Utilizando Técnicas de Mineração de Dados Para Apoiar a Busca Ativa de Famílias em Situação de Vulnerabilidade e Risco Social

Marcos A. P. Terrin[1], **Carlos N. Silla Jr.**[1], **Pedro H. Bugatti**[1]

[1]Programa de Pós-Graduação em Informática (PPGI)
Universidade Tecnológica Federal do Paraná (UTFPR)
Cornélio Procópio – PR – Brasil

marcos.terrin@gmail.com,{carlosjunior,pbugatti}@utfpr.edu.br

Abstract. *In the current Brazilian Government there is a Social Assistance policy that is highly concerned about helping families who might be at social risk and vulnerability. The processes of identification of these families is known as "active search". The task of active search is defined in a document of the Brazilian Ministry of Social Development and Fight Against Hunger. The document provides the main guidelines about how to perform the active search. However, despite the task importance, there are still no tools to help the social assistants with this task. In this work we investigate the use of data mining techniques for this task. The preliminary results show that the classification models used always predict the majority class due to a data imbalance problem. After balancing the data the same classification models achieve 66.66% accuracy.*

Resumo. *No âmbito da Assistência Social, existe a necessidade de se identificar as famílias em situação de vulnerabilidade e risco social, processo esse chamado de "busca ativa". O Ministério do Desenvolvimento Social e Combate à Fome do Brasil orienta que seja realizado o cruzamento de bases de dados como forma de realizar a busca ativa, mas não disponibiliza nenhuma ferramenta para realização desse processo. Este trabalho busca identificar e utilizar técnicas de mineração de dados adequadas para realizar a identificação das famílias em situação de vulnerabilidade e risco social e apoiar a busca ativa. Os resultados preliminares obtidos até o momento demonstraram que na maioria dos casos os modelos prevêem sempre a classe majoritária. Após realizar um balanceamento das amostras foi obtido um resultado de 66.66% das amostras classificadas corretamente.*

1. Introdução

O Plano Brasil sem Miséria foi lançado em junho de 2011 pelo governo federal e seu objetivo é elevar a renda e as condições de bem-estar da população. Para isso, as famílias extremamente pobres que ainda não são atendidas deverão ser localizadas e incluídas de forma integrada nos diversos programas sociais do governo federal (Ex.: Programa Bolsa Família) de acordo com as suas necessidades. Esse plano é direcionado aos brasileiros que vivem em lares cuja renda familiar é de até R$ 70 por pessoa.

De acordo com o Censo 2010 do Instituto Brasileiro de Geografia e Estatística (IBGE), estão nesta situação 16,2 milhões de brasileiros. Para isso, o Plano Brasil sem

Miséria desenvolveu uma nova estratégia chamada **"Busca Ativa"**. O Ministério do Desenvolvimento Social e Combate à Fome (MDS) descreve a Busca Ativa como a procura intencional realizada pela equipe de referência dos Centros de Referência da Assistência Social (CRAS), tendo como objetivo identificar as famílias em situação de vulnerabilidade e risco social. A Busca Ativa contribui para ampliar o conhecimento e a compreensão da realidade social, auxiliar no planejamento municipal, para definição de serviços socioassistenciais a serem ofertados em cada território, para a ação preventiva no território dos CRAS e principalmente levar o Estado até as famílias menos favorecidas inserindo-as nas políticas públicas adequadas [BRASIL 2009].

Do ponto de vista social, a Busca Ativa das famílias em situação de vulnerabilidade social é muito importante para que o Estado possa assistir as famílias mais necessitadas que por algum motivo não procuram de forma espontânea os serviços da assistência social.

No documento de orientações técnicas para os CRAS [BRASIL 2009], o MDS sugere algumas estratégias para realizar a Busca Ativa, tal como, deslocamento da equipe de referência para conhecimento do território, contatos com atores sociais locais (líderes comunitários, associações de bairro, etc.), obtenção de informações e dados provenientes de outros serviços socioassistenciais e setoriais, campanhas de divulgação, distribuição de panfletos, colagem de cartazes e utilização de carros de som.

Outra estratégia de realização da busca ativa proposta pelo MDS é a utilização de dados das famílias do território de atuação do CRAS. Esses dados são proveniente do Cadastro Único, de Programas Sociais e das listagens dos beneficiários do Benefício de Prestação Continuada (BPC), do Programa de Erradicação do Trabalho Infantil (PETI), do Programa Bolsa Família e daqueles que estejam em descumprimento de condicionalidades.

As recomendações de Busca Ativa propostas pelo MDS que dizem respeito ao cruzamento de informações das famílias são baseadas principalmente em informações oriundas do sistema CadÚnico e outros sistemas que acompanham famílias inseridas em programas sociais a nível federal. Apesar das atualizações e inclusões constantes do sistema federal CadÚnico, existem milhares de famílias em extrema pobreza que deveriam, mas ainda não estão incluídas em políticas e programas sociais adequados. Essa situação implica na invisibilidade dessas famílias diante dos municípios, estados e do governo federal. Ao realizar a busca ativa utilizando os dados desses sistemas, o universo de famílias alcançadas fica limitado àquelas que já foram ou estão inseridas em algum programa social. Ou seja, utilizando apenas essas fontes de informação propostas pelo MDS, a Busca Ativa será limitada.

Uma alternativa para realização da Busca Ativa pode ser a utilização de informações coletadas pelo sistema de Informatização da Rede de Serviços da Assistência Social (IRSAS), apresentado na Seção 2. Ao contrário do sistema CadÚnico, utilizado apenas nos CRAS para fins de inclusão dos cidadãos no programa Bolsa Família, o IRSAS é utilizado por todas as unidades da rede de serviços da assistência social do município, em função disto é a porta de entrada de muitas famílias que são atendidas por serviços e que não necessariamente estão inseridas no CadÚnico.

Atualmente não existe nenhuma ferramenta específica para realização da busca

ativa das famílias para os municípios. Este trabalho é o primeiro a tentar automatizar o processo da Busca Ativa através da identificação das famílias em situação de vulnerabilidade e risco social, para isso são utilizadas informações coletadas pelo sistema IRSAS (apresentado na Seção 2), aliadas a técnicas de mineração de dados (apresentadas na Seção 3). Os resultados experimentais são apresentados na Seção 4. Na Seção 5 são apresentadas as conclusões e direções futuras deste trabalho.

2. O IRSAS e a Busca Ativa

Além das estratégias anteriormente apresentadas para a Busca Ativa, o MDS também orienta que seja realizado o cruzamento de bases de dados (extraídas de alguns sistemas estaduais e federais e eventuais dados coletados pelo município) como forma de realizar a Busca Ativa. o MDS garante autonomia para os municípios utilizarem a melhor forma disponível, mas não disponibiliza nenhuma ferramenta para viabilizar esse processo [BRASIL 2013].

O IRSAS é um sistema de informação desenvolvido por uma empresa privada e implantado atualmente em três municípios brasileiros (Cascavel/PR, Londrina/PR e Mogi das Cruzes/SP). Ele funciona de forma integrada em diversas entidades do município, por exemplo, CRAS; Centro de Referência Especializado de Assistência Social (CREAS); Escolas Municipais e Estaduais; Unidades Básicas de Saúde; Conselhos Tutelares; Albergues; APAES; Entidades Sociais da Rede Governamental e não Governamental; Unidades de Acolhimento Institucional; entre outras. Atualmente no município de Cascavel/PR existem 319 unidades utilizando o IRSAS.

O IRSAS pode ser utilizado para coletar informações de uma família que não esteja inserida no CadÚnico, por exemplo, caso uma mãe procure uma organização não governamental (ONG) que atenda pessoas com deficiência auditiva para inserir seu filho em algum programa, a ONG irá cadastrar essa família no IRSAS. Caso essa família não seja encaminhada para o programa Bolsa Família, a mesma não será inserida na base de dados do CadÚnico e continuará invisível para o sistema federal. Desse modo o universo de famílias mantidas na base de dados do IRSAS é maior que as mantidas na base de dados do CadÚnico, conforme ilustrado na Figura 1. Pressupõe-se que as famílias inseridas no CadÚnico já são conhecidas pela assistência social pois precisam ser acompanhadas pela equipe técnica dos CRAS para que sejam inseridas no programa Bolsa Família.

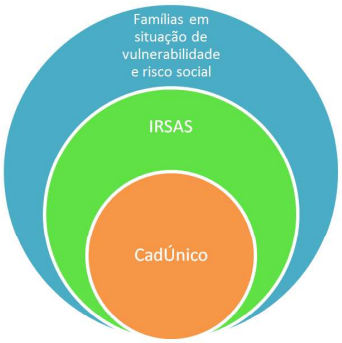

Figura 1. Visibilidade das famílias vulneráveis para os sistemas.

Tendo em vista que o principal objetivo da Busca Ativa é encontrar famílias em situação de vulnerabilidade e risco social que ainda não estejam sendo assistidas devida-

mente pelos órgãos competentes, a realização da Busca Ativa na base de dados do IRSAS possibilita encontrar mais famílias nessa situação do que quando realizada apenas com os dados do CadÚnico e outros sistemas federais e estaduais.

Não existe hoje nenhum sistema disponível para realizar a Busca Ativa das famílias em situação de vulnerabilidade e risco social. Contudo, existe no sistema IRSAS um formulário chamado "Avaliação de Vulnerabilidade e Risco Social" que realiza a avaliação da situação de vulnerabilidade e risco social das famílias em acompanhamento pelos técnicos dos CRAS. Esse formulário é uma ferramenta não automatizada para geração do índice de vulnerabilidade e risco social das famílias. O índice é gerado após o preenchimento de um formulário de avaliação de vulnerabilidade que utiliza dados do cadastro da família e diversas perguntas específicas para realizar o cálculo do índice de vulnerabilidade. Devido ao processo não ser automatizado e demandar a coleta de várias informações sobre a família, a geração da avaliação de vulnerabilidade é feita apenas com uma pequena parte das famílias que já estão em acompanhamento pelos CRAS. Desse modo, atualmente a avaliação de vulnerabilidade não auxilia no processo de Busca Ativa, pois não abrange a geração de classificação para as famílias que não estão sendo ativamente assistidas pela rede de serviços da assistência social do município.

Apesar da Avaliação de Vulnerabilidade e Risco Social do IRSAS ser um instrumento para identificar o nível de vulnerabilidade e risco social das famílias, ele não é obrigatório para que as famílias sejam acompanhadas pelos técnicos do CRAS. Devido a limitações de tempo, recursos humanos e logísticos (transporte, condução) para levar os técnicos até a residência das famílias para coleta dos dados, o formulário de avaliação de vulnerabilidade e risco social do IRSAS acaba não sendo amplamente utilizado.

O formulário de avaliação de vulnerabilidade e risco social foi desenvolvido para medir o nível de vulnerabilidade e risco social das famílias através de diversas questões que possuem uma pontuação vinculada a suas respostas de modo que ao final do preenchimento desse formulário é gerado um índice de vulnerabilidade e risco social da família. Esse índice varia de 0 (zero) a 100 (cem) e gera uma classificação categórica de acordo com a pontuação obtida que pode ser "Baixa" (0 à 15 pontos), "Média" (16 à 30 pontos) ou "Alta" vulnerabilidade (acima de 31 pontos).

O formulário é dividido em duas seções, a primeira seção é referente à avaliação de vulnerabilidade e a segunda seção é referente ao risco social.

A primeira seção do formulário de avaliação de vulnerabilidade e risco social do IRSAS corresponde à avaliação de vulnerabilidade, e é composta por seis partes. A primeira parte avalia os dados de identificação do responsável (raça, sexo, idade, deficiência) e também identifica se algum outro membro da família possui algum tipo de deficiência. A segunda parte avalia as condições habitacionais da família (tipo de logradouro, número de cômodos, condição de moradia, etc.). A terceira parte avalia o acesso ao conhecimento, escolarização do responsável e se existem membros da família em defasagem escolar. A quarta parte avalia as condições de saúde da família (doenças crônicas, pessoas acamadas, gravidez na adolescência, etc.). A quinta parte avalia algumas condições gerais da família (responsável monoparental, se crianças de 0 a 12 anos ficam sozinhas no domicílio, descumprimento de condicionalidades dos programas sociais, etc.). A sexta parte avalia situações de acesso à profissionalização, trabalho e renda da família (ocupação dos

membros, qualificação para o mercado de trabalho, média das despesas fixas e rendas da família). No final é apresentada a pontuação obtida e a classificação (alta, média ou baixa vulnerabilidade).

A segunda seção do formulário (avaliação de risco social) é composta por duas partes. A primeira parte avalia situações de violação de direitos dos membros da família (casos de violências). A segunda parte avalia a situação de membros em cumprimento de medidas judiciais (medidas socioeducativas e situações de acolhimento institucional). No final são apresentados os resultados obtidos através das respostas das questões, é apresentada a pontuação obtida e a classificação (alto, médio ou baixo risco social).

Atualmente a avaliação de vulnerabilidade e risco social do IRSAS tem sido utilizada em cenários como: identificar, dentre as famílias em acompanhamento, quais delas necessitam de maior atenção naquele determinado momento; acompanhamento da evolução da situação de vulnerabilidade e risco social das famílias em acompanhamento; critério de classificação para participação em programas municipais (Ex.: Vale Gás), onde as famílias que obtiverem o maior índice de vulnerabilidade possuem prioridade na participação dos programas; como forma de identificar a vulnerabilidade dos territórios da cidade através do cruzamento dos resultados das avaliações de vulnerabilidade com os endereços das famílias.

Porém é importante ressaltar que a classificação de vulnerabilidade e risco social está disponível apenas para uma pequena parte das famílias cadastradas no IRSAS (348 famílias de 43.016 no município de Cascavel/PR). Apesar de ser um instrumento importante para o acompanhamento da situação de vulnerabilidade do público alvo da assistência social, a Avaliação de Vulnerabilidade e Risco Social é utilizada de forma específica para um pequeno grupo de famílias atendidas deixando de lado uma grande quantidade de cidadãos que possivelmente possam estar em situação de maior vulnerabilidade e risco social do que aqueles que estão sendo acompanhados.

Com o atual cenário em vista, o problema desse trabalho é como realizar a classificação de vulnerabilidade e risco social para 100% das famílias cadastradas no IRSAS utilizando apenas as informações já existentes no cadastro das famílias sem que seja necessário realizar o preenchimento do formulário de Avaliação de Vulnerabilidade e Risco Social existente no IRSAS. Com isso, será possível identificar o nível de vulnerabilidade e risco social de todas as famílias e consequentemente encontrar aquelas com maior índice de vulnerabilidade.

Desse modo, será possível desenvolver uma técnica capaz de apoiar a Busca Ativa das famílias em situação de vulnerabilidade e risco social com base nas informações comuns existentes no cadastro de todas as famílias inseridas no IRSAS sem que seja necessário realizar a visita e o preenchimento do formulário de avaliação de vulnerabilidade e risco social.

3. Protocolo Experimental

3.1. Algoritmos Utilizados

Neste trabalho foram utilizamos dois algoritmos de classificação probabilísticos: o Naive Bayes (NB) [Dougherty 2013] e o *Averaged One-Dependence Estimators* (AODE) [Webb et al. 2005]. O NB é um classificador probabilístico que é baseado na aplicação

do Teorema de Bayes onde existe a suposição de que todos os atributos são independentes entre si [Dougherty 2013]. O desempenho do NB é um tanto surpreendente dado que a suposição de independência entre os atributos é quase sempre irreal [Friedman et al. 1997]. O AODE é um classificador probabilístico que utiliza uma suposição de independência de atributo mais fraca do que o NB, melhorando a acurácia de previsão sem gerar custo computacional indevido. Para manter a eficiência, AODE é restrito ao uso exclusivo de 1 estimador de dependência [Flores et al. 2011].

Foram utilizados os classificadores Bayesianos em função de eles gerarem uma estrutura de interdependência entre os atributos que pode ser útil para avaliar a situação de vulnerabilidade e risco social pelas assistentes sociais, uma vez que esses métodos tornam mais fácil para os usuários compreender a lógica do processo de classificação. Além disto, os classificadores Bayesianos podem ser utilizados para prever qual a probabilidade de uma determinada família estar em situação de vulnerabilidade e risco social o que auxiliaria no processo de busca ativa mesmo se o classificador atingir uma alta acurácia na rotulação das amostras.

3.2. Medidas de Desempenho

O desempenho de um algoritmo de classificação é tipicamente mensurado através da matriz de confusão como a apresentada na Figura 2 (para um problema de 2 classes). A matriz de confusão apresenta os resultados esperados e os obtidos de uma classificação [Kohavi and Provost 1998]. Possui tamanho *LxL*, onde *L* é o número de valores de classes diferentes.

	Predito C+	**Predito C-**
C+	**VP**	**FN**
C-	**FP**	**VN**

Figura 2. Matriz de confusão para duas classes.

Na matriz de confusão, **VP** é o número de exemplos positivos da classe que foram previstos corretamente (Verdadeiros Positivos), **FN** é o número de exemplos negativos da classe que foram previstos incorretamente (Falsos Negativos), **FP** é o número de exemplos positivos da classe que foram classificados incorretamente (Falsos Positivos) e **VN** é o número de exemplos negativos da classe que foram previstos corretamente (Verdadeiros Negativos).

Acurácia é a medida de desempenho geralmente associada com algoritmos de aprendizagem de máquina e é definida como, a porcentagem de amostras positivas e negativas classificadas corretamente sobre a soma de amostras positivas e negativas, Acurácia = *(VP + VN) / (VP + FN + FP + VN)*. No contexto de conjunto de dados balanceado e custo de erro equalizado, é sensato utilizar a taxa de erro como métrica de desempenho. Taxa de erro é 1 - Acurácia. Na presença de conjunto de dados desbalanceados e custo de erro desigual, é mais apropriado utilizar a curva ROC ou outra técnica semelhante [Bowyer et al. 2002].

Na curva ROC o eixo X representa *FP = FP / (VN + FP)* e eixo Y representa *VP = VP / (VP + FN)*. O ponto ideal na curva ROC seria (0,100), que significa que todos os exemplos positivos foram classificados corretamente e nenhum exemplo negativo foi

classificado erroneamente como positivo. A área abaixo da curva ROC (AUC) é uma medida útil para desempenho de classificador de modo que essa medida é independente do critério de seleção e probabilidade a priori [Bowyer et al. 2002].

3.3. Estimadores por amostragem

Existem diversas técnicas de validação por estimadores por amostragem e um dos mais utilizados é a validação cruzada. Em *r-fold cross-validation*, os exemplos são aleatoriamente divididos em *r* partições mutuamente exclusivas chamadas de *folds*. Essas partições possuem tamanho aproximadamente igual a *n/r* amostras. Os exemplos nas partições *(r-1)* são usados para treinamento e a hipótese induzida é testada na partição remanescente. Este processo é repetido *r* vezes, cada vez considerando uma partição diferente para teste. O erro deste estimador é a média dos erros calculados em cada um dos *r folds* [Monard and Baranaukas 2003].

4. Experimentos realizados

A fim de averiguar possíveis soluções para o problema, foram realizados seis experimentos preliminares com os dados rotulados das famílias existentes na base de dados do IRSAS. Para realização dos experimentos foram combinados diferentes atributos, a partir das informações das famílias do município de Cascavel/PR existentes na base de dados do sistema IRSAS resultando em seis conjuntos diferentes de dados. Para cada conjunto de dados foram aplicados dois algoritmos de classificação (Naive Bayes e AODE) na busca de obter a classificação de vulnerabilidade das famílias.

Como o propósito é identificar o nível de vulnerabilidade de 100% das famílias inseridas no IRSAS, foram escolhidos atributos comuns a todas as famílias cadastradas (cadastro básico da família) e descartadas as informações exclusivamente oriundas do formulário de avaliação de vulnerabilidade e risco social, já que apenas uma pequena parte das famílias (0,8%) possuem essas informações disponíveis. A Tabela 1 apresenta a distribuição das amostras utilizadas nos experimentos 1 ao 4 e a Tabela 2 as amostras utilizadas nos experimentos 5 e 6.

Tabela 1. Distribuição das amostras utilizadas nos experimentos 1 ao 4.

Classificação	Quantidade	%
Alta vulnerabilidade	36	10,34
Media vulnerabilidade	306	87,93
Baixa vulnerabilidade	06	1,72

Tabela 2. Distribuição das amostras utilizadas nos experimentos 5 e 6.

Classificação	Quantidade	%
Alta vulnerabilidade	36	50
Media vulnerabilidade	36	50

4.1. Experimento 1 - Utilizando todos os atributos da família

No primeiro experimento foram utilizados diversos atributos da família (256 atributos), sendo esses: atributos comuns de todos os membros da família, atributos do responsável familiar e atributos de todos os dependentes.

A análise dos resultados da Tabela 3 mostra que os algoritmos obtiveram um suposto bom desempenho, visto que os mesmos tiveram altas taxas de acertos. Porém esses resultados não são bons, visto que ao analisar as matrizes de confusão geradas pelos algoritmos (Figura 3), pode-se perceber que os mesmos estão sempre prevendo a classe majoritária, resultando em um valor baixo da medida AUC.

Tabela 3. Resultados do experimento 1 utilizando validação cruzada fator 10.

NaiveBayes			AODE		
Acerto	**Erro**	**AUC**	**Acerto**	**Erro**	**AUC**
87.931%	12.069%	0.636%	87.931%	12.069%	0.627%

Matriz de Confusão – Exp. 1 - Naive Bayes

a	b	C	←	Classificado como
302	0	4	|	a = Média Vulnerabilidade
6	0	0	|	b = Baixa Vulnerabilidade
32	0	4	|	c = Alta Vulnerabilidade

Matriz de Confusão – Exp. 1 - AODE

a	b	C	←	Classificado como
306	0	0	|	a = Média Vulnerabilidade
6	0	0	|	b = Baixa Vulnerabilidade
36	0	0	|	c = Alta Vulnerabilidade

Figura 3. Matrizes de confusão do experimento 1.

Dessa forma, neste primeiro experimento percebeu-se que utilizar apenas os atributos comuns a todas as famílias não é suficiente para resolver o problema em questão.

4.2. Experimento 2 - Utilizando apenas os atributos "chaves" definidos por um especialista do domínio

Este segundo experimento foi realizado com a finalidade de eliminar possíveis atributos redundantes no conjunto dados em relação ao experimento anterior descrito na Seção 4.1. Com base no conhecimento prévio de um dos autores desse trabalho que participou das discussões para criação do formulário de avaliação de vulnerabilidade e risco social do IRSAS, foram selecionados apenas os atributos "chaves" da família (12 atributos).

Os resultados foram bastante similares ao experimento descrito na Seção 4.1. Em uma primeira análise os mesmos também apresentaram taxas de acertos consideravelmente boas como pode ser observado na Tabela 4. Porém, ao analisar as matrizes de confusão dos classificadores (Figura 4), pode-se constatar que novamente a alta taxa de acerto está atrelada ao fato de todos os classificadores sempre preverem a classe majoritária, resultando em um valor baixo da medida AUC.

Com este segundo experimento concluiu-se que mesmo reduzindo a quantidade de atributos no conjunto de dados os algoritmos mantiveram suas taxas de acerto, porém os resultados continuaram não sendo bons em função dos modelos preverem quase sempre a classe majoritária.

Tabela 4. Resultados do experimento 2 utilizando validação cruzada fator 10.

NaiveBayes			AODE		
Acerto	**Erro**	**AUC**	**Acerto**	**Erro**	**AUC**
86.206%	13.793%	0.634%	86.494%	13.505%	0.636%

Matriz de Confusão – Exp. 2 - Naive Bayes				
a	b	C	←	Classificado como
298	1	7	|	a = Média Vulnerabilidade
6	0	0	|	b = Baixa Vulnerabilidade
32	2	2	|	c = Alta Vulnerabilidade

Matriz de Confusão – Exp. 2 - AODE				
a	b	C	←	Classificado como
300	1	5	|	a = Média Vulnerabilidade
6	0	0	|	b = Baixa Vulnerabilidade
35	0	1	|	c = Alta Vulnerabilidade

Figura 4. Matrizes de confusão do experimento 2.

4.3. Experimento 3 - Usando apenas os atributos do formulário de avaliação de vulnerabilidade

Neste terceiro experimento realizou-se uma modificação no conjunto de dados com a finalidade de verificar se utilizando apenas os atributos que são comuns para todas as famílias e ao mesmo tempo também foram utilizados no formulário de avaliação de vulnerabilidade e risco social (12 atributos), é possível realizar a classificação para 100% das famílias.

Utilizando o conjunto de dados com atributos contidos no formulário de avaliação de vulnerabilidade e risco social, obteve-se uma melhora na taxa de acerto para o classificador AODE em relação ao experimento 2 descrito na Seção 4.2. Entretanto, mesmo com um aumento na taxa de acerto os resultados (Tabela 5) obtidos não foram bons porque os modelos continuaram a prever quase sempre a classe majoritária como pode ser observado na Figura 5.

Tabela 5. Resultado do experimento 3 utilizando validação cruzada fator 10.

NaiveBayes			AODE		
Acerto	**Erro**	**AUC**	**Acerto**	**Erro**	**AUC**
85.919%	14.080%	0.625%	86.781%	13.218%	0.633%

Matriz de Confusão – Exp. 3 - Naive Bayes				
a	b	C	←	Classificado como
297	2	7	|	a = Média Vulnerabilidade
6	0	0	|	b = Baixa Vulnerabilidade
32	2	2	|	c = Alta Vulnerabilidade

Matriz de Confusão – Exp. 3 - AODE				
a	b	C	←	Classificado como
300	1	5	|	a = Média Vulnerabilidade
6	0	0	|	b = Baixa Vulnerabilidade
34	0	2	|	c = Alta Vulnerabilidade

Figura 5. Matrizes de confusão do experimento 3.

Novamente os resultados obtidos se mostraram aquém do esperado visto que mesmo utilizando informações do formulário de avaliação de vulnerabilidade (que possui informações específicas sobre as famílias) não foram suficientes para treinar os classificadores.

4.4. Experimento 4 - Utilizando os Atributos do formulário de avaliação em conjunto com os atributos do Experimento 1

Esse experimento foi realizado a fim de averiguar se, utilizando as informações conjuntas do formulário de avaliação de vulnerabilidade com os atributos do conjunto de dados do experimento 1 (apresentado na Seção 4.1), é possível obter melhores resultados. Foram selecionados 259 atributos.

Os resultados obtidos (Tabela 6) elucidam que os algoritmos de classificação tiveram taxas de acertos muito semelhantes as do experimento 1. Porém analisando a matriz de confusão dos algoritmos constatou-se que o problema de quase sempre prever a classe majoritária ainda ocorre para esse novo conjunto de dados (Figura 6).

Tabela 6. Resultado do experimento 4 utilizando validação cruzada fator 10.

NaiveBayes			AODE		
Acerto	**Erro**	**AUC**	**Acerto**	**Erro**	**AUC**
87.931%	12.069%	0.637%	87.931%	12.069%	0.628%

Matriz de Confusão – Exp. 4 - Naive Bayes				
a	b	C	←	Classificado como
302	0	4	|	a = Média Vulnerabilidade
6	0	0	|	b = Baixa Vulnerabilidade
32	0	4	|	c = Alta Vulnerabilidade

Matriz de Confusão – Exp. 4 - AODE				
a	b	C	←	Classificado como
306	0	0	|	a = Média Vulnerabilidade
6	0	0	|	b = Baixa Vulnerabilidade
36	0	0	|	c = Alta Vulnerabilidade

Figura 6. Matrizes de confusão do experimento 4.

O presente experimento mostrou que mesmo combinando todos os dados das famílias com os dados comuns das famílias contidos no formulário de avaliação de vulnerabilidade não foi possível obter bons resultados na predição das classes.

Com intuito de verificar se o desbalanceamento das amostras é o causador da dificuldade para os algoritmos de aprendizado foram realizados dois novos experimentos com amostras balanceadas, que são apresentados nas seções 4.5 e 4.6.

4.5. Experimento 5 - Usando apenas os atributos do formulário de avaliação de vulnerabilidade com classes balanceadas

Neste quinto experimento pretende-se verificar se o problema dos algoritmos sempre preverem a classe majoritária também acontece quando a quantidade de amostras são balanceadas. Neste experimento, foram separadas aleatoriamente 36 amostras da classe “Alta Vulnerabilidade” e 36 amostras da classe “Média Vulnerabilidade”. A classe “Baixa Vulnerabilidade” foi descartada por possuir um número de amostras muito baixo.

Os atributos selecionados para compor esse conjunto de dados são aqueles que são comuns para todas as famílias existentes na base de dados da aplicação e ao mesmo tempo também foram utilizados no formulário de avaliação de vulnerabilidade (12 atributos). Após o balanceamento o conjunto de dados passou a ter 72 amostras.

Os resultados apresentados na Tabela 7 apresentaram uma piora em relação à taxa de acerto. Porém ao analisar a matriz de confusão (Figura 7), pode-se perceber que os algoritmos não estão mais classificando todas as instâncias como sendo da classe majoritária. Nesse experimento foi possível alcançar resultados razoáveis com taxa de acerto igual a 63.888% ao utilizar amostras balanceadas.

4.6. Experimento 6 - Utilizando os Atributos do formulário de avaliação de vulnerabilidade em conjunto com os atributos do Experimento 1 com classes balanceadas

O intuito desse experimento é verificar se utilizando as informações conjuntas do formulário de avaliação de vulnerabilidade com os atributos do conjunto de dados do experi-

Tabela 7. Resultados do experimento 5 utilizando validação cruzada fator 10.

NaiveBayes			AODE		
Acerto	**Erro**	**AUC**	**Acerto**	**Erro**	**AUC**
63.888%	36.111%	0.675%	61.111%	38.888%	0.657%

Matriz de Confusão – Exp. 5 - Naive Bayes			
a	b	←	Classificado como
27	09	|	a = Média Vulnerabilidade
17	19	|	b = Alta Vulnerabilidade

Matriz de Confusão – Exp. 5 - AODE			
a	b	←	Classificado como
26	10	|	a = Média Vulnerabilidade
18	18	|	b = Alta Vulnerabilidade

Figura 7. Matrizes de confusão do experimento 5.

Tabela 8. Resultados do experimento 6 utilizando validação cruzada fator 10.

NaiveBayes			AODE		
Acerto	**Erro**	**AUC**	**Acerto**	**Erro**	**AUC**
66.666%	33.333%	0.725%	66.666%	33.333%	0.726%

Matriz de Confusão – Exp. 6 - Naive Bayes			
a	b	←	Classificado como
27	9	|	a = Média Vulnerabilidade
15	21	|	b = Alta Vulnerabilidade

Matriz de Confusão – Exp. 6 - AODE			
a	b	←	Classificado como
27	9	|	a = Média Vulnerabilidade
15	21	|	b = Alta Vulnerabilidade

Figura 8. Matrizes de confusão do experimento 6.

mento 1 (apresentado na Seção 4.1), é possível obter melhores resultados em comparação ao experimento 5 (apresentado na Seção 4.5) quando a quantidade de amostras é balanceada. Para tanto, foram selecionados 259 atributos.

A análise dos resultados apresentados na Tabela 8 apresenta uma melhora em relação à taxa de classificação do experimento 5 (apresentado na Seção 4.5). Ao analisar a matriz de confusão (Figura 8), pode-se perceber que os algoritmos não estão mais classificando todas as instâncias como sendo da classe majoritária.

No presente experimento foi possível alcançar resultados razoáveis com taxa de acerto de 66.666% e AUC 0.726% ao utilizar amostras balanceadas.

5. Considerações finais e trabalhos futuros

Após analisar os resultados dos experimentos pode-se observar que os modelos utilizados sempre classificavam o registro como membro da classe majoritária em função do conjunto de dados de treinamento possuir amostras desbalanceadas. Após realizar um balanceamento simples das classes, retirando aleatoriamente amostras das classes, foram realizados experimentos adicionais e os resultados, apesar de terem apresentado uma taxa de acerto menor em comparação ao conjunto desbalanceado, apresentou um resultado positivo, pois o classificador passou a predizer mais corretamente as amostras balanceadas e a medida AUC foi melhor do que para os casos desbalanceados. Diante desses fatos conclui-se ser necessária a utilização de técnicas de balanceamento no conjunto de treinamento a fim de obter melhores resultados.

Sabe-se ainda, através do conhecimento tácito adquirido por um dos autores, espe-

cialista no domínio do problema, que existe uma relação de dependência entre os atributos e essa relação deve ser considerada para realizar a classificação de vulnerabilidade e risco social. Diante desse fato, acredita-se que a utilização de modelos de classificação Bayesianos que minimizem a suposição de independência condicional entre os atributos possa obter resultados melhores do que os obtidos até o momento.

Também é importante que o algoritmo utilizado para a classificação possa não apenas classificar a amostra em um dos rótulos de vulnerabilidade, mas que também seja capaz de aferir o grau de probabilidade da família estar em situação de vulnerabilidade. Uma vez que o intuito final da solução é identificar quais famílias precisam ser prioritariamente assistidas. Desse modo, ao invés de obter as famílias rotuladas em uma das classes de vulnerabilidade, será possível obter uma estimativa de probabilidade da família ser vulnerável e consequentemente uma lista das famílias em situação de maior vulnerabilidade. Com essa estimativa será possível fornecer uma lista ordenada de prioridade de atendimento com base na probabilidade de vulnerabilidade, apoiando o processo da busca ativa. Essa lista ordenada das famílias torna-se necessária, uma vez que o município pode não possuir capacidade suficiente de atendimento para atender todas as famílias vulneráveis simultaneamente.

Referências

Bowyer, K. W., Chawla, N. V., Hall, L. O., and Kegelmeyer, W. P. (2002). Smote: Synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research*, 16:321–357.

BRASIL (2009). *Ministério do Desenvolvimento Social e Combate à Fome. Orientações Técnicas: Centro de Referência de Assistência Social - CRAS*. Brasília.

BRASIL (2013). Busca ativa: O que é a busca ativa do plano brasil sem miséria. Disponivel em: `http://www.mds.gov.br/falemds/perguntas-frequentes/superacao-da-extrema-pobreza%20/plano-brasil-sem-miseria-1/busca-ativa/`.

Dougherty, G. (2013). *Pattern Recognition and Classification: An Introduction*. Springer.

Flores, M. J., Gámez, J. A., Martínez, A. M., and Puerta, J. M. (2011). Handling numeric attributes when comparing bayesian network classifiers: Does the discretization method matter? *Applied Intelligence*, 34(3):372–385.

Friedman, N., Geiger, D., and Goldszmidt, M. (1997). Bayesian network classifiers. *Machine Learning*, 29(2-3):131–163.

Kohavi, R. and Provost, F. (1998). Glossary of terms. *Machine Learning*, 30:271–74.

Monard, M. C. and Baranaukas, J. A. (2003). *Sistemas Inteligentes: Fundamentos e Aplicações*, chapter Conceitos Sobre Aprendizado de Máquina, pages 89–114.

Webb, G. I., Boughton, J. R., and Wang, Z. (2005). Not so naive bayes: Aggregating one-dependence estimators. *Machine Learning*, 58(1):5–24.