4

Classificadores Binários, Políticas Públicas Sociais e Dados Desbalanceados

Cinara de Jesus Santos[1], **Vítor Gabriel Barra Souza**[2], **Victor Teixeira de Melo Mayrink**[3], **Henrique Steinherz Hippert** [4], **Marcel de Toledo Vieira**[5]

Resumo

Neste estudo utilizamos uma base de dados de pesquisa vinculada ao desempenho do Programa Bolsa Família (PBF) no ano de 2009. Este programa implica na transferência direta de renda com condicionantes nas áreas de educação, saúde e assistência social, visando atender famílias pobres e extremamente pobres - assim classificadas segundo um determinado valor percapita mensal. Esta base contém informações de cunho financeiro (renda e gastos das famílias), e também grau de instrução dos indivíduos, e elementos descritores do ambiente domiciliar (moradia e entorno). A aplicação dos algoritmos de predição visou averiguar a eficiência desses processos a partir das variáveis que descrevem as famílias, identificando corretamente se estas atendiam ou não ao perfil de beneficiárias do programa. Os algoritmos utilizados foram regressão logística, árvore binária de decisão e rede neural artificial em múltiplas camadas. Diversas medidas de desempenho foram calculadas, a partir da matriz de confusão resultante de cada algoritmo. Os valores encontrados para estas medidas foram baixos frente a uma das classes a serem identificadas. As intervenções aplicadas foram o reembaralhamento aleatório e também super-amostragem da classe minoritária e sub-amostragem da classe majoritária. Embora tenha ocorrido alguma melhora, o desempenho no reconhecimento da classe minoritária permaneceu baixo o que aponta para a necessidade de novos experimentos.

Palavras-chave: Árvores de decisão. Programa Bolsa Família. Classificador. Predição. Regressão logística. Redes neurais.

Abstract

In this study we used a research database focused on the performance of the Bolsa Família Program (PBF) in 2009. In this program, direct income transfer is carried out under conditions of education, health and social assistance, aiming to serve poor and extremely poor families - Thus classified according to a certain monthly per capita value. This database contains financial information (income and family expenses), as well as the degree of education of the individuals, and elements describing the home environment (housing and environment). Applying prediction algorithms, the objective was to ascertain the efficiency of these processes from the variables that describe the families, correctly identifying if they met the beneficiaries profile of the program or not. The algorithms used were logistic regression, binary decision tree and artificial neural network in multiple layers. Several performance measures were calculated from the confusion matrix resulting from each algorithm. The values found for these measures were low compared to one of the classes to be identified. The interventions applied were random re-marshaling and also super-sampling of the minority class and sub-sampling of the majority class. Although some improvement occurred, the performance in the minority class recognition remained low, which indicates the need for new experiments.

Keywords: Decision tree. Bolsa Família Program. Classifier. Prediction. Logistic regression. Artificial neural networks.

[1] UFJF, Programa de Pós-graduação em Modelagem Computacional, E-mail: cinara.dcc@gmail.com

[2] UFJF, Faculdade de Medicina, E-mail: vitor31415@gmail.com

[3] UFJF, Programa de Pós-graduação em Modelagem Computacional, E-mail: mayrink.vtm@gmail.com

[4] UFJF, Departamanto de Estatística/ICE, E-mail: henrique.hippert@ufjf.edu.br

[5] UFJF, Departamanto de Estatística/ICE, E-mail: marcel.vieira@ice.ufjf.br

1 Introdução

Classificadores são separadores de classes que organizam os dados em grupos de casos que apresentem características semelhantes. Estes têm sido usados sobre bases de dados descritoras de ações de assistência social, para explicitar os efeitos destas sobre seu público alvo. Estas bases de dados, em geral, não têm amostras de mesmo tamanho no que diz respeito às classes que as compõem. Quando isto ocorre, dizemos que os dados são desbalanceados.

Este estudo utiliza uma base de dados do Ministério do Desenvolvimento Social e Combate a Fome (MDS), contendo informações sobre beneficiários do Programa Bolsa Família (PBF) - descritores do ambiente domiciliar, grau de instrução dos moradores do domicílio, uso de serviços de saúde pelos mesmos, e informações de cunho financeiro (renda e despesas das famílias). O estudo não visa avaliar o PBF, mas sim o comportamento de classificadores frente a bases de caráter social, pois estas apresentam certas particularidades, como por exemplo, o desbalanceamento das amostras.

Testamos três algoritmos classificadores - regressão logística, árvore binária de decisão e rede neural artificial em múltiplas camadas. O desempenho destes algoritmos foi medido a partir de métricas decorrentes da matriz de confusão para as duas classes objeto de estudo – beneficiários e não beneficiários. Estas medidas avaliam o desempenho do algoritmo em cada grupo - sensibilidade, que representa o índice de acertos da classe de beneficiários, e especificidade, que representa o desempenho na identificação dos não-beneficiários. Como desempenho global foram levantados os valores da acurácia e, média aritmética e média geométrica do acerto dos dois grupos cujos conceitos serão apresentados mais adiante. Como os erros e acertos de uma classe não são complementares dos da outra classe, é importante que ambas sejam corretamente identificadas.

Este artigo é elaborado a partir de trabalho apresentado na 7ª Conferência Sul em Modelagem Computacional (SANTOS *et al.*, 2016). Nesta versão, o texto foi reescrito e ampliado. Novas simulações foram acrescentadas, usando técnicas de reamostragem (sub-amostragem e super-amostragem aleatórias), que visam corrigir o problema de desbalanceamento do banco de dados usado. Em consequência, a seção de Resultados foi reestruturada. Esta versão do artigo inclui também o acréscimo de uma curta apresentação sobre o programa Bolsa-Família, no "Apêndice A", ao final do texto.

2 Literatura disponível

A informação necessária para uma familiarização com a base de dados e com as propostas do programa social que esta descreve pode ser obtida a partir do "Sumário Executivo" que acompanha a base de dados (MDS, 2012), e do Caderno de Estudos nº16, de 2014, do MDS (JANUZZI; QUIROGA, 2014). A base de dados em questão é resultado de pesquisa encomendada pelo MDS, realizada no ano de 2009 com o objetivo de avaliar os efeitos do programa na vida de famílias inicialmente entrevistadas em 2005 e então revisitadas em 2009.

Composta por vários arquivos em separado, a base de dados é bastante extensa e semelhante ao recenseamento quanto à abrangência dos descritores, representados por um grande número de variáveis (1.529 variáveis, após agrupamento dos dados em único arquivo). A base inclui 56.367 indivíduos distribuídos em 11.372 domicílios de 269 municípios em 23 estados da Federação e do Distrito Federal (não constam representantes dos estados do Acre, Roraima

e Tocantins). Dentre as informações coletadas (MDS, 2012), está o grau de instrução dos moradores, informações sobre a saúde dos componentes daquela família (cartão de vacina e grau de nutrição das crianças, se há doentes crônicos ou em outro tipo de tratamento no domicílio), itens de consumo da casa (nas áreas de alimentação, manutenção, entretenimento), condições de saneamento da moradia e do entorno.

Em sua grande maioria, estudos realizados a partir de base de dados relacionada ao PBF pertencem à área de econometria e se baseiam em modelos de regressão a fim medir a influência da ação social sobre a vida dos beneficiários em aspectos ligados a saúde, progressão escolar, capacitação dos indivíduos (empregabilidade) (AMARAL *et al.*, 2012) ou segurança alimentar (MUARETTO *et al.*, 2015). Conforme mencionado por Duarte (2009), programas de transferência direta de renda propõem três frentes: prevenção, enfrentamento e suavização da pobreza, assim como desestimular o trabalho infantil.

Duarte (2009) avaliou o gasto com alimentos em famílias rurais comparando dois grupos, ambos com perfil de beneficiários do PBF, porém, um de fato é contemplado e o outro, apesar de corresponder ao perfil do programa, ainda está na fila de espera, por limitações financeiras de ordem burocrática (MUARETTO *et al.*, 2015). O grupo não-beneficiário foi considerado como grupo de controle enquanto o grupo beneficiário representa o grupo de tratamento, onde se pretende verificar o efeito do programa. As análises foram feitas usando modelo de regressão logística para as estimativas do método denominado "*propensity score*" (probabilidade de seleção) que permite a redução da quantidade das variáveis independentes empregadas no estudo fazendo-se uma comparação das características observáveis de ambos os grupos atentando para o fato de que os indivíduos de cada grupo (beneficiário vs. não beneficiário) necessitam ter características semelhantes para que possam ser comparados.

Gusmão (2012) analisou os efeitos do programa nos municípios de São Gotardo e de Capelinha, ambos em Minas Gerais, no ano de 2009, a partir de indicadores de educação, saúde, renda e emprego, também lançando mão da regressão logística sobre dados qualitativos utilizando os mesmos dois grupos – ambos com perfil do programa, mas um deles na fila de espera, a fim de averiguar o benefício do PBF favoreceu a qualidade de vida das famílias beneficiárias.

Usando dados do Censo 2010 do Instituto Brasileiro de Geografia e Estatística (IBGE), Amaral (2012) avaliou o desempenho escolar da criança vinculando não só ao fato da família ser beneficiária como também a presença da mãe junto à criança. Este estudo também se deu por modelos de regressão logística que estimaram as chances de crianças não estarem na escola, em diferentes limites de renda domiciliares.

O objetivo do estudo aqui apresentado não é a avaliação do Programa Social, mas o comportamento de classificadores frente à base de dados onde há o desbalanceamento e onde as respostas armazenadas nas variáveis não apresentam valores discrepantes se confrontadas as classes que representam.

3 Materiais e métodos

Diferente dos trabalhos mencionados que buscavam desenhar os efeitos do PBF, o objetivo deste trabalho é averiguar o desempenho de classificadores quando aplicados sobre dados de natureza categórica e quantitativa quanto ao reconhecimento do grupo com perfil do programa versus grupo

que não atende ao perfil (não devendo, portanto ser enquadrado como beneficiário).

O primeiro passo foi a realização de uma análise exploratória dos dados, a fim de conhecer os tipos de informações ali contidas, reduzir sua dimensionalidade, e identificar os métodos possíveis a serem aplicados para realização do estudo. Dos algoritmos classificadores foram eleitos as redes neurais artificiais (RNA-MLP), a regressão logística e a árvore binária de decisão. Estes algoritmos devem buscar reconhecer os que recebem o benefício (variável dependente codificada por uma variável *dummy* de valor igual a "1") e os que não o recebem o benefício (variável dependente codificada por "0").

3.1 Base de dados

Os dados foram disponibilizados em diversos arquivos, dos quais alguns traziam as informações referentes a cada indivíduo, outros traziam informações sobre cada domicílio. Originalmente, estes arquivos apresentam as famílias ali cadastradas separadas em três grupos (MDS, 2012):

- famílias beneficiárias do Programa (30%);
- famílias na espera (se enquadram no programa, estão inscritas no CADÚnico – 60%)[6];
- famílias não cadastradas no CADÚnico, pois não possuem as características necessárias a categoria de beneficiários (10%).

Durante a análise exploratória, tomamos contato com as características dos dados e suas anomalias (como dados faltantes, duplicados ou duvidosos), optando por vezes pelo descarte de variáveis assim. As informações na forma categórica necessitaram passar por um tratamento a fim de se tornarem indicadoras, ou seja, atuarem como um fator. Esta ação teve peso considerável nas informações referentes aos indivíduos - como no quesito "*escolaridade*", por exemplo - já que o benefício se dá por domicílio e se fazia necessário que os dados fossem todos assim representados.

Enquanto nos trabalhos onde se buscava averiguar o desempenho do PBF havia a necessidade de semelhança na descrição entre os grupos comparados, o estudo aqui proposto apresenta a necessidade de não semelhança entre os grupos. Assim, o grupo que se encontrava na fila de espera do benefício foi desconsiderado visto que possuem características de beneficiários mas não recebem.

Os dados foram reorganizados em um único arquivo onde cada caso descreve uma família. A amostra passou a contar com 3.254 casos onde o grupo de "famílias beneficiárias" corresponde a 75,38% dos casos, e as "não-beneficiárias" a 24,62%. Os dados são portanto não-balanceados, visto que um grupo é consideravelmente maior que o outro. Isso pode gerar um viés, que tende a negligenciar a classificação do grupo minoritário. É importante, no entanto, que um classificador reconheça ambos os grupos primando pela minimização dos erros.

3.1.1 Escolha das variáveis

A regressão logística foi ferramenta utilizada para a seleção e ordenação das variáveis de interesse, e foi também um dos classificadores empregados no estudo. Este método caracteriza-se pelo uso de um conjunto de variáveis independentes preditoras (também chamadas explanatórias), podendo ser numéricas, categóricas ou ambas, para predizer a ocorrência de um

[6] Estas famílias tem o perfil para participar do programa mas, mediante limitações de recurso destinado ao programa, precisam aguardar o próximo repasse (MUARETTO *et al.*, 2015).

determinado evento representado pela variável dependente binária (representada por "0" ou "1", "falso" ou "verdadeiro", "não" ou "sim"). Como seletor de variáveis elimina aquelas que se mostrarem redundantes ou pouco informativas.

Por se tratar de um número elevado de variáveis, a análise preliminar não considerou todas as variáveis dos arquivos originais. Dentre as mais de 1.000 variáveis originais, selecionamos inicialmente 100 variáveis por serem consideradas as mais relevantes, com base em conhecimentos prévios da literatura (AMARAL et al., 2012; DUARTE, SAMPAIO, SAMPAIO, 2009; e GUSMÃO, TOYOSHIMA, DE PAULA, 2012). A partir dessa escolha parte das variáveis foram reagrupadas e, a partir do modelo de regressão logística na modalidade "*stepwise-forward*" variáveis foram eliminadas e o peso das restantes definido. Este procedimento começa com a escolha da variável independente "X_i" que melhor explica a variável dependente "Y". O próximo passo é escolher uma segunda variável que se mostre mais significativa que a primeira quando adicionada ao modelo. A partir do momento em que a segunda variável entra no modelo, verifica-se a permanência da primeira. Caso permaneça, uma terceira variável é selecionada. Se uma terceira variável entra no modelo, verifica-se a continuidade das duas primeiras no modelo. Novamente, experimenta-se a inclusão de uma nova variável. Caso entre, tenta-se eliminar uma das que já estão no modelo. O procedimento acaba quando não se consegue nem adicionar, nem eliminar variáveis. Esta verificação se dá pelo Critério de Informação de Akaike (AIC), que leva em consideração tanto a complexidade do modelo (definida pelo número de variáveis independentes) quanto o erro de classificação. Quanto menor o seu valor, melhor o modelo encontrado. A partir do momento em que o índice AIC deixou de variar significativamente dá-se por encerrada a seleção de variáveis.

Para definir o ponto de corte da regressão logística, usamos análise da curva ROC (*Receiver Operating Characteristic*) e a menor distância entre sensibilidade (classificação correta dos beneficiários) e especificidade (classificação correta dos não beneficiários), onde o reconhecimento correto do grupo de beneficiários foi associado ao conceito de verdadeiro positivo (VP) e o reconhecimento correto do grupo de não-beneficiários foi associado ao conceito de verdadeiro negativo (VN).

As variáveis eleitas descrevem as seguintes informações:

- quantidade de pessoas no domicílio;
- região geográfica;
- localização do domicílio (urbana ou rural);
- tipo de rua onde se localiza o domicílio (calçada, asfaltada, outro);
- condição de ocupação do domicílio (próprio, alugado, cedido,...);
- material predominante nas paredes externas;
- material predominante no telhado (cobertura externa);
- tipo de escoadouro do banheiro ou sanitário;
- existência de água canalizada dentro do domicílio;
- grau de instrução mais alto entre os membros da família;
- gastos em mensalidades escolares nos últimos 30 dias;
- gastos com saúde para indivíduos até 14 anos;

- gastos com saúde para indivíduos com 15 anos ou mais;
- rendimento percapita (de qualquer fonte que não seja de benefício social);
- gastos com transporte e comunicação;
- gastos com moradia e reformas, mobília, eletrodomésticos e outros artigos para o lar, manutenção da casa;
- gastos com vestuário, higiene pessoal, lazer;
- gastos com alimentos comprados (por domicílio);
- quantos automóveis possui;

3.2 Algoritmos utilizados

Após a definição das variáveis de interesse pela regressão logística (32 eleitas), as observações foram assim divididas para o emprego dos classificadores:

- 70% dos casos para ajuste dos modelos;
- 30% dos casos para teste.

Realizamos 50 rodadas com cada um dos algoritmos, embaralhando os elementos da amostra a cada rodada. Ao final foi tomado o valor médio das métricas de avaliação a partir das matrizes de confusão de cada rodada.

3.2.1 Regressão logística (RL)

O modelo de regressão logística como classificador (GUJARATI; POTTER, 2000) é utilizado quando a variável resposta é binária, com dois resultados possíveis (sim/não, verdadeiro/falso,...). A variável dependente a ser tratada (*é beneficiário?*) assume o valor de "*0*" ou "*1*", que corresponde a "não" e "sim", em função de um conjunto de variáveis independentes. Esta função varia no intervalo de "*0*" a "*1*", segundo uma curva frequentemente chamada de "sigmoidal" ou "curva-S". A fig. 1 mostra um exemplo desta curva.

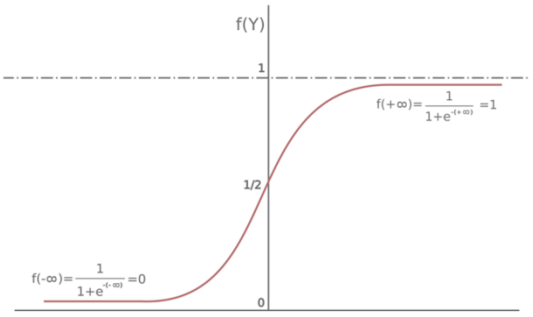

Figura 1 - Regressão Logística (adaptado de FÁVERO *et al*., 2009)

A função que a define é dada por (FÁVERO *et al.*, 2009):

$$f(Y) = \frac{1}{1+e^{-(Y)}} \qquad (1)$$

onde

$$Y = ln\left(\frac{p}{1-p}\right)$$

$$= \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_n X_n \qquad (2)$$

Y é a variável dependente, *p* a probabilidade de ocorrência de evento de interesse, X_i são as variáveis independentes e β_i seus respectivos parâmetros.

De outra forma podemos dizer que a probabilidade da variável dependente *Y* ser igual a "*1*" é condicionada às variáveis explicativas X_i, na forma:

$$P(1) = f(Y = 1|X_1, X_2, \ldots, X_n) = \\ = \frac{1}{1+e^{-(\beta_0+\sum \beta_i X_i)}} \qquad (3)$$

É preciso escolher o ponto de corte na variável *Y* que separa as duas categoria de saída. Não necessariamente este ponto precisa ser "equidistante" de cada classe (no caso, "½", como sinalizado na Figura 1). Esta escolha foi feita com base na análise da curva ROC (*Receiver Operating Characteristic*) e

da menor distância entre a sensibilidade e a especificidade do modelo. O valor escolhido de corte (*cut-off*) foi de 0,739.

3.2.2 Árvore binária de decisão (ABD)

Uma árvore de decisão binária consiste em um conjunto de nós e folhas, onde os nós representam pontos de decisão, e as folhas, as opções disponíveis (cada classe) (MONARD; BARANAUSKAS, 2003). Cada folha pode se tornar um novo nó, até que a separação desejada seja alcançada ou o algoritmo atinja outra condição de parada.

Figura 2 - esquema da árvore de decisão (*elaborado pelos autores*)

A árvore de decisão exibe os resultados de forma hierárquica já que o atributo mais importante é representado na árvore como o primeiro nó, e os atributos menos relevantes são representados nos nós subsequentes.

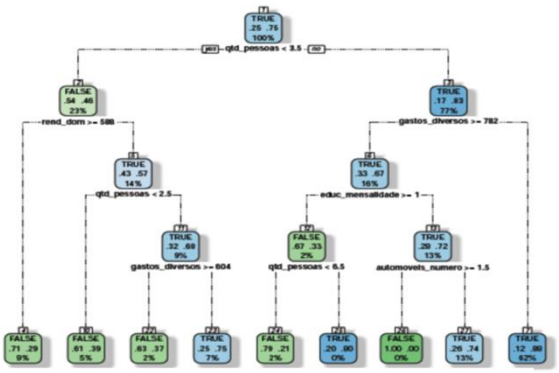

Figura 3 - arvore de decisão gerada pelo código (*elaborado pelos autores*)

Assim, o caminho percorrido na árvore (da raiz à folha de interesse) representa uma regra de classificação. O sucesso desse método se deve ao fato de ser uma técnica extremamente simples, que não requer ajuste de parâmetros, e geralmente alcança um bom índice de acertos. Porém, é necessária uma análise detalhada dos dados de entrada, para garantir bons resultados. A "Figura 3" mostra a árvore de decisão gerada neste estudo.

3.2.3 Redes neurais artificiais (RNA)

Redes Neurais Artificiais (RNA) são algoritmos computacionais (SILVA *et al.*, 2015) que usam um modelo inspirado na estrutura neural de organismos inteligentes, que adquirem conhecimento através da experiência. Seu componente básico é o "neurônio artificial" (fig. 4).

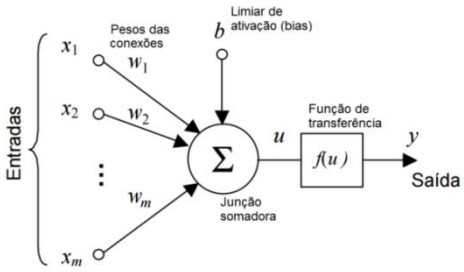

Figura 4 - Representação esquemática de um neurônio artificial *(VON ZUBEN, 2009)*

O neurônio faz primeiro a combinação linear dos dados de entrada (vetor *X*), ponderados pelos pesos *w* associados a cada conexão; o valor obtido é então usado como argumento de uma função não-linear, a função de *ativação* ou de *transferência*, que produz a saída do neurônio, dada na eq. 4:

$$y = f(u) = f(\sum w_i x_i + b) \quad (4)$$

Numa RNA de várias camadas (*Multi-Layer Perceptron*, MLP), como representada na Fig. 5, os neurônios estão organizados em uma ou mais camadas. A saída dos neurônios de cada camada é utilizada como entrada dos neurônios da camada seguinte; esta saída não pode ser usada por neurônios da mesma camada, ou de camadas anteriores (como ocorre em outros tipos de RNA).

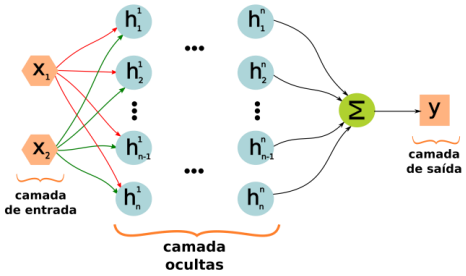

Figura 5 - Esquema de uma RNA-MLP (*elaborado pelos autores*)

A um determinado padrão de entrada *X* corresponde um sinal desejado de saída *d*; no processo de "treinamento" da rede, os pesos *w* são ajustados de forma a aproximar a função que relaciona *d* e *X*. Num problema de classificação, como o do presente estudo, a rede deverá ter apenas um neurônio de saída, produzindo um valor binário, "*0*" ou "*1*". A RNA é uma técnica que necessita de um considerável número de execuções durante o período de treinamento além de uma boa quantidade de elementos, a fim de executar ajustes finos de vários hiper-parâmetros (por exemplo, o número de neurônios, ou o número de camadas), e não permite interpretar claramente a relação entre a entrada e a saída. Sua capacidade de generalização diante de informações incompletas é o que a torna um método bastante utilizado.

3.3 Medição do desempenho dos classificadores

Conforme descritas por Matos (2009), as métricas utilizadas para medir o desempenho de cada algoritmo (sensibilidade, especificidade, acurácia, eficiência, média geométrica e MCC) foram calculadas a partir da "matriz de confusão" (Fig. 5). Nesta matriz, as classificações corretas estão registradas nas células da diagonal principal, e as classificações incorretas nas demais células. Um resultado é chamado "verdadeiro positivo" quando o algoritmo classifica como positivo um caso que de fato é positivo (no caso, reconhece de fato elemento pertencente ao grupo de "beneficiários") e "verdadeiro negativo", quando classifica como negativo um caso que de fato é negativo (reconhece corretamente elemento do grupo de "não-beneficiários"). Um resultado "falso positivo" é aquele em que o algoritmo classificou como positivo um caso que na verdade é negativo ("não beneficiários" classificado como "beneficiário") e "falso negativo", a situação contrária.

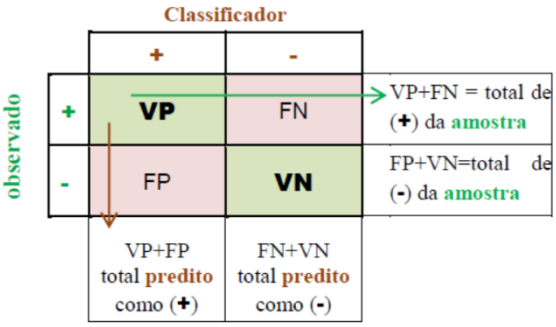

Figura 6 - Matriz de Confusão (*elaborado pelos autores*)

Segue uma breve apresentação dos conceitos das métricas escolhidas para este estudo, e suas respectivas fórmulas (ALBERTO, 2012):

3.3.1 Sensibilidade

Proporção de verdadeiros positivos; ou seja, classificações corretas das famílias beneficiárias.

$$\text{sensibilidade} = \frac{\text{verdadeiro positivo}}{\text{total de positivos}} = \frac{\text{VP}}{\text{VP+FN}} \quad (5)$$

3.3.2 Especificidade [3]

Proporção de verdadeiros negativos; ou seja, classificações corretas das famílias não-beneficiárias.

$$\text{especificidade} = \frac{\text{verdadeiro negativo}}{\text{total de negativos}} = \frac{\text{VN}}{\text{VN+FP}} \quad (6)$$

3.3.3 Acurácia

Proporção de classificações corretas, tanto de casos positivos quanto negativos. Em caso de dados desbalanceados esta medida pode induzir a uma conclusão errônea quanto ao desempenho do algoritmo empregado visto que a classe majoritária encobrirá a falha frente a classe minoritária.

$$\text{Acurácia} = \frac{\text{total de acertos}}{\text{total de dados}} = \frac{\text{VP+VN}}{(\text{VP+FN})+(\text{VN+FP})} \quad (7)$$

3.3.4 Eficiência

É a média aritmética entre sensibilidade e especificidade. Figura como opção de medida onde a quantidade de elementos de cada grupo não se mostra próxima.

$$\text{eficiência} = \frac{\text{sensibilidade} + \text{especificidade}}{2} = \left(\frac{\text{VP}}{\text{VP+FN}} + \frac{\text{VN}}{\text{VN+FP}}\right) * \frac{1}{2} \quad (8)$$

3.3.5 Média geométrica

Ou "*g-mean*", corresponde à média geométrica entre as taxas de verdadeiros positivos (sensibilidade) e verdadeiros negativos (especificidade). Mede o desempenho equilibrado de um classificador em relação às taxas de acertos de ambas as classes, quando o desempenho de ambas as classes é importante (CASTRO; BRAGA, 2011).

$$\text{g} - \text{mean} = \sqrt{\text{sensibilidade} * \text{especificidade}} = \sqrt{\frac{\text{VP}}{\text{VP+FN}} * \frac{\text{VN}}{\text{VN+FP}}} \quad (9)$$

3.3.6 Coeficiente de correlação de Matthews (MCC)

A sigla MCC vem da denominação na língua inglesa – "*Matthews Correlation Coefficient*". Será experimentada como alternativa de medida de desempenho padrão frente a acurácia. Também chamada de "coeficiente Φ", é uma medida de qualidade de classificações binárias que pode ser usada mesmo quando os grupos possuem tamanhos bastante distintos. Retorna um valor no intervalo fechado [-1,+1]: o valor "*+1*" indica uma classificação perfeita, o valor "*0*" indica uma classificação equivalente a que seria feita aleatoriamente, e o valor "*-1*" mostra total discordância entre a classificação feita e observado, conforme ilustrado na Figura 7. Pode ser calculado a partir da matriz de confusão, pela fórmula:

$$\text{MCC} = \frac{(\text{VP}*\text{VN}) - (\text{FP}*\text{FN})}{\sqrt{(\text{VP+FP})(\text{VP+FN})(\text{VN+FP})(\text{VN+FN})}} \quad (10)$$

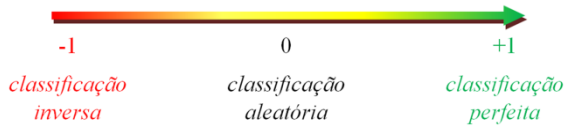

Figura 7 - escala do coeficiente de correlação de Matthews (*elaborada pelos autores*)

3.3.7 Coeficiente de informação de Akaike (AIC)

É uma medida de qualidade de um modelo estatístico que leva em conta tanto a complexidade do modelo (medida pelo número de seus parâmetros) quanto qualidade do ajuste (medida pela verossimilhança). Foi utilizada na escolhas das variáveis do modelo. É definido pela seguinte fórmula:

$$AIC=2*k- 2*ln(L) \quad (11)$$

onde *k* é o número de parâmetros do modelo estatístico e *L* é o valor máximo da função de probabilidade para o modelo estimado. Dado

um conjunto de modelos de candidatos para um problema, o modelo preferido é o que tem o valor mínimo da AIC. Portanto AIC não só premia a qualidade do ajuste, mas também inclui uma penalidade, que é uma função crescente do número de parâmetros estimados. Esta penalidade tem como propósito inibir o super-ajuste (*overfitting*) do modelo, isto é, o aumento excessivo do número de parâmetros livres no modelo (o que tende a melhorar a qualidade do ajuste, independentemente do número de parâmetros livres no processo de geração de dados).

4 Resultados

Da matriz de confusão levantamos a capacidade de classificação de cada método aplicando as métricas anteriormente descritas, a fim de eleger a que melhor represente o desempenho do classificador como um todo.

O resultado inicial pode ser observado na *Tabela 1*, onde a amostra utilizada não sofreu intervenção quanto ao tamanho dos grupos. A RL se mostrou bastante equilibrada (74% - 77%) ao passo que a ABD e a RNA apresentam uma distância considerável entre sensibilidade e especificidade.

Na escolha da métrica global, o MCC se mostrou bastante restritivo – o que aponta para uma forte influência das perdas na classificação de cada grupo (sensibilidade e especificidade) ao passo que acurácia e eficiência se mostram demasiado otimistas frente ao baixo valor de especificidade para a ABD e a RNA. Tais valores podem levar a uma interpretação errônea já que um classificador não pode ser considerado de bom desempenho global se falha em uma das classificações de grupo considerando que ambas são importantes.

Tabela 1 - Medidas de desempenho sem intervenção na representatividade dos grupos

	ABD	**RL**	**RNA**
Sensibilidade	0,9241	0,7414	**0,9405**
Especificidade	0,4360	**0,7658**	0,2510
Acurácia	**0,8037**	0,7083	0,7692
Eficiência	0,6800	**0,7516**	0,5957
G-mean	0,6324	**0,7340**	0,4538
MCC	**0,4285**	0,2484	0,2491

Havendo distância tão acentuada entre a identificação de beneficiários (sensibilidade) e não beneficiários (especificidade), optou-se pela reamostragem dos grupos.

Sub-amostrando a classe majoritária (beneficiários) deixando que a mesma deixe de ser três vezes maior que não beneficiários e passe a ser apenas 10% superior em representatividade, observamos uma perda na sensibilidade para ABD e RNA frente a um baixo ganho na especificidade (vide *Tabela 2*).

Tabela 2 - medidas de desempenho, sub-amostragem de beneficiários

	ABD	**RL**	**RNA**
Sensibilidade	**0,7980**	0,7569	0,7174
Especificidade	0,6076	**0,6929**	0,5839
Acurácia	0,7060	**0,7268**	0,6539
Eficiência	0,7028	**0,7250**	0,6506
G-mean	0,6945	**0,7233**	0,6478
MCC	**0,4522**	**0,4522**	0,3061

(elaborada pelos autores)

Frente a queda da sensibilidade o próximo experimento foi duplicar a classe minoritária. Esta permaneceu menor que beneficiários, mas a distância entre cada grupo quanto a representatividade caiu de 3:1 para 3:2.

Permaneceu a perda na sensibilidade para ABD e RNA, ainda que menor, e pouco ganho na especificidade, conforme apresentado na *Tabela 3*.

Tabela 3 - medidas de desempenho, super-amostragem de não beneficiários

	ABD	RL	RNA
Sensibilidade	**0,8653**	0,7479	0,8433
Especificidade	0,5539	**0,7241**	0,5505
Acurácia	**0,7432**	0,7385	0,7281
Eficiência	0,7096	**0,7360**	0,6969
G-mean	0,6914	**0,7355**	0,6587
MCC	**0,4649**	**0,4649**	0,4063

(elaborada pelos autores)

Intervindo agora no tamanho de ambos os grupos, a *Tabela 4* mostra ainda queda na sensibilidade da ABD e RNA (a maior das três intervenções) porém, maior ganho na especificidade.

Tabela 4 - medidas de desempenho, super-amostragem e sub-amostragem

	ABD	RL	RNA
Sensibilidade	**0,7572**	0,7508	0,6930
Especificidade	0,6505	**0,7149**	0,6742
Acurácia	0,7038	**0,7330**	0,6832
Eficiência	0,7038	**0,7329**	0,6836
G-mean	0,7012	**0,7320**	0,6908
MCC	**0,4667**	**0,4667**	0,3685

(elaborada pelos autores)

Para a escolha da métrica global, observando na Tabela 1, onde há uma diferença considerável entre sensibilidade e especificidade para ADB e RNA, acurácia e eficiência induzem a uma interpretação demasiado otimista frente ao baixo grau de acerto da classe minoritária (especificidade). Dadas as intervenções de representatividade (sub-amostragem e super-amostragem), especificidade e sensibilidade se aproximam, apontando para uma classificação mais equilibrada (ainda deficiente) permitindo que acurácia e eficiência não destoem da classificação individual de cada grupo, permitindo, principalmente na regressão logística que se tome como métrica global qualquer uma das candidatas: acurácia, eficiência ou média geométrica. Em todos os cenários MCC se mostrou extremamente punitivo, não contribuindo como métrica global. Convém então, frente a situação de desequilíbrio, na Tabela 1, observar que a média geométrica permitiu uma melhor percepção quanto ao desempenho do classificador considerando o acerto de ambas as classes. E por isso, figura como melhor opção para métrica global.

Conforme pode ser visto nas tabelas de 1 a 4, os valores observados no desempenho do classificador para a classe minoritária, apesar do ganho, mostram-se baixos, mesmo com as intervenções de representatividade. As tabelas 5 e 6 mostram as variações da sensibilidade e especificidade após cada intervenção.

Tabela 5 - variação da sensibilidade após intervenção na amostra

	ABD	RL	RNA
sub-amostr.	-13,65%	2,09%	-23,72%
super-amostr.	-6,36%	0,88%	-10,33%
ambos	-18,06%	1,27%	-26,32%

(elaborada pelos autores)

Tabela 6 - variação da especificidade após intervenção na amostra

	ABD	RL	RNA
sub-amostr.	39,36%	-9,52%	132,63%
super-amostr.	27,04%	-5,45%	119,32%

ambos	49,20%	-6,65%	168,61%

(elaborada pelos autores)

5 Conclusões

Regras de classificação não estão livres de erro - o que se espera é que um bom procedimento de classificação resulte em poucas classificações errôneas. Os tamanhos diferentes dos dois grupos utilizados neste trabalho nos levam a um problema conhecido, em aprendizado de máquina e mineração de dados, como de "classes desbalanceadas". Este tipo de problema traz dificuldades para a classificação, uma vez que os algoritmos tradicionais normalmente funcionam bem quando as distribuições são equilibradas.

Conforme visto, o desbalanceamento pode induzir a interpretações errôneas dado as elevadas taxas de acurácia, conforme observado na Tabela 1, que confronta os valores apresentados nas tabelas de 2 a 4, onde o desequilíbrio entre as medidas de desempenho de cada grupo não se mostram tão distantes. Ainda assim a identificação de exemplos pertencentes a grupos minoritários fica prejudicada (demonstrado pelos baixos valores encontrados para especificidade).

Desse modo observa-se que apenas a reamostragem simples, realizada de forma aleatória, não trouxe o ganho esperado permanecendo a classe minoritária ainda negligenciada. Comprova-se então que não há um algoritmo único capaz de atender a classificação de ambos os grupos. É importante conhecer o alcance e as limitações de diferentes classificadores e/ou associação dos mesmos. A peculiaridade de cada nicho precisa ser percebida para que o estudo permita apreender conceitos e chegar a um modelo que propicie desempenho razoável ao experimento ou a identificação do fator que impede o sucesso do estudo.

Outra hipótese está na natureza da construção da base de dados utilizada, que não visava a classificação das famílias, mas o acompanhamento de possíveis mudanças na situação das famílias acompanhadas. Neste caso, as informações disponíveis não são suficientes para classificar entre beneficiário ou não, mas para acompanhar uma determinada ação apenas. Outras técnicas, tanto de pré-processamento quanto de classificação podem ser experimentadas a fim de afastar esta dúvida. Assim, para trabalhos futuros pretende-se testar outras técnicas voltadas para dados desbalanceados, tanto no pré-processamento como na codificação do algoritmo classificador (ou associação de mais de um classificador), com o intuito de encontrar adaptações capazes de evitar viés para as classes majoritárias. Dentre as técnicas citadas na literatura utilizada como referência, conforme Batista, Prati e Monard (2004, 2008); e ainda Ramezankhani et al. (2016) e Schiavoni (2015), algumas das opções figuram na introdução de custos de classificação incorreta – que traz o desafio de encontrar os valores de tais custos. Ainda sugerem utilizar formas de reamostragem como o *undersampling* (redução do número de casos da classe majoritária, apesar do risco de acarretar em perda de informação) ou o *oversampling* (replicação de casos da classe minoritária embora possa resultar em *overfitting*) por outros métodos que não uma permutação aleatória, impondo pesos nas escolhas – a dificuldade reside em determinar estes pesos. Castro e Braga (2011) e Batista, Prati e Monard (2004) citam como alternativas de intervenção: Links de Tomek, *Edited Nearest Neighbor Rule* (ENN), Método *Boundary Elimination and Domination Algorithm* (BED), Máquina de Vetor Suporte (*Support Vector Machine*), e ainda algoritmos genéticos.

Apreender o conceito destas ou outras técnicas pode resultar em melhor

desempenho do conjunto amostrado e melhor entendimento de suas peculiaridades.

Referências

ALBERTO, B. L. A. Abordagens de pré-processamento de dados em problemas de classificação com classes desbalanceadas. 2012. Dissertação de Mestrado. Centro Federal de Educação Tecnológica de Minas Gerais.

AMARAL, E.F.L; GONÇALVES, G.Q.; MONTEIRO, V.P.; et al. Avaliação de impactos das condicionalidades de educação do Programa Bolsa Família: uma análise com o censo de 2010. in XVIII Encontro Nacional de Estudos Populacionais, ABEP, Águas de Lindóia/SP–Brasil. 2012.

ANDRADE A.L.S.S.; ZICKER F. Avaliação de testes diagnósticos. In: Andrade A.L.S.S. & Zicker F. (Eds), Métodos de Investigação Epidemiológica em Doenças Transmissíveis. Vol.1. 1997. p.9-30. FNS, OPAS, Brasília, DF.

BARANAUSKAS, J. A. Aprendizado de Máquina Conceitos e Definições. 2007. Notas de aula. Disponível em [http://dcm.ffclrp.usp.br/~augusto/teaching/ami/AM-I-Conceitos-Adicionais-Metricas.pdf]. Acesso em ago. 2016

BATISTA, G. E. A. P. A; PRATI, R. C.; MONARD, M. C. A study of the behavior of several methods for balancing machine learning training data. ACM Sigkdd Explorations Newsletter, v. 6, n. 1, p. 20-29, 2004.

PRATI, R. C.; BATISTA, GEAPA; MONARD, M. C. Curvas ROC para avaliação de classificadores. Revista IEEE América Latina, v. 6, n. 2, p. 215-222, 2008.

BISHOP, C. M. Neural networks for pattern recognition. Oxford University Press, 1995.

BRASIL. Ministério do Desenvolvimento Social e Combate à Fome; Centro de Desenvolvimento e Planejamento Regional. Sumário executivo – avaliação de impacto do Programa Bolsa Família – 2ª Rodada. Brasília, DF: SAGI; IFPRI/Datamétrica Consultoria, Pesquisa e Telemarketing Ltda. 2012.

CAMILO, C. O.; SILVA, J. C. da. Mineração de dados: Conceitos, tarefas, métodos e ferramentas.2009. Universidade Federal de Goiás.

CARVALHO, F. A.T. Aprendizagem Estatística de Dados. 2010. Notas de aula. Disponível em [http://www.cin.ufpe.br/~fatc/AM/AvaliacaoClassificadores.pdf]. Acesso em out.2016

CASTRO, CL de; BRAGA, A. P. Aprendizado supervisionado com conjuntos de dados desbalanceados. Rev. Controle Autom, v. 22, n. 5, p. 441-466, 2011.

CASTRO, L. N.; VON ZUBEN, F. J. Redes Neurais Artificiais. (Notas de aula). Disponível em [ftp://ftp.dca.fee.unicamp.br/pub/docs/vonzuben/ia006_03/topico5_03.pdf]. Acesso em ago. 2016

SANTOS,A.M.; SEIXAS,J.M.; PEREIRA, B.B.; et al. Usando redes neurais artificiais e regressão logística na predição da hepatite A. Rev. Bras. Epidemiol., v. 8, n. 2, p. 117-26, 2005.

DUARTE, G. B.; SAMPAIO, B.; SAMPAIO, Y. Programa Bolsa Família: impacto das transferências sobre os gastos com alimentos em famílias rurais. Revista de economia e sociologia rural, v. 47, n. 4, p. 903-918. 2009

FÁVERO,L.P.L; BELFIORE,P.P.; SILVA, F.L.; et al. Análise de dados: modelagem multivariada para tomada de decisão. São Paulo, Brasil: Campus. 2009.

GONZAGA, A. Métodos de avaliação de Classificadores. 2011. Notas de aula. Disponível em [http://iris.sel.eesc.usp.br/sel 886/Aula_9.pdf]. Acesso em dez.2016

GUJARATI, D. N. Econometria Básica. 3.ed. São Paulo, Brasil: Pearson Makron Books. 2000

GUSMÃO, G. C.; TOYOSHIMA, S. H. ; PAULA, R. Avaliação do Programa Bolsa Família: um estudo de caso no estado de Minas Gerais no ano de 2009. Vozes dos Vales, v. 01, p. 01-31, 2012.

JANNUZZI, P. de M., QUIROGA, J. Síntese das pesquisas de avaliação de programas sociais do MDS. Cadernos de Estudos Desenvolvimento Social em Debate 2011-2014.v. 16. p.1-358. 2014

LIMA, L. M. C. Modelagem de distribuição geográfica para Hydromedusa maximiliani (Mikan, 1820) (Testudines, Chelidae). 2014. Dissertação de Mestrado. Universidade Federal de Juiz de Fora.

MATOS, P. F. LOMBARDI, L. O., CIFERRI, R. R. et al. Relatório técnico "métricas de avaliação". Universidade Federal de São Carlos. 2009. Disponível em [http://gbd. dc. ufscar. br/~pablofmatos/files /ReportMetrica-MatosEtAl.pdf], Acesso em out. 2016.

Ministério do Desenvolvimento Social e Combate à Fome, 2012, Sumário Executivo – Avaliação de Impacto do Programa Bolsa Família – 2ª Rodada. Brasília.

MONARD, M. C. ; BARANAUSKAS, J. A. . Conceitos sobre Aprendizado de Máquina. In: Solange Oliveira Rezende. (Org.). Sistemas Inteligentes - Fundamentos e Aplicações. 1ª ed. Barueri - SP: Editora Manole Ltda., 2003, v. , p. 89-114.

MONARD, M. C. ; BARANAUSKAS, J. A. . Indução de Regras e árvores de Decisão. In: Solange Oliveira Rezende. (Org.). Sistemas Inteligentes - Fundamentos e Aplicações. 1ª ed. Barueri - SP: Editora Manole Ltda., 2003, v. , p. 115-139.

MUNARETTO, L. F., SILVA, J. F., VIANNA, P. H. et al. Um estudo sobre Programa Bolsa Família (PBF): o caso dos municípios que integram a associação dos municípios da zona da produção (AMZOP). In. Anais do IV SINGEP - São Paulo/SP – Brasil. 2015

NETO, Si. B.; NAGANO, M. S.; DA COSTA MORAES, M. B. Utilização de redes neurais artificiais para avaliação socioeconômica: uma aplicação em cooperativas. Revista de Administração da Universidade de São Paulo, v. 41, n. 1, 2006.

OLIVEIRA, S. R. de M. Medidas para Avaliação de Regras e de Modelos de Classificação (Notas de aula). Disponível em [http://www.ime.unicamp.br/~wanders on/Aulas/MT803-Aula10-AprendizadoMaquina-Interestingness.pdf]. Acesso em: nov. 2016

PRATI, R. C.; BATISTA, G.; MONARD, M. C. Curvas ROC para avaliação de classificadores. Revista IEEE América Latina, v. 6, n. 2, pp. 215-222. 2008.

PRETTO, D.; BENDER FILHO, R. Análise da influência dos programas complementares para a emancipação sustentada dos beneficiários vinculados ao programa bolsa família: estudo com ex-beneficiários do município de Santo Ângelo/RS. 2016. Gestão Pública: Práticas e Desafios-ISSN: 2177-1243, v. 8, n. 2.

PRINCIPE, J. C.; EULIANO, N. R.; LEFEBVRE, W. C. Neural and adaptive systems: fundamentals through simulations with CD-ROM. John Wiley & Sons, Inc., 1999.

RAMEZANKHANI, A.; POURNIK, O.; SHAHRABI, F. et al. The impact of

oversampling with SMOTE on the performance of 3 classifiers in prediction of type 2 diabetes. Medical decision making, v. 36, n. 1, p. 137-144, 2016.

RUFINO, H. L. P. Algoritmo de aprendizado supervisionado-baseado em máquinas de vetores de suporte - uma contribuição para o reconhecimento de dados desbalanceados. 2011. Tese de Doutorado. Universidade Federal de Uberlândia - MG

SCHIAVONI, André Spinelli. Um estudo comparativo de métodos para balanceamento do conjunto de treinamento em aprendizado de redes neurais artificiais. 2010. Monografia de Graduação. Universidade Federal de Lavras – MG.

SENNA, M. C. M; BRANDÃO, A. A.; DALT, S. Programa Bolsa Família e o acompanhamento das condicionalidades na área de saúde. Serviço Social & Sociedade, n. 125, p. 148-166, 2016.

SILVA, C. C. S.; VIANNA, R.P.T.; MORAES, R. M. et al. Rede neural artificial e o modelo de apoio à decisão em segurança alimentar nutricional. in Revista de enfermagem UFPE *on-line*, Recife, 9(3):7078-85. 2015. Disponível em [http://www.revista.ufpe.br/revistaenfermagem/index.php/revista/article/download/6317/pdf_7386]. Acesso em 20 de ago. 2016

SOARES, S.; SÁTYRO, N. O Programa Bolsa Família: desenho institucional, impactos e possibilidades futuras. 2009.

SOUZA, F. C. S. Inteligência computacional aplicada na análise e recuperação de portfólios de créditos do tipo non-performing loans. 2015. Dissertação de Mestrado. Universidade Nove de Julho – SP.

SOUZA, F. C. S. de. Métricas de avaliação de modelos de classificação/predição. 2014. Disponível em [https://mineracaodedados.wordpress.com/tag/matriz-de-confusao/]. Acesso em ago. 2016

Wikipedia. Youden's J statistic. Wikipedia, The Free Encyclopedia. Last edition: 22 Nov. 2016. Disponível em [https://en.wikipedia.org/wiki/Youden's_J_statistic]. Acesso em dez. 2016

SANTOS, C. J.; SOUZA, V. G. B.; MAYRINK, V. T. M. et al. Classificadores Binários como Critério de Averiguação em Políticas Públicas. In: **VII Conferência Sul em Modelagem Computacional, 2016.** Rio Grande/RS. 2016. Anais do 7o. MCSul - Conferência Sul em Modelagem Computacional. Rio Grande - RS: FURG, 2016. p. 718-726.

A Programa Bolsa Família

O Programa Bolsa Família (PBF) foi criado em 2003 (MDS, 2012), a partir da lei nº 10.836 de janeiro de 2004, e integrado a outras políticas sociais preexistentes (Programas Fome Zero, Bolsa Escola, Bolsa Alimentação e Auxílio-Gás), com o objetivo de assistir domicílios em situação de pobreza e extrema pobreza. Estes patamares são definidos por valores mensais percapita percebidos pela família. Está no nível de pobreza quem apresenta valor percapita entre R$ 170,00 e R$ 85,01 e em nível de extrema pobreza quem apresenta valor percapita de até R$ 85,00[7]. A família deve estar registrada no Cadastro Único para Programas Sociais do Governo Federal - CADÚnico (MDS, 2012), instrumento que identifica e caracteriza as famílias de baixa renda, no qual constam informações sobre a residência, escolaridade, situação de trabalho, gastos relacionados a saúde e renda de cada membro da família. O cadastro e posterior acompanhamento das

[7] Valor ajustado pelo Decreto-Lei nº 8.794, de 29 de junho de 2016, valores percapita.

famílias são de responsabilidade dos municípios. Tal cadastramento porém não significa inclusão imediata no PBF – é necessário aguardar que o sistema analise as informações contidas no CADÚnico e se verificar de fato há o enquadramento daquela família no perfil do PBF e o município não houver atingido ainda sua cota dentro do repasse recebido.

Existem variantes na composição do benefício de acordo com a presença de gestantes, lactantes e menores de 18 anos, conforme tabela A1

Tabela A 1 - Valores percebidos no PBF

DESCRIÇÃO	***VALOR*** ***limite de parcelas***
Valor básico do benefício (**somente se em situação de extrema pobreza**)	R$ 85,00
variante em caso de gestantes, nutrizes e crianças (menor de 16 anos - VAR)*	Até 5 de R$ 39,00
variante para jovens de 16 a 17 anos (BVJ)*	Até 2 de R$ 46,00
Máximo a ser recebido por uma família	(5 x 39)+(2 x 46)= = R$ 287,00 (pobre) 85+(5 x 39)+(2 x 46) = = R$ 372,00 (extremamente pobre)

**Quantidade de variantes definida pela Lei nº12.512 de 14/10/2011 (5 VAR + 2 BVJ) - valores ajustados em 2016 juntamente com o valor básico do benefício conforme Decreto-Lei nº8.794 de 29/06/2016.*