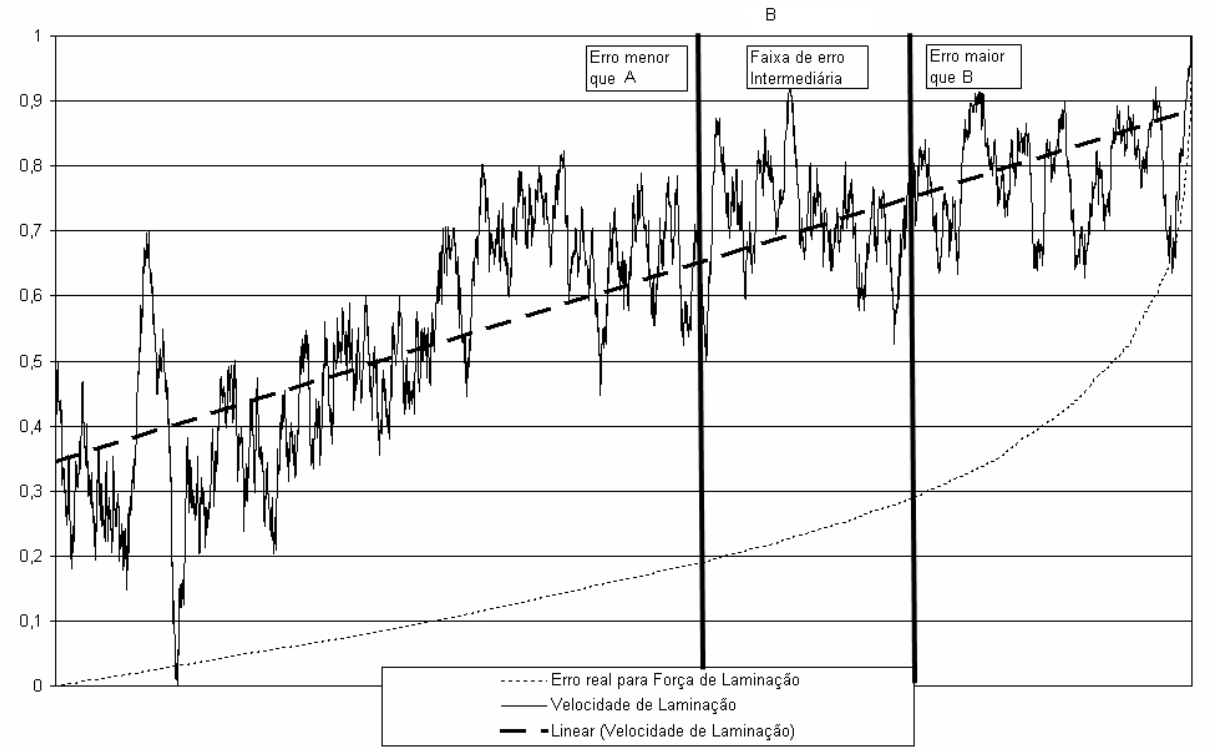

Figura 4.14: Gráfico da Velocidade de Laminação - Cadeira 4, Família de Aço 3

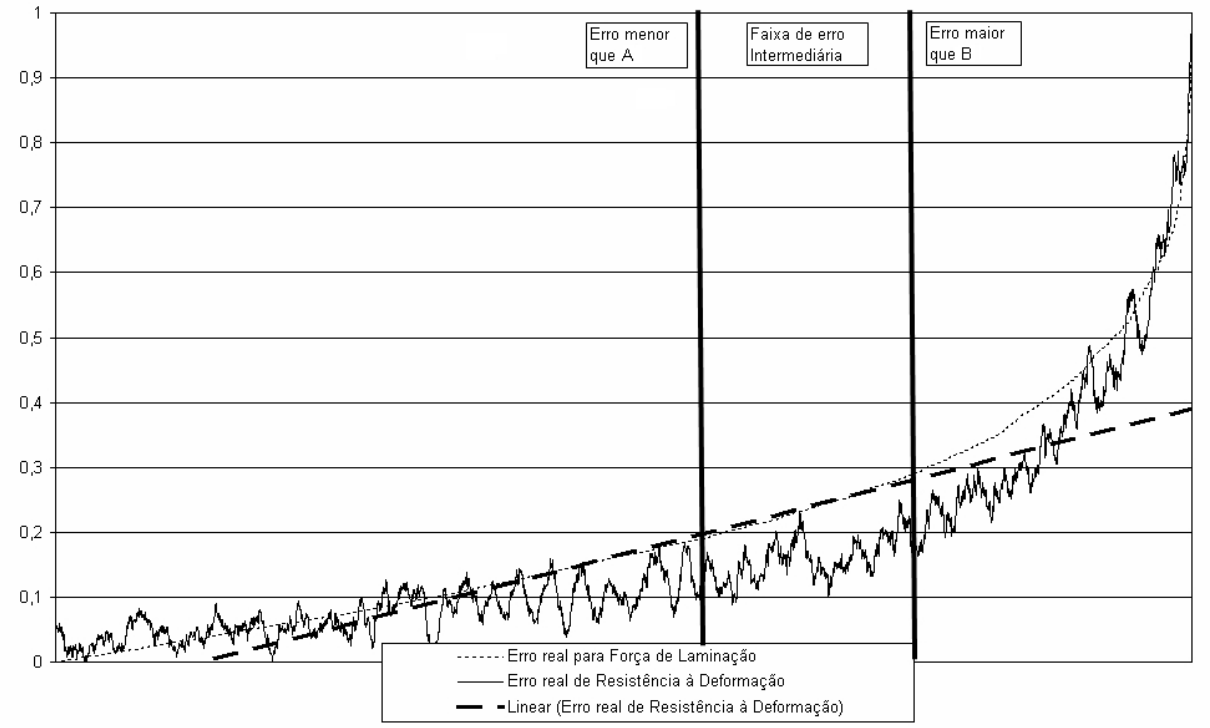

Figura 4.15: Gráfico do Erro de Resistência a Deformação - Cadeira 4, Família de Aço 3

termos de características físico-químicas, o que permitiria a análise em conjunto dessas

duas famílias.

Além de uma análise estatística das variáveis selecionadas por faixa do Erro de Força, apresentada na Tabela 4.9 para a cadeira 4 e no Apêndice C para as demais cadeiras, foram também geradas as taxas de correlação e variância. A tabela 4.10 mostra a Correlação e Variância entre as variáveis selecionadas.

ERRO DE FORÇA DE LAMINAÇÃO - em percentual				
Erro de Força	**Média**	**Máximo**	**Mínimo**	**Desvio Padrão**
OK	4,47	10,00	0,00	2,86
Indefinido	12,32	15,00	10,00	1,44
Ruim	23,01	104,27	15,02	7,02
ERRO DE RESISTÊNCIA À DEFORMAÇÃO - em percentual				
Erro de Força	**Média**	**Máximo**	**Mínimo**	**Desvio Padrão**
OK	5,05	26,45	0,00	3,58
Indefinido	6,92	27,52	0,00	4,67
Ruim	12,28	80,46	0,01	7,25
VELOCIDADE DE LAMINAÇÃO - em m/s				
Erro de Força	**Média**	**Máximo**	**Mínimo**	**Desvio Padrão**
OK	17,79	24,99	5,68	4,24
Indefinido	19,49	25,21	6,10	3,51
Ruim	20,03	24,98	5,84	3,06
COROA TÉRMICA CALCULADA - em microns				
Erro de Força	**Média**	**Máximo**	**Mínimo**	**Desvio Padrão**
OK	205,88	383,00	0,00	79,12
Indefinido	219,99	360,00	0,00	82,98
Ruim	229,12	366,00	0,00	81,44
DESLOCAMENTO AXIAL DOS CILINDROS - em mm				
Erro de Força	**Média**	**Máximo**	**Mínimo**	**Desvio Padrão**
OK	-22,82	150,01	-150,02	75,31
Indefinido	-15,63	150,00	-150,01	74,64
Ruim	0,98	150,01	-150,01	75,83
FORÇA DE FLEXÃO DOS CILINDROS - em KN				
Erro de Força	**Média**	**Máximo**	**Mínimo**	**Desvio Padrão**
OK	736,23	1494,28	101,50	156,04
Indefinido	637,84	1500,70	99,94	222,16
Ruim	583,82	1500,79	97,58	322,63

Tabela 4.9: Análise Estatística por Faixa do Erro de Força - cadeira 4

CORRELAÇÃO	(A)	(B)	(C)	(D)	(E)	(F)	(G)	(H)	(I)	(J)	(K)	(L)	(M)	(N)	(O)	(P)	(Q)	(R)
ERRO_REAL_FOR_LAM_F3 (A)	1,00																	
ERRO_REAL_FOR_LAM_F4 (B)	0,37	1,00																
ERRO_REAL_FOR_LAM_F5 (C)	0,17	0,26	1,00															
ERRO_REAL_RES_DEF_F3 (D)	0,41	0,24	0,16	1,00														
ERRO_REAL_RES_DEF_F4 (E)	0,30	0,62	0,15	0,21	1,00													
ERRO_REAL_RES_DEF_F5 (F)	0,14	0,27	0,28	0,07	0,19	1,00												
VEL_LAM_F3 (G)	0,03	0,26	0,33	0,29	0,02	0,00	1,00											
VEL_LAM_F4 (H)	0,05	0,26	0,34	0,29	0,02	-0,01	0,99	1,00										
VEL_LAM_F5 (I)	0,07	0,27	0,29	0,28	0,02	-0,02	0,97	0,99	1,00									
CRA_TER_CAL_F3 (J)	-0,17	0,09	0,26	-0,12	-0,04	0,09	0,35	0,34	0,32	1,00								
CRA_TER_CAL_F4 (K)	-0,14	0,13	0,28	-0,06	-0,04	0,06	0,46	0,47	0,46	0,96	1,00							
CRA_TER_CAL_F5 (L)	-0,15	0,12	0,27	-0,05	-0,03	0,06	0,45	0,45	0,44	0,95	0,99	1,00						
DES_AXL_CIL_F3 (M)	0,16	0,06	0,00	0,05	0,05	-0,03	0,07	0,09	0,11	-0,46	-0,39	-0,40	1,00					
DES_AXL_CIL_F4 (N)	0,15	0,14	0,06	0,04	0,07	0,02	0,22	0,26	0,28	-0,22	-0,16	-0,18	0,64	1,00				
DES_AXL_CIL_F5 (O)	0,09	0,21	0,22	0,06	0,03	0,02	0,54	0,57	0,59	0,18	0,26	0,22	0,42	0,67	1,00			
FOR_FLE_CIL_F3 (P)	-0,14	-0,16	-0,02	-0,35	-0,09	-0,04	-0,08	-0,09	-0,09	-0,09	-0,12	-0,13	-0,04	0,01	-0,04	1,00		
FOR_FLE_CIL_F4 (Q)	-0,04	-0,25	-0,01	-0,15	-0,09	-0,12	-0,17	-0,12	-0,11	-0,19	-0,19	-0,20	0,13	0,04	-0,01	0,42	1,00	
FOR_FLE_CIL_F5 (R)	0,05	-0,07	-0,39	-0,09	0,02	0,03	-0,29	-0,28	-0,23	-0,27	-0,31	-0,31	0,20	0,15	-0,21	0,10	0,28	1,00

VARIÂNCIA	(A)	(B)	(C)	(D)	(E)	(F)	(G)	(H)	(I)	(J)	(K)	(L)	(M)	(N)	(O)	(P)	(Q)	(R)
ERRO_REAL_FOR_LAM_F3 (A)	33,1																	
ERRO_REAL_FOR_LAM_F4 (B)	18,3	71,9																
ERRO_REAL_FOR_LAM_F5 (C)	7,9	18,1	66,3															
ERRO_REAL_RES_DEF_F3 (D)	14,1	12,2	7,8	35,7														
ERRO_REAL_RES_DEF_F4 (E)	9,6	29,3	6,6	7,1	31,2													
ERRO_REAL_RES_DEF_F5 (F)	4,8	13,5	13,6	2,5	6,2	34,9												
VEL_LAM_F3 (G)	0,1	1,63	2,0	1,3	0,1	0,0	0,5											
VEL_LAM_F4 (H)	0,4	2,7	3,3	2,1	0,1	-0,1	0,9	1,5										
VEL_LAM_F5 (I)	0,7	3,9	4,1	2,9	0,9	-0,2	1,2	2,1	3,0									
CRA_TER_CAL_F3 (J)	-92	72	199	-70	-23	50	24	39	52	8746								
CRA_TER_CAL_F4 (K)	-64	87	185	-31	-17	31	27	46	64	7274	6537							
CRA_TER_CAL_F5 (L)	-55	63	139	-18	-12	22	21	35	48	5659	5078	4032						
DES_AXL_CIL_F3 (M)	40	23	1,7	14	13	-7,0	2,2	4,8	8,7	-1907	-1388	-1122	1950					
DES_AXL_CIL_F4 (N)	67	91	39	20	30	6,9	12	24	37	-1541	-962	-890	2130	5758				
DES_AXL_CIL_F5 (O)	38	123	128	24	14	7,7	27	49	72	1171	1452	995	1293	3558	4935			
FOR_FLE_CIL_F3 (P)	-144	-247	-27	-375	-84	-43	-11	-19	-28	-1468	-1670	-1438	-323	169	-454	31307		
FOR_FLE_CIL_F4 (Q)	-50	-476	-14	-195	-110	-160	-28	-32	-41	-3892	-3504	-2803	1305	759	-224	16503	49868	
FOR_FLE_CIL_F5 (R)	49,3	-114	-589	-99	20	32	-39	-64	-74	-4722	-4662	-3673	1669	2164	-2741	3156	11650	35091

Tabela 4.10: Coeficientes de Correlação e Covariância entre as variáveis selecionadas.

O Coeficiente de Correlação é uma medida do grau de associação da relação linear entre duas variáveis. Essa medida pode variar entre -1 e 1, sendo que quanto mais próximo desses extremos, maior a correlação entre duas variáveis, positiva ou negativamente. Já o Coeficiente de Covariância fornece uma medida não padronizada do grau no qual duas variáveis se movem juntas, e é estimado tomando o produto dos desvios da média para cada variável em cada medida.

Devido a dificuldade inicial de interpretação dos gráficos traçados para essas variáveis, foi também identificada a necessidade de nova geração dos mesmos, porém agora na sua forma bruta, ou seja, sem a aplicação do filtro de média móvel. As figuras 4.16 a 4.20 apresentam os gráficos gerados para a cadeira 4. Os gráficos para as cadeiras 3 e 5 podem ser vistos no Apêndice D.

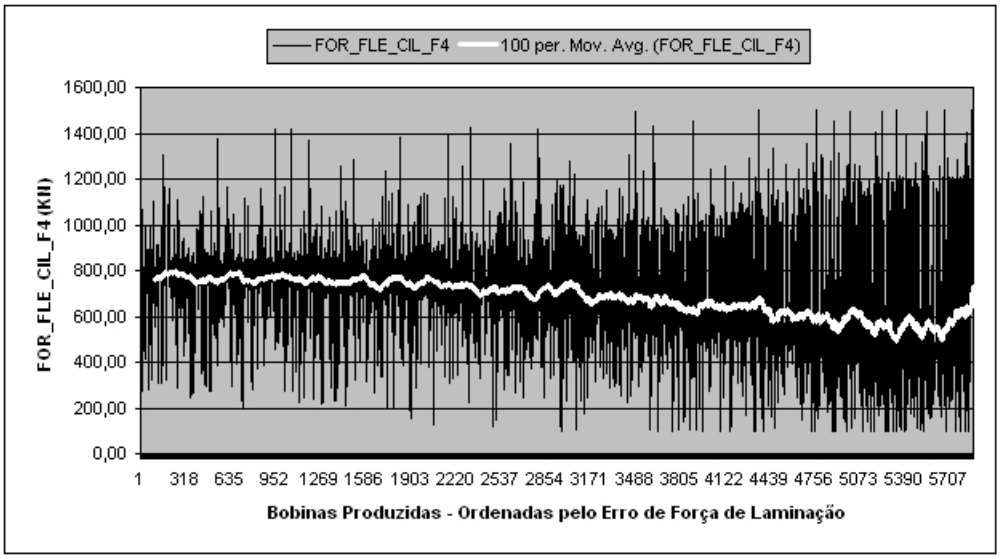

Figura 4.16: Gráfico da Força de Flexão do Cilindros com e sem Filtro de Média Móvel (100 pontos) - cadeira 4, Família de Aço 3 e 9

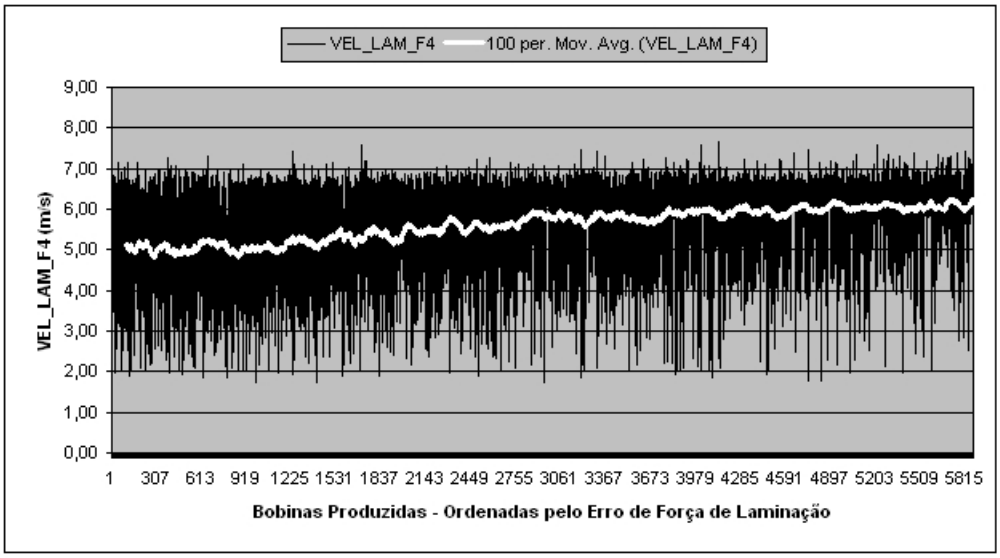

Figura 4.17: Gráfico da Velocidade de Laminação com e sem Filtro de Média Móvel (100 pontos) - cadeira 4, Família de Aço 3 e 9

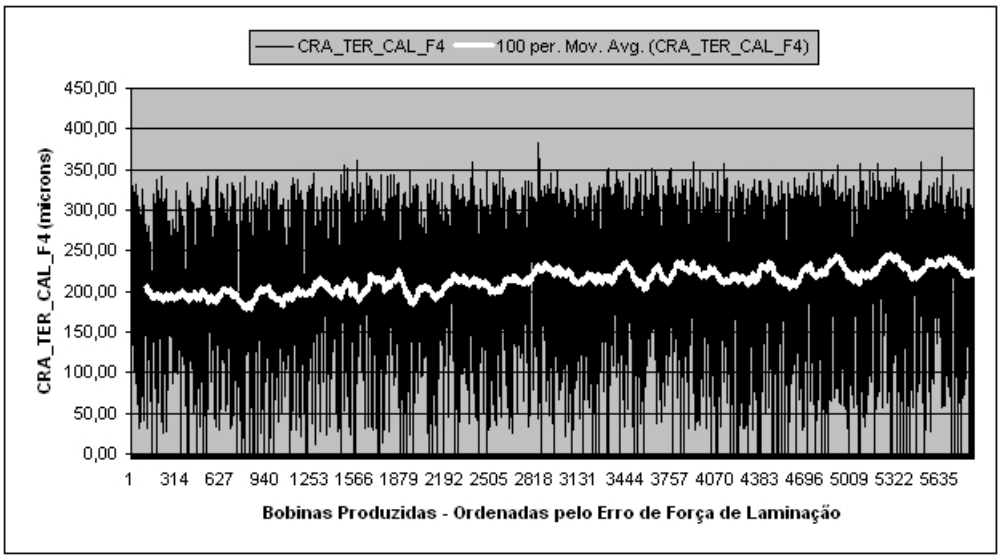

Figura 4.18: Gráfico da Coroa Térmica Calculada com e sem Filtro de Média Móvel (100 pontos) - cadeira 4, Família de Aço 3 e 9

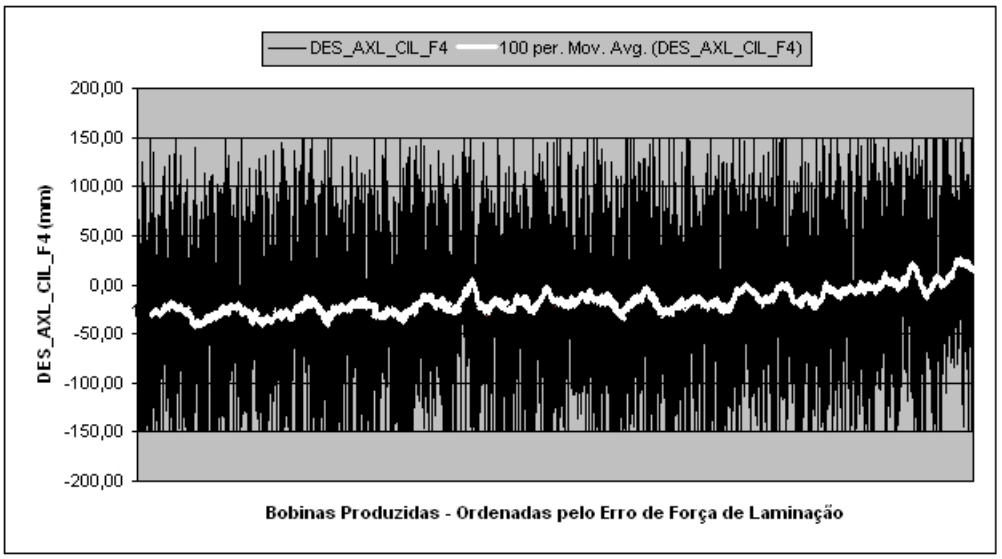

Figura 4.19: Gráfico do Deslocamento Axial dos Cilindros com e sem Filtro de Média Móvel (100 pontos) - cadeira 4, Família de Aço 3 e 9

4.5.1 Regras de Associação

Com o objetivo de validar os resultados obtidos pelo CART, foi aplicado o algoritmo de regras de associação no conjunto final de variáveis selecionadas. Para gerar essas regras, foi necessária a realização de uma discretização prévia (e manual) das variáveis, uma vez que a função de *Association Rules* do *Statistica* não trabalha com dados contínuos. Essa discretização foi feita em torno da média de cada variável dentro das faixas

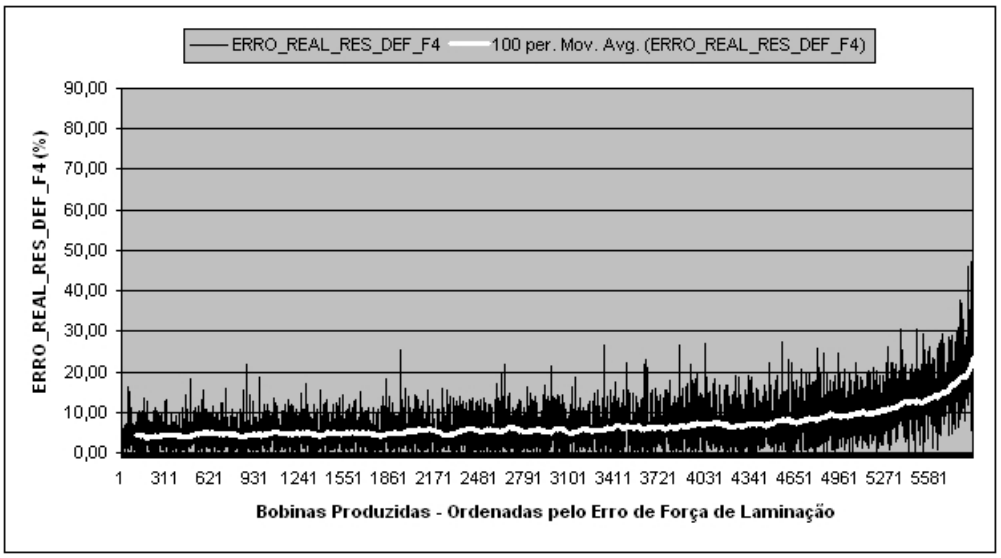

Figura 4.20: Gráfico do Erro de Resistência a Deformação com e sem Filtro de Média Móvel (100 pontos) - cadeira 4, Família de Aço 3 e 9

de erro de Força de Laminação. O Excel© foi utilizado para gerar as novas variáveis categóricas, sendo que cada uma pôde assumir apenas três valores:

- Alto;
- Médio;
- Baixo.

Foram geradas mais de 1000 regras para cada cadeira; porém, para facilitar a análise, as mesmas foram ordenadas pelo coeficiente de Confiança. As Tabelas 4.11 a 4.13 apresentam as regras com mais de 90% de Confiança. No Apêndice E estão listadas mais regras com grau de confiança entre 80% e 90%.

Se	==>	**Então**	**Confiança(%)**	**Cobertura(%)**
ERD_BAIXO, VL_BAIXA, FFC_ALTA, DXL_BAIXA	==>	OK	94,15	10,97
ERD_BAIXO, VL_BAIXA, FFC_ALTA	==>	OK	94,08	19,76
VL_BAIXA, FFC_ALTA, DXL_BAIXA	==>	OK	93,81	12,91
OK, ERD_BAIXO, VL_ALTA, DXL_BAIXA	==>	CT_ALTA	93,49	13,46
ERD_BAIXO, VL_BAIXA, DXL_BAIXA	==>	OK	93,07	13,27
ERD_BAIXO, VL_ALTA, DXL_BAIXA	==>	CT_ALTA	93,07	14,87
FFC_ALTA, DXL_BAIXA	==>	OK	92,12	29,47
FFC_ALTA, CT_ALTA, DXL_BAIXA	==>	OK	91,55	19,57
FFC_ALTA, CT_ALTA	==>	OK	91,47	29,94
ERD_BAIXO, DXL_BAIXA	==>	OK	91,35	27,87
ERD_BAIXO, CT_ALTA, DXL_BAIXA	==>	OK	91,29	21,05
ERD_BAIXO, FFC_ALTA, DXL_BAIXA	==>	OK	91,21	20,68
ERD_BAIXO, VL_BAIXA, CT_ALTA	==>	OK	91,16	11,24
ERD_BAIXO, FFC_ALTA, CT_ALTA	==>	OK	91,14	22,94
ERD_BAIXO, FFC_ALTA, CT_ALTA, DXL_BAIXA	==>	OK	91,12	14,85
FFC_ALTA, VL_ALTA, DXL_BAIXA	==>	OK	90,86	16,25
VL_BAIXA, FFC_ALTA	==>	OK	90,80	23,20
FFC_ALTA, VL_ALTA, CT_ALTA, DXL_BAIXA	==>	OK	90,77	13,23
CT_ALTA, ERD_ALTO	==>	VL_ALTA	90,76	10,37
FFC_ALTA, VL_ALTA, CT_ALTA	==>	OK	90,71	20,80
ERD_BAIXO, VL_BAIXA	==>	OK	90,70	25,41
ERD_BAIXO, FFC_ALTA	==>	OK	90,59	37,56
ERD_BAIXO, VL_ALTA, CT_ALTA, DXL_BAIXA	==>	OK	90,49	13,46
ERD_BAIXO, VL_ALTA, DXL_BAIXA	==>	OK	90,09	14,39
FFC_ALTA	==>	OK	90,00	52,41

Tabela 4.11: Regras de Associação Geradas para as variáveis da cadeira 3.

Se	**==>**	**Então**	**Confiança(%)**	**Cobertura(%)**
ERD_BAIXO, VL_BAIXA, CT_BAIXA, FFC_ALTA	==>	OK	94,63	11,12
ERD_BAIXO, VL_BAIXA, FFC_ALTA	==>	OK	94,39	14,05
ERD_BAIXO, CT_BAIXA, FFC_ALTA	==>	OK	93,17	15,10
ERD_BAIXO, VL_BAIXA, CT_BAIXA	==>	OK	92,54	12,68
ERD_BAIXO, VL_BAIXA, DXL_BAIXA	==>	OK	92,15	10,20
ERD_BAIXO, VL_BAIXA	==>	OK	91,19	17,28
ERD_BAIXO, FFC_ALTA, DXL_BAIXA	==>	OK	90,58	14,08

Tabela 4.12: Regras de Associação Geradas para as variáveis da cadeira 4.

Se	**==>**	**Então**	**Confiança(%)**	**Cobertura(%)**
ERD_BAIXO, VL_BAIXA, CT_BAIXA, FFC_ALTA	==>	OK	93,22	13,59
ERD_BAIXO, VL_BAIXA, FFC_ALTA	==>	OK	93,03	16,84
ERD_BAIXO, VL_BAIXA, DXL_BAIXA, FFC_ALTA	==>	OK	92,73	12,39
CT_ALTA, FFC_BAIXA, DXL_ALTA	==>	VL_ALTA	92,27	16,48
ERD_BAIXO, CT_BAIXA, DXL_BAIXA, FFC_ALTA	==>	OK	91,80	12,21
ERD_BAIXO, CT_BAIXA, FFC_ALTA	==>	OK	91,29	17,86
OK, ERD_BAIXO, VL_BAIXA, CT_BAIXA	==>	FFC_ALTA	91,09	13,59
CT_ALTA, DXL_ALTA	==>	VL_ALTA	90,22	22,63
ERD_BAIXO, CT_ALTA, DXL_ALTA	==>	VL_ALTA	90,18	12,36

Tabela 4.13: Regras de Associação Geradas para as variáveis da cadeira 5.

Observa-se pelas tabelas 4.11 a 4.13 que não foram geradas regras para a classe "Ruim" com um alto grau de Confiança. Como o tamanho da base de dados influencia diretamente na geração das regras[14], foi realizada uma nova tentativa utilizando apenas o conjunto de dados da classe "Ruim". Nessa tentativa foram geradas as seguintes regras para a cadeira 4, descritas na tabela 4.14:

Se	**==>**	**Então**	**Confiança(%)**	**Cobertura(%)**
FFC_BAIXA	==>	Ruim	100,00	70,69
VL_ALTA	==>	Ruim	100,00	69,26
CT_ALTA	==>	Ruim	100,00	63,70
ERD_ALTO	==>	Ruim	100,00	61,16
DXL_ALTA	==>	Ruim	100,00	54,58

Tabela 4.14: Regras de Associação Geradas para faixa de erro "Ruim" - cadeira 4.

Foi também realizada uma nova discretização dos dados, com a utilização do algoritmo *Equal Frequency*[15] disponível no *software Rosetta*. Os dados discretizados foram exportados para o *Statistica*, porém o número de regras geradas com alto grau de confiança (maior do que 80%) foi muito inferior (menos de um terço) à quantidade obtida anteriormente com a discretização manual.

[14] A função de *Association Rules* do *Statistica* utiliza uma implementação do Algoritmo *Apriori* para geração das regras e isso implica no fato de que o tamanho da *flat table* a ser utilizada influencia na geração das regras (Agrawal *et al.*, 1993).

[15] O algoritmo *Equal Frequency* realiza a discretização dos dados de forma que o número de registros em cada classe final da variável que está sendo discretizada seja o mesmo. Por exemplo, uma variável que possui 60 registros e é discretizada em 3 classes, possui os valores limites da discretização ajustados para que cada classe fique com 20 registros.

Capítulo 5

Análise dos Resultados

Durante todo o processo de KDD vários resultados foram obtidos, com a indicação de diversas correlações entre variáveis de processo e o erro de força de laminação. Na seção 4.5 são apresentados apenas os resultados que enquadram-se dentro do conceito de *Data Mining*, ou seja, os que foram considerados como novos e potencialmente úteis. No presente capítulo é feita uma análise mais aprofundada desses resultados.

É importante ressaltar que, além dos gráficos resultantes da utilização do CART, também foram geradas regras de associação na tentativa de se obter informações relevantes sobre as variáveis selecionadas e o erro de força. Porém, avaliando-se essas regras, concluiu-se que as mesmas apenas confirmam os resultados obtidos com as análises do gráficos ou das estatísticas geradas, ou seja, não apresentam nada de novo com relação às correlações entre variáveis.

5.1 Correlação entre Erro de Força e Atuação do Operador

Conforme mencionado na Seção 4.3 e demonstrado nas Figuras 4.7 e 4.8, o primeiro resultado gerado pelo processo de KDD, com potencial de uso pela Usina Siderúrgica, foi a identificação de uma correlação acentuada da atuação do operador com o aumento do erro de força nas cadeiras 4 e 5. Essa constatação foi exposta aos analistas da área de Automação da Siderúrgica e os mesmos levantaram duas hipóteses diante do fato:

- Os operadores, ao atuarem no processo, podem estar aumentando o erro de força, sendo então uma das possíveis causas para que essa variável esteja mais acentuada nas cadeiras 4 e 5;

- A atuação dos operadores pode ser uma conseqüência do erro de força e eles estariam tentando, então, corrigir essa variável, porém não estariam tendo sucesso.

Foi discutido pelos analistas que, nesse momento do processo em que o operador atua, ele pode estar mais preocupado com a estabilidade do processo do que com o erro de força propriamente dito, buscando, por exemplo, evitar que a tira se rompa. Em outras palavras, pode ser mais importante evitar que uma bobina vire sucata do que se preocupar com as questões de qualidade do produto final. Esse resultado foi encaminhado para a equipe técnica responsável pelo processo de Laminação, com o objetivo de validar qual das hipóteses teria mais sentido e avaliar melhor o significado dessa constatação.

5.2 Variáveis significativas indicadas pelo CART

Um segundo resultado importante do trabalho foi a redução do número de variáveis a serem analisadas com relação ao erro de força. Com a utilização do algoritmo CART para gerar a classificação de importância das variáveis utilizadas no modelo de classificação do erro de força, o trabalho manual de análise foi reduzido em cerca de 10 vezes[1]. Isso permitiu uma análise mais aprofundada dessas variáveis, com a geração de estatísticas, gráficos de tendência e regras de associação.

5.2.1 Influência da Força de Flexão dos Cilindros

Dentre as variáveis selecionadas como mais importantes para a análise do erro da força de laminação está a "Força de Flexão dos Cilindros", que obteve nota máxima em todas as cadeiras no coeficiente de importância de classificação, levando em conta todos os tipos de aço analisados. Avaliando-se os gráficos gerados para essa variável (Figuras D.1 e D.3), fica fácil perceber a relação de proporcionalidade inversa entre a mesma e o erro de força; ou seja, quanto menor a "Força de Flexão dos Cilindros" maior é o erro de força. Os analistas da Usina Siderúrgica ficaram inicialmente intrigados com essa forte relação, porém não conseguiram, em uma primeira análise, obter uma explicação para esse resultado. Após uma reflexão mais profunda sobre o assunto, foi solicitado, pelos analistas, o desenho dos gráficos de forma discreta (somente com os pontos), mantendo

[1]No início da aplicação das técnicas de *Data Mining* haviam sido selecionadas cerca de 200 variáveis, após essa etapa, o número de variáveis a serem analisadas resumiu-se a pouco menos de 20.

a ordenação dos dados pelo erro de força de laminação. Esses gráficos, apresentados na Figuras 5.1 a 5.3, confirmaram seguinte suposição:

- Sempre que se modifica a faixa de operação normal da "Força de Flexão dos Cilindros" (em torno de 800 KN), há uma influência negativa na força de laminação.

Essa relação mostra-se evidente nas cadeiras 3, 4 e 5, e a existência de uma maior dispersão dos pontos nas duas últimas pode estar relacionada ao fato de existirem outras influências no erro de força que são mais acentuadas nessas duas cadeiras. A relação de proporcionalidade inversa entre essa variável e o erro de força deve-se ao fato de uma maior quantidade de pontos estar na faixa inferior dos gráficos, considerando o valor de 800 KN como médio.

5.2.2 Correlação entre o Erro de Força e o Erro de Resistência à Deformação

O "Erro de Resistência à Deformação" é outra variável com nota bastante elevada no coeficiente de importância para classificação. Essa variável apresenta um dos maiores índices de correlação com o erro de força, dentre as variáveis selecionadas, conforme apresentado na tabela 4.10. Observa-se no gráfico da Figura 4.15 que o crescimento dessa variável acompanha o erro de força, inclusive na sua característica exponencial. Essa forte

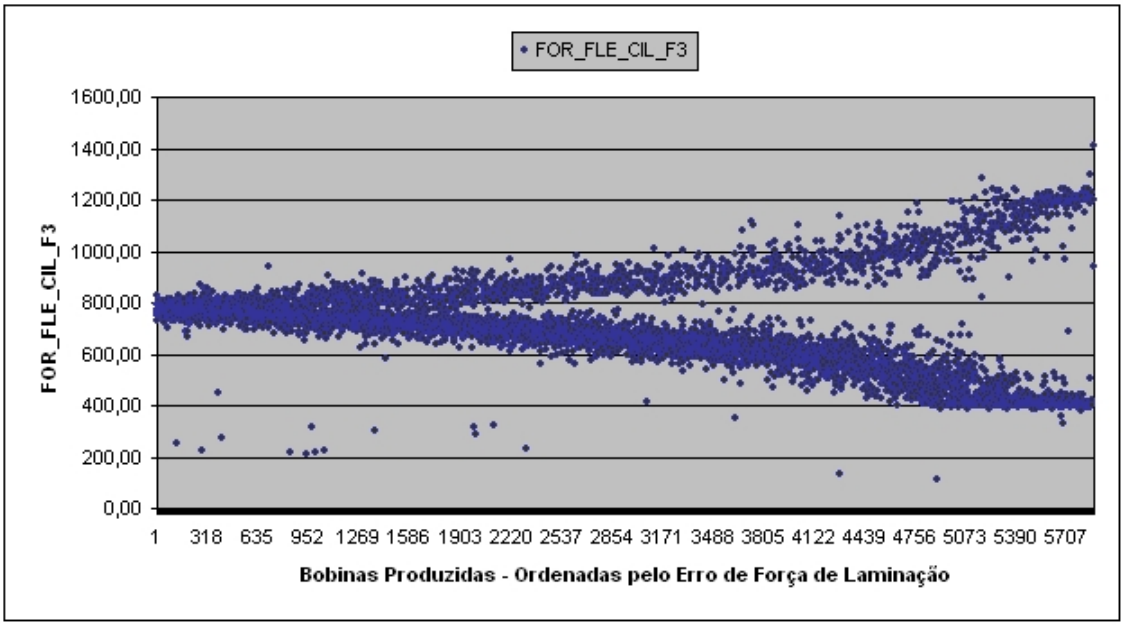

Figura 5.1: Força de Flexão dos Cilindros - Cadeira 3, dados discretos ordenados pelo Erro de Força

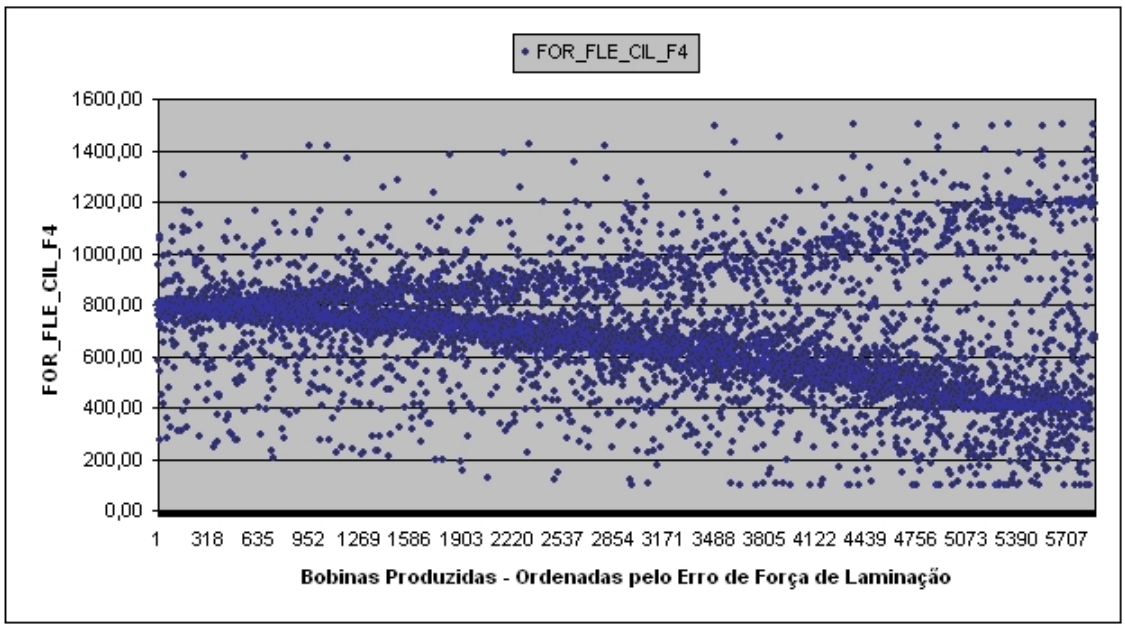

Figura 5.2: Força de Flexão dos Cilindros - Cadeira 4, dados discretos ordenados pelo Erro de Força

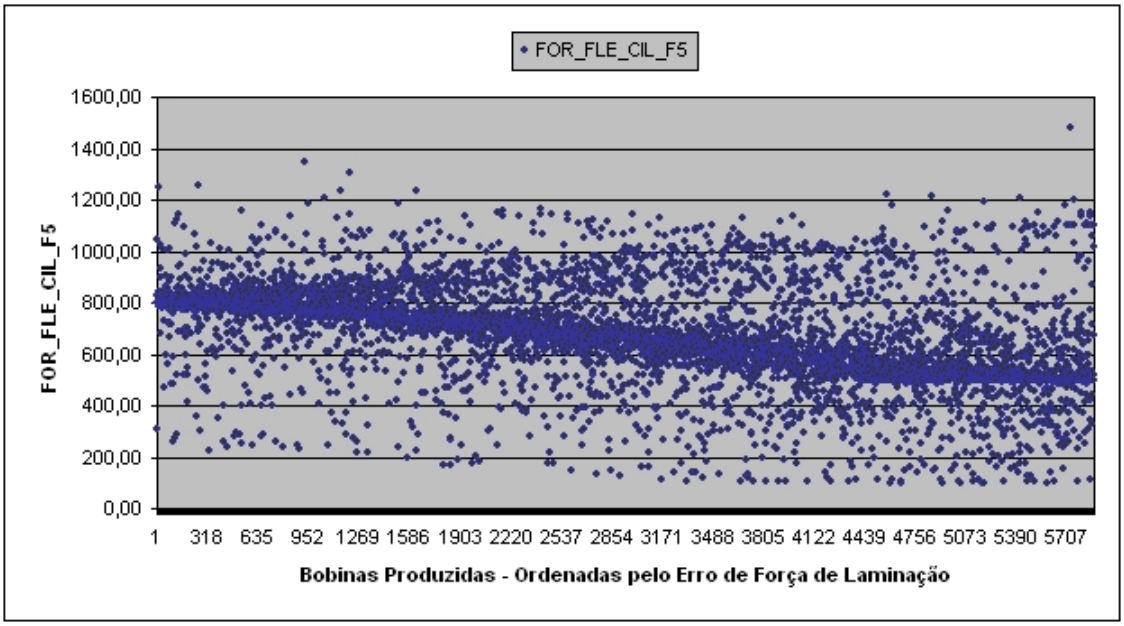

Figura 5.3: Força de Flexão dos Cilindros - Cadeira 5, dados discretos ordenados pelo Erro de Força

correlação, segundo os analistas da Siderúrgica, pode indicar um possível problema no cálculo da "Resistência a Deformação", uma vez que essa medida é utilizada como entrada dentro do modelo de "Força de Laminação". Esse resultado já era, de certa forma, de conhecimento da Usina Siderúrgica, pois trabalhos anteriores já haviam sido realizados

com o objetivo de ajustar o cálculo dessa variável no modelo.

5.2.3 Análise da Coroa Térmica Calculada

Outra variável que também obteve uma elevada nota de importância no modelo de classificação do erro de força, quando considerados todos os tipos de aço, é a "Coroa Térmica Calculada". Analisando-se os gráficos dessa variável, observa-se um comportamento distinto entre a cadeira 3 e as cadeiras 4 e 5: na cadeira 3, a "Coroa Térmica Calculada" apresenta um comportamento levemente inversamente proporcional ao erro de força; já para as outras duas cadeiras, quanto maior a "Coroa Térmica Calculada", maior o erro de força, com uma relação proporcional mais forte. Diante desse resultado, os analistas levantaram a hipótese de que, talvez, o controle de superfície dos cilindros de laminação pode estar melhor ajustado para o conjunto de cadeiras 1, 2 e 3, do que para as cadeiras 4, 5 e 6. Isso deve-se ao fato que quanto melhor esse controle, menor o valor calculado para a Coroa Térmica, o qual tem justamente o objetivo de fazer uma "compensação" no modelo, devido a variações na superfície do cilindro por aumento ou diminuição de temperatura.

5.2.4 Influência da Velocidade e do Deslocamento Axial dos Cilindros

Outras duas variáveis também foram selecionadas como relevantes na análise do erro de força: a "Velocidade de Laminação" e o "Deslocamento Axial dos Cilindros". Através dos gráficos traçados nas Figuras 4.17 e 4.19, observa-se que ambas possuem um comportamento semelhante com relação ao erro de força. Esse comportamento é caracterizado por uma proporcionalidade entre essas variáveis e o erro de força. Essa tendência é válida para todas as cadeiras, porém é mais acentuada para as cadeiras 4 e 5. Com relação à velocidade, segundo os analistas da Usina Siderúrgica, esse comportamento é de certa forma esperado, pois um aumento dessa variável está relacionado com a produção de materiais de espessuras menores e, nesses casos, o erro tende a ser maior. Portanto, esse resultado do processo de KDD é trivial, uma vez que pode ser obtido por outros meios devido ao conhecimento que os analistas já possuíam. Já com relação ao comportamento do "Deslocamento Axial dos Cilindros", os analistas acreditam que essa leve proporcionalidade não é conclusiva e, a princípio, não será considerada como válida.

Capítulo 6

Conclusões e Perspectivas Futuras

6.1 Conclusões

Os resultados apresentados no capítulo anterior comprovam o sucesso da aplicação de KDD à base de dados do processo de Laminação de Tiras a Quente. Entretanto, é importante destacar a grande complexidade e dificuldade de execução do projeto realizado. Pode-se classificar os fatos que justificam essa afirmação naqueles que são inerentes às características do processo de "Descoberta do Conhecimento em Bancos de Dados" e naqueles específicos ao ambiente industrial no qual o projeto foi realizado.

Conforme descrito no Capítulo 2, analisar e extrair informações de uma grande massa de dados não é um processo trivial. No presente trabalho, a manipulação inicial do dados foi uma das etapas mais desafiadoras. Com os recursos de *hardware* que estavam disponíveis, não foi possível a restauração por completo do *backup* do Banco de Dados e, sem os índices, uma simples consulta a uma tabela demorava vários minutos. A Seção 4.2 apresentou a estratégia utilizada para contornar essa situação e viabilizar as manipulações que se fizeram necessárias. Porém, para aplicações futuras, recomenda-se fortemente uma maior disponibilidade de espaço em disco rígido e também a utilização de computadores de performance mais elevada. Outro ponto relevante sobre essa etapa é a confirmação da necessidade de conhecimento na área de Banco de Dados dentro processo de KDD. Apesar da Internet ser uma excelente fonte de conhecimento sobre os SGDBs, questões mais complexas sobre a execução de comandos SQL são de difícil solução e, nesse trabalho, só foram sanadas com a ajuda de especialistas.

Outro ponto inerente ao processo de KDD e que merece destaque nesse trabalho é a grande quantidade de tempo empregada na etapa de análises preliminares. Não é fácil

a familiarização com um banco de dados que possui milhões de registros. A utilização exaustiva do Excel© foi um fator-chave para acelerar e, em alguns momentos até mesmo viabilizar a análise dos dados. Merece destaque aqui a funcionalidade de "tabela dinâmica" que permite de maneira simples a filtragem, agrupamento, correlação e cálculos matemáticos entre as diversas variáveis. Um dos principais resultados do processo de KDD foi, de fato, obtido diretamente nessa etapa do trabalho, na qual se observou a correlação entre a atuação do operador e o aumento do erro de força nas cadeiras de laminação 4 e 5. A iteração com os especialistas nos dados também foi relevante durante essa etapa, levando até à repetição da mesma por completo, conforme descrito ao final da Seção 4.3.

A etapa de *Data Mining* foi, talvez, a mais simples de todo o processo. Porém, é importante destacar que a familiarização inicial com a ferramenta utilizada foi fundamental. O *Statistica*, por se tratar de um pacote completo para análise estatística de dados, possui uma disponibilidade muito grande de algoritmos, e os mesmos possuem várias opções de configuração. A escolha criteriosa pela utilização direta do CART reduziu o esforço a ser gasto nessa etapa. Caso isso não ocorresse, seria necessário um trabalho adicional com a execução de testes dos vários algoritmos disponíveis para a tarefa de classificação. Dentro ainda dessa etapa, destaca-se o esforço para consolidação e apresentação dos resultados, que foi maior do que o tempo gasto com a execução do CART propriamente dito. Questões como filtragem dos dados e geração de gráficos foram realizadas com ajuda do Excel©, devido às limitações do *Statistica*.

Entre os aspectos específicos relativos ao ambiente industrial do qual os dados foram originados, destacam-se os seguintes pontos:

- Dificuldade de acesso aos analistas especialistas nos dados. No ambiente industrial a disponibilidade dos profissionais é algo crítico, uma vez que os mesmos são responsáveis pelo funcionamento adequado dos sistemas que mantêm a planta em operação. Além disso, a realização do trabalho fora do ambiente da Usina dificultou ainda mais a interação necessária entre o "Engenheiro do Conhecimento" que executou o processo de KDD e os especialistas no problema que estava sendo tratado;

- Elevada dispersão dos dados relativos às variáveis de processo. Por se tratarem de medidas diretas dos sensores da planta de Laminação de Tiras a Quente, os dados apresentam, de maneira geral, grande variabilidade em torno dos valores médios, o

que dificulta bastante uma análise visual dos gráficos de tendência. Os mecanismos de filtragem podem minimizar essa questão, mas também podem impedir a observação de alguns comportamentos, como foi o caso da análise preliminar dos gráficos relativos à "Força de Flexão dos Cilindros".

De maneira geral, não observamos muitas dificuldades na aplicação de KDD que sejam específicas do processo industrial, ou seja, a maioria dos problemas enfrentados é inerentes à própria metodologia aplicada, como por exemplo o grande esforço necessário para preparar e limpar uma base de dados com centenas de *gigabytes*.

A tabela 6.1 apresenta uma visão consolidada de cada fase que foi executada dentro do processo de KDD, destacando o esforço gasto em cada etapa, a necessidade de envolvimento dos especialistas no processo industrial e o grau de dificuldade com relação a questões tecnológicas.

Etapa	**Esforço Gasto**	**Participação dos Especialistas**	**Dificuldades Tecnológicas**
Definição do Problema	5 %	Alta	Nenhum
Preparação e Limpeza da Base de Dados	40%	Alta	Elevado
Análises Preliminares	30%	Média	Baixo
Data Mining	10%	Baixa	Médio
Análise dos Resultados	15%	Alta	Nenhum

Tabela 6.1: Visão consolidada das fases do processo de KDD executado.

Os percentuais apresentados com relação ao esforço gasto estão estimados em relação à quantidade total de horas efetivamente gastas, e não à duração total de cada etapa. A fase de "Definição do Problema", por exemplo, apesar de representar apenas 5% do total do esforço gasto, demorou vários meses para ser cumprida. Analisando-se essa tabela e as demais questões anteriormente colocadas, nota-se que a etapa de "Preparação e Limpeza da Base de Dados" foi realmente a mais difícil, utilizando-se como critério as medidas de esforço gasto, necessidade de envolvimento dos especialistas na base de dados e o grau de dificuldades tecnológicas que surgiram e tiveram que ser superadas.

Com relação aos resultados da fase de *Data Mining* propriamente ditos, pode-se afirmar que a correlação entre a "Força de Flexão dos Cilindros" e o "Erro de Força" é um resultado novo, relevante e deve ser usado pelos analistas da Usina Siderúrgica em algum trabalho interno visando uma otimização do processo de Laminação de Tiras a

Quente. A questão da atuação do operador também deve ser encaminhada para maiores estudos. Além desses pontos, duas considerações adicionais mereceram destaque por parte dos analistas da Usina Siderúrgica:

- A grande abrangência do trabalho realizado para tratar de um problema considerado complexo e de difícil solução foi um ponto extremamente positivo. O amplo conjunto de variáveis utilizadas permitiu a análise de várias possíveis influências no erro de força. Isso por si só comprova a capacidade das técnicas de *Data Mining* em lidar com um grande conjunto de atributos e amostras.

- Ainda segundo os analistas, a maneira com a qual os resultados foram apresentados não permitiu uma análise simples e direta da informação extraída. Apesar de não conseguirem indicar qual seria uma forma mais adequada, foi consenso entre eles que os gráficos de tendência apresentados muitas vezes mais dificultaram do que ajudaram nas análises.

6.2 Perspectivas Futuras

Apesar dos resultados positivos obtidos, em aplicações futuras várias questões podem ser revistas para otimizar o trabalho em algumas fases do projeto e também buscar resultados ainda mas expressivos com os algoritmos de *Data Mining*.

Além de uma maior disponibilidade de recursos de *hardware*, um ponto que resultaria em ganhos significativos para o processo de KDD na Usina Siderúrgica, com relação ao tempo de execução do projeto, seria a automatização da geração de *flat tables* a partir dos Bancos de Dados originais. Com o conhecimento adquirido sobre a estrutura dos dados e sobre o formato necessário para gerar as entradas para os algoritmos de *Data Mining*, seria possível desenhar-se um aplicativo que, de forma flexível, conseguisse gerar um *flat table* a partir de uma lista de variáveis a serem analisadas dentro de um intervalo de tempo. Esse aplicativo poderia ser desenvolvido utilizando recursos do próprio SGBD que armazena os dados do processo de produção, ou dentro do *Statistica*, que disponibiliza bibliotecas e interfaces de *software* para viabilizar esse automatismo, incluindo a possibilidade de utilização de técnicas de processamento paralelo ou distribuído[1]. Uma versão

[1]Devido à grande quantidade de dados que podem vir a ser processados, recursos como processamento paralelo, com mais de um processador trabalhando ao mesmo tempo dentro de um computador, ou

mais completa desse aplicativo poderia ainda executar funções de limpeza automática dos dados, baseado no conhecimento adquirido, de modo que o responsável pelo processo de KDD poderia dedicar mais tempo às etapas de Análises Preliminares e também na execução dos algoritmos de *Data Mining*.

Uma questão-chave, que vale de aprendizado para qualquer futura aplicação de KDD dentro de ambientes industriais, é a necessidade de maior interação com os responsáveis pelos dados do processo com os quais se está trabalhando. Conforme citado na seção anterior, a disponibilidade dos analistas dentro do ambiente industrial é algo crítico e deve ser levada em consideração desde a fase de planejamento do projeto. Sem um grande envolvimento desses profissionais, as etapas de "Definição do Problema", "Preparação e Limpeza da Base de Dados" e "Análise dos Resultados" ficam consideravelmente prejudicadas.

Uma última consideração relevante a ser feita é o fato de que não foi objetivo desse trabalho esgotar as possibilidades de análise da massa de dados em questão. O projeto realizado tinha como meta provar a viabilidade da aplicação do processo de KDD em um ambiente industrial. Vários outros projetos podem ser realizados em cima dessa mesma base de dados, utilizando outros algoritmos de classificação ou até mesmo outras técnicas de *Data Mining* para analisar o problema do erro de força, ou até mesmo outras questões que estejam relacionados com esse grande repositório.

Em resumo, destaca-se que os conhecimentos obtidos com os estudos realizados, com a metodologia aplicada e com as dificuldades vencidas em cada etapa do processo, formam uma base de conhecimento que pode ser aplicada novamente em futuros projetos de KDD. A realização desses projetos, dentro do cenário da indústria nacional, é importante para gerar um maior amadurecimento do assunto junto aos responsáveis pelos processos produtivos, e fazer com que a aplicação de técnicas de *Data Mining* e, de forma mais abrangente, a metodologia de KDD, venha a ser uma alternativa interessante para a extração de conhecimento das cada vez maiores bases de dados industriais.

---

distribuído, com mais de um computador executando as tarefas de leitura e movimentação de registros, devem levar a tarefa de geração dos *flat tables* a um ganho considerável de performance.

Apêndice A

Gráficos de Distribuição do Erro de Força por Tipo de Aço

As figuras A.1 a A.6 apresentam gráficos com a distribuição do erro de força para os tipos de aço mais significativos (2, 3, 5, 6, 9 e 10) da base de dados para cada uma das 6 cadeiras de laminação analisadas. Observa-se claramente que, em todos os tipos de aço, a faixa de erro de força **Ruim** vai aumentando da cadeira 1 até a cadeira 4 e 5, sofrendo uma pequena queda na cadeira 6. A faixa **OK** segue comportamento inverso, diminuindo a medida que a cadeira aumenta.

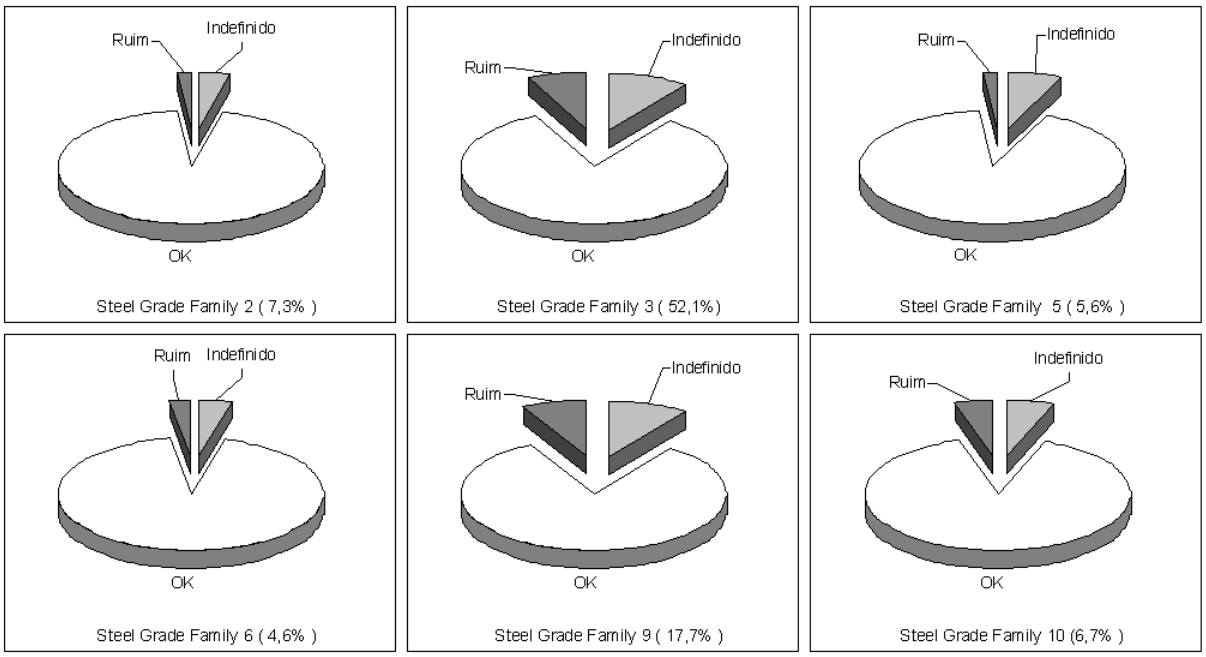

Figura A.1: Distribuição do Erro de Força por Família de Aço - Cadeira 1

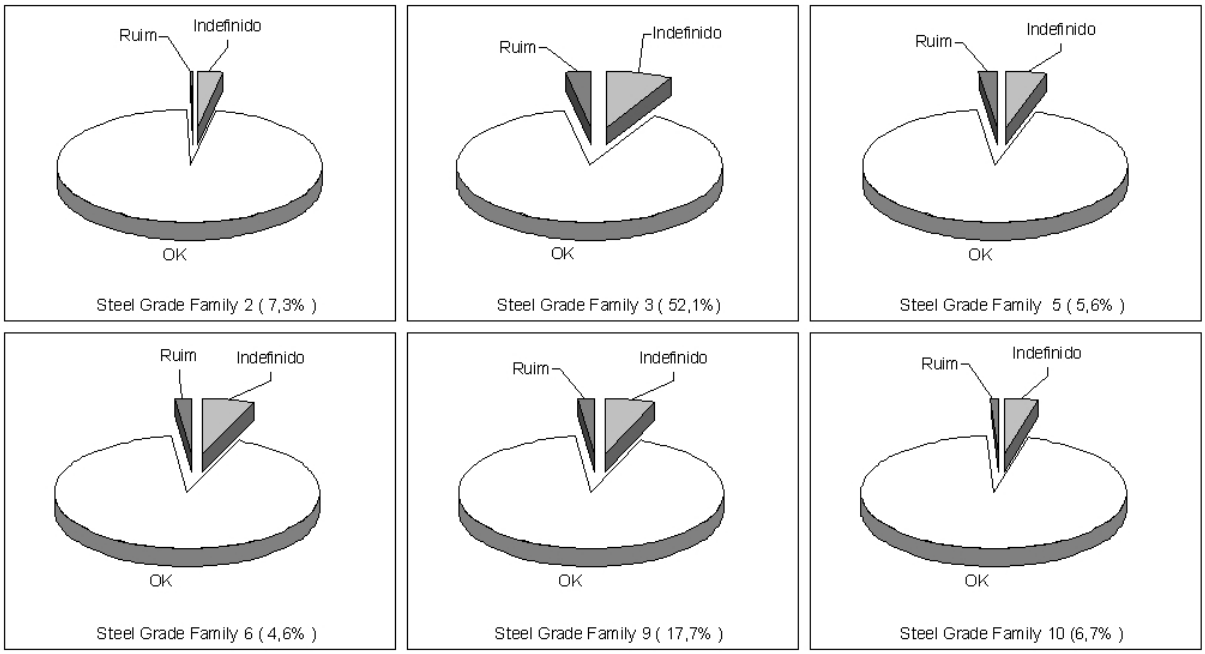

Figura A.2: Distribuição do Erro de Força por Família de Aço - Cadeira 2

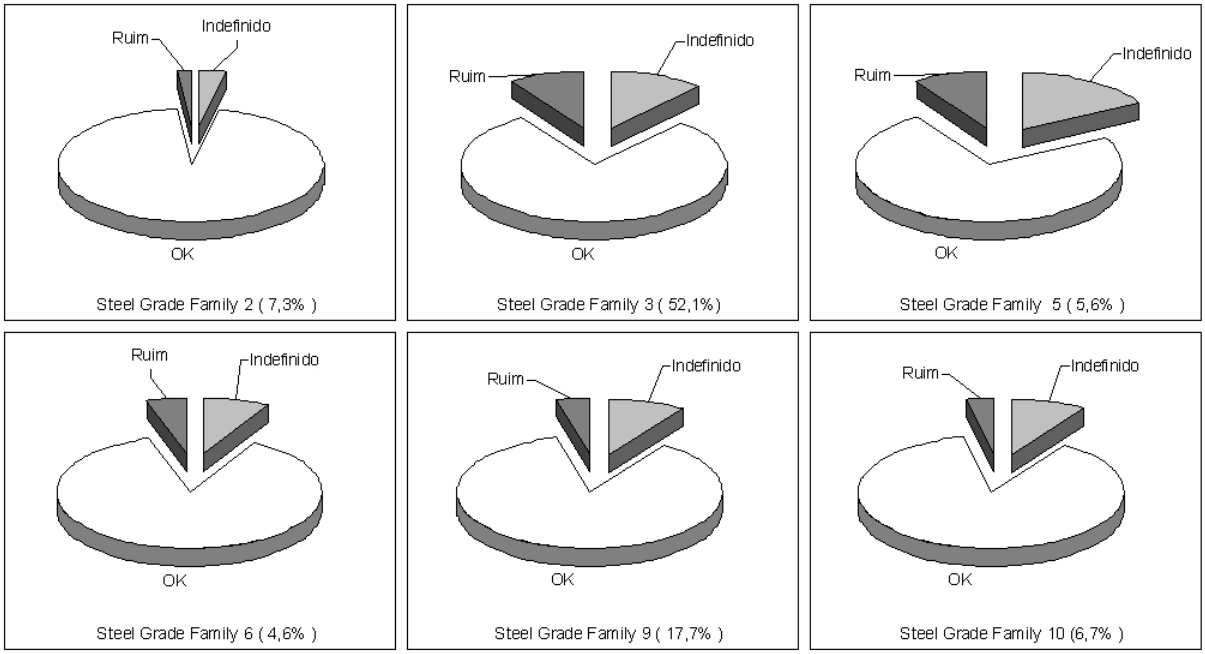

Figura A.3: Distribuição do Erro de Força por Família de Aço - Cadeira 3

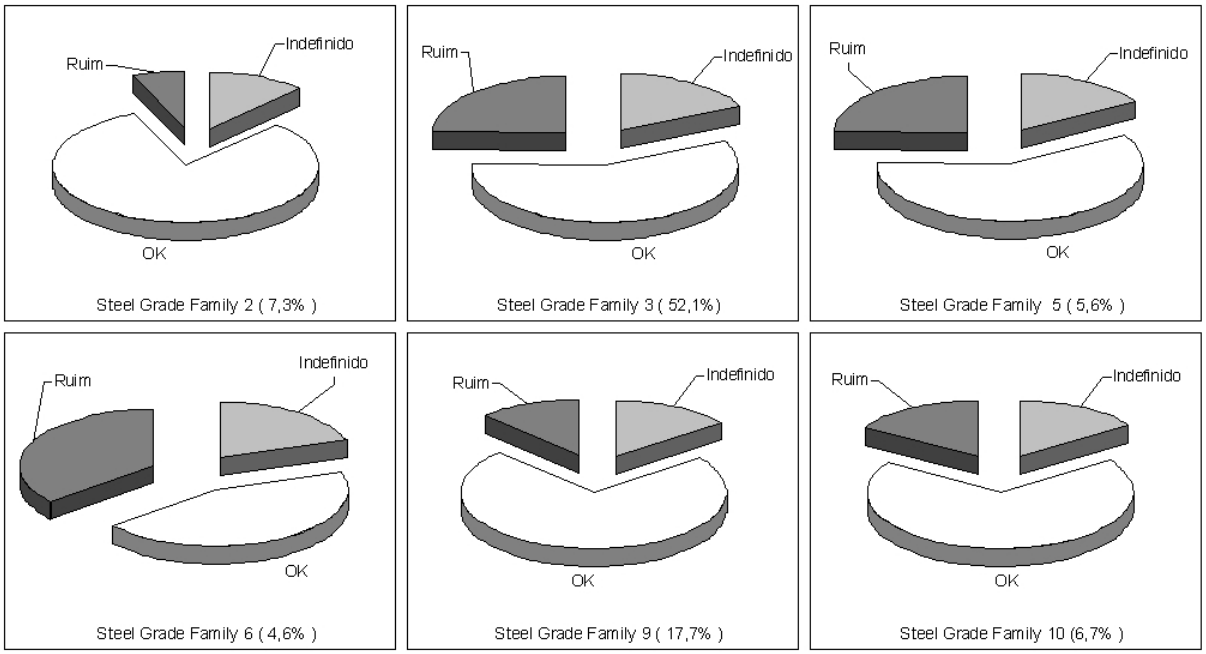

Figura A.4: Distribuição do Erro de Força por Família de Aço - Cadeira 4

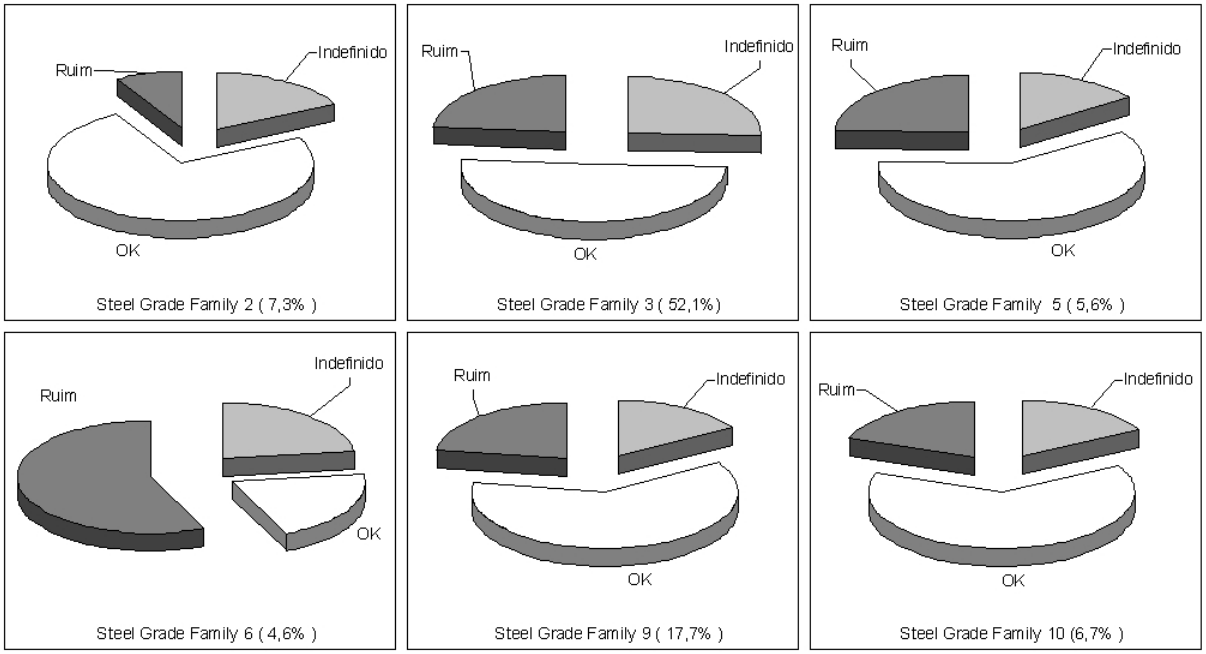

Figura A.5: Distribuição do Erro de Força por Família de Aço - Cadeira 5

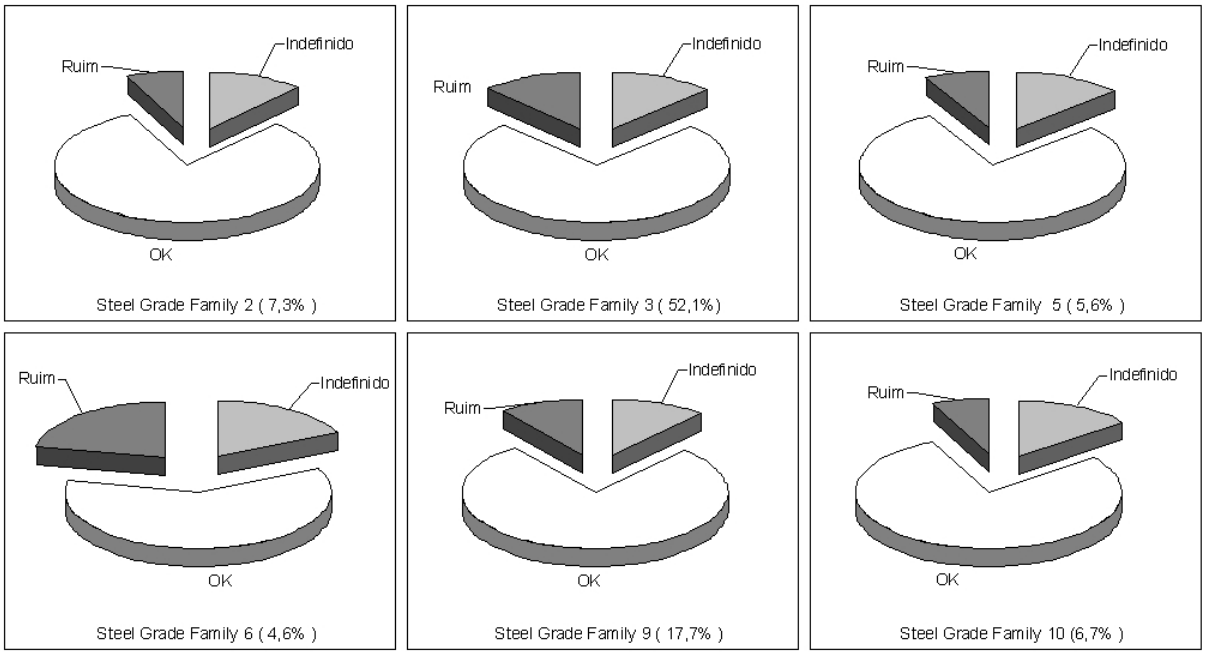

Figura A.6: Distribuição do Erro de Força por Família de Aço - Cadeira 6

Apêndice B

Tabelas de Importância de Variáveis

As tabelas de B.1 a B.3 apresentam a importância das 40 principais variáveis para classificação do erro de força. A implementação do *Statistica* para o algoritmo CART, além do modelo de classificação, gera também um lista de importância de cada um dos atributos da base de dados. Essa lista também pode ser obtida através de uma interpretação da árvore de decisão gerada pelo algoritmo; porém, dependendo do numero de nós, isso pode ser uma tarefa muito penosa.

Avaliando cada uma das tabelas, observa-se que algumas variáveis possuem uma nota alta para praticamente todos os tipos de aço em todas as cadeiras. Entre essas destacam-se:

- FOR_FLE_CIL - Força de Flexão dos Cilindros;
- VEL_LAM - Velocidade de Laminação;
- ERRO_REAL_RES_DEF - Erro de Resistência a Deformação (Valor medido menos o *setup*);
- DES_AXL_CIL - Deslocamento Axial dos Cilindros;
- CRA_TER_CAL - Coroa Térmica Calculada.

Variável	Todos	SGF 2	SGF 3	SGF 9	SGF 10
APR_FOR_BOB_F3	92	68	100	86	58
RES_DEF_MED_F3	91	54	98	73	57
RES_DEF_SET_F3	90	64	90	78	69
ESP_MED_PAS	90	83	88	91	55
ERRO_REAL_RES_DEF_F3	97	67	87	78	77
ERRO_CAL_FOR_LAM_F3	84	59	87	72	52
VEL_LAM_F3	83	71	86	87	62
VAZ_H2O_F3F4	86	55	83	72	63
CAR_EQV	86	63	82	73	44
SET_ESP_F6	80	78	82	72	57
RED_ESP_F3	94	47	81	89	78
FOR_FLE_CIL_F3	100	65	80	93	100
COR_GAP_F3	79	38	80	66	50
APR_FOR_ESP_F3	94	65	80	63	31
RES_DEF_APR_F3	79	100	80	55	54
RES_DEF_CAL_F3	86	45	79	100	58
TEM_TRM_ACA_IN	86	79	79	79	68
DES_AXL_CIL_F3	79	65	78	74	48
CMP_QUI_N	71	43	75	50	50
VAZ_H2O_F2F3	87	80	74	72	40
ERR_ESP_F6	86	74	74	71	84
CMP_QUI_AL	71	55	74	59	36
CMP_QUI_S	79	66	73	62	31
CMP_QUI_AE	78	49	73	55	40
CMP_QUI_C	74	53	72	55	47
CMP_QUI_MN	78	72	71	75	40
FOR_LAM_CAL_F3	82	80	70	78	81
CRA_TER_CAL_F3	91	42	67	86	50
SET_LAR_F6	75	86	67	83	57
CMP_QUI_NI	67	22	66	64	31
CMP_QUI_CR	78	66	64	68	30
ERRO_CAL_RES_DEF_F3	86	67	64	64	68
CMP_QUI_SI	58	35	61	53	31
CMP_QUI_AS	67	26	56	49	31
COR_DES_F3	78	83	54	49	49
DIS_FOR_MED_F3	54	21	54	32	21
CMP_QUI_CU	77	54	51	68	41
CMP_QUI_P	80	76	50	45	45
CAT_ERRO_REAL_RES_DEF_F3	47	64	42	43	47
CAT_ERRO_CAL_FOR_LAM_F3	37	21	41	30	17

Tabela B.1: Tabela de Importância das Variáveis para Cadeira 3

Variável	Todos	SGF 2	SGF 3	SGF 9	SGF 10
ERRO_REAL_RES_DEF_F4	100	100	100	90	100
ERRO_CAL_FOR_LAM_F4	95	64	95	73	86
RES_DEF_CAL_F4	95	72	94	90	90
RED_ESP_F4	95	79	93	100	69
VEL_LAM_F4	91	82	92	92	96
FOR_FLE_CIL_F4	100	93	91	88	94
APR_FOR_BOB_F4	96	64	91	83	85
ESP_MED_PAS	92	69	90	86	87
SET_ESP_F6	87	49	89	80	75
ERRO_CAL_RES_DEF_F4	96	59	89	79	76
FOR_LAM_CAL_F4	98	39	88	77	80
RES_DEF_APR_F4	86	50	85	76	72
COR_DES_F4	88	69	85	68	84
RES_DEF_MED_F4	100	80	84	95	84
DES_AXL_CIL_F4	93	53	84	76	84
CRA_TER_CAL_F4	100	54	84	65	85
APR_FOR_ESP_F4	94	69	84	64	76
TEM_TRM_ACA_IN	92	53	83	81	73
RES_DEF_SET_F4	94	71	83	76	97
SET_LAR_F6	85	73	82	65	49
VAZ_H2O_F3F4	84	83	81	78	65
ERR_ESP_F6	92	67	81	70	61
COR_GAP_F4	94	42	77	86	69
CMP_QUI_S	84	69	77	77	85
CAR_EQV	90	60	75	80	52
CMP_QUI_C	86	71	75	70	66
CMP_QUI_MN	88	55	72	70	54
CMP_QUI_CR	81	33	70	61	77
CMP_QUI_CU	79	49	69	72	69
CMP_QUI_AE	94	48	68	62	88
CMP_QUI_P	73	79	68	58	68
CMP_QUI_AL	90	74	64	64	88
CMP_QUI_N	76	64	64	57	62
CMP_QUI_SI	66	28	63	67	60
CMP_QUI_AS	68	65	60	58	34
DIS_FOR_MED_F4	63	16	57	61	31
CMP_QUI_NI	71	39	55	60	55
CTR_FLX_MAS_F4	43	10	46	52	46
CMP_QUI_CA	44	25	38	50	18
CAT_ERRO_CAL_RES_DEF_F4	49	46	38	40	41

Tabela B.2: Tabela de Importância das Variáveis para Cadeira 4

Variável	Todos	SGF 2	SGF 3	SGF 9	SGF 10
FOR_FLE_CIL_F5	100	89	100	100	80
RES_DEF_MED_F5	97	100	99	89	100
RES_DEF_CAL_F5	96	76	96	80	72
APR_FOR_BOB_F5	97	66	95	67	51
ERRO_CAL_FOR_LAM_F5	99	68	94	87	66
ERRO_CAL_RES_DEF_F5	96	86	94	64	63
ERRO_REAL_RES_DEF_F5	98	74	93	91	71
RES_DEF_SET_F5	98	74	93	78	84
VEL_LAM_F5	93	77	92	75	77
FOR_LAM_CAL_F5	95	73	89	86	63
ESP_MED_PAS	91	70	89	70	62
APR_FOR_ESP_F5	97	83	88	75	72
RED_ESP_F5	100	67	87	78	89
RES_DEF_APR_F5	86	96	86	69	58
ERR_ESP_F6	89	64	84	71	69
COR_GAP_F5	91	76	84	64	64
CAR_EQV	99	69	81	73	57
SET_ESP_F6	92	55	80	69	50
CRA_TER_CAL_F5	90	66	79	78	56
DES_AXL_CIL_F5	95	66	79	76	67
CMP_QUI_S	88	60	79	64	62
TEM_TRM_ACA_IN	93	72	78	70	52
CMP_QUI_MN	95	67	77	71	67
CMP_QUI_P	75	34	77	58	46
CMP_QUI_AE	71	50	76	59	63
CMP_QUI_C	98	67	75	74	53
COR_DES_F5	90	55	74	79	62
CMP_QUI_CR	84	37	73	65	51
CMP_QUI_CU	73	40	73	51	47
CMP_QUI_N	83	62	72	71	59
CMP_QUI_AL	75	40	72	64	44
SET_LAR_F6	84	66	64	58	42
CMP_QUI_NI	66	37	62	44	38
CMP_QUI_SI	69	40	60	55	51
CMP_QUI_AS	77	42	60	52	35
DIS_FOR_MED_F5	65	30	58	50	43
CMP_QUI_CA	50	23	46	44	11
CMP_QUI_MO	55	23	46	38	21
CMP_QUI_NB	52	53	41	36	7
CTR_FLX_MAS_F5	51	20	35	41	32

Tabela B.3: Tabela de Importância das Variáveis para Cadeira 5

Apêndice C

Análise Estatística das Variáveis por Faixa do Erro de Força

As tabelas de C.1 a C.3 apresentam uma análise estatística com valores médios, máximos, mínimos e desvio padrão, para cada uma das variáveis selecionadas, por cadeira e por faixa de erro de força ("OK", "Indefinido" e "Ruim").

ERRO DE FORÇA DE LAMINAÇÃO - em percentual				
Erro de Força	**Média**	**Máximo**	**Mínimo**	**Desvio Padrão**
OK	3,85	9,99	0,00	2,56
Indefinido	12,15	15,00	10,00	1,47
Ruim	20,54	92,37	15,02	6,36
ERRO DE RESISTÊNCIA À DEFORMAÇÃO - em percentual				
Erro de Força	**Média**	**Máximo**	**Mínimo**	**Desvio Padrão**
OK	8,73	29,13	0,00	5,19
Indefinido	11,85	29,20	0,03	6,37
Ruim	15,61	73,32	0,13	8,32
VELOCIDADE DE LAMINAÇÃO - em m/s				
Erro de Força	**Média**	**Máximo**	**Mínimo**	**Desvio Padrão**
OK	12,96	17,68	4,78	2,46
Indefinido	12,99	17,33	4,76	2,18
Ruim	13,17	17,77	5,45	2,03
COROA TÉRMICA CALCULADA - em microns				
Erro de Força	**Média**	**Máximo**	**Mínimo**	**Desvio Padrão**
OK	246,01	415,00	0,00	87,94
Indefinido	223,54	400,00	0,00	102,88
Ruim	192,76	381,00	0,00	117,56
DESLOCAMENTO AXIAL DOS CILINDROS - em mm				
Erro de Força	**Média**	**Máximo**	**Mínimo**	**Desvio Padrão**
OK	-18,54	150,00	-150,01	43,57
Indefinido	-8,89	96,09	-150,00	44,90
Ruim	1,53	125,02	-128,54	44,38
FORÇA DE FLEXÃO DOS CILINDROS - em KN				
Erro de Força	**Média**	**Máximo**	**Mínimo**	**Desvio Padrão**
OK	721,12	1184,51	134,58	117,78
Indefinido	632,87	1282,65	115,75	281,79
Ruim	685,27	1408,25	327,68	369,26

Tabela C.1: Análise Estatística por Faixa do Erro de Força - Cadeira 3

ERRO DE FORÇA DE LAMINAÇÃO - em percentual				
Erro de Força	**Média**	**Máximo**	**Mínimo**	**Desvio Padrão**
OK	4,47	10,00	0,00	2,86
Indefinido	12,32	15,00	10,00	1,44
Ruim	23,01	104,27	15,02	7,02
ERRO DE RESISTÊNCIA À DEFORMAÇÃO - em percentual				
Erro de Força	**Média**	**Máximo**	**Mínimo**	**Desvio Padrão**
OK	5,05	26,45	0,00	3,58
Indefinido	6,92	27,52	0,00	4,67
Ruim	12,28	80,46	0,01	7,25
VELOCIDADE DE LAMINAÇÃO - em m/s				
Erro de Força	**Média**	**Máximo**	**Mínimo**	**Desvio Padrão**
OK	17,79	24,99	5,68	4,24
Indefinido	19,49	25,21	6,10	3,51
Ruim	20,03	24,98	5,84	3,06
COROA TÉRMICA CALCULADA - em microns				
Erro de Força	**Média**	**Máximo**	**Mínimo**	**Desvio Padrão**
OK	205,88	383,00	0,00	79,12
Indefinido	219,99	360,00	0,00	82,98
Ruim	229,12	366,00	0,00	81,44
DESLOCAMENTO AXIAL DOS CILINDROS - em mm				
Erro de Força	**Média**	**Máximo**	**Mínimo**	**Desvio Padrão**
OK	-22,82	150,01	-150,02	75,31
Indefinido	-15,63	150,00	-150,01	74,64
Ruim	0,98	150,01	-150,01	75,83
FORÇA DE FLEXÃO DOS CILINDROS - em KN				
Erro de Força	**Média**	**Máximo**	**Mínimo**	**Desvio Padrão**
OK	736,23	1494,28	101,50	156,04
Indefinido	637,84	1500,70	99,94	222,16
Ruim	583,82	1500,79	97,58	322,63

Tabela C.2: Análise Estatística por Faixa do Erro de Força - Cadeira 4

ERRO DE FORÇA DE LAMINAÇÃO - em percentual				
Erro de Força	**Média**	**Máximo**	**Mínimo**	**Desvio Padrão**
OK	4,76	9,99	0,00	2,92
Indefinido	12,42	15,00	10,01	1,41
Ruim	21,87	93,05	15,01	7,37
ERRO DE RESISTÊNCIA À DEFORMAÇÃO - em percentual				
Erro de Força	**Média**	**Máximo**	**Mínimo**	**Desvio Padrão**
OK	6,21	35,31	0,00	5,00
Indefinido	6,73	30,94	0,00	5,21
Ruim	8,46	55,84	0,00	7,85
VELOCIDADE DE LAMINAÇÃO - em m/s				
Erro de Força	**Média**	**Máximo**	**Mínimo**	**Desvio Padrão**
OK	22,80	30,99	6,60	6,01
Indefinido	26,36	30,69	6,55	4,51
Ruim	26,66	31,42	6,65	4,83
COROA TÉRMICA CALCULADA - em microns				
Erro de Força	**Média**	**Máximo**	**Mínimo**	**Desvio Padrão**
OK	164,55	314,00	0,00	65,42
Indefinido	199,29	296,00	0,00	57,97
Ruim	203,93	326,00	0,00	51,61
DESLOCAMENTO AXIAL DOS CILINDROS - em mm				
Erro de Força	**Média**	**Máximo**	**Mínimo**	**Desvio Padrão**
OK	-6,00	150,01	-150,01	70,72
Indefinido	23,20	150,00	-150,00	64,87
Ruim	28,03	150,00	-150,01	66,60
FORÇA DE FLEXÃO DOS CILINDROS - em KN				
Erro de Força	**Média**	**Máximo**	**Mínimo**	**Desvio Padrão**
OK	748,12	1350,24	126,11	145,86
Indefinido	632,98	1138,15	101,97	187,85
Ruim	561,21	1484,04	99,70	199,25

Tabela C.3: Análise Estatística por Faixa do Erro de Força - Cadeira 5

Apêndice D

Gráficos das Variáveis Selecionadas

Esta seção apresenta os gráficos das 5 variáveis selecionadas como as mais relevantes dentro do processo de análise do erro de força de laminação. As mesmas se encontram "plotadas" levando em consideração apenas os dados de produção das famílias de aço 3 e 9 para as Cadeiras 3, 4 e 5. Todos os gráficos dessa seção estão ordenados pelo erro de força de laminação da Cadeira correspondente e, além dos dados originais, apresentam também os dados com a aplicação do filtro de média móvel com uma janela de 100 registros.

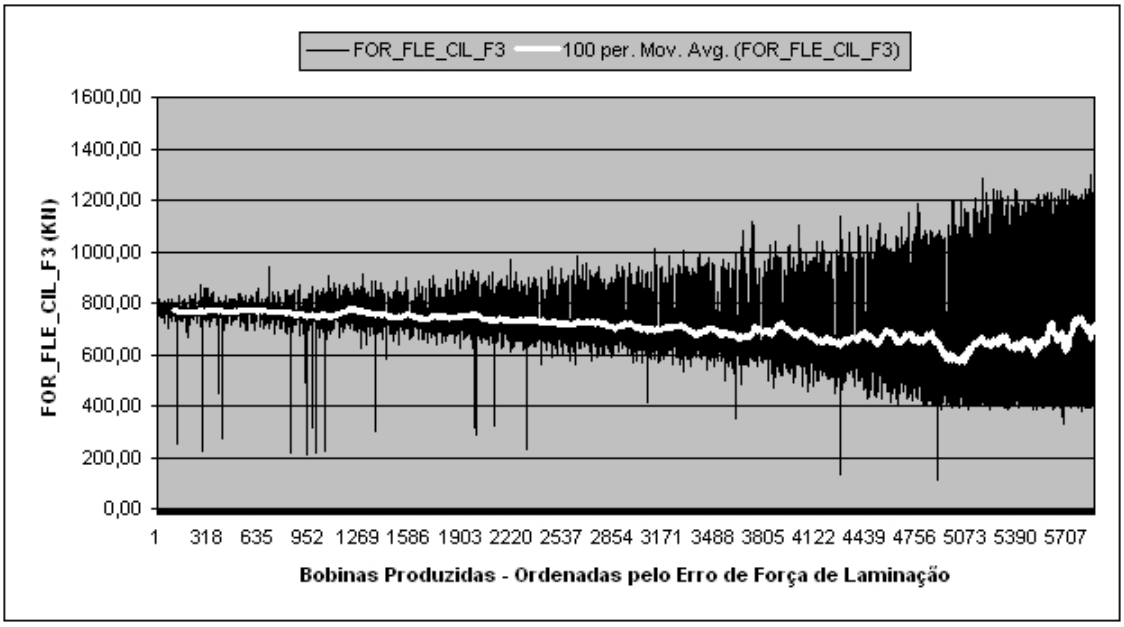

Figura D.1: Gráfico da Força de Flexão do Cilindros com e sem Filtro de Média Móvel (100 pontos) - Cadeira 3, Famílias de Aço 3 e 9

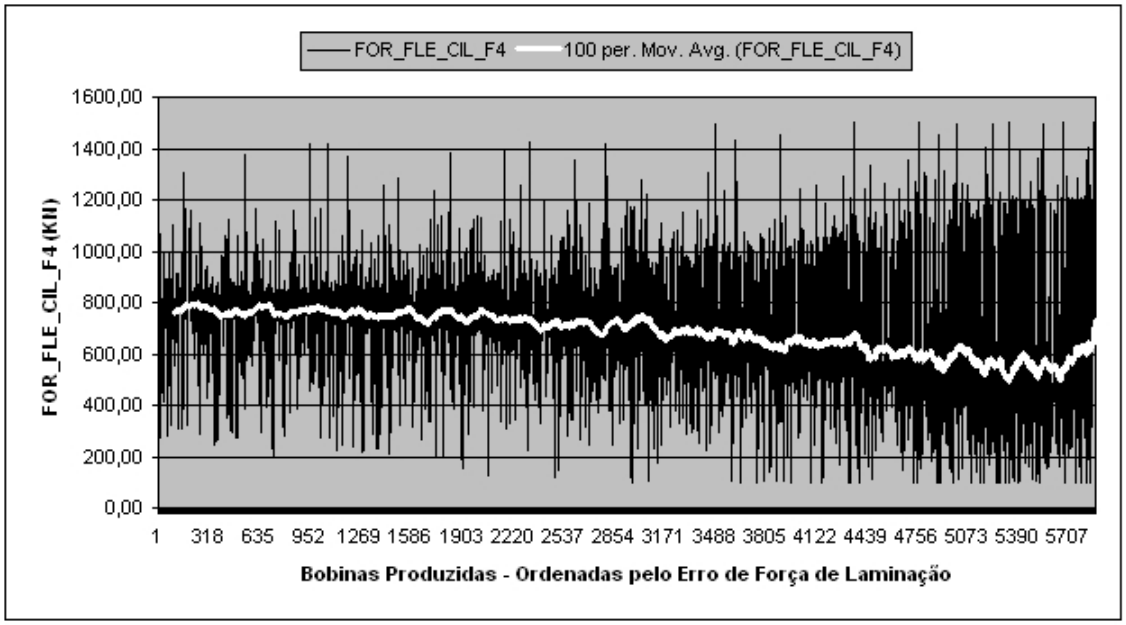

Figura D.2: Gráfico da Força de Flexão do Cilindros com e sem Filtro de Média Móvel (100 pontos) - Cadeira 4, Famílias de Aço 3 e 9

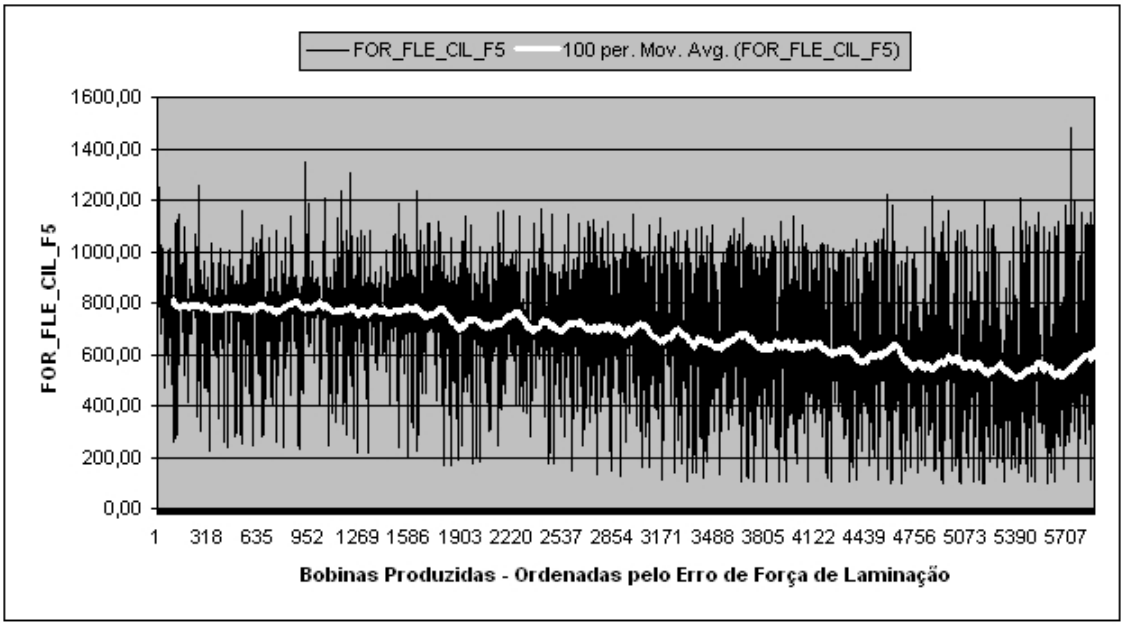

Figura D.3: Gráfico da Força de Flexão do Cilindros com e sem Filtro de Média Móvel (100 pontos) - Cadeira 5, Famílias de Aço 3 e 9

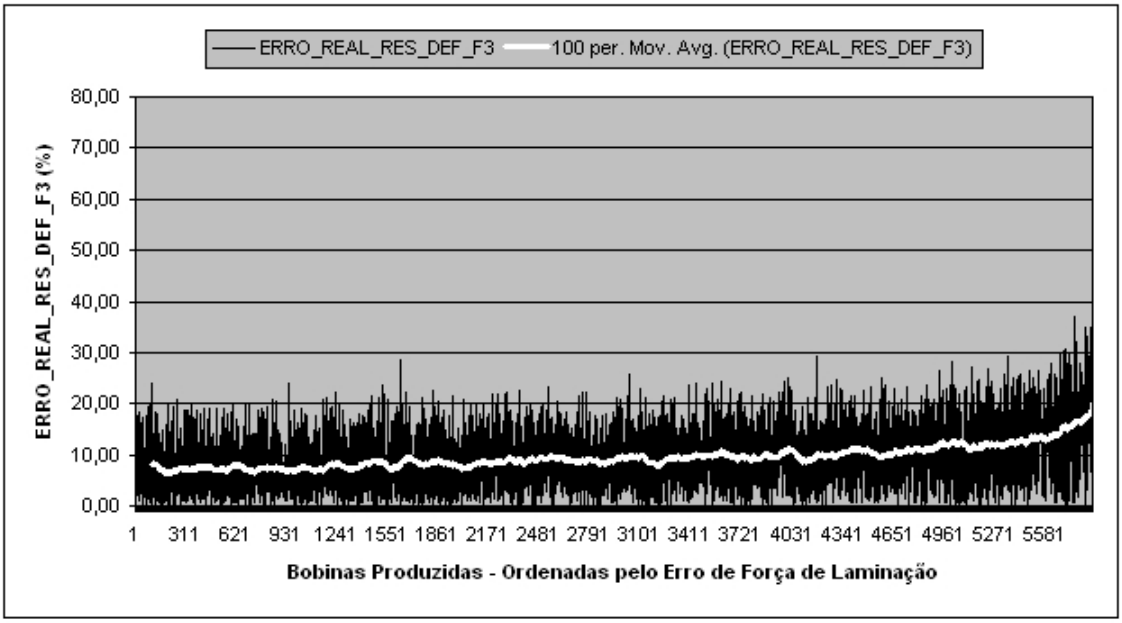

Figura D.4: Gráfico do Erro de Resistência a Deformação com e sem Filtro de Média Móvel (100 pontos) - Cadeira 3, Famílias de Aço 3 e 9

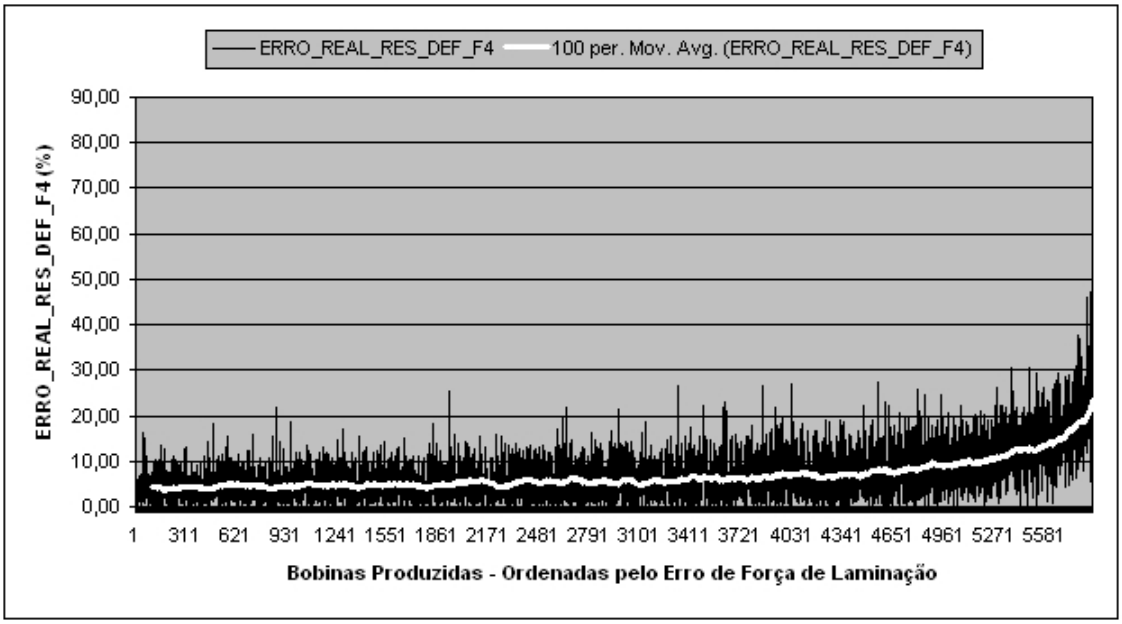

Figura D.5: Gráfico do Erro de Resistência a Deformação com e sem Filtro de Média Móvel (100 pontos) - Cadeira 4, Famílias de Aço 3 e 9

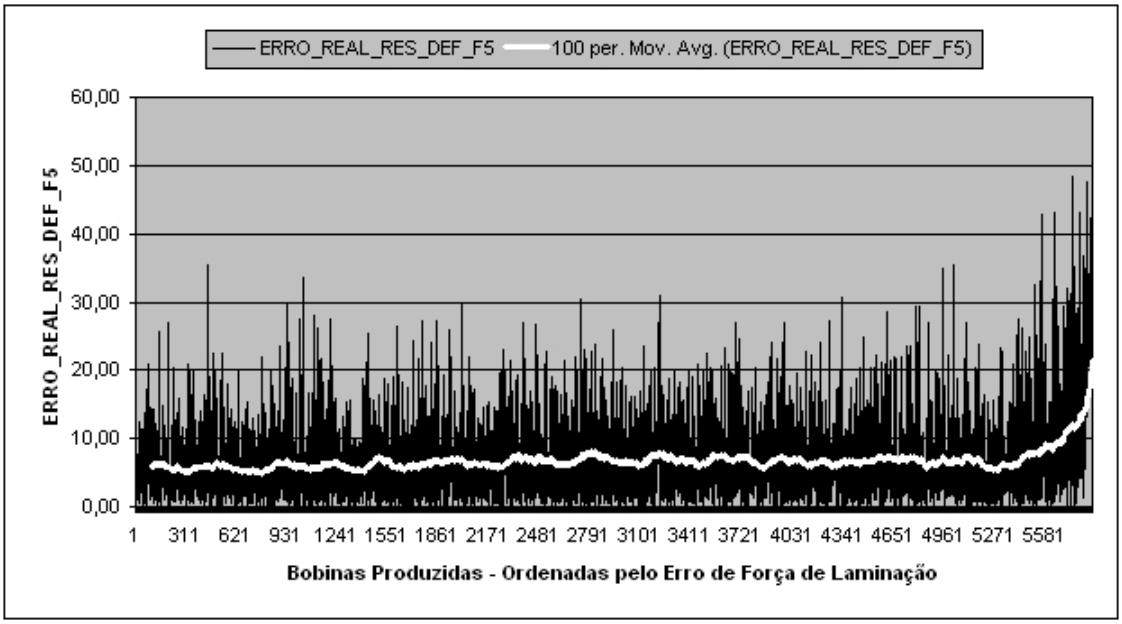

Figura D.6: Gráfico do Erro de Resistência a Deformação com e sem Filtro de Média Móvel (100 pontos) - Cadeira 5, Famílias de Aço 3 e 9

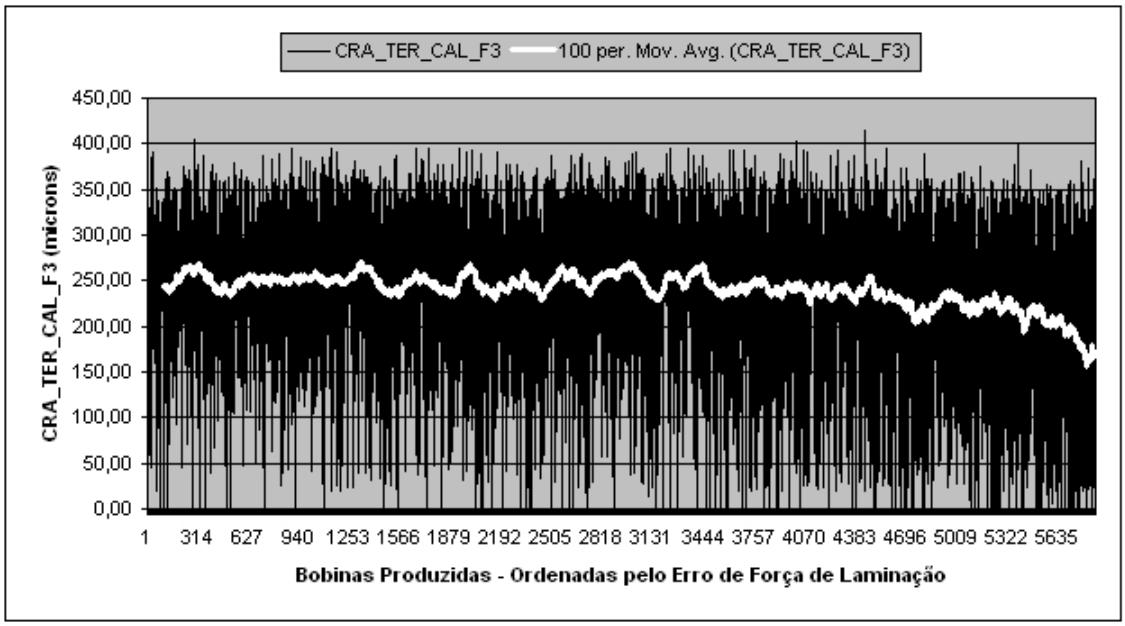

Figura D.7: Gráfico da Coroa Térmica Calculada com e sem Filtro de Média Móvel (100 pontos) - Cadeira 3, Famílias de Aço 3 e 9

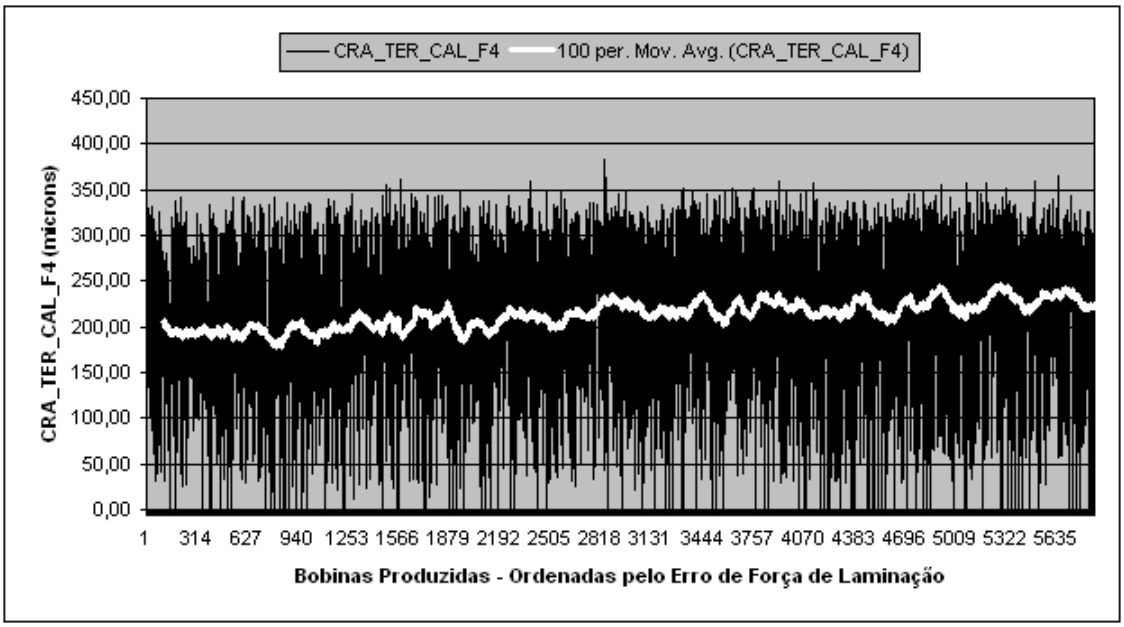

Figura D.8: Gráfico da Coroa Térmica Calculada com e sem Filtro de Média Móvel (100 pontos) - Cadeira 4, Famílias de Aço 3 e 9

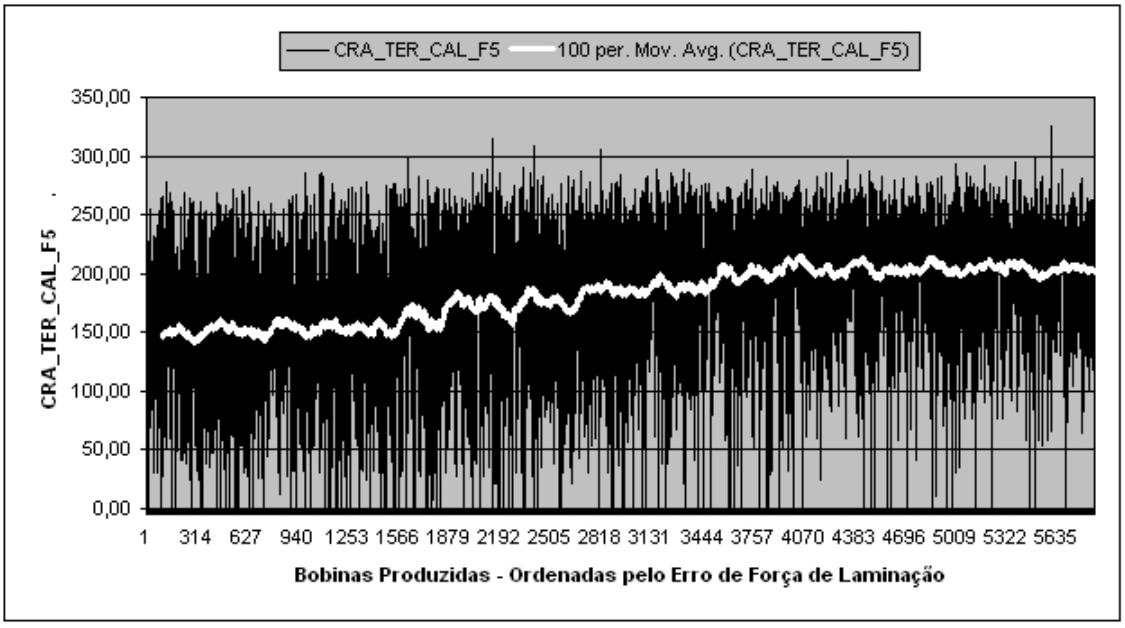

Figura D.9: Gráfico da Coroa Térmica Calculada com e sem Filtro de Média Móvel (100 pontos) - Cadeira 5, Famílias de Aço 3 e 9

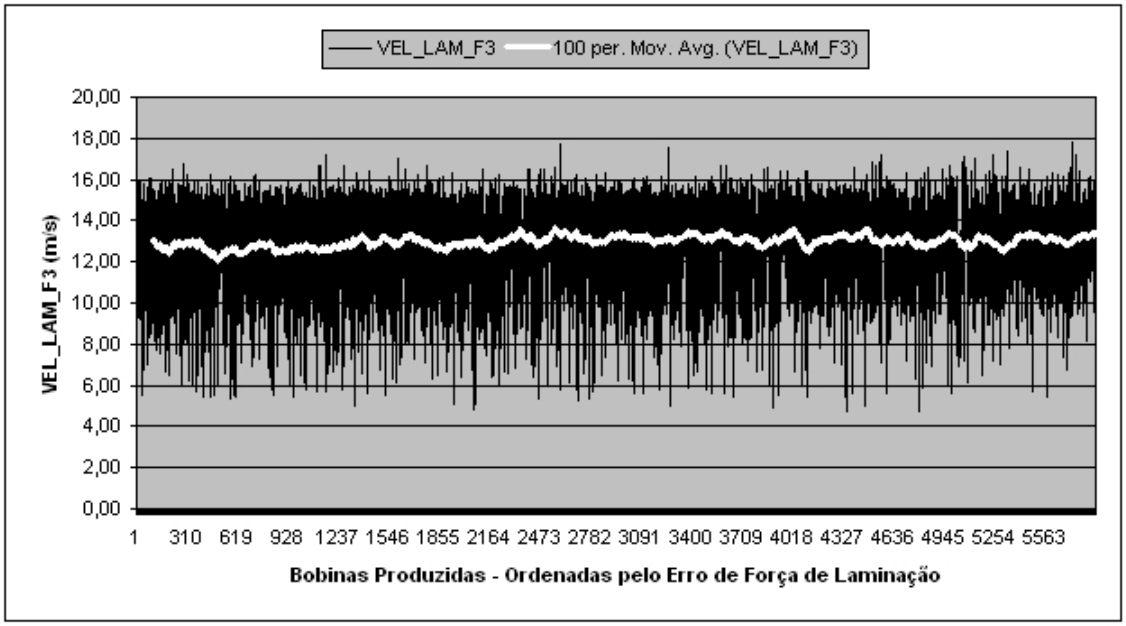

Figura D.10: Gráfico da Velocidade de Laminação com e sem Filtro de Média Móvel (100 pontos) - Cadeira 3, Famílias de Aço 3 e 9

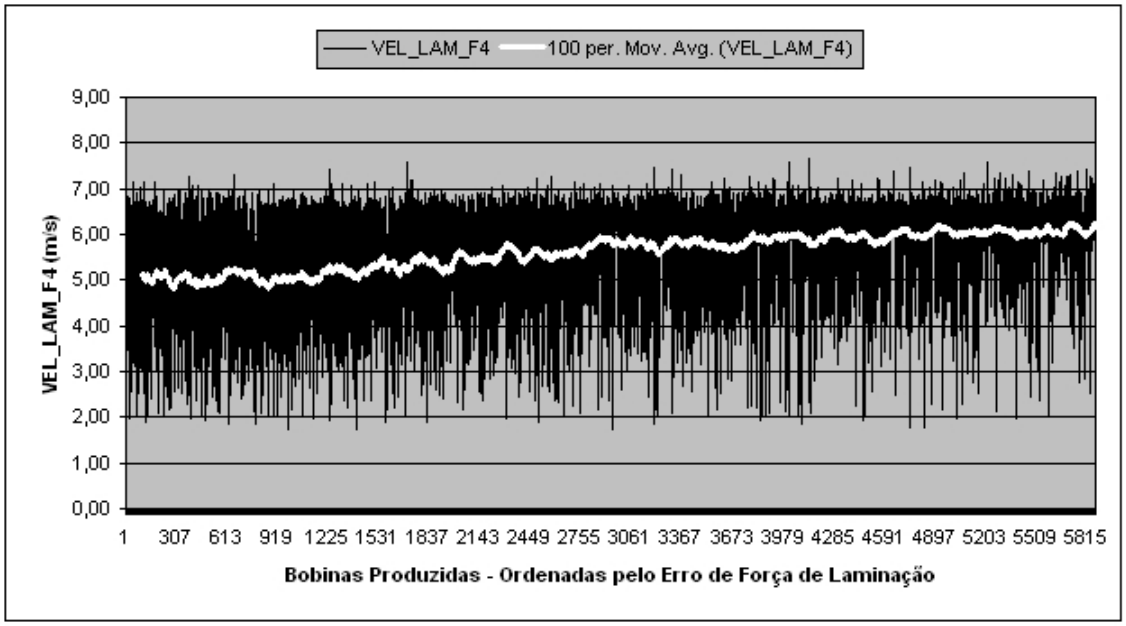

Figura D.11: Gráfico da Velocidade de Laminação com e sem Filtro de Média Móvel (100 pontos) - Cadeira 4, Famílias de Aço 3 e 9

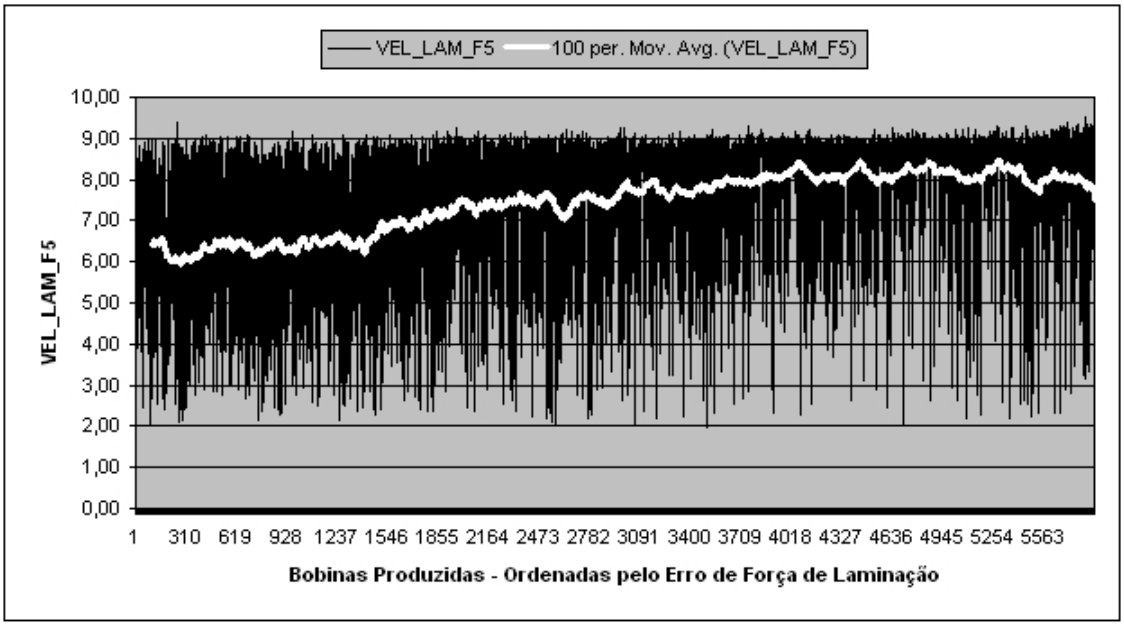

Figura D.12: Gráfico da Velocidade de Laminação com e sem Filtro de Média Móvel (100 pontos) - Cadeira 5, Famílias de Aço 3 e 9

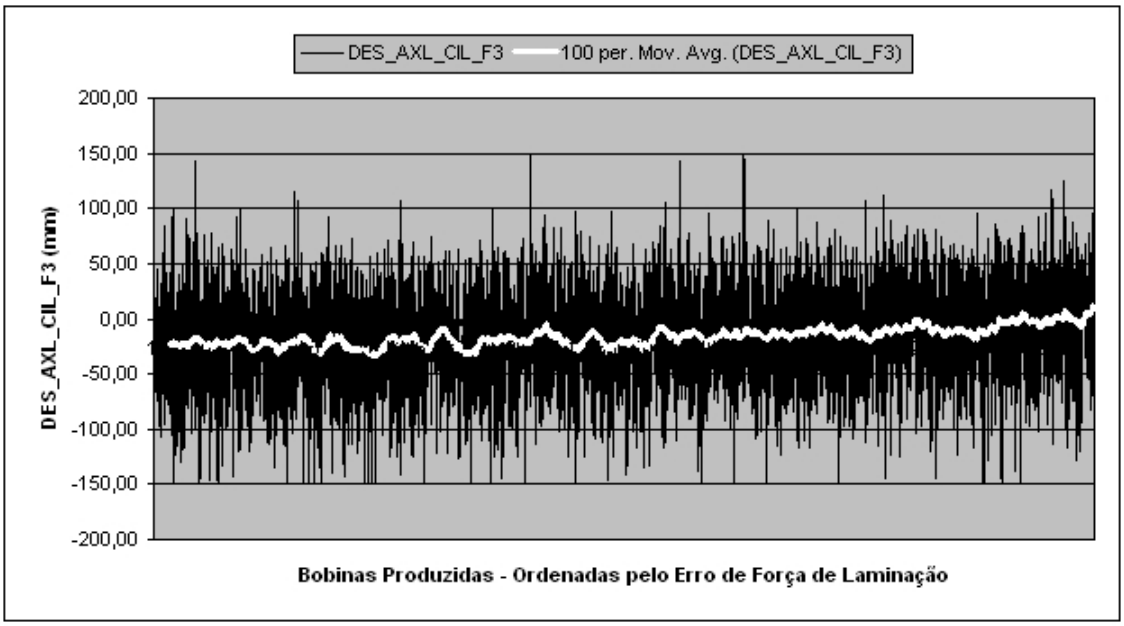

Figura D.13: Gráfico do Deslocamento Axial dos Cilindros com e sem Filtro de Média Móvel (100 pontos) - Cadeira 3, Famílias de Aço 3 e 9

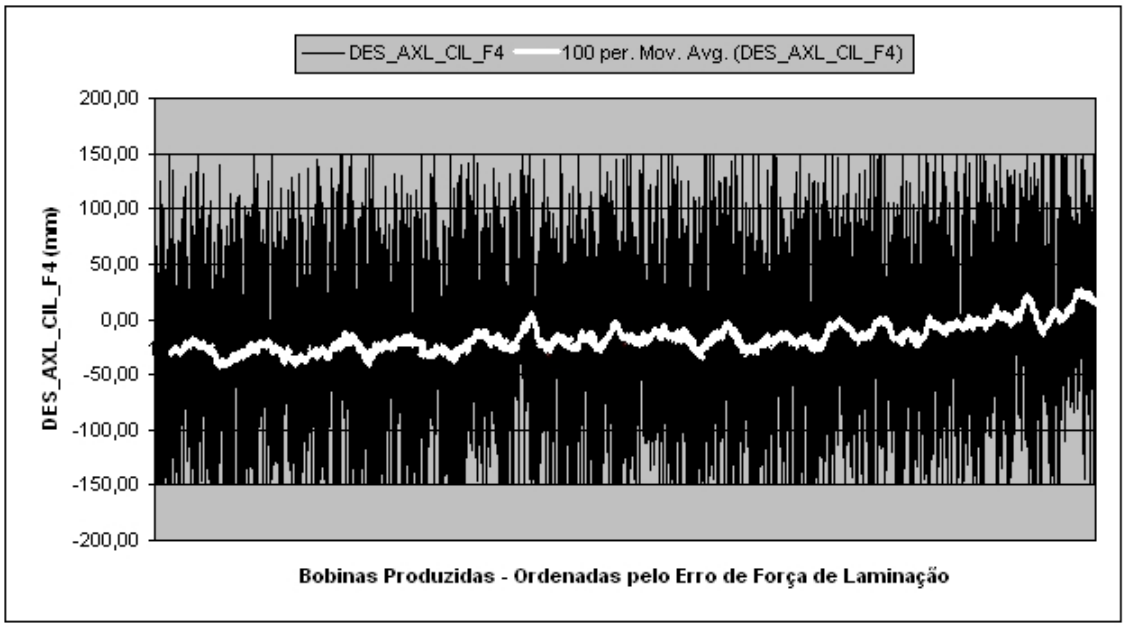

Figura D.14: Gráfico do Deslocamento Axial dos Cilindros com e sem Filtro de Média Móvel (100 pontos) - Cadeira 4, Famílias de Aço 3 e 9

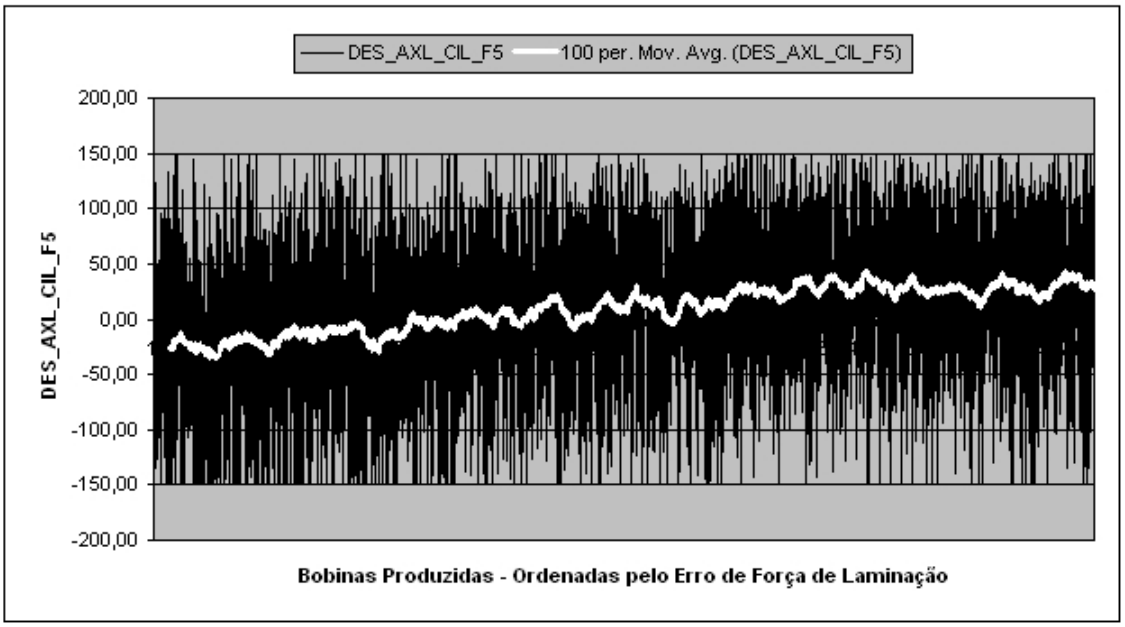

Figura D.15: Gráfico do Deslocamento Axial dos Cilindros com e sem Filtro de Média Móvel (100 pontos) - Cadeira 5, Famílias de Aço 3 e 9

Apêndice E

Regras de Associação Geradas

Na seção 4.5.1 foram apresentadas as regras de associação geradas com grau de confiança superior a 90%. Nesse apêndice apresentam-se mais regras com percentual de confiança entre 80% e 90%. Além dessas, mais regras foram geradas pelo *Statistica*, e as mesmas podem até apresentar algum resultado significativo, porém, não foram consideradas confiáveis (grau de confiança menor que 80%). A tabela E.1 apresenta a legenda para interpretação dessas regras.

Termo	**Significado**
OK	Erro de Força de Laminação menor que A%
Indefinido	Erro de Força de Laminação entre A% e B%
Ruim	Erro de Força de Laminação maior que B%
ERD_BAIXO	Erro de Resistência à Deformação Baixo
ERD_MÉDIO	Erro de Resistência à Deformação Médio
ERD_ALTO	Erro de Resistência à Deformação Alto
VL_BAIXO	Velocidade de Laminação Baixa
VL_MÉDIO	Velocidade de Laminação Média
VL_ALTO	Velocidade de Laminação Alta
CT_BAIXO	Coroa Térmica Calculada Baixa
CT_MÉDIO	Coroa Térmica Calculada Média
CT_ALTO	Coroa Térmica Calculada Alta
DXL_BAIXO	Deslocamento Axial dos Cilindros Baixo
DXL_MÉDIO	Deslocamento Axial dos Cilindros Médio
DXL_ALTO	Deslocamento Axial dos Cilindros Alto
FFC_BAIXO	Força de Flexão dos Cilindros Baixa
FFC_MÉDIO	Força de Flexão dos Cilindros Média
FFC_ALTO	Força de Flexão dos Cilindros Alta

Tabela E.1: Legenda para interpretação das Regras de Associação

Se	==>	Então	Confiança(%)	Cobertura(%)
OK, VL_BAIXA, CT_ALTA	==>	ERD_BAIXO	89,80	11,24
ERD_BAIXO, CT_ALTA	==>	OK	89,76	33,01
FFC_ALTA, VL_ALTA	==>	OK	89,46	28,63
ERD_BAIXO, FFC_ALTA, DXL_ALTA	==>	OK	89,46	14,02
ERD_BAIXO, VL_BAIXA, CT_BAIXA	==>	OK	89,42	11,38
ERD_BAIXO, FFC_ALTA, VL_ALTA, CT_ALTA	==>	OK	89,30	14,36
ERD_BAIXO, VL_ALTA, CT_ALTA	==>	OK	89,28	21,41
ERD_BAIXO	==>	OK	89,13	51,80
ERD_BAIXO, FFC_ALTA, CT_BAIXA	==>	OK	88,83	11,51
ERD_BAIXO, VL_ALTA	==>	OK	87,91	25,89
ERD_BAIXO, VL_BAIXA, DXL_ALTA	==>	OK	87,90	10,27
VL_BAIXA, FFC_ALTA, CT_BAIXA	==>	OK	87,86	11,46
VL_BAIXA, DXL_BAIXA	==>	OK	87,50	16,57
ERD_BAIXO, FFC_ALTA, VL_ALTA	==>	OK	87,11	17,49
ERD_BAIXO, CT_BAIXA	==>	OK	87,00	14,94
ERD_BAIXO, CT_ALTA, FFC_BAIXA	==>	OK	86,78	10,07
FFC_ALTA, CT_BAIXA	==>	OK	86,43	17,36
ERD_BAIXO, DXL_ALTA	==>	OK	86,19	19,88
VL_BAIXA, CT_ALTA	==>	OK	86,17	12,52
FFC_ALTA, DXL_ALTA	==>	OK	85,93	18,21
CT_ALTA, DXL_BAIXA	==>	OK	85,80	31,29
DXL_BAIXA	==>	OK	85,52	44,25
ERD_BAIXO, FFC_BAIXA	==>	OK	85,48	14,24
VL_ALTA, CT_ALTA, DXL_BAIXA	==>	OK	85,40	22,81
OK, VL_BAIXA, FFC_ALTA	==>	ERD_BAIXO	85,17	19,76
OK, VL_BAIXA, FFC_ALTA, DXL_BAIXA	==>	ERD_BAIXO	84,96	10,97

Tabela E.2: Demais Regras de Associação Geradas para as variáveis da Cadeira 3.

Se	**==>**	**Então**	**Confiança(%)**	**Cobertura(%)**
ERD_BAIXO, FFC_ALTA	==>	OK	89,83	27,69
VL_BAIXA, FFC_ALTA, DXL_BAIXA	==>	OK	89,69	12,60
ERD_BAIXO, DXL_ALTA, FFC_ALTA	==>	OK	89,22	11,99
VL_BAIXA, CT_BAIXA, FFC_ALTA	==>	OK	87,75	17,93
CT_BAIXA, FFC_ALTA, DXL_BAIXA	==>	OK	87,68	12,12
OK, ERD_BAIXO, VL_BAIXA, CT_BAIXA	==>	FFC_ALTA	87,65	11,12
VL_BAIXA, FFC_ALTA	==>	OK	87,39	22,07
ERD_BAIXO, CT_BAIXA	==>	OK	86,11	18,90
ERD_BAIXO, FFC_ALTA, VL_ALTA	==>	OK	85,80	11,83
ERD_BAIXO, VL_BAIXA, CT_BAIXA	==>	FFC_ALTA	85,71	11,75
ERD_BAIXO, FFC_ALTA, CT_ALTA	==>	OK	85,32	11,29
CT_BAIXA, FFC_ALTA	==>	OK	84,64	24,13
ERD_BAIXO, DXL_ALTA, CT_ALTA	==>	VL_ALTA	83,40	10,27
FFC_ALTA, DXL_BAIXA	==>	OK	83,33	21,46
VL_BAIXA, CT_BAIXA, DXL_BAIXA	==>	OK	83,19	11,88
VL_BAIXA, DXL_ALTA	==>	CT_BAIXA	83,15	11,85
OK, VL_BAIXA, CT_BAIXA	==>	FFC_ALTA	82,45	17,93
OK, ERD_BAIXO, VL_BAIXA	==>	FFC_ALTA	81,28	14,05
OK, VL_BAIXA, FFC_ALTA	==>	CT_BAIXA	81,25	17,93
VL_BAIXA, DXL_BAIXA	==>	OK	81,20	16,77
ERD_BAIXO, VL_BAIXA, CT_BAIXA	==>	OK, FFC_ALTA	81,11	11,12
CT_BAIXA, DXL_ALTA, FFC_ALTA	==>	OK	81,09	10,59
VL_BAIXA, FFC_ALTA	==>	CT_BAIXA	80,91	20,43
FFC_ALTA	==>	OK	80,61	41,78
DXL_ALTA, CT_ALTA	==>	VL_ALTA	80,56	18,78
CT_ALTA, Ruim	==>	VL_ALTA	80,54	11,00

Tabela E.3: Demais Regras de Associação Geradas para as variáveis da Cadeira 4.

Se	==>	Então	Confiança(%)	Cobertura(%)
VL_BAIXA, CT_BAIXA, FFC_ALTA	==>	OK	89,37	18,05
OK, VL_BAIXA, CT_BAIXA	==>	FFC_ALTA	89,30	18,05
OK, ERD_BAIXO, CT_BAIXA, DXL_BAIXA	==>	FFC_ALTA	89,06	12,21
VL_BAIXA, CT_BAIXA, DXL_BAIXA, FFC_ALTA	==>	OK	89,06	13,31
CT_BAIXA, DXL_BAIXA, FFC_ALTA	==>	OK	88,70	17,11
DXL_ALTA, Ruim	==>	VL_ALTA	88,61	11,13
ERD_BAIXO, VL_BAIXA, CT_BAIXA, DXL_BAIXA	==>	OK	88,45	11,08
DXL_ALTA, Ruim	==>	FFC_BAIXA	88,21	11,08
OK, VL_BAIXA, CT_BAIXA, DXL_BAIXA	==>	FFC_ALTA	87,96	13,31
ERD_BAIXO, VL_BAIXA, CT_BAIXA	==>	OK	87,95	14,92
OK, ERD_BAIXO, VL_BAIXA	==>	FFC_ALTA	87,91	16,84
CT_BAIXA, FFC_ALTA	==>	OK	87,78	25,83
VL_BAIXA, FFC_ALTA	==>	OK	87,75	22,33
CT_ALTA, FFC_BAIXA, Ruim	==>	VL_ALTA	87,67	10,54
OK, CT_BAIXA, DXL_BAIXA	==>	FFC_ALTA	87,46	17,11
VL_BAIXA, DXL_BAIXA, FFC_ALTA	==>	OK	87,43	16,59
OK, ERD_BAIXO, VL_BAIXA, DXL_BAIXA	==>	FFC_ALTA	87,29	12,39
ERD_BAIXO, FFC_BAIXA, DXL_ALTA	==>	VL_ALTA	87,28	12,74
OK, ERD_BAIXO, CT_BAIXA	==>	FFC_ALTA	87,12	17,86
FFC_BAIXA, DXL_ALTA	==>	VL_ALTA	87,06	23,50
ERD_BAIXO, DXL_BAIXA, FFC_ALTA	==>	OK	87,01	16,43
DXL_ALTA, Indefinido	==>	VL_ALTA	86,60	10,35
ERD_BAIXO, Ruim	==>	FFC_BAIXA	86,32	10,32
ERD_BAIXO, VL_BAIXA, CT_BAIXA	==>	FFC_ALTA	85,94	14,58
CT_ALTA, VL_ALTA, Ruim	==>	FFC_BAIXA	85,88	10,54
ERD_BAIXO, VL_BAIXA, DXL_BAIXA	==>	OK	85,71	14,20

Tabela E.4: Demais Regras de Associação Geradas para as variáveis da Cadeira 5.

Referências Bibliográficas

Agrawal, R., Imielinski, T., and Swami, A. (1993). "Mining Association Rules Between Sets of Items in Large Databases". In *Proceedings of the ACM SIGMOD Conference*, pages 207–216, Washington, DC, USA.

Akaike, H. (1974). "New look at the statistical model identification". *IEEE Transactions in Automatic Control*, (19), 716–723.

Apté, C. (1997). "Data Mining: An Industrial Research Perspective". *IEEE Computational Science and Engineering*, **April - June**, 6–9.

Arieh, D. B. and Chopra, M. (1998). "Data Mining Application for Real-Time Distributed Shop Floor Control". *IEEE Computational Science and Engineering*, **1**, 2738–2743.

Aspen PIMS System (1994-2006). Aspen Technology, Inc. `http://www.aspentech.com/products/product.cfm?ProductID=129` - Último acesso em 30/07/2007.

Berry, M. and Linoff, G. (1997). *"Data Mining Techniques - for Marketing, Sales, and Customer Support"*. John Wiley and Sons, New York.

Braga, A. P., Ludemir, T. B., and Carvalho, A. C. P. (2000). *"Redes Neurais Artificiais: Teorias e Aplicações"*. LTC, Belo Horizonte, MG.

Braha, D. and Shmilovici, A. (2002). "Data Mining for Improving a Cleaning Process in the Semiconductor Industry". *IEEE Transactions on Semiconductor Manufacturing*, **15**(1), 91–101.

Breault, J., Goodall, C. R., and Fos, P. J. (2002). "Data Mining a diabetic data warehouse". *Artificial Intelligence in Medicine*, **26**, 37–54.

Breiman, L., Friedman, J., Stone, C. J., and Olshen, R. A. (1984). *"Classification and Regression Trees"*. Wadsworth International Group, Belmont, California.

Cios, K. J., Pedrycz, W., and Swiniarsk, R. M. (1998). *"Data mining methods for knowledge discovery: Rough Set Theory Review"*. Kluwer Academic, Boston.

Cser, L., Korhonen, A. S., Gul, J., Mäntyla, P., Simula, O., Reiss, G., and Ruha, P. (1999). "Data Mining and State Monitoring in Hot Rolling". *IEEE Transactions on Semiconductor Manufacturing*, (3), 529–536.

Dias, C. A. (2002). "Descoberta de Conhecimento em Banco de Dados para Apoio a Tomada de Decisão". Technical report, Governo do Estado de São Paulo - Universidade Estadual Paulista - `http://www.feg.unesp.br/ceie/Monografias/CEIE0206.pdf` - Último acesso em 30/07/2007.

Elsila, U. and Röning, J. (2002). "Knowledge Discovery in Steel Industry Measurements". In *Proceedings of STarting Artificial Intelligence Researchers Symposium*, pages 197–206.

Fayyad, U., Shapiro, G. P., and Smyth, P. (1996). "From data mining to knowledge discovery in databases". *AI Magazine*, **17**(3), 37–54.

Flores, J. (2005). "Patrones de Morosidad para un Producto Crediticio Usando la Técnica de Árbol de Clasificación CART". *Revista de la Facultad de Ingeniería Industrial*, **8**, 29–36.

Gao, L. H., Luo, S. X., Qiu, J. Q., and Li, F. C. (2003). "Data Mining on a Kind of Complex Industrial Process". In *Proceedings of the Second International Conference on Machine Learning and Cybernetics*, pages 105–107. Institute of Electrical and Electronics Engineers (IEEE).

Garcia, A. C. (2004). *"Regras de Classificação - Árvores de Decisão"*. Master's thesis, Instituto de Computação, Universidade Federal Fluminense.

Garcia, S. C. (2000). "O uso de Árvores de Decisão na descoberta de conhecimento na área da saúde". Technical report, UFRGS - Universidade Federal do Rio Grande do Sul. `http://www.inf.ufrgs.br/pos/SemanaAcademica/Semana2000/SimoneGarcia/` - Último acesso em 30/07/2007.

Gardner, S. R. (1998). "Building the Data Warehouse". *Communications of the ACM*, **41**(9), 52–60.

Ge, A. (1999). *"A Neural Network Approach to the Modeling of Blast Furnace"*. Master's thesis, Department of Electrical Engineering and Computer Science, Massachusetts Institute of Technology.

Haapamäki, J. J., Tamminen, S. M., and Röning, J. J. (2005). "Data Mining Methods in Hot Steel Rolling for Scale Defect Prediction". *Artificial Intelligence and Applications*, pages 90–94.

Hand, D., Mannila, H., and Smyth, P. (2001). *"Principles of Data Mining"*. The MIT Press.

Haykin, S. (1999). *"Neural networks - A comprehensive foundation: Self-organizing maps"*. Prentice-Hall, 2nd edition.

Himberg, J., Ahola, J., Alhoniemi, E., Vesanto, J., and Simula, O. (2001). "The Self-Organizing Map as a Tool in Knowledge Engineering". *Pattern Recognition in Soft Computing Paradigm, World Scientific Publishing Company*, **2**, 38–65.

IBM DB2 *Intelligent Miner* (1996). IBM Corporation. `http://www-306.ibm.com/software/data/iminer/` - Último acesso em 30/07/2007.

iDA: *intelligent Data Analyzer* (2002). Information Acumen Corporation, 13570 Grove Dr., 155, Maple Grove, MN 55311-4400. `http:www.infoacumen.com` - Último acesso em 30/07/2007.

Inselberg, A. (1998). "Visual Data Mining with Parallel Coordinates". *Journal of Computational Statistics*, **13**(1), 47–63.

Irizuki, Y., Tsutaki, S., Tani, T., and Furuhashi, T. (1999). "Extration of Operation Know-How of Experienced Operators Using Neural Networks and Its Application to PID and Neuro-Fuzzy Hierarchical Controller". *IEEE Control System Magazine*, **3**, 274–279.

Kusiak, A. (2000). "Decomposition in Data Mining: An Industrial Case Study". *IEEE Transactions on Electronics Packaging Manufacturing*, **23**(4), 345–353.

Lewis, R. (2000). "An Introduction to Classification and Regression Tree (CART) Analysis". In *Proceedings of the 2000 Annual Meeting of the Society for Academy Emergency Medicine*, pages 1–14, San Francisco, California, USA.

Ma, Y. (1998). *"Data Warehousing, OLAP, and Data Mining: An Integrated Strategy for Use at FAA"*. Master's thesis, Department of Electrical Engineering and Computer Science, Massachusetts Institute of Technology.

MacQueen, J. B. (1967). "Some Methods for classification and Analysis of Multivariate Observations". In *Proceedings of the 5th Berkeley Symposium on Mathematical Statistics and Probability*, pages 281–297, Berkeley. University of California Press.

Malone, T. (2005). "Information Technology Essentials - Emerging Technologies". Technical report, MIT - Massachusetts Institute of Technology - `http://ocw.mit.edu/index.html` - Último acesso em 30/07/2007.

Mastrangelo, C. M. (1998). "Data Mining in a Chemical Process Application". *IEEE Computational Science and Engineering*, **1**, 2917–2921.

Menzies, T. and Hu, Y. (2003). "Data Mining For Very Busy People". *IEEE Computer*, **36**(11), 22–29.

Morello, B. C., Michaut, D., and Baptiste, P. (2001). "Knowledge Discovering process for a flexible manufacturing system". *IEEE Transactions on Electronics*, **7**, 651–658.

Navathe, S. and Elmasri, R. (2000). *"Sistemas de Bancos de Dados - Fundamentos e Aplicações"*. Editora LTC.

Ogilvie, T., Swidenbank, E., and Hogg, B. W. (1998). "Use of Data Mining Techniques in the Perfomance Monitoring and Optimisation of a Thermal Power Plant". In *IEE Colloquium on Knowledge Discovery and Data Mining*, pages 7/1–7/4. Institution of Electrical Engineers (IEE).

Oliveira, R. B. T. (2000). *"O Processo de Extração de Conhecimento de Base de Dados Apoiado por Agentes de Software"*. Master's thesis, Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo.

Pawlak, Z. (1991). "Rough sets: Theoretical aspects of reasoning about data". *System Theory, Knowledge Engineering and Problem Solving, Kluwer, Dordrecht*, **9**.

Quinlan, R. J. (1993). *"C4.5: programs for machine learning"*. Morgan Kaufmann, San Francisco, CA.

Roed, G. (1999). *"Knowledge Extraction from Process Data: A Rough Set Approach to Data Mining in Time Series"*. Master's thesis, Knowledge Systems Group, Department of Computer and Information Science, Norwegian University of Science and Technology. Trondheim, Noruega.

Roiger, R. and Geatz, M. (2002). *"Data Mining: A Tutorial-Based primer"*. Addison Wesley.

ROSETTA: *A Rough Set Toolkit for analysis of Data* (1998). Knowledge Systems Group, Dept. of Computer and Information Science, Norwegian University of Science and Technology, Trondheim, Noruega. `http://rosetta.lcb.uu.se/general` - Último acesso em 30/07/2007.

Rätsch, G., Onoda, T., and Müller, K. R. (2001). "Soft margins for AdaBoost". *Machine Learning*, **42**(3), 287–320.

Shearer, C. (2000). "CRISP-DM Model: The New Blueprint for Data Mining". *Journal of Data Warehousing*, **5**(4), 13–22.

Statistica: Data Mining Software (1984-2004). StatSoft, Inc. 2300 East 14th street Tulsa, OK 74104 USA. `http://www.statsoft.com` - Último acesso em 30/07/2007.

Thrun, S., Faloutsos, C., Mitchell, T., and Wasserman, L. (1998). "Automated Learning and Discovery: State-Of-The-Art and Research Topics in a Rapidly Growing Field". Technical report, Carnegie Mellon University, Center for Automated Learning and Discovery.

Two Crows Corporation (1999). "Introduction to Data Mining and Knowledge Discovery". Technical report, 10500 Falls Road. Potomac, Maryland 20854 - U.S.A. - `http://www.twocrows.com/booklet.htm` - Último acesso em 30/07/2007.

Wang, X. Z. (1999). *"Data Mining and Knowledge Discovery for Process Monitoring and Control"*. Springer.

Weka: *Data Mining Software in Java* (2005). Data Mining: Practical machine learning tools and techniques, Witten I. H. e Frank, E. Segunda Edição, Morgan Kaufmann, San Francisco. `http://www.cs.waikato.ac.nz/ml/weka/` - Último acesso em 30/07/2007.

Ye, N. (2003). *"The Handbook of Data Mining"*. Lawrence Eelbaum Associates, London.

Yohannes, Y. and Webb, P. (1999). "Classification and Regression Trees, CART. A user manual for identifying indicators of vulnerability to famine and chronic food insecurity". Technical report, IFPRI - International Food Policy Research Institute, 2033 K Street, N.W. Washington, D.C. 20006-1002.

Yoshida, T. and Touzaki, H. (2000). "A Study on Association among Dispatching Rules in Manufacturing Scheduling Problems". *IEEE Transactions on Electronics Packaging Manufacturing*, **5**, 1355–1360.

Zhong, N. N. and Wang, X. (2003). "Application of Data Mining in Refinery CIMS". In *Proceedings of the Second International Conference on Machine Learning and Cybernetics*, pages 118–122. Institute of Electrical and Electronics Engineers (IEEE).