Pós-Graduação em Ciência da Computação

UMA APLICAÇÃO DE MINERAÇÃO DE DADOS AO PROGRAMA BOLSA ESCOLA DA PREFEITURA DA CIDADE DO RECIFE

por

ROBERTO TABOSA FLORENCIO FILHO

Dissertação de Mestrado

Universidade Federal de Pernambuco
posgraduacao@cin.ufpe.br
www.cin.ufpe.br/~posgraduacao

RECIFE, ABRIL DE 2009

Universidade Federal de Pernambuco

CENTRO DE INFORMÁTICA

PÓS-GRADUAÇÃO EM CIÊNCIA DA COMPUTAÇÃO

ROBERTO TABOSA FLORENCIO FILHO

UMA APLICAÇÃO DE MINERAÇÃO DE DADOS AO PROGRAMA BOLSA ESCOLA DA PREFEITURA DA CIDADE DO RECIFE

Orientador: Prof. Dr. Paulo Jorge Leitão Adeodato

Dissertação apresentada à Pós-Graduação em Ciência da Computação do Centro de Informática da Universidade Federal de Pernambuco como requisito parcial à obtenção do grau de Mestre em Ciência da Computação.

RECIFE, ABRIL DE 2009

Florencio Filho, Roberto Tabosa

Uma aplicação de mineração de dados ao programa bolsa escola da prefeitura da cidade do Recife / Roberto Tabosa Florencio Filho. - Recife: O Autor, 2009.

ix, 100 folhas : il., fig., tab., quadros, gráf.

Dissertação (mestrado) – Universidade Federal de Pernambuco. CIn. Ciência da Computação, 2009.

Inclui bibliografia e apêndice.

1. Mineração de dados. 2. Inteligência artificial. 3. Redes neurais artificiais. I. Título.

006.312 CDD (22. ed.) MEI2010 – 050

AGRADECIMENTOS

Agradeço, a todos aqueles que de alguma forma participaram, me incentivaram e torceram para o êxito desse trabalho. De forma especial, vão meus sinceros agradecimentos:

A Deus, por ter me concedido condições de força, saúde e uma família maravilhosa para conseguir chegar a conclusão deste trabalho.

Ao meu pai, em memória, exemplo de caráter, perseverança e dignidade, que sempre foi minha referência.

À minha esposa pela compreensão e apoio nos momentos mais difíceis e incertos. À minha filha por ela existir e ser uma força natural à minha superação nas dificuldades encontradas.

À minha mãe e irmãos pela torcida e palavras de incentivo.

Ao meu orientador, Professor Paulo Adeodato, pelas chances que me foram dadas, pelas críticas e principalmente pela postura firme na cobrança das atividades.

Ao amigo e chefe, Moacir Antônio Marafon, pela compreensão demandada durante essa jornada.

Ao amigo José Antônio Manso, por ter conseguido junto a Prefeitura da Cidade do Recife e a EMPREL a base de dados para que esse trabalho pudesse ser realizado.

Aos colegas David Fernandes França e Fabiano N. de Carvalho, pela importante troca de conhecimentos durante a disciplina de sistemas de apoio a decisões e mineração de dados.

À Tereza Sato e Rafaela Ávila, pela atenção e disponibilidade sempre que precisei de esclarecimentos sobre o Programa Bolsa Escola da Prefeitura da Cidade do Recife.

RESUMO

UMA APLICAÇÃO DE MINERAÇÃO DE DADOS AO PROGRAMA BOLSA ESCOLA DA PREFEITURA DA CIDADE DO RECIFE

A tarefa de Mineração de Dados envolve um conjunto de técnicas de estatística e inteligência artificial com objetivo de descobrir informações não encontradas por ferramentas usualmente utilizadas para extração e armazenamento de dados em grandes bases de dados. A aplicação da Mineração de Dados pode ser realizada em qualquer área de conhecimento (Ciências Exatas, Humanas, Sociais, Biológica, Saúde, Agrária e outras) proporcionando ganhos de informações e conhecimentos, ora desconhecidos, em qualquer uma delas.

Este trabalho apresenta uma aplicação de mineração de dados ao programa Bolsa Escola da Prefeitura da Cidade do Recife (PCR), particularmente na investigação da situação das famílias beneficiadas, com o objetivo de oferecer à administração municipal uma ferramenta de suporte à decisão capaz de aprimorar o processo de concessão de benefícios. Foi analisada uma massa de dados sócio-econômicos inicialmente de cerca de 60 mil famílias cadastradas no programa. Foi utilizada uma rede neural artificial MultiLayer Perceptron (MLP) para classificar as famílias beneficiadas com base nas suas características sócio-econômicas.

A avaliação de desempenho e resultados obtidos, além da resposta da especialista no domínio de aplicação, demonstram a viabilidade dessa aplicação no processo de concessão do benefício ao Programa Bolsa Escola da Prefeitura da Cidade do Recife.

Palavras-chave: Descoberta de Conhecimento em Bases de Dados, Cross-Industry Standard Process for Data Mining (CRISP-DM), Mineração de Dados, Redes Neurais, Programa Bolsa Escola Municipal (PBEM).

ABSTRACT

A DATA MINING APPLICATION TO THE RECIFE CITY COUNCIL BOLSA ESCOLA PROGRAM.

The Data Mining task involves a set of statistics techniques and artificial intelligence aiming to discover information which were not found by tools usually used for data extration and storage in large databases. The Data Application Mining can be done in any area of knowledge, such as Exact, Human, Social, Biological, Health and Agrarian Sciences, and others, providing information and knowledge gaining, now unknown, in any one of them.

This study presents an application of data mining to the Recife City Council Bolsa Escola Program, particularly in the inquiry of the benefited families' situation, with the objective of offering the municipal administration a decision support tool able to improve the concession process of benefits. A social-economic data mass was initially analyzed in about 60 thousand families registered in the program. An artificial neural net MultiLayer Perceptron (MLP) was used to classify the benefited families based on their social-economic characteristics.

Performance avaliation and the obtained results, besides the specialyst answer in the application mastery, showed its viability in the benefit concession process of the Recife City Council Bolsa Escola Program.

Keywords: Knowledge Discovery in databases, Cross-Industry, Standard Process for Data Mining (CRISP-DM), data mining, neural networks, Council Bolsa Escola Program.

SUMÁRIO

LISTA DE FIGURAS

LISTA DE GRÁFICOS

LISTA DE QUADROS

PRINCIPAIS ABREVIATURAS

CRISP-DM	*Cross-Industry Standard Process for Data Mining*
CSR	Custo Social Representativo
DASE	Diretoria de Apoio Social à Educação
ECA	Estatuto da Criança e do Adolescente
IBGE	Instituto Brasileiro de Geografia e Estatística
IDH	Índice de Desenvolvimento Humano
KDD	*knowledge Discovery in Databases*
KS	*Kolmogorov-Smirnov*
MLP	*Multilayer Perceptron*
MSE	*Mean Squared Error*
ONU	Organização das Nações Unidas
PBEM	Programa Bolsa Escola Municipal
PCR	Prefeitura da Cidade do Recife
PNAD	Pesquisa Nacional por Amostras de Domicílio
PNUD	Programa das Nações Unidas para o Desenvolvimento
RNA	Redes Neurais Artificiais
ROC	*Receiver Operating Characteristics*
SM	Salário Mínimo

1 Introdução

Inicialmente propostos pelo governo federal nos anos 90, os programas de renda mínima e bolsa escola são um dos fenômenos mais significativos no cenário das políticas sociais no Brasil.

Presentes em várias unidades da federação, estes programas têm em comum o enfoque familiar e a vinculação com a educação infantil, bem como a percepção de que a concessão de um benefício monetário surte melhores resultados que as políticas assistenciais tradicionais [AR99].

Isto se dá principalmente porque muitas famílias não têm condições de manter os filhos na escola em razão de sua baixa qualidade de vida e extrema pobreza; e, com renda insuficiente, contam com o trabalho das crianças para seu sustento, caracterizando o trabalho infantil como um dos principais impedimentos à freqüência escolar [CS03].

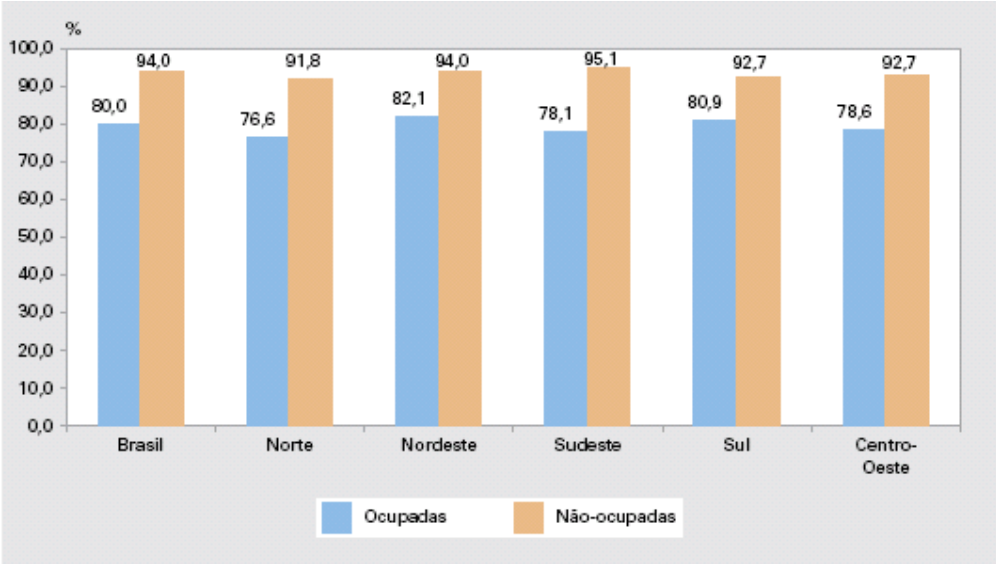
Dados da Pesquisa Nacional por Amostras de Domicílio (PNAD[1]) publicados em 2007 [PNAD07] reforçam esta afirmação, mostrando que em todas as Regiões do Brasil, as taxas de escolarização são maiores em grupos de pessoas que não trabalham (Gráfico 1).

Estudos estatísticos [MS99] comprovam que em geral, as crianças que têm pais analfabetos, sem garantia de renda, alimentação insuficiente, falta de energia, água encanada e esgoto, entre outros problemas de sobrevivência, têm seu aprendizado comprometido, em comparação com outras que não apresentam o mesmo grau de pobreza.

1.1 Motivação e Justificativa

Trabalhos com enfoque social, econômico ou educacional já foram desenvolvidos a cerca do Programa Bolsa Escola em Prefeituras de Estados

[1] Pesquisa Nacional por Amostras de Domicílio. Realizado anualmente pelo Instituto Brasileiro de Geografia e Estatística (IBGE).



Fonte: IBGE, Diretoria de Pesquisas, Coordenação de Trabalho e Rendimento, Pesquisa Nacional por Amostra de Domicílios 2007.

Gráfico 1 - Taxa de escolarização das pessoas de 5 a 17 anos de idade, por situação de ocupação na semana de referência, em 2007 retirado de [PNAD07].

brasileiros; vários outros trabalhos também já foram realizados sobre Mineração de Dados, contudo, não existem estudos específicos utilizando técnica de Mineração de Dados voltada para avaliar a influência de características sócio-econômicas das famílias beneficiadas pelo Programa Bolsa Escola Municipal (PBEM) da Prefeitura da Cidade do Recife (PCR) no sucesso do mesmo.

Apesar do incentivo financeiro, algumas famílias não conseguem manter seus dependentes na escola e, por isso, deixam de ser caracterizados como beneficiários, prejudicando a missão do programa.

Assim, este trabalho foi motivado pelo interesse acadêmico em oferecer aos gestores públicos o aperfeiçoamento do processo decisório de concessão ao benefício como forma de incrementar a eficiência do programa e otimizar o investimento destinado para o mesmo. Para este trabalho foram consideradas as características sócio-econômicas de todos os requerentes apropriados e afastados do Programa. Posteriormente foram criados os rótulos Sucesso e Insucesso em um novo atributo, para

melhor caracterizar cada requerente do Programa conforme será explicado na seção 5.4.2.

1.2 Objetivos do Trabalho

Aperfeiçoar o processo decisório de concessão ao benefício, oferecendo uma ferramenta de suporte à decisão, utilizando método de mineração de dados baseado na análise dos dados cadastrais disponíveis dos requerentes apropriados e afastados, de forma a permitir aos gestores do programa, aprimorar a concessão de benefícios pelo programa Bolsa Escola da Prefeitura da Cidade do Recife.

O resultado obtido assegura aos especialistas no domínio da aplicação, a disponibilização de um importante instrumento como critério de escolha dos novos requerentes apropriados ao benefício.

1.3 Estrutura da Dissertação

Esta dissertação encontra-se estruturada em capítulos distribuídos da forma a seguir relacionada:

- Capítulo 2 - descreve o ambiente social e educacional brasileiro, além de detalhar o funcionamento do Programa Bolsa Escola da Prefeitura da Cidade do Recife.
- Capítulo 3 – apresenta conceitos sobre a descoberta do conhecimento em bases de dados.
- Capítulo 4 – apresenta conceitos sobre Redes Neurais Artificiais.
- Capítulo 5 – descreve, através da metodologia CRISP-DM, os processos de entendimento do negócio, entendimento e preparação dos dados aplicados ao Programa Bolsa Escola.

- Capítulo 6 – descreve, através da metodologia CRISP-DM, os processos de modelagem e avaliação aplicados à base de dados do Programa Bolsa Escola.
- Capítulo 7 - finaliza o trabalho apresentando as conclusões, limitações existentes e sugestões de trabalhos futuros.

2 Conjuntura Econômica, Educação e o Programa Bolsa-Escola

Este capítulo descreve a conjuntura sócio-econômica que envolve a educação e relata o funcionamento e a estrutura existente para apoiar o Programa Bolsa Escola da Prefeitura da Cidade do Recife.

2.1 Conjuntura Econômica e Educação

Programas de inclusão social por transferência de renda são algumas das formas encontradas pelos governantes de minimizar problemas ligados a:

- Exclusão social;
- Necessidade social e
- Pobreza extrema.

2.1.1 Educação - Direito Básico

A educação no Brasil é um direito fundamental garantido a todo cidadão brasileiro pela Constituição de 1988, no seu artigo 6º[Const98].

Os parágrafos I, II e III do artigo 208 (“O dever do Estado com a educação será efetivado mediante a garantia de:”) da Constituição Federal Brasileira, legislam sobre a forma como o Poder Público deve proceder para oferecimento de educação básica de qualidade a todos os cidadãos [Const98]. O Estatuto da Criança e do Adolescente (ECA) Lei 8.069/90, prevê ainda no artigo 54 do inciso VII, parágrafo 3º, obrigatoriedade do Poder Público, em fornecer o devido acesso ao ensino fundamental através de “programas suplementares de material didático-escolar, transporte, alimentação e assistência à saúde” [ECA09].

Este direito, nas últimas décadas, vem ocupando local de destaque entre os direitos humanos, sendo considerado um direito essencial e indispensável para o exercício da cidadania. Esse local de destaque no âmbito do poder público dá-se essencialmente por três motivos:

- Nenhum dos outros direitos seja civil, político, econômico e social pode ser devidamente exercido sem um mínimo de educação;
- A educação reduz as desigualdades sociais que é, segundo o Banco Mundial [TWB44], um dos fatores de impacto para redução da pobreza;
- A educação é uma das três dimensões que compõe a fórmula de cálculo do Índice de Desenvolvimento Humano (IDH) para classificação da Organização das Nações Unidas (ONU) conforme é demonstrado na seção 2.1.5 [PNUD06].

Apesar de todos os compromissos feitos pelos governantes por meio de instrumentos internacionais, a educação básica no Brasil, de maneira geral, necessita de incentivos para sua efetiva melhoria e, programas sociais por transferência de renda é um tipo de ação que vem sendo adotada.

2.1.2 Exclusão Social

A exclusão social diz respeito a pessoas ou grupos desfavorecidos. É a fase extrema do processo de marginalização sendo esse, o ponto máximo atingível onde o individuo, de forma gradativa, se afasta da sociedade em conseqüência de rupturas consecutivas com a mesma [Cos98].

Nem sempre uma única característica determina a situação de exclusão social, por exemplo, a situação em que uma família de baixa renda pode ser considerada pobre, mas estar integrada a um grupo comunitário. Uma situação (pobreza) pode direcionar à outra (exclusão social), mas não necessariamente levá-la até esta.

2.1.3 Necessidade Social

Karl Max[2] descreveu a necessidade social como algo que só se pode afirmar através da pressão exercida sobre os indivíduos, com objetivo de que as decisões deles tenham uma determinada orientação. Nesse contexto, a necessidade social está fortemente ligada à necessidade da inclusão social e da garantia dos direitos básicos sociais garantidos por lei[3]: educação, saúde, trabalho, lazer, dentre outros.

2.1.4 Pobreza Extrema

Nessas condições encontram-se indivíduos cujas necessidades básicas para sobrevivência humana não são satisfeitas.

Em um panorama melhorado tem-se a pobreza moderada, quando as necessidades básicas são satisfeitas, mas os indivíduos continuam privados de necessidades básicas sociais como entretenimento, saúde e educação.

A metodologia utilizada para definição de pobreza no Brasil, tem o salário mínimo como referência. São consideradas as classificações [SEPLAG08]:

- Famílias extremamente pobres, aquelas com renda per capita de até ¼ do salário mínimo;
- Famílias pobres, aquelas com renda per capita de até ½ do salário mínimo;

2.1.5 Índice de Desenvolvimento Humano

O Índice de Desenvolvimento Humano (IDH) é uma medida comparativa que contempla três dimensões: riqueza, educação e longevidade. Foi criado em 1990 pelo

2 Karl Heinrich Marx – Nascido em Tréveris na Alemanha, ao dia 5 de maio de 1818, foi um intelectual e revolucionário alemão, fundador da doutrina comunista moderna. O pensamento de Marx influencia até hoje várias áreas de conhecimento.

3 Direitos Básicos presentes no Art. 6º da Constituição Federal de 1988.

economista paquistanês Mahbub ul Haq. Posteriormente teve a colaboração do economista indiano Amartya Sen, ganhador do Prêmio Nobel de Economia de 1998 e vem sendo utilizado desde 1993 pelo Programa das Nações Unidas para o desenvolvimento do seu relatório anual. Os países membros da ONU, anualmente recebem uma classificação de acordo com o IDH [HDI08].

O último relatório anual do IDH foi publicado em 2008 utilizando como fonte de dados para o levantamento, dados da população atualizados de 2006. Segundo a pesquisa, o Brasil está na septuagésima colocação dentre 179 países, com IDH de 0,807 em uma medida que vai de 0 a 1. Ainda segundo o relatório, o país possui estimativa de vida em torno de setenta e dois anos de idade e taxa de alfabetização de adultos (acima dos quinze anos de idade) de 89,6% [HDR08].

A Pesquisa Nacional por Amostra de Domicílios mais recente, com dados do ano de 2007, mostra que, a nível Brasil, no grupo de crianças e adolescentes de 6 à 14 anos, o percentual de freqüência escolar foi de 97%, sendo que 79,2% correspondem ao ensino da rede pública. A pesquisa revela ainda que as maiores taxas de escolaridades estão nas faixas de idade de 7 à 9 anos e 10 à 14 anos com 98,1% e 97,4%, respectivamente [PNAD07] o que demonstra a grande concentração de crianças que, de alguma forma, estão comprometidas com os estudos e por outro lado, uma janela de oportunidade do poder público em gerar melhores condições para que essas crianças e jovens não abandonem seus estudos.

Dessa forma, fica claro enxergar a concessão do benefício do Programa Bolsa Escola da Prefeitura da Cidade do Recife como um Programa de Inclusão Social, uma vez que, além da transferência de renda propiciando melhores condições econômicas, atende às necessidades sociais realizando dezenas de trabalhos sócio-educativos com as famílias beneficiadas[4].

[4] Sistemas Sociais Básicos: São agrupados em cinco domínios – o social, o econômico, o institucional, o territorial e o das referências simbólicas.

2.2 O Programa Bolsa Escola

O Programa Bolsa Escola da Prefeitura da Cidade do Recife é um programa de transferência de renda e inclusão social, criado pela Lei 16.302 e regulamentado pelo Decreto 17.665 em 1997, conduzido pela Diretoria de Apoio Social à Educação (DASE), vinculada à Secretaria de Educação, Esporte e Lazer, tendo como principal objetivo atender às famílias carentes, sendo este atendimento condicionado a matrícula e permanência das crianças em escola municipal do Recife.

2.2.1 Objetivo do Programa Bolsa Escola

O objetivo maior do Programa é reduzir os índices de evasão escolar nas escolas municipais. As famílias beneficiadas recebem um incentivo financeiro para motivar a permanência de suas crianças na escola. O valor do benefício é de metade de um salário mínimo vigente (atualmente duzentos e trinta e dois reais e cinqüenta centavos) para quem tem uma criança na escola ou um salário mínimo vigente (atualmente quatrocentos e sessenta e cinco reais) para quem tem duas ou mais crianças na escola. Sendo esse valor pago apenas pela Prefeitura da Cidade do Recife. Neste contexto o valor do benefício caracteriza-se como uma compensação financeira para garantir a permanência das crianças na escola. Atualmente são investidos em média, 1,9 milhão de reais por mês para pagamento das famílias beneficiadas [PCR08].

2.2.2 Funcionamento do Programa Bolsa Escola

As informações presentes nesta seção foram obtidas através de entrevistas e relatórios entregues pela Diretoria de Apoio Social à Educação (DASE) da PCR.

Todos os critérios para inclusão de uma família no programa devem ser satisfeitos. São eles:

- Ter o requerente realizado o cadastro;
- Famílias com crianças em idade de seis a quatorze anos e onze meses;
- Famílias com renda per capita de até 1/3 do salário mínimo vigente;

- Crianças matriculadas que estudem em escolas da Rede Municipal do Recife;

- Famílias que morem no Recife há pelo menos cinco anos consecutivos;

Em 2001 foram inscritas 60.451 famílias. Os dados destas famílias são utilizados até o momento para realização de novas inclusões ao benefício, não havendo novas inscrições, uma vez que, existem várias famílias a serem contempladas e que esperam pelo benefício.

O processo de concessão ao benefício segue uma série de etapas que são descritas a seguir. Os dados informados no momento do cadastro do interessado (requerente) são armazenados em base de dados em que o sistema SEBE[5], a partir desses dados, gera uma pontuação[6] e os requerentes considerados mais necessitados recebem uma pontuação mais alta e os menos necessitados recebem uma pontuação mais baixa. Algumas das características do requerente e suas respectivas pontuações estão listadas no Apêndice A. Existindo disponibilidade financeira, os maiores pontuados são listados e visitados para confirmação das informações. Caso continue com o perfil, é feita por uma comissão, uma avaliação final da família candidata, para finalmente proceder a efetivação dos requerentes classificados que, com isso, passam a receber mensalmente o benefício conquistado. A inclusão do requerente passará sua situação de "pontuado" para "apropriado".

Para permanecer no Programa, além de continuar satisfazendo os critérios de inclusão o requerente precisa estar morando com a criança (dependente), justificar as faltas na escola, acompanhar o desenvolvimento da criança na escola, usar adequadamente o benefício e participar das atividades do Programa.

---

[5] Sistema desenvolvido em 2001 para o Programa Bolsa Escola Municipal (PBEM), através de parceria entre a Secretaria de Educação, Esporte e Lazer (SEEL) e a Empresa de Informática do Município (EMPREL).

[6] A pontuação do requerente pelo SEBE é dada pela soma das pontuações de cada característica sócio-econômica cadastrada e confirmada após visita do entrevistador. Algumas características são mostradas no Apêndice A.

Se algum dos critérios deixar de ser satisfeito, a família (requerente) será afastada do Programa sendo sua situação no banco de dados alterada de “apropriado” para “afastado”. A exceção está para o caso de requerentes com mais de um dependente e algum(ns) desses dependentes deixa(m) de satisfazer às condições de permanência no Programa devido a idade (ao completar 14 anos e 12 meses); para essa situação, apenas o(s) dependente(s) é afastado do Programa, mas a família (requerente) continua recebendo o benefício. A estrutura e funcionamento da DASE são descritos na Ilustração 1:

Organograma:

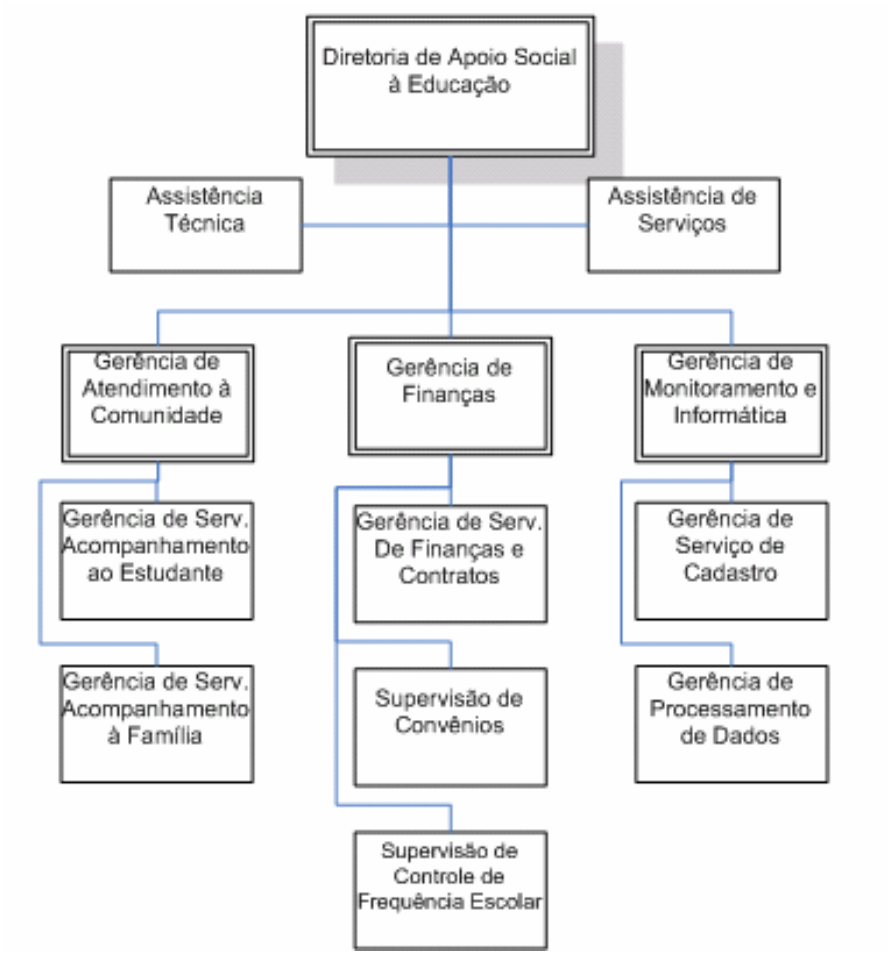


Ilustração 1 - Organograma da estrutura organizacional para controle do Bolsa Escola

Atribuições:

Gerência de Atendimento à Comunidade – responsável pela inclusão e acompanhamento das famílias beneficiadas, avaliando a situação sócio-econômica e a obediência aos critérios estabelecidos para permanência no Programa, realizando visitas domiciliares e atendimento aos requerentes e beneficiários na própria Diretoria.

Gerência de Serviço de Acompanhamento ao Estudante – realiza visitas domiciliares aos requerentes e beneficiários, para inclusão, acompanhamento dos dependentes, averiguação de denúncias e atendimento à solicitação dos Conselhos Tutelares, Escolas Municipais, Juizado de Menores, ONG´s, entre outros.

Gerência de Serviço de Acompanhamento à Família – realiza atendimento ao público, prestando informações sobre os critérios do Programa, sobre a importância do bom uso do benefício concedido, atualizando informações cadastrais e posicionando os requerentes sobre situação atual do seu cadastro.

Gerência de Finanças – responsável pela folha de pagamento dos beneficiários do Programa e todas as etapas que precedem o pagamento, além do controle mensal da freqüência escolar.

Gerência de Serviços de Finanças e Contratos – realiza alterações na folha de pagamento por motivo de inclusão de novos beneficiários e desligamentos; realiza também alterações no valor concedido ao beneficiário e solicita bloqueios e desbloqueios à Caixa Econômica Federal mensalmente.

Supervisão de Convênios – realiza validação das alterações efetuadas na folha de pagamento acompanhando a situação dos beneficiários recebedores do benefício de forma complementar (Programas Bolsa Família e Bolsa Escola) e quando necessário, visitando-os.

Supervisão de Controle de Freqüência Escolar – Controla e acompanha a freqüência escolar municipal.

Gerência de Monitoramento e Informática – Responsável pela área de informática e armazenamento dos dados cadastrais do Programa através dos Sistema Bolsa Escola (SEBE) e Folha de Pagamento (FGET), bem como geração de relatórios bimestrais sobre o Programa Bolsa Escola Municipal (PBEM).

Gerência de Serviço de Cadastro – Realiza controle e arquivamento dos cadastros de todos os requerentes e beneficiários inscritos no Programa, elaborando técnicas de arquivamento para facilitar extração dos dados quando solicitados.

Gerência de Serviço de Processamento de Dados – Realiza digitação de dados cadastrais do PBEM nos sistemas, além de ser responsável pelo acompanhamento da manutenção e instalação de micros e programas de informática.

Assistências Técnica e de Serviços – Prover apoio direto a Diretoria de Apoio Social à Educação.

O fluxo de processos do Programa Bolsa Escola da Prefeitura da Cidade do Recife é mostrado na Ilustração 2.

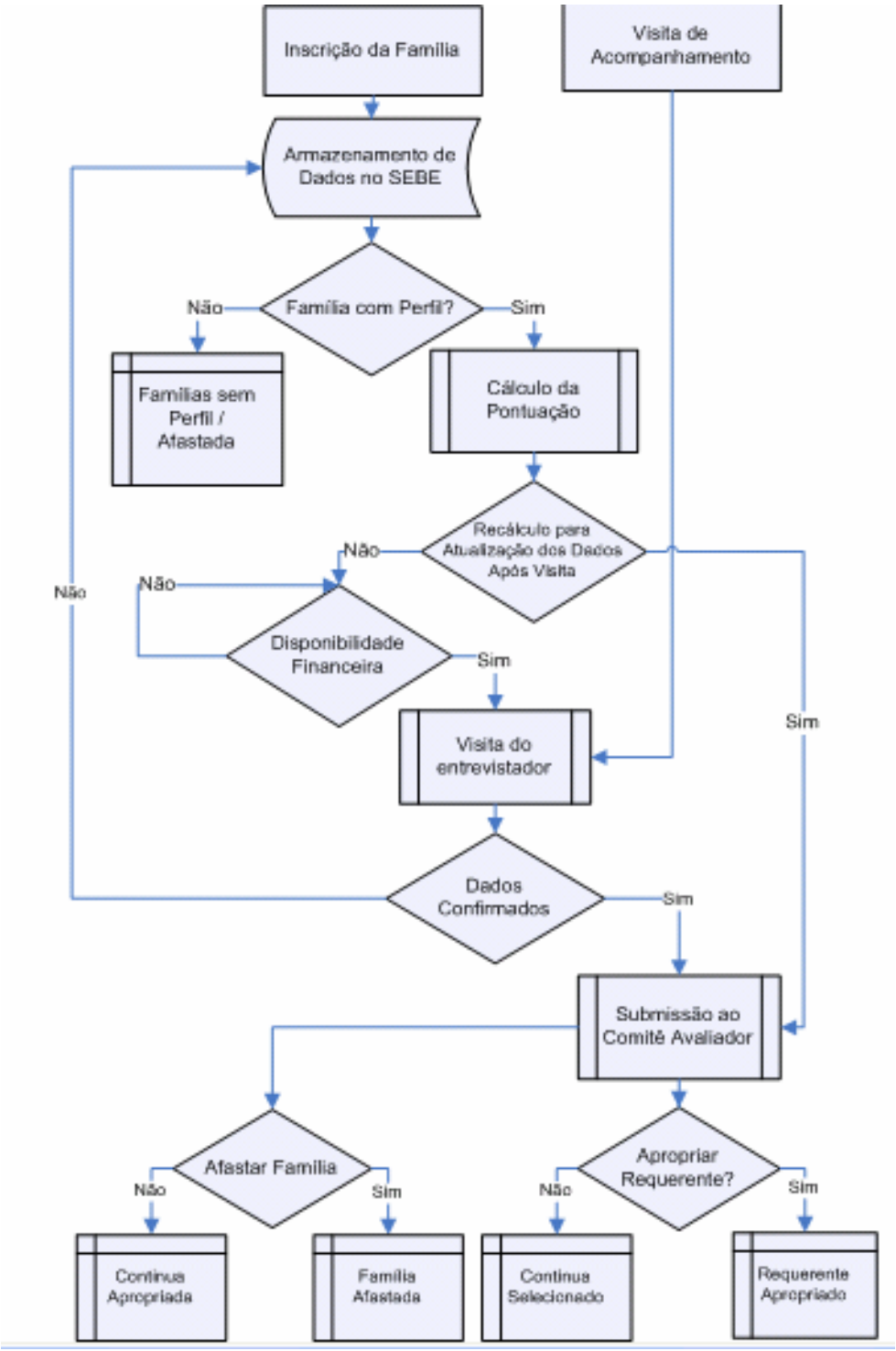


Ilustração 2 - Fluxo de processos do Programa Bolsa Escola da PCR

3 Descoberta de Conhecimento em Bases de Dados

Neste capítulo são abordadas as fases para o processo de descoberta do conhecimento através de bancos de dados.

3.1 Introdução

O processo de descoberta do conhecimento em bancos de dados foi definido por Fayyad, juntamente com outros pesquisadores, em 1996 como sendo "*um processo, de várias etapas, não trivial*[7]*, interativo*[8] *e iterativo*[9] *para identificação de padrões compreensíveis, válidos, novos e potencialmente úteis a partir de grandes conjuntos de dados*" [FSS96]. O processo de *knowledge Discovery in Databases* (KDD) foi definido em cinco fases: seleção de dados, pré-processamento dos dados, transformação, mineração de dados e interpretação dos resultados. As fases de todo o processo de KDD são mostradas na Ilustração 3 e descritas nas próximas seções deste capítulo.

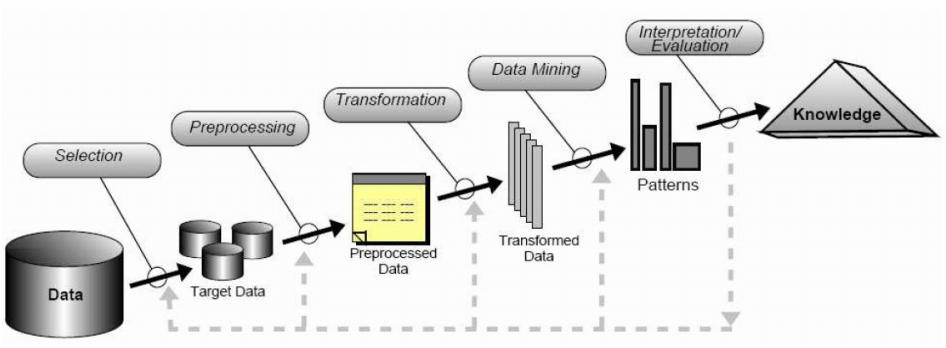


Ilustração 3 - Fases do processo de descoberta do conhecimento em bancos de dados [FPS96]

[7] Não trivial devido a complexidade envolvida no problema.

[8] Interativo por exigir a intervenção do usuário.

[9] Iterativo pelo fato de cada fase de todo o processo poder ser executada várias vezes.

3.2 Fases do KDD

3.2.1 Seleção dos Dados

Nesta fase é realizada a escolha das fontes de dados, podendo ser fontes internas ou externas da empresa ou mesmo fonte de conhecimento tácito dos especialistas. Estas fontes podem ser oriundas de arquiteturas de bancos de dados diversos, internet, planilhas e documentos originados de aplicativos diversos ou mesmo documentos manuscritos que deverão ser digitalizados.

3.2.2 Pré-processamento dos Dados

Nesta fase são realizados ajustes nos dados de forma a eliminar as inconsistências existentes. São tratados casos de erros de preenchimento dos campos, dados fora do padrão esperado e ausência dos mesmos. Essa fase corrige a base de dados evitando a tentativa de análise pelo algoritmo de mineração. Demanda cerca de oitenta por cento do esforço total no processo de KDD.

A utilização de um data warehouse facilita esta etapa por ser inerente a utilização de rotinas de extração, tratamento e carga dos dados o que leva a uma maior nível de consistência dos dados, porém um projeto de KDD não é inviabilizado pela ausência daquele.

Alguns exemplos de tratamento dos dados seriam:

- Valores nulos podem ser substituídos estatisticamente pela média, quando tratar-se de dados numéricos; substituídos por um rótulo ou pela moda, quando os dados são categóricos. Os valores nulos ou ausentes podem ainda ser preenchidos manualmente ou terem suas tuplas descartadas e por fim, pode-se optar por preenchimentos através de métodos de mineração de dados.

- Padronização de campos; datas ou substituição das datas por períodos (dias, meses, outros.).

- Valores fora dos padrões podem ser corrigidos através da exclusão de tuplas que possuem valores fora do domínio ou ainda, optar-se pela correção desses valores de forma manual ou via Structured Query Language (SQL).

- Amostras em excesso; considerar um maior nível de granularidade.

A seleção dos atributos mais relevantes para o objetivo da mineração também faz parte do escopo desta etapa. Esta tarefa tem objetivo de otimizar o tempo de processamento do algoritmo de mineração dos dados. Dois métodos principais, além do conhecimento do especialista no domínio, são utilizados para seleção de atributos [AVL99]:

- Múltiplas iterações: utiliza os métodos de Forward Selection e Backward Elimination. O primeiro deles é iniciado com um conjunto vazio de atributos e de forma iterativa seleciona sucessivamente um atributo por vez, utilizando como limiar a melhora da medida da qualidade do resultado.

- Iteração simples: utiliza algoritmos com fundamentos estatísticos de probabilidades de distribuição.

A fase de pré-processamento sugere ainda o "enriquecimento" dos dados. Trata-se da criação de novas informações com base no conhecimento dos especialistas no domínio e dos dados, de forma a melhorar a qualidade das informações já existentes, possivelmente, aumentando o desempenho do algoritmo de mineração.

3.2.3 Transformação

Nesta fase é realizada a transformação dos dados, visando à necessidade do algoritmo a ser utilizado na fase de mineração. No caso de uma Rede Neural Artificial, por exemplo, os dados necessitam estar em uma representação numérica.

A transformação dos dados pode ser feita de várias formas:

- Transformação Direta – é a substituição direta de valores numéricos por categóricos. Exemplo variável sexo:

  0 -> M

  1 -> F

- Discretização – é o mapeamento contínuo em intervalos; divisão de uma variável numérica, em intervalos pré-definidos com ajuda do especialista no domínio. Exemplo variável idade:

  0 à 18 anos -> FaixaEtaria1

  19 à 26 anos -> FaixaEtaria2

  27 à 45 anos -> FaixaEtaria3

Observa-se uma imprescindível participação do especialista no domínio, definindo o critério para criação das classes.

- Binarização – é a transformação de valores categóricos em valores representados por 0 ou 1. Exemplo variável estado civil:

  Casado -> 001

  Solteiro -> 010

- Normalização – é a transformação dos valores de cada atributo de forma a gerar novos valores em intervalos de [-1;1] ou [0;1].

3.2.4 Mineração de Dados

Pode ser considerada a fase mais importante do processo de KDD. É a fase em que, de fato, acontece a transformação dos dados em conhecimento a ser utilizado pelo especialista no domínio da aplicação. Essa transformação se dá pela escolha apropriada do algoritmo que será capaz de extrair de forma eficiente, conhecimento escondido na base de dados para a tarefa proposta. O Quadro 1 apresenta algumas tarefas possíveis com KDD e as respectivas técnicas.

Tarefa de KDD	Técnicas de Mineração de Dados
Associação	Estatística e Teoria dos Conjuntos
Classificação	Estatística, Algoritmos Genéticos, Redes Neurais e Árvore de Decisão
Agrupamento de Padrões	Redes Neurais e Estatística
Previsão de Séries Temporais	Redes Neurais, Lógica Nebulosa e Estatística

Quadro 1 - Tarefas em KDD e técnicas utilizadas

Algumas dessas técnicas, por serem as mais utilizadas são descritas a seguir:

3.2.5 Interpretação dos Resultados

Nesta fase é realizada a tarefa de interpretar os resultados da mineração de dados para o problema proposto.

Padrões são mostrados como resultados. Esses resultados devem ser aferidos e validados junto ao especialista no domínio da aplicação de forma a verificar a viabilidade do modelo construído.

4 Redes Neurais Artificiais

Neste capítulo são abordados os principais conceitos que envolvem a utilização de redes neurais artificiais.

4.1 Introdução

Os primeiros estudos sobre redes neurais se deram na década de 40, quando alguns pesquisadores contribuíram de forma significativa para evolução e uso desse conceito:

- McCulloch e Pitts em 1943 [McP43], quando sugeriram que as redes neurais seriam máquinas computacionais.
- Hebb em 1949, quando definiu a primeira regra de aprendizagem auto-organizada.
- Rosenblatt em 1958, quando propôs o primeiro modelo de aprendizagem supervisionada com o perceptron[10].

Segundo Haykin [Hay99], uma rede neural é um processador distribuído paralelamente e constituído por unidades de processamento simples, que possuem propensão natural para armazenar conhecimento extraído e torná-lo disponível para uso, sendo semelhante ao cérebro humano em dois aspectos:

- O conhecimento é adquirido pela rede a partir do seu ambiente através de um processo de aprendizagem. Nesse processo, um conjunto de exemplos é apresentado à rede, que obtém as características para representação da informação fornecida que serão utilizadas para gerar respostas ao problema.
- Pesos sinápticos são utilizados para armazenamento dos conhecimentos adquiridos.

[10] É a forma mais simples de uma rede neural usada para a classificação de padrões linearmente separáveis (padrões que estão em lados opostos de um hiperplano).

As Redes Neurais Artificiais (RNAs), são sistemas paralelos distribuídos compostos por unidades de processamento simples, que são os neurônios artificiais e que por sua vez, calculam determinadas funções matemáticas. Esses neurônios podem estar dispostos em uma ou mais camadas e interligados por conexões que estão associadas a pesos; esses pesos armazenam conhecimento "percebido" pelo modelo e ponderam a entrada recebida por cada um dos diversos neurônios da RNA [BCL07]. Pode ser agrupada em camadas de entrada, intermediárias (ou escondidas) e de saída.

Os neurônios (artificiais) e suas sinapses são chamados de Perceptron. Cada neurônio realiza a média ponderada dos sinais recebidos na entrada, sendo que os pesos ficam armazenados nas sinapses e o resultado desta ponderação, utilizado como entrada para função de ativação [Cor02]. Por último, a saída desta função de ativação é a resposta (resultado) do neurônio (Ilustração 4).

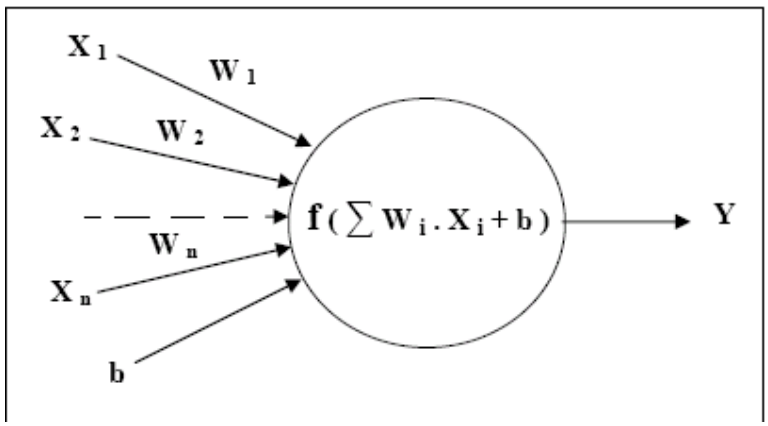


Ilustração 4 - Neurônio artificial de um Perceptron (Excerto [BCL07]).

A capacidade das Redes Neurais Artificiais em aprender e generalizar[11] a informação aprendida, de atuar como "mapeadores universais de funções multivariáveis, com custo computacional que cresce apenas linearmente com o número de variáveis" [BCL07] foram os motivos para sua escolha no problema de mineração de dados do PBEM.

[11] Capacidade de generalizar, segundo Haykin (1999), se refere a capacidade da Rede Neural Artificial produzir saídas adequadas para entradas que não estavam presentes durante o treinamento.

4.2 Processo de Aprendizagem

Uma característica importante das RNAs é a capacidade de aprendizado através de amostras. O procedimento utilizado para realizar o processo de aprendizagem é chamado de algoritmo de aprendizagem, função essa que é a de modificar os pesos sinápticos da rede de forma ordenada para atingir um objetivo desejado [Hay99], ou seja, busca pelo melhor desempenho da rede segundo critérios estabelecidos anteriormente.

O aprendizado é definido como um processo pelo qual os parâmetros livres de uma rede neural são ajustados por meio de uma forma continuada de estimulo pelo ambiente externo, sendo o tipo específico de aprendizado definido pela maneira particular como ocorrem os ajustes dos parâmetros livres [MM70].

Esta definição do processo de aprendizagem implica a seguinte seqüência de eventos [Hay99]:

1. A rede neural é estimulada por um ambiente.
2. A rede neural sofre modificações em seus parâmetros livres como resultado desta simulação.
3. A rede neural responde de uma forma nova para o ambiente, devido às mudanças que ocorreram em sua estrutura interna.

Existe uma variedade de algoritmos de aprendizagem de uma rede neural, cada um oferece vantagens próprias. Os algoritmos de aprendizagem diferem uns dos outros no modo em que a adaptação a um peso sináptico de um neurônio é formulado. Outro fator a ser considerado é a maneira pela qual uma rede neural, composta por um conjunto de neurônios interligados, diz respeito ao seu ambiente.

Cinco regras de aprendizagem são conhecidas [Hay99], a saber:

- Aprendizagem por correção de erros;
- Aprendizagem baseada em memória;
- Aprendizado hebbiana;

- Aprendizagem competitiva e
- Aprendizagem Boltzman.

Todas são utilizadas para o processo de desenho das redes neurais. Pode-se dividi-las basicamente na forma de seu aprendizado como, supervisionado ou não-supervisionado que diz respeito a utilização ou não de um “professor” .

No estudo do aprendizado supervisionado, uma disposição essencial é um "professor" capaz de fornecer exatamente correções à rede, estimulando suas entradas por meio de padrões de entrada e observando a saída calculada pela mesma comparando com a saída desejada. Sendo a resposta da rede a função dos valores atuais dos pesos, estes são então ajustados de forma a aproximar a saída da rede com a saída que se deseja [Hay99]. A redução da diferença entre a saída da rede e a saída esperada, é realizada através de pequenos ajustes nos pesos a cada etapa do treinamento (Ilustração 5).

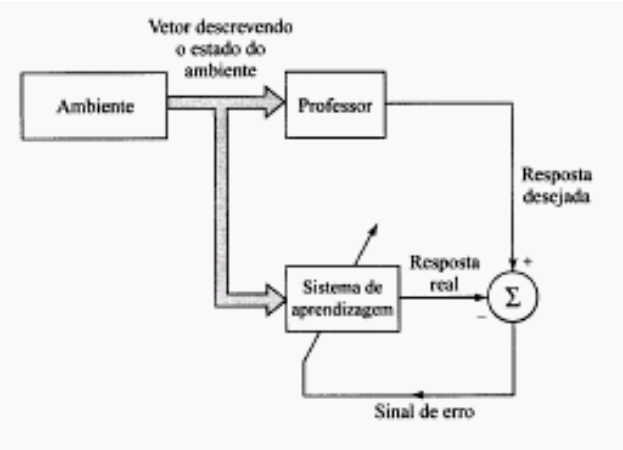


Ilustração 5 - Esquema do aprendizado supervisionado (Excerto [Hay99]).

No estudo do aprendizado não-supervisionado, não existe um “professor” (supervisor) para acompanhamento e correções no processo de aprendizagem, ou seja, não existem exemplos rotulados da função a ser aprendida pela rede neural.

O aprendizado não-supervisionado pode ser dividido em:

- Aprendizagem por reforço: o aprendizado é realizado através da intensa interação com o ambiente, com objetivo de minimizar um índice escalar de desempenho. Na aprendizagem por reforço, um crítico converte um sinal de reforço primário recebido do ambiente em sinal de reforço com melhor qualidade (sinal heurístico) (Ilustração 6).

Vantagem da aprendizagem por reforço – interação com o ambiente, desenvolvendo a capacidade de aprender, utilizando apenas os resultados de sua experiência.

Desvantagens da aprendizagem por reforço – não existe um professor que forneça uma resposta desejada em cada passo do processo de aprendizagem, o que otimizaria o resultado das etapas do treinamento. Outra desvantagem é que o atraso na geração do sinal de reforço primário, resulta que a máquina de aprendizagem deva decidir individualmente a cada ação na seqüência de passos de tempo que levam ao resultado final [Hay99].

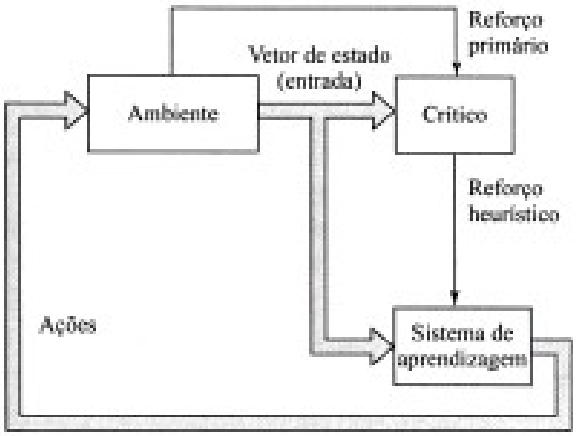


Ilustração 6 - Esquema do aprendizado por reforço (Excerto [Hay99]).

- Aprendizagem não-supervisionada (auto-organizada): nesse tipo de aprendizagem, não existe um "crítico" para supervisionar o aprendizado como no aprendizado por reforço. Os parâmetros livres da rede são otimizados em relação a uma medida independente da tarefa, definida

para o aprendizado da rede. Esta desenvolve habilidades de formar representações internas (Ilustração 7).

Ilustração 7 - Esquema do aprendizado não-supervisionado (Excerto [Hay99]).

A forma mais comum de utilização das RNAs é o aprendizado por meio de um conjunto de dados [BCL07]. Algumas tarefas básicas de aprendizagem para RNAs são [Hay99]:

- Associação de Padrões: tarefa que assume duas formas:
  - Auto-associação, onde uma rede neural armazena um conjunto de padrões que são apresentados repetidamente à rede. Logo em seguida é apresentado uma descrição distorcida do que foi armazenado originalmente. A tarefa é encontrar o padrão particular. Nessa forma é utilizada a aprendizagem não-supervisionada.
  - Heteroassociação, onde um conjunto arbitrário de padrões de entrada é associado a um outro conjunto arbitrário de padrões de saída. Nessa forma é utilizada a aprendizagem supervisionada.
- Reconhecimento de Padrões: processo pelo qual um padrão recebido é atribuído a uma classe dentre um número predeterminado de classes.
- Aproximação de Funções: processo pelo qual um padrão é apresentado às entradas e saídas da rede e para cada apresentação os pesos são adaptados de forma a mapear as relações de entrada e saída. Nessa tarefa é utilizada a aprendizagem supervisionada.

- Controle: processo pelo qual é mantida em condição controlada uma parte importante de um sistema.
- Filtragem: processo que utiliza-se de algoritmo com função de "*extrair informação sobre uma determinada grandeza de interesse a partir de um conjunto de dados ruidosos*".
- Formação de Feixe: processo utilizado para "*distinguir entre as propriedades espaciais de um sinal-alvo e o ruído de fundo*".

4.3 Redes Perceptron de Única Camada

Foi em 1958, através do trabalho de Frank Rosenblatt, que surgiu o conceito de aprendizado em redes neurais artificiais. Esse modelo proposto, era composto por uma estrutura de rede tendo unidades básicas de neurônios e por uma regra de aprendizado[Ros58]. Tornou-se conhecido como perceptron. Rosenblatt demonstrou que um *neurônio McCulloch e Pitts (MCP)*[12] treinado com o algoritmo de aprendizado do perceptron, sempre converge caso o problema em questão seja linearmente separável [Ros58].

O trabalho de Rosenblatt foi desacreditado, quanto à capacidade computacional do modelo, por causa das criticas feitas por Minsky e Papert na década de 70[MiP69]. Em 1982, Hopfield publicou trabalho sobre redes neurais e o algoritmo back-propagation, o que fez com que novos trabalhos utilizando redes neurais surgissem com novo impulso.

O perceptron, quando utilizado um neurônio, consegue apenas realizar classificação de padrões com somente duas classes. Se expandida a camada de saída do perceptron para incluir mais de um neurônio, pode-se realizar classificação de mais de duas classes, contudo, as classes devem ser linearmente separáveis (Ilustração 8) para

[12] MCP – Warren McCulloch e Walter Pitts, em 1943, foram os pioneiros na proposta de um modelo artificial de um neurônio biológico.

que o perceptron funcione corretamente [Hay99], o que em problemas do "*mundo-real*"[13] é menos provável de ocorrer.

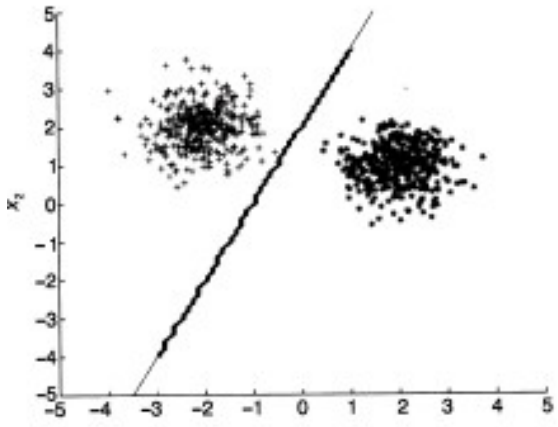


Ilustração 8 - Representação gráfica de classes linearmente separáveis (Excerto [Hay99]).

4.4 Redes Perceptron de Múltiplas Camadas

As redes neurais de múltiplas camadas, com camadas intermediárias formadas por neurônios com função de ativação sigmoidal[14] são denominadas *Multilayer Perceptron* (MLP). Diferentemente das redes com única camada, são capazes de resolver problemas de características não-lineares, através de funções de ativação de cada neurônio da rede e da composição de sua estrutura em camadas sucessivas [BCL07]. As MLPs de única camada intermedirária (escondida) são suficientes para aproximar qualquer função contínua [Cyb89].

As camadas em uma rede MLP são dispostas em seqüência, uma após a outra. A Ilustração 9 mostra o modelo de uma rede MPL típica com uma camada intermediária. O processamento realizado por cada neurônio de uma determinada camada é definido pela combinação dos processamentos realizados pelos neurônios da camada anterior que estão conectados a eles.

[13] Problema do mundo-real: diz respeito a situações reais que podem ser vistas como problemas ou oportunidades.

[14] A função de ativação mais frequente em redes MLP é a função sigmóide, dada por: y = 1 / (1+exp(x)).

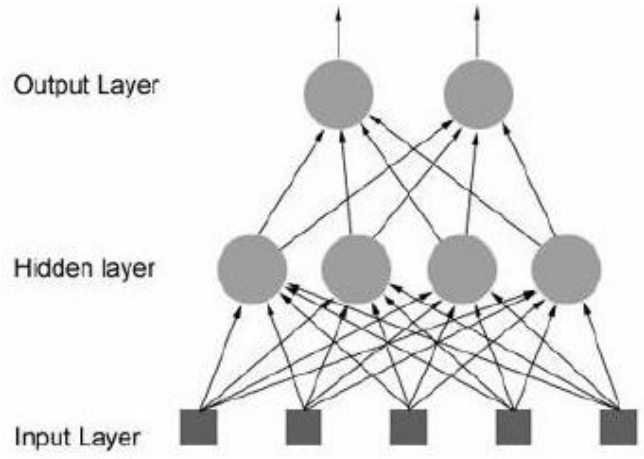


Ilustração 9 - Rede MLP com uma camada intermediaria (Excerto [Hay99])

O algoritmo mais utilizado para treinamento das RNAs é o Backpropagation [Cru07]. Por ser supervisionado, utiliza pares de entrada e saída (x, yd) para, por meio de um mecanismo de correção de erros, ajustar os pesos da rede. O treinamento ocorre em duas fases, em que cada fase percorre a rede em um sentido. Essas duas fases são chamadas de fase *forward* e *backward*. A fase *forward* é utilizada para definir a saída da rede para um dado padrão de entrada. A fase *backward* utiliza a saída desejada e a saída fornecida pela rede para atualizar os pesos de suas conexões [BCL07]. Ou seja, no algoritmo *Backpropagation*, em cada ciclo de aprendizagem, os erros obtidos pelas diferenças entre as saídas da rede e os valores de treino são propagados no sentido contrário (para trás), ocorrendo depois um ajuste dos pesos das conexões. O treino termina usualmente quando se obtém o mínimo erro à saída, no caso de regressão, ou o mínimo de classificações erradas [Cru07].

O algoritmo *Backpropagation* permite o ajuste automático dos pesos das sinapses.

O cálculo do erro para a saída de um neurônio é dado por:

ei (k) = yi (k) – di (k)

em que:

i-> neurônio

ei -> erro do neurónio i

yi -> saída do neurónio i

di -> saída desejada do neurónio i

k -> entrada em causa

O erro total é:

$$e(k) = \sum_{i=1}^{N} \frac{1}{2} e_i^2(k)$$

em que N é o número de neurônios.

Após o cálculo do erro, atualiza-se os pesos das sinapses [Fah98]:

$$\Delta w_{ij} = w_{ij}(k+1) + w_{ij}(k) = -\eta \frac{\partial e(k)}{\partial w_{ij}}$$

em que:

wij -> peso da sinapse ij;

η -> controle da aprendizagem (de 0 a 1)

Os pesos são então ajustados de acordo com a fórmula:

$$w_{ij}(k+1) = w_{ij}(k) - \eta . y_i(k) . e_j(k) ,$$

onde k indica a iteração atual do algoritmo. Ou seja, o peso de uma sinapse wij será resultado da diferença do peso anterior e do produto do fator de controle da aprendizagem pela saída do neurônio i (início da sinapse) e pelo erro do neurônio j (fim da sinapse). O fator de controle da aprendizagem deve aumentar ou diminuir o efeito do erro no peso da sinapse, que pode levar a uma convergência mais rápida ou mais demorada, conforme o seu valor for mais próximo de 1 ou de 0 [Cru07]. Então tem-se duas situações:

- Se a convergência for mais rápida, poderá cair num mínimo local ou nunca convergir devido à variação dos pesos.

- Uma busca mais lenta poderá levar ao mínimo global, porém demandará bastante tempo.

A escolha deste valor é fundamental, pois permitirá o controle direto da capacidade de generalização da rede [ZuG93]. Um novo termo é acrescido para um maior controle de processo; o termo momentum, onde gera um novo peso pela fórmula::

$$w_{ij}(k+1) = w_{ij}(k) - \eta.y_i(k).e_j(k) + \mu\Delta wij(k-1)$$

A diferença está na introdução de um termo que corresponde ao ajuste da iteração anterior aferida de um fator μ, chamado de inércia (momentum), que ajuda à convergência para um mínimo global, pois permite que parte do ajuste da iteração anterior vá refletir-se no ajuste atual [Cru07].

Um ponto importante a se observar é o número de iterações a serem executadas pela rede para aprendizagem. Um número excessivo de iterações pode levar ao *overfitting*[15].

Deve-se conseguir um equilíbrio entre a precisão e a generalização, para parada do treinamento, através de critérios, como [ZuG93]:

- Erro máximo;
- Gradiente de erro;
- Número de iterações;
- Validação cruzada,

[15] Situação em que a RNA se adapta muito bem aos casos de aprendizagem, mas responderá mal a outros.

5 Mineração de Dados PBEM – Entendimento do Negócio, Entendimento e Preparação dos Dados

Neste capítulo são abordadas as três primeiras etapas da metodologia selecionada para o projeto de mineração de dados do PBEM.

5.1 Introdução

A metodologia empregada neste trabalho foi a CRISP-DM (*Cross-Industry Standard Process for Data Mining*). Embora esta possua uma desvantagem no que diz respeito ao reuso do conhecimento em relação à metodologia DMLC[16] [WiH00] conforme mostrado no Quadro 2 [Cun05] com resumo das principais metodologias para projetos de mineração de dados, para o trabalho de mineração do Programa Bolsa Escola optou-se pela metodologia CRISP-DM pelo fato do reuso do conhecimento, nesse caso, não ser relevante e por ser a metodologia mais difundida no Mundo[17].

Ano	Metodologia	Iterativo	Interativo	Organizado em Fases	Processos Extras	Reuso Conhecimento	Visão Projeto	Levantamento Requisitos	Entendimento Negócio
1996	Fayyad	A	M	A	N	N	N	N	N
1996	CRISP-DM	A	M	A	N	N	A	M	A
1996	Brachman	A	M	A	M	N	N	M	M
1997	Klemettine	A	M	A	N	N	N	N	A
1998	Feldens	A	M	A	N	N	N	N	N
2003	DMLC	A	M	A	N	M	A	M	A

N= não aborda; M= menciona, mas sem detalhes; A= aborda em detalhes.

Quadro 2 - Comparativo das principais metodologias para projetos de mineração de dados (Retirado de [Cun05]).

[16] DMLC (Data Mining Life Cycle) é uma metodologia para projetos de mineração de dados, proposta com objetivo de trazer melhorias à metodologia CRISP-DM no que diz respeito ao reuso do conhecimento e documentação dos recursos envolvidos.

[17] Metodologia mais utilizada no mundo:.CRISP-DM- 51%. Fonte: Pesquisa da KDnuggets realizada em julho de 2002. Disponível em http://www.kdnuggets.com/polls/2002/methodology.htm. Acesso em março de 2009.

O CRISP-DM é uma metodologia não proprietária, desenvolvida em 1996 por um consórcio internacional de empresas, que está estruturada em torno de tarefas e objetivos para cada uma das fases de um projeto de mineração de dados [Cri96].

A metodologia é composta de 6 fases (Ilustração 10):

Entendimento do negócio: nesta fase é desenvolvida uma visão clara das necessidades a serem satisfeitas;

Entendimento dos dados: nesta fase são identificados os dados disponíveis e onde se encontram;

Preparação dos dados: nesta fase os dados devem ser preparados de acordo com o método a ser utilizado na etapa de modelagem;

Modelagem: nesta fase são utilizados algoritmos de inteligência artificial para o problema a ser tratado.

Avaliação: nesta fase é avaliado o resultado encontrado comparando-se com o objetivo do projeto de mineração.

Distribuição dos resultados: nessa fase é disponibilizado o resultado do projeto aos interessados.

Ilustração 10 - Fases dos Processos da Metodologia CRISP-DM.

As fases de entendimento do negócio, entendimento e preparação dos dados e, suas aplicações à mineração de dados do PBEM são descritas nas próximas sessões. As fases de modelagem e avaliação estão no Capítulo 6. A seção do entendimento dos dados teve seu conteúdo reduzido para evitar redundâncias, uma vez que, foi tema do Capítulo 2. A fase de distribuição não faz parte do escopo deste trabalho, limitando-se a entrega documental ao especialista no domínio.

5.2 Entendimento do Negócio

Para entendimento do negócio - funcionamento do PBEM - foram realizadas as atividades a seguir discorridas:

- Quatro visitas ao especialista no domínio;
- Troca de e-mail´s para dúvidas e esclarecimentos a cerca do problema;

- Acompanhamento em visitas realizadas pela equipe do Programa Bolsa Escola às famílias beneficiadas pelo Programa ou que poderão ser beneficiadas;
- Definição do sucesso do PBEM.

Destaque para a penúltima atividade em que foram realizadas duas visitas de acompanhamento do beneficiário e três visitas para possível inclusão de novos beneficiários.

Pela observação dos fatos *in loco*, conversa com os requerentes ao benefício e com os especialistas, pôde-se constatar que o benefício do PBEM é utilizado basicamente para 3 situações distintas:

- Para garantir uma situação mínima adequada para manter a criança na escola;
- Para sobrevivência da família. Um exemplo está em imagens, tiradas durante uma das visitas, de moradia onde residem seis pessoas (um adulto, um adolescente e quatro crianças) (Apêndice B).
- De forma indevida.

Como o objetivo deste trabalho é aperfeiçoar o processo decisório de concessão do benefício através da análise de dados dos requerentes "apropriados" (requerentes que atualmente recebem o benefício) e dos "afastados" (requerentes que em algum momento receberam o benefício, mas foram afastados por motivos diversos), o montante inicial de registros (instâncias) a serem analisados é 10.770 (dez mil, setecentos e setenta).

Conversas com a especialista no domínio do PBEM levaram à definição de uma nova classe alvo identificando os casos de sucesso do Programa. Segundo a especialista no domínio do PBEM o sucesso do programa é reconhecido como a continuidade do requerente no Programa, excluindo-se quaisquer afastamentos que não sejam os afastamentos considerados naturais, que são a chegada à idade de quatorze anos e doze meses do dependente e a obtenção de renda per capita maior que 1/3 do salário mínimo.

Dessa forma, para esse trabalho foi necessário identificar a causa do afastamento para informar à RNA quais os “bons” e “maus” participantes do Programa. Por este motivo foi criada variável alvo TIPO como será detalhado na seção 5.4.2.

O ponto de operação do sistema proposto é mostrado na Ilustração 11. O ponto de operação foi escolhido observando-se a não existência de dados a posteriori (para este trabalho diz respeito a próxima atividade após confirmação dos dados dos requerentes) e como ferramenta de apoio ao processo decisório do comitê avaliador.

5.3 Entendimento dos Dados

5.3.1 Fonte dos Dados

Para entendimento dos dados foram realizadas as seguintes atividades:

- Sete marcações de visitas aos especialistas dos dados, contudo apenas três bem sucedidas (concretizadas);
- Troca de e-mail´s para dúvidas e esclarecimentos a cerca dos dados;
- Pesquisa exploratória dos dados.

Para esta última tarefa, vale ressaltar a dificuldade em encontrar respostas através da pesquisa no banco de dados, uma vez que, existem muitas tabelas (entidades) inconsistentes (deixaram de ser atualizadas em um dado momento; são atualizadas, mas não retratam a realidade dos fatos ou mesmo nunca foram utilizadas), as descrições dos campos não permitem um entendimento de seu conteúdo e por fim a inexistência de um dicionário de dados para todas as entidades (só existe para entidade tbRequerente).

Os dados foram obtidos em setembro de 2008 contendo o registro de todas as famílias inscritas no PBEM até então – 60.451 (sessenta mil, quatrocentos e cinqüenta e um registros) famílias e ocupa espaço de 728 MB.

A base de dados do bolsa escola é composta de 184 tabelas relacionadas, sendo as principais a tbRequerente (possui dados do requerente) e a tbDependente (possui

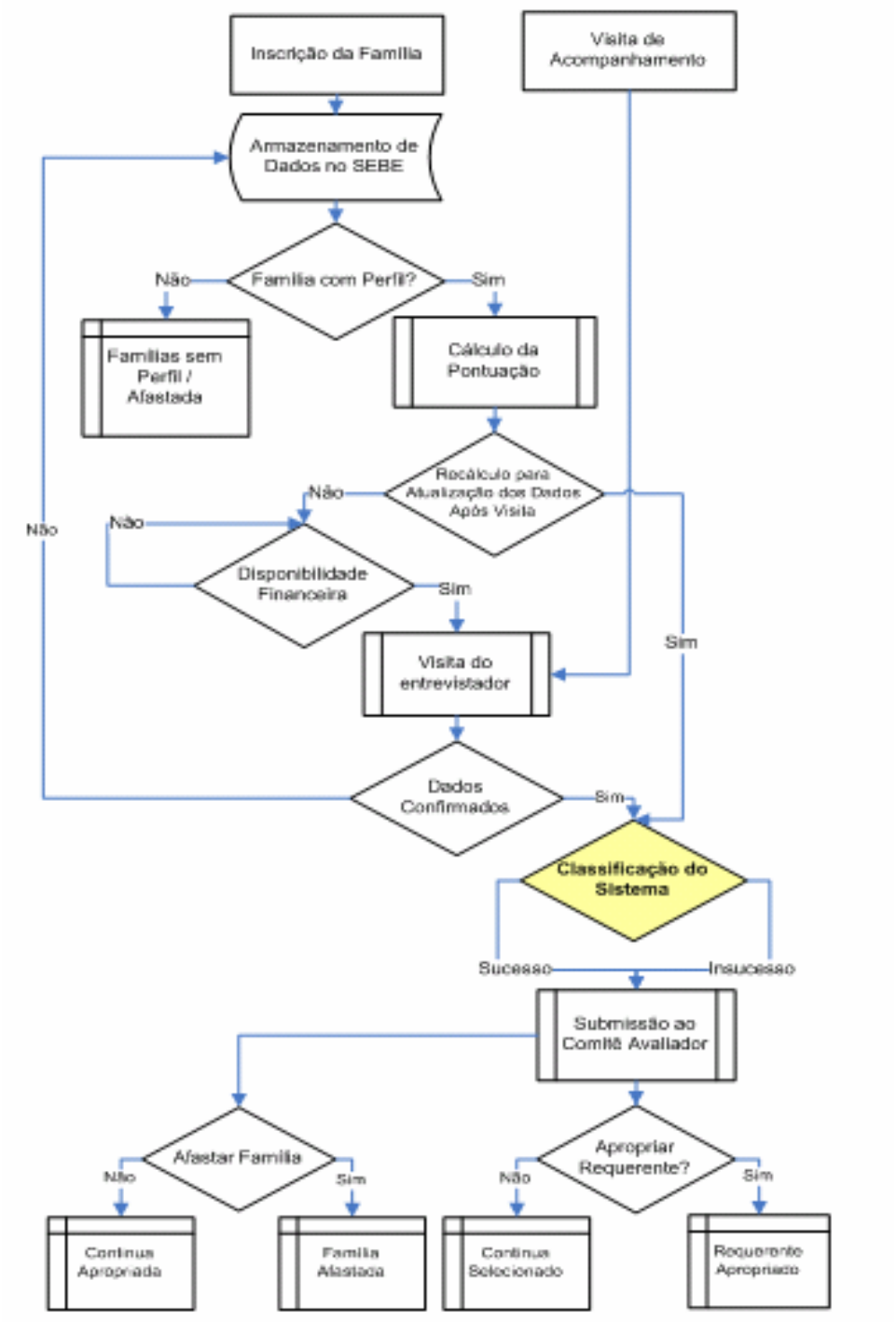


Ilustração 11 - Identificação do ponto de operação do sistema proposto.

dados dos dependentes de cada requerente). A primeira delas é a mais importante neste trabalho que tem nível de granularidade no “requerente”. O modelo entidade-relacionamento do banco de dados está no Apêndice C.

5.3.2 Base de Cadastro dos Requerentes

Essa base armazena os dados cadastrais dos requerentes ao PBEM fornecidos no momento da inscrição para o Programa na SEDE da Prefeitura do Recife. Esta entidade possui 102 atributos que armazenam informações diversas do requerente como, nome (nrequenome), sexo (crequesexo), data de nascimento (drequenasc), logradouro (erequeende), dados do cônjuge (nrequencon, crequescon, drequednco, cestadufco, drequechco, crequeidco, crequeorco, cestadufic, drequeemic, cgrauicodc, ccbotbcodc, ccbotbcoc2, csittrcodc, drequedesc, frequesinc e csitescodc) e dados econômicos sociais (drequecheg, cescivcodi, cgrauicodi, ccbotbcodi, ccbotbcodo, vrequerend, vrequeperc, crequeener, crequeagua, crequesani, arequenuco, arequememb, frequebfam e outras). O dicionário de dados da entidade tbRequerente está no Apêndice D.

Foi realizado levantamento, para classe alvo, sobre o preenchimento dos principais campos sinalizados pela especialista no domínio que influenciam na escolha do beneficiado.

Preenchimento dos dados quanto à existência de energia elétrica:

Crequeener (Energia Elétrica)		
Categoria	Qtde.	%
Sim	1.112	96%
Não	52	4%
Total	1.164	100%

Quadro 3 - Preenchimento dos dados quanto à existência de energia elétrica.

Preenchimento dos dados quanto à existência de água encanada:

Crequeagua (Água Encanada)		
Categoria	Qtde.	%
Sim	959	82%
Não	205	18%
Total	1.164	100%

Quadro 4 - Preenchimento dos dados quanto à existência de água encanada.

Preenchimento dos dados quanto à existência de saneamento básico:

Crequesani (Saneamento Básico)		
Categoria	Qtde.	%
Sim	876	75%
Não	288	25%
Total	1.164	100%

Quadro 5 - Preenchimento dos dados quanto à existência de saneamento básico.

Preenchimento dos dados quanto ao número de cômodos existentes na moradia:

Arequenuco (Número de Cômodos)		
Categoria	Qtde.	%
1	162	14%
2	216	19%
3	182	16%
4 ou mais	604	52%
Total	1.164	100%

Quadro 6 - Preenchimento dos dados quanto ao nº de cômodos da moradia.

Preenchimento dos dados quanto ao número de membros da família:

Arequememb (Número de Membros da Família)		
Categoria	Qtde.	%
Até 3	175	15%
4 à 7	797	68%
8 à 11	174	15%
12 ou mais	18	2%
Total	1.164	100%

Quadro 7 - Preenchimento dos dados quanto ao nº de membros da família.

Preenchimento dos dados quanto ao valor da renda da família:

Vrequerend (Valor da Renda da Família)		
Categoria	Qtde.	%
Até 207,5	612	53%
207,51 à 415	443	38%
415,1 ou mais	109	9%
Total	1.164	100%

Quadro 8 - Preenchimento dos dados quanto ao valor da renda da família.

Preenchimento dos dados quanto ao valor da renda per capita da família:

Vrequeperc (Valor da Renda per Capita)		
Categoria	Qtde.	%
Até 100	1.136	98%
101 à 200	26	2%
201 ou mais	2	0%
Total	1.164	100%

Quadro 9 - Preenchimento dos dados quanto ao valor da renda per capita da família.

Preenchimento dos dados quanto ao indicador de recebimento do benefício Bolsa-família do Governo Federal:

Frequebfam (Indicador do Bolsa-Família)		
Categoria	Qtde.	%
Recebe	960	82%
Não Recebe	204	18%
Total	1.164	100%

Quadro 10 - Preenchimento dos dados quanto ao indicador de recebimento do benefício Bolsa-família.

Quando considerada toda base dos requerentes beneficiados e ex-beneficiados pelo Programa (Quadro 11), observa-se um percentual muito abaixo do esperado, uma vez que o Programa Bolsa-Família do Governo Federal contempla em Recife cerca de 100.000 (cem mil) famílias[18] e o perfil sócio-econômico dos participantes de ambos os Programas são semelhantes. A especialista no domínio da aplicação concorda com esse fato e acredita que possa haver inconsistência na base de dados.

Frequebfam (Indicador do Bolsa-Família)		
Categoria	Qtde.	%
Recebe	2.588	24%
Não Recebe	8.182	76%
Total	10.770	100%

Quadro 11 - Preenchimento dos dados quanto ao indicador de recebimento do benefício Bolsa-família para beneficiados e ex-beneficiados.

Preenchimento dos dados quanto ao estado civil do requerente:

Cescivcodi (Estado Civil do Requerente)		
Categoria	Qtde.	%
Casado	136	12%
Divorciado	2	0%
Separado	102	9%
Solteiro	500	43%
União Simples	344	30%
Viúvo	80	7%
Total	1.164	100%

Quadro 12 - Preenchimento dos dados quanto ao estado civil do requerente.

[18] Dados oficiais do Governo Federal obtidos de http://www.mds.gov.br/adesao/mib/matrizview.asp?IBGE=2611606

Preenchimento dos dados quanto ao grau de instrução do requerente:

Cgrauicodi (Grau de Instrução do Requerente)		
Categoria	Qtde.	%
Analfabeto	236	20%
Ens. Fundam Incompl.	126	11%
Ens. Fundam Compl.	704	60%
Ens. Médio Incompl.	55	5%
Ens. Médio Compl.	40	3%
Superior Incompl.	1	0%
Superior Compl.	2	0%
Não Informado	-	0%
Total	1.164	100%

Quadro 13 - Preenchimento dos dados quanto ao grau de instrução do requerente.

Preenchimento dos dados quanto ao tipo de ocupação da moradia:

Ccbotbcodo (Ocupação da Moradia)		
Categoria	Qtde.	%
ALUGADO	226	19%
ARRENDADO	2	0%
CEDIDO	153	13%
FINANCIADO	1	0%
INVASÃO	238	20%
NAOTEM	1	0%
OUTRA	1	0%
PRÓPRIO	542	47%
Total	1.164	100%

Quadro 14 - Preenchimento dos dados quanto ao tipo de ocupação da moradia.

Preenchimento dos dados quanto ao tipo de vedação da moradia:

Ctpvedcodi (Tipo de Vedação)		
Categoria	Qtde.	%
ADOBE	5	0%
ALVENARIA	910	78%
MADEIRA	155	13%

MADEIRITE	41	4%
MATERIALAPROVEITADO	1	0%
NAOTEM	1	0%
OUTRO	1	0%
TAIPANAOREVESTIDA	44	4%
TAIPAREVESTIDA	6	1%
Total	1.164	100%

Quadro 15 - Preenchimento dos dados quanto ao tipo de vedação da moradia.

Preenchimento dos dados quanto ao tipo de piso da moradia:

Ctppiscodi (Tipo de Piso)		
Categoria	Qtde.	%
CERAMICA	68	6%
CONTRAPISO	361	0,31%
PALAFITAS	1	0%
TERRABATIDA	190	16%
TIJOLO/CIMENTO	544	47%
Total	1.164	100%

Quadro 16 - Preenchimento dos dados quanto ao tipo de piso da moradia.

Preenchimento dos dados quanto ao tipo de teto da moradia:

Ctpcobcodi (Tipo de Teto)		
Categoria	Qtde.	%
LAJE	242	21%
NAOTEM	5	0%
PLASTICO/LONA	169	15%
TELHACERAMICA	509	44%
ZINCO/BRASILIT	239	21%
Total	1.164	100%

Quadro 17 - Preenchimento dos dados quanto ao tipo de teto da moradia.

Preenchimento dos dados quanto ao tipo de gleba da terra:

Cgletecodi (Gleba da Terra -Porção de terra)		
Categoria	Qtde.	%
ARRENDATÁRIO	175	15%
POSSEIRO	270	23%
PROP. PARENTE	163	14%
PROP. PATRÃO	30	3%
PROPRIETÁRIO	526	45%
Total	1.164	100%

Quadro 18 - Preenchimento dos dados quanto ao tipo de gleba da moradia.

Preenchimento dos dados quanto ao tipo de construção da moradia:

Ctpconcodi (Tipo de Construção)		
Categoria	Qtde.	%
EM ACABAMENTO	322	28%
INICIADA	56	5%
OUTROS	69	6%
PARALISADA	283	24%
PRONTA	345	30%
SEM INFORMACAO	89	8%
Total	1.164	100%

Quadro 19 - Preenchimento dos dados quanto ao tipo de construção da moradia.

Preenchimento dos dados quanto à idade do requerente:

Crequeidad (Idade do Requerente)		
Categoria	Qtde.	%
ATÉ 30 ANOS	45	4%
31 À 50 ANOS	838	72%
51 À 70 ANOS	255	22%
71 OU MAIS	26	2%
Total	1.164	100%

Quadro 20 - Preenchimento dos dados quanto à idade do requerente.

Atributo criado a partir de atributo que contem a data de nascimento do requerente (seção 5.4.2).

Preenchimento dos dados quanto à idade dos dependentes:

Qreque0A15 (Número de Requerentes que Possuem Dependentes até 15 anos)		
Categoria	Qtde.	%
0	149	13%
1	400	34%
2	315	27%
3 ou mais	300	26%
Total	1.164	100%

Quadro 21 - Preenchimento dos dados quanto à idade dos dependentes para os requerentes que possuem dependentes até 15 anos.

Atributo criado (seção 5.4.2) a partir de atributo que contem a data de nascimento do dependente na entidade tbDependente.

Preenchimento dos dados quanto à idade dos dependentes:

QrequeAp16 (Número de Requerentes que Possuem Dependentes a partir de 16 anos)		
Categoria	Qtde.	%
0	203	17%
1	278	24%
2	288	25%
3 ou mais	395	34%
Total	1.164	100%

Quadro 22 - Preenchimento dos dados quanto à idade dos dependentes para os requerentes que possuem dependentes a partir de 16 anos.

Atributo criado (seção 5.4.2) a partir de atributo que contem a data de nascimento do dependente na entidade tbDependente.

Preenchimento dos dados quanto à idade dos dependentes:

nrequeIddc (Idade do Cônj. Do Requerente)		
Categoria	Qtde.	%
ATÉ 15 ANOS	1.015	51,36%
A PARTIR DOS 16 ANOS	961	48,63%
Total	1.976	100%

Quadro 23 – Quantidade de dependentes por faixa etária.

Preenchimento dos dados quanto à idade dos cônjuges dos requerentes:

nrequeIddc (Idade do Cônj. Do Requerente)		
Categoria	Qtde.	%
ATÉ 30 ANOS	20	2%
31 À 50 ANOS	1.021	88%
51 À 70 ANOS	110	9%
71 OU MAIS	13	1%
Total	1.164	100%

Quadro 24 - Preenchimento dos dados quanto à idade dos cônjuges dos requerentes.

Preenchimento dos dados quanto ao sexo do requerente:

Crequesexo (Sexo do Requerente)		
Categoria	Qtde.	%
F	997	86%
M	167	14%
Total	1.164	100%

Quadro 25 - Preenchimento dos dados quanto ao sexo do requerente

5.4 Preparação dos Dados

Por não existir um Data Warehouse para a base do bolsa escola e com isso rotinas para extração, tratamento e carga adequados dos dados, um elevado nível de

inconsistências e outliers da base foi encontrado, o que demandou bastante tempo para preparação dos dados à modelagem.

A tarefa inicial foi agrupar todos os dados sócio-econômicos do requerente em uma única estrutura desnormalizada que passou a conter 128 colunas. Para isso as seguintes atividades foram executadas:

- Substituição na tabela tbRequerente dos campos com códigos pelas respectivas descrições; em um data warehouse seria trazer as descrições das tabelas de dimensão para a tabela de fatos.
- Padronização das descrições de cada campo, retirando erros de digitação e variações desnecessárias. O tratamento de inconsistências foi realizado atribuindo-se a moda de cada atributo. O problema dos dados ausentes foi resolvido calculando-se o valor médio para cada atributo numérico de cada classe e substituindo os valores ausentes por esse valor médio calculado.
- Criação de campos com informações de forma a tornar mais eficiente o resultado da mineração;
- Eliminação das variáveis menos significativas;
- Preparação das variáveis para o algoritmo de mineração a ser utilizado para modelagem.

5.4.1 Filtragem e Limpeza dos Dados

Como mencionado na seção 5.2, para o objetivo do trabalho, 10.770 registros (instâncias) seriam utilizados na mineração de dados.

Inicialmente, oito registros foram excluídos como parte da limpeza dos dados. O motivo deve-se a inconsistência de dados, pois todos os códigos dos requerentes da tabela tbRequerente devem constar da tabela tbHistoricoDependente na qual constam os motivos dos afastamentos, dessa forma a base ficou com 10.762 registros. Tal observação foi comunicada aos especialistas dos dados.

Quanto às colunas (atributos), o primeiro passo foi eliminar as colunas com relação direta (código e descrição) e aquelas que identificavam unicamente o registro na tabela [CCK00]. Abaixo relação das colunas retiradas e respectivos motivos:

Coluna	Descrição	Motivo da Retirada
CESCOLCODI	Código da Escola	Foi retirado, uma vez que, optou-se pela análise por Bairros
CBAIRRCODI	Código do Bairro	Foi deixado campo com descrição dos Bairros
CNACIOCODI	Código da Nacionalidade	Foi retirado devido a alta insidência de uma única classe
CPROPFCODI	Código do Proprietário do Fone	Foi retirado devido ao baixo grau de confiança da informação, conforme conversado com a especialista no domínio
CESCIVCODI	Código do Estado Civil do Requerente	Foi deixado campo com descrição do estado civil
CGRAUICODI	Código do Grau de Instrução do Requerente	Foi deixado campo com descrição do grau de instrução
CCBOTBCODI	Código da Profissão	Foi deixado campo com descrição da profissão
CCBOTBCODO	Código da Ocupação	Foi deixado campo com descrição do tipo de ocupação
CSITTRCODI	Código da Situação do trabalho	Foi deixado campo com descrição da situação do trabalho
CSITESCODI	Código da Situação Especial	Foi retirado devido a baixa ocorrência (1% da base)
CGRAUICODC	Código do Grau de Instrução do cônjuge do Requerente	Foi deixado campo com descrição do grau de instrução do conjuge do requerente
CCBOTBCODC	Código da Profissão da cônjuge	Foi deixado campo com descrição do profissao do cônjuge
CCBOTBCOC2	Código da Ocupação do cônjuge	Foi deixado campo com descrição da ocupção do cônjuge
CSITTRCODC	Código da Situação no Mercado de Trabalho do Cônjuge	Foi retirado pela redundancia com a descrição da ocupação
CSITESCODC	Código da Situação Especial do Cônjuge	Foi retirado devido a baixa ocorrência (1,8% da base)
COCUMOCODI	Código da Ocupação da Moradia	Foi deixado campo com descrição do tipo de ocupação da moradia
CTPVEDCODI	Código do Tipo de Vedação	Foi deixado campo com descrição do tipo de vedação
CTPPISCODI	Código do Tipo de Piso	Foi deixado campo com descrição do tipo de piso
CTPCOBCODI	Código do Tipo de Cobertura	Foi deixado campo com descrição do tipo de cobertura
CGLETECODI	Código da Gleba da Terra	Foi retirado devido a redundância com a descrição do tipo de ocupação

CTPCONCODI	Código do Tipo da Construção	Foi deixado campo com descrição do tipo de construção
CSTATUCODI	Código do Status do Requerente	Foi retirado por conter dados a posteriori
CORIGECODI	Código da origem da inscrição	Foi retirado , uma vez que, segundo a especialista no domínio, não é relevante
AREQUECTPS	Número da Carteira Profissional	Não relevante para o problema
AREQUEFOLH	Número da Folha da CN	Não relevante para o problema
AREQUEFONE	Número do Telefone de Contato	Não relevante para o problema
AREQUEISOC	Número da Inscrição Social (PIS/PASEP/SUS)	Não relevante para o problema
AREQUENCPF	Número do CPF	Foi retirado por identificar unicamente um registro na tabela
AREQUENLIV	Número do Livro da CN	Não relevante para o problema
AREQUESECP	Série da CP	Não relevante para o problema
AREQUETECN	Número do Termo da Certidão de Nascimento	Não relevante para o problema
CESTADCOCP	Sem Informação	Foi retirado , uma vez que, 100% dos registros são de PE e N/A(não avaliado - não preenchido)
CESTADCODC	Estado Emissor da CN	Não relevante para o problema
CESTADCODI	Estado Expedidor da CI	Foi retirado , uma vez que, 98% dos registros são de PE e N/A(não avaliado - não preenchido)
CESTADCODM	UF de Moradia	Foi retirado , uma vez que, 100% dos registros são de PE
CMUNICICEO	Código do Município de Origem	Preenchimento não padronizado
CMUNICICEP	Código do Município de Residência	Foi retirado , uma vez que, 100% dos registros são de Recife
crequecodi	Código do Requerente	Foi retirado por identificar unicamente um registro na tabela
CREQUEDDDF	DDD do Telefone	Foi retirado , uma vez que, 100% dos registros são de PE e N/A(não avaliado - não preenchido)
CREQUEIDCO	Número da Carteira de Identidade do Cônjuge	Foi retirado por ser código único, além de 68% serem N/A
CREQUEINSC	Inscrição Antiga do Requerente	Foi retirado por identificar unicamente um registro na tabela
CREQUENUID	Número da Carteira de Identidade	Foi retirado por identificar unicamente um registro na tabela
CREQUENUME	Número da Moradia	Preenchimento não padronizado
CREQUEPONT	Pontuação	Foi retirado por conter dados a posteriori
CREQUETESC	Tipo da Escola (E-Estadual, M-Municipal)	Foi retirado , uma vez que, 100% dos são escolas municipais do Recife

CREQUETIPO		Foi retirado devido ao alto nº de não preenchimento (60%)
CREQUEZONA	Zona Residencial	Foi retirado , uma vez que, 100% dos dados pertencem a uma mesma classe
CTPDOCCODI		Foi retirado , uma vez que, 100% dos dados pertencem a uma mesma classe
DREQUEANFA	Ano de Fabricação do Veículo	Foi retirado , uma vez que, 100% dos dados pertencem a uma mesma classe
DREQUEAPRO	Data utilizada para procedimento interno do sistema	Não relevante para o problema
DREQUECHCO	Data de Chegada do Cônjuge	Foi retirado , uma vez que, 100% dos dados pertencem a uma mesma classe
DREQUECHEG	Data de Chegada	Não relevante para o problema
DREQUECTPS	Data de Emissão CP	Foi retirado , uma vez que, 100% dos dados pertencem a uma mesma classe
DREQUEDESC	Data de Desemprego do Cônjuge	Foi retirado , uma vez que, 100% dos dados pertencem a uma mesma classe
DREQUEDESE	Data do Desemprego	Foi retirado , uma vez que, 100% dos dados pertencem a uma mesma classe
DREQUEDNCO	Data de Nascimento do Cônjuge	Foi retirado porém, antes foi utilizado para gerar nova variável - idade do cônjuge
DREQUEEMCN	Data de Emissão da CN	Não relevante para o problema
DREQUEEMIC	Data de Emissão da Ci do Cônjuge	Não relevante para o problema
DREQUEEMII	Data de Emissão da CI	Foi retirado , uma vez que, 100% dos dados pertencem a uma mesma classe
DREQUEEXCE		Foi retirado , uma vez que, 100% dos dados pertencem a uma mesma classe
DREQUEEXPO	Data utilizada para procedimento interno do sistema	
DREQUEINPG		Foi retirado , uma vez que, 100% dos dados pertencem a uma mesma classe
DREQUEINSC	Data de Inscrição	Não relevante para o problema
DREQUENASC	Data de Nascimento	Foi retirado porém, antes foi utilizado para gerar nova variável - idade do requerente
DREQUEPONT	Data utilizada para procedimento interno do sistema	Não relevante para o problema
DREQUESELE	Data da seleção do requerente	Não relevante para o problema
DREQUETRAN		Foi retirado , uma vez que, 100% dos dados pertencem a uma mesma classe
EREQUECOMP	Complemento do endereço	Preenchimento não padronizado
EREQUEENDE	Logradouro	Preenchimento não padronizado
EREQUEINFC	Informações Complementares do Requerente	Preenchimento não padronizado
EREQUENCEP	Número do CEP	Optou-se pela análise por Bairros

EREQUEREFE	Ponto de Referência	Preenchimento não padronizado
ESITESDSCC	Descrição de Situação especial do cônjuge	Foi retirado , uma vez que, 100% dos dados pertencem a uma mesma classe
ESITTRDSCC	Descrição da Situação de Trabalho do Cônjuge	Foi retirado , uma vez que, 80% dos dados não preenchidos.
ESTATUDESC	Descrição da Situação do Requerente no Programa	Foi retirado por conter dados a posteriori
FREQUEPONT	Flag para requerentes pontuados no Programa	Foi retirado , uma vez que, 100% dos dados pertencem a uma mesma classe
FREQUEVEIC	Indicador de Veículo (C/M)	Foi retirado , uma vez que, 100% dos dados pertencem a uma mesma classe
FREQUEVISI	Flag de visita do entrevistador	Dado não confiável
motivoAfasta	Motivo do afastamento do requerente	Foi retirado porém, antes foi utilizado para gerar nova variável - tipo do afastamento
NREQUECART	Nome do Cartório da CN	Não relevante para o problema
NREQUEENTR	Nome do Entrevistador	Não relevante para o problema
NREQUENCON	Nome do Cônjuge	Não relevante para o problema
NREQUENOME	Nome do Requerente	Não relevante para o problema
NREQUENOMM	Nome da Mãe	Não relevante para o problema
NREQUENOMP	Nome do Pai	Não relevante para o problema
TREQUEULAT	Data utilizada para procedimento interno do sistema	Não relevante para o problema
VREQUEVAVE	Valor da Renda Per-capta	Foi retirado , uma vez que, 100% dos dados não preenchidos.

Quadro 26 - Relação de colunas retiradas da base e respectivos motivos.

5.4.2 Criação de Campos

Foram criados novos atributos durante o processo de agregação dos dados, com objetivo de ganhar informação para modelagem na RNA.

> CREQUEIDAD: Idade do requerente. Criado a partir do atributo DREQUENASC (data de nascimento).

> QREQUEQTDP: Quantidade de Dependentes do Requerente.

> QREQUE0A15: Quantidade de Dependentes até 15 anos de idade

> QREQUEAP16: Quantidade de Dependentes a partir de 16 anos de idade

> NREQUEIDDC: Idade do Cônjuge do Requerente

> TIPO: Tipo do requente. Variável Alvo. Criada com objetivo de marcar o requerente como "Sucesso" ou "Insucesso" (classe alvo), segundo definição da especialista no domínio. 89,18% das instâncias foram classificadas como "Sucesso" e 10,82% como "Insucesso" no PBEM.

5.4.3 Seleção de Variáveis Mais Relevantes

Após eliminação de oitenta e sete variáveis e criação de outras seis, com objetivo de confirmar a relevância de todas as quarenta e seis variáveis (atributos) restantes (sendo dezesseis numéricas e trinta categóricas), foram utilizadas técnicas para avaliar a importância de cada uma. Todas as técnicas tiveram como referência a variável alvo. As três técnicas utilizadas, além do conhecimento tácito[19] do especialista do PBEM, são descritas nas próximas sessões.

- **Curvas ROC**

Gráficos ROC (*Receiver Operating Characteristics*) constituem uma ferramenta útil para a visualização e avaliação de modelos de classificação. Eles também são utilizados para se avaliar como um sistema de aprendizado é capaz de ordenar os exemplos, permitindo uma análise independente do limiar de classificação. A análise ROC provê uma avaliação mais rica do que simplesmente avaliar o modelo de classificação a partir de uma única medida [RP05]. Também conhecido como curvas ROC [ProF98], representam os falsos positivos ou *fp rate* (no eixo das abscissas) e verdadeiros positivos ou *tp rate* (no eixo das ordenadas). Mostram a relação das taxas de falsos positivos (FP) e verdadeiros positivos (VP) através da variação de um limiar. Esta relação prediz o comportamento dos classificadores independentemente dos custos

[19] Conhecimento tácito: conhecimento subjetivo e inerente às habilidades de uma pessoa.

e da distribuição das classes. Para cada ponto de corte, a sensibilidade e o complemento da especificidade (1 – especificidade) são calculados e colocados em cada eixo respectivo de um gráfico bidimensional [FAW05]. Os gráficos ROC foram também utilizados para avaliação de desempenho do modelo experimentado (seção 6.2.2). Exemplos de gráficos ROC de duas variáveis são mostrados nos gráficos 2 e 3.

Gráfico 2 - Curva ROC da variável CREQUEIDAD (idade do requerente).

Para a variável contínua que indica a idade do requerente, CREQUEIDAD, a Curva ROC demonstra ser este um bom classificador por se distanciar da linha diagonal, que representa um classificador aleatório.

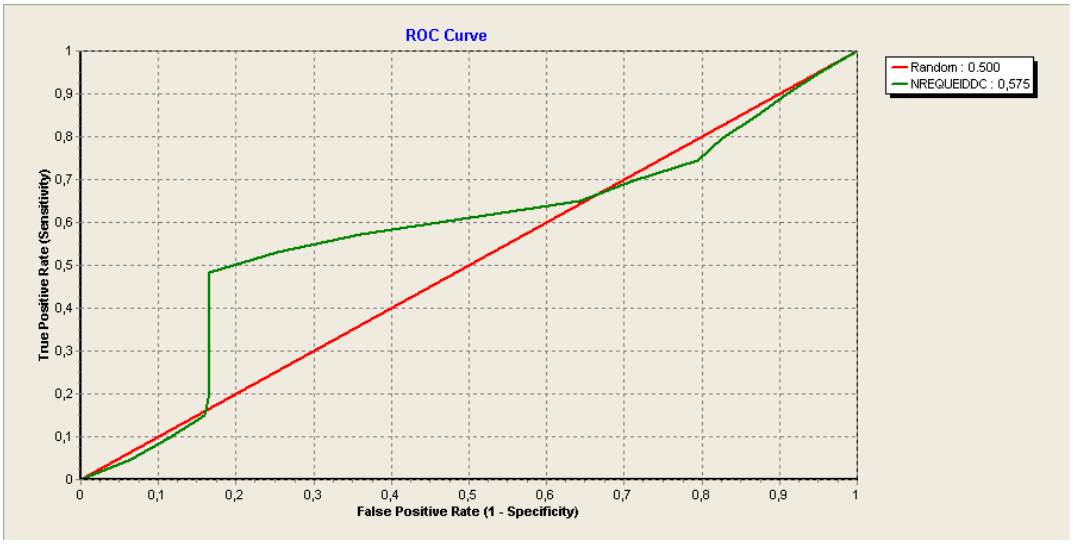


Gráfico 3 - Curva ROC da variável NREQUEIDDC (Idade do cônjuge do requerente).

Para a variável contínua que indica a idade do cônjuge do requerente, NREQUEIDDC, a Curva ROC também demonstra ser este um bom classificador, exceto em *fp rate* < 0,17 e *fp rate* > 0,7 com momentos de classificações piores (curva abaixo da diagonal) que as de classificadores aleatórios, ou seja, com poucas evidências fazem classificações negativas cometendo erros *false positive*.

O ranking das principais variáveis para curva ROC é:

Nrequeiddc (idade do cônjuge do requerente) = 0,575

Crequeidad (idade do requerente) = 0,563

Vrequerend (valor da renda da família) = 0,533

Arequememb (número de membros da família) = 0,519

QrequeA15 (quantidade de dependentes até 15 anos de idade) = 0,495

QrequeAP16 (quantidade de dependentes a partir dos 16 anos de idade) = 0,485

- **Ganho de Informação**

O Ganho de Informação ou *Information Gain Attribute Ranking* é uma técnica utilizada para seleção de variáveis que considera cada variável individualmente, ordenando-as com base em suas capacidades preditivas. Nessa técnica, cada atributo Ai (sendo A um atributo de um conjunto de dados) da base de dados é associado a um valor correspondente ao ganho de informação, na forma [Mer07]:

$$GI_i = E(C) - E(C|A_i). \quad \text{Onde:}$$

GI – Ganho de Informação;

E(C) – Cálculo da entropia do atributo classe antes de observado o atributo A;

E(C|A) - Cálculo da entropia do atributo classe após observado o atributo A.

Sendo:

$$E(C) = -\sum_{c \in C} p(c) log_2 p(c),$$

(retirado de [Mer07])

onde *p(c)* é a probabilidade da ocorrência da classe *C* na base de dados e

$$E(C|A) = -\sum_{a \in A} p(a) \sum_{c \in C} p(c|a) log_2 p(c|a),$$

(retirado de [Mer07])

sendo *p(a)* a probabilidade de *a* ocorrer na base de dados; *p(c|a)* a probabilidade da classe *c* ocorrer após ocorrência do valor do atributo *a*.

Antes de submeter os dados à técnica, foi necessário converter os campos categóricos em números binários, transformação essa que foi utilizada também para fase

que antecede a modelagem, conforme descrito na seção 5.4.5. O resultado do ranking das quarenta e cinco variáveis (a variável TIPO é a variável alvo) utilizando a técnica *Information Gain Attribute Ranking* é mostrado no Quadro 27 - Resultado do Information Gain Attribute Ranking às variáveis.

Ordem	Pontuação	Variável	Descrição
1	0.022397	CREQUEIDAD	Idade do requerente
2	0.008351	CREQUESEXO	Sexo do requerente
3	0.004378	EORIGEDESC	Descrição da origem do requerente
4	0.003544	QREQUE0A15	Quantidade de dependentes até 15 anos de idade.
5	0.003209	ECBOTBDSCC	Ocupação do cônjuge
6	0.002715	CREQUESCON	Sexo do cônjuge
7	0.002712	FREQUEBFAM	Beneficiado pelo bolsa-família
8	0.002676	ECBOTBDESC	Descrição da profissão
9	0.002263	EBAIRRNESC	Bairro da escola
10	0.00191	EBAIRRNOME	Bairro do requerente
11	0.001877	ESITTRDESC	Descrição da situação de trabalho
12	0.001859	ECBOTBDSCO	Descrição da ocupação
13	0.001646	EGRAUIDESC	Grau de instrução do requerente
14	0.001452	AREQUEMEMB	Número de membros da família
15	0.001451	QREQUEAP16	Quantidade de dependentes a partir de 16 anos.
16	0.001215	VREQUEREND	Valor da renda da família
17	0.001125	NREQUEIDDC	Idade do cônjuge do requerente
18	0.001116	CREQUEORCO	Órgão expedidor da CI(carteira de identidade) do Cônjuge
19	0.000934	NESCIVNOME	Descrição do estado civil
20	0.000917	CESTADUFIC	Estado emissor da CI do cônjuge
21	0.000917	NESTADNMIC	Estado emissor da CI do requerente
22	0.000906	ESITESDESC	Descrição de situação especial
23	0.000888	CREQUEORGA	Órgão da CI
24	0.000883	NESTADNOMC	UF do proprietário
25	0.000799	EGRAUIDSCC	Grau de instrução do cônjuge
26	0.000794	EPROPFDESC	Proprietário do fone
27	0.000695	ETPVEDDESC	Tipo de vedação
28	0.000574	CTPLGDCODI	Tipo de logradouro
29	0.000563	CESTADCODN	UF de moradia
30	0.000471	NESTADNMCO	Estado do cônjuge
31	0.000471	CESTADUFCO	Estado de origem do cônjuge
32	0.000453	EOCUMODESC	Ocupação da moradia
33	0.000371	CREQUEFECH	Tipo de ocupação da moradia requerente

Ordem	Pontuação	Variável	Descrição
34	0.000329	ETPPISDESC	Tipo do piso
35	0.000304	ETPCOBDESC	Tipo da cobertura
36	0.000303	EGLETEDESC	Gleba da terra
37	0.000197	FREQUESINE	Indicador de inscrição no SINE
38	0.000158	ETPCONDESC	Situação da construção da moradia
39	0.000143	CREQUEAGUA	Indicador de água encanada
40	0.000126	CREQUEENER	Indicador de energia elétrica
41	0.000119	FREQUESINC	Indicador de Inscrição no SINE do Cônjuge
42	0.000103	CREQUESANI	Indicador de saneamento básico
43	0	VREQUEPERC	Valor de renda per capta da família
44	0	AREQUENUCO	Número de cômodos
45	0	QREQUEQTDP	Quantidade de dependentes do requerente

Quadro 27 - Resultado do Information Gain Attribute Ranking às variáveis.

- **Coeficiente de Pearson**

O coeficiente de correlação de Pearson mede o grau da correlação entre duas variáveis e a direção dessa correlação, podendo ser positiva ou negativa. Assume valores entre -1 e 1. Sendo ***r*** o coeficiente de Pearson, quando ***r = 1*** significa que existe uma correlação positiva perfeita entre as duas variáveis; quando ***r = -1*** significa que existe uma correlação negativa perfeita, o que quer dizer que quando uma variável aumenta a outra diminui na mesma proporção. Por outro lado, quanto mais próximo o ***r*** for de zero, menos dependentes linearmente são as duas variáveis [HaK06].

O cálculo do coeficiente de correlação de Pearson é realizado da forma abaixo apresentada:

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2} \cdot \sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}},$$

Sendo $x_1, x_2, \ldots, x_n$e $y_1, y_2, \ldots, y_n$ valores medidos de ambas as variáveis e

$$\bar{x} = \frac{1}{n} \cdot \sum_{i=1}^{n} x_i \quad \text{e} \quad \bar{y} = \frac{1}{n} \cdot \sum_{i=1}^{n} y_i$$

sendo as médias aritméticas de ambas as variáveis.

A análise pelo método de correlação de Pearson foi realizada para as variáveis numéricas, sendo que, algumas das variáveis foram binarizadas (crequesexo - sexo do requerente, frequesine – indicador de inscrição no SINE, crequeener – indicador de energia elétrica, crequeagua – indicador de água encanada, crequesani – indicador de saneamento básico) ou discretizadas (egrauidesc – grau de instrução do requerente, egrauidscc – grau de instrução do cônjuge) antes de serem submetidas à análise . O apêndice E apresenta os resultados encontrados para as dezesseis variáveis submetidas ao método. O Quadro 28 mostra as maiores correlações encontradas.

Variável 1	Variável 2	Correlação
QREQUEQTDP (quantidade de dependentes do requerente)	AREQUEMEMB (número de membros da família)	,785(**)
VREQUEPERC (valor da renda *per capta* da família)	VREQUEREND (valor da renda da família)	,774(**)
QREQUEQTDP (quantidade de dependentes do requerente)	QREQUE0A15 (quantidade de dependentes até 15 anos)	,592(**)

Quadro 28 - Variáveis com maiores níveis de correlação.

- **Escolha das variáveis**

A utilização da curva ROC para seleção das variáveis mais relevantes, demonstrou que variáveis como CREQUEIDAD (idade do requerente) e

NREQUEIDDC (idade do cônjuge do requerente) são bons classificadores para a variável alvo TIPO.

A utilização do Ganho de Informação para seleção das variáveis mais relevantes, demonstrou a capacidade preditiva de cada uma das quarenta e cinco variáveis considerando a variável TIPO como variável alvo.

A utilização do Coeficiente de correlação de Pearson para seleção das variáveis mais relevantes, demonstrou o grau de correlação entre cada uma das dezesseis variáveis numéricas.

Com base nos três tipos de análises realizadas, observou-se a menor relevância da variável QREQUEQTDP (quantidade de dependentes, independentemente do recebimento ou não do benefício, do requerente) para classificação do problema. Trata-se de uma variável criada como objetivo de otimizar o processo de modelagem através da RNA, porém a análise de correlação dos dados, utilizando o método de Pearson demonstrou o elevado grau de correlação (0,785) desta variável com a variável AREQUEMEMB (quantidade de membros da família), sendo que, a QREQUEQTDP pelo método do ganho de informação, teve uma das piores classificações, dessa forma optou-se pela retirada da mesma. A escolha pela retirada de apenas uma variável se dá pela opção em deixar que a RNA, com sua robustez e poder de generalização, trate adequadamente as variáveis restantes no modelo experimentado.

5.4.4 Estatística Descritiva dos Dados

Definidos os atributos (variáveis) a serem modelados, foi realizado levantamento exploratório dos dados através de regras estatísticas para as duas categorias de variáveis: numéricas e categóricas.

- **Variáveis Numéricas (contínuas)**

Foram levantados valores de mínimo, máximo, média, desvio-padrão e desvio médio-padrão para as 15 variáveis numéricas: AREQUEMEMB, AREQUENUCO,

CREQUEAGUA, CREQUEENER, CREQUEIDAD, CREQUESANI, CREQUESEXO, EGRAUIDESC, EGRAUIDSCC, FREQUESINE, NREQUEIDDC, QREQUE0A15, QREQUEAP16, VREQUEPERC e VREQUEREND.

Atributo	Descrição	Valor Mínimo	Valor Máximo	Média	Desvio-Padrão
AREQUEMEMB	Número de membros da família	2	20	5,3387	2,019
AREQUENUCO	Número de cômodos	1	7	3,5326	1,642
CREQUEAGUA	Indicador de água encanada	0	1	0,836	0,3703
CREQUEENER	Indicador de energia elétrica	0	1	0,951	0,2158
CREQUEIDAD	Idade do requerente	15	93	41,9088	9,0253
CREQUESANI	Indicador de saneamento básico	0	1	0,7656	0,4237
CREQUESEXO	Sexo do requerente	0	1	0,0839	0,2773
EGRAUIDESC	Grau de instrução do requerente	0	9	3,0375	1,356
EGRAUIDSCC	Grau de instrução do cônjuge do requerente	0	8	1,2542	1,7025
FREQUESINE	Indicador de inscrição no SINE	0	1	0,0424	0,2014
NREQUEIDDC	Idade do cônjuge do requerente	5	99	44,5337	6,4377
QREQUE0A15	Quantidade de dependentes até 15 anos.	0	11	1,8319	1,4934
QREQUEAP16	Quantidade de dependentes a partir de 16 anos.	0	13	2,1212	1,6268
VREQUEPERC	Valor renda per capta da família	0	302	43,4878	27,2032
VREQUEREND	Valor da renda da família	0	1580	217,51	137,292

Quadro 29 - Exploração de dados das variáveis numéricas.

- **Variáveis Categóricas (discretas)**

Foram levantados valores de Gini, quantidade e percentuais de cada classe das 30 variáveis categóricas: CREQUEORGA, CREQUESCON, CESTADUFCO, CREQUEORCO, CESTADUFIC, FREQUESINC, CESTADCODN, CREQUEFECH, FREQUEBFAM, CTPLGDCODI, EBAIRRNOME, EPROPFDESC, NESTADNOMC, NESCIVNOME, ECBOTBDESC, ECBOTBDSCO, ESITTRDESC, ESITESDESC, NESTADNMCO, NESTADNMIC, ECBOTBDSCC, EOCUMODESC, ETPVEDDESC, ETPPISDESC, ETPCOBDESC, EGLETEDESC, ETPCONDESC, EORIGEDESC, EBAIRRNESC e tipo. Alguns resultados são mostrados no Quadro 30. O resultado da análise estatística de algumas das trinta variáveis categóricas encontra-se no Apêndice F.

Resultados

Atributo	Gini	Distribuição

Values	Count	Percent	Histogram
CONTRAPISO	3466	32,21 %	
TIJOLO/CIMENTO	5072	47,13 %	
TERRABATIDA	1498	13,92 %	
CERAMICA	700	6,50 %	
PALAFITAS	26	0,24 %	

ETPPISDESC (Tipo do piso) — 0,6506

Values	Count	Percent	Histogram
LAJE	2183	20,28 %	
TELHACERAMICA	5085	47,25 %	
ZINCO/BRASILIT	2070	19,23 %	
PLASTICO/LONA	1396	12,97 %	
NAOTEM	28	0,26 %	

ETPCOBDESC (Tipo de cobertura) — 0,6818

Values	Count	Percent	Histogram
ARRENDATÁRIO	1567	14,56 %	

EGLETEDESC (Tipo de gleba da terra – porção de terra) — 0,7074

		PROPRIETÁRIO	4725	43,90 %	
		PROPPARENTE	1831	17,01 %	
		POSSEIRO	2385	22,16 %	
		PROPPATRÃO	254	2,36 %	
ETPCONDESC (Situação da moradia)	0,7614	**Values**	**Count**	**Percent**	**Histogram**
		EMACABAMENTO	2795	25,97 %	
		PRONTA	3264	30,33 %	
		PARALISADA	2815	26,16 %	
		SEMINFORMAÇÃO	827	7,68 %	
		INICIADA	512	4,76 %	
		OUTROS	549	5,10 %	

Quadro 30 - Exploração de dados das variáveis categóricas.

Para todo conjunto dos requerentes classificados como sucesso ou insucesso há uma predominância das características de moradia dos mesmos. A maior parte das construções possuem piso de cimento, cobertura de telha cerâmica, sendo os requerentes ou cônjuges os proprietários da porção de terra e a moradia encontra-se com a construção finalizada.

5.4.5 Discretizações e Normalizações

Em seguida, procedeu-se a codificação dos campos categóricos em numéricos para compatibilizar com a entrada da rede neural.

- **Discretização dos Dados**

Através da análise dos histogramas foram definidos intervalos de modo a facilitar a modelagem pela RNA. Segundo Kamber [HaK06], algumas técnicas para discretização dos dados podem ser utilizadas, dentre elas a igualdade de freqüência. Em uma igualdade de frequência em histograma, as barras são criadas de modo que, a freqüência de cada segmento seja constante.

Atributos como CREQUEORGA (Órgão expedidor da carteira de identidade do requerente), CTPLGDCODI (código do tipo do logradouro), NESTADNOMC (Estado de origem do cônjuge) e ESITESDESC (código da situação especial de dependentes) por terem natureza categórica, foram discretizados levando-se em consideração a ocorrência (*count*) de cada classe conforme mostrado no Apêndice G.

- **Normalização dos Dados**

Foram normalizados todos os atributos numéricos no intervalo [0,1] por interpolação linear que consiste em considerar os valores mínimo e máximo de cada atributo no ajuste de escala.

Segundo Kamber [HaK06], as técnicas de normalizações mais utilizadas são:

- Normalização Linear: nessa técnica, a distribuição dos valores do atributo segue uma função linear, preservando a relação entre os dados originais.

  $x_{normalizado} = (x - x_{min}) / (x_{max} - x_{min})$

  onde x é o valor normalizado e correspondente ao valor inicial x do atributo; x_{min} é o valor mínimo dentre todos os valores do atributo e x_{max} é o valor máximo dentre todos os valores do atributo.

- Normalização pela Soma: Nessa técnica, divide-se o valor da variável pela soma dos valores do atributo.

  $X_{normalizado} = x / \sum x$

- Normalização pelo Máximo: Nessa técnica, o maior valor do atributo é tomado para o fator de normalização.

  $X_{normalizado} = x / x_{max}$

- Normalização pelo Score: Nessa técnica, o eixo central é deslocado para a média dos valores do atributo e normaliza em função do desvio-padrão.

$$x_{normalizado} = (x - \bar{x}) / \sigma$$

- **Conversão dos Atributos Categóricos em Numéricos**

Os atributos categóricos EGRAUIDESC e EGRAUIDSCC foram transformados em numéricos para compatibilizar com a entrada da RNA:

Classes	Valor
Analfabeto	1
Ate a 4ª Serie incompleta do ensino fundamental	2
Com 4ª serie completa do ensino fundamental	3
Com ensino fundamental completo	4
De 5ª a 8ª Serie incompleta do ensino fundamental	5
Ensino medio completo	6
Ensino medio incompleto	7
NÃO INFORMADO	0
Superior completo	8
Superior incompleto	9
N/A	0

Quadro 31 - Valores para cada classe das variáveis EGRAUIDESC e EGRAUIDSCC.

Os N valores do domínio do atributo são representados com N bits. cada categoria é discreta a um valor de 1 até N e é representado por uma string de dígitos binários.

Para esta tarefa foi utilizada a técnica de representação binária por temperatura [AVL99]. O código de temperatura é utilizado com maior frequência quando os valores discretos estão relacionados de algum modo. Se, por exemplo, uma variável de natureza discreta puder assumir valores (analfabeto, até a 4ª série incompleta do ensino fundamental, superior incompleto, superior completo); deseja-se que a diferença entre analfabeto e até a 4ª série incompleta do ensino fundamental seja menor, enquanto que a diferença entre analfabeto e superior completo seja a maior possível. O código termômetro para analfabeto será representado como 1000, enquanto que superior completo será 1111 e superior incompleto 1110. A similaridade dos valores do domínio é verificada utilizando-se a Distância de *Hamming* que, conta para cada posição *i* da cadeia de bits se, são iguais ou diferentes. Se diferentes, soma-se uma unidade a um

contador. No exemplo observado, a distância de *Hamming* entre os valores analfabeto e superior completo é igual a três já que os segundo, terceiro e quarto bits são diferentes.

Os campos categóricos CREQUEAGUA, CREQUEENER, CREQUESANI e FREQUESINE foram convertidos em números binários, por relação direta, sendo cada categoria convertida em um atributo que pode assumir valor 1, se a instância possui a categoria correspondente, ou 0 caso contrário.

6 Mineração de Dados PBEM – Modelagem e Avaliação

Neste capítulo são abordadas as fases de modelagem e avaliação de desempenho segundo a metodologia do CRISP-DM, além do custo para erros e acertos do modelo.

6.1 Modelagem

Para realizar a avaliação dos critérios de concessão de crédito ao programa Bolsa Escola, foi utilizado o modelo de rede neural Multilayer Perceptron (MLP), treinada com o algoritmo backpropagation [Hay99], [DGR86], modelo utilizado com sucesso em aplicações de problemas de classificação de padrões. Dentre as características mais atrativas deste tipo de rede neural é possível destacar a excelente capacidade de generalização, a simplicidade de operação da rede e o fato da mesma, ser um aproximador universal de funções [KMH89].

Antes do início do treinamento da RNA, foi desenvolvido em linguagem de programação "C", programa para realizar a separação dos dados, dividindo-se aleatoriamente todo o conjunto de dados em dez subconjuntos (folds) disjuntos, respeitando a mesma proporção dos dados para as classes da variável alvo "Tipo" (90% Sucesso e 10% Insucesso) [HaK06]. Cada um dos dez subconjuntos criado foi utilizado como conjunto de testes (sendo estes uma amostra independente dos dados utilizados na construção do modelo) e os outros nove subconjuntos foram reunidos em conjuntos de treinamento e validação. O processo foi repetido por dez vezes, onde cada um dos 10 conjuntos gerados foi utilizado para avaliar o desempenho da RNA. Esse procedimento é conhecido como *K-fold CrossValidation* (Validação Cruzada com K Conjuntos Estratificados).

Em etapa subseqüente, verificou-se a possibilidade de fornecer maior estabilidade ao sistema pelo método de Monte Carlo utilizando um número de iterações

que, para o trabalho foram consideradas cinco. Dessa forma, o procedimento descrito acima foi refeito por mais quatro vezes de forma a obter cinco iterações resultantes. O resultado de todas as iterações gerou um total de cinqüenta amostras para treinamento, avaliação e teste a serem submetidas à RNA. Dessa forma foram obtidas cinco avaliações para cada instância da base disponível. Posteriormente, foi calculada a mediana dos escores e com isso obteve-se o desempenho final do sistema.

6.1.1 N-Fold Cross-Validation

Este método divide o conjunto dos dados em n subconjuntos sendo que, um subconjunto é separado e utilizado como conjunto de teste, enquanto todos os outros são utilizados no processo de treinamento da RNA. Este processo é repetido n vezes com cada um dos n subconjuntos sendo utilizado como conjunto de teste apenas uma vez [KOH95]. Um vantagem desse método é que todos os exemplos do conjunto de dados são utilizados para treinamento e teste (Ilustração 12).

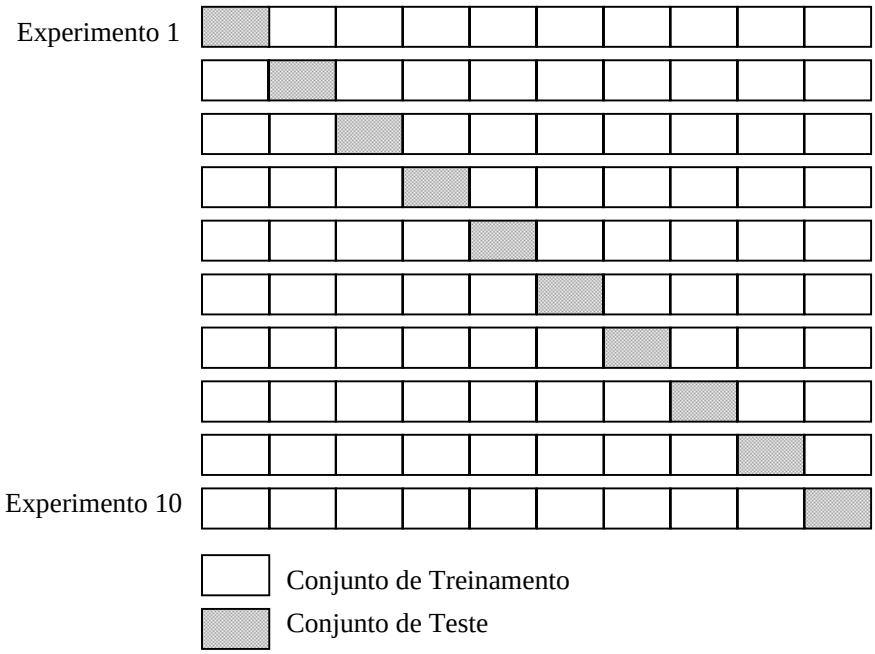


Ilustração 12 - Método Cross-Validation com 10 folds

6.1.2 Monte Carlo

Monte Carlo é o procedimento que investiga as distribuições aleatórias de várias estatísticas. “Área da matemática que está preocupada em experiências com números aleatórios” [HaH64]. Este método é utilizado principalmente em duas aplicações, uma é para avaliar a distribuição aleatória empírica de uma estatística e outra é para estudar os efeitos de violar suposições que estão escondidas em algumas estatísticas.

O objetivo principal é, a partir de simulações, criar um ambiente no qual a informação sobre ações alternativas possíveis possa ser conseguida através de experimentação, de maneira rápida e a um baixo custo.

Para este trabalho o método Monte Carlo é utilizado para resolver um problema[20] através de uma série de simulações aleatórias. Sendo a precisão do resultado final dependente do número de tentativas. Tal equilíbrio entre a precisão do resultado e o tempo computacionalmente viável levou a escolha do número de cinco iterações para aplicação das simulações.

6.1.3 Treinamento e Topologia da Rede

Em etapa final para a modelagem está a construção do modelo da RNA a ser utilizado para o problema de classificação binária. Foram levados em conta, problemas que podem surgir no treinamento, com escolhas inapropriadas dos parâmetros utilizados no modelo (Quadro 32).

Várias combinações na taxa de aprendizado, momentum, critério de parada, número de camadas e neurônios nas camadas intermediárias foram experimentados. Para escolha final de tais combinações, foram considerados o percentual de acerto para classe alvo e tempo de treinamento realizados em um dos cinqüenta subconjuntos de dados.

[20] Problema: Proporcionar maior estabilidade ao sistema.

Configuração	Maior extremo	Menor Extremo
Taxa de Aprendizado	Instabilidade em torno da melhor solução	Demora excessiva no treinamento
Critério de Parada	Reduz o risco de mínimo local, acelera o treinamento e aumenta o risco de instabilidade	Sem efeito de supressão do risco de mínimo local.
Nº ciclos	Má generalização e demora excessiva no treinamento	Não atinge o conhecimento

Quadro 32 - Riscos dos valores extremos para treinamento da RNA.

A rede neural utilizada no experimento deste trabalho foi a rede MLP com uma única camada escondida, visto que segundo Cybenko [Cyb89], estas redes podem generalizar qualquer função linearmente contínua. A partir de experimentos primários, conforme discorrido em parágrafo anterior, optou-se por 5 neurônios na camada escondida, utilizando a ferramenta WEKA[21]. Como algoritmo de aprendizagem foi utilizado o *Backpropagation*, com taxa de aprendizagem igual a 0,001, momentum 0,1, e critério de parada baseado na quantidade de épocas: 10.000 e no incremento do *Mean Squared error* (erro quadrado médio) referente ao conjunto de validação como métrica.

Para realização dos cinqüenta experimentos, resultantes do *ten-fold cross-validation* e do método de Monte Carlo, foram necessários quatro computadores (cada um com capacidade de processamento distinta) executando o treinamento da RNA vinte e quatro horas por dia, durante cerca de doze dias.

Considerando que o problema foi modelado como categórico e em 2 classes, a resposta de cada neurônio pode ser conjugada por meio de uma transformação linear em uma única grandeza escalar (o escore) que está definido no domínio contínuo entre 0 e 1, com o zero representando o "Sucesso" e a unidade (100%) significando o "Insucesso". Essa representação da resposta por uma grandeza escalar possibilita um monitoramento muito mais refinado do desempenho da rede.

[21] WEKA. Ferramenta de mineração de dados em software livre. <http://www.cs.waikato.ac.nz/ml/weka> Acesso em Novembro/2008.

6.1.4 Erro Quadrado Médio

O erro quadrado médio, no campo da estatística, de um estimador é uma das formas de quantificar o montante pelo qual um estimador difere o verdadeiro valor da quantidade a ser estimada. A diferença ocorre devido à aleatoriedade ou simplesmente pelo motivo do estimador não levar em conta as informações que poderiam gerar uma estimativa mais exata [MFD74].

O erro da camada de saída da RNA é dado $E_p=1/2 \sum_{j=1} (t_{pj} - O_{pj})^2$, onde O_{pj} é a resposta ao neurônio e n_j e t_{pj} é a saída desejada.

6.2 Avaliação

A transformação da saída da rede multilayer perceptron em uma resposta contínua para o problema binário da decisão (escore) demanda um limiar (ponto de corte) para definir a separação das classes [SBRN04]. O ponto de corte utilizado foi obtido pelo KS2Máx. As métricas aqui usadas para medição do desempenho são o gráfico do teste estatístico de Kolmogorov-Smirnov (KS2) [Con99], a curva ROC [ProF98], o coeficiente de GINI [Hoff80] e a matriz de confusão [Kan03].

6.2.1 Kolmogorov-Smirnov

Também conhecido como KS, o teste *Kolmogorov-Smirnov* [Con99] compara, através de função de distribuição acumulada, dados oriundos de duas distribuições indicando, nesse trabalho, os sucessos e insucessos.

Em problemas de classificação binária a curva do KS2 é a diferença entre duas funções de distribuição acumuladas tendo a pontuação como variável independente. Uma distribuição contém a pontuação da classe positiva e, a outra, da classe negativa. Tem sido utilizada para avaliar o desempenho de classificadores binários indicando a diferença entre as distribuições de maus e bons exemplos, ou seja, ele avalia o quão bem

o modelo consegue distinguir os exemplos classificados como bons dos exemplos classificados como maus [Con99].

A partir do KS é possível avaliar o desempenho do classificador a partir de pontos de corte diferentes, de forma a fazer com que a decisão do final do classificador seja tomada de forma mais suave e otimizada. O ponto ótimo de corte é o ponto de distância máxima entre as duas classes, que significa garantir o maior percentual de acerto.

Logo, para este trabalho, o ponto de corte utilizado foi o KS2Max. O ponto de corte do KS2Max foi utilizado também para matriz de confusão apresentada na seção 6.2.4

Pelo Gráfico 4, nota-se que existe um ponto de maior separação entre as distribuições dos escores das ocorrências de sucesso e insucesso. Esse ponto indica que o sistema é capaz de prever o insucesso com uma boa margem de acerto. A medição do KS2Max para o conjunto de teste ficou em de 0,3536 e a área (AUC) foi de 0,2470.

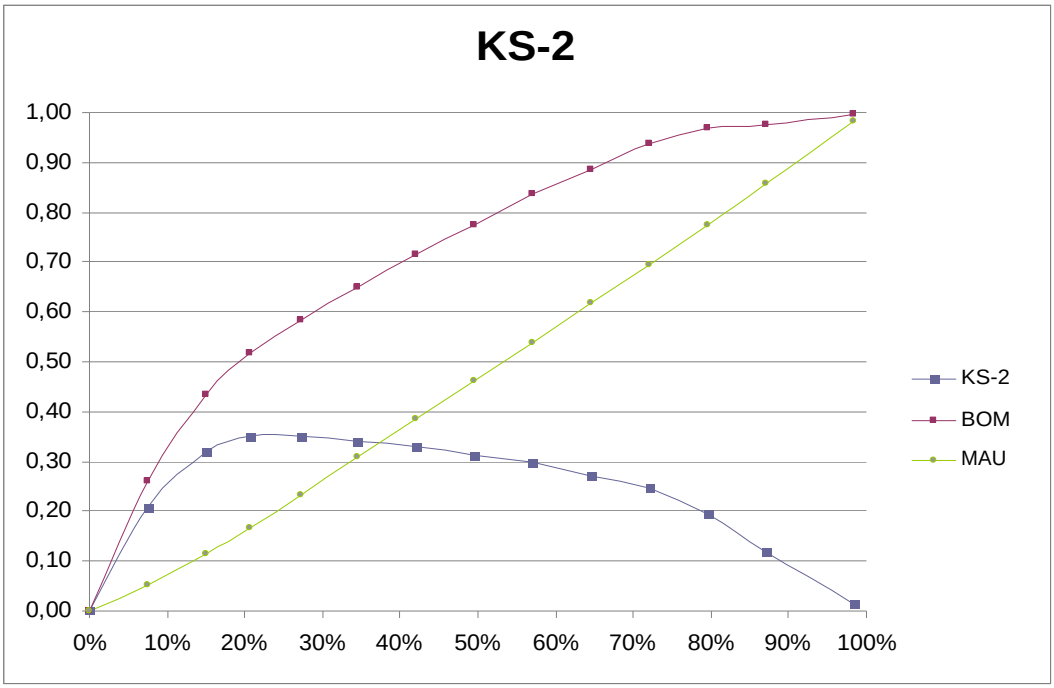


Gráfico 4 - Gráfico KS para RNA.

6.2.2 Gráfico ROC

O gráfico ROC (*Receiver Operating Characteristics*) além de utilizado para avaliação e seleção das variáveis mais significativas, foi utilizado também para aferição do desempenho do modelo experimentado.

As curvas ROC mostram a relação entre o Erro I e a sensibilidade (conceitos estes que serão vistos na seção 6.2.4) o que corresponde à relação das taxas de falsos positivos (FP) e verdadeiros positivos (VP) [ProF98]. Esta relação prediz o comportamento dos classificadores independentemente dos custos e da distribuição das classes. Para cada ponto de corte, a sensibilidade e o complemento da especificidade (1 – especificidade) (conceitos estes que serão vistos também na seção 6.2.4) são calculados e colocados em cada eixo respectivo de um gráfico bidimensional. O Gráfico 5 ilustra a curva ROC do trabalho.

O resultado obtido para este trabalho foi uma área de 0,7470, resultado esse que representa uma maior probabilidade de acerto para classificação de exemplos aleatórios de verdadeiros positivos [FAW05].

Para o trabalho, o resultado diz respeito a maior probabilidade de acerto quanto à classificação das instâncias consideradas insucesso no PBEM.

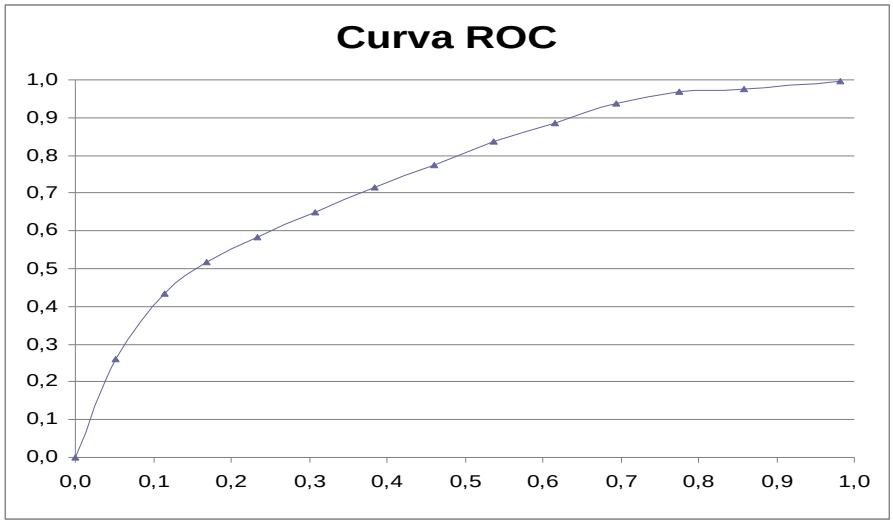


Gráfico 5 - Gráfico da Curva ROC..

6.2.3 Coeficiente de GINI

O Coeficiente de Gini é uma medida de desigualdade desenvolvida pelo estatístico italiano Corrado Gini. É baseado na curva de *Lorenz* e comumente utilizada para calcular a desigualdade de distribuição de renda, mas pode ser usada para qualquer distribuição. Numericamente, varia entre 0 e 1, onde 0 corresponde à completa igualdade de renda (onde todos têm a mesma renda) e 1 corresponde à completa desigualdade (onde uma pessoa tem toda a renda, e as demais nada têm) [Hoff80]. O resultado obtido para este trabalho foi uma área de 0,4405 (Gráfico 6.3), sendo o eixo x correspondente à população e o y correspondente a classe alvo (insucesso).

Para o trabalho, o resultado demonstra uma significativa qualidade do classificador em classificar as instâncias como sucesso ou insucesso.

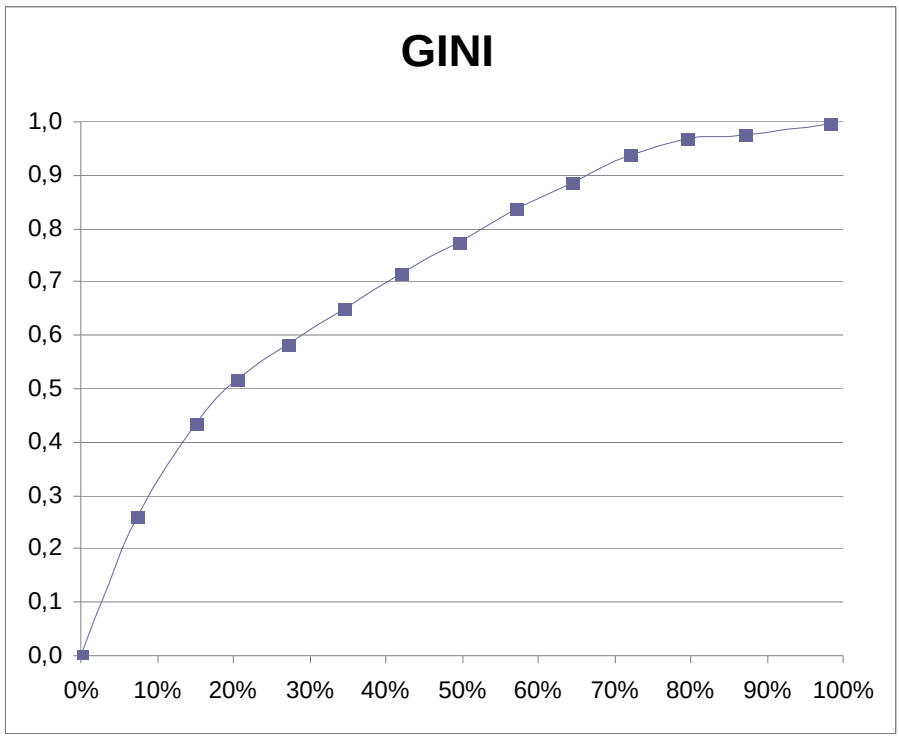


Gráfico 6 - Gráfico da Curva de Lorenz.

6.2.4 Matriz de Confusão

Uma matriz de confusão compara os casos reais desejados com a classificação realizada pelo sistema [Kan03] e por este motivo, é utilizada para avaliação da *performance* dos classificadores. O resultado da classificação de uma determinada instância resulta em uma de quatro situações possíveis. Considerando o acerto como *positive* e o erro como *negative*, se uma instância é *positive* e é classificada pelo sistema como tal, tem-se uma situação de *true positive*; se uma instância é *positive* e é classificada pelo sistema como *negative*, tem-se uma situação de *false negative*. De forma similar, se uma instância é *negative* e é classificada pelo sistema como *negative*, tem-se uma situação de *true negative* e se uma instância é *negative* e é classificada como *positive* tem-se a situação de *false positive.* Tomando como base essas quatro situações possíveis para cada instância, pode-se montar uma matriz de confusão (Quadro 33) que servirá para extrair métricas como, especificidade, sensibilidade e acurácia.

Real / Previsto	***Positivo***	***Negativo***
Positivo	*True Positive*	*False Positive*
Negativo	*False Negative*	*True Negative*

Quadro 33 - Matriz de Confusão

TP (*True Positive*) – Número de previsões acertadas para os casos realmente positivos.

FN (*False Negative*) – Número de previsões erradas para os casos realmente positivos.

TN (*True Negative*) – Número de previsões acertadas para os casos realmente negativos.

FP (*False Positive*) – Número de previsões erradas para os casos realmente negativos.

Para este trabalho:

TPR (True Positive Rate). Número de instâncias classificadas corretamente como insucesso com relação a todos os insucessos da base.

FNR (False Negative Rate). Número de instâncias classificadas erroneamente como sucesso com relação a todos os insucessos da base, complementar ao TPR.

TNR (True Negative Rate). Número de instâncias classificadas corretamente como sucesso com relação a todos os sucessos da base.

FPR (False Positive Rate). Número de instâncias classificadas erroneamente como insucesso com relação a todos os sucessos da base, complementar ao TNR.

A partir da matriz de confusão, pode-se obter duas medidas de erro:

- Erro I: corresponde ao percentual dos requerentes que são casos de sucesso e são classificados como insucesso. É dado pela fórmula FP/(FP+TN). Existe ainda uma outra medida que é equivalente a esta.
    - Especificidade – mede a taxa de acertos do classificador sobre os requerentes que representam o sucesso. É dada pela fórmula TN/(FP+TN).
- Erro II: corresponde ao percentual dos requerentes que são casos de insucesso e são classificados como sucesso. É dado pela fórmula FN/(FN+TP). Existe ainda uma outra medida que é equivalente a esta.
    - Sensibilidade – mede a taxa de acertos do classificador sobre os requerentes que representam o insucesso. É dada pela fórmula TP/(FN+TP).

A taxa de acerto total do classificador é dada pela medida da acurácia (*accuracy*) através da fórmula (TP+TN)/(TP+FN+TN+FP).

A matriz de confusão para este trabalho é representada no Quadro 34.

	Insucesso	**Sucesso**
Insucesso	701	2.173
Sucesso	463	7.425

Quadro 34 - Matriz de Confusão para o trabalho

As respectivas taxas de erro e acerto da matriz de confusão são apresentadas no Quadro 35.

	Insucesso	**Sucesso**
Insucesso	TPR=0,60	FPR=0,23
Sucesso	FNR=0,40	TNR=0,77

Quadro 35 - Taxas de erros e acertos da matriz de confusão

As medidas, para o trabalho, extraídas a partir da matriz de confusão são mostradas no Quadro 36.

Erro I = 0,23	**Erro II** = 0,4
Especificidade = 0,77	**Sensibilidade** = 0,6
Especificidade = 1 – 0,23 = 0,77	**Sensibilidade** = 1 – Erro II = 1 – 0,4 = 0,6
Acurácia = 0,685 ou **68,5%**	

Quadro 36 - Outras medidas de desempenho

A taxa de acerto total (68,5%), demonstra o bom desempenho do classificador para o PBEM.

6.2.5 Custo

A análise de custos foi interpretada levando-se em consideração que a média de dependentes por família requerente é de 2.7, ou seja, mais de um dependente por família requerente. Como o valor do benefício é de um salário mínimo (SM) vigente para famílias com mais de uma criança na escola, consideramos este valor para a

análise. Por outro lado, também foi considerado o custo social representativo[22] (CSR) como o valor a ser considerado por ocasião de erros na operação de classificação. O Quadro 37 relaciona os custos envolvidos com a classificação.

Classificação	Custo Associado
TP - Classificar corretamente como "Insucesso"	0,00/Req.
FP - Classificar erroneamente como "Insucesso"	CSR
TN - Classificar corretamente como "Sucesso"	1 SM/Req.
FN - Classificar erroneamente como "Sucesso"	1 SM/Req. + CSR

SM: Salário Mínimo Vigente
CSR: Custo Social Representativo

Quadro 37 - Custos envolvidos com a classificação da RNA.

Outra forma de analisar os custos envolvidos é fazendo uso da Teoria dos Jogos. A Teoria dos Jogos é a ciência da tomada de decisões em situações de interdependência [Fian06]. Ao contrário de uma decisão unilateral, aqui se trata de decidir levando-se em conta decisões de outro(s), envolvido(s) em um mesmo problema de decisão. Os custos não foram associados por motivo de não ter-se um valor para o custo social representativo definido no trabalho.

Sendo assim, a matriz de resultados, também denominada de matriz de payoffs, possui representação conforme mostrado no Quadro 38.

		Sistema	
		T	**F**
Real	**P**	TP(0,00/Req.)	FP(-CSR)
	N	TN(1 SM/Req.)	FN(-1 SM/Req.- CSR)

Quadro 38 - Matriz com resultados dos custos envolvidos.

[22] Custo Social Representativo: conceito sugerido como forma de mensurar os possíveis custos sociais envolvidos em não beneficiar no PBEM um dependente que é caso de sucesso ou beneficiar um dependente que é caso de insucesso. Seria, por exemplo, o custo para sociedade em ter um jovem envolvido com a criminalidade.

7 Considerações Finais

A motivação para esta dissertação partiu da disponibilidade dos dados sobre o Programa Bolsa Escola da Prefeitura da Cidade do Recife, bem como, a possibilidade de aplicar processo de mineração de dados de forma a gerar benefício social significativo.

7.1 Conclusões

O processo de concessão do benefício uma vez apoiado pelo modelo proposto, poderá gerar uma redução de perdas aos cofres públicos, uma vez que, a decisão para concessão a um requerente será antecipada pelo modelo utilizado para classificar as instâncias como sendo “Sucesso” ou “Insucesso”.

A Diretoria de Apoio Social à Educação da Prefeitura da Cidade do Recife em conjunto com a EMPREL[23], forneceram a base de dados e esclarecimentos necessários para início e continuidade do trabalho.

Através da metodologia CRISP-DM foram abordadas todas as etapas do projeto de mineração de dados com exceção da etapa de distribuição que não faz parte do escopo deste trabalho.

O entendimento do negócio foi amplamente discutido com a equipe da Diretoria de Apoio Social à Educação da Prefeitura da Cidade do Recife que estava sempre disponível a dirimir dúvidas que viessem a aparecer.

O entendimento dos dados compostos por uma base com 184 tabelas demandou um tempo considerável, uma vez que, os funcionários responsáveis pelos dados estavam envolvidos em outros projetos não sendo possível a elucidação de dúvidas de forma rápida. A principal fonte de extração das variáveis foi tabela tbRequerente com todos os

[23] Empresa Municipal de Informática do Recife.

dados cadastrais do requerente ao benefício do Programa Bolsa Escola Municipal; para esta tabela foi feito levantamento estatístico dos preenchimentos de diversos campos.

Na fase de preparação dos dados, as entidades tbDependente e tbEvolucaoFamilia também foram utilizadas para saber o número de dependentes por requerentes que recebem o benefício e o motivo do afastamento, respectivamente. Na preparação dos dados foram tratados os outliers e missing data; foram também criadas novas variáveis e através de três técnicas distintas, avaliadas para seleção ao processo de modelagem.

Para etapa de modelagem, foi escolhida técnica de inteligência artificial utilizando Redes Neurais Artificiais para modelagem do problema. Esta etapa foi também exaustiva, demandando vários dias, principalmente por ter-se optado pelo método de Monte Carlo com objetivo de fornecer maior estabilidade ao sistema; foram feitos vários testes variando combinações de configurações da Rede Neural Artificial antes de se definir o modelo final da RNA a ser aplicado aos cinqüenta subconjuntos.

A última etapa do CRISP-DM no trabalho, a avaliação de desempenho, foi mostrada através do teste de Komogorov-Smirnov, curvas ROC e Lorenz, além de modelo para aferir os custos de escolhas certas e erradas do classificador utilizando a Teoria dos Jogos. Os resultados das avaliações foram satisfatórios e demonstraram o bom desempenho dos classificadores para prever os Insucessos da base.

7.2 Validação

O objetivo do trabalho foi alcançado por demonstrar o bom desempenho dos classificadores utilizando técnica de mineração de dados, com redes neurais artificiais, e pela aceitação por parte da especialista no domínio da aplicação. O resultado vem confirmar algumas constatações empíricas[24] como a influência sócio-econômica sobre os casos de sucesso e insucesso no PBEM.

[24] Constatações a partir de experiências vivenciadas.

7.3 Limitações

Existiram limitações do modelo construído, principalmente, pela ausência de dados atualizados em várias tabelas, o que impossibilitou a utilização de outras variáveis que pudessem representar um ganho para rede neural artificial. Outra limitação é a falta de um dicionário de dados para as diversas entidades que compõem o banco de dados, além de uma falta de documentação do sistema SEBE para entendimento inicial dos dados o que pode ter ocasionado o não aproveitamento de outras informações relevantes ao problema.

7.4 Trabalhos Futuros

Os resultados obtidos viabilizam a realização de outros trabalhos complementares futuros com a base do Programa Bolsa Escola da Prefeitura Municipal do Recife como:

- O desenvolvimento de software que implemente o modelo construído, de forma a apoiar o comitê avaliador em tempo real na decisão sobre a concessão ou não ao benefício;
- Correlação com bases de dados de outras Secretarias e Órgãos como forma de melhor avaliar os motivos dos afastamentos (saúde, segurança pública e outras) e criar variáveis que poderão melhorar a performance da RNA;
- Utilização de outras técnicas de mineração de dados com objetivo de avaliar e encontrar melhores níveis de acerto para classe alvo Insucesso.
- Estudo para levantamento do custo social representativo proposto neste trabalho.

8 Referências Bibliográficas

[AR99] AMARAL, C; RAMOS, S. Programas de Renda Mínima e Bolsa-Escola: Panorama Atual e Perspectivas. (1999). Disponível em: < http://www.iets.org.br/biblioteca/Programas_de_renda_minima_e_Bolsa_Escola.pdf> Acesso em: 10 nov. 2008

[CS03] Cardoso, E.; Souza, A.P. The impact of income transfers on child labor and school attendance in Brazil. São Paulo: USP, 2003. http://www.econ.fea.usp.br/elianacardoso/ECONBRAS/cardoso-souza.pdf. Acesso em setembro/2008.

[PNAD07] Pesquisa Nacional por Amostra de Domicílio. http://www.ibge.gov.br/home/estatistica/populacao/trabalhoerendimento/pnad2007/comentarios2007.pdf. Acessado em 20 de Março de 2009.

[MS99] MAMBRINI, J.; CÉSAR, C. C.; SOARES, J. F. Fatores Determinantes do Desempenho dos Alunos Mineiros no SAEB de 1995. In: 1a. Jornada Latino-americana de Estatística Aplicada. Anais, São Carlos, Brasil. Agosto. 1999..<http://www.fae.ufmg.br/game/saeb95.pdf>. Acesso em novembro/2008.

[Const98] Constituição da República Federativa do Brasil de 1998. Sítio: http://www.planalto.gov.br/ccivil_03/constituicao/constitui%C3%A7ao.htm. Acessado em 10 de Março de 2009.

[ECA09] Estatuto da Criança e do Adolescente. Sítio: http://www.planalto.gov.br/ccivil_03/leis/l8069.htm. Acessado em 10 de Março de 2009.

[TWB44] Sítio do Banco Mundial. Mundialhttp://www.worldbank.org/. Acessado em 10 de Março de 2009.

[PNUD06] Programa das Nações Unidas para o Desenvolvimento http://www.pnud.org.br/pobreza_desigualdade/reportagens/index.php?id01=2388&lay=pde. Acessado em 10 de Março de 2009.

[Cos98] COSTA, Alfredo B. (1998). Exclusões Sociais. Lisboa: Edição Gradiva (Série Cadernos Democráticos).

[SEPLAG08] Secretaria do Planejamento e Gestão do Estado de Santa Catarina. http://www.seplag.rs.gov.br/principal.asp?conteudo=indicadores&act=view&cod_menu=132&cod_indicador=13&cod_menu_esq=123. Acessado em 10 de Março de 2009.

[HDI08] Human Development Indices. (2008).http://hdr.undp.org/en/media/HDI_2008_EN_Tables.pdf Acessado em 27 de Março de 2009.

[HDR08] Human Development Reports (Relatório de Desenvolvimento Humano) http://hdrstats.undp.org/2008/countries/country_fact_sheets/cty_fs_BRA.html. Acessado em 27 de Março de 2008.

[PCR08] Sítio da Prefeitura da Cidade do Recife. http://www.recife.pe.gov.br/2008/08/21/bolsa-escola_recebe_novos_beneficiarios_163535.php. Acesso em dezembro/2008.

[FSS96] Fayyad, U., G. Piatetsky-Shapiro, and P. Smyth, (1996). *The KDD process for extracting useful knowledge from volumes of data*. Commum. ACM,

[FPS96] Fayyad, U., Piatetsky-Shapiro, G., & Smyth, P. (1996). Knowledge Discovery and Data Mining: Towards a Unfying Framework. *In Proceeding of The Second International Conference on Knowledge Discovery and Data Mining (KDD´96).*

[AVL99] Aurélio M.; Vellasco M.; Lopes, Carlos H..(1999) Descoberta de Conhecimento e Mineração de Dados. IAC – Laboratório de Inteligência Computacional Aplicada. PUC-Rio.

[McP43] W.S. McCulloch and W. Pitts. (1943). A logical calculus of the ideas immanent in nervous activity. Bulletin of Mathematical Biophysic.

[Hay99] Haykin, Simon..(1999). Neural Networks: A Compreensive Foundation. 2nd Edition. Prentice Hall International, Inc.. New Jersey, USA.

[BCL07] Braga, Antônio de Pádua; Carvalho, André Ponce; Ludermir, Tereza Bernard. Redes Neurais Artificiais: Teoria e Aplicações. Rio de Janeiro: LTC, 2007.

[Cor02] Cortez P., "Modelos Inspirados na Natureza Para a Previsão de Séries Temporais", Tese de Doutorado, Universidade do Minho, 2002.

[MM70] J.M. Mendel and R. W. McLaren. (1970). *Adaptative, learning, and pattern recognition systems; theory and applications*, chapter Reinforcement-learning control and pattern recognition systems. NW.

[Ros58] F.Rosenblatt. (1958). The perceptron: A probabilistic model for information storage and organization in the brain. Psychol.

[MiP69] M. Minsky and S. Papert. (1969). Perceptrons: An introduction to computacional geometry. MIT Press, Massachusetts.

[Cyb89] G. Cybenko.(1989). Approximation by Superpositions of a sigmoide function. Mathematics of Control, Signals and System.

[Cru07] Cruz A.(2007). “Data Mining via Redes Neuronais Artificiais e Máquinas de Vetores de Suporte”. Dissertação de Mestrado.Universidade do Minho.

[Fah98] Fahlman S.(1998). “Faster Learning Variations on Back-Propagation: An Empirical Study”, In D. Touretzky G. H. and Sejnowski, T., editors, Proceedings of Connectionist Models Summer School. Los Altos CA, USA. Morgan Kaufmann Publishers.

[ZuG93] Zupan J. and Gasteiger J.(1993). “Neural Networks For Chemists: An Introduction”, VCH, New York.

[WiH00] Wirth, R. and J. Hipp. (2000). CRISP-DM: Towards a standard process model for data minig,in In Proceedings of the Fourth International Conference on the Parctical Application of Knowledge Discovery and Data Mining (PADD00).

[Cun05] Cunha R. (2005). “Metodologia para Desenvolvimento de Soluções em Mineração de Dados: Um Estudo Prático em Diagnóstico de Falhas”. Dissertação de Mestrado. UFPE.

[Cri96] CRISP-DM. (1996). CRoss Industry Standard Process for Data Mining. Disponível em: http://www.crisp-dm.org/Overview/index.htm. Acessado em: 12 nov 2008.

[CCK00] CHAPMAN, P; CLINTON, J.; KERBER, R; KHABAZA, T.; REINARTZ, T.; SHEARER, C.; WIRTH, R. CRISP-DM 1.0 Step by Step Data Mining Guide. SPSS Inc. 2000.

[RP05] R. C. PRATI AND P. FLACH, “ROCCER: an algorithm for rule learning based on ROC analysis”, in Proceedings of the Nineteenth International Joint Conference on Artificial Intelligence (IJCAI'2005), Edinburgh (UK), 2005.

[ProF98] PROVOST F., & FAWCETT, T. Robust classification systems for imprecise environments. In Proc. 15th Nat. Conf. on Artificial Intelligence. 1998. < http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.36.5123> Acesso em Novembro/2008.

[FAW05] Fawcett, T. (2005). An introduction to roc analysis. Pattern Recognition Letters.

[Mer07] Merschmann L. (2007). “Classificação Probabilística Baseada em Análise de Padrões”. Tese de Doutorado. Niterói, RJ.

[HaK06] HAN, J., KAMBER, M.; Data Mining: concepts and Techniques. USA: Morgan Kaufmann; Second Edition, 2006

[DGR86] D.E. RUMELHART, G.E. HINTON, AND R.J. WILLIAMS, "Learning internal representations by error propagation", MIT Press Computational Models of Cognition and Perception Series, vol. 1, 1986.

[KMH89] K. HORNIK, M. STINCHCOMBE, AND H. WHITE, "Multilayer feedforward networks are universal approximators", Neural Network, vol. 2, no. 5, 1989.

[KOH95] KOHAVI, R.; A study of cross-validation and bootstrap for accuracy estimation and model selection. *Proceedings of the Fourteenth International Joint Conference on Artificial Intelligence:* 1137–1143. (Morgan Kaufmann, San Mateo).

[HaH64] Hammersley, J.M e Handscomb, D. C. (1964). *Monte Carlo methods*. Methuen, London.

[MFD74] Mood, A., F. Graybill, D. Boes (1974). *Introduction to the Theory of Statistics (p. 229)* (3 ed.). McGraw-Hill.

[SBRN04] ROBERTO A.F. SANTOS, PAULO J.L. ADEODATO ADRIAN L. ARNAUD, RODRIGO C.L.V. CUNHA, GERMANO C. VASCONCELOS, DOMINGOS S.M.P. MONTEIRO - Uma Aplicação de Mineração de Dados na Manutenção de Redes de Telefonia. SBRN, 2004.

[Con99] CONOVER, W.J., Practical Nonparametric Statistics, 3rd ed., John Wiley & Sons, New York, USA, 1999.

[Hoff80] Hoffman, R. (1980). *Estatística para Economistas*. Pioneira, São Paulo, Brasil.

[Kan03] Kantardzic, M. (2003). *Data Mining: Concepts, Models, Methods and Algorithms.* John Wiley & Sons.

[Fian06] FIANI, R.(2006) Teoria dos Jogos. 2a. Edição. Rio de Janeiro: Elsevier.

Apêndice A

Possuir Bens

Descrição dos Bens	Pontuação
TV P/B	0
LIQUIDIFICADOR	0
RÁDIO	0
FOGÃO A GÁS	0
NÃO TEM	0
NÃO TEM	0
TV CORES	-5
APAR. DE SOM	-5
BICICLETA	-5
MÁQ. DE COSTURA	-10
GELADEIRA	-10
TELEFONE	-10
VÍDEO CASSETE/ DVD	-10
CARROÇA	-10
MÁQ. DE LAVAR	-50
BARRACA	-80
OUTRO LOTE	-100
ESTAB. COMERCIAL	-150
OUTRA CASA	-200

Quantidade de Dependentes

Quantidade	Pontuação
8	490
7	420
6	350
5	280
4	210
3	140
2	70
1	0

Tipo de Cobertura

Cobertura	Pontuação
PLASTICO/LONA	50
ZINCO/BRASILIT	25
TELHA CERAMICA	10
LAJE	-25
NAO TEM	-25

Tipo de Construção

Construção	Pontuação
PARALISADA	20
INICIADA	15
EM ACABAMENTO	10
SEM INFORMAÇÃO	0
OUTROS	0
PRONTA	-15

Tipo de Piso

Piso	Pontuação
PALAFITAS	50
TERRA BATIDA	25
CONTRAPISO	15
TIJOLO/CIMENTO	-5
CERAMICA	-5

Tipo de Vedação

Vedação	Pontuação
MADEIRITE	25
MADEIRA	25
OUTRO	0
ALVENARIA	-25
ADOBE	-25
TAIPA NAO REVESTIDA	-25
NÃO TEM	-25

Tipo de Ocupação da Moradia

Tipo da Moradia	Pontuação
ALUGADO	100
FINANCIADO	100
ARRENDADO	100
INVASÃO	50
CEDIDO	0
PRÓPRIO	0
NÃO TEM	0
OUTRA	0

Renda Familiar

Faixa da Renda(R$)	Pontuação
0 à 30	100
31 à 60	75
61 à 90	50
91 à 120	25
121 à 150	0
151 à 180	-100

Situação Especial

Situação Especial	Pontuação
PORTADOR DE DEFICIÊNCIA INCAPAZ DE PROVER SUSTENTO	50
PORTADOR DE DOENÇA CRÔNICA	50
TAIPA REVESTIDA	-25
MATERIAL APROVEITADO	-25

Situação de Trabalho

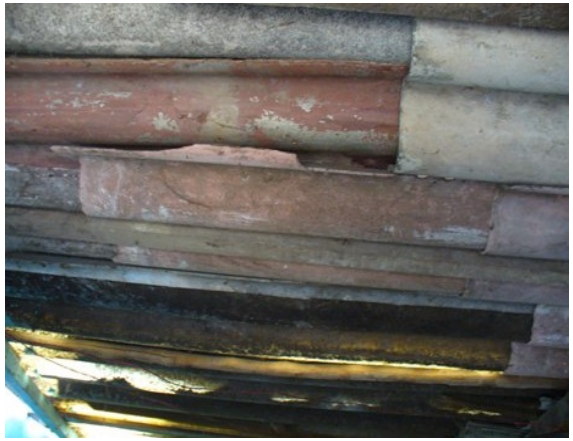
Situação	Pontuação
DESEMPREGADO	50
BISCATEIRO	50
NAO TRABALHA	50
ASSALARIADO SEM CARTEIRA DE TRABALHO	30
ASSALARIADO COM CARTEIRA DE TRABALHO	25
TRABALHADOR RURAL	25
OUTRA	25
PENSIONISTA/APOSENTADO/BPC	0
EMPREGADOR RURAL	0
AUTONOMO COM PREVIDENCIA SOCIAL	-10
EMPREGADOR	-25
AUTONOMO SEM PREVIDENCIA SOCIAL	-40
CRIAÇA/ADOLESCENTE SOB PROTEÇÃO ESPECIAL	50
CRIANÇA/ADOLESCENTE SOB MEDIDA SÓCIO-EDUCATIVA	50
CRIANÇAS DE 0 A 6 ANOS COM DESNUTRIÇÃO	50
CRIANÇA/ADOLESCENTE SOB RISCO PESSOAL/SOCIAL (TRAB.INF.)	50

Apêndice B

Possuir Bens (geladeira / pontuação=-10)

Tipo de Vedação (Taipa Não Revestida / Pontuação=-25)

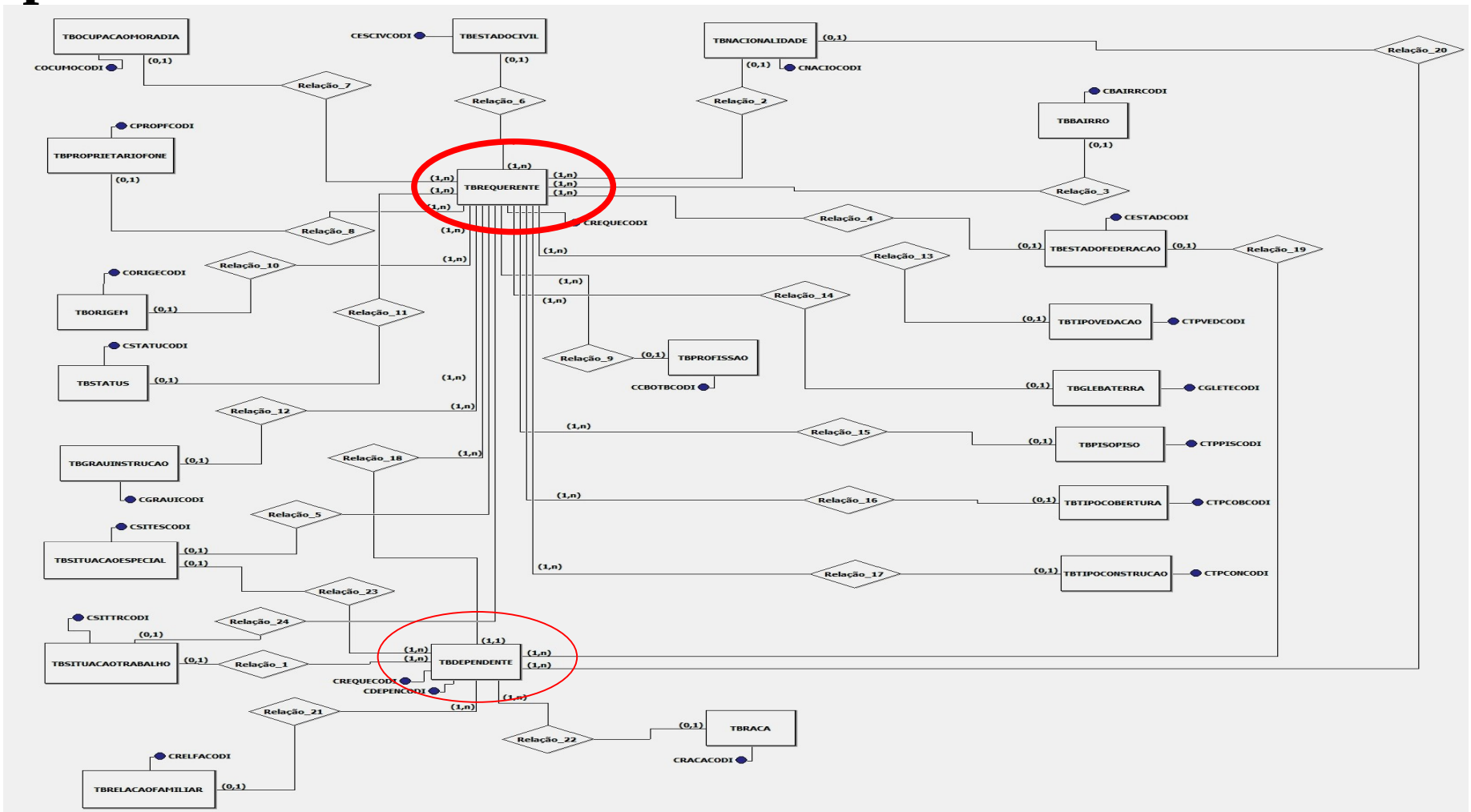
Tipo de Cobertura (Brasilit / pontuação=25)



Tipo de Piso (Terra Batida / Pontuação = 25)

Tipo de Vedação (Taipa Não Revestida / Pontuação=-25)

Apêndice C


TBOCUPACAOMORADIA
COCUMOCODI
(0,1)
Relação_7
CESCIVCODI
TBESTADOCIVIL
(0,1)
Relação_6
TBNACIONALIDADE
(0,1)
CNACIOCODI
Relação_2
Relação_20
CBAIRRCODI
TBBAIRRO
(0,1)
Relação_3
CPROPFCODI
TBPROPRIETARIOFONE
(0,1)
TBREQUERENTE
(1,n)
CREQUECODI
Relação_8
Relação_4
CESTADCODI
TBESTADOFEDERACAO
Relação_19
Relação_13
Relação_10
CORIGECODI
TBORIGEM
(0,1)
Relação_11
Relação_14
CSTATUCODI
TBSTATUS
(0,1)
Relação_9
TBPROFISSAO
CCBOTBCODI
TBTIPOVEDACAO
CTPVEDCODI
TBGLEBATERRA
CGLETECODI
Relação_12
Relação_15
TBPISOPISO
CTPPISCODI
TBGRAUINSTRUCAO
CGRAUICODI
Relação_18
Relação_5
Relação_16
TBTIPOCOBERTURA
CTPCOBCODI
CSITESCODI
TBSITUACAOESPECIAL
Relação_23
Relação_17
TBTIPOCONSTRUCAO
CTPCONCODI
CSITTRCODI
Relação_24
TBSITUACAOTRABALHO
Relação_1
TBDEPENDENTE
(1,1)
CREQUECODI
CDEPENCODI
Relação_21
Relação_22
TBRACA
CRACACODI
CRELFACODI
TBRELACAOFAMILIAR

Apêndice D

#	FK	PK	Field Name	Field Type	Domain	Size	Scale	Subtype	Array	Not Null	Charset	Collate	Description
1		PK1	CREQUECODI	INTEGER						☒			Código do Requerente
2			CREQUEINSC	VARCHAR		10				☐	NONE	NONE	Inscrição Antiga do Requerente
3			CESCOLCODI	CHAR		3				☐	NONE	NONE	Código da Escola
4			CREQUETESC	CHAR		1				☐	NONE	NONE	Tipo da Escola (E-Estadual, M-Municipal)
5			NREQUENOME	VARCHAR		40				☒	NONE	NONE	Nome do Requerente
6			CREQUESEXO	CHAR		1				☒	NONE	NONE	Sexo do Requerente
7			DREQUENASC	TIMESTAMP						☒			Data de Nascimento
8			EREQUEENDE	VARCHAR		40				☒	NONE	NONE	Logradouro
9			CREQUENUME	VARCHAR		5				☒	NONE	NONE	Número da Moradia
10			EREQUECOMP	VARCHAR		20				☐	NONE	NONE	Complemento do endereço
11	FK		CBAIRRCODI	SMALLINT						☒			Código do Bairro
12			EREQUENCEP	CHAR		8				☒	NONE	NONE	Número do CEP
13			EREQUEREFE	VARCHAR		40				☐	NONE	NONE	Ponto de Referência
14			CMUNICICEP	INTEGER						☒			Código do Município de Residência
15	FK		CESTADCODM	CHAR		2				☒	NONE	NONE	UF de Moradia
16			CREQUEZONA	CHAR		1				☒	NONE	NONE	Zona Residêncial
17	FK		CNACIOCODI	SMALLINT						☒			Código da Nacionalidade
18			CMUNICICEO	INTEGER						☐			Código do Município de Origem
19			DREQUECHEG	TIMESTAMP						☒			Data de Chegada
20			CREQUEDDDF	VARCHAR		3				☐	NONE	NONE	DDD do Telefone
21			AREQUEFONE	VARCHAR		8				☐	NONE	NONE	Número do Telefone de Contato

22	FK	CPROPFCODI	SMALLINT			☐			Código do Proprietário do Fone
23		NREQUENOMP	VARCHAR	50		☐	NONE	NONE	Nome do Pai
24		NREQUENOMM	VARCHAR	50		☐	NONE	NONE	Nome da Mãe
25		AREQUEISOC	CHAR	14		☐			Número da Inscrição Social (PIS/PASEP...
26		CREQUENUID	VARCHAR	15		☐	NONE	NONE	Número da Carteira de Identidade
27		CREQUEORGA	VARCHAR	5		☐	NONE	NONE	Órgão da CI
28	FK	CESTADCODI	CHAR	2		☐	NONE	NONE	Estado Expedidor da CI
29		DREQUEEMII	TIMESTAMP			☐			Data de Emissão da CI
30		AREQUENCPF	VARCHAR	11		☐	NONE	NONE	Número do CPF
31		AREQUETECN	VARCHAR	10		☐			Número do Termo da Certidão de Nasci...
32		AREQUENLIV	VARCHAR	10		☐			Número do Livro da CN
33		AREQUEFOLH	VARCHAR	10		☐			Número da Folha da CN
34		DREQUEEMCN	TIMESTAMP			☐			Data de Emissão da CN
35	FK	CESTADCODC	CHAR	2		☐	NONE	NONE	Estado Emissor da CN
36		NREQUECART	VARCHAR	40		☐	NONE	NONE	Nome do Cartório da CN
37		AREQUECTPS	INTEGER			☐			Número da Carteira Profissional
38		AREQUESECP	INTEGER			☐			Série da CP
39		DREQUECTPS	TIMESTAMP			☐			Data de Emissão CP
40	FK	CESCIVCODI	SMALLINT			☒			Código do Estado Civil
41	FK	CGRAUICODI	SMALLINT			☐			Código do Grau de Instrução
42	FK	CCBOTBCODI	SMALLINT			☐			Código da Profissão
43	FK	CCBOTBCODO	SMALLINT			☐			Código da Ocupação
44	FK	CSITTRCODI	SMALLINT			☒			Código da Situação de Trabalho
45		DREQUEDESE	TIMESTAMP			☐			Data do Desemprego
46		FREQUESINE	CHAR	1		☐	NONE	NONE	Indicador de Inscrição no SINE (S/N)
47	FK	CSITESCODI	SMALLINT			☐			Código da Situação Especial
48		NREQUENCON	VARCHAR	40		☐	NONE	NONE	Nome do Cônjuge
49		CREQUESCON	CHAR	1		☐	NONE	NONE	Sexo do Cônjuge
50		DREQUEDNCO	TIMESTAMP			☐			Data de NAscimento do Cônjuge
51	FK	CESTADUFCO	CHAR	2		☐	NONE	NONE	Estado de origem do Cônjuge
52		DREQUECHCO	TIMESTAMP			☐			Data de Chegada do Cônjuge
53		CREQUEIDCO	VARCHAR	15		☐	NONE	NONE	Número da Carteira de Identidade do Cô...
54		CREQUEORCO	VARCHAR	5		☐	NONE	NONE	Órgão Expedidor da CI do Cônjuge
55	FK	CESTADUFIC	CHAR	2		☐	NONE	NONE	Estado Emissor da CI do Cônjuge
56		DREQUEEMIC	TIMESTAMP			☐			Data de Emissão da Ci do Cônjuge
57	FK	CGRAUICODC	SMALLINT			☐			Código do Grau de Instrução do Cônjuge
58	FK	CCBOTBCODC	SMALLINT			☐			Código da Profissão do Cônjuge
59	FK	CCBOTBCOC2	SMALLINT			☐			Código da Ocupação do Cônjuge
60	FK	CSITTRCODC	SMALLINT			☐			Código da Situação no Mercado de Trab...
61		DREQUEDESC	TIMESTAMP			☐			Data de Desemprego do Cônjuge
62		FREQUESINC	CHAR	1		☐	NONE	NONE	Indicador de Inscrição no SINE do Cônju...
63	FK	CSITESCODC	SMALLINT			☐			Código da Situação Especial do Cônjuge

63	F	CSITESCODC	SMALLINT				☐			Código da Situação Especial do Conjuge
64	F	COCUMOCODI	SMALLINT				☒			Código da Ocupação de Moradia
65	F	CTPVEDCODI	SMALLINT				☒			Código do Tipo de Vedação
66	F	CTPPISCODI	SMALLINT				☒			Código do Tipo de Piso
67	F	CTPCOBCODI	SMALLINT				☒			Código do Tipo de Cobertura
68		CREQUEENER	CHAR		1		☒	NONE	NONE	Indicador de Energia Elétrica (S/N)
69		CREQUEAGUA	CHAR		1		☒	NONE	NONE	Indicador de Água Encanada (S?N)
70		CREQUESANI	CHAR		1		☒	NONE	NONE	Indicador de Instalação Sanitária (S/N)
71	F	CGLETECODI	SMALLINT				☐			Código da Gleba da Terra
72	F	CTPCONCODI	SMALLINT				☒			Código do Tipo da Construção
73		AREQUENUCO	SMALLINT				☒			Número de Cômodos
74		AREQUEMEMB	SMALLINT				☒			Número de Membros da Família
75		VREQUEREND	SMALLINT				☒			Valor da Renda da Família
76		VREQUEPERC	SMALLINT				☒			Valor da Renda Per-capta
77		FREQUEVEIC	CHAR		1		☐	NONE	NONE	Indicador de Veículo (C/M)
78		DREQUEANFA	CHAR		4		☐	NONE	NONE	Ano de Fabricação do Veículo
79		VREQUEVAVE	SMALLINT				☐			Valor atual do veículo
80		EREQUEINFC	VARCHAR		5000		☐	NONE	NONE	Informações Complementares do Requer...
81		DREQUEINSC	TIMESTAMP				☒			Data de Inscrição
82	F	CSTATUCODI	SMALLINT				☒			Código do Status
81		DREQUEINSC	TIMESTAMP				☒			Data de Inscrição
82	F	CSTATUCODI	SMALLINT				☒			Código do Status
83		NREQUEENTR	VARCHAR		40		☐	NONE	NONE	Nome do Entrevistador
84		CREQUEPONT	SMALLINT				☒			Pontuação
85		TREQUEULAT	TIMESTAMP				☒			Data da última atualização
86		CESTADCOCP	CHAR		2		☐	NONE	NONE	
87		FREQUEVISI	CHAR		1		☐	NONE	NONE	
88		CTPDOCCODI	SMALLINT				☐			
89		CESTADCODN	CHAR		2		☐	NONE	NONE	
90		DREQUEPONT	TIMESTAMP				☐			
91		CREQUETIPO	CHAR		1		☐	NONE	NONE	
92		DREQUESELE	TIMESTAMP				☐			
93		DREQUETRAN	TIMESTAMP				☐			
94	F	CORIGECODI	CHAR		3		☐	NONE	NONE	
95		CREQUEFECH	CHAR		1		☐	NONE	NONE	
96		DREQUEEXPO	TIMESTAMP				☐			
97		DREQUEINPG	TIMESTAMP				☐			
98		FREQUEPONT	CHAR		1		☐	NONE	NONE	
99		DREQUEEXCE	TIMESTAMP				☐			
100		DREQUEAPRO	TIMESTAMP				☐			
101		FREQUEBFAM	CHAR		1		☐	NONE	NONE	
102		CTPLGDCODI	CHAR		8		☒	NONE	NONE	

Apêndice E

	CREQUESEXO	FREQUESINE	CREQUEENER	CREQUEAGUA	CREQUESANI	AREQUENUCO	AREQUEMEMB	VREQUEREND	VREQUEPERC
CREQUESEXO	1	,066(**)	0,01	0,015	0,011	0,006	,048(**)	,073(**)	,039(**)
FREQUESINE	,066(**)	1	0,005	0,008	,027(**)	0,018	-0,018	,028(**)	,044(**)
CREQUEENER	0,01	0,005	1	,351(**)	,093(**)	,077(**)	0,015	,064(**)	,059(**)
CREQUEAGUA	0,015	0,008	,351(**)	1	,242(**)	,153(**)	,049(**)	,119(**)	,098(**)
CREQUESANI	0,011	,027(**)	,093(**)	,242(**)	1	,295(**)	,060(**)	,177(**)	,143(**)
AREQUENUCO	0,006	0,018	,077(**)	,153(**)	,295(**)	1	,164(**)	,240(**)	,148(**)
AREQUEMEMB	,048(**)	-0,018	0,015	,049(**)	,060(**)	,164(**)	1	,263(**)	-,292(**)
VREQUEREND	,073(**)	,028(**)	,064(**)	,119(**)	,177(**)	,240(**)	,263(**)	1	,774(**)
VREQUEPERC	,039(**)	,044(**)	,059(**)	,098(**)	,143(**)	,148(**)	-,292(**)	,774(**)	1
EGRAUIDESC	0,013	,088(**)	0,008	,029(**)	,045(**)	,049(**)	-,069(**)	,056(**)	,088(**)
EGRAUIDSCC	,103(**)	-0,007	0,009	,022(*)	,059(**)	,082(**)	,138(**)	,159(**)	,057(**)
CREQUEIDAD	,123(**)	-,045(**)	0,016	,035(**)	,041(**)	,116(**)	,080(**)	,082(**)	,056(**)
QREQUEQTDP	,038(**)	-0,01	,025(**)	,058(**)	,083(**)	,159(**)	,785(**)	,341(**)	-,146(**)
QREQUE0A15	0,006	0,004	0,007	-0,01	-0,018	-,050(**)	,456(**)	,089(**)	-,188(**)
QREQUEAP16	,040(**)	-0,016	,025(*)	,080(**)	,118(**)	,238(**)	,536(**)	,334(**)	-0,005
NREQUEIDDC	-,074(**)	-,028(**)	0,016	,032(**)	,039(**)	,112(**)	,051(**)	,027(**)	0,004

Apêndice E

	EGRAUIDESC	EGRAUIDSCC	CREQUEIDAD	QREQUEQTDP	QREQUE0A15	QREQUEAP16	NREQUEIDDC
CREQUESEXO	0,013	,103(**)	,123(**)	,038(**)	0,006	,040(**)	-,074(**)
FREQUESINE	,088(**)	-0,007	-,045(**)	-0,01	0,004	-0,016	-,028(**)
CREQUEENER	0,008	0,009	0,016	,025(**)	0,007	,025(*)	0,016
CREQUEAGUA	,029(**)	,022(*)	,035(**)	,058(**)	-0,01	,080(**)	,032(**)
CREQUESANI	,045(**)	,059(**)	,041(**)	,083(**)	-0,018	,118(**)	,039(**)
AREQUENUCO	,049(**)	,082(**)	,116(**)	,159(**)	-,050(**)	,238(**)	,112(**)
AREQUEMEMB	-,069(**)	,138(**)	,080(**)	,785(**)	,456(**)	,536(**)	,051(**)
VREQUEREND	,056(**)	,159(**)	,082(**)	,341(**)	,089(**)	,334(**)	,027(**)
VREQUEPERC	,088(**)	,057(**)	,056(**)	-,146(**)	-,188(**)	-0,005	0,004
EGRAUIDESC	1	,088(**)	-,182(**)	-,041(**)	,023(*)	-,071(**)	-,113(**)
GRAUIDSCC	,088(**)	1	-,046(**)	,184(**)	,026(**)	,200(**)	-,114(**)
CREQUEIDAD	-,182(**)	-,046(**)	1	0,01	-,311(**)	,298(**)	,366(**)
QREQUEQTDP	-,041(**)	,184(**)	0,01	1	,592(**)	,673(**)	0,002
QREQUE0A15	,023(*)	,026(**)	-,311(**)	,592(**)	1	-,198(**)	-,178(**)
QREQUEAP16	-,071(**)	,200(**)	,298(**)	,673(**)	-,198(**)	1	,166(**)
NREQUEIDDC	-,113(**)	-,114(**)	,366(**)	0,002	-,178(**)	,166(**)	1

Tabela 5.26 – Resultado da Análise de Correlação pelo Coeficiente de Pearson.

Apêndice F

Attribute	Gini	Distribution

Values	Count	Percent	Histogram
SSP	9478	88,07 %	
SDS	987	9,17 %	
DPT	1	0,01 %	
SP	3	0,03 %	
ITB	7	0,07 %	
SSD	1	0,01 %	
SPP	1	0,01 %	
RFB	3	0,03 %	
ITP	1	0,01 %	
SSO	2	0,02 %	
RGN	1	0,01 %	
MEX	1	0,01 %	
DS	1	0,01 %	

CREQUEORGA 0,2154

		N/A	268	2,49 %	
		AP	1	0,01 %	
		DPD	1	0,01 %	
		SSS	1	0,01 %	
		OE	1	0,01 %	
		SDSP	1	0,01 %	
		SSPQ	1	0,01 %	
		SEPC	1	0,01 %	

		Values	Count	Percent	Histogram
CTPLGDCODI	0,3060	R	8910	82,79 %	
		EST	50	0,46 %	
		1ATV	28	0,26 %	
		FAV	1	0,01 %	
		AV	709	6,59 %	
		TV	689	6,40 %	
		SUB	10	0,09 %	
		CRG	60	0,56 %	
		ROD	16	0,15 %	

		VL	39	0,36 %
		AT	66	0,61 %
		LD	30	0,28 %
		BC	59	0,55 %
		ESC	35	0,33 %
		LRG	3	0,03 %
		2ATV	21	0,20 %
		3ATV	11	0,10 %
		PRL	1	0,01 %
		PC	7	0,07 %
		6ATV	1	0,01 %
		2ASUB	2	0,02 %
		5ATV	1	0,01 %
		LOT	1	0,01 %
		SIT	3	0,03 %
		1ASUB	1	0,01 %
		Q	1	0,01 %
		4ATV	3	0,03 %

		C	1	0,01 %	
		MRO	1	0,01 %	
		3ASUB	1	0,01 %	
		PTE	1	0,01 %	

		Values	Count	Percent	Histogram
NESTADNOMC	0,3918	N/A	7908	73,48 %	
		PERNAMBUCO	2812	26,13 %	
		PARAIBA	13	0,12 %	
		ALAGOAS	9	0,08 %	
		RIOGRANDEDONORTE	2	0,02 %	
		SAOPAULO	6	0,06 %	
		PARA	1	0,01 %	
		BAHIA	3	0,03 %	
		RIODEJANEIRO	3	0,03 %	
		PIAUI	1	0,01 %	
		MINASGERAIS	1	0,01 %	
		CEARA	2	0,02 %	
		SERGIPE	1	0,01 %	

ESITESDESC	0,0424	Values	Count	Percent	Histogram
		N/A	10530	97,84 %	
		PORTADORDEDOENÇACRÔNICA	104	0,97 %	
		PORTADORDEDEFICIÊNCIAINCAPAZDEPROVERSUSTENTO	121	1,12 %	
		CRIANÇA/ADOLESCENTESOBMEDIDASÓCIOEDUCATIVA	5	0,05 %	
		CRIANÇA/ADOLESCENTESOBPROTEÇÃOESPECIAL	1	0,01 %	
		CRIANÇASDE0A6ANOSCOMDESNUTRIÇÃO	1	0,01 %	

		Values	**Count**	**Percent**	**Histogram**
ETPPISDESC	0,6506	CONTRAPISO	3466	32,21 %	
		TIJOLO/CIMENTO	5072	47,13 %	
		TERRABATIDA	1498	13,92 %	
		CERAMICA	700	6,50 %	
		PALAFITAS	26	0,24 %	

		Values	Count	Percent	Histogram
ETPCOBDESC	0,6818	LAJE	2183	20,28 %	
		TELHACERAMICA	5085	47,25 %	
		ZINCO/BRASILIT	2070	19,23 %	
		PLASTICO/LONA	1396	12,97 %	
		NAOTEM	28	0,26 %	

		Values	Count	Percent	Histogram
EGLETEDESC	0,7074	ARRENDATÁRIO	1567	14,56 %	
		PROPRIETÁRIO	4725	43,90 %	
		PROPPARENTE	1831	17,01 %	
		POSSEIRO	2385	22,16 %	
		PROPPATRÃO	254	2,36 %	

		Values	Count	Percent	Histogram
ETPCONDESC	0,7614	EMACABAMENTO	2795	25,97 %	
		PRONTA	3264	30,33 %	

		PARALISADA	2815	26,16 %
		SEMINFORMAÇÃO	827	7,68 %
		INICIADA	512	4,76 %
		OUTROS	549	5,10 %

Apêndice G

Attribute	Gini	Distribution

Values	Count	Percent	Histogram
SSP	9478	88,07 %	
OUTROS	1284	11,93 %	

CREQUEORGA 0,2101

Values	Count	Percent	Histogram
R	8910	82,79 %	
OUTROS	454	4,22 %	
AV	709	6,59 %	
TV	689	6,40 %	

CTPLGDCODI 0,3043

Values	Count	Percent	Histogram
N/A	7908	73,48 %	
PERNAMBUCO	2812	26,13 %	
OUTROS	42	0,39 %	

NESTADNOMC 0,3918

Values	Count	Percent	Histogram
N/A	10530	97,84 %	
OUTROS	232	2,16 %	

ESITESDESC 0,0422

Dissertação de Mestrado Profissional apresentada por **Roberto Tabosa Florêncio Filho** Pós-Graduação em Ciência da Computação do Centro de Informática da Universidade Federal de Pernambuco, sob o título, **"Uma Aplicação de Mineração de Dados ao Programa Bolsa Escola da Prefeitura da Cidade do Recife"**, orientada pelo **Prof. Paulo Jorge Leitão Adeodato** e aprovada pela Banca Examinadora formada pelos professores:

Prof. Germano Crispim Vasconcelos
Centro de Informática / UFPE

Dr. Adrian Lucena Arnaud

Prof. Paulo Jorge Leitão Adeodato
Centro de Informática / UFPE

Visto e permitida a impressão.
Recife, 23 de abril de 2009.

Prof. FRANCISCO DE ASSIS TENÓRIO DE CARVALHO
Coordenador da Pós-Graduação em Ciência da Computação do
Centro de Informática da Universidade Federal de Pernambuco.