Universidade Federal de Campina Grande

Centro de Engenharia Elétrica e Informática

Curso de Graduação em Engenharia Elétrica

PAULO PIRES FERNANDES NETO

RELATÓRIO DE ESTÁGIO INTEGRADO

Campina Grande, Paraíba
2014

PAULO PIRES FERNANDES NETO

RELATÓRIO DE ESTÁGIO INTEGRADO

Relatório de Estágio Integrado submetido à Unidade Acadêmica de Engenharia Elétrica da Universidade Federal de Campina Grande como parte dos requisitos necessários para a obtenção do grau de Bacharel em Ciências no Domínio da Engenharia Elétrica.

Orientador:
Professor Dr. Hiran de Melo

Campina Grande, Paraíba
2014

PAULO PIRES FERNANDES NETO

RELATÓRIO DE ESTÁGIO INTEGRADO

Relatório de Estágio Integrado submetido à Unidade Acadêmica de Engenharia Elétrica da Universidade Federal de Campina Grande como parte dos requisitos necessários para a obtenção do grau de Bacharel em Ciências no Domínio da Engenharia Elétrica.

Aprovado em ____ de _____ de 2014

BANCA EXAMINADORA

_____

Prof. Dr. Hiran de Melo

Universidade Federal de Campina Grande

Orientador

_____

Professor Convidado

Universidade Federal de Campina Grande

Avaliador

À Deus, toda honra e toda glória.

AGRADECIMENTOS

A Deus, por ter me dado forças e iluminado meus caminhos para que eu pudesse concluir mais uma etapa da minha vida.

A meus pais e minha irmã, por serem tão dedicados, amorosos e por ter me dado condições para me tornar o profissional e o homem que sou.

A toda a família Fernandes e Peixoto por todo apoio e suporte.

Ao meu grande amigo, parceiro e praticamente um irmão, Raul, por ter me acompanhando muitos anos da minha vida, inclusive os de vida acadêmica.

Ao professor e orientador Dr. Hiran de Melo, pela contribuição no desenvolvimento deste trabalho, por ser um grande amigo além de um mestre em passar conhecimento, onde pude conhece-lo além da vida acadêmica e conviver no dia-dia.

Ao Professor e coorientador Dr. José Carlos Reston Filho pela oportunidade que me concedeu, pela satisfação em me passar conhecimento e me permitir crescer profissionalmente.

Àqueles, que não por menor importância, não foram citados, mas também tiveram grande contribuição na realização desse grande sonho de ser engenheiro eletricista.

LISTA DE ILUSTRAÇÕES

SUMÁRIO

1. INTRODUÇÃO

O Objetivo deste relatório é fazer uma descrição das atividades desenvolvidas durante o Estágio Integrado realizado na empresa IDAAM Educação Superior LTDA, no período de novembro de 2013 à abril de 2014, correspondendo uma carga horária de 660 horas.

A disciplina de Estágio Integrado é oferecida aos estudantes do Curso de Graduação em Engenharia Elétrica pela Universidade Federal de Campina Grande com o intuito de fazer o aluno concluinte realizar atividades em empresas, e dessa forma prepará-lo melhor para o mercado de trabalho.

No decorrer do estágio foram oferecidas oportunidades de aprender e aplicar conhecimentos sobre mineração de dados, para classificação, segmentação e predição. Foram elaborados treinamentos em softwares de cunho importante para o minerador de dados tais como o pacote estatístico e de business intelligence STATSOFT Statistica e um treinamento aprofundado do software IBM SPSS Modeler. Foram tidas oportunidades de aprender mais sobre estatística descritiva e inferencial. Também houveram atividades relacionadas a extração de conhecimento de base de dados a partir de clusterização e a partir de algoritmos supervisionados.

Foi realizado um treinamento específico para o forecast de séries temporais, criação de modelos híbridos com 2 ou mais algoritmos, enfatizando o uso de Redes Neurais Artificiais e modelos ARIMA. Foi feito um treinamento para avaliação da qualidade, robustez e capacidade de generalização de modelos, melhoria de desempenho de modelos a partir do pré-processamento de dados e o uso de Principal Components Analysis no pré-processamento de dados.

O aluno também teve oportunidade em cursar disciplinas que a Instituição oferece, para que o mesmo possa desenvolver melhores ferramentas e obter uma maior experiência para iniciação nas indústrias do parque industrial de Manaus.

1.1 Objetivos do Estágio

Os objetivos do estágio são:

- Apresentar ao aluno o mercado de trabalho oferecendo a oportunidade de se ter uma transição menos impactante da vida estudantil para a vida profissional.
- Oferecer uma oportunidade de conhecimento teórico e prático sobre uma área bastante importante no meio industrial e para qualquer empresa nos dias de hoje, que é a mineração de dados.
- Possibilitar ao profissional recém-formado a oportunidade de ser inserido no vasto parque industrial que Manaus oferece.

2. A EMPRESA

A empresa IDAAM Educação Superior LTDA foi fundada em 2006, pelo Engenheiro Eletricista José Carlos Reston Filho. Situa-se na avenida Djalma Batista, no segundo andar do Shopping Plaza, em Manaus, Amazonas.

A empresa é consolidada há mais de 8 anos no mercado local e já tendo formado milhares de alunos que hoje estão empregados nas maiores empresas de Manaus, a Pós-Graduação IDAAM é uma das maiores instituições de pós-graduação da cidade.

A empresa possui vários fatores de destaque no mercado, tais como:

- Possibilidade de Dupla Certificação;
- Qualidade, preço, boa localização e estacionamento seguro;
- Professores em sua maioria Mestres e Doutores com ampla experiência profissional;
- Metodologias inovadoras, com uso de Dinâmicas e Jogos;
- Conteúdos atualizados com as exigências do mercado de trabalho;
- Acesso ao Mestrado Internacional em parceria com a Universidade do Minho (Portugal)
- Programa de competências comportamentais afinado com as novas tendências mundiais;
- Material Didático próprio;

2.1 Missão da Empresa

A prioridade da instituição é oferecer a melhor formação em pós-graduação em Manaus, aliando o conhecimento teórico com a correspondente aplicação prática, visando o aumento da empregabilidade de seus alunos.

Em todos os cursos oferecidos pela Pós-Graduação IDAAM utiliza-se material didático próprio, investem na formação de uma equipe de professores com alta formação acadêmica e ampla vivência no mercado de trabalho, além de oferecer uma infraestrutura completa, com localização privilegiada e metodologias inovadoras usadas nas melhores instituições do mundo, como jogos empresariais, Aprendizagem Baseada em Projetos, entre outras.

A Pós-Graduação IDAAM atualmente possui o maior número de alunos matriculados em pós-graduação de Manaus, oferecem uma das melhores estruturas de ensino. Na área de engenharia a instituição oferece cursos de pós graduação em Segurança do trabalho, Qualidade 6 sigma e Lean Manufacturing.

3. ATIVIDADES DESENVOLVIDAS

O estagiário esteve sob orientação do Engenheiro José Carlos Reston Filho, aprendeu e trabalhou na área de mineração de dados, mais precisamente, algoritmos de classificação, predição e de clusterização com aplicação no forecast de séries temporais em geral. Além do treinamento no software IBM SPSS Modeler, foi tida a oportunidade de cursar disciplinas na instituição, com a finalidade do aluno desenvolver ferramentas e mais experiência para iniciação nas indústrias que o parque industrial de Manaus oferece.

3.1 Revisão Teórica de Estatística descritiva e inferencial

Primeiramente, foi feita uma revisão em Estatística descritiva, pelo fato de que todo o estudo em mineração de dados requer a utilização de suas ferramentas, tais como, medidas de posição em geral, medidas de dispersão em geral, distribuição normal, regressão linear, correlação, coeficiente de correlação de Pearson, erro médio, correlação e regressão linear múltipla, séries temporais, entre outras.

3.2 Apresentação das Técnicas de Mineração de Dados para classificação, segmentação e predição

A mineração de dados pode ser considerada como uma parte do processo de descoberta do conhecimento em banco de dados (KDD – Knowledge Discovery in Database), onde sua principal tarefa é transformar dados de baixo nível, em informações, ou conhecimento de alto nível. A mineração de dados, é uma das etapas desse processo e pode ser definida como a extração de conhecimento em base dados, com a finalidade de se encontrar padrões e tendências no modelo observado.

O conhecimento que se alcança com a mineração de dados têm se mostrado bastante útil em diversas áreas. Abaixo esta apresentada algumas áreas nas quais a mineração de dados é aplicada de forma satisfatória:

- Retenção de clientes: identificação de perfis para determinados produtos, venda cruzada;
- Bancos: identificar padrões para auxiliar no gerenciamento de relacionamento com o cliente;
- Cartão de Crédito: identificar segmentos de mercado, identificar padrões de rotatividade;
- Cobrança: detecção de fraudes;
- Telemarketing: acesso facilitado aos dados do cliente;
- Eleitoral: identificação de um perfil para possíveis votantes;
- Medicina: indicação de diagnósticos mais precisos;
- Segurança: na detecção de atividades terroristas e criminais;
- Auxílio em pesquisas biométricas;
- RH: identificação de competências em currículos;
- Tomada de Decisão: filtrar as informações relevantes, fornecer indicadores de probabilidade
- Engenharia: Predição de cargas elétricas, ventos, armazenamento de agua em reservatórios de uma companhia hidrelétrica, geração hidráulica e geração térmica, etc.

O uso da mineração de dados permite, por exemplo, que:

- Um supermercado melhore a disposição de seus produtos nas prateleiras, através do padrão de consumo de seus clientes;
- Uma companhia de marketing direcione o envio de mensagens promocionais, obtendo melhores retornos;
- Uma empresa aérea possa diferenciar seus serviços oferecendo um atendimento personalizado;
- Empresas planejem melhor a logística de distribuição dos seus produtos, prevendo picos nas vendas;
- Empresas possam economizar identificando fraudes;
- Agências de viagens possam aumentar o volume de vendas direcionando seus pacotes a clientes com aquele perfil;

A mineração de dados possui várias etapas, opera diferentes qualidades de dados e diversos algoritmos de extração de informações. Um roteiro bem elaborado para a mineração de dados, garante que todas as questões críticas e pontos importantes sejam contemplados de forma que o minerador de dados não se perca em meio as complexidades e consiga atingir os objetivos desejados.

Além do método adequado, o processo de mineração de dados envolve tarefas e algoritmos para a extração de novos conhecimentos. Entre as técnicas de mineração de dados, destacam-se algumas, tais como: associação, classificação, regressão, clusterização, estimação e predição. Entre os vários algoritmos utilizados, os mais usuais são: redes neurais artificiais, árvores de decisão, algoritmos genéticos, lógica nebulosa e estatística.

3.2.1 Técnicas de mineração

3.2.2 Classificação

A classificação é uma das mais utilizadas técnicas de mineração de dados, que visa identificar a qual classe um determinado registro pertence. Essa técnica pode ser utilizada tanto para entender dados existentes, tanto para prever como novos dados irão se comportar. São comuns as tarefas de classificação de clientes em baixo, médio e alto risco de empréstimo bancário, de clientes potencialmente consumidores de um determinado produto a julgar pelo seu perfil, entre outras.

Os algoritmos de árvores de decisão são os mais utilizados para fins de classificação.

3.2.3 Estimação

A estimação é similar a classificação, porém é usada quando o registro é identificado por um valor numérico e não um categórico. Estimar algum índice é determinar seu valor mais provável diante dos outros índices semelhantes sobre os quais se tem conhecimento. Então, podemos estimar o valor de uma determinada variável

analisando-se os valores das demais. Por exemplo, um conjunto de registros contendo os valores mensais gastos por diversos tipos de consumidores e de acordo com os hábitos de cada um. Após a análise dos dados, o modelo é capaz de dizer qual será o valor gasto por um consumidor novo.

Os algoritmos de regressão e as redes neurais são bastante utilizados nesses casos.

3.2.4 Previsão

A previsão está associada à avaliação de um valor futuro de uma variável a partir de dados históricos do seu comportamento no passado. Pode-se prever, por exemplo, o valor de uma ação de uma determinada empresa três meses adiante, podemos também prever futuras vendas de um determinado produto para o planejamento e controle da produção, etc.

Os algoritmos utilizados para previsão são as redes neurais artificias, regressão linear, árvores de decisão, modelos ARIMA, Filtros de Kalman, entre outros.

3.3 Estudo e aplicação do software IBM SPSS Modeler

O IBM SPSS Modeler é um pacote de software abrangente no estilo workbench de mineração de dados que fornece uma maneira de construir modelos preditivos através de uma interface gráfica com o usuário, utilizando diversas técnicas avançadas de modelagem. Além disso, também tem a capacidade para executar várias operações de pré-processamento de dados para melhor preparar os dados a serem modelados.

O pacote também oferece suporte na forma de facilidades de visualização, processamento estatístico, navegação e suporte periférico para acesso e manipulação de dados que permitem o desenvolvimento rápido e fácil de experimentos de mineração de dados.

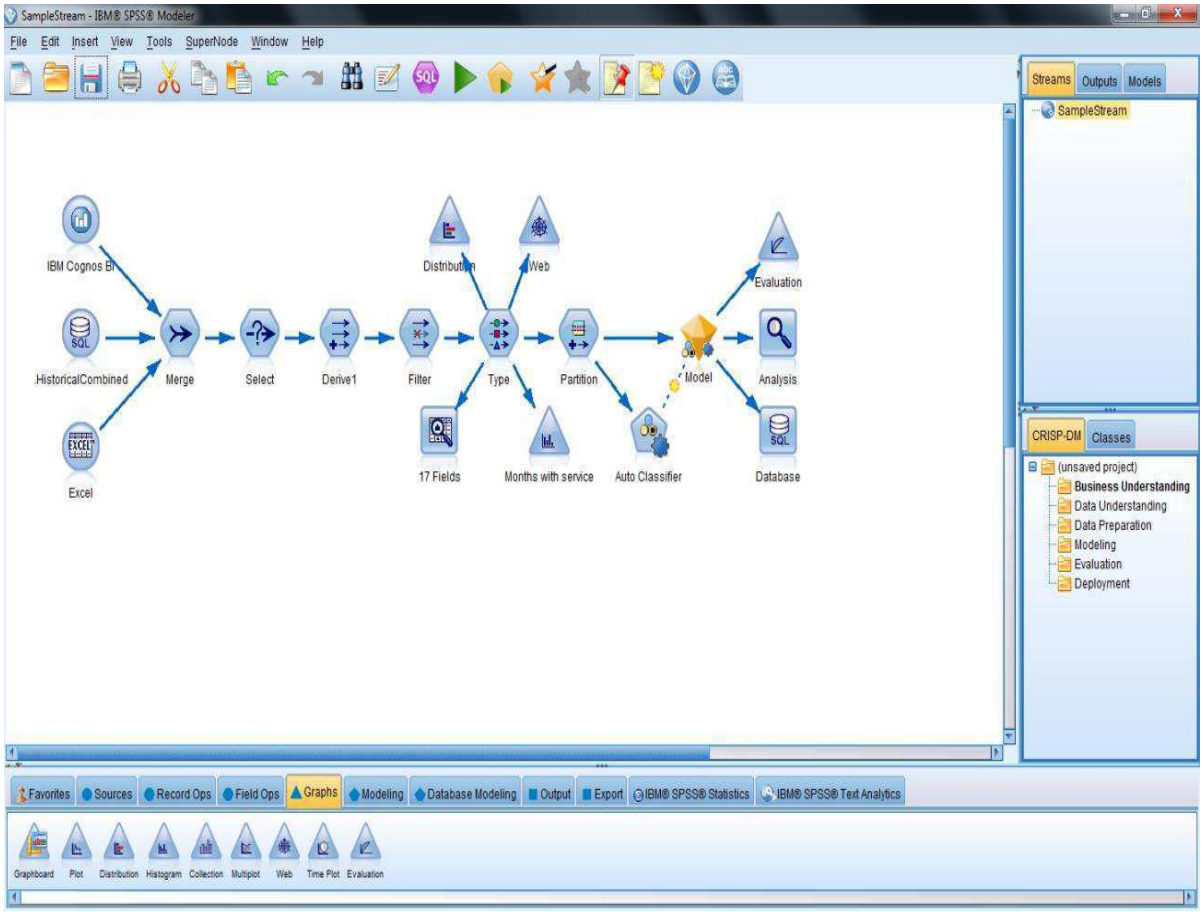

Figura 1- Interface do software IBM SPSS Modeler.

O modelo de mineração de dados usado pelo IBM SPSS Modeler é o CRISP-DM. A sigla significa Cross-Industry Standard Process for Data Mining (CRISP-DM). O método é composto de seis fases, que abordam os principais pontos da mineração de dados. Estas seis fases formam um processo cíclico e cobrem todas as etapas da mineração de dados, inclusive a fase da inclusão dos resultados.

A figura 2 abaixo mostra o modelo CRISP-DM e suas seis fases. A sequência de fases se da de forma não-linear, podendo ocorrer a transição para diferentes fases. As setas indicam as mais importantes e mais frequentes dependências entre as fases.

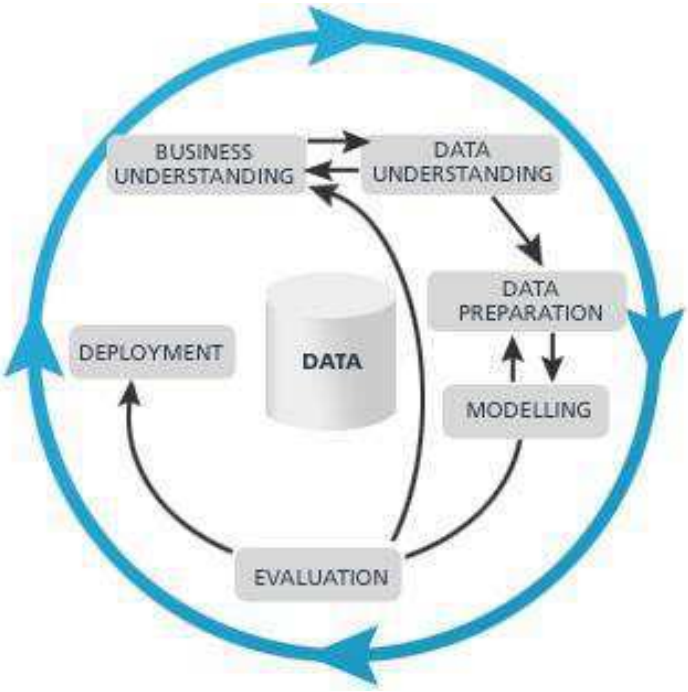

Figura 2 - Processo Cíclico CRISP-DM.

Entendimento dos Problemas: Nessa etapa, o foco é entender qual o objetivo que se deseja atingir com a mineração de dados. O entendimento do problema irá ajudar nas próximas etapas.

Entendimento dos Dados: As fontes fornecedoras dos dados podem vir de diversos locais e possuírem diversos formatos. Após definir os objetivos, é necessário conhecer os dados visando:

- Descrever de forma clara o problema;
- Identificar os dados relevantes para o problema em questão;
- Certificar-se de que as variáveis relevantes para o projeto não são interdependentes.

Preparação dos Dados: Devido às diversas origens possíveis, é comum que os dados não estejam preparados para que os métodos de Mineração de Dados sejam aplicados diretamente. Dependendo da qualidade desses dados, algumas ações podem ser necessárias. Este processo de limpeza dos dados geralmente envolve filtrar, combinar e preencher valores vazios.

Modelagem: É nesta fase que as técnicas (algoritmos) de mineração serão aplicadas. A escolha da(s) técnica(s) depende dos objetivos desejados.

Avaliação: Considerada uma fase crítica do processo de mineração, nesta etapa é necessária a participação de especialistas nos dados, conhecedores do negócio e

tomadores de decisão. Diversas ferramentas gráficas são utilizadas para a visualização e análise dos resultados (modelos).

Testes e validações, visando obter a confiabilidade nos modelos, devem ser executados (cross validation, suplied test set, use training set, percentage split) e indicadores para auxiliar a análise dos resultados precisam ser obtidos (matriz de confusão, índice de correção e incorreção de instâncias mineradas, estatística kappa, erro médio absoluto, erro relativo médio, precisão, F-measure, dentre outros).

Distribuição: Após executado o modelo com os dados reais e completos é necessário que os envolvidos conheçam os resultados.

Durante o estágio, foi feito um treinamento contínuo no uso do IBM SPSS Modeler, para que toda a teoria sobre mineração de dados fosse aplicada, e assim pudessemos realizar projetos e trabalhos na área.

3.4 Forecast de Séries Temporais

Uma série temporal é uma coleção de observações feitas sequencialmente ao longo do tempo. A característica mais importante deste tipo de dados é que as observações vizinhas são dependentes e estamos interessados em analisar e modelar esta dependência. Enquanto em modelos de regressão, por exemplo, a ordem das observações é irrelevante para a análise, em séries temporais a ordem dos dados é crucial. Vale notar também que o tempo pode ser substituído por outra variável como espaço, profundidade, etc.

Como a maior parte dos procedimentos estatísticos foi desenvolvida para analisar observações independentes o estudo de séries temporais requer o uso de técnicas especificas. Dados de séries temporais surgem em vários campos do conhecimento como Economia (preços diários de ações, taxa mensal de desemprego, produção industrial), Medicina (eletrocardiograma, eletroencefalograma), Epidemiologia (número mensal de novos casos de meningite), Meteorologia (precipitação pluviométrica, temperatura diária, velocidade do vento), Engenharia (variação das cargas de energia ao longo o tempo), etc.

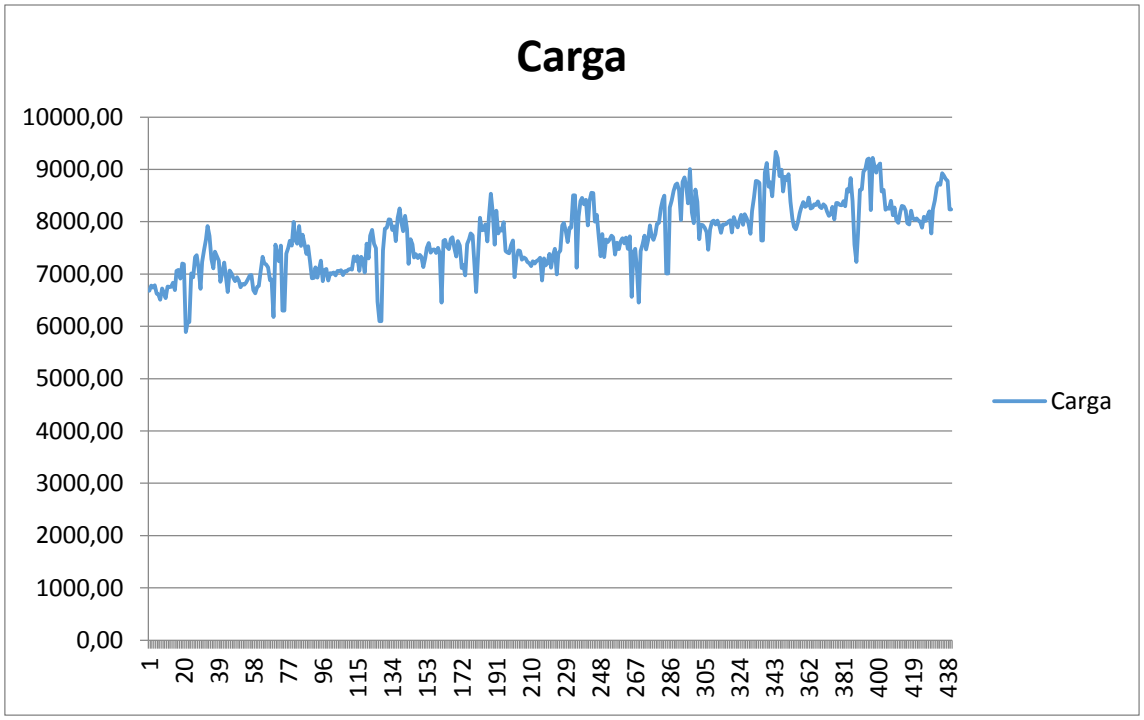

Figura 3 - Série temporal de cargas de energia elétrica.

Durante o estágio, foi possível realizar trabalhos para a predição de séries temporais, tais como a previsão de cargas de energia elétrica das regiões brasileiras, fazer predição para controle de produção na empresa Motor Honda, que se localiza em Manaus e também foi realizada a predição anual de matrículas dos alunos na instituição Pós Graduação IDAAM, etc.

3.5 Extração de conhecimento de bases de dados a partir de clusterização e algoritmos supervisionados

A análise de agrupamentos tem por finalidade formar grupos ou elementos similares entre si (clusters). Um agrupamento, ou cluster, é uma coleção de registros que apresentam semelhanças entre si, entretanto, diferente dos outros registros de demais agrupamentos. A diferença entre agrupamento e classificação, é que na classificação as classes são pré-definidas pelo pesquisador, enquanto no agrupamento, não são.

Na análise de agrupamentos, os grupos ou classes são constituídos com base na semelhança dos dados, então o minerador de dados terá a terefa de analisar as classes resultantes e avaliar se estas terão alguma utilidade. A análise de agrupamentos é uma técnica normalmente preliminar, utilizada quando não se sabe nada, ou quase nada,

sobre os dados. Agrupar ou segmentar um mercado, é um forma usual de análise de agrupamentos, onde consumidores são reunidos em classes representantes dos segmentos desse mercado.

Os algoritmos utilizados para agrupamentos, geralmente são algoritmos estatísticos específicos para esta finalidade, porém as redes neurais e os algoritmos genéticos são utilizados para estes fins, onde nas redes neurais a clusterização irá fazer parte do aprendizado não supervisionado. A figura 4 abaixo, mostra a formação de 3 grupos (clusters).

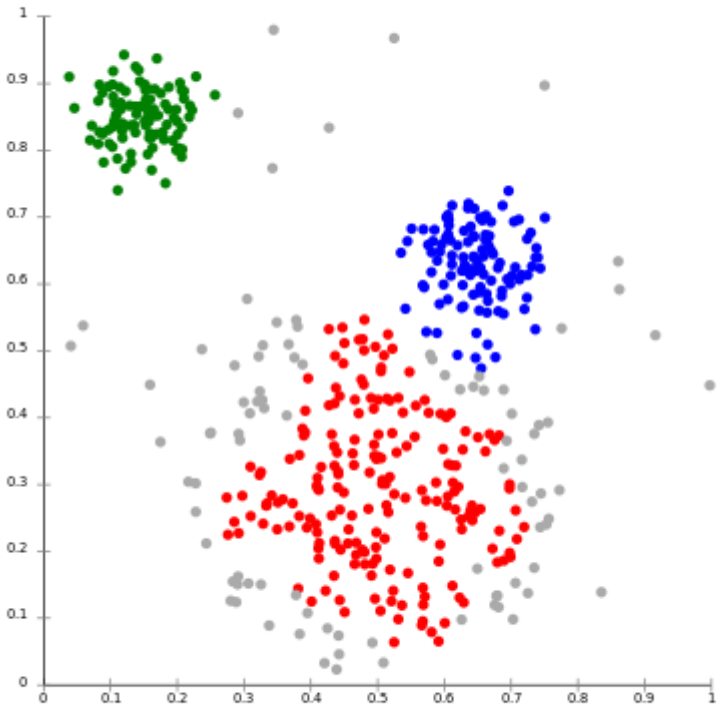

Figura 4 - Registros agrupados em três clusters.

Foi feito um trabalho com a finalidade de validar algoritmos utilizados para extrair padrões que possam auxiliar na tomada de decisões e posteriormente realizar predições de matrículas de alunos da instituição Pós Graduação IDAAM, onde a idéia central se da em prever dias de grande volume de matrículas muito mais do que o quantitativo de matrículas (vendas) de um dia. Assim criamos grupos de clusters que possuem tamanhos variados.

Assim podemos estudar se um grande volume de vendas é precedido por dois ou três dias de vendas baixas. Ou se um dia de vendas grandes influenciam os dias posteriores.

Portanto foi utilizado o algoritmo K-means, separados em 6 clusters. O banco de dados nos fornece a quantidade de matrículas feitas a cada dia da semana, de segunda-feira à sábado, durante todo o ano, de janeiro a dezembro, no caso temos 365 registros. Abaixo é ilustrado na figura 5, como ficou o processo de clusterização do dados feita a partir do algoritmo K-means do software IBM SPSS Modeler.

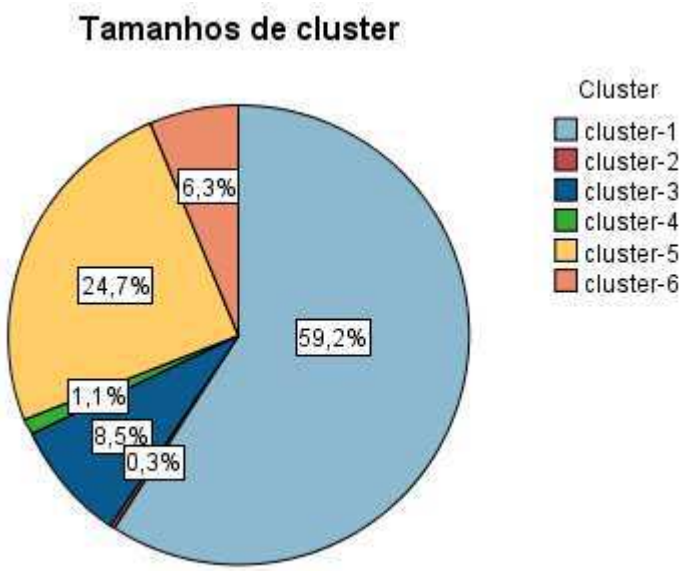

Figura 5 - Clusterização dos dados a partir do algoritmo K-means.

Valor	Proporção	%	Contagem
cluster-1		59.18	216
cluster-2		0.27	1
cluster-3		8.49	31
cluster-4		1.1	4
cluster-5		24.66	90
cluster-6		6.3	23

Figura 6 - Distribuição do K-means.

A figura 6 mostra a proporção dos clusters gerados, onde cada cluster significa a quantidade de matrículas realizadas no dia, cluster - 1 correspodem a 0, 1, 2, 3 matrículas, cluster - 2 corresponde a 29 matrículas, cluster - 3 correspondem a 9, 10, 11 e 12 matrículas, cluster - 4 corresponde a 19, 20, 21 e 22 matrículas, cluster - 5 correspondem a 4, 5, 6, 7, 8 matrículas e cluster - 6 correspondem a 13, 14, 15, 16, 17 e 18 matrículas. Os números que não aparecem, como 23, 24, 25, 26, 27 e 28 se da pelo

fato de não estarem registrados no banco de dados, por exemplo, não existe nenhum dia que houveram 23 matrículas.

Nota-se que, para a maioria dos atributos, tem-se maior concentração de registros para uma classe e um número muito pequeno pertencente à outras classes. Esta má distribuição tem influência no resultado obtido pelos algoritmos de Mineração de Dados, dificultando a indução de regras.

O modelo clusterizado será a entrada de um bloco de partição, onde optamos separar 75% das amostras para treinamento e 25% para teste, o bloco de partição será a entrada para o bloco do algoritmo de árvores de decisão C5.0. A stream pode ser visualizada na figura 7 a seguir.

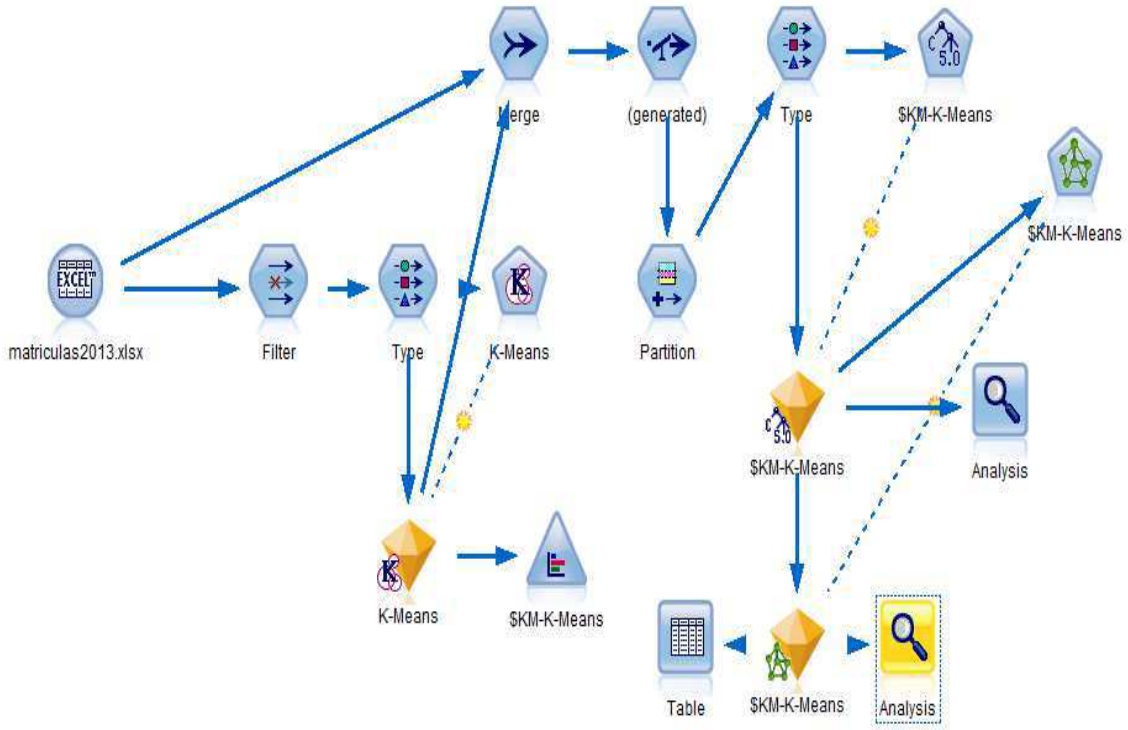

Figura 7 – Stream canvas do projeto com a modelagem do algoritmo C5.0.

Os resultados para o algoritmo C5.0, com o objetivo de obter uma melhoria na classificação gerada pelo K-means, dos dados da matrícula, podem ser visualizados na firura 8 abaixo.

$KM-K-Means		cluster-1	cluster-2	cluster-3	cluster-4	cluster-5	cluster-6
cluster-1	Count	137	0	3	0	20	0
	Row %	85.625	0.000	1.875	0.000	12.500	0.000
cluster-2	Count	0	154	0	0	0	0
	Row %	0.000	100.000	0.000	0.000	0.000	0.000
cluster-3	Count	0	0	160	0	0	0
	Row %	0.000	0.000	100.000	0.000	0.000	0.000
cluster-4	Count	0	0	0	169	0	0
	Row %	0.000	0.000	0.000	100.000	0.000	0.000
cluster-5	Count	16	0	0	0	145	0
	Row %	9.938	0.000	0.000	0.000	90.062	0.000
cluster-6	Count	0	0	0	0	0	165
	Row %	0.000	0.000	0.000	0.000	0.000	100.000

Figura 8 – Resultados para o algoritmo C5.0.

Percebemos que para o cluster – 1, a capacidade de predição dos resultados chega a 85,62% corretamente, para o cluster – 2, o resultado é de 100%, assim como para o cluster – 3, cluster – 4 e cluster – 6. Para o cluster – 5 a capacidade de predição pode chegar a 90,06%.

Uma das grandes opções do software, é o poder de junção de vários modelos ou algoritmos diferentes com o objetivo de maximizar os resultados, neste caso, valores de predição das matrículas.

A etapa a seguir é usar os valores previstos pelo algoritmo C5.0 como entradas de uma rede neural de multiplas camadas (RNMC). A rede neural (figura 9) possui 5 entradas, que são, semana, mês, dia, a previsão do algoritmo C5.0 ($C-$KM-K-Means) e os valores de confiança do modelo C5.0 ($CC-$KM-K-Means), valores esses que podem variar de 0 a 1, quanto mais próximo de 1, mais confiáveis são os valores. A camada intermediária possui 7 neurônios, onde esses neurônios foram gerados automaticamente pelo software, obedecendo um critério de exatidão mínimo de 97%, e a camada de saída, será o nosso alvo, que é o modelo clusterizado, realizado pela função K-means.

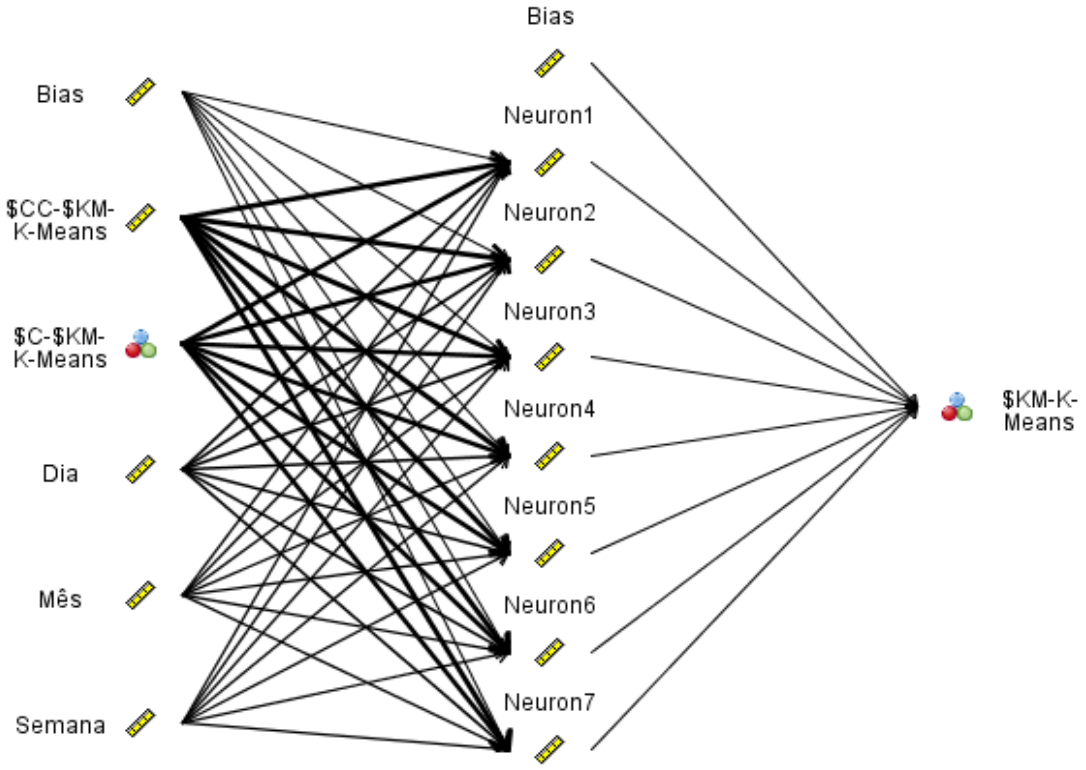

Figura 9 – Rede Neural de multiplicas camadas.

Foi estabelecido, antes de gerar o modelo final da rede neural, que seriam utilizados 75% das informações, para a fase de treinamento da rede neural e 25% para a fase de testes. A figura 10 abaixo mostra uma comparação de resultados do K-means + C5.0, onde o resultado final se apresenta um taxa de acerto de 96,07% para fase de treinamento e 94,48% para fase de testes e o modelos final K-means + C5.0 + RNMC, onde a exatidão é de 98,82% para fase de treinamento e 98,72% para fase de testes.

Results for output field $KM-K-Means

Comparing $C-$KM-K-Means with $KM-K-Means

'Partition'	1_Training		2_Testing	
Correct	930	96,07%	308	94,48%
Wrong	38	3,93%	18	5,52%
Total	968		326	

Comparing Agreement with $KM-K-Means

'Partition'	1_Training		2_Testing	
Correct	923	98,82%	309	98,72%
Wrong	11	1,18%	4	1,28%
Total	934		313	

Figura 10 – Resultados finais com o uso da RNMC.

Finalmente, foi concluida a tarefa de validar os algoritmos e a criação de um modelo bom o bastante, para que se quisermos fazer predições das matrículas, poderemos ter um modelo confiável para se trabalhar.

3.6 Criação de modelos híbridos com dois ou mais algoritmos

Os métodos híbridos representam uma técnica já bastante difundida na literatura especializada e que demonstra a viabilidade de sua utilização, visando especialmente extrair as melhores características de modelos distintos, em favor da obtenção dos melhores resultados.

Durante este estágio, foram feitos trabalhos utilizando o modelo ARIMA de Box & Jenkins, combinados com a Rede Neural Perceptron Multicamadas, via algoritmo backpropagation, responsável pela previsão de cargas, valendo citar como suas principais características a facilidade de solução de problemas complexos e trabalhar bem com a não-linearidade.

Para fins de previsão de cargas elétricas, os modelos híbridos têm sido bastante difundidos, com resultados satisfatórios em relação a outros já descritos na literatura especializada. Os resultados dos trabalhos demonstraram de forma clara a viabilidade de se combinar métodos distintos, no intuito de extrair as melhores características de cada

modelo, que no caso foram visando a precisa previsão de cargas de curto prazo de forma que foram desenvolvidos para captarem as características lineares e não-lineares das séries temporais.

Um dos trabalhos realizados durante o estágio foi a predição de estoque de motos da Moto Honda – Manaus, o banco de dados nos fornece informações de vendas mensais entre janeiro de 2005 à agosto de 2013, o modelo aplicado com um modelo híbrido conjugando ARIMA + Rede Neural Perceptron multicamadas, na figura 11 abaixo, podemos visualizar o gráfico comparando o modelo previsto e o modelo original da série temporal.

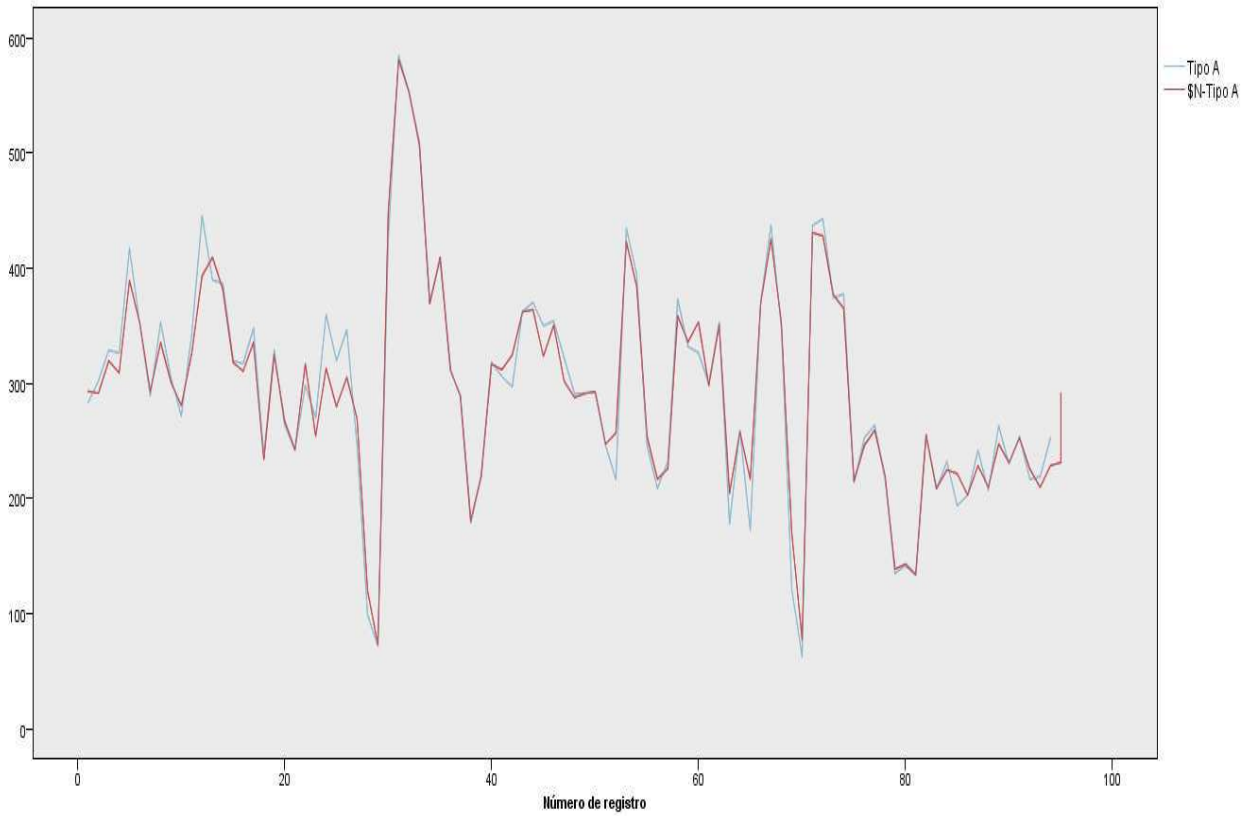

Figura 11 - Comparação do modelo previsto (vermelho) e o modelo original (azul).

Os resultados são conclusivos em apontar o modelo híbrido como um método efetivo para a predição futura das vendas de motos da empresa, ou até uma planejamento da produção, que no trabalho proposto foram previstos 5 passos (meses) a frente, foram atingidos um coeficiente de correlação de Pearson de 0,98 e um erro médio de 1,2% apenas.

A importância de fazer a previsão do controle da produção de uma determinada indústria, nesse caso a Moto Honda, é prever futuras vendas de um determinado produto, para um melhor planejamento da produção, o que nos possibilitaria minimizar

as chances de ocorrer uma baixa produção e possivelmente o estoque de produtos não atender a grande demanda de vendas de determinado mês, ou então, em situação análoga, podemos prever uma superprodução e evitar o excesso de estoque, evitando assim adicionais custos para a empresa.

3.7 Avaliação da qualidade, robustez e capacidade de generalização de modelos

O setor elétrico é de fundamental importância ao desenvolvimento econômico do país, dessa maneira o mercado fica cada vez mais competitivo e exigem técnicas cada vez mais precisas para predição de cargas de energia para garantir a segurança quanto ao fornecimento de energia, entre outros fatores.

Durante a fase de modelagem do processo de mineração de dados, podemos usar diferentes tipos de modelos e algoritmos seja qual for nosso objetivo. Um dos projetos do estágio, foi o de prever cargas de energia, com a finalidade de facilitar a tamada de decisões, melhor planejamento do fluxo de potência, maximização de lucros, minimizar perdas, minimizar erros, possibilidade de planejar manutenções preventivas, melhoria da segurança e um melhor planejamento de expansão do setor elétrico.

Uma das técnicas de mineração para a predição de cargas de energia, são as redes neurais artificiais. O software IBM SPSS Modeler, nos oferece uma gama de ferramentas para a avaliação da qualidade, além de que as redes neurais possuem uma grande capacidade de generalização, ou seja, possui a capacidade de identificar padrões nos dados para os quais nunca foram treinados. Também possui uma robustez computacional ao lidar com erros no conjunto de treinamento

O software IBM SPSS Modeler, para métodos de avaliação do modelo de redes neurais, conta com ferramentas como desvio padrão, erro médio, erro médio absoluto, correlação linear de Pearson, erro máximo, erro mínimo, elementos visuais gráficos como comparação entre o modelo original e o modelo previsto, entre outros.

3.8 Uso de Principal Components Analysis no pré-processamento de dados

O uso da função Principal Components Analysis nos fornece poderosas técnicas de redução de dados para reduzir a complexidade de seus dados. Duas abordagens semelhantes, porém distintos são fornecidos.

Uma delas é a Análise de Componentes Principais (PCA) que encontra combinações lineares dos campos de entrada que fazem o melhor trabalho de capturar a variação em todo o conjunto de campos, onde os componentes são ortogonais (perpendiculares) entre si. O PCA se concentra em toda variância, incluindo tanto compartilhada quanto variância única.

A outra abordagem é a análise fatorial, ela tenta identificar conceitos subjacentes, ou fatores, que explicam o padrão de correlações dentro de um conjunto de campos observados. A análise fatorial se concentra em apenas variância compartilhada. Essa variância é exclusiva para campos específicos que não são considerados na estimativa do modelo. Vários métodos de análise fatorial são fornecidos pelo nó Factor / PCA.

Para ambas as abordagens, o objetivo é encontrar um pequeno número de campos derivados que resumem as informações do conjunto original de campos.

Os pontos fortes desse método são que a análise fatorial e PCA pode efetivamente reduzir a complexidade de seus dados sem sacrificar grande parte do conteúdo da informação. Estas técnicas podem ajudar a construir modelos mais robustos que são executados mais rapidamente do que seria possível com os dados em seu estado normal.

Um dos projetos duranto o estágio, foi o de testar um modelo híbrido utilizando técnicas de predição de séries temporais. O modelo preditivo era composto pelas técnicas ARIMA(p,d,q) + Redes Neurais de Mútiplas Camadas (RNMC). O objetivo do modelo híbrido é o de conjugar as melhores caracteristicas de cada técnica fazendo com que ocorro uma maximização dos resultados da previsão.

O projeto consistia em predição de cargas de energia para a região nordeste do Brasil. A série histórica de valores de carga de energia foi obtida junto ao Operador Nacional do Sistema – ONS a partir de dados coletados do Informativo Preliminar Diário da Operação – IPDO. A série está registrada em base semanal e na unidade de

medida Mega Watt Médio (MWMed) e correspondem a um período de 04/01/2003 a 29/12/2009.

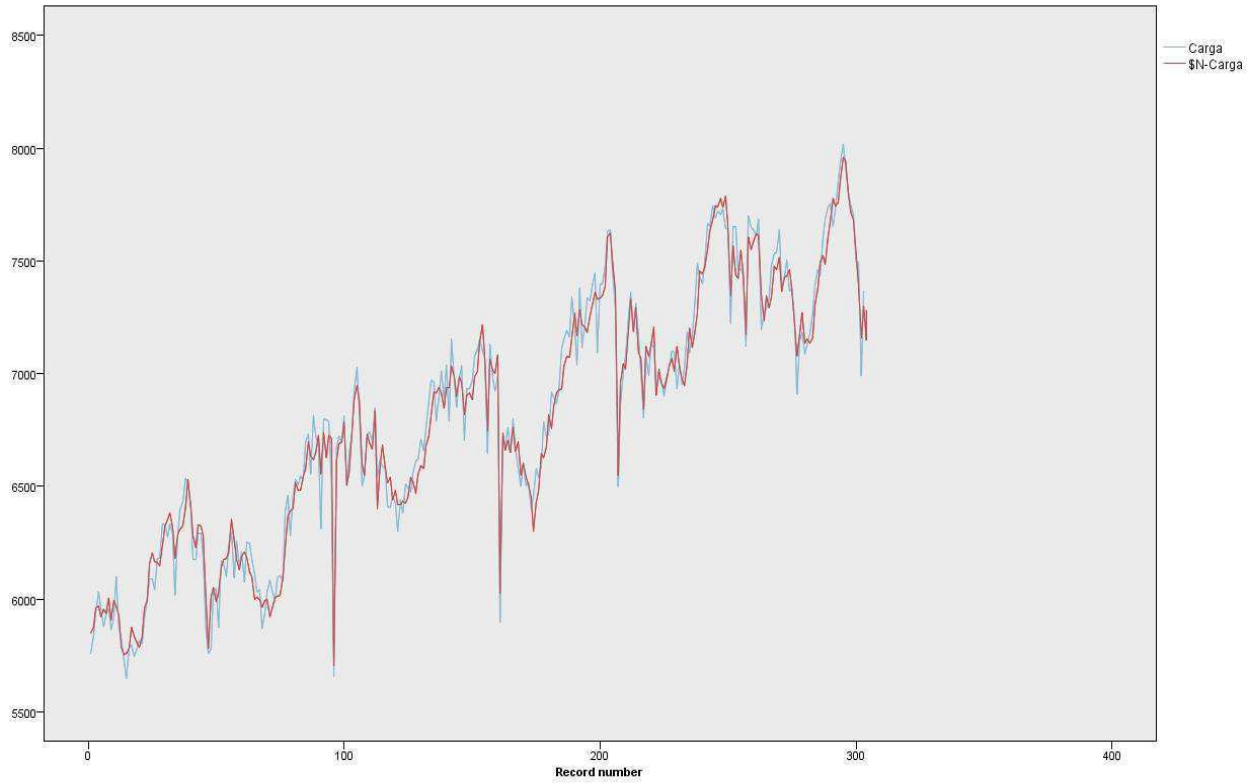

Figura 12 – Previsão de cargas de energia com o uso do Factor / PCA.

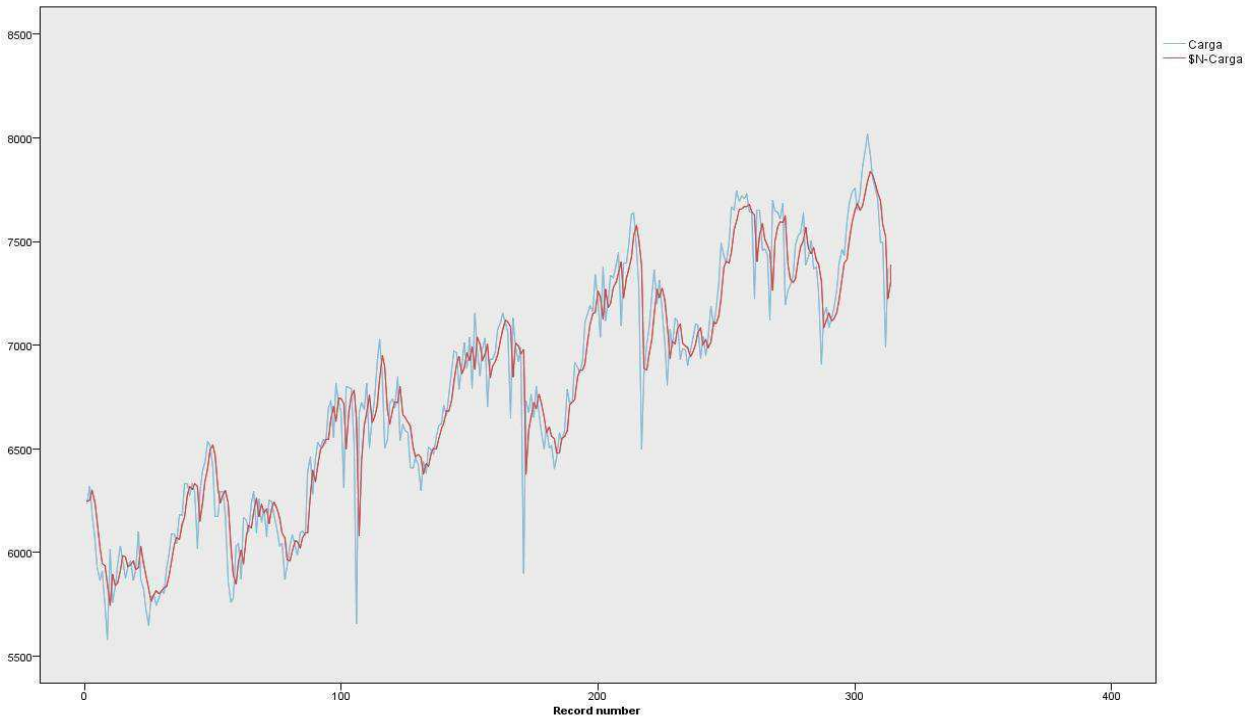

Figura 13 – Previsão de cargas de energia sem o uso do Factor / PCA.

As figuras 12 e 13 mostram a comparação entre a série temporal de cargas de energia real e a série prevista, com e sem o uso do factor / PCA, respectivamente, para o pré processamento de dados do modelo. Para a previsão sem o uso do PCA obtemos um índice de correlação linear de 0,952, enquanto com o uso do PCA, se obteve um índice de 0,988, para ambos os casos em uma previsão de 5 passos (semanas) a frente. Percebe-se que com o uso do nó PCA obtemos resultados superiores fazendo com que seu uso seja de crucial importância para o pré-processamento de dados.

3.9 Disciplinas cursadas

Durante o estágio integrado, foram tidas oportunidades de cursar disciplinas oferecidas pela instituição, na área de Engenharia de Produção com ênfase em Lean Manufacturing. O curso é uma especialização de importância fundamental para formar profissionais aptos ao planejamento e controle da produção através da filosofia da manufatura enxuta, baseada no Sistema Toyota de produção (Toyota Production System). Neste sistema de produção a mentalidade enxuta (lean thinking) tem enfoque na identificação e eliminação de desperdícios, a manutenção apenas das atividades que agregam valor aos produtos e serviços, a gestão visual e inteligente dos estoques baseada em Kanban e na produção puxada (pull system), além das ferramentas próprias de Lean.

As disciplinas oferecidas neste curso e que foram cursadas pelo estagiário foram:

Ciclo Básico:

- Liderança e Desenvolvimento Profissional (36h)
- Criatividade e Aprendizagem Baseada em Problemas (36h)
- Gerenciamento de Projetos (24h)
- Métodos quantitativos (24h)
- Metodologia Científica (36h)
- Competências Comportamentais (24h - 4 Palestras)

Disciplinas Específicas:

- Lean Manufacturing (36h)
- Manutenção Produtiva Total (36h)
- Gestão Visual (24h)
- Value Stream Map (24h)
- Planejamento e Controle da Produção (36h)
- Lean Office (24h)

4. CONSIDERAÇÕES FINAIS

O estágio realizado na empresa IDAAM Educação Superior LTDA. proporcionou ao aluno novos conhecimentos adquiridos em uma área que é de crucial importância para a tomada de decisões em qualquer empresa, que é a mineração de dados. Também proporcionou a consolidação dos conhecimentos adquiridos durante toda a graduação de Engenharia Elétrica, já que a realização dos projetos necessitava de uma base de conhecimentos prévia por parte do estagiário em várias disciplinas, tais como Álgebra linear, probabilidade, processos estocásticos, gerenciamento de energia elétrica, distribuição de energia elétrica, entre outras. Além disso, o estagiário pôde ter contato diretamente com engenheiros eletricistas na empresa, profissionais atuantes na área de projetos de consultoria.

Além disso, o estagiário teve oportunidades de ganhar experiência em áreas como gerenciamento de projetos, processos de Lean Manufacturing, qualidade seis sigma, processos Supply Chain, entre outros. Essas áreas de conhecimento são ferramentas de crucial importância para quem deseja atuar na indústria.

Deste modo, o estágio serviu para proporcionar um grande crescimento profissional e pessoal, além do contato direto com profissionais com grande experiência na área e o desenvolvimento de novos conhecimentos em áreas como a Engenharia da Produção.

5. REFERÊNCIAS BIBLIOGRÁFICAS

BOX, G.E.; JENKINS, G.M**. Times series analysis: forecasting and control.** San Francisco: Holden - Day, 1976. 575 p.

Cross Industry Standard Process for Data Mining (CRISP-DM) Disponível em: <http://www.crisp-dm.org/>. Acesso em: 12 de jan. de 2014.

HAYKIN, S. (1999) **Neural networks: a comprehensive foundation.** 2 ed. New Jersey: Prentice-Hall, 1999, p.842.

IBM SPSS Modeler Disponível em: <http://www-03.ibm.com/software/products/pt/spss-modeler/> Acesso em: 10 de jan. de 2014.

U. M. FAYYAD, et al. (Eds.) **Advances in Knowledge Discovery and Data Mining**, AAAI Press/MIT Press: 1996.

ZHANG, G. P. **Time series forecasting using a hybrid ARIMA and neural network model. Neurocomputing,** Atlanta, v. 50: 2003, p. 159-175.

Manual IBM SPSS Modeler Training: **Predictive Modeling with IBM SPSS Modeler.**

ANEXO – Software IBM SPSS Modeler

A1. Sobre o IBM SPSS Modeler

O software IBM® SPSS Modeler é um conjunto de ferramentas de mineração de dados que lhe permitem desenvolver rapidamente modelos preditivos utilizando o conhecimento do problema a que se quer resolver e implantá-los em operações de negócios para melhorar a tomada de decisão. Concebido em torno do modelo CRISP-DM padrão da indústria, SPSS Modeler suporta todo o processo de mineração de dados, a partir de dados para melhores resultados de negócios ou problemas.

O SPSS Modeler oferece uma variedade de métodos de modelagem tiradas de aprendizado de máquina, inteligência artificial, e as estatísticas. Os métodos disponíveis na paleta Modeling permitem obter novas informações a partir de seus dados e desenvolver modelos preditivos. Cada método tem alguns pontos fortes e é mais adequada para determinados tipos de problemas.

A2. Visão geral do software

Como uma aplicação de mineração de dados, o IBM® SPSS Modeler oferece uma abordagem estratégica para encontrar relações úteis em grandes conjuntos de dados. Em contraste com os métodos estatísticos mais tradicionais, você não precisa necessariamente saber o que você está procurando quando você começa. Você pode explorar os dados, ajuste de diferentes modelos e investigar diferentes relações, até encontrar informações úteis.

Uma visão geral da interface do software pode ser observada na figura abaixo, onde percebe-se que o software fornece ao usuário uma maneira de contruir modelos para determinados objetivos através de uma interface gráfica com o usuário.

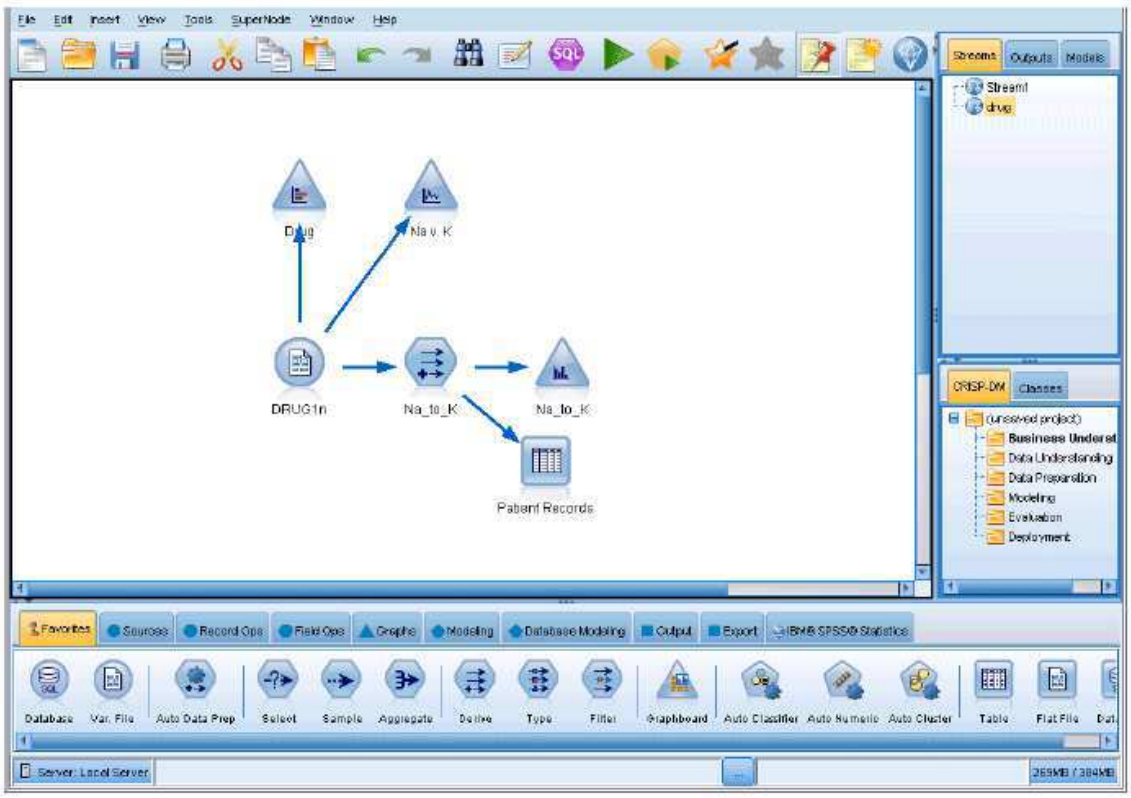

Figura 14 – Interface do sofware.

Em cada ponto do processo de mineração de dados, a interface é fácil de usar, o IBM® SPSS Modeler convida os seus conhecimentos de negócio específico. Algoritmos de modelagem, tais como previsão, classificação, segmentação e detecção de associação, assegurar modelos potentes e precisos. Os resultados do modelo podem ser facilmente implantados e lidos em bancos de dados, IBM® SPSS Statistics, e uma grande variedade de outras aplicações.

Trabalhar com SPSS Modeler é um processo de três etapas de trabalho com dados.

- Em primeiro lugar, você lê os dados no SPSS Modeler.
- Em seguida, você executa os dados por meio de uma série de manipulações.
- Finalmente, você envia os dados para um destino.

Esta seqüência de operações é conhecida como *data stream*, porque os fluxos de dados vão passando registro por registro da fonte através de cada manipulação e, finalmente, para o destino, ou um modelo ou tipo de saída de dados.

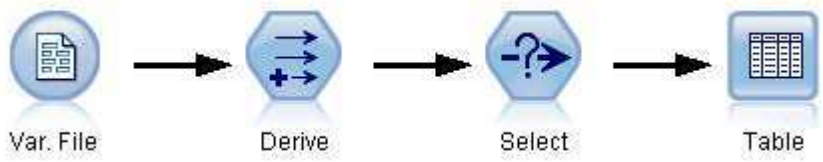

Figura 15 – Sequência de uma data stream básica.

A3. IBM SPSS Modeler stream canvas

A *stream canvas* é a maior área da janela do IBM ® SPSS Modeler e é o lugar onde você vai construir e manipular *data streams*.

Streams são criados pelo desenho de diagramas de operações de dados relevantes para o seu objetivo na tela principal da interface. Cada operação é representado por um ícone ou nó, e os nós estão ligados entre si em uma stream que representa o fluxo de dados através de cada operação. Você pode trabalhar com múltiplos fluxos de uma só vez em SPSS Modeler, ou na mesma tela corrente ou abrindo uma nova tela corrente. Durante a sessão, as streams são armazenadas no gerenciador de streams, no canto superior direito da janela do SPSS Modeler.

A4. Paleta de nós ou funções do software

A maioria das ferramentas de modelagem e de manipulação dos dados no IBM® SPSS Modeler se encontra na paleta de nós, na parte inferior da janela abaixo da stream canvas.

Por exemplo, na guia da 'Record Ops' (figura 16) contém os nós que você pode usar para executar operações sobre os registros de dados, como a seleção, fusão e acrescentar.

Para adicionar nós para a tela, clique duas vezes em ícones da Paleta de nós ou arraste e solte-os na tela. Você, então, pode conectá-los e criar uma stream, o que representa o fluxo de dados.

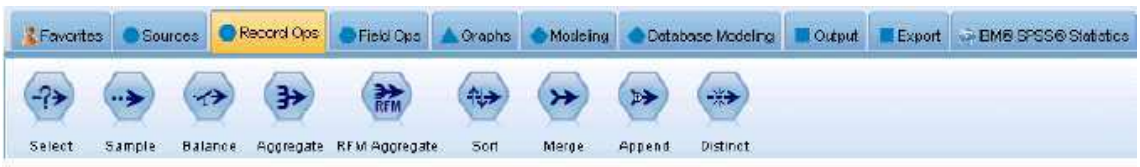

Figura 16 – Record Ops na Paleta de nós.

Cada guia da paleta contém uma coleção de nós relacionados utilizados para diferentes fases das operações das streams , tais como:

- *Sources* – São nós que trazem dados para o SPSS Modeler.
- *Record Ops* – Nós que realizam operações em registros de dados , como a seleção , fusão, e acrescentar .
- *Field Ops* – Nós que executam operações em campos de dados , como filtrar , derivar novos campos e determinar o nível de medida para determinados campos.
- *Graphs* – Nós que apresentam graficamente os dados antes e depois da modelagem. Os gráficos incluem gráficos como histogramas, nós web e gráficos de avaliação.
- *Modeling* – Nós que usam os algoritmos de modelagem disponíveis no SPSS Modeler , como redes neurais, árvores de decisão, algoritmos de agrupamento e sequenciamento de dados.
- *Database modeling* – Nós que usam os algoritmos de modelagem disponíveis no Microsoft SQL Server, IBM DB2 e Oracle.
- *Output* – Nós que produzem uma variedade de saída para dados , gráficos e os resultados do modelo que podem ser vistos no SPSS Modeler.
- *Export* – Nós que produzem uma variedade de saídas que pode ser vistas em aplicações externas, tais como a coleta de dados do IBM® SPSS ou Excel.
- *SPSS Statistics* – Nós que importam dados de, ou exportam dados para , IBM ® SPSS ® Statistics, ou outros softwares que o minerador desejar, bem como executam procedimentos SPSS Statistics.

A5. Uso do nó C5.0

O software contém quatro diferentes algoritmos para a contrução de uma árvore de decisão (mais geralmente conhecido como indução de regras), são eles: C5.0, CHAID, Quest, e C&R Tree (árvores de classificação e regressão). Eles são semelhantes na medida em que podem construir uma árvore de decisão através de dados de forma recursiva de divisão em subgrupos definidos pelos campos de previsão como eles se relacionam com o alvo.

Usaremos o nó C5.0 para criar um modelo de indução de regras. Ele contém o modelo de indução de regras em qualquer árvore de decisão ou conjunto de regras de formato. Pelo default do software, o nó C5.0 é rotulado com o nome do campo de saída. O modelo C5.0 pode ser pesquisado e as previsões podem ser feitas através da passagem de novos dados através dele na Stream canvas.

Antes da stream ser usada pelo nó C5.0 ou, essencialmente, qualquer nó na paleta, os níveis de medição de modelagem de todos os campos usados no modelo deve ser instanciada (ou no nó *source* ou um nó do *type*). Isso porque todos os nós de modelagem usam essas informações para criar os modelos.

A figura 17 abaixo mostra como é uma stream simples usando o nó C5.0.

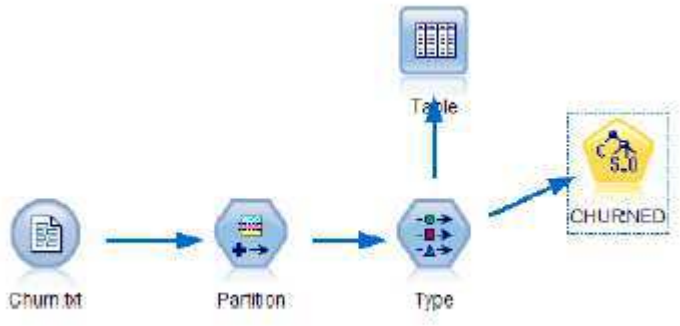

Figura 17 – Stream usando o nó C5.0

A figura 18 abaixo mostra as opções que podem ser usadas para a modelagem do nó C5.0.

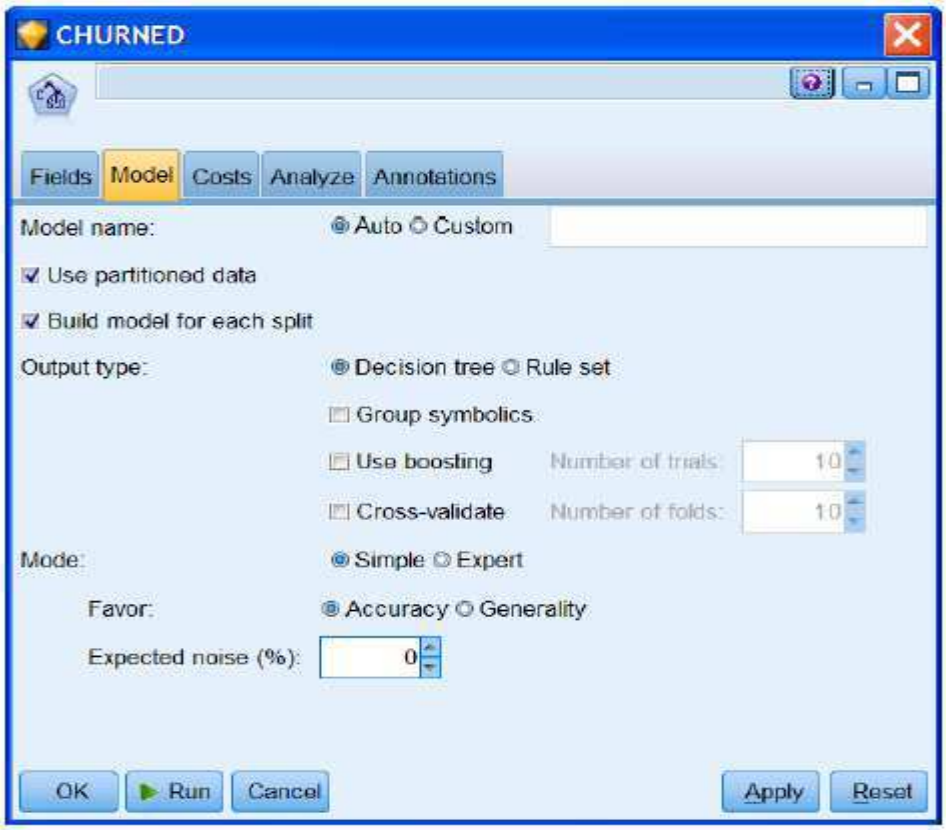

Figura 18 – Opções do nó C5.0.

A opção Model name permite que você defina o nome tanto para o C5.0 e nós resultantes da regra C5.0 . A forma (*decision tree* ou *rule set*) do modelo resultante é selecionado usando o tipo de saída: option.

A opção *Use partitioned data* é verificada para que o nó C5.0 fará uso do campo *Partition* criado pelo nó de partição no início da stream. Sempre que esta opção estiver marcada , apenas os casos o nó de partição atribuído à amostra de Treinamento será utilizado para a construção do modelo, o resto dos casos será realizada para fins de testes e / ou de validação . Se nada for feito , o campo será ignorado e o modelo será treinado em todos os dados. Por default, usamos a configuração padrão para o nó de partição , por isso 50% dos casos será usado para treinamento e 50% para o teste.

A opção *Build model for each split* permite que você use uma única stream para construir modelos separados para cada valor possível de um tipo de dado, podendo ser categorical ou continuous input field, que é especificado como campo de divisão na guia Fields tab ou nó Type. Com o campo split, você pode facilmente construir o modelo

mais apropriado para cada valor possível nos campos em apenas uma única execução da stream.

A opção *Cross-validation* fornece uma maneira de validar a precisão dos modelos C5.0 quando há poucos registros nos dados que permitam uma amostra de teste separado. Ele faz isso dividindo os dados em N subgrupos de igual tamanho e se encaixa modelos N . Cada modelo usa (N- 1) dos subgrupos de formação, em seguida, aplica-se o modelo resultante para o subgrupo restante e registra a precisão. Figuras da precisção são agrupados ao longo dos subgrupos de validação N e este sumário estatístico estima a melhor precisão do modelo aplicados aos novos dados. Uma vez que modelos N estão aptos , a validação de N-folds exigem mais recursos e relatam a estatística de precisão , mas não apresenta as árvores de decisão N ou conjuntos de regras. Por padrão N, o número de folds , é definido como 10.

Para um campo preditor que tem sido definido como categórica , C5.0 forma, normalmente, um ramo por valor no conjunto. No entanto, marcando a caixa de seleção *Group Symbolic*, o algoritmo pode ser definido de modo que ele encontra agrupamentos sensatos dos valores dentro do campo, reduzindo assim o número de regras . Este é muitas vezes desejável . Por exemplo, em vez de ter uma regra por região do país, os valores simbólicos do grupo podem produzir uma regra como:

Região [do Sul, Centro-Oeste] ...

Região [Nordeste, Norte] ...

Uma vez treinados, C5.0 constrói uma árvore de decisão ou conjunto de regras que podem ser usados para previsões. No entanto, também podem ser instruídos para construir uma série de modelos alternativos para os mesmos dados, seleccionando a opção de *Boosting*. Com esta opção, quando se faz uma previsão, é consultado cada um dos modelos alternativos antes de tomar uma decisão. Isso muitas vezes pode fornecer uma previsão mais precisa, mas leva mais tempo para treinar. Além disso, o modelo resultante é um conjunto de previsões de árvore de decisão e o resultado é determinado pelo voto, o que não é simples de interpretar.

O algoritmo pode ser configurado para favorecer a Precisão nos dados de treinamento (por default) ou Generalidade de outros dados.

O nó C5.0 vai tratar automaticamente erros (ruído) dentro dos dados e, se conhecido, você pode informar ao spss Modeler a proporção esperada de dados ruidosos ou errados. Esta opção é raramente usada.