UMA ANÁLISE DE CANCELAMENTOS EM TELEFONIA UTILIZANDO MINERAÇÃO DE DADOS

Daniel Frankowicz de Andrade

DISSERTAÇÃO SUBMETIDA AO CORPO DOCENTE DA COORDENAÇÃO DOS PROGRAMAS DE PÓS-GRADUAÇÃO DE ENGENHARIA DA UNIVERSIDADE FEDERAL DO RIO DE JANEIRO COMO PARTE DOS REQUISITOS NECESSÁRIOS PARA A OBTENÇÃO DO GRAU DE MESTRE EM CIÊNCIAS EM ENGENHARIA CIVIL.

Aprovada por:

_____
Prof. Alexandre Gonçalves Evsukoff, DSc.

_____
Prof. Nelson Francisco Favilla Ebecken, DSc.

_____
Prof. Elton Fernandes, DSc.

_____
Prof. Angelo Maia Cister, DSc.

RIO DE JANEIRO, RJ - BRASIL
SETEMBRO DE 2007

ANDRADE, DANIEL FRANKOWICZ DE

Uma análise de cancelamentos em telefonia utilizando mineração de dados [Rio de Janeiro] 2007.

VII, 69 p. 29,7 cm (COPPE/UFRJ, M. Sc., Engenharia Civil, 2007)

Dissertação - Universidade Federal do Rio de Janeiro, COPPE

1. Análise de Churn em Telefonia Móvel

2. Data Mining

3. Marketing de Relacionamento

I. COPPE/UFRJ II. Título (série)

AGRADECIMENTOS

Agradeço a Deus, sem cuja permissão nada teria sido iniciado, quanto menos concluído;

Agradeço ao meu orientador Alexandre Evsukoff, que soube explorar em mim as experiências de trabalhar no marketing de uma empresa de telefonia e soube lidar com as minhas limitações de tempo. Agradeço também bela proposta de dissertação e pelo total alinhamento desta com as minhas atividades profissionais, gerando oportunidades futuras dentro da empresa.

Agradeço ao Angelo Cister Maia pela bela base de dados que me disponibilizou. Através desta oportunidade pude testar a metodologia objeto da minha dissertação, e avaliar os resultados na visão do segmento de telecomunicações.

Agradeço à minha esposa Vivian e aos meus pais, que me ajudaram nos momentos em que mais precisei e que me acompanharam em todas as fases deste mestrado.

Agradeço ao amigo Hugo Azevedo, que foi meu professor na PUC, e me motivou a realizar as presentes análises.

Obrigado a todos.

Resumo da Dissertação apresentada à COPPE/UFRJ como parte dos requisitos necessários para a obtenção do grau de Doutor em Ciências (M. Sc.)

UMA ANÁLISE DE CANCELAMENTOS EM TELEFONIA UTILIZANDO MINERAÇÃO DE DADOS

Daniel Frankowicz de Andrade

Setembro/2007

Orientador: Alexandre Gonçalves Evsukoff

Programa: Engenharia Civil

O objetivo desta tese foi apresentar uma metodologia para mineração de dados para a análise de churn em telefonia móvel.

Através da análise de uma base de clientes onde parte deles evadiu a empresa (churn), desenvolvi modelos que facilitam a identificação deste perfil de cliente. Complementei a análise comparando estudos supervisionados e não superfisionados neste tipo de problemas.

Ao final, complementando o desenvolvimento da metodologia, realizei algumas simulações no SAS, utilizando vários modelos, mostrando que independente do modelo utilizado a metodologia apresentada é muito eficiente na obtenção dos resultados desejados.

Abstract of Dissertation presented to COPPE/UFRJ as a partial fulfilment of the requirements for the degree of Master of Science (M. Sc.)

A CHURN ANALISYS IN TELCOM USING DATA MINING

Daniel Frankowicz de Andrade

September/2007

Advisor: Alexandre Gonçalves Evsukoff

Department: Civil Engineering

The goal of this thesis is to introduce a methodology for churn analysis in telco mobile enterprise.

Through the analysis of a telco customer database, where part of them left the company (churn), I worked on a methodology to make it easy to identify these characteristics. Complementing the analysis I compare the approach for this kind of temporal database with supervised models and not supervised models.

At last, complementing the presented methodology, I run soma SAS applications, using many different models and structures, to show that these methods helps, at least the most important models, to achieve good results.

ÍNDICE

Capítulo 1 – Introdução

Originalmente, o *marketing* foi concebido como ramo de aplicações econômicas, para estudar os canais de distribuição de produtos. Posteriormente, tornou-se uma disciplina de gestão dedicada a incrementar as vendas. Mais tarde, passou a ser considerado como ciência aplicada ao estudo do comportamento do cliente, a fim de entender o processo de compra e venda, envolvido no *marketing* de produtos e serviços.

Devido à concorrência e sofisticação do cliente, o conceito do *marketing* evoluiu para uma orientação no sentido de satisfazer às necessidades do cliente. Sob este conceito do *marketing*, oposto à orientação tradicional para vendas, o produto é uma variável a ser realizada e modificada em resposta às mudanças nas exigências do cliente.

Assim sendo, as novas empresas, as marcadas pelos novos ventos da economia globalizada, se vêem obrigadas a mudar seu comportamento perante os novos e exigentes consumidores.

A necessidade de disputar o consumidor e mantê-lo fiel à sua marca/produto fez com que as empresas revissem seus conceitos, voltados quase que exclusivamente para seu produto e/ou marca, e voltassem seus olhos mais atentamente para os consumidores. Nunca conceitos, tais como CRM (*Costumer Relationship Management*), retenção de clientes (lealdade de seus clientes), análise de *database marketing*, *data mining* e segmentação estiveram tão em voga quanto nos dias de hoje.

Por essa razão, considera-se relevante pesquisar, analisar e discutir sistemas integrados como alternativa para o crescimento e desenvolvimento empresarial. Acredita-se que inovar pode, paradoxalmente, voltar à atenção para o cliente em busca de informações que possam ampliar o negócio de empresa. Sabe-se ainda, que a lealdade (fidelidade) é um fator importante nesse mercado sem limites, uma vez que a inteligência e saber onde buscar a informação necessária são de fundamental importância.

Abordagem Metodológica

O presente trabalho uniu duas áreas de conhecimento que são o marketing, na concepção administrativa e quantitativa, e a área de mineração de dados (*datamining*).

É apresentado por tema o CRM analítico; como **sujeito de estudo:** análise de cancelamentos (Análise de *Churn*); **delimitação do estudo:** modelagem matemática através do uso de técnicas de mineração de dados, para melhor compreensão do perfil do cliente, campanhas de retenção do cliente e previsão de evasão voluntária da base de dados das empresas; **objeto de** estudo: análise de cancelamentos (Análise de Churn) em uma empresa de telefonia celular (Data Mining e CRM (*Customer Relationship Management*)).

Como objetivos a serem alcançados tem-se a classificação, previsão e identificação de clientes que por ventura possam evadir da base de dados, ocorrendo, desta forma, desconforto financeiro para empresa.

Ao final deste trabalho ter-se-á condições de propor uma forma de uso sistemático para a classificação e previsão de clientes que possam ser alvos de ação de marketing (ação de retenção), dentro da base de dados de qualquer empresa que possua concorrentes a sua altura. O **objetivo principal** é reconhecer clientes que, por insatisfação com a empresa, desejam sair da mesma, ou, simplesmente, trocar de empresa por não considerarem a atual digna de seus préstimos.

Este trabalho aproximou-se ao máximo da realidade, através de sugestões bibliográficas especializadas, base de dados verdadeira, análise dos dados e definição das diversas técnicas para formatação dos dados e extração de conhecimento.

As metodologias utilizadas para a classificação foram as técnicas de Redes Neurais, em seus diversos aspectos quanto a sua calibragem, assim como geração de regras de associação e filtragem dos dados, como pré-processamento dos dados.

Através de bibliografia especializada sugerida e selecionada, foram desenvolvidos os primeiros capítulos sobre estado da arte da telefonia no Brasil, teoria de *churn*, mineração de dados e gerência de relacionamento com o cliente (customer relationship management -CRM).

Em seguida, o trabalho contará com a descrição do experimento de campo, onde procurará aplicar os conhecimentos das técnicas de *cleaning* (limpeza dos dados) e a classificação dos dados.

A pesquisa foi realizada em uma base real de dados de uma empresa de telefonia celular, que guarda o direito de não ser identificada, após ter sido minuciosamente limpa (processo de *cleaning* dos dados) e devidamente auditada quanto robustez dos dados e sua veracidade no mercado.

Após a análise dos resultados da pesquisa, foi feita uma avaliação de todo o processo experimental, com o objetivo de levantar os pontos fracos e fortes, para então se chegar a conclusões e recomendações para futura aplicabilidade deste processo.

Estrutura da Tese

A tese tem por estrutura os capítulos: Capítulo 1 - apresenta os principais conceitos de marketing envolvidos na presente dissertação; Capítulo 2 – fala a respeito da teoria de cancelamentos de clientes em empresas (análise de *churn*); Capítulo 3 - retrata o experimento e sua análise, problemas apresentados e como foram resolvidos; Capítulo 4 – demonstra os resultados da metodologia apresentada no capítulo anterior e por fim é apresentado o capítulo 5 – com a conclusão da análise e sugestões para estudos posteriores.

Capítulo 2 - Análise de Cancelamentos (“CHURN”)

2.1 Definição

O estudo sobre *churn* é extremamente importante, pois grande parte das organizações, telecomunicações e informática estão sob a mesma responsabilidade. E, a cada dia que passa, fica mais difícil separá-las. Mas afinal, o que é o *churn*?

Na informática, *churn* é utilizado, por alguns autores, para expressar a renovação acelerada dos produtos ou a conhecida obsolescência programada. Mas o *churn*, objeto deste estudo, que tem como base o *marketing*, trata da perda de clientes sofrida por uma empresa para a concorrência, ou seja, é uma medida da infidelidade dos clientes. Este é o conceito que está mais em pauta entre as empresas de telecomunicação ou em qualquer outra empresa de serviços. No Brasil, com a desregulamentação que as telecomunicações sofreram, este é um dos setores que mais vive este fenômeno.

Outros setores da economia também têm que administrar o *churn*. Os bancos e as administradoras de cartão de crédito são dois exemplos bem conhecidos. Para os consumidores, esta vasta gama de opções significa maior liberdade de escolha. Portanto, seja na telefonia fixa, móvel, comunicação de dados ou internet, podem mudar de fornecedor com facilidade.

Conseqüentemente, os fornecedores tratarão de traçar estratégias específicas, de acordo com o comportamento dos consumidores e terão que ir mais longe, atingindo níveis maiores de segmentação e diferenciação, para alcançar seu objetivo de fidelidade. Aplicações de *business intelligence* e Customer Relantionship Management (CRM) são estratégias para tais indústrias.

Para os consumidores, isso significa uma vitória. Hoje, podem comparar propostas, exigir um tratamento de qualidade, ter um suporte e assistência técnica cada vez melhores.

2.2 Causas do Churn

Há três tipos de *churn*: involuntário, voluntário e inevitável.

Involuntário - quando o usuário deixa de pagar pelo serviço e tem seu

fornecimento cancelado. Os motivos pelos quais o cliente deixa de pagar podem ser os mais diversos, como desemprego, falta de capital suficiente para se manter entre outros.

Voluntário - quando o cliente decide mudar de fornecedor, seduzido por campanhas de marketing e/ou promoções.

Inevitável - quando o usuário vem a falecer ou muda-se para uma localidade não atendida pelo fornecedor.

Até há pouco tempo atrás, a cultura das empresas era “conquistar o cliente a qualquer preço”. Atualmente, a cultura é “reter o cliente” e, claro, conquistar o bom cliente. Conquistar um cliente do tipo alta-rotatividade, com propensão compulsiva ao *churn pode* não ser um bom negócio; melhor deixá-lo com a concorrência.

É provado que o custo para se manter um cliente é muito menor que o de se conquistar um novo cliente. Para se evitar o *churn*, empregam-se ferramentas de mineração de dados e estatística multivariada. Estas ferramentas permitem que se analise o banco de dados com informações do perfil histórico de cada usuário e que se determine quais clientes são leais, quais são propensos ao *churn* e quais são realmente de alto valor para a empresa.

Com base nessas informações, a empresa toma atitudes não só reativas, em relação aos clientes que desistiram do serviço, mas, principalmente, ações pró-ativas, ao identificar os bons clientes e selecionar planos especiais para garantir sua fidelização, evitando, com isso, a evasão dos clientes que agregam altos valores para a empresa.

Não se pode assumir automaticamente que uma taxa de *churn* alto é ruim, bem como uma taxa de *churn* baixa é boa. Tudo depende do teatro de guerra onde se está travando a luta de mercado. Em mercados altamente competitivos ou em transição, quando se deseja manter uma participação elevada na publicidade, o *churn* alto pode ser o custo de tal estratégia.

O baixo *churn* também deve ser considerado no contexto estratégico. É lucrativo, mas pode significar pouca agressividade de marketing, revelando-se (ou não) prejudicial em longo prazo, já que tende a inibir a entrada de novos consumidores para cobrir os cancelamentos naturais ao longo do tempo.

2.3 Implicações

Atualmente, a análise de *churn* tem se tornado um desafio para empresas de vários

ramos, porém poucas reconhecem suas reais implicações. O setor de Telecomunicações é o pioneiro na busca por ferramentas e sistemas que ajudem nessa análise.

Como é feita a análise do *churn*? Basicamente, a análise de *churn* é uma análise de dados históricos que permite prever que clientes poderão vir a deixar a empresa, com tomada de ações, por parte da empresa, que evitem esse *churn*. Na maior parte dos casos, o estudo concentra-se no comportamento e /ou perfil do cliente.

Quando se diz que a taxa de *churn* de uma empresa varia de 25% a 30% anualmente, deseja-se realmente dizer que essa empresa está perdendo de 25% a 30% dos clientes que se encontram na sua base de dados. Isso significa que ela está perdendo clientes que fazem ou já fizeram negócios com ela e que, por algum motivo, não ficaram satisfeitos e evadiram de sua base de dados.

Isso levanta dois pontos. O primeiro ponto é que a análise de *churn* tem que ser detalhada o suficiente para descobrir as variáveis que fazem com que esses clientes insatisfeitos evadam-se e o porquê da insatisfação.

O problema dessa análise é que o volume de dados a ser analisado é muito grande e, normalmente, o banco de dados dessas empresas é muito extenso, o que dificulta ainda mais o acesso às informações e à consolidação dessas informações num sistema de maneira eficiente e rápida.

Aliado ao problema de volume de dados, tem-se, também, o problema da falta de conhecimento, tanto teórico quanto técnico, por parte das empresas. A maioria das empresas não possui o mínimo de conhecimento em exploração de dados e análise estatística, pré-requisitos importantes para poder prever o *churn*. Uma vez descobertas as variáveis, as empresas podem tomar as devidas ações para evitar o *churn*.

Outro ponto é que muitos desses clientes são clientes que têm um grande potencial de gasto com a empresa. Portanto, são clientes valiosos que a empresa está perdendo e que, provavelmente, nunca voltarão. Uma análise detalhada do *churn* permite às empresas determinarem aqueles clientes que apresentam um comportamento que indica que eles podem se tornar "*churners*" e quais desses "*churners*" são clientes que merecem uma atenção especial. Essas informações contribuem para que as empresas melhorem os seus programas de fidelização dos clientes e reduzam os seus custos em marketing, uma vez que há queda no crescimento do mercado e a alta competitividade.

Com o aquecimento da competição, operadoras de telefonia buscam, nas soluções de CRM, os caminhos para engordar a carteira de clientes e aumentar suas receitas.

Com a abertura e a desregulamentação que as telecomunicações sofreram,

mundialmente, este é um dos setores que mais vivem o fenômeno *churn*. Outros setores da economia já aprenderam a administrar o "*churn*" há tempos. Os bancos e as administradoras de cartão de crédito são dois exemplos bem conhecidos.

No Brasil, os clientes ainda não têm a mais vasta gama de opções, porém já começam a sentir a liberdade de escolha. Seja na telefonia fixa, móvel, comunicação de dados, Internet, etc., troca-se de fornecedor com facilidade.

A grande migração para os provedores de Internet gratuitos é um exemplo recente dessa facilidade. Muitas pessoas usam um "*web-mail*" para escolher e trocar de provedor de acesso quando quiser e, como o comportamento das pessoas e das corporações é diferente, os fornecedores vão traçar estratégias específicas. Eles terão que atingir níveis maiores de segmentação e diferenciação, para alcançar seu objetivo de fidelidade. Os clientes devem se preparar para serem alvos de planos de marketing bastante ousados.

Com isso, o cliente tem muito o que comemorar, pois viveu anos e anos sem opções, com poucas empresas oferecendo produtos e serviços tão essenciais. Hoje se pode comparar propostas, exigir um tratamento de qualidade, ter um suporte e assistência técnica cada vez melhores. Muitos ainda têm contratos antigos e vínculos fortes, mas, com o tempo, essa inércia deve ser superada.

A possibilidade de uma migração fácil manterá os fornecedores em estado de alerta. E, se mesmo assim, houver algum problema sério, seja de natureza técnica ou econômica, ou ainda por qualquer desgaste na relação, a resposta do cliente tem grandes chances de ser: “Adeus”. Nesse caso, o maior desafio das empresas será o de segurar e manter o seu cliente na sua base de dados.

Hoje, a palavra é relacionamento, ou seja, conhecer o usuário e possuir alternativas de soluções para ele. Antigamente, o atendimento se resumia, apenas, na recepção ao cliente. Esse tipo de iniciativa, cujo objetivo é aperfeiçoar os canais de atendimento, representa apenas um dos elementos que compõem uma política de CRM. A primeira corrida das operadoras envolveu as centrais de atendimento devido às exigências da Anatel e, por isso, elas passaram a associar CRM somente ao *call center*. Iniciar um projeto abordando os canais de atendimento é ótimo, mas se não houver a noção do todo, ele se torna ineficiente.

Há dez anos, o *call center* era uma central de reclamações e, agora, começa a evoluir para um serviço de interação. Uma postura como essa pode auxiliar as operadoras a reduzir o número de queixas nos Procons. Quando a companhia estabelece

boa relação com seu cliente, em caso de algum problema, ele vai procurar os canais da empresa antes de buscar os órgãos de defesa do consumidor.

A integração dos sistemas é, portanto, um ponto-chave para o sucesso das ações de CRM.

No Brasil, aproximadamente, um a cada oito projetos, não alcança bons resultados. Os aspectos que exigem cuidado são a interligação com as diversas arquiteturas da companhia e a personalização do produto. O principal problema na adoção dos processos de CRM é a falta de iniciativas internas para a integração entre os canais de marketing, vendas e serviços e, principalmente, entre os sistemas de *back office* e *front office* das corporações. Por isso, o foco das operadoras, hoje, é a integração. Elas ainda não possuem arquiteturas interligadas, especialmente em função dos sistemas legados, gerando também perda operacional.

Mas os benefícios do CRM não se limitam à conquista e à retenção de clientes. Eles também se traduzem em redução de custos e aumento da receita média por assinante. Com o conceito, a companhia economiza, porque atira no alvo certo. Além disso, a guerra de preços apresenta alguns limites, por isso, o diferencial será o atendimento.

As operadoras começaram a perceber a importância dessa ação, principalmente as de telefonia móvel. Há uma certa acomodação por parte das prestadoras de serviços de linhas fixas e isso ocorre devido à falta de concorrência. Com o aumento da competitividade, esse quadro será alterado e todas correrão em busca de soluções de CRM.

A previsão de que, provavelmente, os clientes mudarão de fornecedor e a determinação de incentivos eficazes, do ponto de vista de custo, para persuadi-los a continuar são iniciativas muito difíceis para a maioria das empresas de telecomunicações. Os volumes de dados necessários são muito grandes, dependendo da operadora, e, freqüentemente, difíceis de acessar e consolidar por meio de ferramentas convencionais. Muitas razões levam um assinante a abandonar sua operadora de telecomunicações. Os custos para se conquistar um novo cliente já foram razoavelmente dimensionados por estas empresas. Mas na telefonia fixa é uma incógnita. Os *business plans* das operadoras locais já estão sendo modificados para adicionar mais um campo de custos: o custo do *churn*. Isto significa, a médio prazo, uma diminuição da margem de lucro operacional das empresas de telefonia local.

Alguns clientes buscam serviços de melhor qualidade, novas tecnologias ou serviços avançados. Uma grande maioria procura as melhores tarifas, pois a tendência é

que cada dia haja menor diferenciação quanto à qualidade de serviço entre as operadoras de telecomunicações. Outros assinantes, simplesmente, mudam de operadoras por não poderem pagar suas contas, clientes inadimplentes que dão origem ao *churn* involuntário. Alguns são seduzidos por campanhas de marketing, promoções e uma infinidade de prêmios, até mesmo em dinheiro, pelas operadoras. Os três últimos tipos de clientes são, efetivamente, os grandes responsáveis pelo *churn*.

Do ponto de vista do consumidor brasileiro, que passou anos sem opções com poucas empresas oferecendo produtos e serviços tão essenciais, a comemoração é inevitável. Hoje, é possível comparar propostas e exigir um tratamento de qualidade, sem falar no suporte técnico cada vez melhor. É verdade que muitos consumidores ainda têm contratos antigos e vínculos fortes, mas, com o tempo, essa inércia deverá ser superada. A possibilidade de uma migração mais fácil manterá os fornecedores em estado de alerta. Se o consumidor tiver algum problema de natureza técnica, econômica ou, ainda, desgaste na relação (cliente-empresa), a decisão de deixar a empresa será muito mais fácil.

Predizer que é provável que clientes saiam e persuadir os "*churn*ers" a permanecerem é empreendimento extremamente difícil para a maioria das companhias de telecomunicações. Os volumes de dados necessários são enormes e, freqüentemente, os *data-marts* foram segmentados para que cada setor possa ver seu cliente de uma maneira, tornando a consolidação dos dados difícil. E o que faltando para muitas organizações se adequarem a essa nova postura do consumidor é, simplesmente, a perícia para suportar a mineração complexa dos dados e as tarefas de análise preditiva, que são essenciais no *churn*.

A implementação do projeto, que demandou a estruturação de um departamento de análise, trabalhou no lançamento do Business Highway, que, voltado para os pequenos negócios, oferece três números telefônicos em uma só linha. Dados sobre clientes - alvo para o produto que estava espalhado entre diversos departamentos, o que exigiu sua integração em um novo *data mart*.

Começou, então, a fase de análise e modelagem dos dados, na qual identificaram-se questões relativas à qualidade, e a equipe familiarizou-se com a distribuição dos dados, eliminando atributos que não estavam fortemente vinculados à compra do *Business Highway*. No jargão do profissional que trabalha com banco de dados, significa normalizar a base de dados, gerando chaves estrangeiras, possibilitando uma maior facilidade na obtenção dos dados, que se encontram em banco de dados

relacionais e/ou voltados para objetos. Dessa forma, um simples *query*, através de uma linguagem apropriada (SQL), permite aos funcionários levantarem a informação necessária e a medirem a força de cada atributo individual com relação à propensão do cliente - alvo de comprar o produto.

Depois da análise, o time rapidamente construiu e testou uma série de modelos experimentais, a partir das árvores decisórias, e ofereceu aos departamentos de marketing e vendas uma lista com os "melhores prospects", facilitando a identificação dos clientes com maior potencial de rentabilidade e, também, aqueles que exigem muita atenção, apesar de comprarem pouco. No futuro, a companhia poderá buscar a definição de padrões indicadores de tendência a churn.

2.4 Modelo Preditivo de Churn

A transformação de informação em conhecimento é baseada em técnicas estatísticas. A informatização dos setores de serviços gerou grandes bases de dados, mas a disponibilização desta informação não significou, necessariamente, maior conhecimento e qualidade nos serviços.

Na era do conhecimento, a estatística e a mineração de dados possuem papel importante, além de serem ferramentas poderosas, e passam a ser vistas como um poderoso método de gestão. O foco da qualidade está em constante evolução. Passaram a era dos produtos, dos processos, dos clientes. Hoje, o foco passa a ser o conhecimento através de uma abordagem mais abrangente.

Os modelos matemáticos são adotados sistematicamente no meio acadêmico, auxiliando na validação de hipóteses, principalmente, no processo de indução. O objetivo dos modelos matemáticos é tornar a pesquisa cientifica a mais eficiente possível.

O uso de modelos matemáticos na indústria ocidental tem uma história de altos e baixos . Vistos muito mais como uma ferramenta, auxiliaram na indústria japonesa e, posteriormente, na indústria ocidental. Porém, pouco influenciaram no comportamento gerencial e na forma de gestão. A estatística deveria ser vista como um meio para se obter conhecimento e aumentar a chance de tomar decisões corretas, auxiliando na obtenção de conhecimento sobre os processos, permitindo ver os mesmos sem preconceitos, opiniões ou pré-julgamentos.

O modelo preditivo de churn é um modelo matemático/estatístico construído com base em um grupo de variáveis, que, por conhecimento prévio do negócio ou estudo específico, demonstraram ser relevantes para explicação do evento que se deseja prever. É um modelo baseado no comportamento do cliente, que exibe a probabilidade de abandono e o valor que representa para empresa. Com esses dados, a operadora pode fazer uma campanha de retenção com uma perda muito pequena, além de exigir investimentos menores nessa ação.

O principal objetivo da fórmula é atribuir, para cada variável (ou categoria desta), coeficientes (ou pesos) que estão relacionados ao poder que cada variável possui de explicar o evento em questão. Variáveis muito importantes para a conclusão do modelo devem receber os maiores pesos, enquanto que as mais secundárias receberão pesos inferiores.

Esta fórmula matemática pode ser delineada por conhecimento empírico do negócio (pesos atribuídos arbitrariamente), ou através de técnicas estatísticas de modelagem de dados ou, ainda, auxiliado pelo especialista. É claro que, na maioria das vezes, os modelos que usam técnicas estatísticas possuem índices de acerto maiores do que aqueles que são baseados em conhecimentos e intuições pessoais. Esse fenômeno deve-se a grande quantidade de dados a serem analisados. Em momento algum, pode-se isolar especialista. O conhecimento que o especialista possui é de fundamental importância, principalmente, para sistemas e/ou modelos matemáticos híbridos.

2.5 Fidelização

A fidelização de clientes anda preocupando empresários, executivos e consultores de marketing. Ações, siglas e conceitos, como CRM, Marketing One-to-One, *Database Marketing,* Programas de Fidelização, Programas de Recompensa, *Datamarketing Behavior* e tantos outros provocam mais confusão no mercado. A confusão de conceitos é real. Há quem considere, por exemplo, que programa de fidelização e programa de recompensas (milhagem, cupom, etc.) são a mesma coisa. Recompensa, como se sabe, é baseada no princípio behaviorista de estímulo – resposta, uma vez cessado o estímulo, a resposta tende a desaparecer. Ao passo que fidelização é construída sobre os alicerces do relacionamento e da confiança mútua, que pode se utilizar ou não de programa de recompensas.

Quando se coloca para a empresa a questão do gerenciamento do relacionamento

com o cliente, não se está sugerindo mecanismos de estímulo-resposta, mas sim, maneiras de compreender em profundidade as características, hábitos, desejos, necessidades e potencialidades do cliente, com o objetivo de se desenvolver ações voltadas ao fortalecimento dos vínculos de relacionamento, "encantá-lo" e motivá-lo a manter-se fiel ao seu negócio.

As empresas devem se preocupar em desenvolver estratégias de fidelização de clientes pelas seguintes razões:

- A competição cada vez mais acirrada entre empresas, grupos e países.
- Qualidade e produtividade como fatores decisivos de competitividade.

Exigência de agilidade nas decisões.

1. Poder financeiro crescente dos consumidores.
2. Redução da lealdade às marcas.
3. Importância crescente dos serviços.
4. Sociedade regida pela informação.

Diante desse quadro, não é mais admissível que a empresa fique à margem do que acontece com os seus clientes em geral e com cada grupo de clientes, em particular. Ignorar o cliente, desconhecer as causas de seu comportamento e deixar de monitorar o seu relacionamento com a empresa é candidatar-se à obsolescência. Muitas empresas gastam fortunas tentando escrever sua missão. É preciso ficar claro que a principal missão da empresa é interferir na decisão do cliente em tornar-se fiel.

> "O que dá vida ao modelo de empresa baseada na fidelidade não é a oferta de utilidades imediatas, mas a criação de valor para os clientes, condição fundamental em todas as empresas bem-sucedidas. Como efeito, a fidelidade mede de forma confiável se a empresa gera valor: os clientes continuam comprando dela ou preferem outra empresa. Como causa, a fidelidade aumenta as *receitas e a participação no mercado, o crescimento sustentável permite atrair e conservar os melhores funcionários e os investidores fiéis viram sócios"*. Frederick Reichheld

Capítulo 3 – DESCRIÇÃO DOS DADOS

3.1 Problema

Através deste experimento buscou-se identificar e predizer clientes que possuam inclinação e/ou tendência a evadirem da base de dados de uma empresa de telefonia móvel. A população é composta por determinado segmento de clientes. Esses clientes possuem perfil de pessoa jurídica e que possuem, em média 4 linhas telefônicas móveis e são considerados heavy users , isto é, clientes que retêm um alto índice de minutagem ao mês.

O experimento inicia buscando-se um melhor entendimento da base de dados, tentando identificar:

- O Comportamento das variáveis
- Variáveis relevantes ao processo de classificação
- Possíveis valores faltantes (missing values)
- Possibilidades para redução de variáveis e dimensões

As variáveis tinham por definição a seguinte descrição:

V1: Adicional em roaming - O cliente paga ao receber ligação;

V2: Chamada internacional em roaming;

V3: De móvel para fixo;

V4: De móvel para móvel mesma operadora;

V5: De móvel para móvel de outras operadoras;

V6: De móvel para fixo efetuando DDD;

V7: De móvel para outro móvel mesma operadora em roaming;

V8: De móvel para móvel corporativo;

V9: De móvel para móvel de outras operadoras em roaming;

V10: Serviços diversos 0800, 0300, auxílio a lista.

A estratégia de escolha dos registros foi a de escolher 12 meses completos sendo tirada a fotografia de status, ativo ou inativo, no 13º mês. Dessa forma, a massa de dados foi escolhida entre os meses de agosto de 2004 a julho de 2005, tendo sido retirado o status ativo ou inativo, para o mês de agosto de 2005. Apesar de atender perfeitamente às necessidades do experimento acadêmico, trazendo situações de

desbalanceamento à base de dados, no mercado de trabalho ter-se-ia buscado o status de ativo ou inativo no 14º mês, visto que para empresas de grande porte, como as de telefonia, o tempo gasto para coleta, tratamento e disponibilização das informações toma boa parte do 13º mês, tempo fundamental no processo de tomada de decisão para reter os clientes com perfil de evasão.

3.2 Características Temporais

Uma das principais características ao se analisar bases de dados ligados ao segmento de telefonia é a complexidade no trato de bases que variam no tempo. Na mineração de dados clássica tem-se duas dimensões para serem consideradas, a dimensão das variáveis ou atributos, e a das amostras ou registros. Em uma base de telefonia, além de analisarmos as variáveis que caracterizam um cliente, e de termos uma quantidade enorme de clientes a serem considerados para a análise, tem-se a evolução desta relação ao longo do tempo.

Clusterizar series temporais é um dos mais importantes problemas, que vem atraindo considerável interesse, no conjunto de opções para mineração de dados temporais (temporal data mining). Aplicações práticas, tanto na indústria como no campo científico, consistem na identificação de clientes com características similares de crescimento, agrupamento de elementos de rede com histórico similar de falhas, e identificação de grupos de pacientes com mesma progressão.

Séries temporais podem ser grandes de três formas, gerando desafio para a análise: o comprimento/tamanho da série temporal; o número de amostras ou registros; e o número de atributos registrados para cada unidade de tempo. Séries temporais podem ser extremamente longas dependendo da granularidade das métricas/medidas e da disponibilidade das informações. No segmento de telecomunicações existem diversas informações disponíveis, com longos períodos históricos disponíveis para análise. O número de amostras ou data points podem representar centenas de milhões de clientes residenciais, com várias variáveis registradas para cada um em determinado momento no tempo.

Boa parte da literatura na Mineração de Dados Temporais trata de séries temporais univariadas (uma componente). Séries temporais com uma única variável são tipicamente mapeadas em vetores de parâmetros com tamanho fixo através de modelos de Fourier, wavelet ou ARMA (Auto Regressive Moving Average), onde após a

modelagem os vetores são clusterizados. Estes métodos, contudo, não são apropriados para séries temporais multivariadas. Fourier ou wavelet são indefinidos ou de implementação computacional cara em mais de duas ou três dimensões. Modelagens ARMA necessitam de premissas de linearidade que torna restritivo o seu uso. Modelos não lineares para parametrização temporal multivariada existem, mas sua computação normalmente requer múltiplos passos de reinderização (rendering), inviável para dados massivos (volumosos).

Neste experimento buscou-se desenvolver uma metodologia para tratar problemas de bases multivariadas que variam no tempo. Enquanto as abordagens tradicionais tentam eliminar uma das dimensões: tempo, atributos ou registros, neste experimento desenvolveu-se novas variáveis que sintetizam três características principais da evolução temporal para cada variável, ou seja, em uma base de dados onde pretendemos estudar algum fenômeno que num determinado momento do tempo se destaca do restante da base, este movimento de separação pode ser observado através da mudança de comportamento de algumas das variáveis, seja se afastando da média histórica com uma inclinação ou tendência, pode apresentar um novo padrão de volatilidade ou pode alterar seu patamar de valor médio, conforme estudo a seguir.

Para completar o experimento realizou-se modelagem não supervisionada para apresentar algumas abordagens da literatura e para exemplificar o que pode ser observado quando não se tem a classe de controle.

3.3 Estatísticas descritivas básicas

O primeiro passo na etapa de reconhecimento de uma base de dados é analisar as medidas descritivas da mesma, tanto da base como um todo, como se separando os grupos dos ativos e inativos.

As métricas utilizadas nesta análise foram a média, desvio padrão, mínimo, máximo, quantidade de registros e percentil, abrindo esta informação por classe (churn ou não churn), mês e variável.

Inicialmente identificou-se tratar de uma base não balanceada, onde apenas 7% dos registros correspondem à clientes que evadiram a base da empresa. Este desbalanceamento deve ser considerado no momento das modelagens.

Através desta primeira análise descritiva observa-se que para algumas variáveis existe clara separação entre os clientes ativos e inativos. Esta análise aponta também valores máximos bastante discrepantes, podendo significar a presença de outliers. Para dar-se continuidade às comparações entre as variáveis faz-se necessário normalizá-las, visto que apresentam ordens de grandeza bastante diferentes.

Analisando-se os histogramas das variáveis originais (ANEXO I), calculado a partir da soma de cada variável para os 12 meses, observa-se os histogramas das variáveis V1, V6, e V9 estão bastante concentradas nos valores próximos de zero. Em parte isso se deve ao grande número de outliers presentes nestas variáveis. Aparentemente, com exceção da variável V8, as distribuições apresentam características exponenciais, o que pode indicar que uma boa opção para se normalizar estas variáveis seja a aplicação da função logarítmica, visto que as variáveis são muito assimétricas à direita.

Como o experimento trata de variáveis referentes aos minutos trafegados por cada cliente, para cada tipo de chamada, tem-se valores que variam de zero a milhares de minutos em um mês. O perfil de distribuição destas variáveis concentra-se em baixos valores e com uma maior dispersão entre os valores maiores que a média do que em relação aos valores abaixo da média, não representando uma distribuição normal. Tal característica faz-se relevante no momento em que pode impactar os resultados de alguns modelos estatísticos que pressupõem distribuições normais.

3.4. Evolução das variáveis ao longo dos meses

Para se estudar a evolução das variáveis ao longo dos meses foram considerados alguns métodos. O primeiro método foi estudar a evolução mensal da média das variáveis (ANEXO II). Neste modelo utilizou-se as dimensões *variável* versus *tempo*, utilizando como valor a média dos clientes.

Através destas análises conclui-se que as medias da base são bem parecidas com as dos ativos, o que é razoável visto que apenas 7% dos clientes da base estão inativos no 13º mês.

Em seguida realizou-se a análise gráfica (ANEXO II) da evolução das variáveis no tempo. Para isso inicialmente considerou-se a evolução da média das variáveis.

Para facilitar o trato com os meses da amostra denominou-se m1 o primeiro mês da amostra, agosto de 2004, e assim por diante até m12, julho de 2005.

Os gráficos em anexo sugerem vários características da base. Para serem consideradas no tratamento desta base, precisarão ser confirmadas através de novas análises.

A primeira característica identificada foi que o mês de outubro de 2004 (m3) provavelmente foi preenchido com valores errados visto as discrepâncias apresentadas para este mês principalmente nas variáveis V2, V4, V5, V7, V8 e V10. Como não se trata de um mês com características de sazonalidade, como por exemplo, os meses referentes ao período de férias escolares, tudo leva a crer que este mês precisará ser tratado separadamente.

Observa-se também através destes gráficos que o grupo dos inativos apresenta comportamentos parecidos com a média global até o nono mês (abril de 2005). A partir deste mês há uma clara diferenciação do comportamento dos ativos com os inativos. Em geral os clientes inativos no 13º mês apresentam aumento de tráfego em algumas das variáveis entre o 9º mês e o 11º mês. Ou seja, pelos gráficos da evolução mensal das médias dos clientes por variáveis observa-se uma aparente ruptura neste mês. Tal fato será melhor explorado posteriormente.

Como hipóteses inicial pode-se considerar que os clientes propensos a deixar a base aumentam o seu volume de tráfego aproximadamente quatro meses antes de efetivamente deixarem a base, sendo que em alguns casos no último mês o consumo cai bastante em função dos clientes já estarem deixando de usar os serviços da operadora.

Aprofundando mais esta análise observam-se três tipos de comportamento. O comportamento descrito acima ocorre claramente nas variáveis V1, V2. Ambas as variáveis referem-se ao tráfego em roaming, ou seja, fora da sua localidade contratada. A variável V2 refere-se à roaming internacional, ou seja, o cliente saiu do país. Este pode ser um dos motivos do cancelamento do serviço: Mudança de estado ou país. Em outras variáveis o cliente reduz o consumo mais cedo, como nas variáveis V3, V4, V5, V8. Estas variáveis se referem ao tráfego móvel-móvel e ao móvel fixo. Já nas demais variáveis o consumo aumenta no último mês, podendo indicar que este grupo de clientes gasta mais na véspera de deixar a base. Outra opção é o 12º mês também estar com problemas de preenchimento.

Neste momento do experimento estas hipóteses não passam de suposições que para serem consideradas precisam ser estudadas mais a fundo.

É relevante considerar que apesar do efeito visual e da aparente informação apresentada sobre a base, esta análise considera a média dos clientes, desta forma são impactadas por outliers inerentes a cada variável e a cada mês.

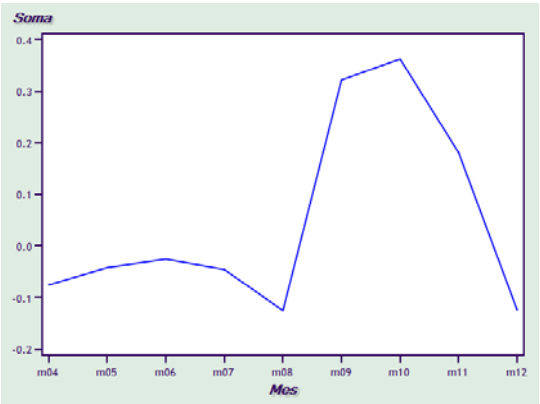
Até o momento temos dois itens que precisam ser analisados no processo de limpeza da base: o 3º mês (outubro de 2004) e os clientes outliers. Estes dois itens podem estar relacionados ou não, de qualquer forma devem ser tratados separadamente ao modelo preditivo.

Para se entender melhor a evolução temporal das variáveis e a sua relação ao longo do tempo, tentando identificar possíveis rupturas temporais e confirmar as observações anteriores, utilizou-se três modelos de análise:

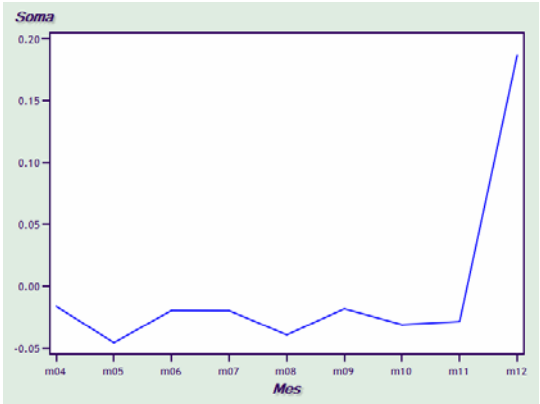
- Evolução da média standarizada;
- Evolução do determinante da matriz covariância;
- Análise dos AutoValores.

Para as três análises considerou-se o período posterior ao mês m3 que apresenta inconsistências e pode comprometer as análises.

Para a primeira análise normalizou-se as variáveis aplicando a função logaritmo (ln) e standarizou-se as variáveis calculando a média das variáveis por cliente por mês, e em seguida a média dos clientes por mês. Esta análise, além de reforçar a ruptura no 9º mês, indica que pode haver algum problema na base no 12º mês.



INATIVOS

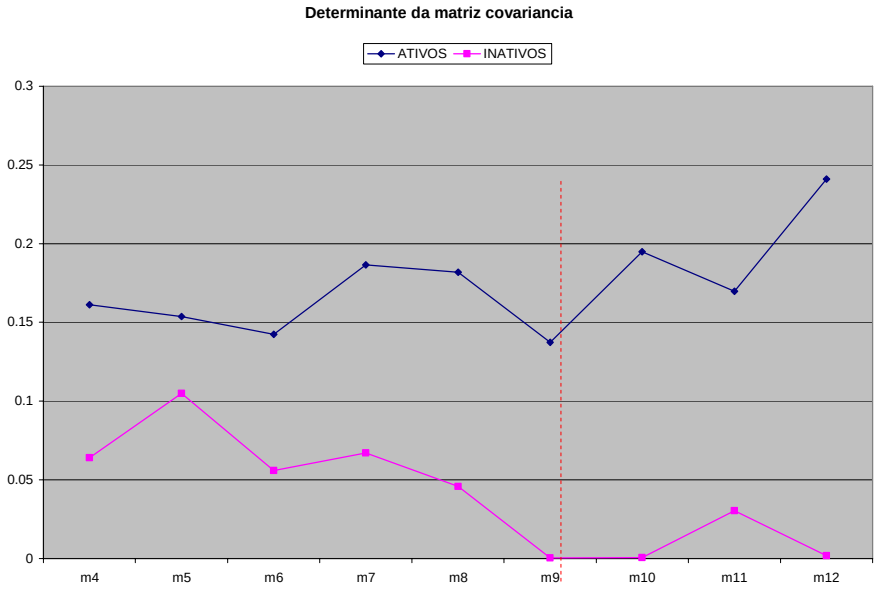


ATIVOS

Esta análise trata de uma visão sintética da base contemplando a média.

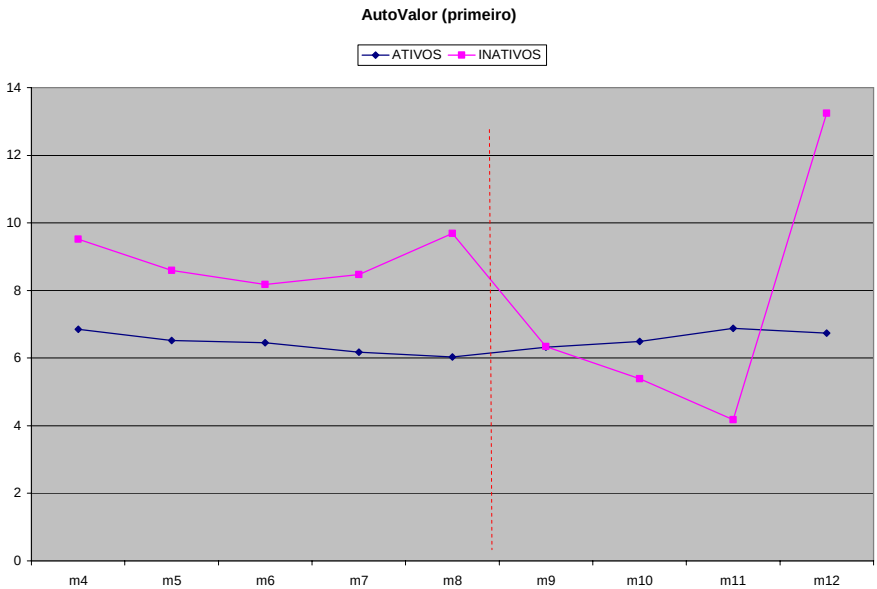
Analisando-se a evolução do determinante da matriz covariância observou-se que as curvas (ativos x inativos) apresentam comportamento semelhante nos primeiros meses, com patamares diferentes, até o 7º mês, quando as curvas passaram a apresentar tendências diferentes. Enquanto a curva dos ativos apresenta tendência de leve crescimento, a dos inativos apresenta clara tendência de queda, aumentando a diferença

de patamar entre as duas curvas. A partir do 9º mês a diferença de comportamento das curvas é bem clara.



Esta análise trata de uma visão sintética da base contemplando a variância.

Analisando-se os autovalores referentes à variância das variáveis transformadas através da análise de componentes principais, dando destaque ao primeiro e maior auto valor, observa-se uma clara ruptura do comportamento dos inativos em relação aos ativos do 8º para o 9º mês.



Esta análise trata de uma visão sintética da base contemplando a concentração da variância.

3.5 Variáveis Sintetizadas

A partir das análises da evolução temporal das 10 variáveis pode-se criar novas variáveis sintetizadas que podem explicar de forma diferente o experimento. Considerar-se-á 3 variáveis sintetizadas: Patamar (média dos meses 8 ao 11), Tendência (inclinação da evolução temporal através da interpolação por mínimos quadrados) e Volatilidade (desvio padrão).

Através de gráficos box-plot e das médias foi observado comportamento diferenciado entre o grupo dos cliente ativos e dos inativos. Por vezes eles apresentavam comportamento semelhante, mas, ou volume de tráfego mensal era constantemente inferior, gerando diferenciação de *patamar* de tráfego, ou, enquanto o grupo dos clientes ativos apresenta um comportamento homogêneo ao longo dos meses do ano, o grupo dos clientes inativos apresenta oscilações no seu volume de tráfego, gerando uma volatilidade acima da média dos clientes, ou uma tendência de crescimento ou queda de tráfego. Tais observações levaram a esta transformação dos dados, enriquecendo a análise e trazendo novas soluções para o problema estudado.

Analisando-se a evolução temporal da média das variáveis (ANEXO III) observou-se para que para algumas variáveis os grupos dos clientes ativos e inativos aparecem em patamares diferentes. Em geral o grupo dos clientes inativos apresenta um patamar menor que o dos ativos durante boa parte do período observador. Nos últimos 4 a 5 meses os clientes inativos mudam seu patamar em relação aos ativos, em geram indo para um patamar mais alto.

A *tendência* é calculada para cada variável considerando-se a inclinação da reta que melhor explica os valores dos meses 8 ao 11, utilizando regressão linear.

Estas novas 10 variáveis sintetizadas (ANEXO IV) apresentam características bastante diferentes das observadas até então.

Através da análise das variáveis de tendência observa-se que as variáveis TD_V1, TD_V6, TD_V7, e TD_V10 separam perfeitamente as classes dos ativos e inativos. As demais variáveis separam pouco ou nada as classes. Diante deste cenário pode-se considerar somente as variáveis “boas” (listadas acima) para os próximos modelos.

As variáveis de *volatilidade* são calculadas a partir do desvio padrão de cada variável nos 4 meses analisados.

As variáveis que mais separam as classes individualmente são as variáveis DP_V3, DP_V4, DP_V5 e DP_V8.

3.6 Variáveis Derivadas

Neste capítulo buscou-se criar/identificar novas variáveis, derivadas das variáveis originais, que possam agregar valor ao estudo e ajudar na modelagem do problema.

Foram estudadas variáveis derivadas (ANEXO VI) conceituais do ponto de vista do negócio de telefonia, e variáveis provenientes do comportamento estudado das variáveis.

Para cada uma das variáveis criadas foram estudadas os seus desdobramentos nas variáveis sintetizadas: patamar(média), tendência(inclinação) e volatilidade(desvio padrão).

A primeira variável estudada foi o somatório das variáveis V3, V4 e V5, que são variáveis relacionadas ao tráfego VC1 (tráfego móvel local). Esta nova variável foi normalizada através da função logaritmo (ln) e consolidada para os meses 8, 9, 10 e 11.

A segunda variável é com posta pela soma das variáveis V7 e V9. Pelo sentido do negócio de telefonia, ambas referem-se ao tráfego roaming nacional.

A terceira variável é uma equação onde os elementos do numerador apresentam tendência de crescimento nos meses estudados, enquanto as variáveis do denominador apresentam queda no mesmo período: (V1+v7)/(v4+v6+v10)

A quarta variável, no mesmo propósito e construção da variável anterior é simplesmente V7/V6.

Pela análise dos histogramas (ANEXO VI) a variável de tendência V7/V6 (VD4tend) foi a que apresentou melhor separação entre as classes.

Pela análise scatter plot das variáveis duas a duas não foi observado nenhuma combinação de 2 variáveis que separe as classes, o que não significa que possam haver outras combinações de mais variáveis que separem melhor as classes.

A redução de 12 variáveis para 8 componentes principais indica que existem algumas variáveis correlacionadas que podem ser explicadas em 98% dos casos por estas componentes. A ACP não é uma análise que considere a separação das classes.

O gráfico da projeção das variáveis transformadas pela ACP não aponta uma separação das classes.

3.7 Comparação com a abordagem na literatura

Boa parte da literatura na Mineração de Dados Temporais trata de séries temporais univariadas (uma componente). Séries temporais com uma única variável são tipicamente mapeadas em vetores de parâmetros com tamanho fixo através de modelos de Fourier, wavelet ou ARMA (Auto Regressive Moving Average), onde após a modelagem os vetores são clusterizados. Estes métodos, contudo, não são apropriados para séries temporais multivariadas. Fourier ou wavelet são indefinidos ou de implementação computacional cara em mais de duas ou três dimensões. Modelagens ARMA necessitam de premissas de linearidade que torna restritivo o seu uso. Modelos não lineares para parametrização temporal multivariada existem, mas sua computação normalmente requer múltiplos passos de reinderização (rendering), inviável para dados massivos (volumosos).

Para dar prosseguimento às análises faz-se necessário eliminar uma das 3 dimensões (tempo, variáveis e clientes) da base. Diante desta situação, em estudo realizado por CISTER, ANGELO MAIA [1], consolidou-se a dimensão das variáveis na sua média aritmética para cada mês. Assim analisou-se a base de clientes x tempo onde a variável analisada era a média aritmética das variáveis apresentadas na base.

No presente experimento a análise considerará uma base de clientes x variáveis, onde o tempo será tratado segundo alguns cenários (ANEXO VII). Posteriormente considerar-se-á outras variáveis derivadas.

Inicialmente considerou-se uma base constituída pela soma dos meses m9, m10 e m11, meses onde a maioria dos clientes churn (inativos no 13º mês) aumentam o volume de tráfego.

Analisando-se os histogramas destas variáveis conclui-se que não existe uma separação perfeita entre as duas classes em cada variável. As variáveis que melhor separam as classes são as variáveis V3, V4, V5 e V8.

Realizando-se análise das variáveis duas a duas (scatter plot) não se identifica combinação de duas variáveis que separe as duas classes.

A análise de Componentes principais destas variáveis atinge 99% com 4 componentes. Com as duas principais componentes o modelo é explicado em aproximadamente 70%. Realizando esta mesma análise com as variáveis normalizadas

observou-se que para se atingir 90% de explicação do modelo são necessárias 7 componentes.

A ACP não possui relação com a classificação, mas sim dá indícios para uma possível redução de variáveis.

O gráfico da projeção das 2 principais componentes principais não apresenta separação das classes.

A análise boxplot das variáveis originais aponta uma dispersão grande dos dados (muitos outliers).

Pela análise box-plot vas variáveis mês a mês separadas pelas classes observa-se que o grupos dos clientes ativos apresenta comportamento homogêneo ao longo dos meses, enquanto o grupo dos inativos apresenta alterações de comportamento.

O tratamento dado às bases temporais na literatura em muitos casos consiste na elilminação da

Capítulo 4 – RESULTADOS OBTIDOS

Neste capítulo procurou-se demonstrar que a metodologia de identificação do grupo de clientes propensos ao cancelamento (churn) facilita a grande maioria das abordagens e modelagens existentes, tanto supervisionadas como não supervisionadas.

O modelos estudados exemplificam diferentes abordagens matemáticas possíveis, cada uma com suas vantagens e desvantagens, sendo que para se comparar os resultados apresentados considerou-se o critério de classificação por percentil. Neste critério prioriza-se os clientes a serem contatados mediante uma limitação de recursos da empresa, seja financeira, pessoal ou tempo, informando o percentual de sucesso neste contato.

4.1 Modelagem Supervisionada (no SAS)

Através da modelagem supervisionada espera-se a obtenção de resultados mais precisos, visto que os padrões de saída são dados ao sistema para que a rede aprenda associar os estímulos corretamente.

Para não fugir do foco desta dissertação no estudo da metodologia apresentada, e para não repetir desenvolvimentos disponíveis no mercado com sistemas de modelagem e datamining, utilizou-se o sistema SAS Miner para demonstrar os resultados obtidos.

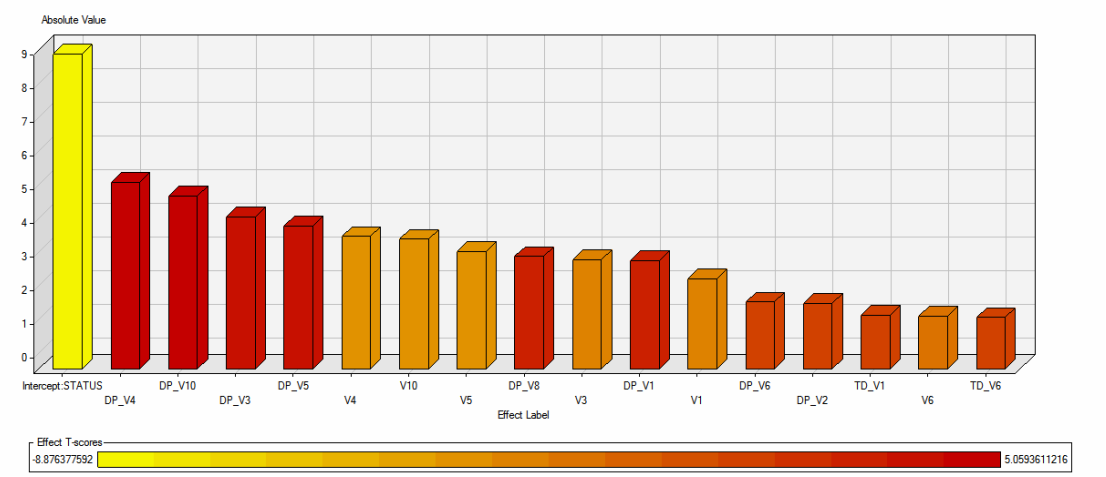
Inicialmente foram simulados modelos básicos no SAS Miner, sem utilizar as variáveis sintetizadas elaboradas nesta dissertação.

Ao se buscar as formas ótimas de normalização pelo SAS foi proposto a utilização da função logarítmica, mesmo resultado obtido mediante análise dos histogramas, confirmando os critérios adotados anteriormente.

Considerando a base normalizada (ln) dos meses 8 à 11, e considerando as variáveis sintetizadas (patamar, tendência e volatilidade), obteve-se os seguintes resultados para as respectivas análises:

a) Regressão Logística

Inicialmente realizou-se a modelagem das variáveis originais não normalizadas, onde aplicando Regressão Logística o modelo proposto pelo SAS considerou 16 variáveis como principais para a modelagem.



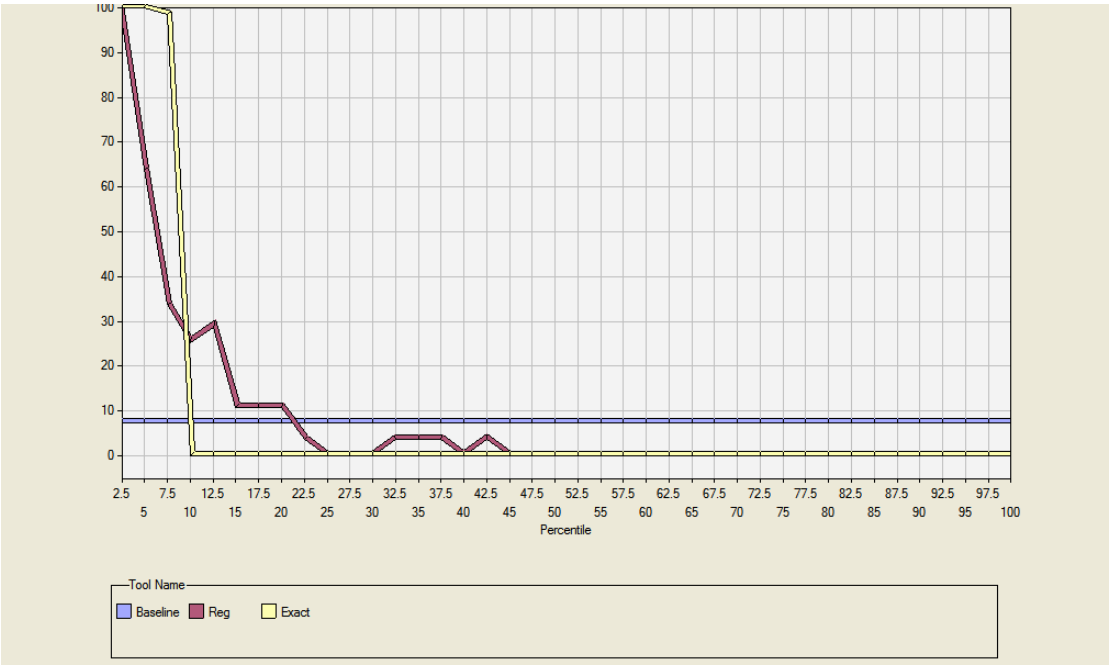
As principais variáveis consideradas foram a Volatilidade de V4, V10, V3 e V5. O grande número de variáveis deve-se, em parte, como mecanismo matemático para compensar as variáveis não normalizadas.

Table of F_STATUS by I_STATUS

F_STATUS(From: STATUS)
I_STATUS(Into: STATUS)
Frequency,
Percent,
Row Pct,
Col Pct,

	-1	1	Total
-1	46	37	83
	4.14	3.33	4.46
	55.42%	44.58%	
	79.31%	3.51%	
1	12	1,017	1029
	1.08	91.46	92.54
	1.17%	98.83%	
	20.69%	96.49%	
Total	58	1,054	1112
	5.22	94.78	100

Pela ausência de uma matriz de custos, o modelo foi afetado pelo desbalanceamento da base, sem contar que não considerou que é mais importante acertar os clientes churn do que errar os clientes não churn.

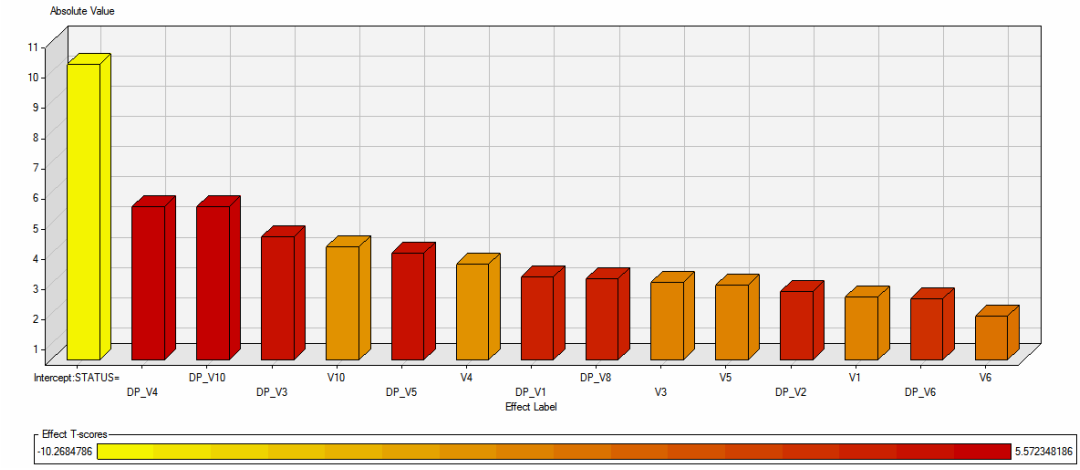


O modelo de regressão logística, que considerou todas as variáveis, apresentou resultado muito bom. Para o percentil 10 o modelo consegue garantir 30% de clientes churn.

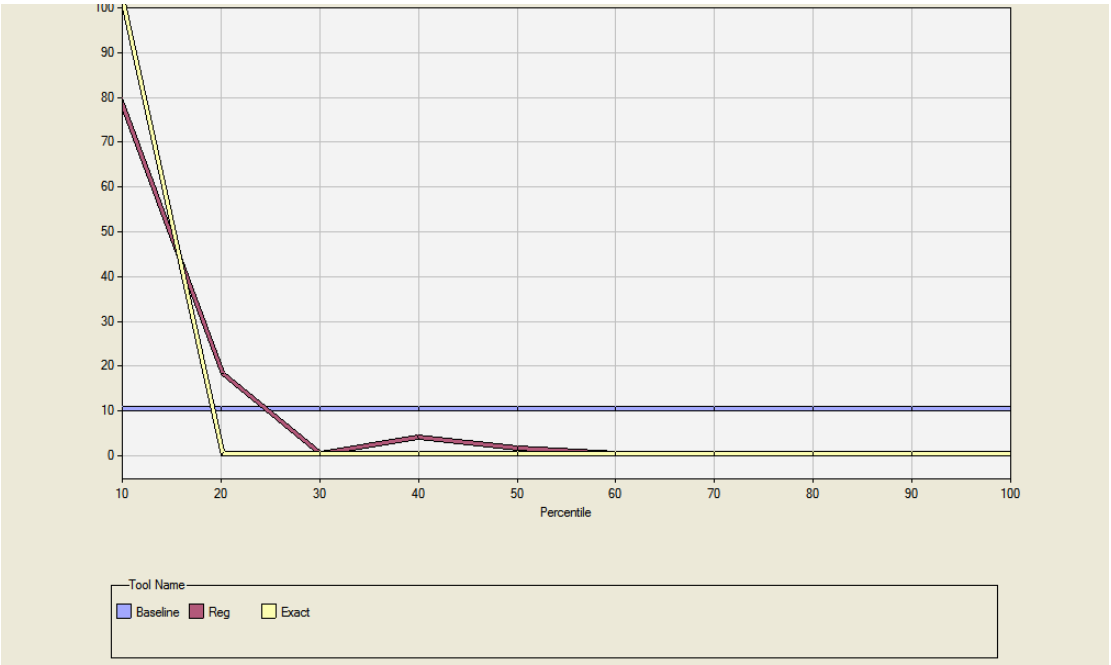
b) Regressão Logística Stepwise (profit/loss , entry 0,3 stay 0,05)

O objetivo da regressão, também chamada de identificação de modelos, é o desenvolvimento de um modeo numérico de um processo real. Neste caso as variáveis de entradasão utilizadas pelo modelo para realizar uma estimativa das variáveis de saída de forma a minimizar o critério de erro, que é utilizado para o ajuste do modelo.

Neste modelo foi utilizada a regressão logística Stepwise com critério de significância 0,3 para entrada e 0,05 para saída.



Como era esperado para uma regressão stepwise, menos variáveis foram selecionadas.



c) Árvore de Decisão

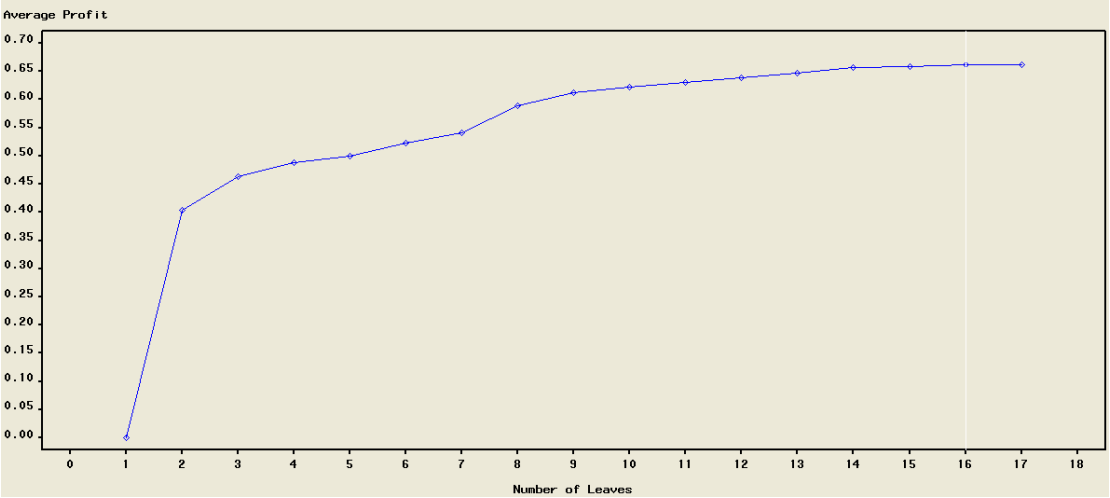
As árvores de decisão são uma evolução das técnicas que apareceram durante o desenvolvimento das disciplinas de machine learning. Essa análise trabalha, testando automaticamente, todos os valores do dado para identificar aqueles que são fortemente associados com os itens de saída selecionados para exame. Os valores que são encontrados com forte associação são os prognósticos chaves ou fatores explicativos, usualmente chamados de regras sobre o dado.

As árvores de decisão são meios de representar resultados de DM na forma de árvore, e que lembram um gráfico organizacional horizontal. Dados um grupo de dados com numerosas colunas e linhas (registros com as devidas variáveis), uma ferramenta de árvore de decisão pede ao usuário para escolher uma das colunas como objeto de saída, e aí mostra o único e mais importante fator correlacionado com aquele objeto de saída, como o primeiro ramo (nó) da árvore de decisão. Os outros fatores são subseqüentemente classificados como nós do(os) nó(s) anterior(es). Isso significa que o usuário pode, rapidamente, ver qual o fator que mais direciona o seu objeto de saída e entender o porquê do fator ter sido escolhido. Uma boa ferramenta de AD vai, também, permitir que o usuário explore a árvore de acordo com a sua vontade, do mesmo modo que ele poderá encontrar grupos alvos que lhe interessem mais, e aí, ampliar o dado exato associado ao seu grupo alvo. Os usuários podem, também, selecionar os dados

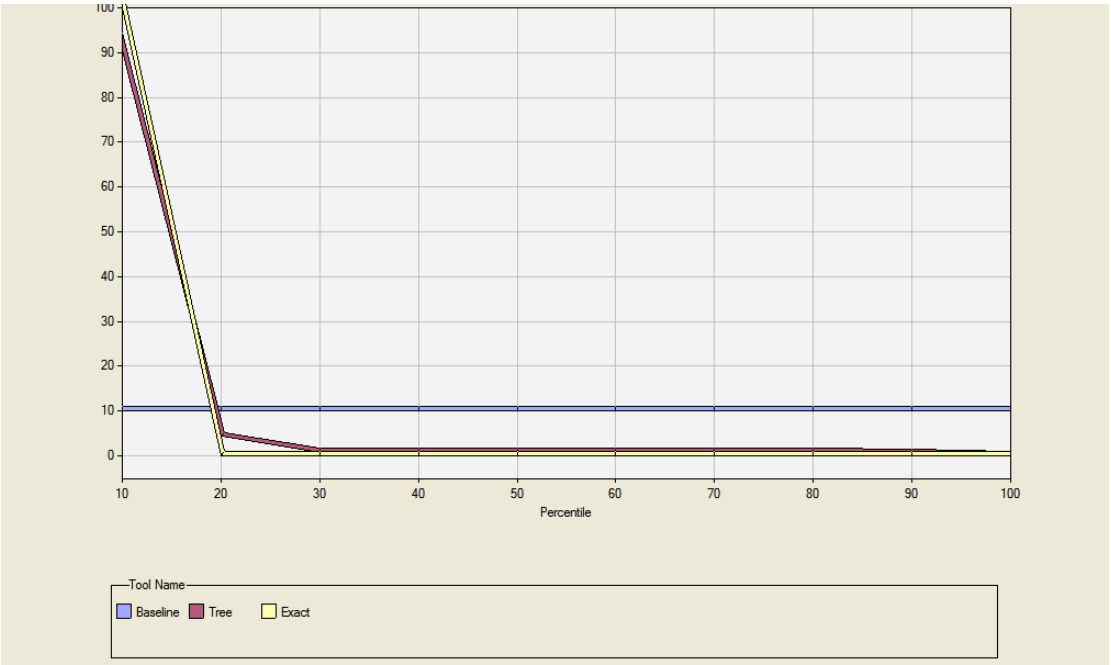
fundamentais em qualquer nó da árvore, movendo-o pra dentro de uma planilha ou outra ferramenta para análise posterior.

As árvores de decisão são, quase sempre, usadas em conjunto com a tecnologia de Indução de Regras, mas são únicas no sentido de apresentar os resultados da Indução de Regras num formato com priorização. Então, a regra mais importante é apresentada na árvore como o primeiro nó e as regras menos relevantes são mostradas nos nós subseqüentes. As vantagens principais das árvores de decisão são que elas fazem decisões levando em consideração as regras que são mais relevantes, além de serem compreensíveis para a maioria das pessoas. Ao escolher e apresentar as regras em ordem de importância, as árvores de decisão permitem aos usuários ver, na hora, quais fatores mais influenciam os seus trabalhos.

%Capture Response/non cumulative



A modelagem utilizando árvore de decisão aponta para um resultado próximo do ideal, sugerindo over training.

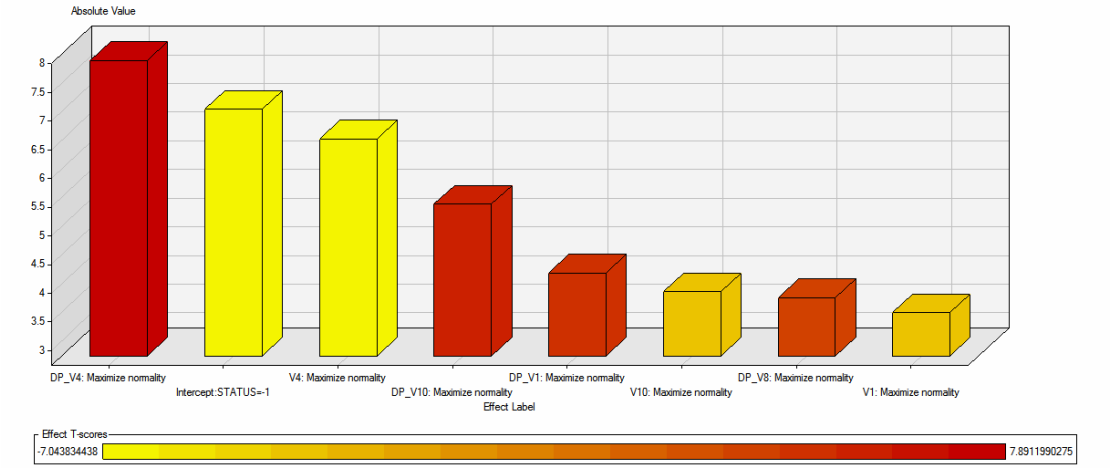


O gráfico com a taxa de acerto por percentil apresenta um resultado de classificação muito bom, muito próximo do ideal, facilitando para a empresa realizar ações de retenção para um pequeno grupo de clientes propensos ao cancelamento.

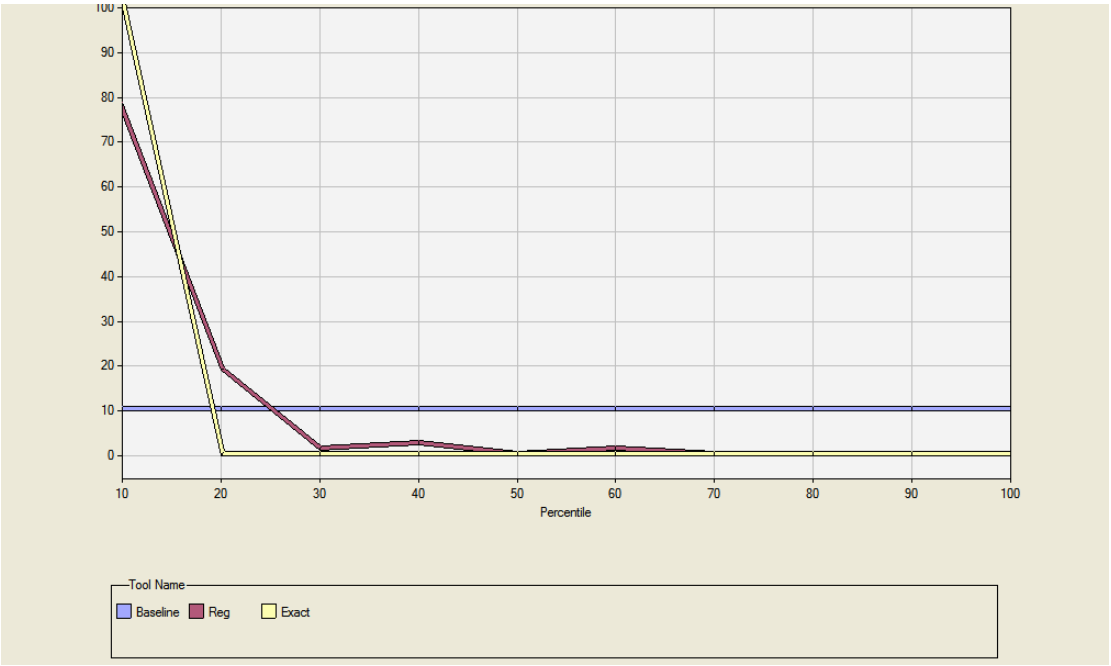
d) Regressão Stepwise com as variáveis transformadas

Foram sugeridas variáveis normalizadas, na sua maioria através da função log, ou seja, a avaliação do sistema para a melhor forma de normalização dos dados é similar à adotada nas análises anteriores..

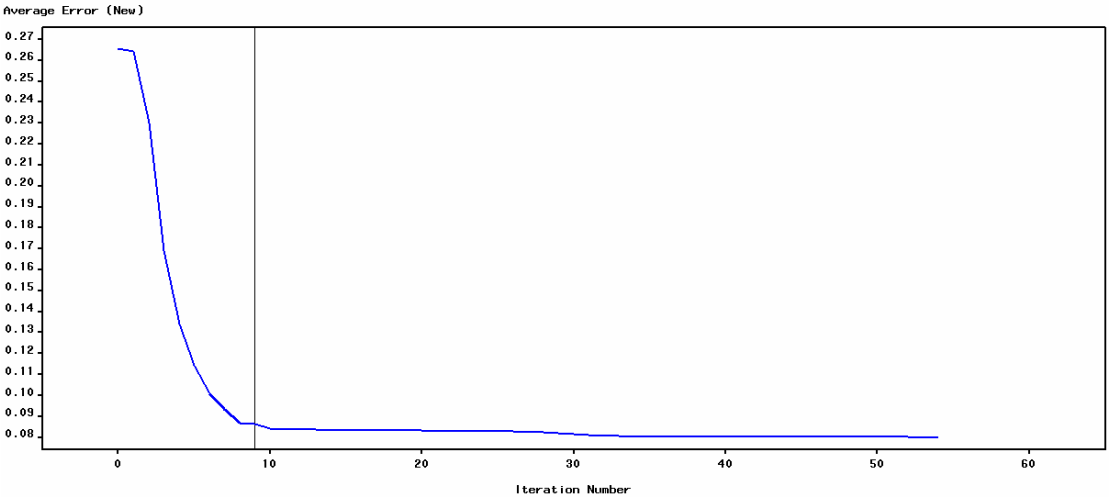
Tanto as médias como os desvios padrão foram normalizados. As variáveis de tendência foram sugeridas para continuar como estavam.

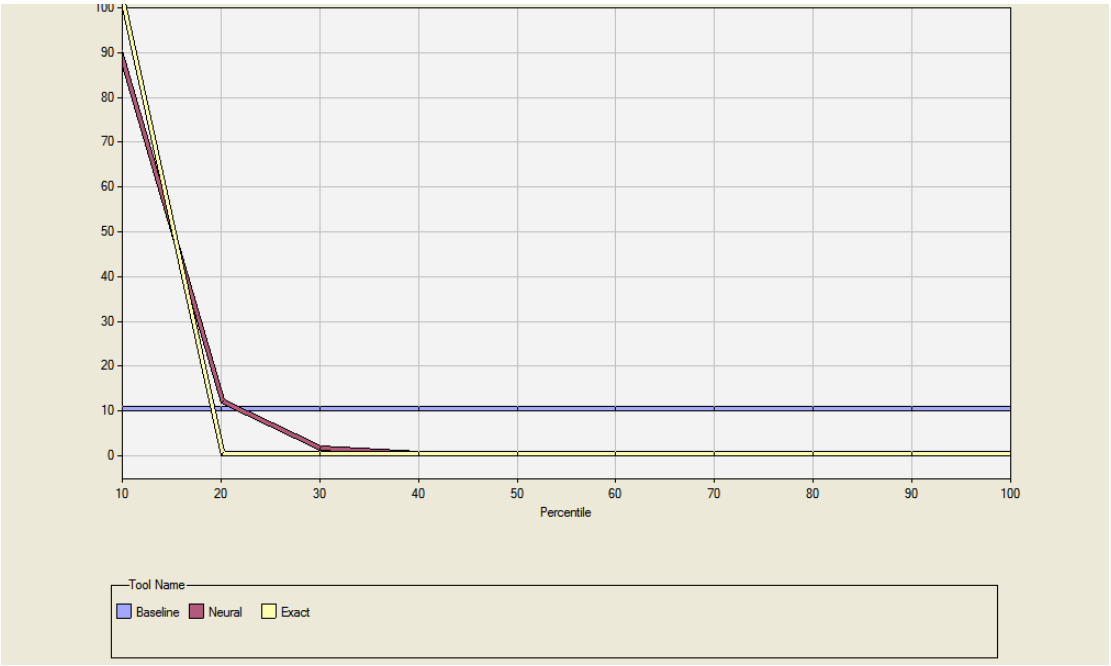


Após utilização das variáveis transformadas, o modelo stepwise selecionou menos variáveis.

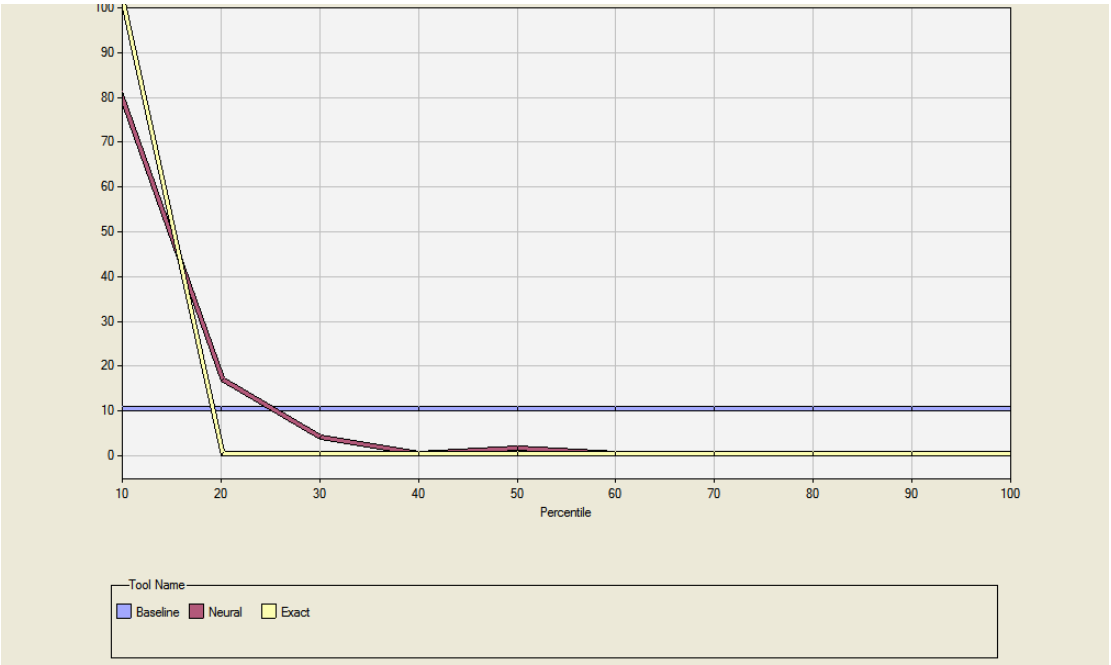


e) Redes Neurais Multilayer Perceptron sobre vars transformadas





f) RNA considerando apenas as principais variáveis



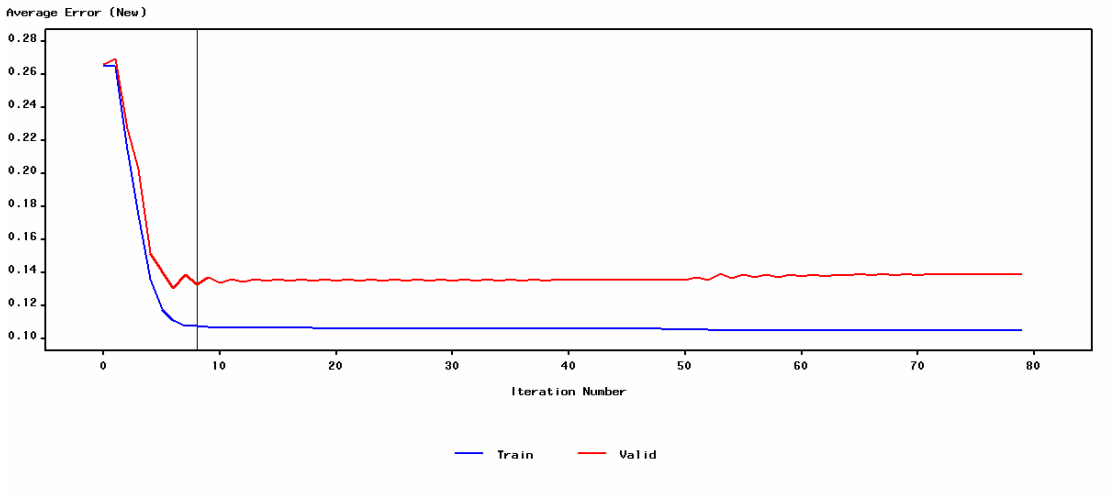
Variáveis consideradas: DP_V4, V4, DP_10, DP_1, V10, DP_V8 e V1.

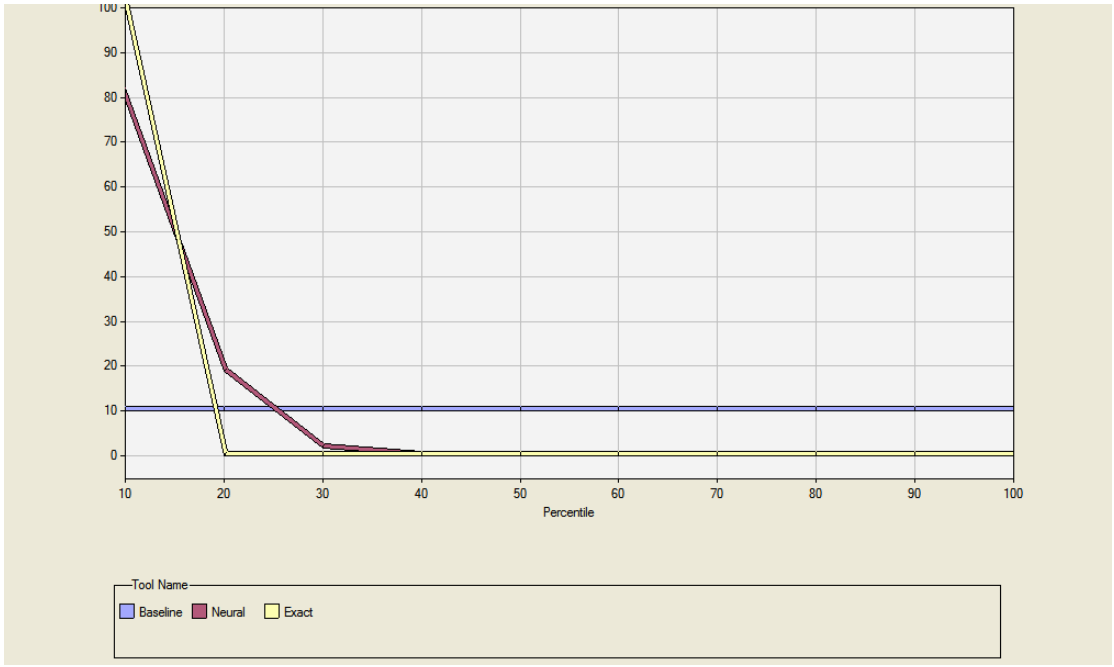
A vantagem de usar menos variáveis é a maior robustez do modelo. Na prática do mercado prefere-se assumir resultados piores de classificação para poder ter modelos mais ágeis e robustos.

g) Data Partition (Test 70% - Validation 30%)

Apesar do pequeno número de registros de clientes inativos, realizou-se esta modelagem considerando amostra de 30% para validação.

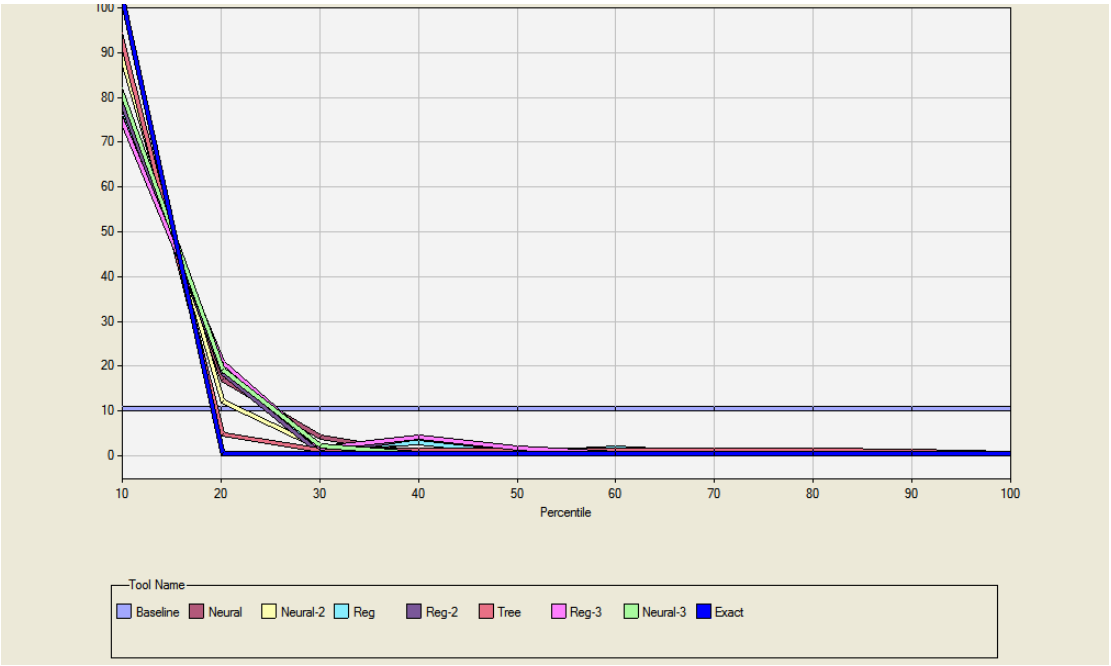
h) RNA Stepwise com validação





i) Comparativo dos modelos

A Árvore de Decisão apresentou o melhor resultado no percentil 10, seguido pela rede neural (g) e pela regressão linear.



Todos os modelos apresentaram resultados muito bons, sinal que a base selecionada e o trabalho de preparação dos dados, utilizando as variáveis sintetizadas, trouxe para o modelo um grau de acerto próximo do ideal.

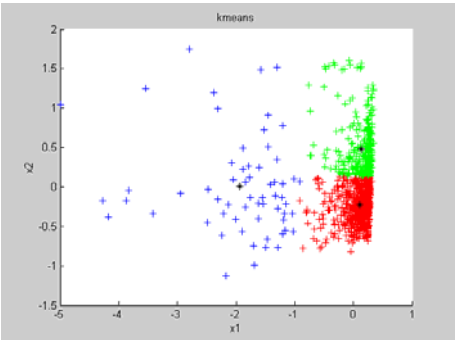
4.2 Modelagem não supervisionada

Na modelagem não supervisionada nenhum tipo de sinal ou informação é fornecido ao sistema. O sistema aprende por si as propriedades interessantes dos estímulos de entrada. O treinamento não-supervisionado é mais plausível em sistemas de aprendizagem cujo o sistema é biológico. Desenvolvido por Kohonen e outros pesquisadores, o sistema não necessita de vetores alvos (*target*). Nenhuma comparação é predeterminada para saída. O conjunto de treinamento consiste, somente, nos vetores de entrada. O algoritmo de treinamento modifica seu peso sinápticos para produzir os vetores de saída que são consistentes, i.e., mapeiam vetores que são suficientemente similares produzindo, então, os padrões de saída [WASSERMAN,1989].

A Classificação não supervisionada apresenta desvantagens em relação à supervisionada, pois o modelo não sabe a priori quem são os cliente churn e não churn, limitando-se a identificar grupos de clientes com características semelhantes e as movimentações de cleintes entre os clusters (observado nos gráficos pela movimentação dos centróides).

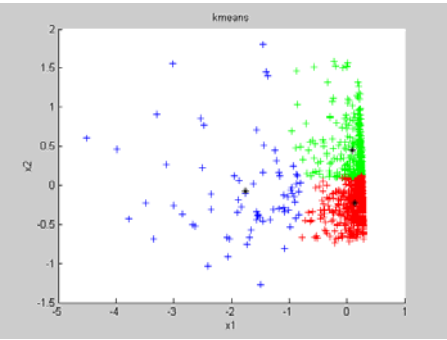
Na análise abaixo utilizou-se o modelo de kmeans para os meses m8, m9, m10 e m11.

M8



-1.9419 0.1040 0.1390
0.0113 -0.2245 0.4848

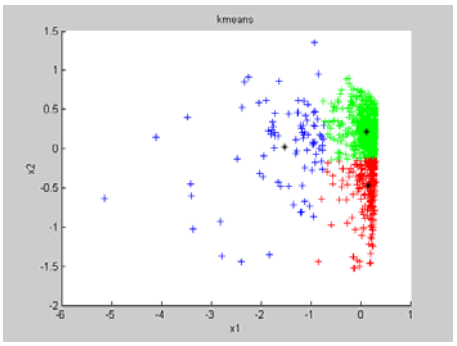
M9



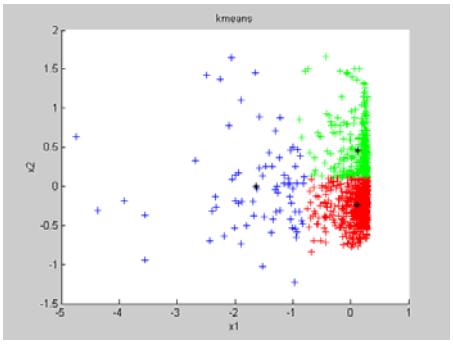
0.1397 0.0937 -1.7522
-0.2305 0.4518 -0.0731

M10

M11



0.1476 0.1218 -1.5223
-0.4718 0.2199 0.0221



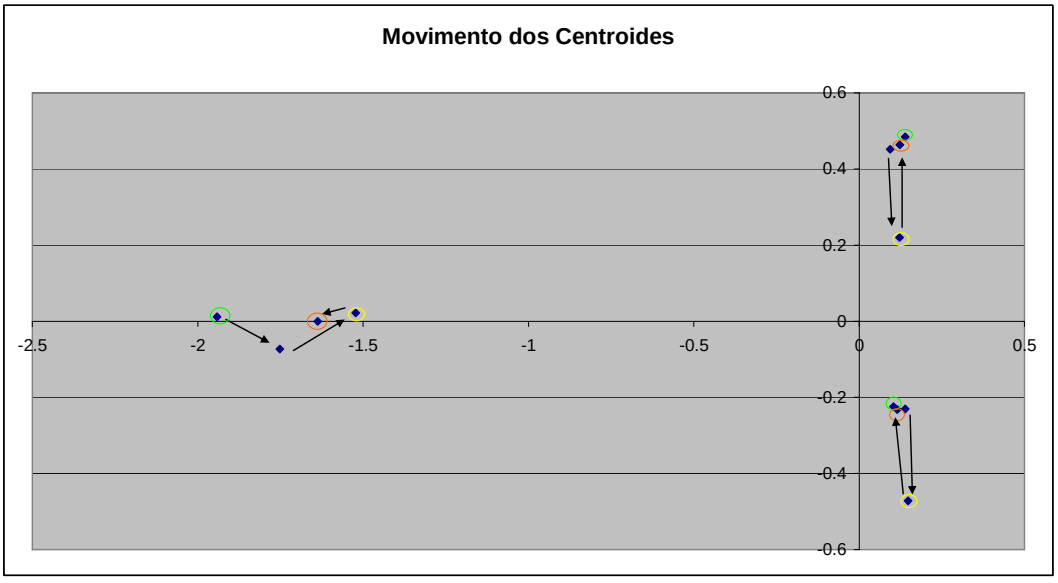
0.1150 0.1229 -1.6368
-0.2320 0.4635 -0.0005

Características: 3 clusters utilizando Análise de Componentes Principais (ACP):

Considerando-se a movimentação dos centróides identificou-se um maior afastamento no mês m10, o que pode apondar um mês de ruptura na evolução temporal dos dados. Nos capítulos anteriores observamos que de fato existe a ruptura temporal, e que apesar da análise não supervivionada não separar os grupos churn dos não churn, pode estar concentrando os clientes propensos a cancelamento em um ou alguns grupos, facilitando a abordagem pela empresa na tentativa de retê-los.

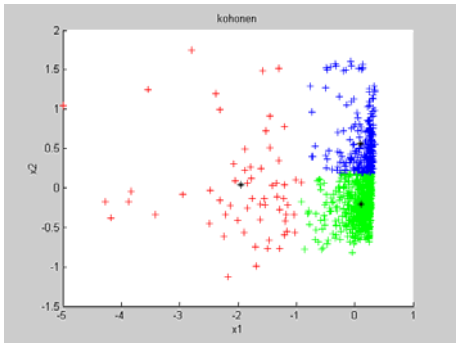
Quadro com a evolução no tempo dos centróides:

Mês 8			Mês 9			Mês 10			Mês 11		
-1.9419	0.104	0.139	0.1397	0.0937	-1.7522	0.1476	0.1218	-1.5223	0.115	0.1229	-1.6368
0.0113	-0.2245	0.4848	-0.2305	0.4518	-0.0731	-0.4718	0.2199	0.0221	-0.232	0.4635	-0.0005



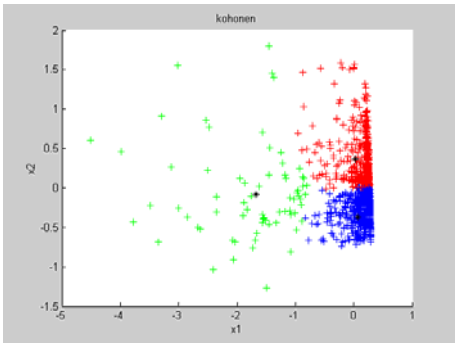
Através do modelo de classificação não supervisionada kohonen obteve-se resultados similares.

M8



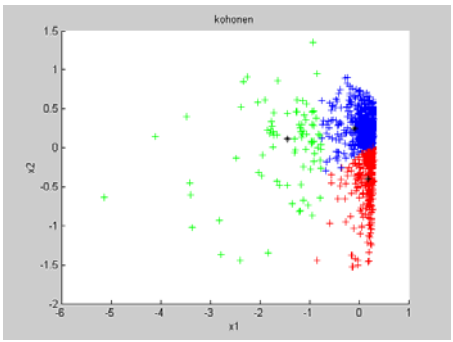
-1.9556 0.1050 0.0966
0.0392 -0.2080 0.5649

M9



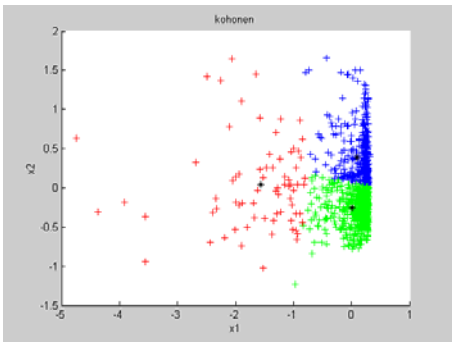
0.0270 -1.6799 0.0632
0.3707 -0.0793 -0.3682

M10



0.1831 -1.4429 -0.0880
-0.4030 0.1115 0.2444

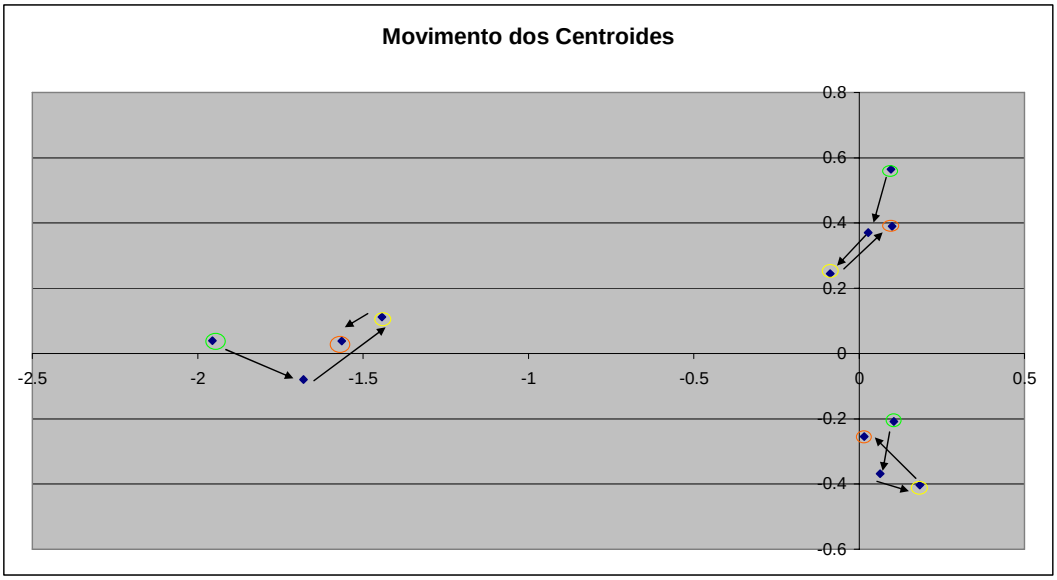
M11



-1.5649 0.0154 0.0995
0.0386 -0.2543 0.3899

Movimentação dos Centróides:

Mês 8			Mês 9			Mês 10			Mês 11		
-1.9556	0.105	0.0966	0.027	-1.6799	0.0632	0.1831	-1.4429	-0.088	-1.5649	0.0154	0.0995
0.0392	-0.208	0.5649	0.3707	-0.0793	-0.3682	-0.403	0.1115	0.2444	0.0386	-0.2543	0.3899



A análise de agrupamentos por kohonen é bastante similar à realizada utilizando kmeans. O mês 10 representa o maior afastamento em relação ao centróide do mês 8, sendo que o centróide do mês 11 volta a se aproximar do centróide do mês 8.

Capítulo 5 – CONCLUSÕES

Neste experimento analisou-se uma base de dados de telefonia, caracterizada por um determinado segmento de clientes heavy users (clientes com alto volume de minutos trafegados ao mês) de telefonia móvel, que possuam inclinação e/ou tendência a evadirem da base de dados da empresa.

A analise das características descritivas e temporais da base truxe conclusões importantes para a evolução do experimento, tais como a ruptura temporal, permitindo a utilização de um menor período de tempo para a análise ao invés da utilização dos 12 meses coletados, e a determinação das variáveis relevantes para a classificação dos grupos (propensos ao churn dos não propensos ao churn).

O comportamento médio da base analisada apresenta uma ruptura temporal entre o oitavo e o nono mês em algumas variáveis, apontando que as próprias variáveis, uma combinação delas ou algumas das suas características poderiam separar o grupo de clientes propensos ao cancelamento do grupo de clientes não propensos.

Através de três análises distintas (evolução temporal da média estandarizada, evolução temporal do determinante da matriz covariância e da análise de Auto Valores) confirmou-se as conclusões obtidas através da análise das médias no que diz respeito ao momento da ruptura temporal, facilitando o processo de análise com um período menor de meses necessários para a etapa de classificação.

Para a classificação das variáveis foi utilizada uma metodologia de predição através de 3 variáveis sintetizadas, patamar, tendência e volatilidade, onde, mais do que resolver um problema comum no mercado de telecomunicações, buscou-se consolidar a metodologia para análise genérica de problemas similares de predição em bases de dados temporais multivariados.

Cada uma das características utilizadas na sintetização de novas variáveis demonstrou extrema eficiência na representação da separação dos grupos que, ou estão com valores médios diferentes (patamares diferentes), ou apresentam algum comportamento de afastamento ou aproximação (comportamento de tendências diferentes) ou apresentam oscilações e volatilidades diferentes.

A análise de bases de dados temporais multivariados apresenta grande grau de dificuldade devido ao grande número de dimensões, o que extrapola os modelos básicos de análise multivariada.

Na literatura existem diversas formas de abordagem para estes casos, como por exemplo a consolidação das variáveis em uma única variável.

Analisando-se soluções não supervisionadas para a base estudada identificou-se uma movimentação dos centróides, significando que, na ausência de controle na modelagem (classe dos churn x não churn) podem auxiliar na identificação dos clientes que mudaram de cluster, indicando mudança no perfil de consumo, e assim disparando um alerta para ações de fidelização.

Os modelos utilizados demonstraram robustez e eficiência ao não necessitarem das informações contidas no último mês de amostra, o que é muito importante para análises reais, onde não há tempo hábil para se realizar a análise e tomar ações de retenção nos clientes candidatos a churn.

BIBLIOGRAFIA

1. AGRAWAL, R.,SRIKANT, R. ,."Fast algorithms for mining association rules", VLDB-94/1994.

2. BOGMANN, Itzhak Meir. Marketing de Relacionamento.São Paulo: Nobel, 2000.

3. CISTER, Angelo M.; EBECKEN, Nelson F. F. - "CRM through DM: a case study" – Third International Conference on Data Mining- DATA MINING III, Bolongna, Italy-2002.

4. CISTER, Angelo M.; EBECKEN, Nelson F. F. - "A screening tool based on neural networks" Conference on Data Mining- DATA MINING IV, Rio de Janeiro, Brazil-2003.

5. CISTER, Angelo M. - Mineração de dados para a análise de atrito em telefonia móvel [Rio de Janeiro] 2005.

6. CORRÊA, H. L.; GIANESI, I.G.N.; CAON, M. Planejamento, programação e controle da produtividade: MRPII/ERP conceitos, uso e implantação. São Paulo: Ed. Atlas S.A., 1997.

7. COUTINHO, Fernando V.; Datamining; www.dwbrasil.com.br/2003.

8. FAYYAD, Usama; et al - Advances in Knowledge Discovery and Data Mining - Mit Press 1a. Ed. 1996.

9. GROTH, Robert – Data Mining: Building Competitive Advantage – Prentice Hall PTR , 2000

10. HAN, Jiawei; KAMBER, Micheline, - Data Mining: Concepts and techniques- Morgan Kaufmann publishers, San Francisco – CA, 2001.

11. JONHSON, R. A., WICHERN, D. – Applied Multivariate Statistical Analysis, Prentice Hall: Upper Saddle River, New Jersey, 3ª. ed. .

12. KOTLER, P. & ARMSTONG, G. - Princípios de Marketing. - 7ª ed. Rio de Janeiro: Prentice Hall do Brasil, 1997.

13. KOTLER, Phillip. - Administração de marketing: análise, planejamento Implementação e controle. - 8.ª ed., SP: Atlas, 1998.

14. MATTISON, Rob. – Telecom Churn Management: The Golden Opportunity – APDG Plubkishing, 2001.

15. MCKENNA, Regis. - Marketing de Relacionamento: estratégias bem sucedidas para a era do cliente - 1ª ed. Rio de Janeiro: Campus. (1993). (trabalho original publicado em 1991).

16. NASCIMENTO, José Augusto. Programa de Fidelização e Clube de clientes. São Paulo: 1996 (seminário diretorial)

17. NOGUEIRA, Carlos F., - "Metodologia de Valorização de Clientes Utilizando Mineração de Dados" - Tese de Doutorado, COPPE-Cívil/UFRJ – Programa de Engenharia Cívil, jun. 2004.

18. PEPPERS, Don ; ROGERS, Martha. -One to One Manager: Real - World Lessons in Customer Relationship Management - The. New York: Currency/Doubleday,1999.

19. PESSOA, Luis Adauto F. C. - "Aprendizado Não-supervisionado em Redes Neurais" - Dissertação de Mestrado, COPPE-Sistemas/UFRJ - Programa de Engenharia de Sistemas, 1990.

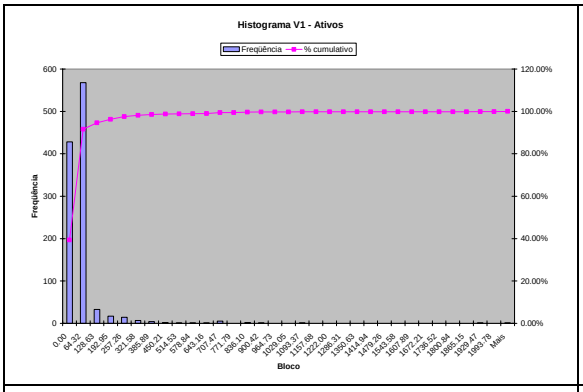
20. PINHEIRO, Carlos A. R. – "Redes Neurais para prevensão de inadimplência em Operadoras de Telefonia" – Tese de Doutorado, COPPE-Cívil/UFRJ – Programa de Engenharia Cívil, jul. 2005.

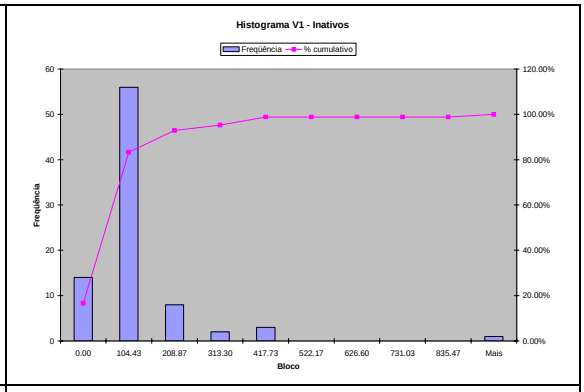
21. RUST, Roland T.; ZEITHAML, Valarie A.; LEMON, Katherine N. – Driving Customer Equity: How Customer lifetime Value is Reshaping – Free Press, 2000

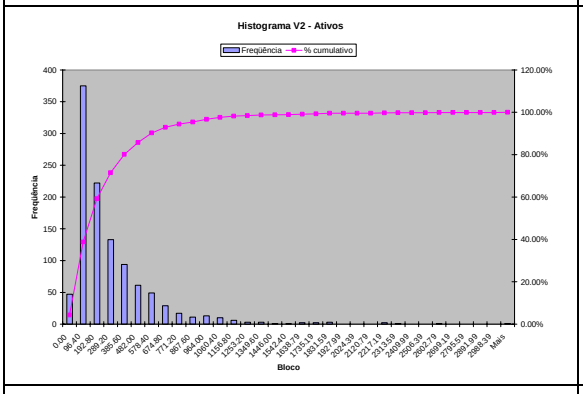
22. SHEPPARD, David Associates Inc.- The New Direct. Marketing: How to implement a profit-driven Database Marketing Strategy - a Handbook for Direct Marketing Company and Users of Direct Marketing Methods. Homewood, IL. Dow Jones-Irwin., 1990

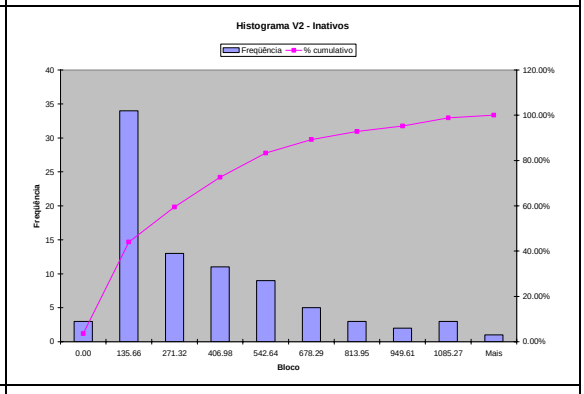
23. VAVRA, Terry G. Marketing de Relacionamento. São Paulo: Atlas, 1996.

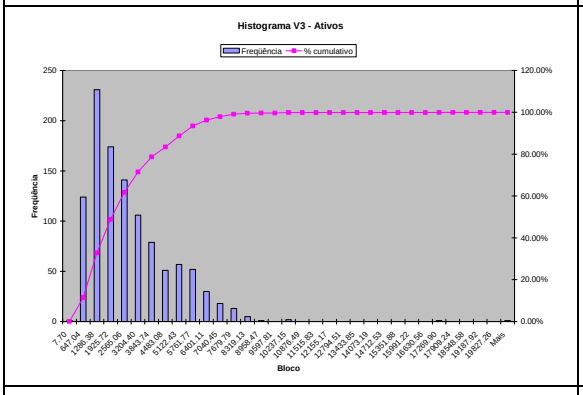
ANEXO I – Histogramas variáveis originais

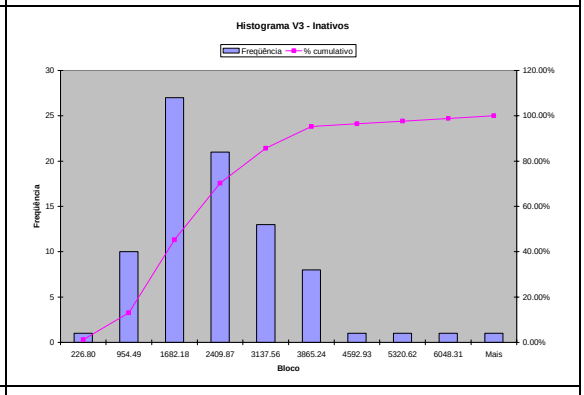


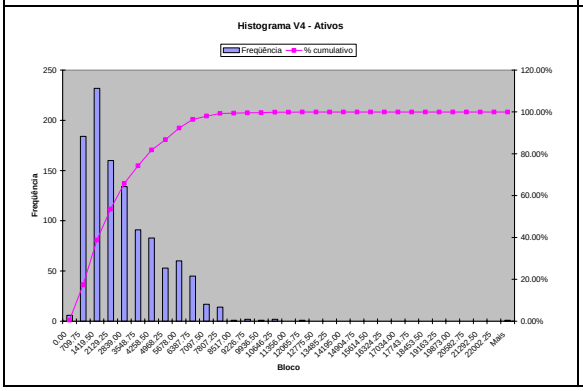


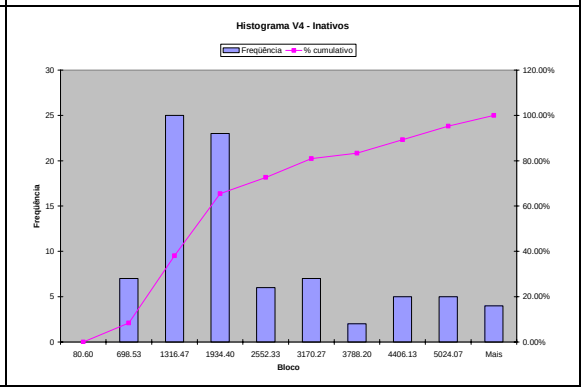


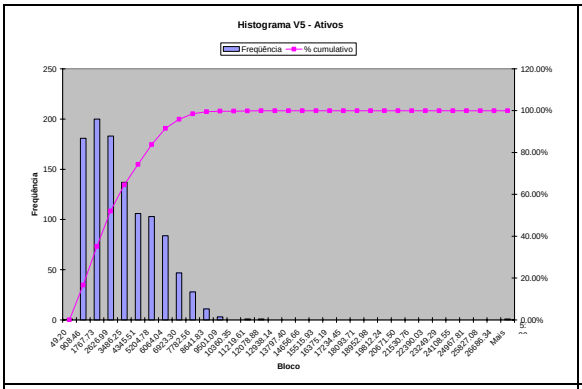


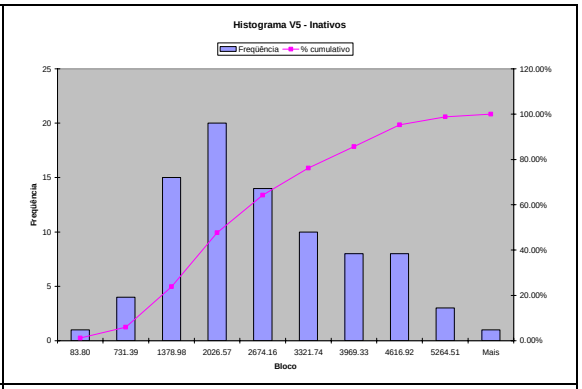


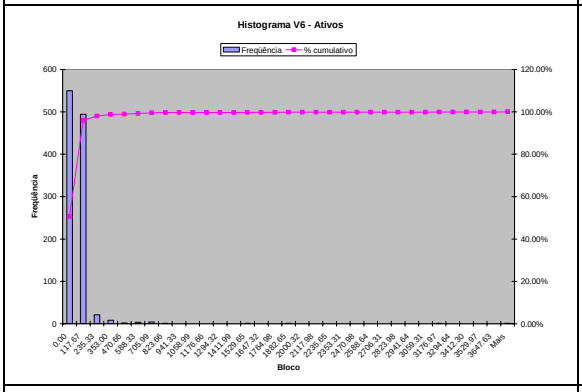


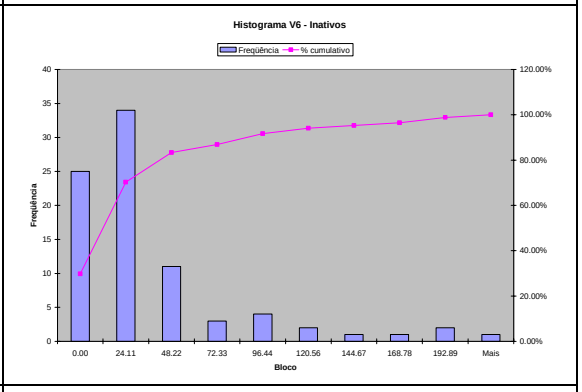


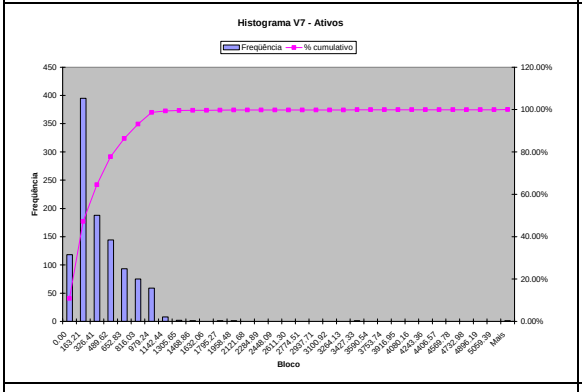


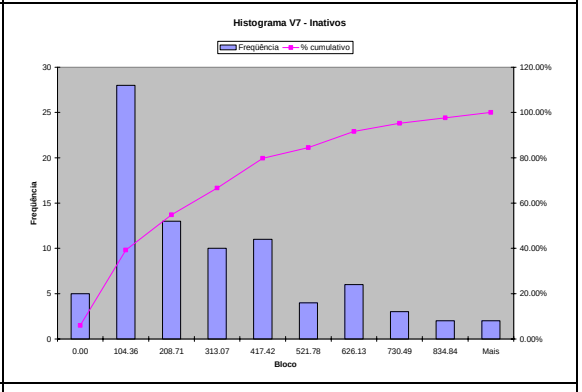

Histograma V5 - Ativos
Freqüência
% cumulativo
Frequência
Bloco

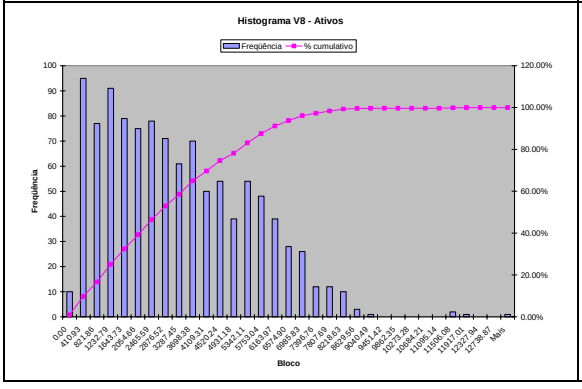

Histograma V5 - Inativos
Freqüência
% cumulativo
Frequência
83.80
731.39
1378.98
2026.57
2674.16
3321.74
3969.33
4616.92
5264.51
Mais
Bloco

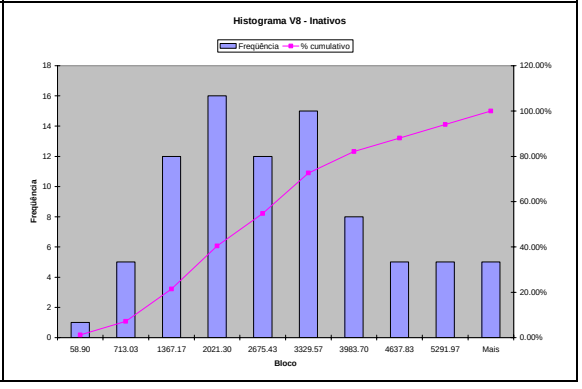

Histograma V6 - Ativos
Freqüência
% cumulativo
Frequência
Bloco

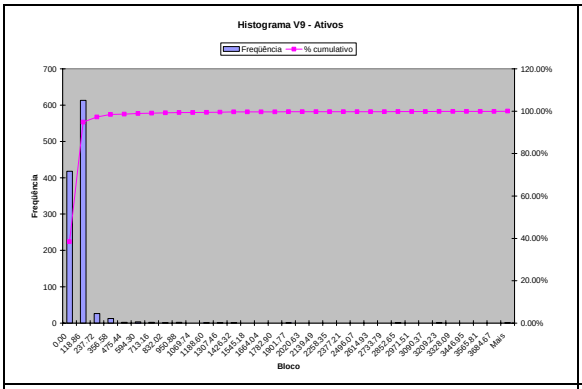

Histograma V6 - Inativos
Freqüência
% cumulativo
Frequência
0.00
24.11
48.22
72.33
96.44
120.56
144.67
168.78
192.89
Mais
Bloco

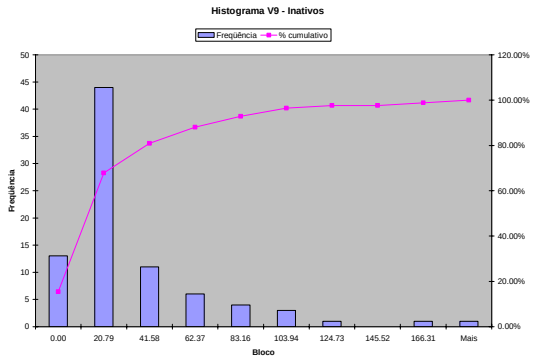

Histograma V7 - Ativos
Freqüência
% cumulativo
Frequência
Bloco

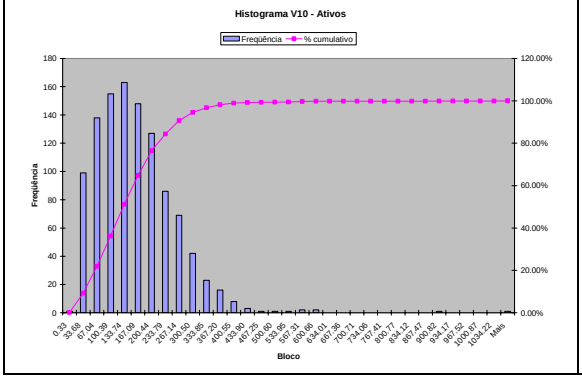

Histograma V7 - Inativos
Freqüência
% cumulativo
Frequência
0.00
104.36
208.71
313.07
417.42
521.78
626.13
730.49
834.84
Mais
Bloco

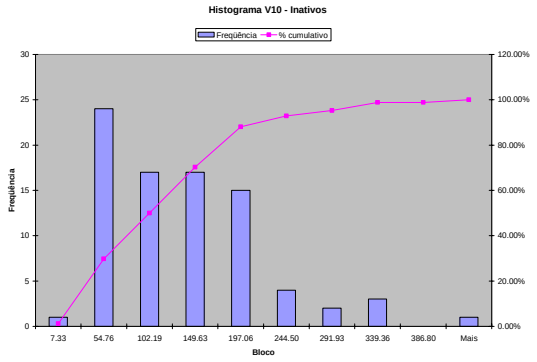

Histograma V8 - Ativos
Freqüência
% cumulativo
Frequência
Bloco


Histograma V8 - Inativos
Freqüência
% cumulativo
Frequência
58.90
713.03
1367.17
2021.30
2675.43
3329.57
3983.70
4637.83
5291.97
Mais
Bloco

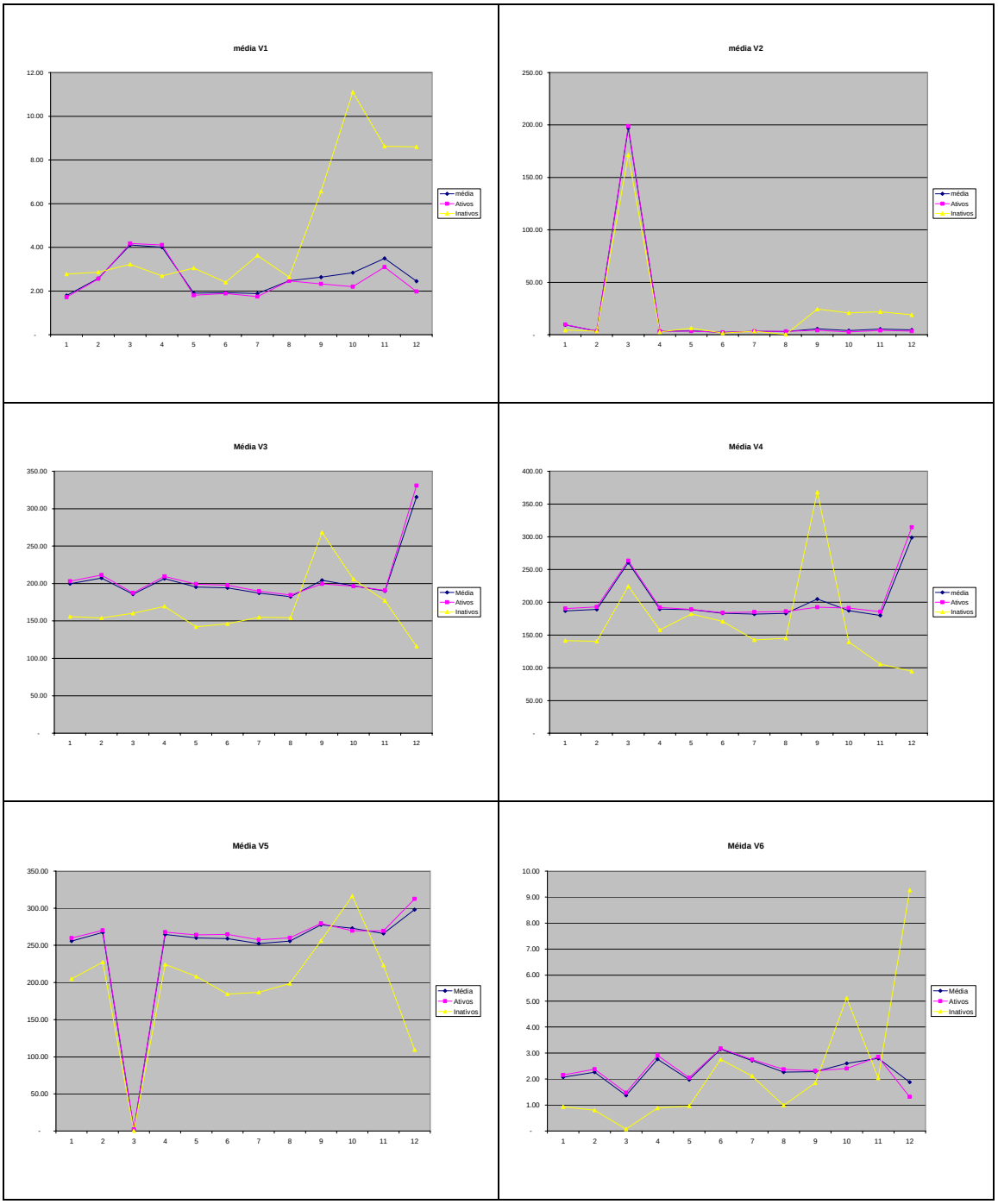

Histograma V9 - Ativos
Freqüência
% cumulativo
Freqüência
Bloco
0.00
118.86
237.72
356.58
475.44
594.30
713.16
832.02
950.88
1069.74
1188.60
1307.46
1426.32
1545.18
1664.03
1782.89
1901.75
2020.61
2139.48
2258.34
2377.20
2496.07
2614.93
2733.79
2852.65
2971.51
3090.37
3209.23
3328.09
3446.95
3565.81
3684.67
Mais

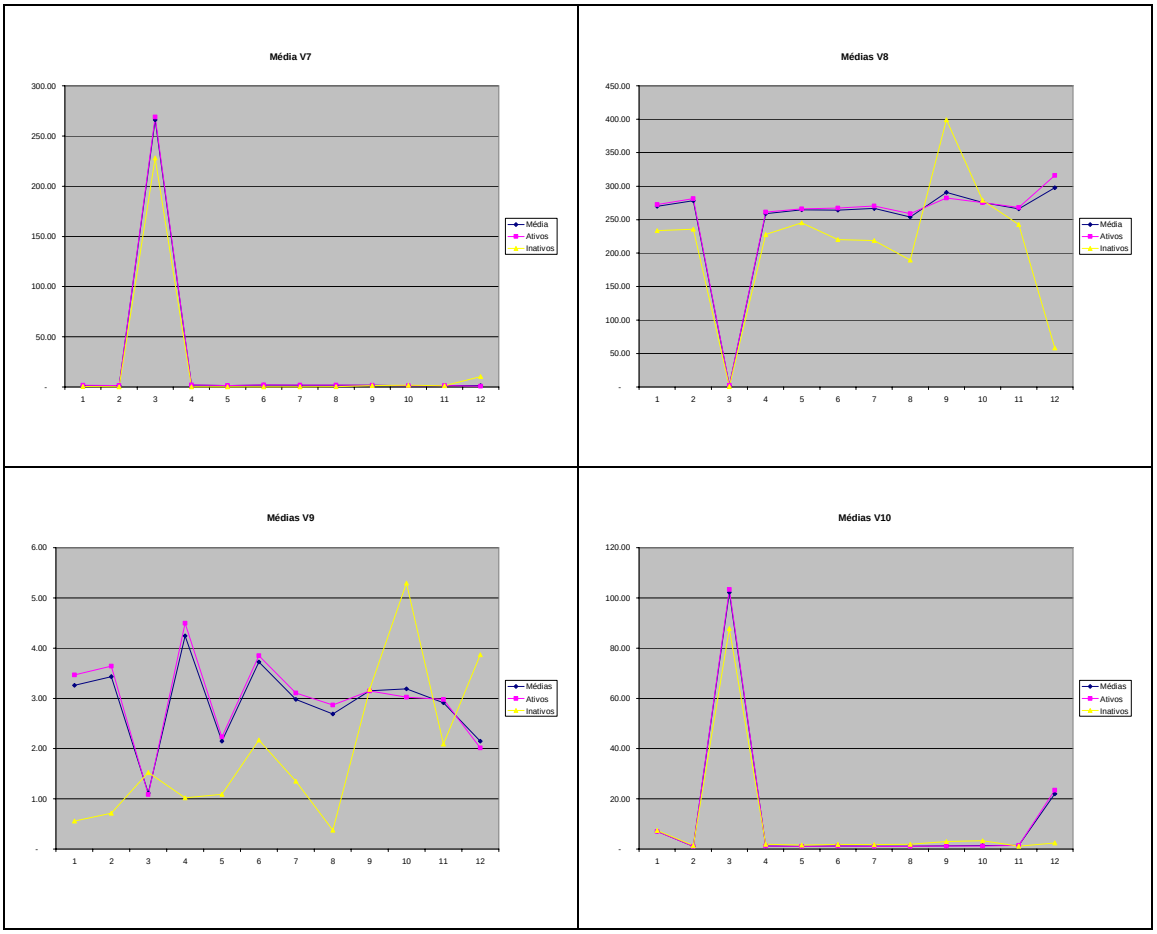

Histograma V9 - Inativos
Freqüência
% cumulativo
Freqüência
Bloco
0.00
20.79
41.58
62.37
83.16
103.94
124.73
145.52
166.31
Mais


Histograma V10 - Ativos
Freqüência
% cumulativo
Freqüência
Bloco
0.33
30.68
61.04
100.39
133.74
167.09
200.44
233.79
267.14
300.50
333.85
367.20
400.55
433.90
467.25
500.60
533.95
567.30
600.66
634.01
667.36
700.71
734.06
767.41
800.77
834.12
867.47
900.82
934.17
967.52
1000.87
1034.22
Mais

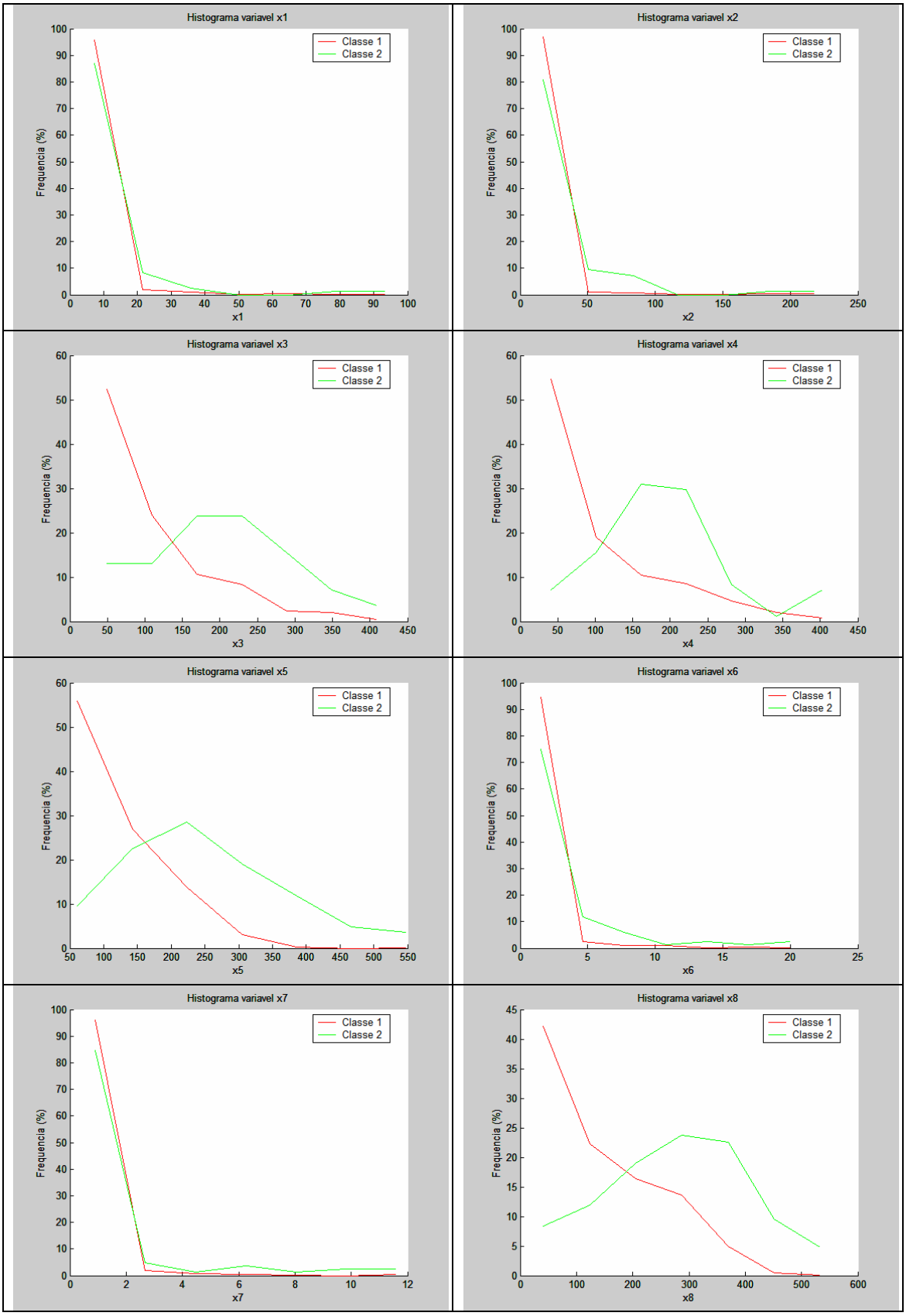

Histograma V10 - Inativos
Freqüência
% cumulativo
Freqüência
Bloco
7.33
54.76
102.19
149.63
197.06
244.50
291.93
339.36
386.80
Mais

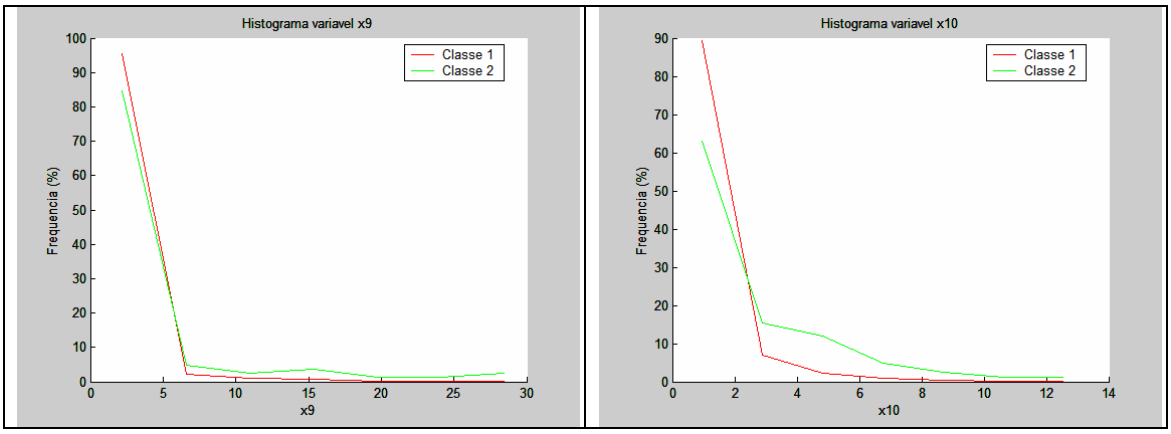
ANEXO II – Evolução temporal das medias das variáveis

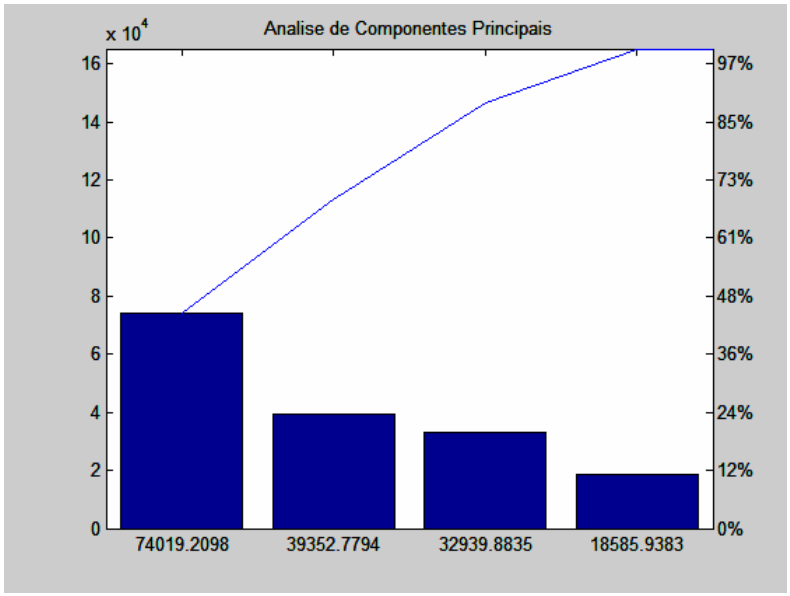



Média V7
300.00
250.00
200.00
150.00
100.00
50.00
-
1 2 3 4 5 6 7 8 9 10 11 12
Média
Ativos
Inativos
Médias V8
450.00
400.00
350.00
300.00
250.00
200.00
150.00
100.00
50.00
-
1 2 3 4 5 6 7 8 9 10 11 12
Média
Ativos
Inativos
Médias V9
6.00
5.00
4.00
3.00
2.00
1.00
-
1 2 3 4 5 6 7 8 9 10 11 12
Médias
Ativos
Inativos
Médias V10
120.00
100.00
80.00
60.00
40.00
20.00
-
1 2 3 4 5 6 7 8 9 10 11 12
Médias
Ativos
Inativos

ANEXO III – Variável Sintetizada: Patamar




Histograma variavel x9
Classe 1
Classe 2
Frequencia (%)
x9
Histograma variavel x10
Classe 1
Classe 2
Frequencia (%)
x10


x 10^4
Analise de Componentes Principais
97%
85%
73%
61%
48%
36%
24%
12%
0%
74019.2098
39352.7794
32939.8835
18585.9383

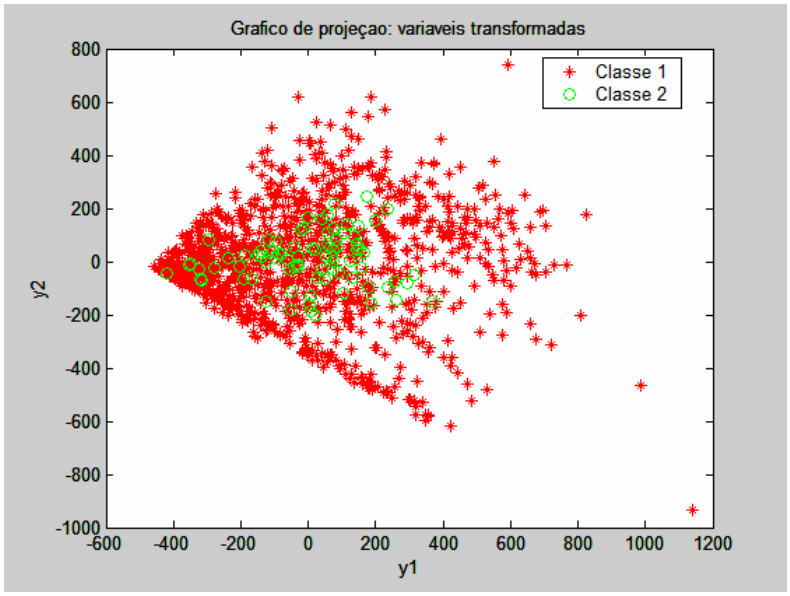
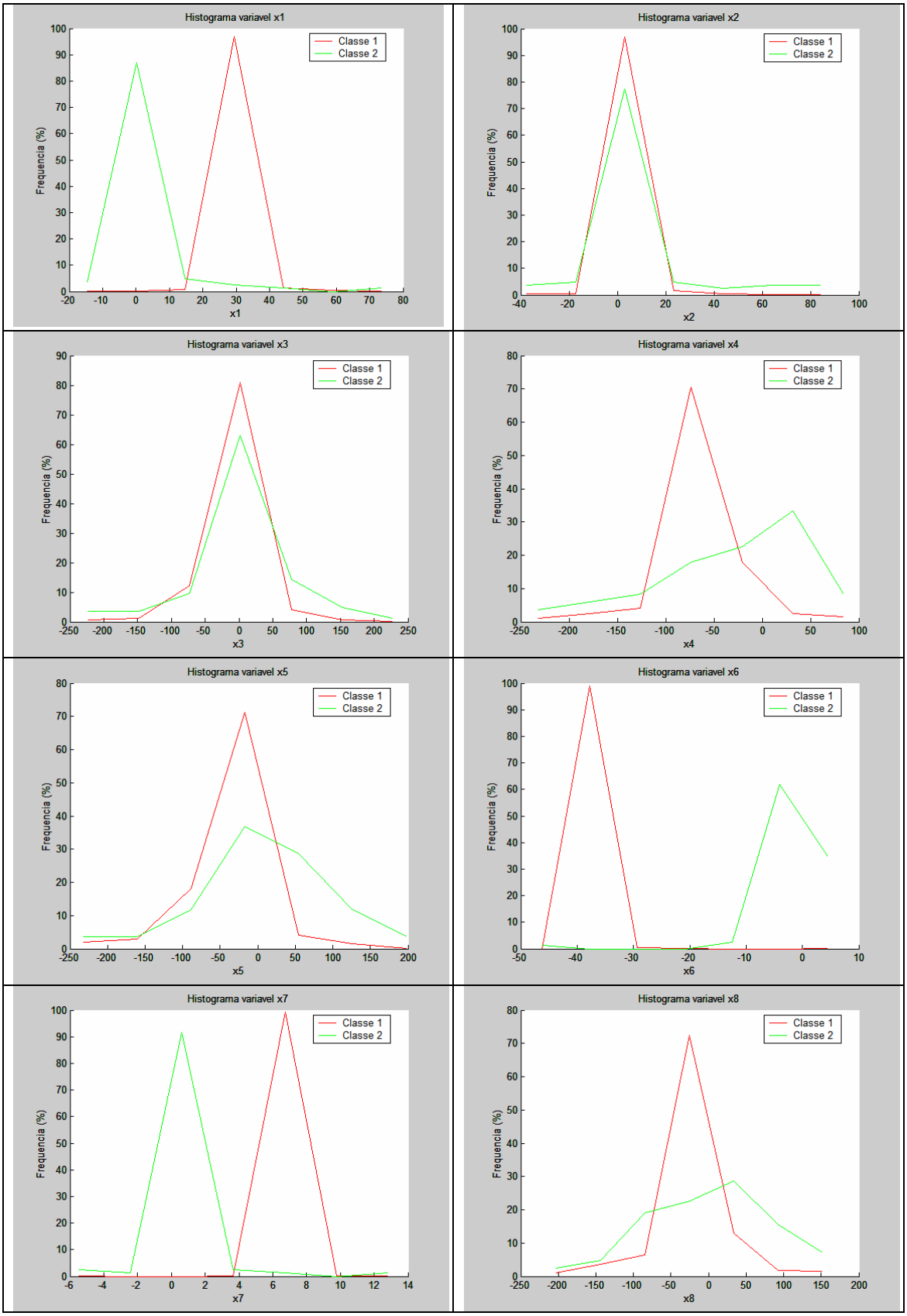
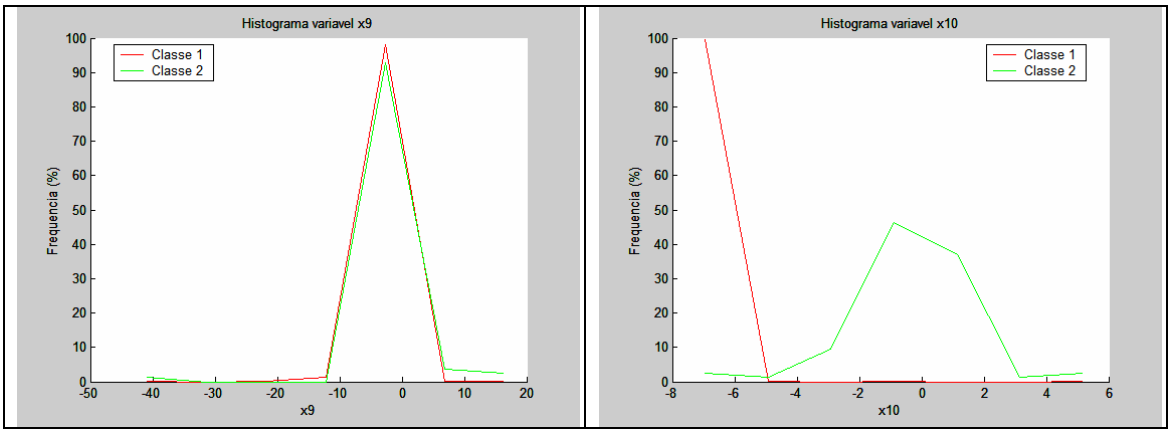
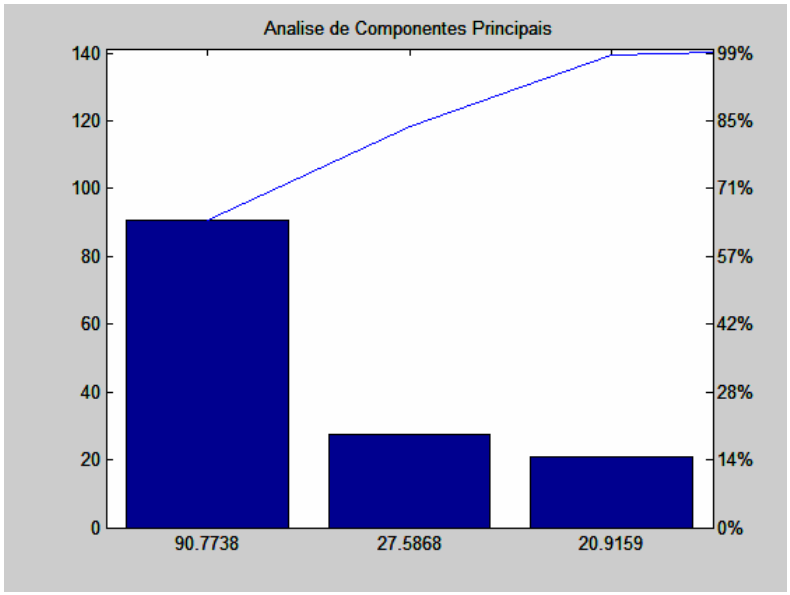


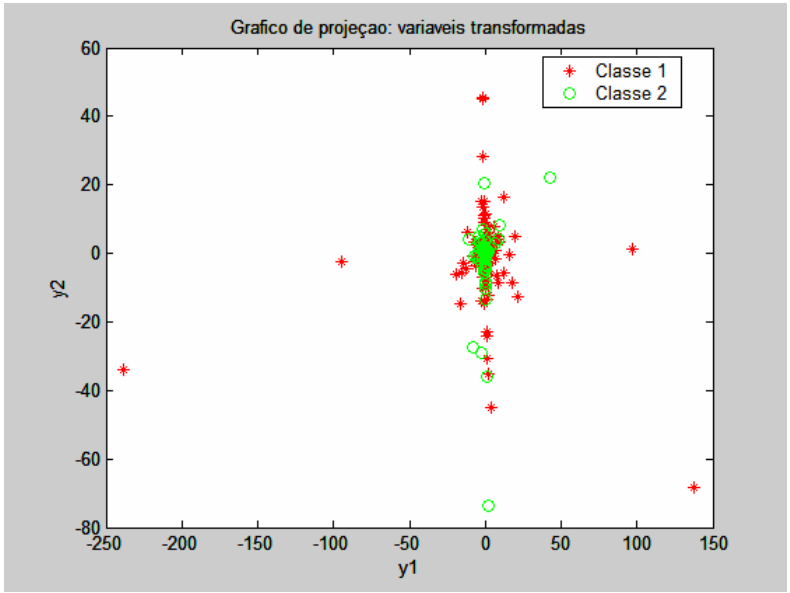
Grafico de projeçao: variaveis transformadas
Classe 1
Classe 2
y2
y1

ANEXO IV – Variável Sintetizada: Tendência

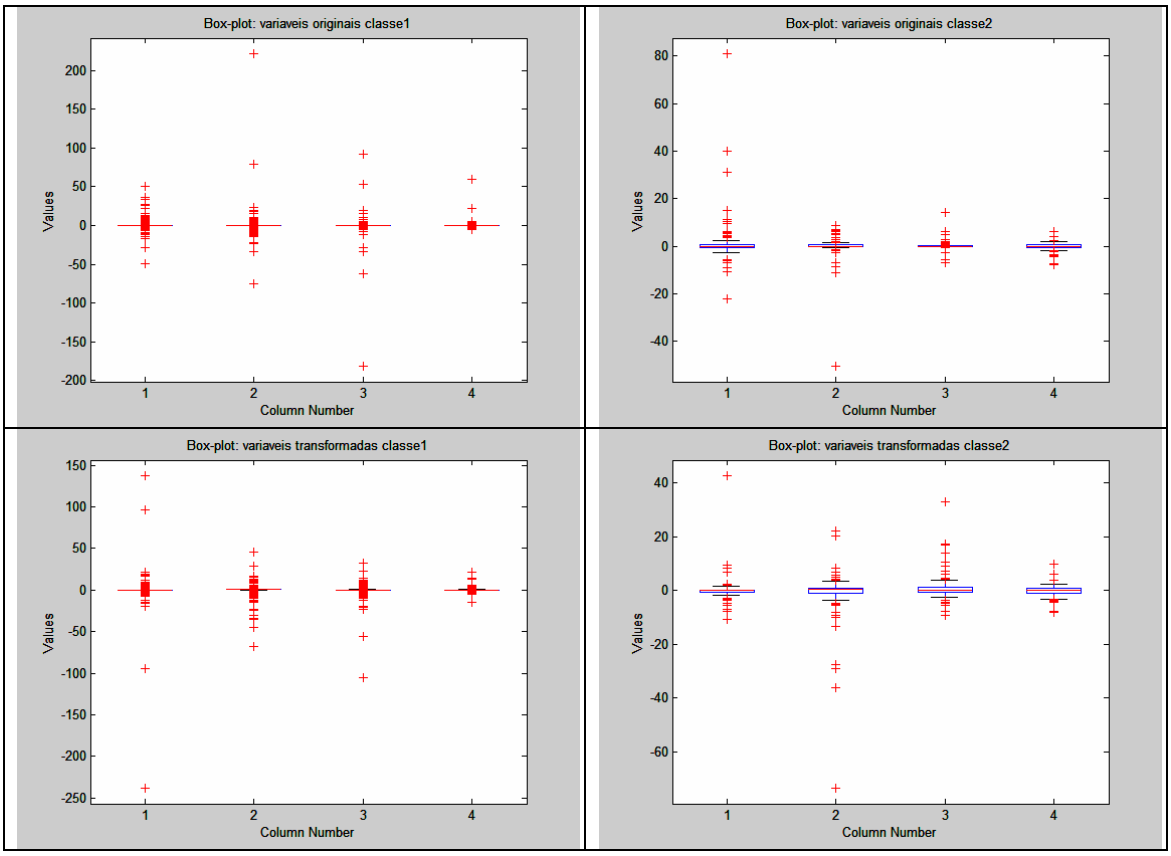




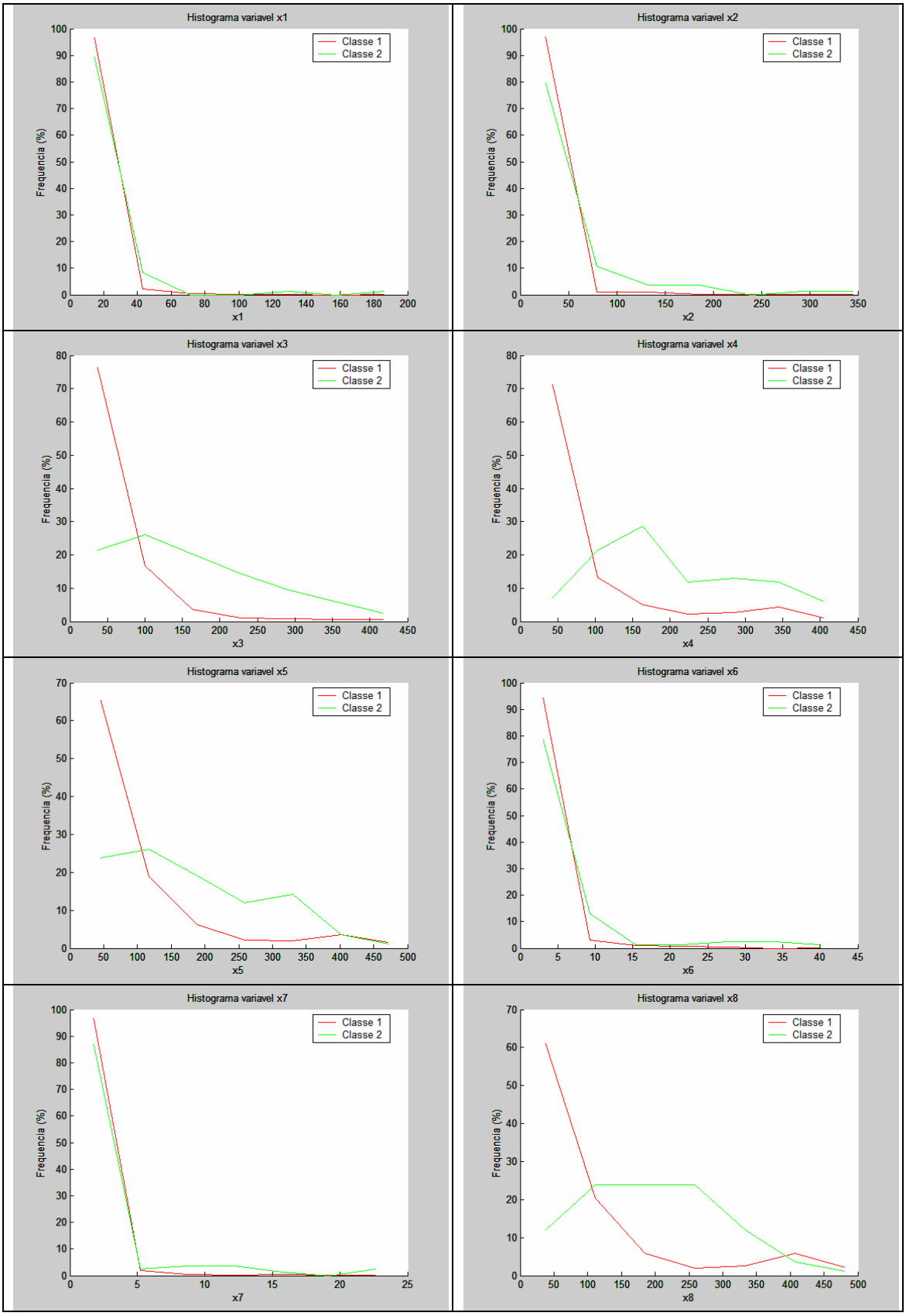


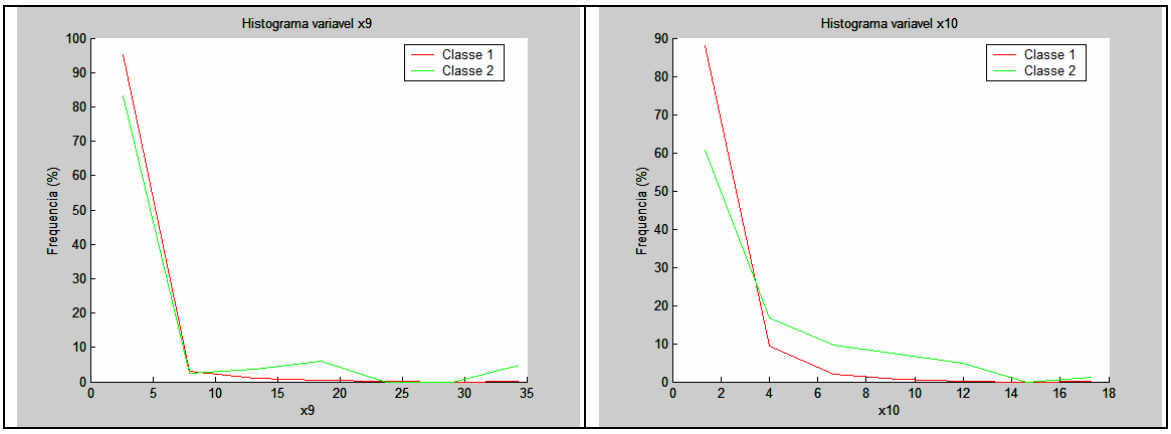


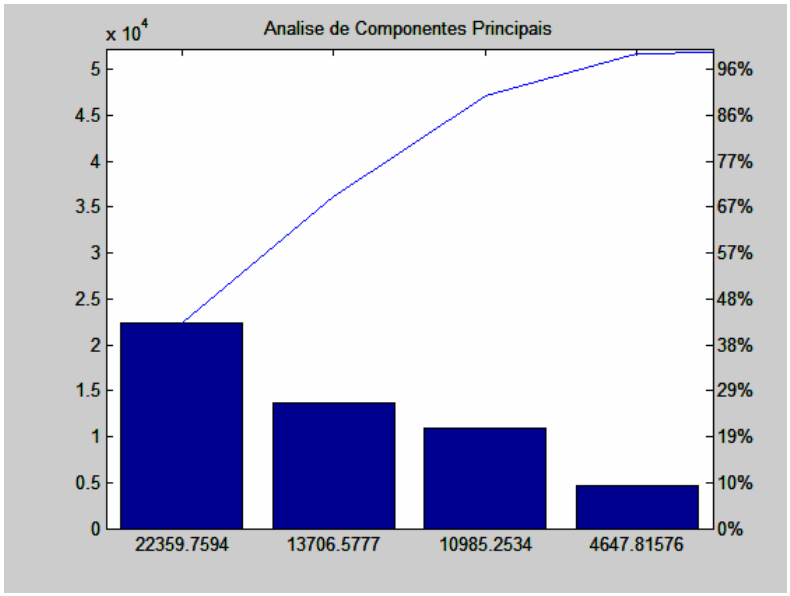
BOXPLOT das classes separadas:


Box-plot: variaveis originais classe1
Values
Column Number
Box-plot: variaveis originais classe2
Values
Column Number
Box-plot: variaveis transformadas classe1
Values
Column Number
Box-plot: variaveis transformadas classe2
Values
Column Number

ANEXO V – Variável Sintetizada: Volatilidade




Histograma variavel x9
Classe 1
Classe 2
Frequencia (%)
x9
Histograma variavel x10
Classe 1
Classe 2
Frequencia (%)
x10


x 10^4
Analise de Componentes Principais
96%
86%
77%
67%
57%
48%
38%
29%
19%
10%
0%
22359.7594
13706.5777
10985.2534
4647.81576

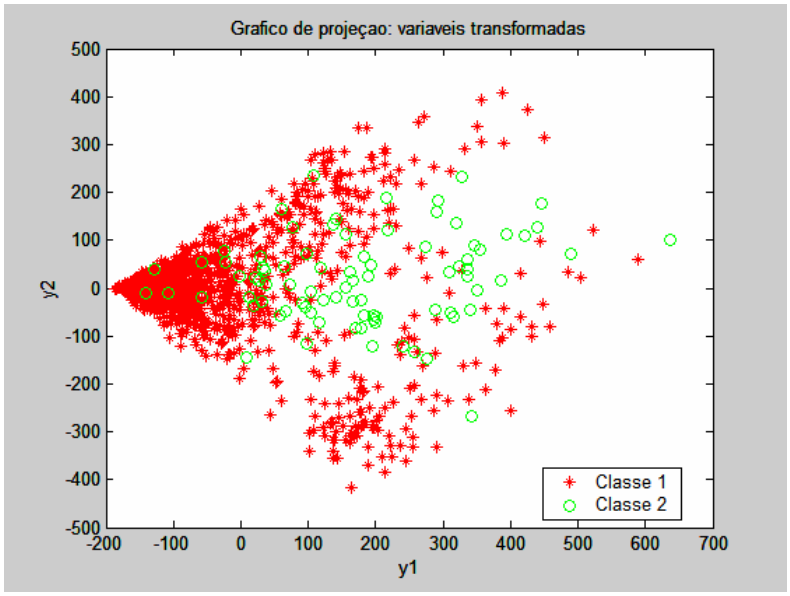
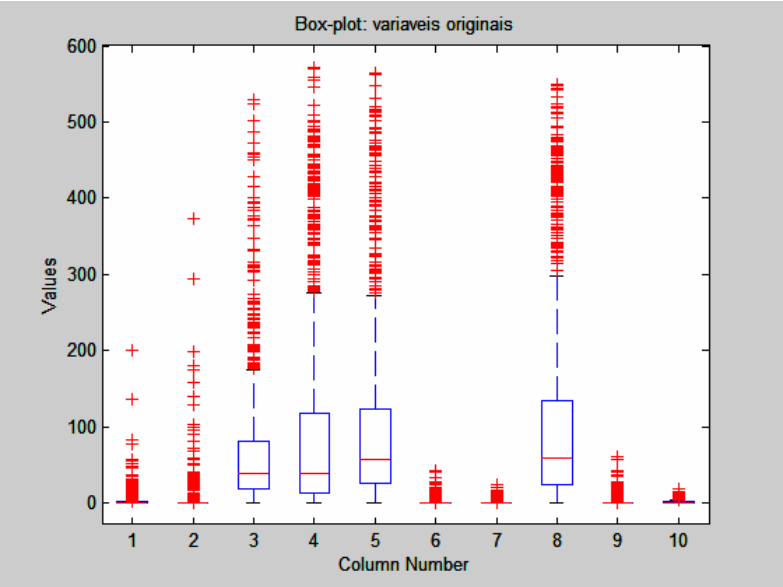
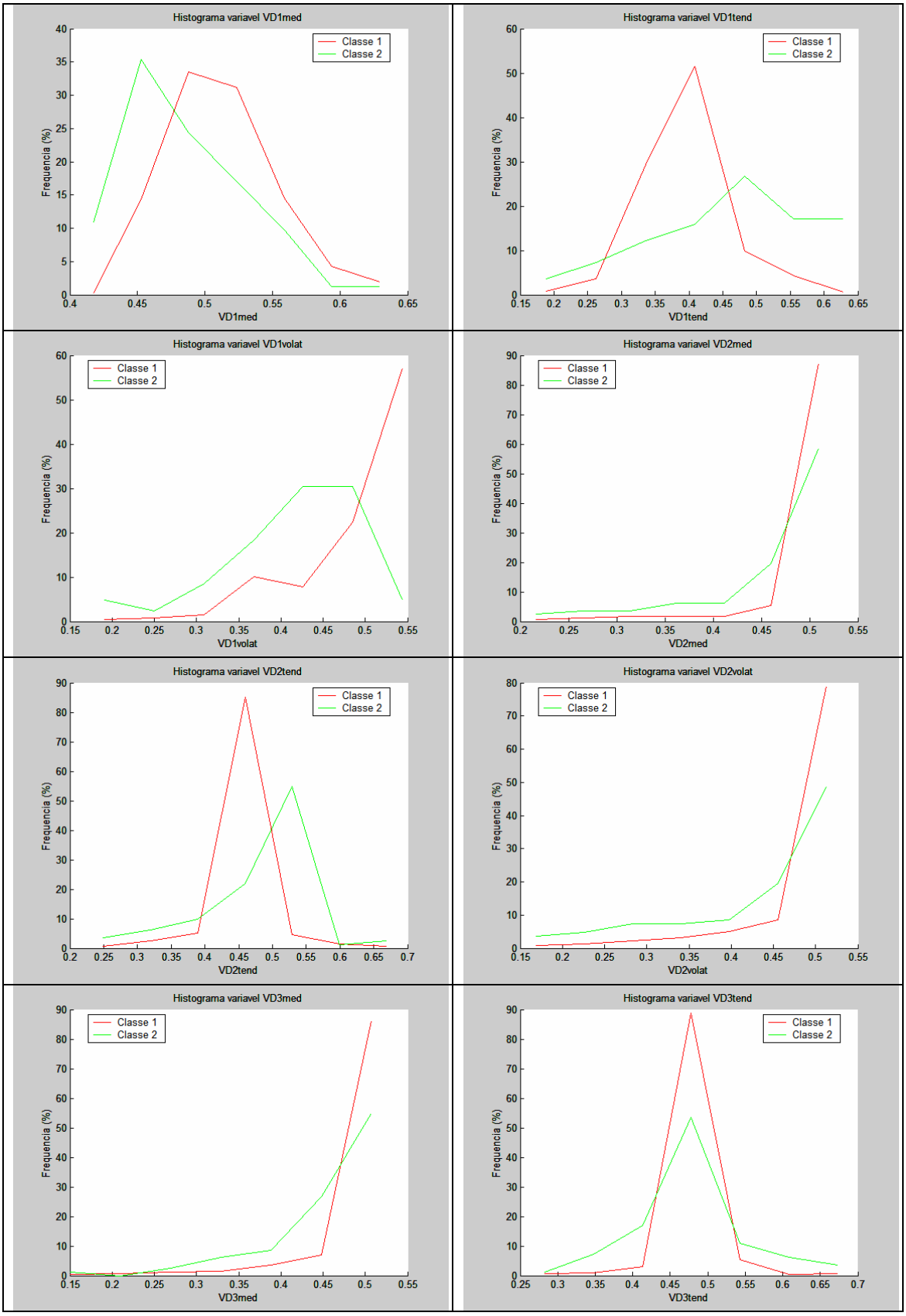
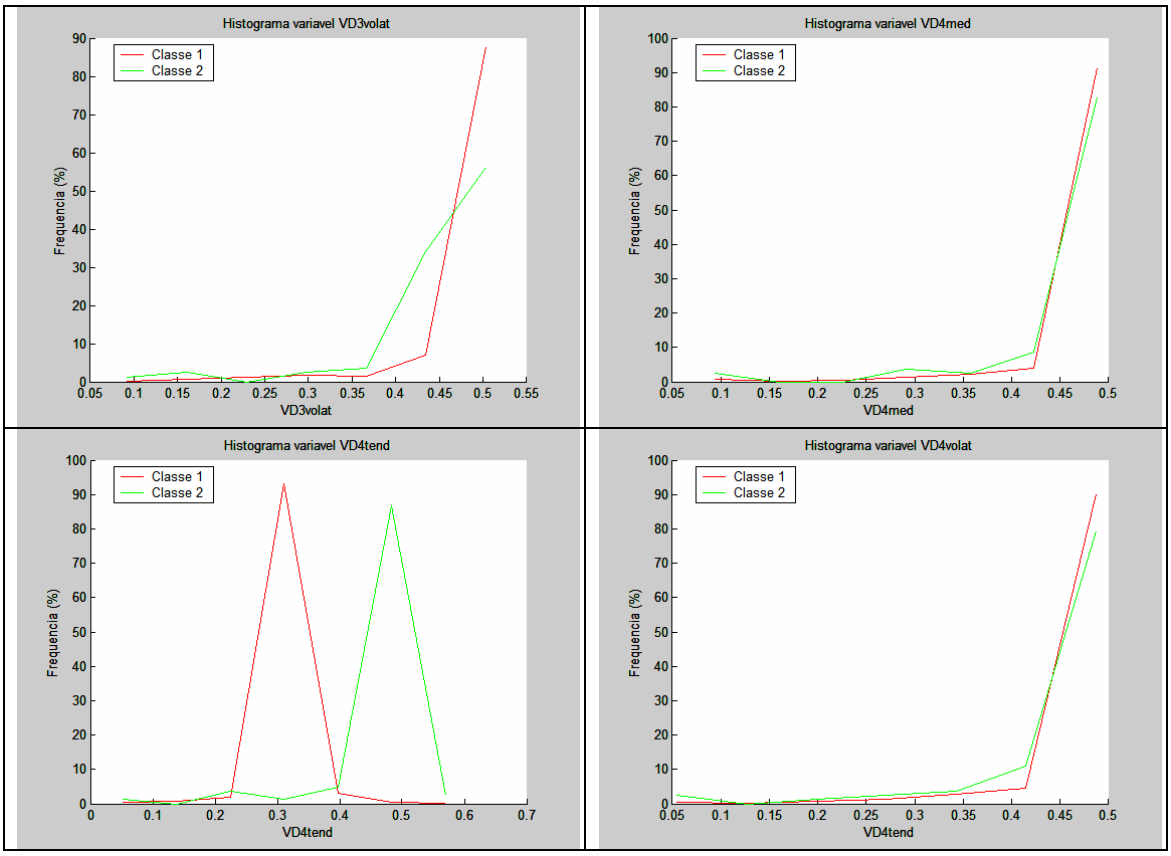


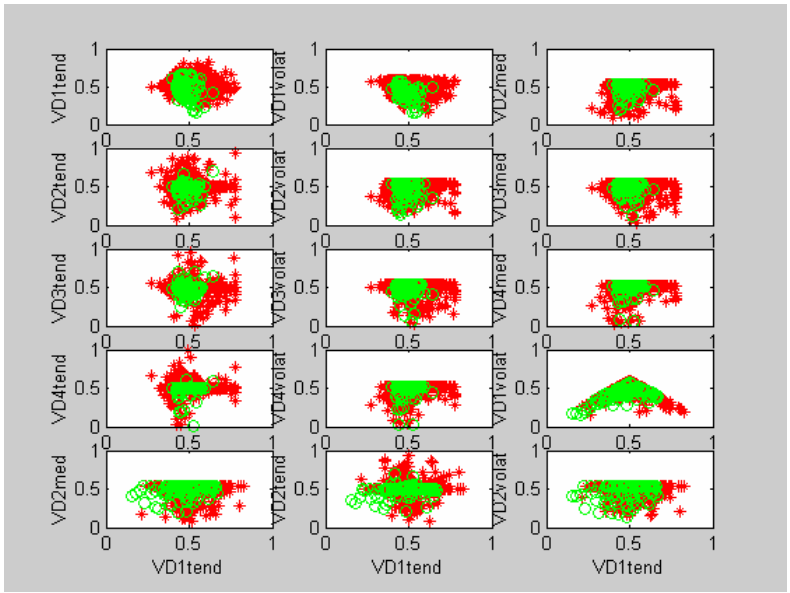
Grafico de projeçao: variaveis transformadas
y2
y1
Classe 1
Classe 2

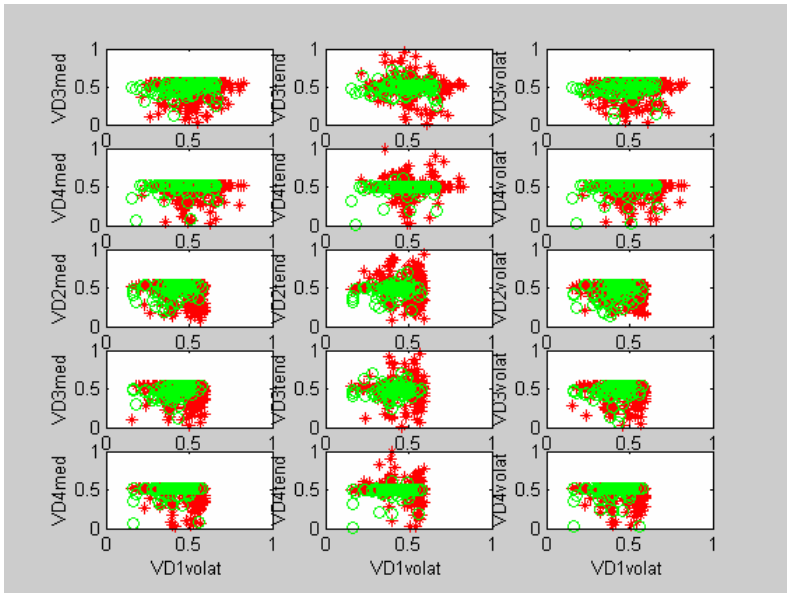

Box-plot: variaveis originais
Values
Column Number
600
500
400
300
200
100
0
1
2
3
4
5
6
7
8
9
10

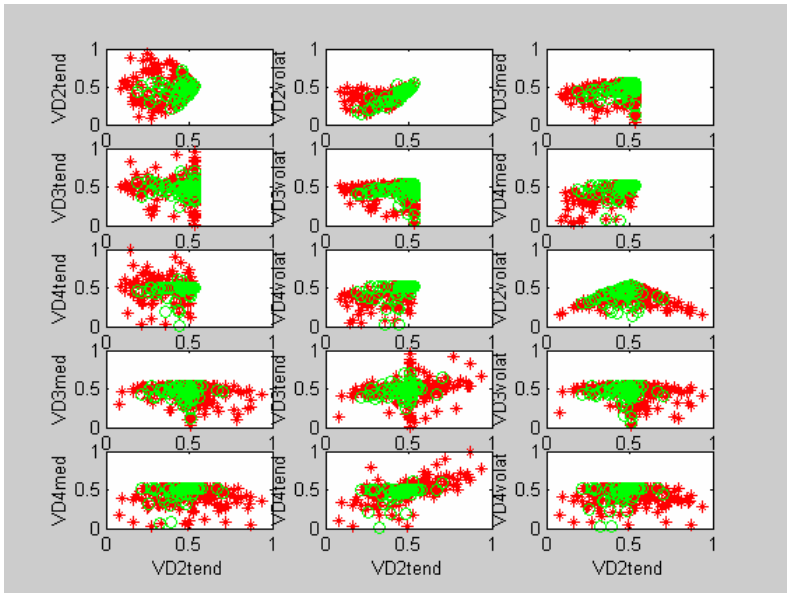
ANEXO VI – Variável Derivadas

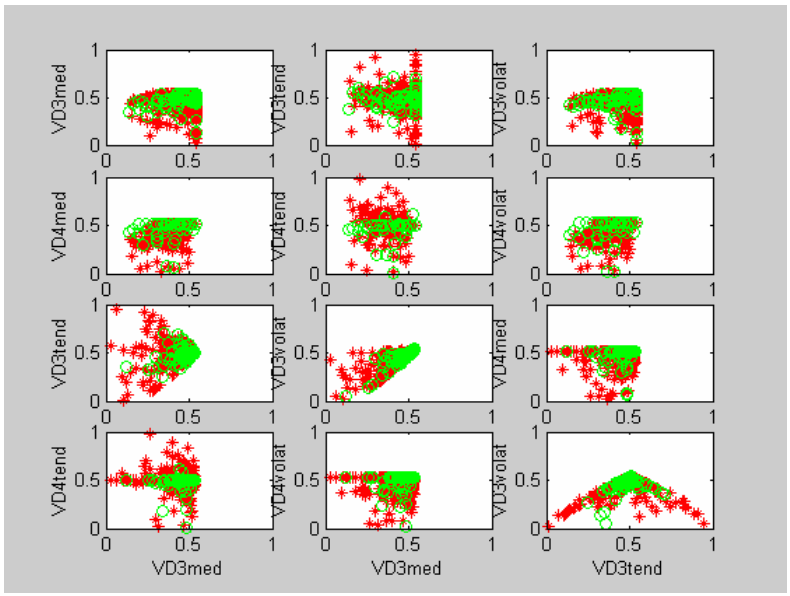


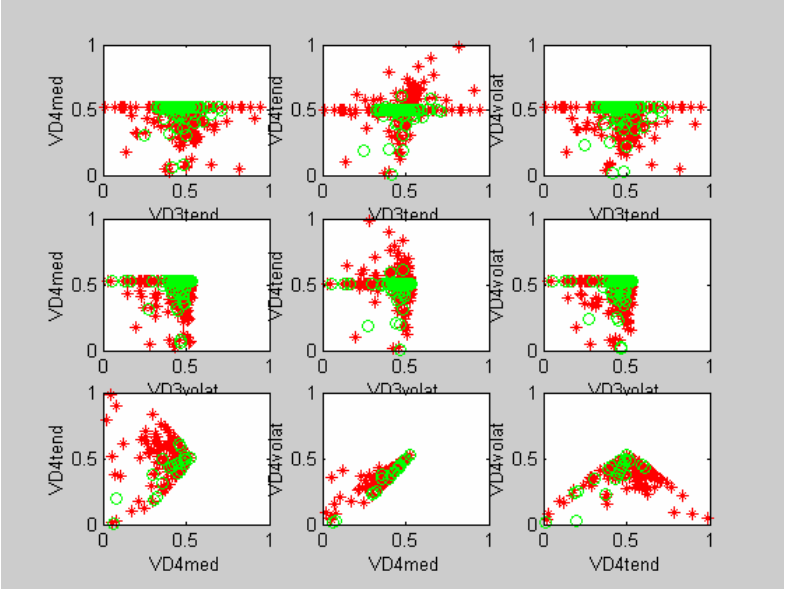

Histograma variavel VD3volat
Classe 1
Classe 2
Frequencia (%)
VD3volat
Histograma variavel VD4med
Classe 1
Classe 2
Frequencia (%)
VD4med
Histograma variavel VD4tend
Classe 1
Classe 2
Frequencia (%)
VD4tend
Histograma variavel VD4volat
Classe 1
Classe 2
Frequencia (%)
VD4tend

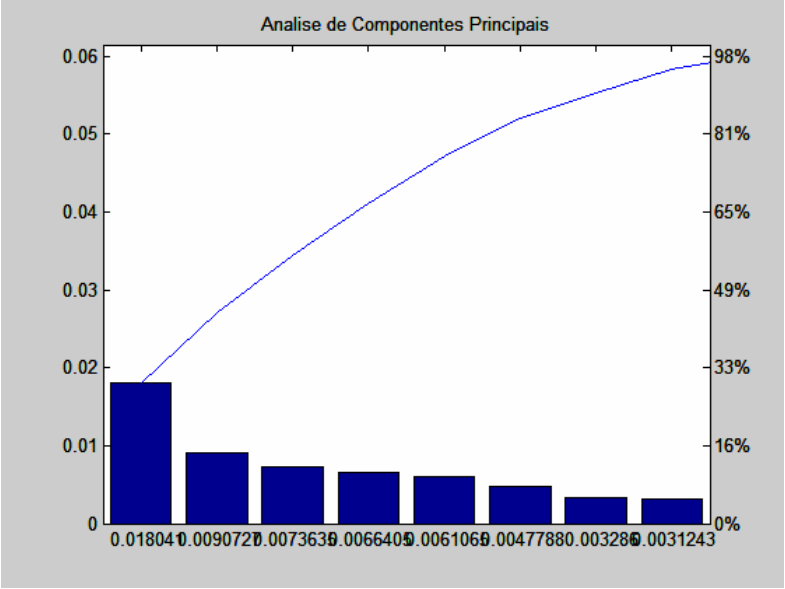

VD1tend
VD1volat
VD2med
VD2tend
VD2volat
VD3med
VD3tend
VD3volat
VD4med
VD4tend
VD4volat
VD1volat
VD2med
VD2tend
VD2volat
VD1tend
VD1tend
VD1tend


VD3med
VD4med
VD2med
VD3med
VD4med
VD3tend
VD4tend
VD2tend
VD3tend
VD4tend
VD3volat
VD4volat
VD2volat
VD3volat
VD4volat
VD1volat
VD1volat
VD1volat


VD2tend
VD3tend
VD4tend
VD3med
VD4med
VD2volat
VD3volat
VD4volat
VD3tend
VD4tend
VD3med
VD4med
VD2volat
VD3volat
VD4volat
VD2tend
VD2tend
VD2tend


VD3med
VD4med
VD3tend
VD4tend
VD3tend
VD4tend
VD3volat
VD4volat
VD3volat
VD4volat
VD4med
VD3volat
VD3med
VD3med
VD3tend


VD4med
VD4tend
VD4volat
VD3tend
VD3volat
VD4med
VD4tend


Analise de Componentes Principais
0.06
0.05
0.04
0.03
0.02
0.01
0
98%
81%
65%
49%
33%
16%
0%
0.018041
0.0090727
0.0073639
0.0066405
0.0061065
0.0047788
0.003286
0.0031243

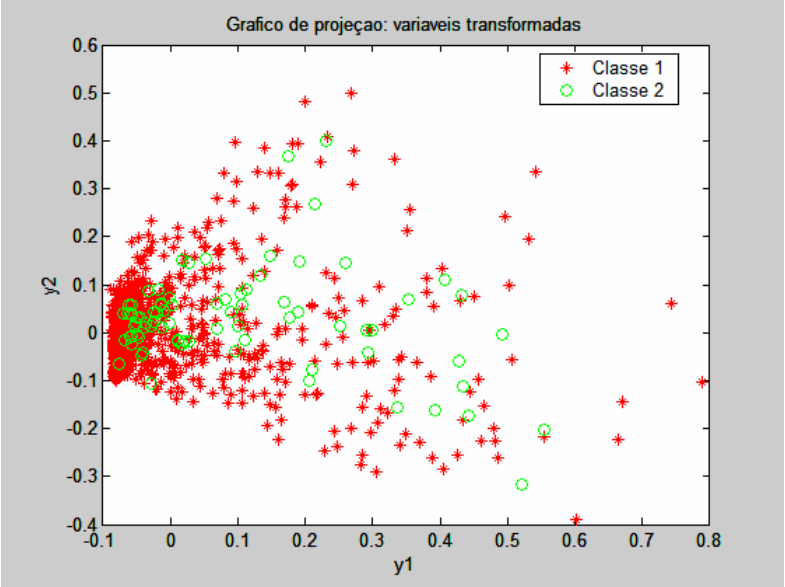
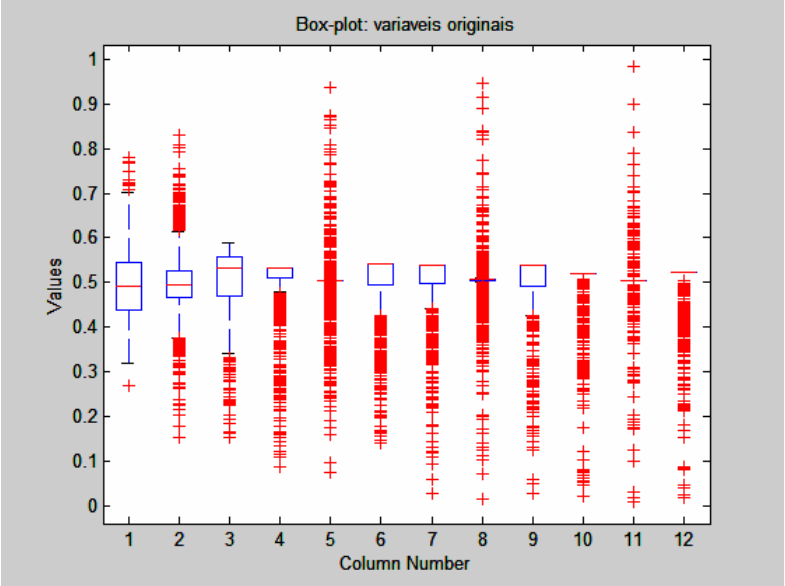
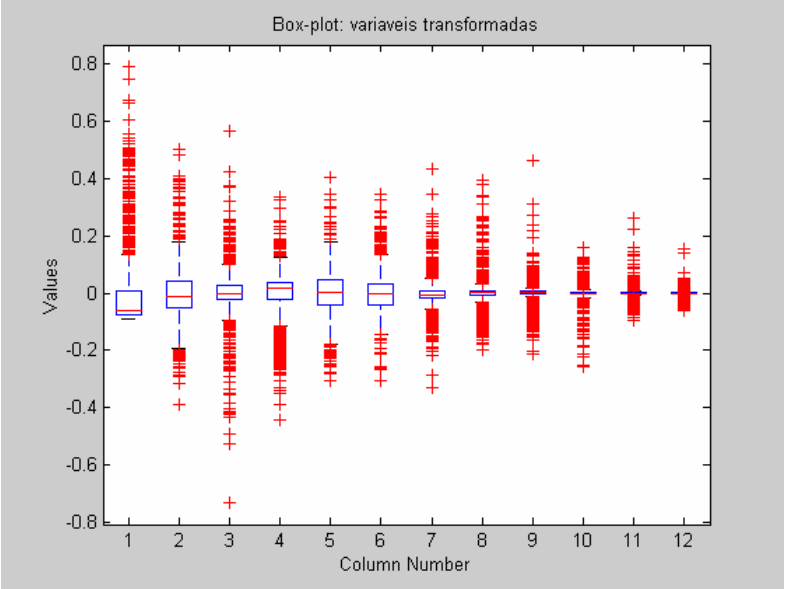
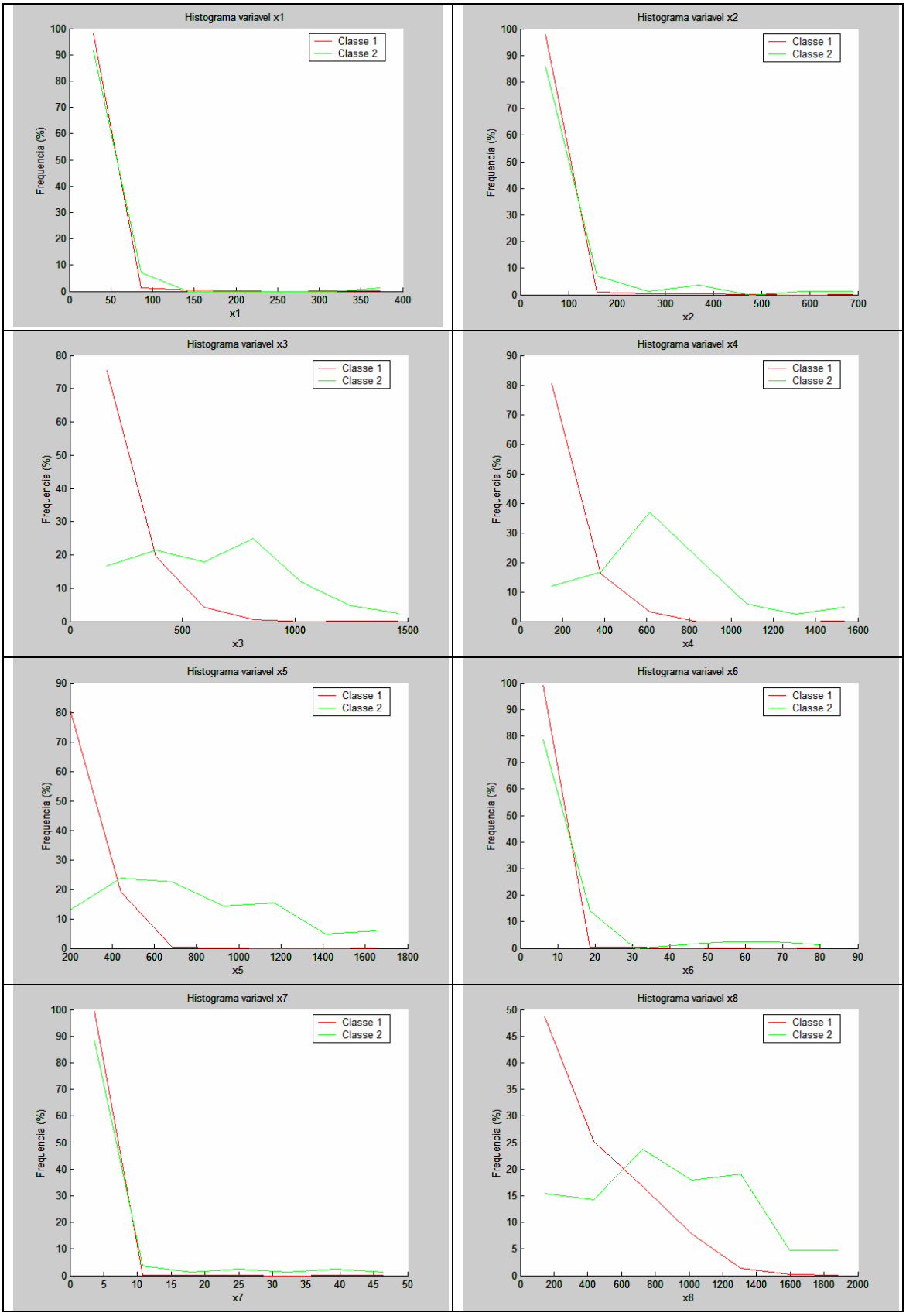


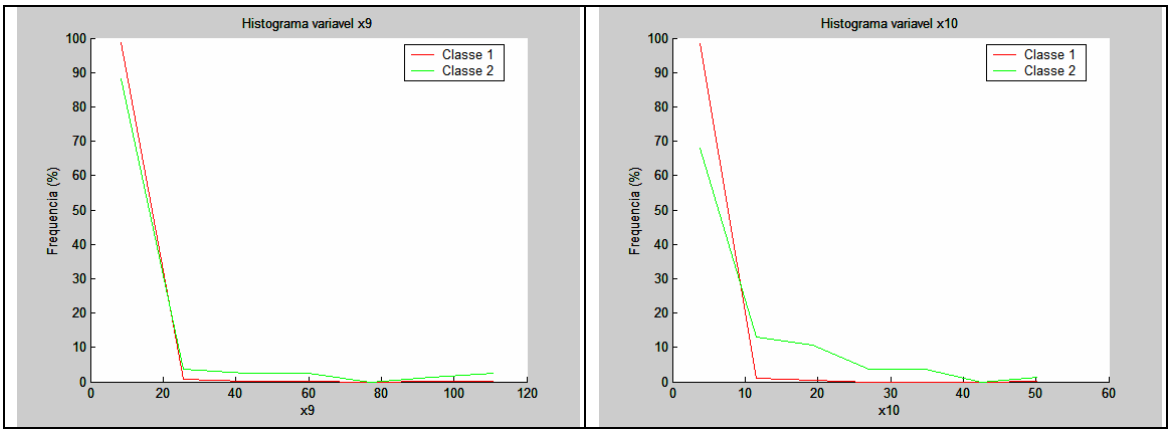
Grafico de projeçao: variaveis transformadas
Classe 1
Classe 2
y1
y2

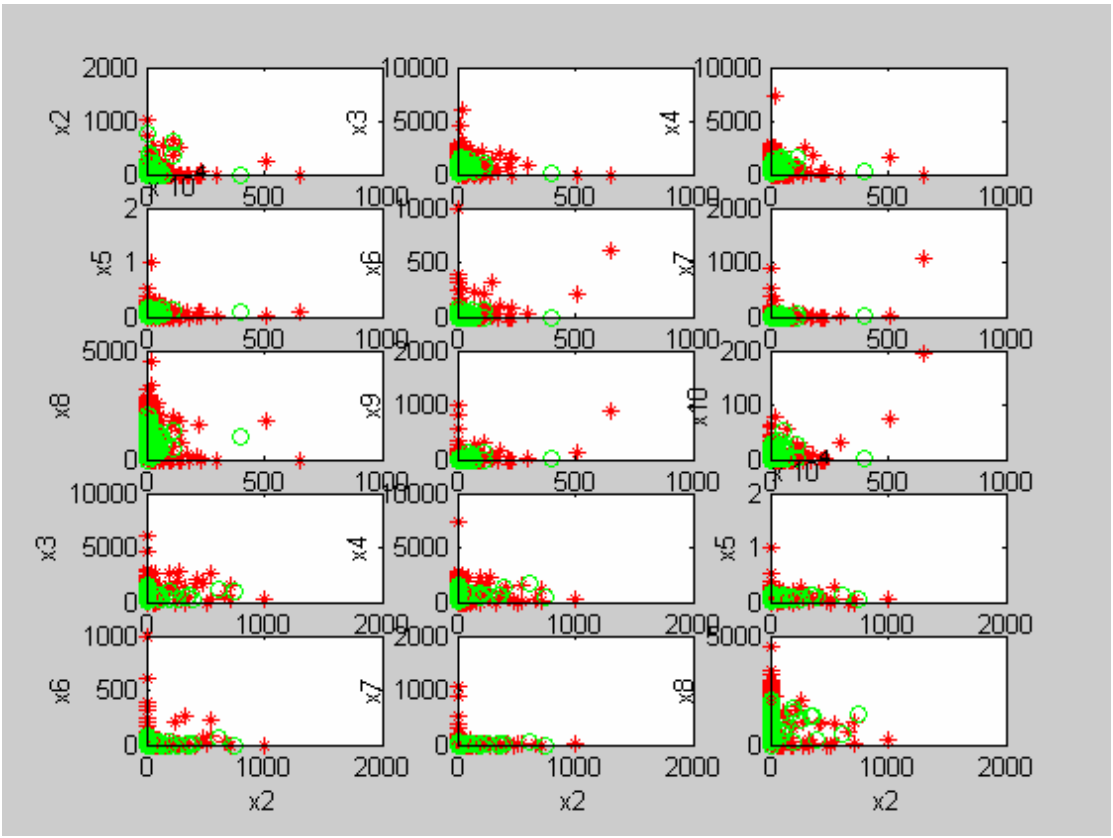

Box-plot: variaveis originais
Values
Column Number

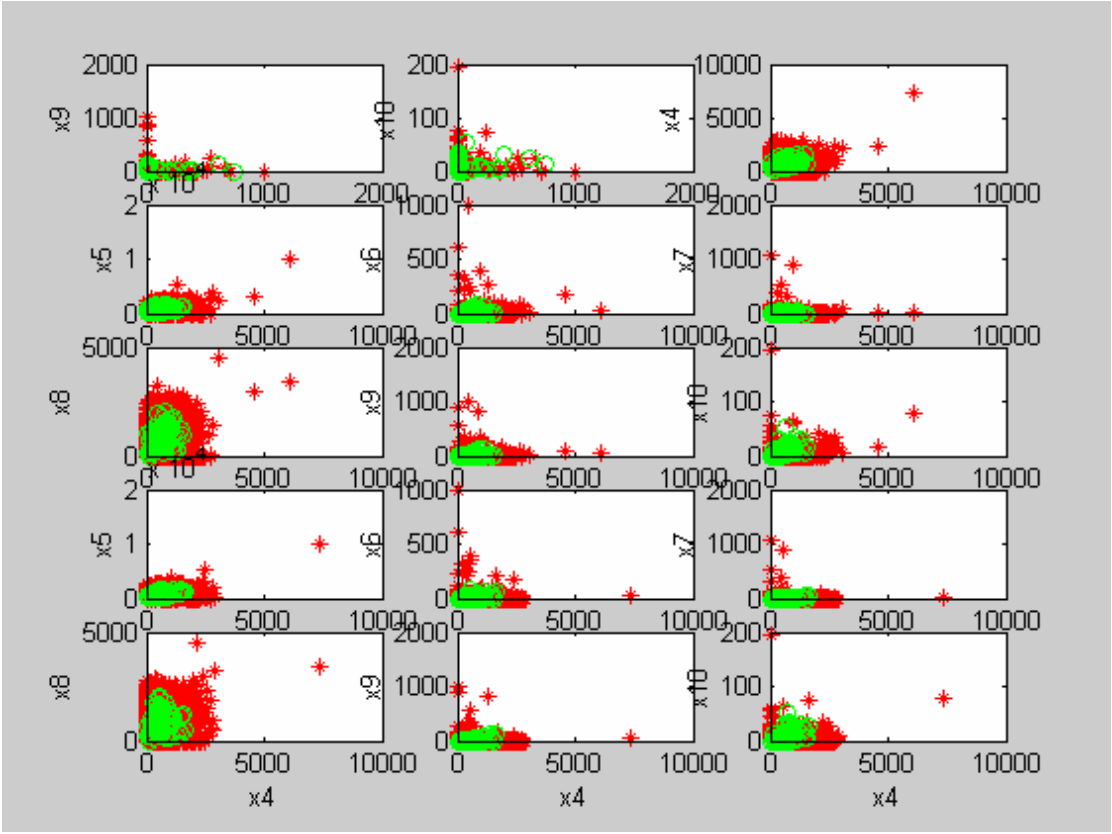

Box-plot: variaveis transformadas
Values
Column Number
0.8
0.6
0.4
0.2
0
-0.2
-0.4
-0.6
-0.8
1
2
3
4
5
6
7
8
9
10
11
12

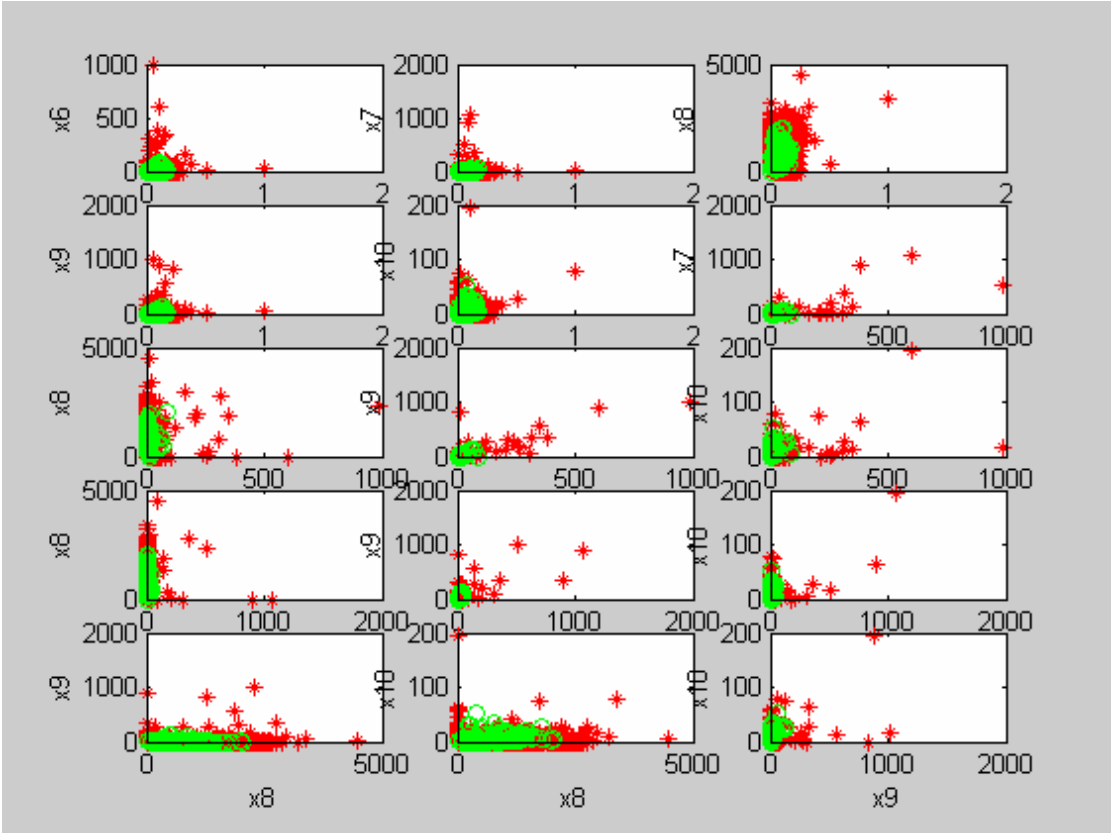
ANEXO VII – Análise com tratamento temporal

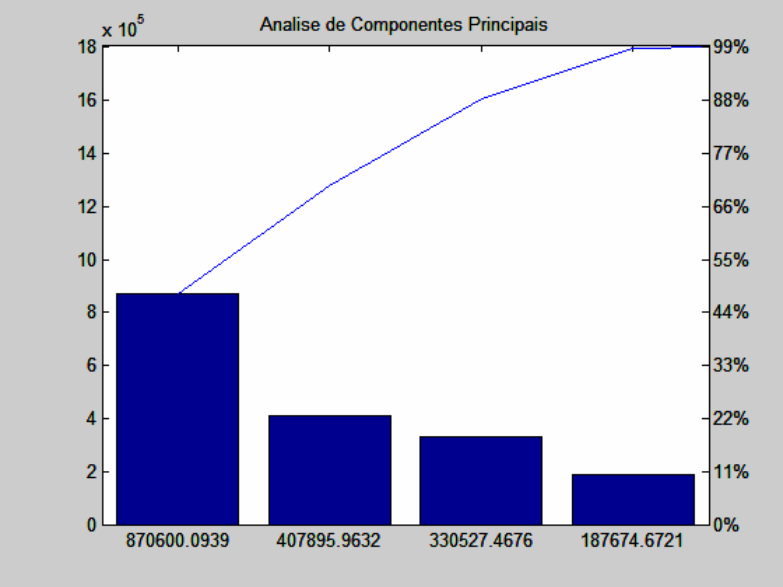



Histograma variavel x9
Classe 1
Classe 2
Frequencia (%)
x9
Histograma variavel x10
Classe 1
Classe 2
Frequencia (%)
x10


x2
x3
x4
x5
x6
x7
x8
x9
x10
x2
x2
x2


x4
x4
x4


x8
x8
x9


Analise de Componentes Principais
x 10^5
18
16
14
12
10
8
6
4
2
0
99%
88%
77%
66%
55%
44%
33%
22%
11%
0%
870600.0939
407895.9632
330527.4676
187674.6721

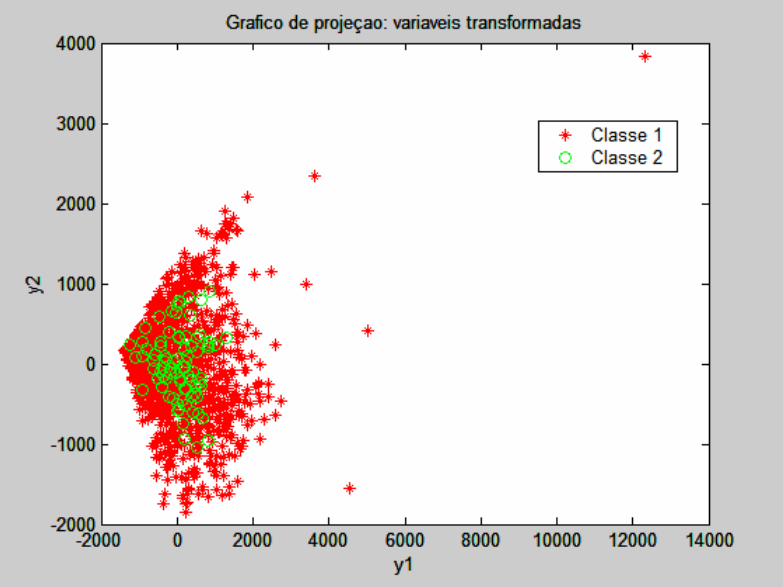
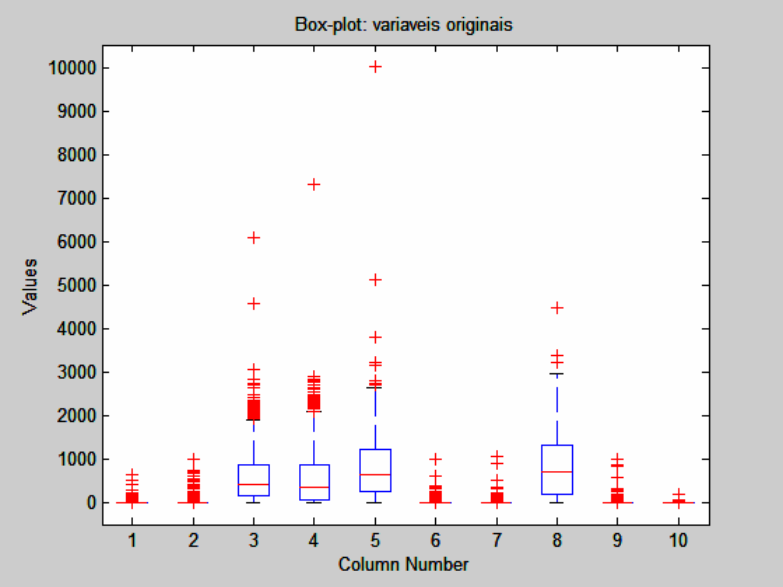
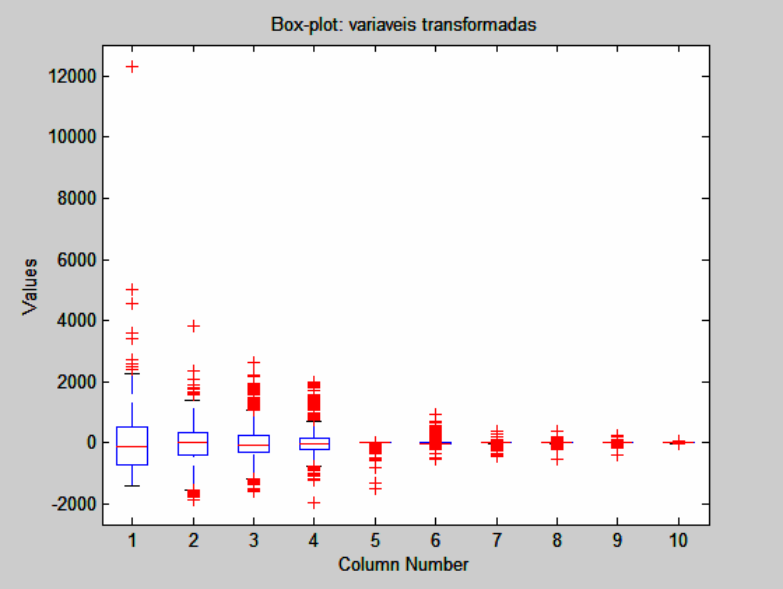
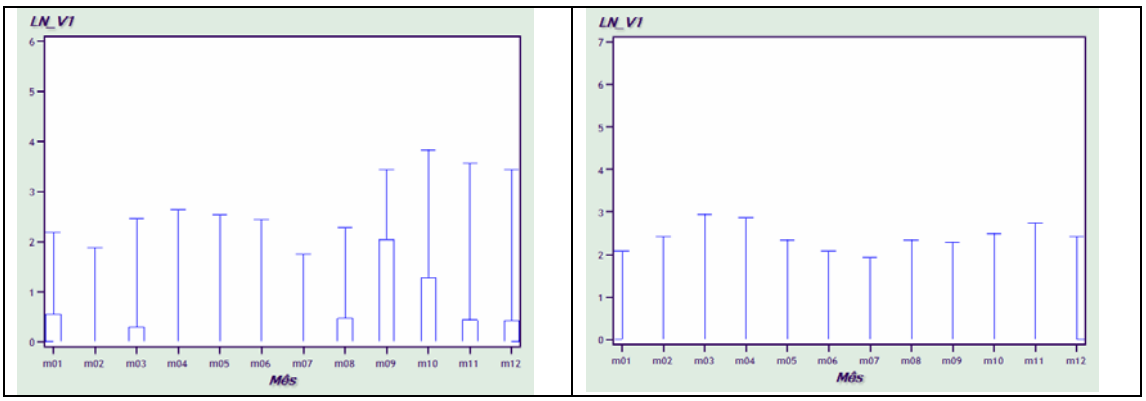


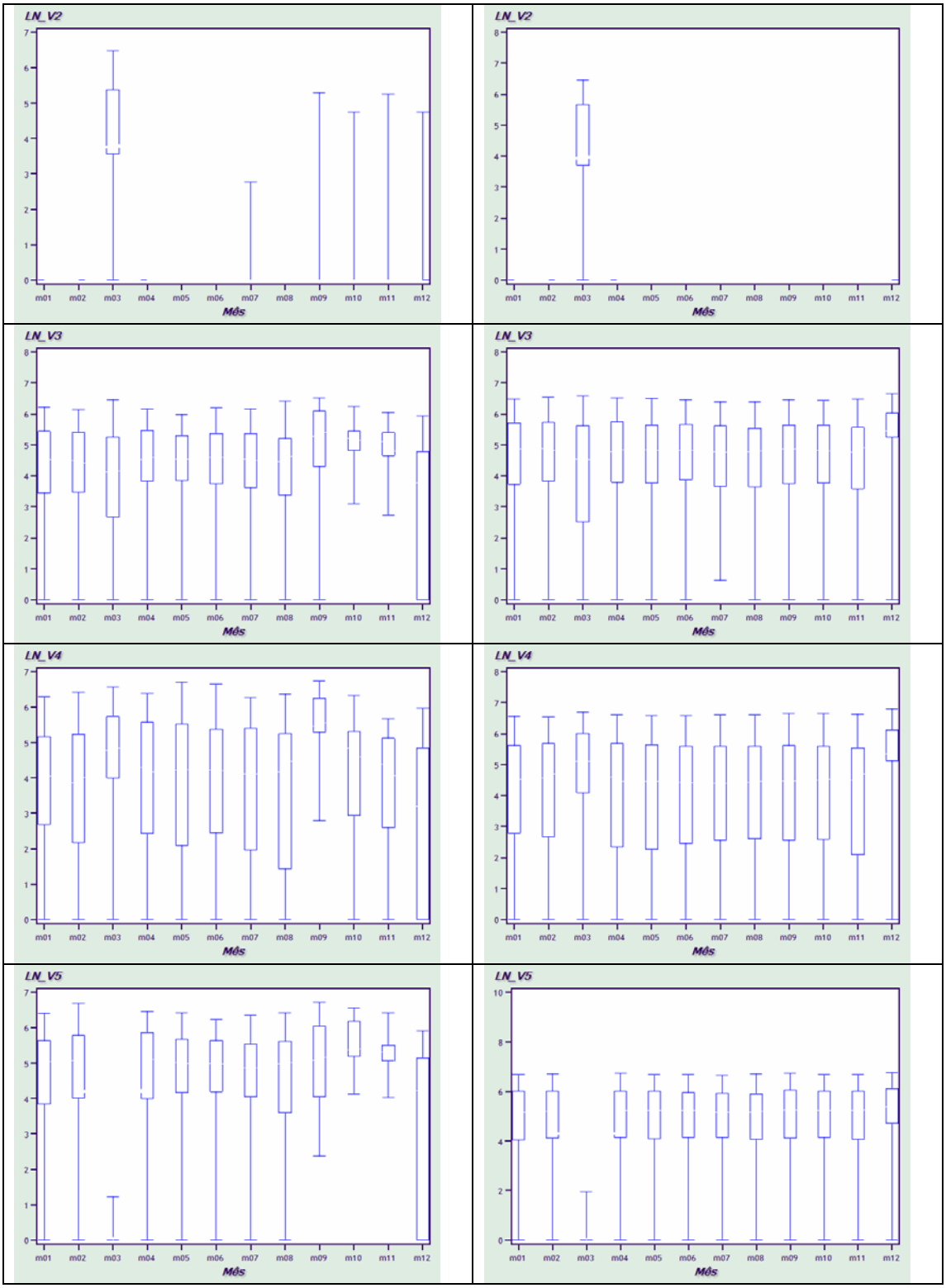
Grafico de projeçao: variaveis transformadas
Classe 1
Classe 2
y2
y1
4000
3000
2000
1000
0
-1000
-2000
-2000
0
2000
4000
6000
8000
10000
12000
14000

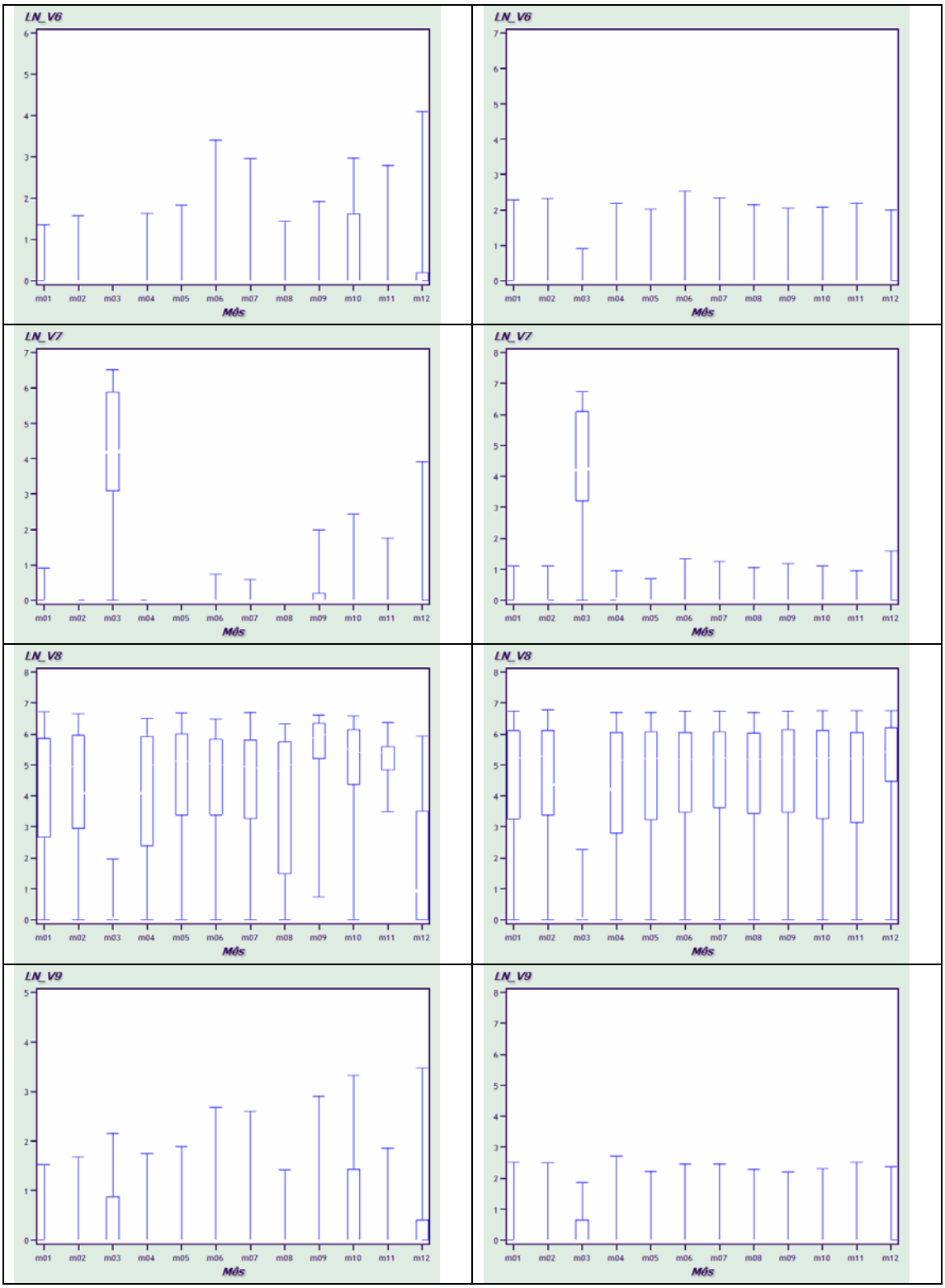


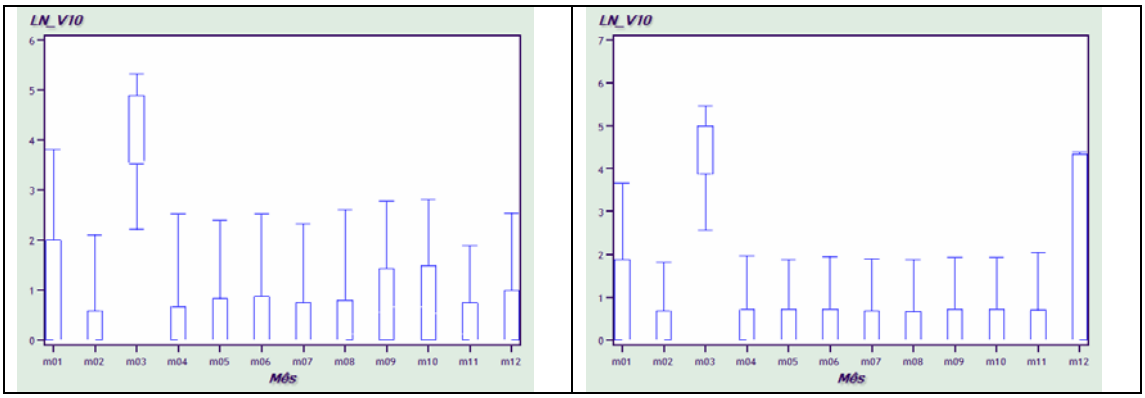


INATIVO ATIVO




LN_V2
LN_V2
LN_V3
LN_V3
LN_V4
LN_V4
LN_V5
LN_V5
m01 m02 m03 m04 m05 m06 m07 m08 m09 m10 m11 m12
Mês


LN_V6
LN_V6
LN_V7
LN_V7
LN_V8
LN_V8
LN_V9
LN_V9
m01 m02 m03 m04 m05 m06 m07 m08 m09 m10 m11 m12
Mês


LN_V10
6
5
4
3
2
1
0
m01
m02
m03
m04
m05
m06
m07
m08
m09
m10
m11
m12
Mês
LN_V10
7
6
5
4
3
2
1
0
m01
m02
m03
m04
m05
m06
m07
m08
m09
m10
m11
m12
Mês

Livros Grátis

(http://www.livrosgratis.com.br)

Milhares de Livros para Download:

Baixar livros de Literatura
Baixar livros de Literatura de Cordel
Baixar livros de Literatura Infantil
Baixar livros de Matemática
Baixar livros de Medicina
Baixar livros de Medicina Veterinária
Baixar livros de Meio Ambiente
Baixar livros de Meteorologia
Baixar Monografias e TCC
Baixar livros Multidisciplinar
Baixar livros de Música
Baixar livros de Psicologia
Baixar livros de Química
Baixar livros de Saúde Coletiva
Baixar livros de Serviço Social
Baixar livros de Sociologia
Baixar livros de Teologia
Baixar livros de Trabalho
Baixar livros de Turismo