**UNIVERSIDADE FEDERAL DE JUIZ DE FORA**

**FACULDADE DE ENGENHARIA**

**PROGRAMA DE PÓS-GRADUAÇÃO EM MODELAGEM COMPUTACIONAL**

**Cinara de Jesus Santos**

# Avaliação do uso de classificadores para verificação de atendimento a critérios de seleção em programas sociais

**Juiz de Fora**

**2017**

**Cinara de Jesus Santos**

# Avaliação do uso de classificadores para verificação de atendimento a critérios de seleção em programas sociais

Dissertação apresentada ao Programa de Pós-graduação em Modelagem Computacional, da Universidade Federal de Juiz de Fora como requisito parcial a obtenção do grau de Mestre em Modelagem Computacional. Área de concentração: Métodos Numéricos Aplicados

Orientador: Prof. D.Sc. Henrique Steinherz Hippert

Co-orientador: Prof. PhD Marcel de Toledo Vieira

**Juiz de Fora**

**2017**

Ficha catalográfica elaborada através do programa de geração
automática da Biblioteca Universitária da UFJF,
com os dados fornecidos pelo(a) autor(a)

Santos, Cinara de Jesus.
Avaliação do uso de classificadores para verificação de atendimento a critérios de seleção em programas sociais / Cinara de Jesus Santos. -- 2017.
87 f.

Orientador: Henrique Steinherz Hippert
Coorientador: Marcel de Toledo Vieira
Dissertação (mestrado acadêmico) - Universidade Federal de Juiz de Fora, Faculdade de Engenharia. Programa de Pós Graduação em Modelagem Computacional, 2017.

1. algoritmos classificadores. 2. classificadores binários. 3. árvore binária de decisão. 4. regressão logística. 5. redes neurais artificiais. I. Hippert, Henrique Steinherz, orient. II. Vieira, Marcel de Toledo, coorient. III. Título.

**Cinara de Jesus Santos**

# Avaliação do uso de classificadores para verificação de atendimento a critérios de seleção em programas sociais

Dissertação apresentada ao Programa de Pós-graduação em Modelagem Computacional, da Universidade Federal de Juiz de Fora como requisito parcial a obtenção do grau de Mestre em Modelagem Computacional. Área de concentração: Métodos Numéricos Aplicados

Aprovada em 07 de março de 2017

BANCA EXAMINADORA

Prof. D.Sc. Henrique Steinherz Hippert - Orientador
Universidade Federal de Juiz de Fora

Prof. PhD Marcel de Toledo Vieira – Co.orientador
Universidade Federal de Juiz de Fora

Prof. D.Sc. Ricardo da Silva Freguglia
Universidade Federal de Juiz de Fora

Prof. D.Sc. Fabrízzio Condé de Oliveira
Universidade Salgado de Oliveira

*Dedico este trabalho a amigos e familiares que compreenderam minha ausência e amorosamente me incentivaram na continuidade de meus esforços.*

## AGRADECIMENTOS

Agradeço primeiramente a Deus por providenciar as ferramentas e suportes necessários em nossa caminhada;

Aos familiares e amigos de longa data que compreenderam minha ausência e amorosamente me incentivaram na continuidade de meus esforços;

Aos meus mentores Henrique e Marcel pela paciência, ensinamentos e dedicação que possibilitaram o desenvolvimento deste trabalho;

Ao nosso coordenador Rafael Bonfim, pela presteza e dedicação; As meninas da secretaria, Samantha, Renata, Nathália e Adriana pela prontidão em nos orientar e providenciar nossas diversas solicitações;

Aos professores do Programa pelo aprendizado adquirido e conhecimentos transmitidos;

Aos tecnicos do PPGMC por tornarem possível este trabalho;

Aos amigos que fiz nesses dois anos e que me presentearam com sua presteza intelectual assim como com sua humanidade, em especial a Daniela Schimitz, Camila, Jesuliana, Yulia, Taís Medeiros, Stephanie, Raphael Cordeiro, Rafael Veiga, Érica, Letícia, Artur, Emmanuel e Vitor Gabriel – muito obrigada pelo suporte tanto acadêmico quanto pessoal;

À CAPES pelo auxílio financeiro.

*"Se não puder voar, corra. Se não puder correr, ande. Se não puder andar, rasteje, mas continue em frente de qualquer jeito."*

*Martin Luther King*

## RESUMO

Classificadores são separadores de grupos que mediante determinadas características organiza os dados agrupando elementos que apresentem traços semelhantes, o que permite reconhecimento de padrões e identificação de elementos que não se encaixam. Esse procedimento de classificação e separação pode ser observado em processos do cotidiano como exames (clínicos ou por imagem), separadores automáticos de grãos na agroindústria, identificador de probabilidades, reconhecedores de caracteres, identificação biométrica - digital, íris, face, etc. O estudo aqui proposto utiliza uma base de dados do Ministério do Desenvolvimento Social e Combate a Fome (MDS), contendo informações sobre beneficiários do Programa Bolsa Família (PBF), onde contamos com registros descritores do ambiente domiciliar, grau de instrução dos moradores do domicílio assim como o uso de serviços de saúde pelos mesmos e informações de cunho financeiro (renda e gastos das famílias). O foco deste estudo não visa avaliar o PBF, mas o comportamento de classificadores aplicados sobre bases de caráter social, pois estas apresentam certas particularidades. Sobre as variáveis que descrevem uma família como beneficiária ou não do PBF, testamos três algoritmos classificadores - regressão logística, árvore binária de decisão e rede neural artificial em múltiplas camadas. O desempenho destes processos foi medido a partir de métricas decorrentes da chamada matriz de confusão. Como os erros e acertos de uma classe não são os complementares da outra classe é de suma importância que ambas sejam corretamente identificadas. Um desempenho satisfatório para ambas as classes em um mesmo cenário não foi alçado - a identificação do grupo minoritário apresentou baixa eficiência mesmo com reamostragem seguida de reaplicação dos três processos classificatórios escolhidos, o que aponta para a necessidade de novos experimentos.



**Palavras-chave:** Palavra-chave. Algoritmos classificadores. Predição. Regressão logística. Árvores de decisão. Redes neurais artificiais. Classificadores binários.

## ABSTRACT

Classifiers are group separators that, by means of certain characteristics, organize the data by grouping elements that present similar traits, which allows pattern recognition and the identification of elements that do not fit. Classification procedures can be used in everyday processes such as clinical or imaging exams, automatic grain separators in agribusiness, probability identifiers, character recognition, biometric identification by thumbprints, iris, face, etc. This study uses a database of the Ministry of Social Development and Fight against Hunger (MDS), containing information on beneficiaries of the Bolsa Família Program (PBF). The data describe the home environment, the level of education of the residents of the household, their use of public health services, and some financial information (income and expenses of families). The focus of this study is not to evaluate the PBF, but to analyze the performance of the classifiers when applied to bases of social character, since these have certain peculiarities. We have tested three classification algorithms - logistic regression, binary decision trees and artificial neural networks. The performance of these algorithms was measured by metrics computed from the so-called confusion matrix. As the probabilities of right and wrong classifications of a class are not complementary, it is of the utmost importance that both are correctly identified. A good evaluation could not be archive for both classes in a same scenario was not raised - the identification of the minority group showed low efficiency even with resampling followed by reapplication of the three classificatory processes chosen, which points to the need for new experiments.



**Keywords:** Classification algorithms. Prediction. Logistic regression. Decision trees. Artificial neural networks. Binary classifiers.

## LISTA DE ABREVIATURAS E SIGLAS

| | |
|---|---|
| AIBF | Análise de Impacto do Programa Bolsa Família |
| AIC | critério de informação de Akaike |
| AUC | área sob a curva |
| BVE | benefício variável de caráter extraordinário |
| CADÚnico | Cadastro Único para Programas Sociais do Governo Federal |
| CEF | Caixa Econômica Federal |
| IBGE | Instituto Brasileiro de Geografia e Estatística |
| MCC | Matthew's correlation coefficient |
| MDS | Ministério do Desenvolvimento Social e Combate a Fome |
| MLP | multilayer perceptron |
| PBF | Programa Bolsa Família |
| RL | regressão logística |
| RNA | rede neural artificial |
| ROC | Receiver Operating Characteristic |
| SIGPBF | Sistema de Gestão do Programa Bolsa Família |
| SMOTE | Synthetic Minority Oversampling Technique |

SUMÁRIO

# LISTA DE ILUSTRAÇÕES

## LISTA DE QUADROS

## LISTA DE TABELAS

# 1 Introdução

Classificadores são separadores de grupos que, a partir de determinadas características, organizam os dados agrupando elementos que apresentem traços semelhantes – o que permite reconhecimento de padrões e identificação de elementos que não pertencem aos grupos. Exemplos do uso de classificadores podem ser encontrados em vários processos do cotidiano, como nos resultados de exames clínicos (por imagem ou não), classificadores de textos, sistemas reconhecedores de caracteres, sistemas de biometria, sistemas de classificação automática na indústria, sistemas de reconhecimento de áreas por fotos de satélite entre outros. Como ferramenta de apoio ao planejamento, classificadores são empregados no processo de filtragem de informações relevantes produzindo indicadores de probabilidade.

O estudo aqui proposto utiliza uma base de dados adquirida de pesquisa prévia encomendada pelo Ministério do Desenvolvimento Social e Combate a Fome (MDS), no ano de 2009, pesquisa essa sobre o desempenho do Programa Bolsa Família (PBF), que realiza transferência direta de renda com condicionantes nas áreas de educação, saúde e assistência social, visando beneficiar famílias pobres e extremamente pobres. Esta base de dados inclui registros descritores do ambiente domiciliar, do grau de instrução dos moradores do domicílio, do uso de serviços de saúde, e de informações de cunho financeiro (renda e gastos das famílias), conforme descrito no sumário executivo que acompanha os dados [1], disponíveis no site do MDS. Uma rápida explanação sobre o PBF pode vista no *"apêndice A"*.

Neste âmbito, temos então duas classes a serem identificadas – a de beneficiários e a de não beneficiários do Programa. Como a base de dados é muito extensa (mais de 1.500 variáveis), convém que sejam eleitas parte destas variáveis para a execução do estudo. Isto porque nem todas as variáveis contribuem igualmente para o modelo do classificador - algumas contribuem pouco ou muito pouco. Esta ação resulta em um modelo o mais simples possível sem comprometer sua função de representar o caso em estudo.

Baseando-se na literatura disponível inicialmente foram escolhidas 60 variáveis (listadas nos quadros D.1 e D.2 - apêndice D) que sofreram transformações resultando então em 34 (processo descrito no item 3.1.3), e estas submetidas a três algoritmos classificadores – regressão logística, árvore binária de decisão e rede neural artificial.

O desempenho destes algoritmos foi avaliado a partir de métricas decorrentes da matriz de confusão, que registra em suas linhas e colunas os erros e acertos da predição. Como os erros e acertos de uma classe não são informações complementares da outra, é importante que ambas as classes sejam corretamente identificadas – tanto a de *"beneficiários"* como a de *"não beneficiários"* para que se possa avaliar o comportamento de classificadores aplicados sobre bases de caráter social pois estas apresentam algumas particularidades, como o fato de serem desbalanceadas - quando uma (ou mais classes) apresenta um tamanho desproporcional (muito maior ou muito menor) em relação as demais. Nesta situação os algoritmos tradicionais geram modelos que falham no reconhecimento de classes poucos representadas (classes minoritárias). Uma vez encontrado o algoritmo classificador que permita uma correta leitura dos grupos, pode-se partir para a construção de cenários onde, dada uma coleção de informações a respeito do objeto de estudo ao longo de um período, é possível vislumbrar seu comportamento em períodos adiante - assim como o histórico hidrográfico permite o planejamento de produção de energia de uma hidrelétrica ou de um investimento agrário em uma nova cultura ou novo local de plantio. Da mesma forma também possibilitaria projeções de ações de cunho social dado o acesso a registros sobre estas ações em um dado período de tempo.

Assim, ressalta-se que não há neste estudo interesse em averiguar a eficácia do PBF - o objetivo aqui se restringe ao comportamento dos classificadores diante de dados oriundos de políticas públicas sociais visto que dados desta natureza são desbalanceados, ou seja, o propósito é identificar sua capacidade ao identificar as famílias que recebem o benefício e as que não o recebem (informação esta passível de comparação a partir de registro disponível na base de dados), a fim de averiguar o desempenho desses algoritmos no que diz respeito a percepção de cada grupo e a partir daí, levantar as implicações para a determinação de cenários quando de posse de um histórico dos dados.

Nas seções a seguir procurou-se apresentar as influências a que está submetido este trabalho e seu desenvolvimento. Na seção 2 são citados outros trabalhos baseados em algoritmos classificadores aplicados sobre informações do PBF e bases de dados não equilibradas, onde se observa que um grupo (por vezes mais de um grupo) possui representatividade superior aos demais e no que isso implica.

Na seção 3 é apresentada a base de dados utilizada neste estudo, e na seção 4 as transformações aplicadas na mesma e a apresentação dos algoritmos classificadores utilizados. Assim como rápido conceito das métricas de desempenho aplicadas sobre estes classificadores.

A seguir, na seção 5, observa-se quais os valores alcançados para as métricas aplicadas, seguido da escolha da que melhor representa o desempenho frente a natureza dos dados utilizados. As conclusões estão dispostas na seção 6 .

E por fim, na seção 7, é apresentada a expectativa para continuidade do estudo a partir de sugestão encontradas nas referências aqui utilizadas para aprofundar a investigação sobre o desempenho de classificadores para dados desbalanceados.

# 2 Literatura Disponível

## 2.1 Classificadores

O uso de classificadores vem da necessidade de separação das informações a fim de facilitar decisões e sua escolha está diretamente relacionada ao tipo de dado a ser utilizado na aplicação. De acordo com a literatura disponível há diversas propostas de associação ou modificação de algoritmos ou ainda, ajustes simultâneos em partes da amostra na busca de melhores resultados, como relatado, por exemplo, das referências [2], [3], [4] e [5].

## 2.2 Trabalhos anteriores realizados com dados do Programa Bolsa Família

Em sua grande maioria, estudos realizados a partir de base de dados relacionada ao Programa Bolsa Família (PBF) pertencem a área de Econometria e geralmente usam modelos de regressão linear para analisar a qualidade de vida alcançada por meio do PBF em aspectos ligados a saúde [6], progressão escolar ([7] e [8]), capacitação dos indivíduos ([9] e [10]), ou segurança alimentar ([11] e [12]).

Conforme mencionado na referência [11], programas de transferência direta de renda buscam três frentes: prevenção, enfrentamento e suavização da pobreza; além do desestímulo do trabalho infantil. Ainda neste trabalho é avalidado 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 e burocráticas do sistema de distribuição de renda. O grupo de não-beneficiários foi considerado como grupo de controle, ou seja, como referência quanto a neutralidade, já que busca medir as mudanças trazidas pelo recebimento do PBF, enquanto o grupo de beneficiários representam o grupo de tratamento, onde houve a intervenção (recebimento do benefício) e pretende-se verificar seu efeito. As análises foram feitas usando modelo de regressão logística para as estimativas do método denomidado *"propensity score"* 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 necessitam ter características semelhantes

- através de variáveis observáveis para garantir que ambos os grupos sejam comparáveis.

Na referência [8] Gusmão 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 para dois grupos – ambos com perfil do programa mas um deles na fila de espera, afim de averiguar o quanto a entrada no PBF favoreceu a qualidade de vida das famílias beneficiárias. Essa espera se dá porque, ainda que a informação levantada sobre a família aponte para um possível beneficiário, o dinheiro não é liberado imediatamente por questões burocráticas e orçamentárias - a família contemplada precisa entrar no planejamento do próximo repasse, gerando uma fila de contemplados até o efetivo recebimento do benefício.

Usando dados do Censo 2010 do Instituto Brasileiro de Geografia e Estatística (IBGE), Amaral [7] avaliou o desempenho escolar da criança vinculando não só ao fato da família ser beneficiária como também a presença da mãe junto a criança. Este estudo também se deu por modelos logísticos que estimaram as chances de crianças não estarem na escola, em diferentes limites de renda domiciliar.

A pesquisa de Neto [10] tem enfoque socioeconômico – implementando classificação financeira e de risco de cooperativas pecuárias no estado do Paraná. A seleção das variáveis se deu pela aplicação do *"teste-T"* e de regressão logística sobre dados do ano de 1999 provenientes do Sistema de Acompanhamento de Cooperativas (SAC) do Estado do Paraná. O algoritmo classificatório utilizado foi uma rede neural artificial (RNA) do tipo mapas de auto-organização (*Self-Organizing Feature Maps* — SOFM) a fim de separar as cooperativas em diversos grupos tendo como critério suas características similares.

Silva [12] avalia segurança alimentar usando registro de 287 famílias residentes em São José dos Ramos, no interior do estado da Paraíba, com aplicação de questionário próprio, e aplicando RNA do tipo MLP feedforward.

## 2.3 Dados desbalanceados

Um conjunto de dados é tido como desbalanceado quando apresenta classes onde o numero de elementos que a compõem é muito diferente das demais. São situações como manifestação de alguma doença onde, inicialmente, o número de portadores é muito menor do que o número de sadios ou, o número de fraudes em

operações financeiras em relação as ações idôneas, ou ainda o número de pacotes perdidos em uma transmissão remota de dados. Em todos esses exemplos e em outros, queremos que o grupo minoritário seja cada vez menor e, que o grupo contrário cresça cada vez mais. Na maioria das vezes o foco do estudo depende da correta detecção dessa minoria pois é ela quem determina o custo da falha de uma ação - o custo da não eficácia de um medicamento, o custo da falha de algoritmo de segurança, o custo da falha na transmissão de dados. E portanto, define se o processo vale a pena ou não. Os algoritmos tradicionais não têm conseguido fornecer esta classificação de forma satisfatória em situação de dados desbalanceados [13].

Em sua grande maioria a classe minoritária apresenta um desempenho muito baixo, pois o grande volume de elementos da classe majoritária induzem os métodos classificadores a uma determinada conformação exatamente pelo volume de exemplos. Daí as propostas de tratamento dos dados, o pré-processamento, em uma das classes ou em ambas como reamostragem e outros procedimentos, antes de aplicar o classificador e/ou modificações nos algoritmos de classificação. A referência [13] ressalta a proposição de intervenções – aplicar aos dados reamostragem por super-amostragem na classe minoritária e/ou sub-amostragem na classe majoritária, de forma aleatória ou ponderada. Já para os classifcadores há propostas de inserção de custos diferenciados para cada classe, ou alteração de kernels, e outras técnicas. O problema reside em como levantar os valores dos parâmetros para essas intervenções. Assim, em [13] temos:

> *Na abordagem de pré-processamento de dados, o objetivo é balancear o conjunto de treinamento através de mecanismos de reamostragem de dados no espaço de entrada, que incluem sobreamostragem da classe minoritária, subamostragem da classe majoritária ou a combinação de ambas as técnicas (Japkowicz, 2000b; Laurikkala, 2001; Estabrooks et al., 2004; Batista et al., 2005).*
>
> *A sobreamostragem é baseada na replicação de exemplos preexistentes (sobreamostragem com substituição) ou na geração de dados sintéticos. No primeiro caso, a seleção de exemplos a serem replicados pode ser aleatória (sobreamostragem aleatória) ou direcionada (sobreamostragem informativa). Com relação à geração de dados sintéticos, a técnica de interpolação é comumente usada. Por exemplo, no conhecido método SMOTE (Synthetic Minority Oversampling Technique), proposto em Chawla et al. (2002), para cada exemplo positivo $x_i$ , novos exemplos artificiais são criados entre os segmentos de reta que ligam $x_i$ aos seus k vizinhos mais próximos.*

*A sub-amostragem envolve a eliminação de exemplos da classe majoritária. Os exemplos a serem eliminados podem ser escolhidos aleatoriamente (subamostragem aleatória) ou a partir de alguma informação a priori (subamostragem informativa).*

*...Apesar das técnicas de subamostragem e sobreamostragem possuírem o mesmo propósito, elas introduzem diferentes características ao novo conjunto de treinamento que podem algumas vezes, dificultar o aprendizado (Drummond and Holte, 2003; Mease et al., 2007; He and Garcia, 2009)...*

# 3 Material

Em um estudo que envolva simulações, é importante definir seu ambiente (onde o problema ocorre), e delimitar de que maneira é gerado e porque persiste (que fatores o alimentam, representando-os por variáveis) para assim criarmos uma estrutura que nos permita perceber suas implicações (quais os efeitos gerados), ou seja, é preciso definir um modelo, um sistema capaz de representar o "objeto" em estudo para vislumbrar seus efeitos.

Uma característica frequente de várias bases de dados reais é o desbalanceamento das classes [2]. Este fato pode comprometer o desempenho do algoritmo que assume a base de dados como uma distribuição equilibrada entre os grupos e por isso, o custo por uma classificação errada ser o mesmo para todas as classes, o que não é verdade – é muito mais grave um exame não acusar uma doença ou um patógeno do que alarmar um paciente sadio que na verdade não possui doença ou distúrbio algum, visto que o caso contrário pode custar até mesmo uma vida.

Muitos dos sistemas tradicionais de classificação não estão preparados para aprender conceitos que reconheçam ambas as classes com precisão sob estas condições. Como resultado obtém-se alta precisão de classificação para a classe majoritária e grande negligência no que se refere classe minoritária.

## 3.1 Tratamento dos dados

O tratamento prévio dos dados se faz necessário antes de sua utilização no estudo pretendido, qualquer que seja, para que se possa averiguar a natureza dos dados, sua distribuição e possíveis anomalias. Ainda que os dados recebidos já tenham sido utilizados em outra pesquisa, isso não garante que estejam prontos para uso imediato. De acordo com o foco do estudo faz-se necessário prepará-los – o que, por vezes implica em uma limpeza, agrupamento informações e/ou transformações de parte das informações e, talvez, selecionar parte de seus descritores principalmente quando estes se mostram bastante extensos (na base de dados em questão, são mais de 1.000 variáveis).

Tal procedimento é geralmente referenciado como pré-processamento e implica nas seguintes etapas (ou parte delas) [14]:

**Limpeza** - retirar inconsistências como registros incompletos e valores equivocados (e/ou suspeitos). A limpeza pode se dar pela remoção do registro inconsistente, atribuição de valores padrões, ou técnicas de agrupamento em busca de valores melhores ou ainda a imputação de dados, que consiste na substituição de dados faltantes por valores estimados plausíveis, com o objetivo de "completar" os bancos de dados e possibilitar a análise com todos os dados em estudo.

**Integração dos dados** - algumas vezes os dados não estão em uma única forma, contendo arquivos que trazem informação focando em grupo e outros que trazem informações individualizadas para cada elemento do grupo, ou ainda sob forma de texto, planilha e outras mídias. Neste caso, recomenda-se uma análise cuidadosa dos dados em busca de redundâncias, dependências entre as variáveis e valores conflitantes.

**Transformação dos dados** - considerando que em alguns casos não é possível trabalhar com valores textuais, ou valores inteiros que precisam se tornar decimais, e que alguns programas trabalham apenas com valores numéricos e outros permitem trabalhar com valores categóricos, algumas vezes se faz necessário transformar valores categóricos em valores numéricos ou transformar valores inteiros em decimais (e vice versa) ou ainda inserir valores numéricos correspondentes aos valores textuais para o devido processamento dos dados (lembrando que não convém eliminar a variável original).

**Redução dos dados** - por vezes nos deparamos com um volume considerável de informações não sendo todas de suma relevância para o processo a ser aplicado. Ou mesmo, tal volume comprometer o algoritmo escolhido. Convém então realizar a redução do conjunto de dados (sejam seus descritores ou o tamanho da amostra) cuidando para não comprometer a representatividade dos dados originais.

### *3.1.1 Arquivos da base de dados*

A base de dados utilizada é do ano de 2009, disponibilizada pelo MDS e descreve 11.372 famílias em 269 municípios de 23 estados da federação e do Distrito Federal (não constam os Estados do Acre, Roraima ou Tocantins). É constituída por cinco arquivos, conforme descrito no quadro 3.1.

As variáveis que compõem os arquivos da base de dados representam as informações coletadas a partir de um questionário aplicado por empresa contratada

pelo MDS, a Datamétrica Consultoria, Pesquisa e Telemarketing Ltda[1], que tinha como objetivo acompanhar a vida de famílias previamente selecionadas e suas condiçoes de vida após ingressar no PBF. Estas variáveis trazem informações domiciliares, bem como características sociais, educacionais, econômicas, de saúde e de antropometria dos moradores. Conta com seções de perguntas sobre:

1. características do domicílio;
2. características dos moradores, migração e antropometria;
3. educação (dados gerais e dados sobre gastos com educação);
4. saúde (dados gerais; dados de mulheres entre 10 e 49 anos de idade, dados sobre agentes de saúde; dados sobre gastos com saúde, e dados sobre saúde da criança);
5. trabalho e trabalho infantil;
6. rendimentos;
7. gastos individuais (gastos com transporte público e particular e com comunicações e gastos com alimentação fora de casa;
8. gastos coletivos do domicílio;
9. alimentos e bebidas - alcoólicas e não alcoólicas, adquiridos para consumo no domicílio;
10. inventário de bens duráveis (itens presentes no domicílio de propriedade dos moradores, ou alugados - animais e implementos agrícolas; e propriedades em posse de jure – legalizada – ou de fato – não legalizada);
11. avaliação das condições de vida – envolvimento com a comunidade: trabalho voluntário, cooperativas, etc.;
12. benefícios (PBF; e informação de benefícios que recebe ou já recebeu para cada morador do domicílio);
13. acesso a crédito, inclusão bancária e educação financeira;
14. percepção sobre pobreza, bem-estar e confiança; e
15. choques[2] e mecanismos de longo prazo.

[1] maiores informações no sumário executivo [1]

[2] se o entrevistado já passou por dificuldades em função de terceiros ou em função de catástrofes naturais como inundações, seca, pragas agrícolas, etc.

Quadro 3.1: Arquivos da Base de Dados Utilizada

| **arquivo** | **assunto** |
|---|---|
| domicilios.sav | Localização do domicílio, quantidade cômodos, se em zona urbana ou rural, se há saneamento, etc. |
| individuo.sav | quantas pessoas por família, informações sobre educação, saúde, empregabilidade, previdência, etc. |
| gastos coletivos.sav | gastos fixos com moradia, vestuário, transporte, serviços e outros gastos. |
| alimentos.sav | gastos com alimentação |
| beneficios.sav | se cadastrado ou não em algum projeto social, dados sobre benefícios recebidos, período de recebimento, etc. |

*fonte: elaboração própria*

Originalmente, as informações destes arquivos se apresentaram divididas em três categorias:

- famílias beneficiárias do Programa (30%) - apresentam características compatíveis com ;
- famílias não beneficiárias mas inscritas no CADÚnico (60%);
- famílias não cadastradas no CADÚnico[3], (10%).

### *3.1.2 Dificuldades em geral*

Dado que alguns arquivos traziam as informações referentes a cada indivíduo e outros informações sobre cada domicílio, o primeiro passo foi a realização de uma análise exploratória, a fim de conhecer as características dos dados e corrigir anomalias como dados faltantes, duplicados ou duvidosos, ou descartar variáveis que não contribuíam para o estudo. Algumas variáveis eram do tipo categóricas textuais, necessitando, portanto, que atuassem como fator.

Em alguns arquivos da base de dados as informações para cada domicílio (ou indivíduo) se apresentam em mais de uma linha estando portanto na forma "*item vs. domicilio*" ou "*item vs. indivíduo*". Podemos dizer que os dados estavam agrupados de acordo com o domicílio (1ª referência) e em cada grupo de domicílio estavam organizados os itens de interesse (alimentos ou gastos rotineiros de uma casa). Para a junção dos arquivos foi necessário que estes estivessem em função de apenas uma

[3]CADÚnico é a ferramenta utilizada pelo governo para identificar os indivíduos que se enquadram nos programas de políticas sociais - vide *apêndice A*

referência - domicílios. Daí, as transformações iniciais primaram pela junção dos arquivos considerando:

- Junção dos arquivos caracterizados por domicílio;
- Junção de dois blocos de dimensão distinta (domicílio vs. indivíduos);

Na junção dos arquivos caracterizados por domicílio, o arquivo "*domicilios.sav*" descreve em cada linha cada moradia quanto ao acesso a serviços como saneamento, coleta de lixo, entorno da moradia, calçamento. Já para os arquivos "*gastos coletivos.sav*" e "*alimentos.sav*", os domicílios são descritos por um grupo de linhas. Estas linhas se referem a cada objeto tratado para aquela família. Era necessário que cada linha se referencie a um único domicílio, contendo todas as informações, no caso, cada informação se tornaria uma variável e estas, distribuídas nas colunas do arquivo. Sendo 11.372 domicílios o resultado após a junção deverá conter 11.372 linhas. O arquivo "*gastos coletivos.sav*" trata dos gastos fixos de um domicílio, como IPTU, aluguel, luz, água, etc. contabilizando 54 itens de interesse por cada um dos domicílios cadastrados. Cada linha descreve o gasto ali tratado tendo como referência o identificador do domicilio e o identificador do gasto em questão (54 linhas para cada domicílio) como ilustrado na figura 3.1.

| GASTOS COLETIVOS | | |
|---|---|---|
| **Id domicílio** | **Descrição** | **Valor** |
| 0001 | aluguel do imóvel | $ |
| 0001 | energia elétrica | $ |
| : | : | : |
| 0002 | vestuário adulto masculino | $ |
| : | : | : |
| : | : | : |

**54 itens para o domicílio 0001**

**54 itens para o domicílio 0002**

Figura 3.1: Necessidade de redimensionamento de "gastos coletivos"
*Fonte: elaboração própria*

O arquivo necessita sofrer um ajuste de dimensão, para que cada linha (lançamento) represente um domicílio. Este procedimento promove o rearranjo do arquivo fazendo com que cada linha de gasto referentes ao "*domicilio x*" se transforme em variável e as linhas passem a descrever o domicílio.

(54 itens de estudo) vs. (11.372 domicílios) $\Rightarrow$ 614.088 linhas

Para isso são necessárias duas variáveis de referência – identificador do domicílio como índice primário, e identificador do gasto como índice secundário. A transformação resultante está ilustrada na figura 3.2:

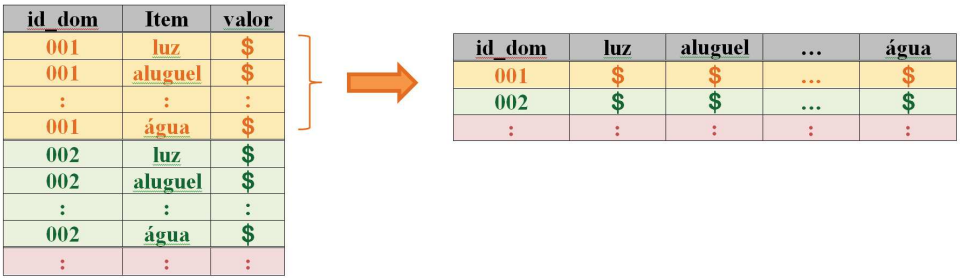


Figura 3.2: Redimensionamento de dados
*Fonte: elaboração própria*

O mesmo ocorre com o arquivo "*alimentos.sav*", que trata dos alimentos presentes em cada domicílio. Novamente o identificador do domicílio como índice primário, e identificador do alimento consumido como índice secundário nos permitirá rearranjar o arquivo de modo que cada linha descreva um domicílio.

(65 alimentos) vs. (11.372 domicílios) $\Rightarrow$ 739.180 linhas

O arquivo "*beneficios.sav*" já traz suas informações em função do domicílio, portanto bastou anexá-lo aos outros dois previamente tratados. O arquivo "*individuos.sav*" traz informações quanto a aspectos antropométricos, saúde (presença de doença crônica ou não, uso dos serviços de saúde,etc), grau de instrução, entre outras informações num total de 56.367 indivíduos. Este sofreu transformações que permitissem considerar suas informações referenciando-se ao domicílio a que pertence o indivíduo, gerando então variáveis equivalentes a um determinado grupo de indivíduos pertencentes ao domicílio "$x_i$". Durante o processo de redimensionamento de alguns arquivos, algumas variáveis foram descartadas visto que seu conteúdo poderia ser identificado pelo preenchimento ou não de outras variáveis. Sendo do tipo binário ("sim" ou "não"), para assumir o valor "sim", determinado grupo de variáveis receberam valores, do contrário este mesmo grupo de variáveis não apresentou informações.

Conforme mencionado no item 3.1.1, a base de dados apresenta três grupos, porém, para o experimento considerou-se apenas dois deles - famílias "beneficiárias", e famílias "não-beneficiárias", descartando assim o outro grupo que possue características

de beneficiários mas ainda se encontram na espera para recebimento do benefício. Após o descarte de observações comprometidas pela ocorrência de dados faltantes, a amostra então 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%. Aplicando-se os classificadores sobre esta amostra, sem maiores intervenções, deparamo-nos com um viés decorrente da classe de maior frequência – uma tendência a classificar os dados do grupo majoritário e negligenciar o grupo minoritário. Se o grupo caracterizado como beneficiário (mas ainda na espera) fosse anexado isso tornaria o tamanho dos grupos ainda mais distante visto que este grupo e os beneficiários possuem as mesmas características - não estão recebendo o benefício meramente por limitações orçamentárias por um período. Além disso, a proporção 3:1 considerando somente o grupo de beneficiários (2.453 casos) e não beneficiários (801 casos) passaria a 9:1 (2.453 + 4.820 famílias que atendem a situação de beneficiários vs. 801 famílias não-beneficiárias), agravando ainda mais o desequilíbrio entre os grupos frente aos classificadores.

### *3.1.3 Escolha das variáveis*

A regressão logística foi a ferramenta utilizada para a seleção e ordenação das variáveis de interesse, e foi também um dos classificadores empregados no estudo. Em geral, a regressão é usada para predição - prever o valor de *"y"* a partir do valor de *"x"* - e estimar o quanto *"x"* influencia ou modifica *"y"*.

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, foram escolhidas 64 variáveis por serem consideradas as mais relevantes[4], com base em conhecimentos prévios da literatura (algumas citadas na seção 2). Em seguida, parte das variáveis foram reagrupadas e, a partir do modelo de regressão logística na modalidade *"stepwise-forward"* variáveis foram eliminadas (por se tratar de variáveis redundantes ou pouco informativas) 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 necessidade da permanência da primeira variável. Caso permaneça, uma terceira variável é selecionada. Se uma terceira variável entra no modelo, verifica-se a continuidade das duas anteriores. Novamente, experimenta-se a inclusão de uma nova variável. Caso entre, tenta-se eliminar uma das que já estão no modelo. O procedimento acaba

[4]listadas nos quadros D.1 e D.2 no apêndice D

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 verdadeiros positivos) e especificidade (classificação correta dos verdadeiros negativos), 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).

Uma listagem das variáveis eleitas encontram-se no apêndice D e abordam 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);
- 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 ;
- quanto gastou 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, limpeza da casa;

- gastos com vestuário, higiene pessoal, lazer;
- gastos com alimentos comprados por família;
- quantos automóveis possui;

# 4 Métodos

## 4.1 Algoritmos de Classificação

O processo de classificação pode ser supervisionado ou não-supervisionado. Na classificação supervisionada o padrão de classe existente na amostra é conhecido - o classificador aprende pelo exemplo (treinamento) a fim de estabelecer um padrão e então identificar novos elementos de acordo com suas classes (teste). Na classificação não-supervisionada não há uma referência prévia para os padrões de treinamento, de maneira que, os algoritmos têm que "identificar"uma estrutura nos dados que permita dividi-los em grupos. A escolha dos casos para treinamento e teste e os parâmetros dos algoritmos de classificação interferem no desempenho dos mesmos.

Os algoritmos classificadores eleitos para o experimento são do tipo supervicionados. São eles a rede neural artificial de múltiplas camadas (RNA-MLP), a regressão logística (RL) e a árvore binária de decisão (ABD). 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 uma variável *dummy* de valor igual a "0"). Os valores finais para cada um dos algoritmos classificadores são a média de 50 reamostragens por permutação a fim de afastar a hipótese de mera obra do acaso. O algoritmo ABD utilizou-se de todas as variáveis disponíveis do arquivo original ao passo que a RL e RNA utilzaram as 32 variáveis obtidas na primeira aplicação de regressão logística (escolha das variáveis). Todos os algoritmos utilizaram as mesmas amostras.

### *4.1.1 Arvore Binária de Decisão (ABD)*

Uma árvore binária de decisão consiste em um conjunto de nós e folhas, onde os nós representam pontos de decisão, e as folhas, as opções disponíveis (cada classe). Cada folha pode se tornar um novo nó (figura 4.1), até que a separação desejada seja alcançada ou o algoritmo atinja outra condição de parada. 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, de modo que o caminho percorrido na árvore (da raiz à folha de interesse) representa uma regra de classificação [15]. O sucesso desse método se deve ao fato de ser uma técnica extremamente simples, que geralmente alcança um bom

índice de acertos. O próprio algoritmo de aprendizado faz a seleção dos atributos considerados relevantes. Funcionam bem com amostras de grande tamanho e muitas variáveis.

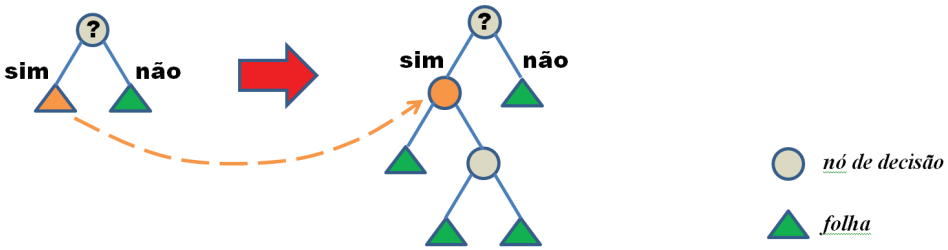


Figura 4.1: Esquema de uma árvore binária de decisão
*Fonte: elaboração própria*

De acordo com as opções de implementação do "software" a ser utilizado, pode-se informar que variáveis utilizar . O modelo será criado a partir da amostra de treino e este será aplicado sobre a amostra-teste. De modo simples podemos dizer que uma árvore de decisão é o número mínimo de perguntas que devem ser respondidas para avaliar a probabilidade de tomar uma decisão correta, na maioria das vezes - permite abordar o problema de uma forma estruturada e sistemática para chegar a uma conclusão.

A figura 4.2 mostra a árvore de decisão gerada neste estudo.

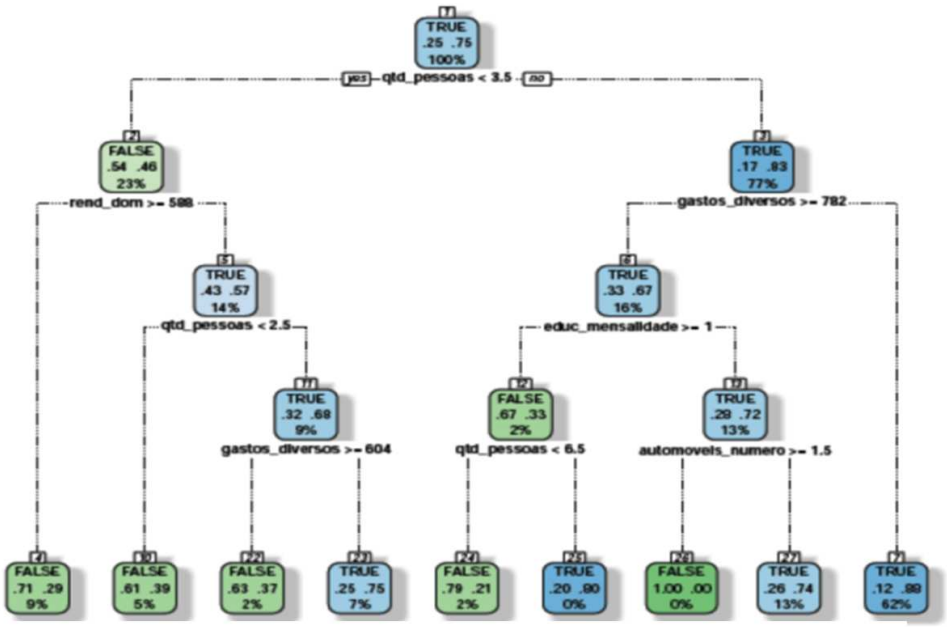


Figura 4.2: árvore de decisão gerada neste estudo
*Fonte: elaboração própria*

### *4.1.2 Regressão Logística (RL)*

A regressão logística é uma ferramenta estatística que permite descrever uma variável dependente a partir do ajuste de um conjunto de outras variáveis ditas independentes (por vezes chamadas de explicativas ou preditoras) que podem ser do tipo numéricas, categóricas ou ambas. Este modelo de regressão é utilizado quando a variável resposta é binária, ou seja, apresenta dois resultados possíveis (sim/não, verdadeiro/falso,...) variando no intervalo de "0" a "1", segundo uma curva chamada "sigmoidal" ou "curva-S" conforme mostrada na Figura 4. A função que a define é dada por [16]:

$$f(Y) = \frac{1}{1 + exp^{-(Y)}} \tag{4.1}$$

onde

$$Y = ln\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_n X_n \tag{4.2}$$

| | |
|---|---|
| $Y$ | *variável dependente*; |
| $p$ | *probabilidade de ocorrência de evento de interesse*; |
| $(1\text{-}p)$ | *probabilidade de não-ocorrência de evento de interesse*; |
| $X_i$ | *variáveis independentes (ou explicativas)*; |
| $\beta_i$ | *coeficientes das variáveis explicativas*; |

O modelo de regressão logística, conforme equação 4.2, é o logaritmo natural da probabilidade de ocorrência num grupo, dividido pela probabilidade de ocorrência no outro grupo. Os coeficientes $\beta_i$ são estimados pelo modelo de regressão, representado pela equação 4.2, e indicam a importância de cada variável independente na ocorrência do evento. Se $\beta_i$ é positivo, significa que a probabilidade de ocorrência do evento aumenta, quando a variável independente $X_i$ aumenta. Se $\beta_i$ é negativo, a probabilidade de ocorrência do evento diminui, quando a variável independente $X_i$ diminui.

É preciso escolher o ponto de corte na variável Y que separa as duas categorias de saída – uma referenciada pelo valor "0" e outra, pelo valor "1". Não necessariamente este ponto precisa ser "equidistante" de cada classe, conforme ilustrado na figura 4.3. No presente estudo 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 para ponto de corte (*cut-off*) foi de 0,739.

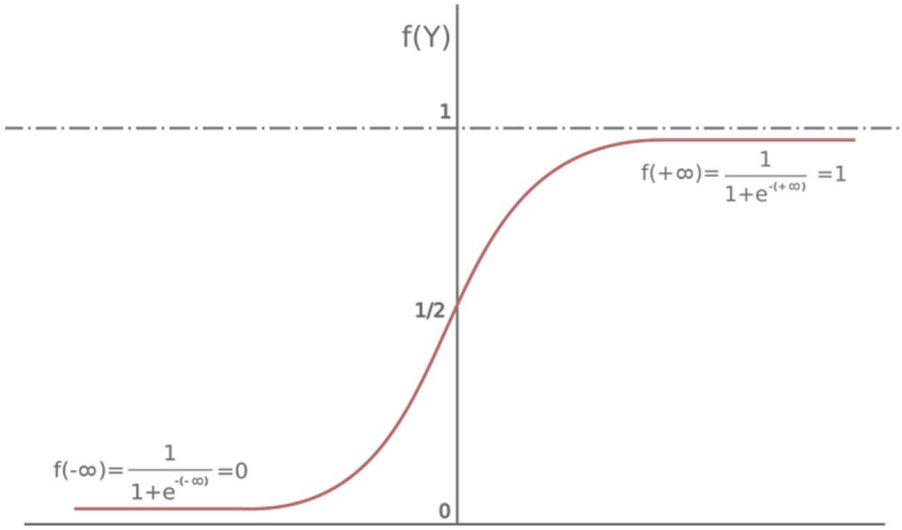


Figura 4.3: Função logística
*Fonte: adaptado de [16]*

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:

$$P(1) = f(Y = 1|X_1, X_2, ..., X_n) = \frac{1}{1 + e^{-(\beta_0 + \Sigma\ \beta_i X_i)}} \tag{4.3}$$

Daí, a regra de classificação de regressão logística para discriminação de dois grupos, é a seguinte:

- Se P(Y=1) for maior que o ponto de corte, então será classificada como Y=1;
- caso contrário, Y=0

#### 4.1.2.1 Coeficiente de Informaçao de Acaike (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:

$$AIC = 2 * k - 2 * \ln(L) \tag{4.4}$$

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).

### *4.1.3 Rede Neural Artificial (RNA)*

O modelo matemático que conhecemos hoje para representar um neurônio artificial foi idealizado por Warren S. McCulloch e Walter H. Pitts no ano de 1943 [17]. É uma técnica baseada no comportamento dos neurônios dos seres vivos. De maneira geral, uma rede neural pode ser vista como um conjunto de unidades de entrada "$x_i$" e saída "y" conectadas por camadas intermediárias e cada ligação possui um peso associado "$w_i$" conforme ilustrado na figura 4.4. A soma ponderada das entradas é submetida a uma função de ativação "$f(u)$" (também referenciada como função de transferência) que determina se a soma é maior que um valor numérico - o limiar do neurônio (bias) - se sim, o neurônio é ativado (valor "1") caso contrário, é desativado (valor "0") – representado pela equação 4.5.

$$y = f(u) = f(\Sigma w_i x_i + b) \tag{4.5}$$

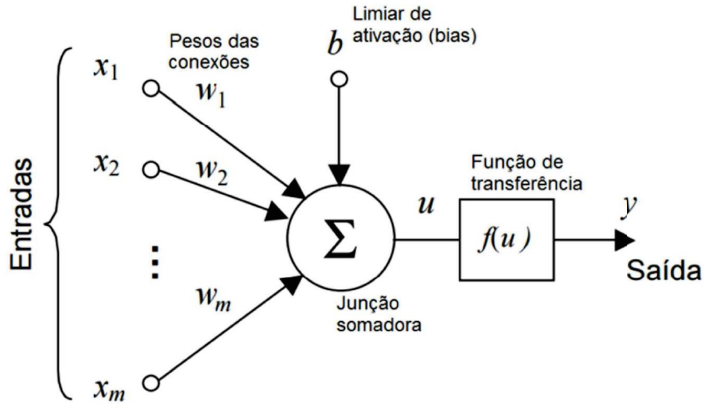


Figura 4.4: Representação esquemática de um neurônio artificial
*Fonte: referência [18]*

Numa RNA de múltiplas camadas (Multi-Layer Perceptron, MLP) os neurônios estão organizados em duas ou mais camadas na configuração feedforward onde as camadas estão organizadas de tal modo que os neurônios de uma camada estimulam todos os neurônios da camada seguinte. Isso implica que nenhum neurônio pode estimular um neurônio da mesma camada ou de camadas anteriores como ocorre em outros tipos de RNA. O esquema representado na figura 4.5 mostra uma rede de *"k"* entradas com *"p"* camadas ocultas podendo cada uma conter de 1 até *"n"* neurônios.

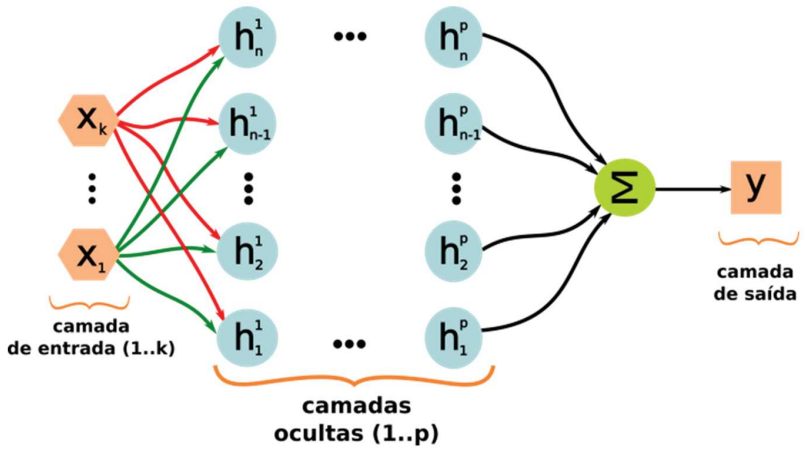


Figura 4.5: Esquema de uma rede neural multicamada
*Fonte: adaptado de [17]*

O número de camadas ocultas e neurônios contidos em uma rede é representado por um colchete onde a quantidade de elementos equivale a quantidade de camadas e o valor de cada elemento dentro do mesmo informa o numero de neurônios de cada camada. Assim a notação $[x_1, x_2, ..., x_k]$ representa,

$x_i$ é a quantidade de neurônios na camada *"i"*

$i$ é a quantidade de camadas que há na rede

Pode-se considerar que a RL é um caso particular de uma RNA com um único neurônio na camada oculta.

Foi utilizada no experimento a configuração "*Multi Layer Perceptron*" [32, 8], com critério de parada por validação cruzada. Foram testados cada uma das configurações de aprendizado citados no quadro 4.1 e a melhor resposta foi "*levenberg-Marquardt backpropagation*", que corresponde a configuração "*default*". Os valores alcançados em cada configuração RNA pode ser visto no apêndice C.

Quadro 4.1: algoritmos de aprendizado testados na RNA

| **sigla** | **nome da metodologia** |
|---|---|
| trainbfg | BFGS quasi-Newton backpropagation |
| trainbr | Bayesian regularization backpropagation |
| traincgb | Conjugate gradient backpropagation with Powell-Beale restarts |
| traincgf | Conjugate gradient backpropagation with Fletcher-Reeves updates |
| traincgp | Conjugate gradient backpropagation with Polak-Ribiére updates |
| traingd | Gradient descent backpropagation |
| traingda | Gradient descent with adaptive learning rate backpropagation |
| traingdm | Backpropagation de gradiente decrescente com momentum |
| traingdm | Gradient descent with momentum backpropagation |
| traingdx | Gradient descent with momentum and adaptive learning rate backpropagation |
| **trainlm** | **Levenberg-Marquardt backpropagation** |
| trainoss | One-step secant backpropagation |
| trainrp | Resilient backpropagation |
| trainscg | Scaled conjugate gradient backpropagation |

*Fonte: elaboração própria*

## 4.2 Medidas de desempenho para classificadores binários

As métricas aqui levantadas para medir o desempenho de cada algoritmo classificador foram calculadas a partir da "matriz de confusão" [2] (figura 4.6). 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 casos como positivo em concordância com o observado (neste estudo, o grupo de "beneficiários") e "verdadeiro negativo", quando classifica casos como negativo também em concordância com o observado (neste estudo, o 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.

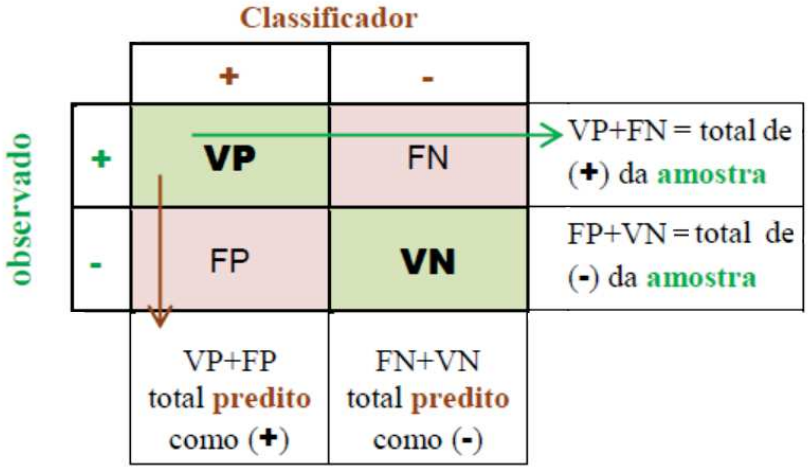


Figura 4.6: Matriz de Confusão
*Fonte: adaptado de [2]*

Segue uma breve apresentação dos conceitos de métricas baseadas na matriz de confusão que serão aplicados sobre os três algoritmos classificadores e suas respectivas fórmulas. Como medidas de desempenho individual temos, atrelado ao conceito de "positisvos" a sensibilidade, confiabilidade positiva, suporte, cobertura, *f-measure*, taxa de erro positiva. Atrelado ao conceito de "negativo", a especificidade, confiabilidade negativa e taxa de erro negativa. Como medidas de desempenho global, acurácia, eficácia, g-mean, kappa, younden e MCC.

### *4.2.1 Sensibilidade*

Também referenciado como *"recall"*, é a proporção de verdadeiros positivos (predições positivas corretas) em relaçao ao total de positivos da amostra, ou seja, classificações corretas das famílias beneficiárias. Varia entre "0" e "1" [19].

$$sensibilidade = \frac{verdadeiros\ positivos}{total\ de\ positivos\ da\ amostra} = \frac{VP}{VP + FN} \tag{4.6}$$

É uma medida de desempenho do classificador frente a uma das classes. Sendo o caso em questão o reconhecimento de pertencimento entre duas classes, os dados majoritários são normalmente atrelados ao conceito de "positivo" e portanto de sensibilidade.

### 4.2.2 Confiabilidade positiva

Também referenciada como *"precisão"* ou ainda valor predito positivo (VPP). É a probabilidade da classe vinculada ao verdadeiro positivo (majoritária) ser identificada. É a taxa de classificação positiva correta frente aos valores preditos como positivos [19].

$$VPP = \frac{predição\ positiva\ correta}{predição\ positiva} = \frac{VP}{VP + FP} \tag{4.7}$$

### 4.2.3 Suporte

Ou *"frequência"*, é a medida do número de exemplos corretamente identificados como verdadeiro positivo ([4] e [18]).

$$suporte = \frac{verdadeiro\ positivo}{total} = \frac{VP}{(VP + FN) + (VN + FP)} \tag{4.8}$$

### 4.2.4 Cobertura

Medida do número de exemplos preditos correspondente a classe definida como verdadeiros positivos frente ao total de elementos da amostra. Taxa de Predição desta classe ([4] e [18]).

$$cobertura = \frac{predição\ positiva}{total} = \frac{VP + FP}{(VP + FN) + (VN + FP)} \tag{4.9}$$

### 4.2.5 F-measure

Também chamada *"F-score"*, é uma média ponderada entre confiabilidade positiva (precisão) e sensibilidade. Está no intervalo de "0" a "1". O resultado *F-Measure* é um indicativo de que, quanto mais próximo de "1", melhor é o desempenho e resultados mais próximos de "0" demonstram desempenho ruim para a classe atrelada ao conceito de "positivo" ([13] e [20]).

$$f - measure = \frac{(1 + \beta) * precisão * sensibilidade}{\beta^2 * precisão + sensibilidade} = \frac{(1 + \beta) * \frac{VN}{VN+FP} * \frac{VP}{VP+FN}}{\beta^2 * \frac{VN}{VN+FP} + \frac{VP}{VP+FN}} \tag{4.10}$$

onde,

$$\beta = \frac{sensibilidade}{precisão} \tag{4.11}$$

### 4.2.6 Especificidade

Proporção de verdadeiros negativos (predições negativas corretas) em relaçao ao total de negativos da amostra, ou seja, classificações corretas das famílias não-beneficiárias. Varia entre “0” e “1” [19].

$$especificidade = \frac{verdadeiros\ negativos}{total\ de\ negativos\ da\ amostra} = \frac{VP}{VP + FN} \tag{4.12}$$

É uma medida de desempenho do classificador frente a uma das classes. No caso, a classe minoritária fica atrelada ao conceito de “negativo”, e portanto, especificidade. E por ser minoritária, por vezes é negligenciada se os dados são do tipo desbalanceado pois a classe majoritária provoca um viés nos algoritmos de classificação se não é levado em consideração durante a escolha da configuração do classificador o fato dos dados serem desbalanceados.

### 4.2.7 Confiabilidade negativa

Ou valor predito negativo (VPN), é a probabilidade da classe vinculada ao verdadeiro negativo ser identificada [19].

$$VPP = \frac{predição\ negativa\ correta}{predição\ negativa} = \frac{VN}{VN + FN} \tag{4.13}$$

### 4.2.8 Acurácia

Medida de desempenho global, é a proporção de classificações corretas, tanto de casos positivos quanto negativos. Em caso de dados desbalanceados e, se for de interesse que mais de uma classe seja corretamente identificada, esta medida pode induzir a uma conclusão errônea quanto ao desempenho do algoritmo empregado visto que a classe majoritária encobrirá o baixo desempenho frente a classe minoritária .Varia entre “0” e “1” [19].

$$acurácia = \frac{total\ de\ acertos}{total\ de\ elementos\ da\ amostra} = \frac{VP + VN}{(VP + FN) + (VN + FP)} \tag{4.14}$$

### 4.2.9 Eficiência

Média aritmética da sensibilidade e especificidade. Normalmente usada quando a quantidade de elementos apresenta considerável diferença entre grupos. Traz uma medida de desempenho global menos equivocada que a acurácia visto que a deficiência da classe minoritária promove um deslocamento do valor em questão, ficando entre a sensibilidade e a especificidade. Varia entre "0" e "1" [19].

$$eficiência = \frac{sensibilidade + especificidade}{2} \tag{4.15}$$

### 4.2.10 Média geométrica

Ou "g-mean", foi proposta por Kubat et al. (1998) e 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 [10].

$$g - mean = \sqrt{sensibilidade * especificidade} = \sqrt{\left(\frac{VP}{VP + FN}\right)} \tag{4.16}$$

### 4.2.11 Índice Kappa

Definido como uma medida de associação usada para descrever e testar o grau de concordância entre predito e observado na classificação ([20] e [21]). Este índice varia de "0"a "1"e pode ser interpretado conforme a tabela 4.1:

$$\kappa = \frac{total\ de\ acertos - proporção\ de\ acertos\ esperada}{total\ de\ amostras - proporção\ de\ acertos\ esperada} =$$

$$= \frac{(VP + VN) - \left\{\frac{[(VP+FN)*(VN+FP)]+[(FP+VN)*(FN+VN)]}{(VP+FN)+(VN+FP)}\right\}}{[(VP + FN) + (VN + FP)] - \left\{\frac{[(VP+FN)*(VN+FP)]+[(FP+VN)*(FN+VN)]}{(VP+FN)+(VN+FP)}\right\}} \tag{4.17}$$

onde "$\kappa$ " é o coeficiente Kappa.

Tabela 4.1: Interpretação dos valores do Índice kappa

| **Valor de kappa** | $> 0,20$ | 0,21 - 0,40 | 0,41 - 0,60 | 0,61 - 0,80 | 0,81 - 1,00 |
|---|---|---|---|---|---|
| **qualidade do classificador** | ruim | fraca | boa | muito boa | excelente |

Fonte: referência [20]

### *4.2.12 Índice de Youden*

Proposto em 1950 por Youden como uma solução prática para relacionar a sensibilidade e a especificidade em testes diagnósticos, este índice tem como objetivo medir a performance geral de testes diagnósticos. Também configura boa opção utilizado para determinar ponto de corte entre sensibilidade e especificidade [22].

$$Younden = sensibilidade + especificidade - 1 = \left(\frac{VP}{VP+FN}\right) + \left(\frac{VN}{VN+FP}\right) - 1 \quad (4.18)$$

É a maior distância perpendicular entre um ponto da curva ROC e a diagonal do Espaço ROC, conforme mostra a figura 4.7.

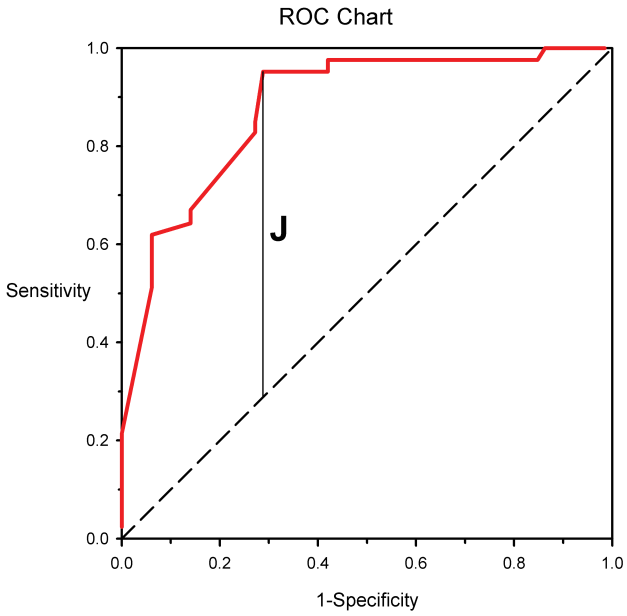


Figura 4.7: Coeficiente Youden(J)

*Fonte: referencia [22]*

### 4.2.13 Coeficiente de correlação de Matthews (MCC)

A sigla MCC vem da denominação desta métrica em inglês - Matthews Correlation Coefficient, também referenciada como coeficiente . 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], onde os valores *"+1"* indica uma classificação perfeita, *"0"* indica uma classificação equivalente a que seria feita aleatoriamente, e *"-1"* uma classificação imprópria, invertida [19].


Figura 4.8: escala do coeficiente de correlação de Matthews
*Fonte: elaboração própria*

Pode ser calculado a partir da matriz de confusão, pela fórmula:

$$MCC = \frac{(VP * VN) - (FP * FN)}{\sqrt{(VP + FP) + (VP + FN) + (VN + FP) + (VN + FN)}} \tag{4.19}$$

Se qualquer uma das quatro somas no denominador for zero, o denominador pode ser arbitrariamente fixado em *"1"*, resultando em MCC igual a zero.

### 4.2.14 Curva ROC

Por vezes o uso de gráficos e/ou diagramas permite uma melhor visualização da informação. A curva ROC (*Receiver Operating Characteristic*) é uma ferramenta utilizada para avaliar desempenho e acurácia de um modelo de classificação e serve como ferramenta na decisão de onde estabelecer o ponto de corte. Tal ponto é identificado através da construção desta curva onde são calculados os valores de sensibilidade e falsos positivos encontrados na amostra, chamado espaço ROC [23]. Alguns pontos no espaço ROC merecem destaque - na figura 4.9 vemos a marcação desses pontos e sua descrição no quadro 4.2.

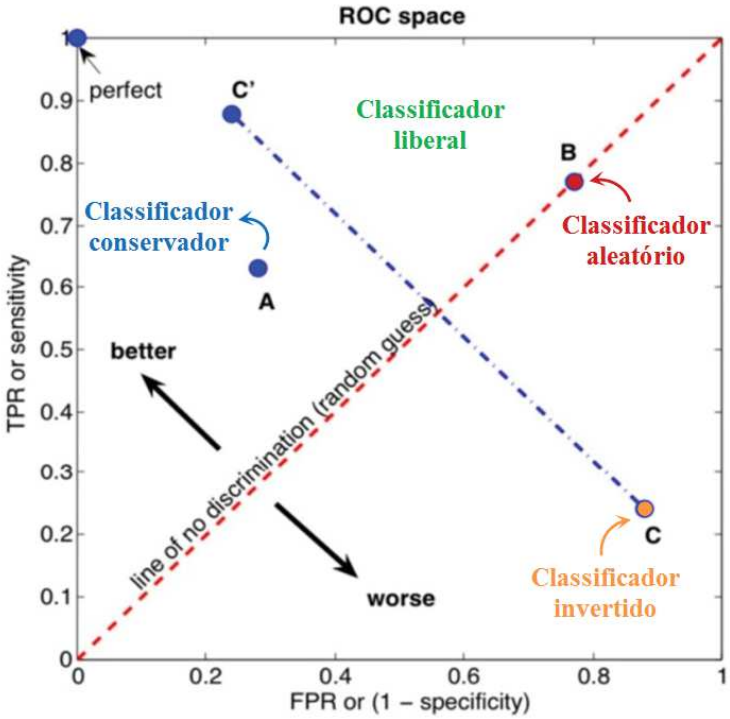


Figura 4.9: Análise sob a curva ROC
*Fonte: Adaptado da referencia [23]*

Quanto mais próxima a curva estiver do canto superior esquerdo do espaço destinado a representação gráfica, melhor será o desempenho do método avaliado. O quadro 4.2 descreve os tipos de classificadores identificados segundo a curva ROC conforme pontos sinalizados na figura 4.9.

A referência [24] apresenta de forma bastante clara o significado de pontos que marcam as extremidades do espaço ROC:

> *... o ponto (0,0) representa a estratégia de nunca classificar um exemplo como positivo. Modelos que correspondem a esse ponto não apresentam nenhum falso positivo, mas também não conseguem classificar nenhum verdadeiro positivo. A estratégia inversa, de sempre classificar um novo exemplo como positivo, é representada pelo ponto (100%,100%). O ponto (0,100%) representa o modelo perfeito, i.e., todos os exemplos positivos e negativos são corretamente classificados. O ponto (100%,0) representa o modelo que sempre faz predições erradas. Modelos próximos ao canto inferior esquerdo podem ser considerados "conservativos": eles fazem uma classificação positiva somente se têm grande segurança na classificação. Como consequência, eles cometem poucos erros falsos positivos, mas frequentemente têm baixas taxas de verdadeiros positivos. Modelos próximos ao canto superior direito podem ser considerados "liberais": eles predizem a classe positiva com maior freqüência, de tal maneira que classificam a maioria dos exemplos positivos corretamente, mas, geralmente, com altas taxas de falsos positivos.*

Quadro 4.2: Tipos de classificadores segundo a curva ROC

| **Tipo de Classificador** | **Descrição** |
|---|---|
| Classificador Liberal | acima da diagonal vermelha, entre os pontos C' e B;<br>boa classificação de VN, mas, com muitos FP; |
| Classificador Invertido | apresenta resultados abaixo da diagonal do espaço ROC (classificador C);<br>contém informação sobre a Classe, mas de forma errada;<br>se a saída for negada, o ponto passa para a metade superior do espaço ROC (classificador C') |
| Classificador Conservador | nas imediações do ponto A;<br>boa classificação de VP e com poucos FP; |
| Classificador Aleatório (Random Guessing) | apresenta resultados sobre a diagonal do espaço ROC (na figura, o Classificador B);<br>não tem nenhuma informação sobre a Classe;<br>O classificador não discrimina as Classes. |

Fonte: adaptado da referência [23]

### *4.2.15 Área sob a curva (AUC)*

A área sob a curva (AUC) é a representação numérica do desempenho de um algoritmo no Espaço ROC. Quanto mais o valor se aproximar de "1" melhor o desempenho do classificador. A proximidade da diagonal implica em uma área em torno de "0,5" o que aponta para uma predição aleatória, conforme assinalado na tabela 4.2 – classificador aleatório (quadro 4.2) [23].

Na figura 4.10 temos o espaço ROC com marcação das direções de desempenho e na figura 4.11, a curva ROC de uma amostra classificada por RNA cuja área sob a curva alcançou os valores 0,84905; 0,79178 e 0,79064 durante as etapas de treino, validação e teste respectivamente.

Tabela 4.2: Valores de referência para avaliaćão segundo AUC

| Valor AUC | 1 | 0,9 | 0,8 | 0,7 | 0,6 | 0,5 |
|---|---|---|---|---|---|---|
| Interpretação | previsão perfeita | excelente previsão | boa previsão | previsão medíocre | previsão pobre | previsão aleatória |

*Fonte: referência [25]*

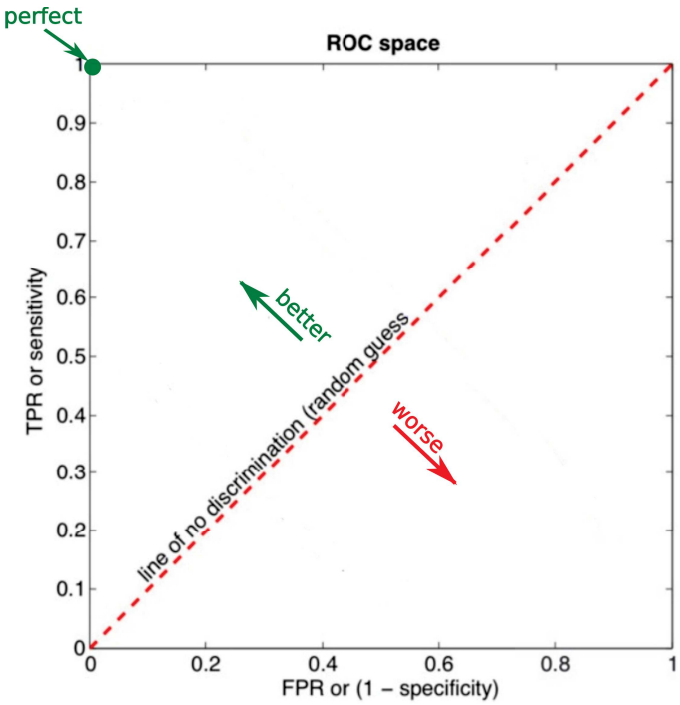


Figura 4.10: Sinalização de desempenho no espaço ROC

*Fonte: adaptado da referência [23]*

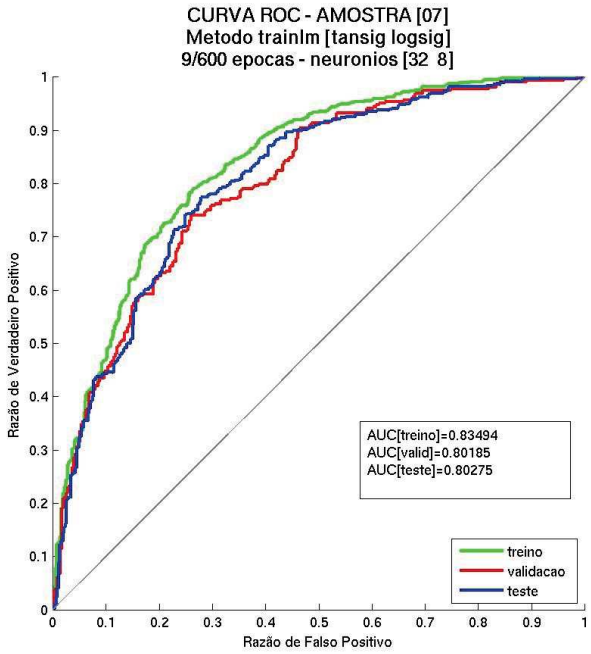


Figura 4.11: curva ROC - RNA sem intervenção

*Fonte: elaboração própria*

### 4.2.16 Taxa de erro positiva

Consiste na proporção de casos negativos que foram incorretamente classificados como positivos [26].

$$tx\_err_{pos} = \frac{identificados \quad erroneamente \quad como \quad positivos}{total \quad de \quad negativos \quad da \quad amostra} = \frac{FP}{FP + VN} \tag{4.20}$$

### 4.2.17 Taxa de erro negativa

Consiste na proporção de casos positivos que foram incorretamente classificados como negativos [26].

$$tx\_err_{neg} = \frac{identificados \quad erroneamente \quad como \quad negativos}{total \quad de \quad positivos \quad da \quad amostra} = \frac{FN}{FN + VP} \tag{4.21}$$

### 4.2.18 Taxa de erro global

Proporção de casos incorretamente classificados considerando toda a amostra [26].

$$tx\_err = \frac{identificados \quad erroneamente}{total \quad da \quad amostra} = \frac{FP + FN}{(FP + VN) + (FN + VP)} \tag{4.22}$$

## 4.3 Sobre-amostragem e sub-amostragem

São técnicas de pré-processamento que buscam equilibrar uma amostra desbalanceada. Subamostragem (ou *undersampling*) remove elementos da classe majoritária e sobreamostragem (ou *oversampling*) inclui elementos na classe minoritária. Tanto uma ação como outra busca um nível adequado de balanceamento. Ambas podem ser utilizadas de forma aleatória ou por meio de algum critério de seleção. Aqui foram aplicadas tanto a sobre-amostragem como a sub-amostragem na modalidade aleatória.

O risco da aplicação da subamostragem reside no risco de acarretar em perda de informação. Já para a sobreamostragem, pode-se deparar com o efeito adverso de um *overfitting*[1].

[1] quando o modelo fica superajustado e perde a capacidade de generalização.

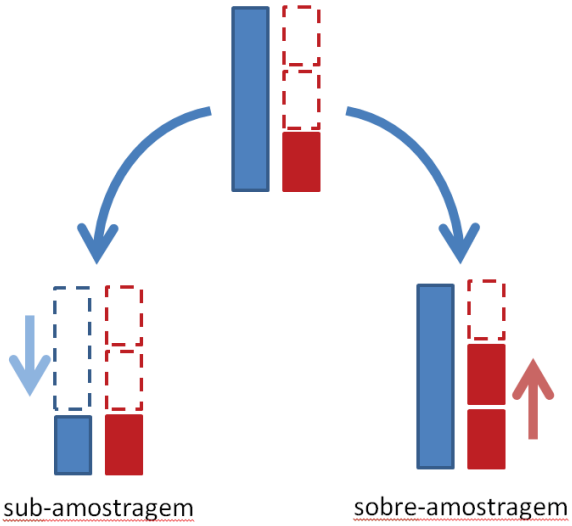


Figura 4.12: sobre-amostragem & sub-amostragem em dados desbalanceados
*Fonte: elaboração própria*

# 5 Discutindo resultados alcançados

Conforme pode ser visto na tabela 5.1, os valores observados nas métricas aplicadas apresentavam valores baixos no que diz respeito ao desempenho do classificador sobre a classe minoritária. Desse modo, o classificador como um todo se mostra deficiente já que estamos considerando que a identificação de ambos os grupos é importante para o estudo. Convém lembrar que os dados utilizados tinham como objetivo acompanhar uma ação já implementada (distribuição direta de renda sob condicionantes) e não para análise quanto a necessidade do benefício ou não. Apesar das mais de 1.000 variáveis presentes na base de dados original, uma hipótese seria a falta de mais informações relevantes para que o classificador reconhecesse cada grupo. Isso talvez se deva ao fato de boa parte das informações coletadas serem somente declaradas, principalmente no que diz respeito ao conjunto de variáveis referentes a ganhos financeiros, muitas vezes vazio ou incompleto.

A escolha dos casos no processo de reamostragem foi aleatória. Existem processos de reamostragem direcionados e associação de procedimentos classificatórios que não foram implementados, conforme comentando na sessão 6.

## 5.1 Comportamento observado

Neste estudo é importante que ambas as classes sejam reconhecidas pelo sistema classificador já que, apesar de se tratar de apenas duas classes, estas não são complementares. O que se observa, porém, da figura 5.1 a 5.3, são os valores preditos de cada classe (beneficiários vs. não beneficiários) bastante dispersos e/ou sobrepostos. No *eixo-y* marca-se a classificação dos grupos - os valores de referência, em círculos azuis, assumem valor *"0"* se *"não-beneficiários"* ou valor *"1"* se *"beneficiários"*. Para a RNA, observa-se que os valores preditos extrapolaram este intervalo (vide figura 5.3) - predição dos "beneficiários" em cruzes verdes, e predição dos "não beneficiários" em triângulos vermelhos. O *eixo-x* identifica a quantidade de valores preditos a partir da amostra-teste. Junto a isso, o fato das classes apresentarem quantidade diferente de elementos, na proporção 3:1 aponta como possível motivo quanto ao desempenho do classificador junto a classes minoritária não se mostrar satisfatório.

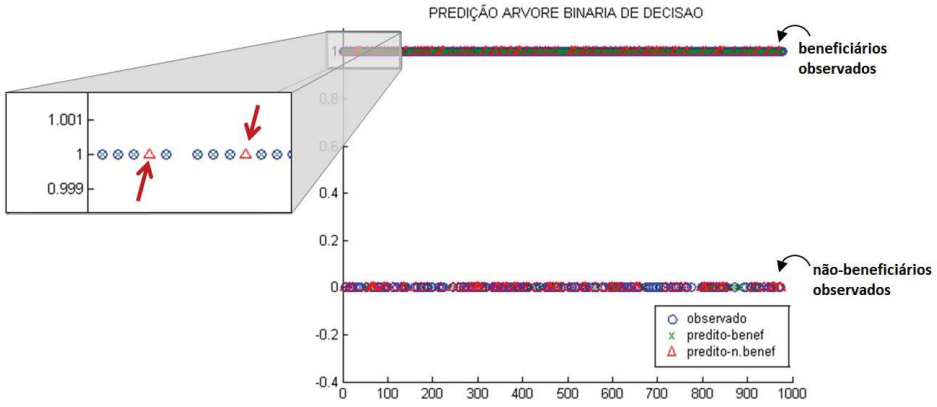


Figura 5.1: Dispersão dos dados simulados na ABD

*fonte: elaboração própria*

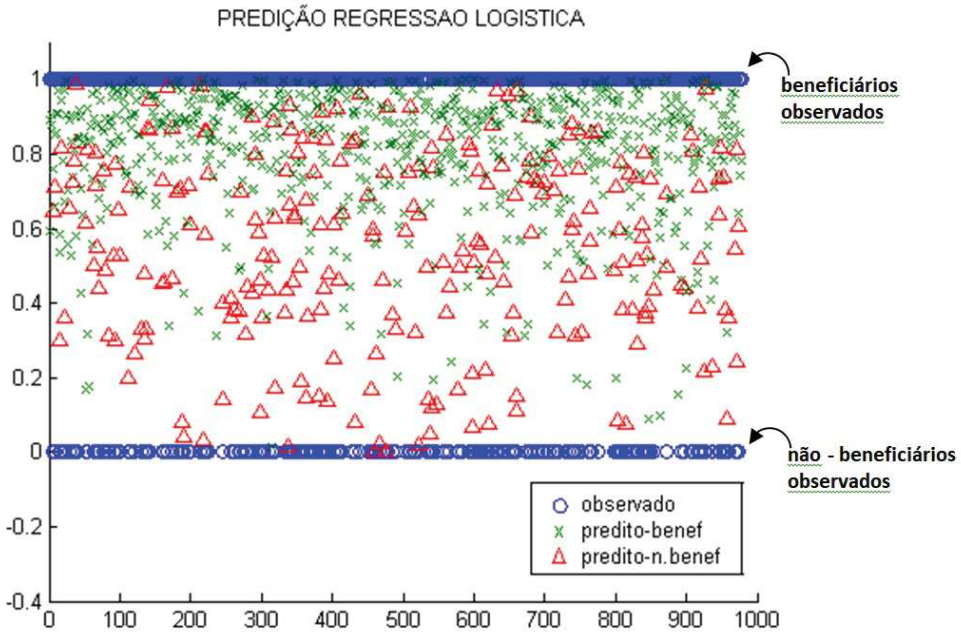


Figura 5.2: Dispersão dos dados simulados na RL

*fonte: elaboração própria*

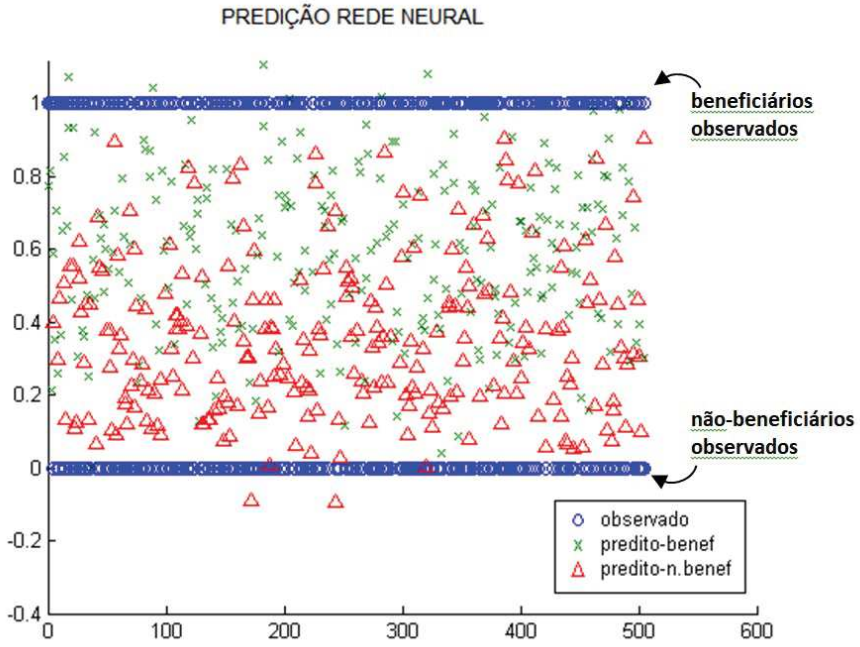


Figura 5.3: Dispersão dos dados simulados na RNA
*fonte: elaboração própria*

Da matriz de confusão levantamos a capacidade de classificação de cada método aplicando várias das métricas (descritas na sessão 4.2) a fim de eleger a que melhor represente o desempenho do classificador como um todo. Lembrando que, como métrica individual temos a *sensibilidade* e a *especificidade*. Das métricas globais algumas foram bastante restritivas - influenciadas pelo baixo desempenho da classificação da classe minoritária, como foi o caso dos coeficientes Matthews, Youden e kappa. Ao passo que outras foram fortemente influenciadas pela sensibilidade, e portanto pela classe majoritária (acurácia, AUC, *f-measure* e eficiência), podendo levar a uma interpretação errônea já que um classificador não pode ser considerado de bom desempenho se falha na identificação de um dos grupos, quando é necessário considerar o desempenho de ambas as classes. Este fato pode ser observado na tabela 5.1, na aplicação da ABD e RNA, usando a amostra sem intervenção, onde os valores de acurácia, eficiência e *f-measure* são bastante elevados e destoam dos valores encontrados para especificidade, com exceção da regressão logística, onde sensibilidade e especificidade resultaram em números próximos - mesmo com intervenção (reamostragem) na amostra e se mantém na proximidade dos 70% - e portanto, para RL as métricas eficiência, *g-mean*, acurácia, *f-measure* e AUC, para o classificador como um todo, não destoam dos valores de cada classe (sensibilidade e especificidade). Já os valores para o MCC, Youden e Kappa são muito restritivos e foram gravemente penalizados pelos erros de classificação para os não-beneficiários.

Tabela 5.1: Valores das médias das métricas aplicadas aos classificadores antes e após intervenção na amostra em 50 conjuntos de teste

| | Sem Intervenção | | | Beneficiários (sub-amostrado) | | | Não-beneficiários (super-amostrado) | | | Intervenção em ambos os grupos | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | **ABD** | **RL** | **RNA** | **ABD** | **RL** | **RNA** | **ABD** | **RL** | **RNA** | **ABD** | **RL** | **RNA** |
| Sensibilidade | 0,9241 | 0,7414 | **0,9405** | **0,7980** | 0,7569 | 0,7174 | **0,8653** | 0,7479 | 0,8433 | **0,7572** | 0,7508 | 0,6930 |
| Especificidade | 0,4360 | **0,7658** | 0,2510 | 0,6076 | **0,6929** | 0,5839 | 0,5539 | **0,7241** | 0,5505 | 0,6505 | **0,7149** | 0,6742 |
| Suporte | 0,6965 | **0,7349** | 0,7071 | **0,4159** | 0,3316 | 0,3743 | **0,5267** | 0,2832 | 0,5131 | **0,3780** | 0,3581 | 0,3458 |
| Cobertura | 0,8355 | 0,6704 | **0,8931** | **0,6039** | 0,4781 | 0,5730 | **0,7014** | 0,3913 | 0,6894 | **0,5530** | 0,5009 | 0,5094 |
| Conf. Pos. (VPP) | 0,6536 | 0,7083 | **0,7935** | 0,6902 | **0,6929** | 0,6598 | 0,7515 | 0,7241 | **0,7526** | **0,6838** | 0,7149 | 0,6834 |
| Conf. Neg.(VPN) | **0,8765** | 0,7658 | 0,5386 | 0,7369 | **0,7569** | 0,6515 | 0,7273 | **0,7479** | 0,6583 | 0,7301 | **0,7508** | 0,6862 |
| Acurácia | **0,8037** | 0,7083 | 0,7692 | 0,7060 | **0,7268** | 0,6539 | **0,7432** | 0,7385 | 0,7281 | 0,7038 | **0,7330** | 0,6832 |
| F-measure | **0,8338** | 0,7371 | 0,7935 | 0,7200 | **0,7525** | 0,6053 | **0,7808** | 0,7435 | 0,7115 | **0,7572** | 0,7508 | 0,6374 |
| Área sob a curva | **0,8011** | **0,8011** | 0,7071 | **0,7964** | **0,7964** | 0,7043 | **0,8081** | **0,8081** | 0,7728 | **0,8099** | **0,8099** | 0,7441 |
| Eficiência | 0,6800 | **0,7516** | 0,5957 | 0,7028 | **0,7250** | 0,6506 | 0,7096 | **0,7360** | 0,6969 | 0,7038 | **0,7329** | 0,6836 |
| G-mean | 0,6324 | **0,7340** | 0,4538 | 0,6945 | **0,7233** | 0,6478 | 0,6914 | **0,7355** | 0,6587 | 0,7012 | **0,7320** | 0,6908 |
| MCC | **0,4285** | 0,2484 | 0,2491 | **0,4522** | **0,4522** | 0,3061 | **0,4649** | **0,4649** | 0,4063 | **0,4667** | **0,4667** | 0,3685 |
| Índice de Youden | 0,3602 | **0,4285** | 0,2081 | 0,4055 | **0,4498** | 0,3177 | 0,4192 | **0,4720** | 0,4047 | 0,4077 | **0,4657** | 0,3846 |
| Kappa | 0,3234 | **0,5480** | 0,1683 | 0,3527 | **0,4351** | 0,2881 | 0,3653 | **0,5112** | 0,3686 | 0,3690 | **0,4506** | 0,3691 |
| Taxa de erro global | **0,1963** | 0,2917 | 0,2308 | 0,2940 | **0,2732** | 0,3461 | **0,2568** | 0,2615 | 0,2719 | 0,2962 | **0,2670** | 0,3168 |
| Taxa de erro positiva | 0,5640 | **0,2463** | 0,7490 | 0,3924 | **0,2702** | 0,4161 | 0,4461 | **0,1913** | 0,4495 | 0,3495 | **0,2752** | 0,3258 |
| Taxa de erro negativa | 0,0759 | 0,2586 | **0,0595** | **0,2020** | 0,2431 | 0,2826 | **0,1347** | 0,2521 | 0,1567 | **0,2428** | 0,2492 | 0,3070 |

**valores médios após 50 rodadas - melhores valores destacados*
fonte: elaboração própria

Nas sessões a seguir serão apresentadas em separado as métricas resultantes (tabela 5.1) de cada um dos algoritmos classificadores.

### *5.1.1 Árvore Binária de Decisão*

A amostra foi dividida em treinamento e teste sendo a quantidade de elementos de 2.278 (70%) e 976 (30%) respectivamente. A classificação de cada grupo apresentou desempenho bastante distinto (tabela 5.2) - enquanto a *sensibilidade* alcançada permaneceu entre 76% e 92%, a *especificidade* foi consideravelmente baixa, entre 43% e 65%. Ainda assim, a eficiência (que é a média do desempenho dos dois grupos) figura um desempenho muito otimista frente a especificidade, ao passo que MCC, Youden e Kappa se mostram muito punitivos mesmo para uma especificidade de 43%.

Observando na tabela 5.2 a distância entre *sensibilidade* e *especificidade*, a melhor medida de desempenho aponta para a média geométrica (*g-mean*), principalmente quando esta diferença se torna considerável como na aplicação do classificador sem intervenção na amostra (tabela 5.2 – 1ª coluna) que, tanto antes como depois da intervenção resultou em um valor nada satisfatório devido a perda entre 30% e 40% (vide valor de *g-mean*).

Tabela 5.2: Média das métricas de desempenho para a árvore binária de decisão

| | **Sem intervenção (2.453 vs. 801)** | **Sub-amostr. *(beneficiários)* (881 vs. 801)** | **Super-amostr. *(não-benef.)* (2.453 vs. 2.163)** | **Intervenção em ambos os grupos (1602 vs. 1602)** |
|---|---|---|---|---|
| Sensibilidade | ***0,9240*** (0,0134) | 0,7980 (0,0544) | 0,8653 (0,0253) | 0,7572 (0,0346) |
| Especificidade | 0,4343 (0,0469) | 0,6076 (0,0536) | 0,5539 (0,0445) | ***0,6505*** (0,0329) |
| Acurácia | ***0,8024*** (0,0129) | 0,7060 (0,0185) | 0,7432 (0,0110) | 0,7038 (0,0155) |
| *F-measure* | ***0,9240*** (0,0134) | 0,7980 (0,0544) | 0,8653 (0,0253) | 0,7572 (0,0346) |
| Área sob a curva | 0,8011 (0,0147) | 0,7964 (0,0168) | 0,8081 (0,0096) | ***0,8099*** (0,0116) |
| Eficiência | 0,6791 (0,0204) | 0,7028 (0,0187) | ***0,7096*** (0,0140) | 0,7038 (0,0150) |
| *G-mean* | 0,6324 (0,0327) | 0,6945 (0,0198) | 0,6914 (0,0204) | ***0,7012*** (0,0151) |

Tabela 5.2: Média das métricas de desempenho para a árvore binária de decisão (continuação)

| | **Sem intervenção (2.453 vs. 801)** | **Benef. (sub-amostr.) (881 vs. 801)** | **Não-benef. (super-amostr.) (2.453 vs. 2.163)** | **Intervenção em ambos os grupos (1602 vs. 1602)** |
|---|---|---|---|---|
| MCC | 0,4243 (0,0267) | 0,4522 (0,0324) | 0,4649 (0,0231) | ***0,4667*** (0,0251) |
| Índice de Youden | 0,3584 (0,0410) | 0,4055 (0,0371) | ***0,4192*** (0,0280) | 0,4077 (0,0301) |
| Kappa | 0,3234 (0,0429) | 0,3527 (0,0469) | 0,3653 (0,0392) | ***0,3690*** (0,0344) |
| Taxa de erro global | ***0,1976*** (0,0129) | 0,2940 (0,0185) | 0,2568 (0,0110) | 0,2962 (0,0155) |
| Taxa de erro positiva | 0,5657 (0,0469) | 0,3924 (0,0536) | 0,4461 (0,0445) | ***0,3495*** (0,0329) |
| Taxa de erro negativa | ***0,0760*** (0,0134) | 0,2020 (0,0544) | 0,1347 (0,0253) | 0,2428 (0,0346) |

*$nota_1$: valores médios após 50 rodadas - destaque para as métricas de melhor valor entre os classificadores

*$nota_2$: no corpo da tabela, os números entre parênteses são o desvio-padrão

*$nota_3$: no cabeçalho, os valores entre parênteses são o número de elementos para beneficiários e não-beneficiáios respectivamente

*fonte: elaboração própria*

### *5.1.2 Regressão Logística*

Foi utilizada no processo de seleção das variáveis que compõem o modelo a fim de eliminar as que se mostraram redundantes ou de pouca significância. A partir daí, as 32 variáveis escolhidas e as amostras foram as mesmas para todos os algoritmos classificadores - a amostra, composta por 3.254 elementos entre benficiários e não-beneficiários, foi dividida em treinamento (2.278 - 70%) e teste (976 - 30%). Como classificador, observando-se a medida *g-mean*, seu desempenho ficou entre 73% e 75% (tabela 5.3). Porém esta faixa ainda representa um indicativo baixo de desempenho (perda de quase 30%) e assim, não será considerado satisfatório. As métricas acurácia, eficiência e *f-measure*, nesse classificador, não destoam do valor de desempenho individual encontrado para cada grupo (*sensibilidade* e *especificidade*) estando próximos também da média geométrica (*g-mean*) que permaneceu como opção prioritária de análise de desempenho global por continuar figurando em valor condizente com o grau de acertabilidade dos dois grupos (*sensibilidade* e *especificidade*). Os valores para o MCC, Youden e kappa continuam bastante restritivos.

Tabela 5.3: Média das métricas de desempenho para a regressão logística

| | **Sem intervenção (2.453 vs. 801)** | **Sub-amostr. *(beneficiários)* (881 vs. 801)** | **Super-amostr. *(não-benef.)* (2.453 vs. 2.163)** | **Intervenção em ambos os grupos (1602 vs. 1602)** |
|---|---|---|---|---|
| Sensibilidade | ***0,7618*** (0,0215) | 0,7569 (0,0400) | 0,7479 (0,0269) | 0,7508 (0,0337) |
| Especificidade | 0,7079 (0,0337) | 0,6929 (0,0413) | ***0,7241*** (0,0293) | 0,7149 (0,0330) |
| Acurácia | ***0,7483*** (0,0144) | 0,7268 (0,0164) | 0,7385 (0,0122) | 0,7330 (0,0126) |
| *F-measure* | ***0,7618*** (0,0215) | 0,7569 (0,0400) | 0,7479 (0,0269) | 0,7508 (0,0337) |
| Área sob a curva | 0,8011 (0,0147) | 0,7964 (0,0168) | 0,8081 (0,0096) | ***0,8099*** (0,0116) |
| Eficiência | 0,7349 (0,0153) | 0,7250 (0,0163) | ***0,7360*** (0,0115) | 0,7329 (0,0125) |
| *G-mean* | 0,7340 (0,0158) | 0,7233 (0,0169) | ***0,7355*** (0,0117) | 0,7320 (0,0128) |
| MCC | 0,4243 (0,0267) | 0,4522 (0,0324) | 0,4649 (0,0231) | ***0,4667*** (0,0251) |
| Índice de Youden | 0,4697 (0,0304) | 0,4498 (0,0326) | ***0,4720*** (0,0230) | 0,4657 (0,0251) |
| Kappa | ***0,5480*** (0,0429) | 0,4351 (0,0429) | 0,5112 (0,0355) | 0,4506 (0,0356) |
| Taxa de erro global | ***0,2517*** (0,014) | 0,2732 (0,0164) | 0,2615 (0,0122) | 0,2670 (0,0126) |
| Taxa de erro positiva | ***0,1123*** (0,0135) | 0,2702 (0,0229) | 0,1913 (0,0164) | 0,2752 (0,0198) |
| Taxa de erro negativa | 0,5043 (0,0251) | 0,2753 (0,0265) | 0,3507 (0,0248) | ***0,2569*** (0,0209) |

*$nota_1$: valores médios após 50 rodadas - destaque para as métricas de melhor valor entre os classificadores
*$nota_2$: no corpo da tabela, os números entre parênteses são o desvio-padrão
*$nota_3$: no cabeçalho, os valores entre parênteses são o número de elementos para beneficiários e não-beneficiáios respectivamente

*fonte: elaboração própria*

### *5.1.3 RNA - MLP*

A amostra foi dividida em treinamento, validação e teste sendo os valores de 1.595 (49%), 683 (21%) e 976 (30%) respectivamente. Da tabela 5.4 temos que, com a amostra original (1ª coluna), no que diz respeito a identificação de cada grupo, esta apresentou a maior das sensibilidades (94%), mas a pior das especificidades (25%). Acurácia, eficiência, *f-measure* induzem a um desempenho que não cabe ao classificador pela deficiência constatada frente a classe minoritária; e MCC, Youden e kappa, como nos outros algoritmos classificadores aqui aplicados, se mostram

excessivamente restritivos.

Após sub-amostragem, os ganhos do grupo minoritário ainda foram baixos - especificidade alcançou 58% e a sensibilidade caiu para 72%, algo que não era esperado. Novamente, quando a predição de ambas as classes é próxima, como na reamostragem de ambos os grupos, acurácia, eficiência, *f-measure* se mostram coerentes (próximo de 70%) e MCC, Youden e Kappa permanecem muito pessimistas (entre 24% e 41%).

Mesmo com o recurso da reamostragem, a AUC se manteve entre 70% e 77% mas induz a assumir um desempenho muito otimista face a especificidade de 25% (1ª coluna). *G-mean* continua apresentando valores coerentes após reamostragem. Por isso, também nesse classificador é escolhida como melhor opção de avaliação global, frente aos valores encontrados para cada grupo (sensibilidade e especificidade).

Tabela 5.4: Média das métricas de desempenho para a rede neural

| | **Sem intervenção (2.453 vs. 801)** | **Sub-amostr. *(beneficiários)* (881 vs. 801)** | **Super-amostr. *(não-benef.)* (2.453 vs. 2.163)** | **Intervenção em ambos os grupos (1602 vs. 1602)** |
|---|---|---|---|---|
| Sensibilidade | ***0,9405*** (0,0369) | 0,7174 (0,0754) | 0,8433 (0,0518) | 0,6930 (0,0674) |
| Especificidade | 0,2510 (0,1484) | 0,5839 (0,1355) | 0,5505 (0,1915) | ***0,6742*** (0,1039) |
| Acurácia | ***0,7692*** (0,0201) | 0,6539 (0,0532) | 0,7281 (0,0537) | 0,6832 (0,0723) |
| *F-measure* | ***0,7935*** (0,0296) | 0,6598 (0,0562) | 0,7526 (0,0612) | 0,6834 (0,0818) |
| Área sob a curva | 0,7071 (0,0670) | 0,7043 (0,0749) | ***0,7728*** (0,0837) | 0,7441 (0,0909) |
| Eficiência | 0,5957 (0,0585) | 0,6506 (0,0564) | ***0,6969*** (0,0761) | 0,6836 (0,0722) |
| *G-mean* | 0,4538 (0,1964) | 0,6478 (0,0713) | 0,6587 (0,1768) | ***0,6908*** (0,0746) |
| MCC | 0,2491 (0,1255) | 0,3061 (0,1116) | ***0,4063*** (0,1506) | 0,3685 (0,1443) |
| Índice de Youden | 0,2081 (0,1229) | 0,3177 (0,1067) | ***0,4047*** (0,1584) | 0,3846 (0,1470) |
| Kappa | 0,1683 (0,1077) | 0,2881 0,1185) | 0,3686 (0,1508) | ***0,3691*** (0,1585) |

Tabela 5.4 Média das métricas de desempenho para a rede neural

| | **Sem intervenção (2.453 vs. 801)** | **Sub-amostr. *(beneficiários)* (881 vs. 801)** | **Super-amostr. *(não-benef.)* (2.453 vs. 2.163)** | **Intervenção em ambos os grupos (1602 vs. 1602)** |
|---|---|---|---|---|
| Taxa de erro global | ***0,2308*** (0,0201) | 0,3461 (0,0532) | 0,2719 (0,0537) | 0,3168 (0,0723) |
| Taxa de erro positiva | 0,7490 (0,1484) | 0,4161 (0,1355) | 0,4495 (0,1915) | ***0,3258*** (0,1039) |
| Taxa de erro negativa | ***0,0595*** (0,0369) | 0,2826 (0,0754) | 0,1567 (0,0518) | 0,3070 (0,0674) |

*$nota_1$: valores médios após 50 rodadas - destaque para as métricas de melhor valor entre os classificadores

*$nota_2$: no corpo da tabela, os números entre parênteses são o desvio-padrão

*$nota_3$: no cabeçalho, os valores entre parênteses são o número de elementos para beneficiários e não-beneficiáios respectivamente

*fonte: elaboração própria*

### *5.1.4 Considerações finais*

O melhor classificador será aquele que apresentar menos perdas e estas perdas precisam estar em uma faixa aceitável. Conforme registrado nas tabelas 5.2, 5.3 e 5.4, houve melhora na métrica *especificidade* frente ao resultado observado na amostra original, mas ainda assim as perdas são de cerca de 30%, o que é ainda um percentual bastante alto. Conforme mensionado anteriormente, a amostra original é composta por 3.254 domicílios de várias regiões do país, onde 2.453 são beneficiários e 801 não. As métricas de desempenho individual (*sensibilidade* e *especificidade*) sinalizam grande distância, e por isso confrontou-se a aplicação dos classificadores na amostra original e após três intervenções na amostra:

**sub-amostragem da classe majoritária** - feita de forma aleatória, foi fixada em 881 domicílios, 10% a mais que a classe minoritária. Em todos os três algoritmos de classificação houve uma perda na métrica da sensibilidade e apenas no classificador RL houve perda na especificidade, de 2,119%.

**sobre-amostragem da classe minoritária** - também de forma aleatória a quantidade de domicílios foi replicada passando de 801 para 2.163, o que resultou em um acréscimo de 170%. Conforme a tabela E.1, RL não apresentou melhora tão acentuada como os outros dois algoritmos na métrica *especificidade* (2,288%) mas se

mostrou como método mais equilibrado frente a natureza desbalanceada dos dados (tabela 5.1) e se manteve estável, tendo variado de 0,6029 a 0,7241. As perdas na *sensibilidade* formam menos acentuadas que na sub-amostragem.

**sobre-amostragem e sub-amostragem** - nesta intervenção, tanto a classe majoritária foi sub-amostrada em 1.602 elementos quanto a classe minoritária foi sobre-amostrada até atingir 1.602 elementos. Com os grupos equiparados foi o melhor desempenho para ABD e RN - 49,781% e 168,606% respectivamente. Os valores alcançados podem ser conferidos na tabela 5.1.

Dada a importância de reconhecimento das duas classes, as referências de controle do classificador precisam ser a métrica individual especificidade, por se tratar da medida de reconhecimento da classe minoritária, e a métrica global *g-mean* que se mostrou a mais coerente frente as medidas de cada grupo. Uma vez que *g-mean* é resultado do produto das duas medidas individuais de cada grupo, quanto mais próximas estas estiverem, melhor é o valor desta métrica.

Também devido as métricas individuais, no algoritmo classificador RNA encontram-se as maiores distâncias entre estas duas medidas - o que influencia também o valor da área sob a curva (AUC), que se mostra mais baixo sob a aplicação deste classificador.

# 6 Conclusões

O desempenho destes algoritmos foi avaliado a partir de métricas decorrentes da matriz de confusão, que registra em suas linhas e colunas os erros e acertos da predição. Como os erros e acertos de uma classe não são informações complementares da outra, é importante que ambas as classes sejam corretamente identificadas – tanto a de *"beneficiários"* como a de *"não-beneficiários"* para que se possa avaliar o comportamento de classificadores aplicados sobre bases de caráter social pois estas apresentam algumas particularidades, como o fato de serem desbalanceadas. Nesta situação os algoritmos tradicionais geram modelos que falham no reconhecimento de classes poucos representadas (classes minoritárias).

Os resultados obtidos apontam para a necessidade de uma abordagem diferenciada tendo em vista a deficiência na identificação de uma das classes. Existem na literatura relatos semelhantes, que atribuem tais resultados ao fato das técnicas tradicionais de classificação maximizarem a precisão (item 2.3 deste trabalho) em relação ao conjunto de dados, supondo estes com distribuição equilibrada - as classes em estudo com um número equilibrado de elementos [13]. A precisão se refere a razão de valores preditos relacionados ao conceito de "positivos" *versus* valores observados de fato. Esta classe "positiva" corresponde a classe majoritária, os "beneficiários". Ocorre que a classe minoritária não tem a mesma distribuição nem a mesma representatividade (figura 4.12). Se considerados todos os domicílios da base de estudo com características de beneficiários [1] a proporção seria de 9:1 e não 3:1 aumentando ainda mais a desproporcionalidade entre os grupos.

Conforme observado no presente estudo (tabela 5.1), apenas a reamostragem simples, realizada de forma aleatória, não trouxe o ganho esperado permanecendo a classe minoritária ainda negligenciada (perdas em torno de 30%). Comprova-se então que não há um algoritmo único capaz de atender a todas as tarefas, visto que a intervenção que melhorou a métrica *especificidade* foi a mesma que penalizou a *sensibilidade*. O quadro 6.1 apresenta estes desempenhos segundo as intervenções sofridas na amostra enquanto o quadro 6.2 organiza esse desempenho segundo o classificador.

[1] antes de submeter a amostra aos algoritmos de classificação, foi descartado um grupo com características de beneficiários mas na espera do recebimento do benefício

Quadro 6.1: Desempenho dos classificadores segundo as intervenções na amostra

| | SUB-AMOSTRAGEM | | | SOBRE-AMOSTRAGEM | | | AMBOS | | |
|---|---|---|---|---|---|---|---|---|---|
| | **ABD** | **RL** | **RNA** | **ABD** | **RL** | **RNA** | **ABD** | **RL** | **RNA** |
| Sensibilidade | -13,636% | -0,643% | -23,721% | -6,353% | -1,825% | -10,335% | -18,052% | -1,444% | -26,316% |
| Especificidade | 39,903% | -2,119% | 132,629% | 27,539% | 2,288% | 119,323% | 49,781% | 0,989% | 168,606% |
| Acurácia | -12,014% | -2,873% | 42,750% | -7,378% | -1,310% | 45,152% | -52,891% | -52,145% | -28,206% |
| *F-measure* | -13,636% | -0,643% | 22,882% | -6,353% | -1,825% | 63,107% | -25,996% | -6,156% | 54,396% |
| Área sob a curva | -0,587% | -0,587% | 52,667% | 0,874% | 0,874% | 94,474% | -63,026% | -66,671% | 77,367% |
| Eficiência | 3,490% | -1,347% | 71,182% | 4,491% | 0,150% | 119,014% | -18,569% | -31,841% | 82,412% |
| *G-mean* | 9,820% | -1,458% | -47,065% | 9,330% | 0,204% | -27,436% | -44,734% | -62,507% | -3,352% |

*fonte: elaboração própria*

Quadro 6.2: Desempenho de cada classificador frente às intervenções na amostra

| | ÁRVORE | | | REGRESSÃO LOGÍSTICA | | | REDE NEURAL | | |
|---|---|---|---|---|---|---|---|---|---|
| | **Under** | **Over** | **Both** | **Under** | **Over** | **Both** | **Under** | **Over** | **Both** |
| Sensibilidade | -13,636% | -6,353% | -18,052% | -0,643% | -1,825% | -1,444% | -23,721% | -10,335% | -26,316% |
| Especificidade | 39,903% | 27,539% | 49,781% | -2,119% | 2,288% | 0,989% | 132,629% | 119,323% | 168,606% |
| Acurácia | -12,014% | -2,873% | 42,750% | -7,378% | -1,310% | 45,152% | -52,891% | -52,145% | -28,206% |
| *F-measure* | -13,636% | -0,643% | 22,882% | -6,353% | -1,825% | 63,107% | -25,996% | -6,156% | 54,396% |
| Área sob a curva | -0,587% | -0,587% | 52,667% | 0,874% | 0,874% | 94,474% | -63,026% | -66,671% | 77,367% |
| Eficiência | 3,490% | -1,347% | 71,182% | 4,491% | 0,150% | 119,014% | -18,569% | -31,841% | 82,412% |
| *G-mean* | 9,820% | -1,458% | -47,065% | 9,330% | 0,204% | -27,436% | -44,734% | -62,507% | -3,352% |

*nota: **under**sampling (sub-amostragem) **over**sampling (sobre-amostragem) **both** ⇒ sobre-amostragem e sub-amosragem
*fonte: elaboração própria*

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. Uma possibilidade é o fato de os valores médios de faixas de renda serem diferentes de uma região para outra. Pela tabela 6.1 a distribuição de beneficiários e não-beneficiários é a seguinte:

Tabela 6.1: Distribuição dos domicílios segundo a região geográfica

| **Região** | **Não-benef.** | | **Beneficiários** | | **Total** | |
|---|---|---|---|---|---|---|
| N | 130 | (4,00%) | 432 | (13,28%) | 562 | (17,27%) |
| NE | 259 | (7,96%) | 850 | (26,12%) | 1.109 | (34,08%) |
| S | 44 | (1,35%) | 144 | (4,43%) | 188 | (5,78%) |
| SE | 297 | (9,13%) | 746 | (22,93%) | 1.043 | (32,05%) |
| CO | 71 | (2,18%) | 281 | (8,64%) | 352 | (10,82%) |
| | **801** | (24,62%) | **2.453** | (75,38%) | **3.254** | (100,00%) |

*fonte: elaboração própria*

Técnicas, tanto de pré-processamento quanto de classificação precisam ser experimentadas a fim de possibilitar o atendimento de novos nichos. Principalmente em bases de dados onde os valores das variáveis de cada grupo são próximos, como ocorre nesta base de dados - lembrando que variáveis que, segundo a literatura poderiam ser de grande importância apresentaram muitas lacunas como dados referentes ao índice de massa corpórea (IMC), que poderia ser calculado e identificar a situaçao nutricional, mas não pode ser levantada pela falta de dados inerentes a este cálculo. Assim como dados referente a fontes de renda, que também apresentou consideráveis lacunas.

# 7 Trabalhos Futuros

O foco deste estudo não visa avaliar o desempenho do PBF, mas o comportamento de classificadores aplicados sobre bases de dados de caráter social, pois estas apresentam certas particularidades como o desbalanceamento. O propósito é averiguar a capacidade dos algoritmos classificadores quanto a identificação das famílias que recebem o benefício e as que não o recebem (informação esta passível de comparação a partir de variável disponível na base de dados).

Uma vez encontrado o algoritmo classificador que permita uma correta leitura dos grupos, é possível a construção de cenários onde, dada uma coleção de informações a respeito do objeto de estudo ao longo de um período, é possível vislumbrar seu comportamento em períodos adiante. Assim como o histórico hidrográfico permite o planejamento de produção de energia de uma hidrelétrica ou de um investimento agrário em uma nova cultura ou novo local de plantio. Da mesma forma também possibilitaria projeções de ações de cunho social dado o acesso a registros sobre estas ações em um dado período de tempo.

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 classes majoritárias. Dentre as técnicas citadas na literatura utilizada como referência ([3], [24], [27] e [28]), algumas das opções figuram na introdução de custos de classificação incorreta – que traz o desafio de encontrar os valores de tais custos. Ou utilizar formas de re-amostragem 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. As referências [13] e [3] citam como alternativas de intervenção os links de Tomek, edited nearest neighbor rule (ENN), método boundary elimination and domination algorithm (BED), máquina de vetor suporte (support vector machine – SVM), 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

[1] DO DESENVOLVIMENTO SOCIAL E COMBATE À FOME; CENTRO DE DESENVOLVIMENTO E PLANEJAMENTO REGIONAL., B. M., *Sumário executivo – avaliação de impacto do Programa Bolsa Família – 2ª Rodada*, Tech. rep., Ministério do Desenvolvimento Social e Combate à Fome; Centro de Desenvolvimento e Planejamento Regional, jun 2012.

[2] ALBERTO, B., *Abordagens de pré-processamento de dados em problemas de classificação com classes desbalanceadas.*, Master's Thesis, Centro Federal de Educação Tecnológica de Minas Gerais (Mestrado em Modelagem Matemática e Computacional), aug 2012.

[3] BATISTA, G., PRATI, R., MONARD, M., "A study of the behavior of several methods for balancing machine learning training data", *ACM Sigkdd Explorations Newsletter*, v. 6, n. 1, pp. 20–29, 2004.

[4] MONARD, M. C. ; BARANAUSKAS, J. A., "Conceitos sobre aprendizado de máquina", In: *Sistemas Inteligentes - Fundamentos e Aplicações*, 1st ed., chap. 4, pp. 89–114, Editora Manole Ltda, 2003.

[5] RUFINO, H. L. P., *Algoritmo de aprendizado supervisionado-baseado em máquinas de vetores de suporte - uma contribuição para o reconhecimento de dados desbalanceados*, Ph.D. Thesis, Universidade Federal de Uberlândia, sep 2011.

[6] SENNA, M. C. M., BRANDÃO, A. A.AND DALT, S., "Programa Bolsa Família e o acompanhamento das condicionalidades na área de saúde", *Serviço Social & Sociedade*, , n. 125, pp. 148–166, jan 2016.

[7] AMARAL, E., GONÇALVES, G., MONTEIRO, V., SANTOS I.J., SANTOS, A., "Avaliação de Impactos das Condicionalidades de Educação do Programa Bolsa Família: uma Análise com o Censo de 2010", *Anais do XVIII Encontro Nacional de Estudos Populacionais*, pp. 16p., nov 2012.

[8] 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", *Revista Vozes do Vale. Ano I*, v. 1, n. 03, pp. 1–31, may 2012.

[9] MUNARETTO, L. F., E. A., "Um estudo sobre Programa Bolsa Família (PBF): o caso dos municípios que integram a associação dos municípios da zona da produção (AMZOP)", *IV Simpósio Internacional de Gestão de Projetos, Inovação e Sustentabilidade (IV SINGEP)*, nov 2015.

[10] NETO, S. B., NAGANO, M. S., DA COSTA MORAES, M. B., "Utilização de redes neurais artificiais para avaliação socioeconômica: uma aplicação em cooperativas", *Revista de Administração*, v. 41, n. 1, pp. 59–68, jan 2006.

[11] DUARTE, G., 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, pp. 903–918, oct 2009.

[12] SILVA, C. C. S. E. A., "Rede neural artificial e o modelo de apoio à decisão em segurança alimentar nutricional", *Revista de enfermagem UFPE*, v. 9, n. 3, pp. 7078–7085, mar 2015.

[13] CASTRO, C.L. ; BRAGA, A., "Aprendizado supervisionado com conjuntos de dados desbalanceados." *Revista Controle e Automação*, v. 22, n. 5, pp. 441–466, sep 2011.

[14] CAMILO, C.O. ; SILVA, J., *Mineração de Dados: Conceitos, Tarefas, Métodos e Ferramentas*, Tech. rep., Universidade Federal de Goias Instituto de Informática, aug 2009.

[15] MONARD, M. C. ; BARANAUSKAS, J. A., "Indução de Regras e Árvores de Decisão", In: *Sistemas Inteligentes - Fundamentos e Aplicações*, 1st ed., chap. 5, pp. 115–139, Editora Manole Ltda, 2003.

[16] FÁVERO, L.P.L; BELFIORE, P., SILVA, F., B.L., C., *Análise de Dados - Modelagem Multivariada para Tomada de Decisões*. 1st ed. Campus Editora, 2009.

[17] PRINCIPE, J. C., EULIANO, N. R., LEFEBVRE, W. C., *Neural and adaptive systems: fundamentals through simulations*. Wiley New York, 1999.

[18] BARANAUSKAS, J., *Aprendizado de máquina conceitos e definições*, 2007.

[19] SOUZA, F. C. S. D., "Métricas de avaliação de modelos de classificação/predição." internet, 2014.

[20] MATOS, P. F. E. A., *Relatório técnico "Métricas de Avaliaçao"*, Tech. rep., Universidade Federal de São Carlos (UFScar), sep 2009.

[21] ANDRADE, A.L.S.S., Z. F., "Avaliação de testes diagnósticos". In: *Métodos de Investigação Epidemiológica em Doenças Transmissíveis*, 1, 1997.

[22] WIKIPEDIA, "Youden's J statistic — Wikipedia, The Free Encyclopedia", 2016.

[23] GONZAGA, A., "Métodos de avaliação de Classificadores", 2011.

[24] 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.

[25] *Modelagem de distribuição geográfica para Hydromedusa maximiliani (Mikan, 1820)(Testudines, Chelidae)*, Master's Thesis, Universidade Federal de Juiz de Fora - ICB - Programa de Pós-graduação em Ciências Biológicas: Comportamento e Biologia Animal, feb 2014.

[26] DE ASSIS TENÓRIO DE CARVALHO, F., "Aprendizagem Estatística de Dados", 2010.

[27] "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, pp. 137–144, jan 2016.

[28] SCHIAVONI, A. S., *Um estudo comparativo de métodos para balanceamento do conjunto de treinamento em aprendizado de redes neurais artificiais*, Master's Thesis, Universidade Federal de Lavras – MG, mar 2015.

[29] PRETTO, D., BENDER FILHO, R., "Análise da influência dos programas complementares para a emancipação sustentada dos benefíciários vinculados ao programa bolsa família: estudo com ex-beneficiários do município de Santo Ângelo/RS", *Gestão Pública: Práticas e Desafios*, v. 8, n. 2, pp. 19p., oct 2015.

[30] *Bolsa Família - Transferência de Renda e Apoio à Família no Acesso à Saúde, à Educação e à Assistência Social*, publicação técnica da Secretaria de Avaliação e Gestão da Informação (SAGI / MDS) 30, Secretaria de Avaliação e Gestão da Informação (SAGI / MDS), mai 2015.

[31] DO DESENVOLVIMENTO SOCIAL E COMBATE À FOME; CENTRO DE DESENVOLVIMENTO E PLANEJAMENTO REGIONAL, B. M., DE PESQUISA SOBRE POLÍTICAS ALIMENTARES (IFPRI2), I. I., DATAMÉTRICA CONSULTORIA, P. E. T. L., *Questionário 2009 – avaliação de impacto do Programa Bolsa Família – 2ª Rodada*, Tech. rep., Ministério do Desenvolvimento Social e Combate à Fome; Centro de Desenvolvimento e Planejamento Regional, 2009.

# APÊNDICE A - Política Social de Transferência de Renda

## Sobre o Programa Bolsa Família

O Programa Bolsa Família (PBF) foi criado em 2003, a partir da lei nº 10.836, passando a integrar outras políticas sociais preexistentes (Programas Fome Zero, Bolsa Escola, Bolsa Alimentação e Auxílio-Gás). A família deve estar cadastrada no Cadastro Único para Programas Sociais do Governo Federal (Cadastro Único ou CADÚnico), instrumento que identifica e caracteriza as famílias de baixa renda onde constam informações como características da residência, identificação de cada pessoa, escolaridade, situação de trabalho e renda [29]. Podem se inscrever no Cadastro Único:

- Famílias com renda mensal de até meio salário mínimo por pessoa;
- Famílias com renda mensal total de até três salários mínimos; ou
- Famílias com renda maior que três salários mínimos, desde que o cadastramento esteja vinculado à inclusão em programas sociais numa das três esferas do governo.

Pessoas que moram sozinhas podem se cadastrar (famílias unipessoais), assim como pessoas que vivem em situação de rua - sozinhas ou com a família. O acompanhamento da família e a inserção da mesma no CADÚnico são de responsabilidade dos municípios.

Os vulneráveis-alvo do programa são crianças (de 0 e 15 anos), gestantes e nutrizes, além de jovens (de 16 a 17 anos). O programa visa assistir domicílios em situação de pobreza (entre R$ 85,01 e R$ 170,00 percapita) e extrema pobreza (até R$ 85,00 percapita).

O recebimento se dá através de cartão magnético, emitido pela Caixa Econômica Federal e varia de acordo com a necessidade da família cadastrada além do cumprimento de algumas condicionantes, podendo ter em sua composição parte ou todos os elementos descritos a seguir:

1. **Benefício Básico:** R$ 85,00 concedidos apenas a famílias extremamente pobres, com renda per capita igual ou inferior a R$ 85,00 - desde a implantação do PBF este valor vem representando cerca de 10% do salário mínimo a partir de 2012;

2. **Benefício Variável (BV):** R$ 39,00 concedidos pela existência na família de crianças de zero a 15 anos, gestantes e/ou nutrizes – limitado hoje a cinco dessas parcelas por família - uma parcela representa hoje cerca de 0,6% do salário mínimo;

3. **Benefício Variável Vinculado ao Adolescente (BVJ):** R$ 46,00 concedidos pela existência na família de jovens entre 16 e 17 anos – limitado a até dois parcelas por família ainda que haja mais de dois jovens na mesma, representa hoje algo em torno de 0,25% do salário mínimo, tendo sido implementado em 2007;

4. **Benefício Variável de Caráter Extraordinário (BVCE):** valor calculado caso a caso, onde a premissa é que famílias que recebiam dos Programas Bolsa Escola, Bolsa Alimentação, PNAA e Auxílio-Gás, que, na data de ingresso no Programa Bolsa Família, exceda o limite máximo de R$ 45,00 - esse benefício será mantido até a cessação das condições de elegibilidade de cada um dos benefícios que lhe deram origem;

5. **Benefício para a Superação da Extrema Pobreza na Primeira Infância (BSP):** criado pela Medida Provisória nº 570, em 14 de maio de 2012, trata-se de uma complementação de renda destinada às famílias já beneficiadas pelo PBF que possuem, em sua composição familiar, crianças de 0 a 6 anos de idade e mesmo recebendo os demais benefícios, permanecem em situação de extrema pobreza, ou seja, renda familiar mensal inferior a R$ 85,00 por pessoa.

O recebimento do benefício somente ocorre se observadas algumas condições (e, por isso, chamadas condicionalidades) que as famílias beneficiárias se comprometem a cumprir. Estão relacionadas à saúde, educação e assistência social. A condicionalidade de saúde se refere a famílias com crianças de até 7 anos, sendo solicitado o preenchimento do cartão de vacina, acompanhamento do crescimento e desenvolvimento (curva nutricional), e as mulheres na faixa de 14 a 44 anos, gestantes ou que estiverem amamentando (nutrizes) devem fazer o pré-natal e acompanhamento pós-parto e observar tanto a sua saúde como a do seu bebê. A condicionalidade de educação visa incentivar a frequência mínima de crianças e adolescentes (de 6 a 17 anos) sendo 85% para crianças e adolescentes de até 15 anos e 75% para os jovens de 16 e 17 anos. A condicionalidade de assistência social busca impedir que crianças e adolescentes até 15 anos entrem ou permaneçam no trabalho infantil. Para tal, aqui é exigida dessas crianças e adolescentes frequência mensal mínima de 85% da carga horária nos Serviços de Convivência e Fortalecimento de Vínculos (SCFV) do Programa de

Erradicação do Trabalho Infantil (PETI). O quadro A.1 traz de forma concisa as condicionantes a serem atendidas, a que área de concentração pertencem e a legislação que a regulamenta.

Caso a família não atenda as condicionalidades, o benefício pode ser bloqueado, suspenso, ou até cancelado. A família que descumprir alguma condicionalidade pela primeira vez, receberá só uma advertência. Na segunda vez, terá seu beneficio bloqueado por 30 dias. Na terceira vez, acarretará em uma suspensão do benefício por 60 dias. Na quarta vez, o benefício é suspenso por mais 60 dias. E quando chega ao quinto descumprimento da condicionalidade o beneficio é cancelado. As etapas estão ilustradas no próprio site do MDS pelo esquema da figura A.1.

Tabela A.1: Valores percebidos no PBF

| **descriao** | **valor** |
|---|---|
| Valor básico do benefício **(somente se em situação de extrema pobreza)** | R$ 85,00 |
| variante - gestantes, nutrizes e crianças **(menor de 16 anos - VAR)*** | Até 5 de R$ 39,00 |
| variante - jovem 16 a 17 anos (**BVJ**)* | Até 2 de R$ 46,00 |
| Máximo a ser recebido por uma família | (5*39)+(2*46)=R$ 287,00 ***(pobre)*** <br><br> 85+[(5 ∗ 39) + (2 ∗ 46)]=R$ 372,00 ***(extremamente pobre)*** |

**Quantidade de variantes definida pela Lei nº 12.512 de 14/10/2011 (5 VAR + 2 BVJ) valores ajustados em 2016 juntamente com o valor básico do benefício conforme Decreto-Lei nº 8.794 de 29/06/2016.*

*Fonte: referência [25]*

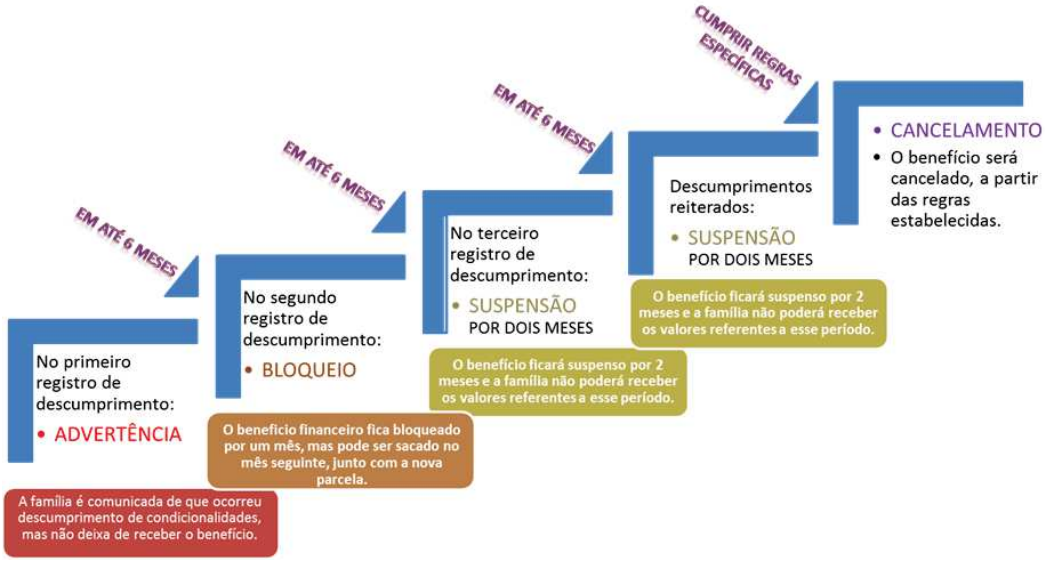


Figura A.1: consequências do não cumprimento das condicionantes

*Fonte: http://mds.gov.br*

O acompanhamento quanto ao não-cumprimento das condicionalidades é feito pelas três esferas do governo via agentes de saúde, pela escola onde a criança vinculada ao programa e via atendimento de agentes sociais alimentando-se um sistema chamado Sistema de Gestão do Programa Bolsa Família (SIGPBF). A partir daí, são implementadas ações de acompanhamento das famílias em descumprimento, consideradas em situação de maior vulnerabilidade social.

Conforme a referência [30], nos municípios e nos estados, a participação e o controle social do Bolsa Família são exercidos pelos Conselhos de Assistência Social (CMAS ou CEAS). Porém ainda existem alguns municípios em que essa função está a cargo de Instâncias de Controle Social (ICS) exclusivas. Tanto os Conselhos quanto as ICS devem ter composição paritária, ou seja, devem ter o mesmo número de representantes do governo e da sociedade civil. Os Conselhos podem colaborar para o bom funcionamento do programa, contribuindo, por exemplo, para o acompanhamento das condicionalidades e acompanhando a gestão local, para que o público-alvo seja efetivamente atendido. Também podem apoiar a integração entre o Bolsa Família e outras políticas que promovam oportunidades para as famílias. É importante que os Conselhos estimulem a participação, em suas reuniões, de beneficiários do Bolsa Família. Para garantir a transparência na implementação do programa e assegurar que os benefícios cheguem às famílias que preencham os requisitos definidos em lei para acesso ao Bolsa Família, o controle social é articulado com instrumentos de fiscalização. O Ministério do Desenvolvimento Social e Combate à Fome executa a fiscalização do programa por meio de sua equipe técnica e submete a avaliação de

suas ações à auditoria dos órgãos de controle, como a Controladoria-Geral da União (CGU), o Tribunal de Contas da União (TCU) e os ministérios públicos federal e estaduais.

A família com dificuldades para cumprir as condicionantes deve procurar o Centro de Referência de Assistência Social (Cras), o Centro de Referência Especializada de Assistência Social (Creas) ou a equipe de assistência social do município para que não corra o risco de ter o benefício bloqueado, suspenso ou até mesmo cancelado.

Quadro A.1: Resumo das condicionantes do PBF

| **Área de Concentração** | **Condicionalidades/ Público Alvo** | **Previsão Legal** |
|---|---|---|
| Educação | Crianças de 06 a 15 anos de idade devem ter frequência escolar mensal mínima de 85% da carga horária. Já os adolescentes de 16 e 17 anos devem ter frequência mínima de 75%. | Lei nº 10.836 de 09 de janeiro de 2004. Decreto nº 5.209, de 17 de setembro de 2004, Portaria interministerial MEC/MDS nº 3.789, de 17 de novembro de 2004 e Portaria nº 251, de 12 de dezembro de 2012. |
| Saúde | As gestantes e nutrizes devem comparecer às consultas de pré-natal a assistência ao puerpério. Já as crianças menores de 07 anos de idade deverão cumprir o calendário de vacinação e realizar o acompanhamento do seu crescimento e desenvolvimento. | Lei nº 10.836 de 09 de Janeiro de 2004; Decreto nº 5.209, de 17 de setembro de 2004, e Portaria nº 251, de 12 de dezembro de 2012. |
| Assistência Social | As crianças e adolescentes de até 15 anos de idade, em risco ou retiradas do trabalho infantil, exige-se a frequência mínima de 85% da carga horária relativa aos Serviços de Convivência e Fortalecimento de Vínculos – SCFV. | Portaria MDS nº 666, de 28 de dezembro de 2005; Portaria nº 251, de 12 de dezembro de 2012. |

*Fonte: referência [29]*

## Valores das componentes do benefício do Programa Bolsa Família até o ano de 2016

| Número de gestantes, nutrizes, crianças e adolescentes de até 15 anos | Número de jovens de 16 e 17 anos | Tipo de benefício | 2016 | 2015 | 2014 | 2013 | 2012 | 2011 | 2010 | 2009 | 2008 | 2007 | 2006 | 2005 | 2004 | 2003 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 0 | 0 | Básico | R$ 85,00 | R$ 77,00 | R$ 70,00 | R$ 70,00 | R$ 68,00 | R$ 68,00 | R$ 68,00 | R$ 68,00 | R$ 62,00 | R$ 58,00 | R$ 50,00 | R$ 50,00 | R$ 50,00 | R$ 50,00 |
| 1 | 0 | Básico + 1 variável | R$ 124,00 | R$ 112,00 | R$ 102,00 | R$ 102,00 | R$ 90,00 | R$ 90,00 | R$ 90,00 | R$ 90,00 | R$ 82,00 | R$ 76,00 | R$ 65,00 | R$ 65,00 | R$ 65,00 | R$ 65,00 |
| 2 | 0 | Básico + 2 variáveis | R$ 163,00 | R$ 147,00 | R$ 134,00 | R$ 134,00 | R$ 112,00 | R$ 112,00 | R$ 112,00 | R$ 112,00 | R$ 102,00 | R$ 94,00 | R$ 80,00 | R$ 80,00 | R$ 80,00 | R$ 80,00 |
| 3 | 0 | Básico + 3 variáveis | R$ 202,00 | R$ 182,00 | R$ 166,00 | R$ 166,00 | R$ 134,00 | R$ 134,00 | R$ 134,00 | R$ 134,00 | R$ 122,00 | R$ 112,00 | R$ 95,00 | R$ 95,00 | R$ 95,00 | R$ 95,00 |
| 4 | 0 | Básico + 4 variáveis | R$ 241,00 | R$ 217,00 | R$ 198,00 | R$ 198,00 | R$ 156,00 | R$ 156,00 | R$ 134,00 | R$ 134,00 | R$ 122,00 | R$ 112,00 | R$ 95,00 | R$ 95,00 | R$ 95,00 | R$ 95,00 |
| 5 | 0 | Básico + 5 variáveis | R$ 280,00 | R$ 252,00 | R$ 230,00 | R$ 230,00 | R$ 178,00 | R$ 178,00 | R$ 134,00 | R$ 134,00 | R$ 122,00 | R$ 112,00 | R$ 95,00 | R$ 95,00 | R$ 95,00 | R$ 95,00 |
| 0 | 1 | Básico + 1 BVJ | R$ 131,00 | R$ 119,00 | R$ 108,00 | R$ 108,00 | R$ 101,00 | R$ 101,00 | R$ 101,00 | R$ 101,00 | R$ 92,00 | R$ 58,00 | R$ 50,00 | R$ 50,00 | R$ 50,00 | R$ 50,00 |
| 1 | 1 | Básico + 1 variável + 1 BVJ | R$ 170,00 | R$ 154,00 | R$ 140,00 | R$ 140,00 | R$ 123,00 | R$ 123,00 | R$ 123,00 | R$ 123,00 | R$ 112,00 | R$ 76,00 | R$ 65,00 | R$ 65,00 | R$ 65,00 | R$ 65,00 |
| 2 | 1 | Básico + 2 variáveis + 1 BVJ | R$ 209,00 | R$ 189,00 | R$ 172,00 | R$ 172,00 | R$ 145,00 | R$ 145,00 | R$ 145,00 | R$ 145,00 | R$ 132,00 | R$ 94,00 | R$ 80,00 | R$ 80,00 | R$ 80,00 | R$ 80,00 |
| 3 | 1 | Básico + 3 variáveis + 1 BVJ | R$ 248,00 | R$ 224,00 | R$ 204,00 | R$ 204,00 | R$ 167,00 | R$ 167,00 | R$ 167,00 | R$ 167,00 | R$ 152,00 | R$ 142,00 | R$ 95,00 | R$ 95,00 | R$ 95,00 | R$ 95,00 |
| 4 | 1 | Básico + 4 variáveis + 1 BVJ | R$ 287,00 | R$ 259,00 | R$ 236,00 | R$ 236,00 | R$ 189,00 | R$ 189,00 | R$ 167,00 | R$ 167,00 | R$ 152,00 | R$ 142,00 | R$ 95,00 | R$ 95,00 | R$ 95,00 | R$ 95,00 |
| 5 | 1 | Básico + 5 variáveis + 1 BVJ | R$ 326,00 | R$ 294,00 | R$ 268,00 | R$ 268,00 | R$ 211,00 | R$ 211,00 | R$ 167,00 | R$ 167,00 | R$ 152,00 | R$ 142,00 | R$ 95,00 | R$ 95,00 | R$ 95,00 | R$ 95,00 |
| 0 | 2 | Básico + 2 BVJ | R$ 177,00 | R$ 161,00 | R$ 146,00 | R$ 146,00 | R$ 134,00 | R$ 134,00 | R$ 134,00 | R$ 134,00 | R$ 122,00 | R$ 58,00 | R$ 50,00 | R$ 50,00 | R$ 50,00 | R$ 50,00 |
| 1 | 2 | Básico + 1 variável + 2 BVJ | R$ 216,00 | R$ 196,00 | R$ 178,00 | R$ 178,00 | R$ 156,00 | R$ 156,00 | R$ 156,00 | R$ 156,00 | R$ 142,00 | R$ 76,00 | R$ 65,00 | R$ 65,00 | R$ 65,00 | R$ 65,00 |
| 2 | 2 | Básico + 2 variáveis + 2 BVJ | R$ 255,00 | R$ 231,00 | R$ 210,00 | R$ 210,00 | R$ 178,00 | R$ 178,00 | R$ 178,00 | R$ 178,00 | R$ 162,00 | R$ 94,00 | R$ 80,00 | R$ 80,00 | R$ 80,00 | R$ 80,00 |
| 3 | 2 | Básico + 3 variáveis + 2 BVJ | R$ 294,00 | R$ 266,00 | R$ 242,00 | R$ 242,00 | R$ 200,00 | R$ 200,00 | R$ 200,00 | R$ 200,00 | R$ 182,00 | R$ 172,00 | R$ 95,00 | R$ 95,00 | R$ 95,00 | R$ 95,00 |
| 4 | 2 | Básico + 4 variáveis + 2 BVJ | R$ 333,00 | R$ 301,00 | R$ 274,00 | R$ 274,00 | R$ 222,00 | R$ 222,00 | R$ 200,00 | R$ 200,00 | R$ 182,00 | R$ 172,00 | R$ 95,00 | R$ 95,00 | R$ 95,00 | R$ 95,00 |
| 5 | 2 | Básico + 5 variáveis + 2 BVJ | R$ 372,00 | R$ 336,00 | R$ 306,00 | R$ 306,00 | R$ 244,00 | R$ 244,00 | R$ 200,00 | R$ 200,00 | R$ 182,00 | R$ 172,00 | R$ 95,00 | R$ 95,00 | R$ 95,00 | R$ 95,00 |
| | | LIMITADOR - GESTANTES, NUTRIZES E CRIANÇAS (MENOR DE 16 ANOS) | 5 | 5 | 5 | 5 | 5 | 5 | 3 | 3 | 3 | 3 | 3 | 3 | 3 | 3 |
| | | LIMITADOR - JOVEM 16 A 17 ANOS (BVJ) | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 0 | 0 | 0 | 0 | 0 |
| | | LIMITADOR - BSP | 1 | 1 | 1 | 1 | 1 | | | | | | | | | |
| gestantes, nutrizes, idade < 15 anos | | VARIAVEL | R$ 39,00 | R$ 35,00 | R$ 32,00 | R$ 32,00 | R$ 22,00 | R$ 22,00 | R$ 22,00 | R$ 22,00 | R$ 20,00 | R$ 18,00 | R$ 15,00 | R$ 15,00 | R$ 15,00 | R$ 15,00 |
| 16 < idade < 17 | | BVJ | R$ 46,00 | R$ 42,00 | R$ 38,00 | R$ 38,00 | R$ 33,00 | R$ 33,00 | R$ 33,00 | R$ 33,00 | R$ 30,00 | R$ 30,00 | R$ - | R$ - | R$ - | R$ - |
| <6 anos, PER CAPITA menor que limite p/ extrema pobreza | | BSP | R$ 70,00 | R$ 70,00 | R$ 70,00 | R$ 70,00 | R$ 70,00 | | | | | | | | | |
| **VALOR PERCAPITA** QUE DEFINE [**POBREZA**] OU [**EXTREMA POBREZA**] | | POBREZA | R$ 170,00 | R$ 154,00 | R$ 140,00 | R$ 140,00 | R$ 140,00 | R$ 140,00 | R$ 140,00 | R$ 140,00 | R$ 137,00 | R$ 120,00 | R$ 100,00 | R$ 100,00 | R$ 100,00 | R$ 100,00 |
| | | | 19% | 20% | 19% | 21% | 23% | 26% | 27% | 30% | 33% | 32% | 29% | 33% | 38% | 42% |
| | | EXTREMA POBREZA | R$ 85,00 | R$ 77,00 | R$ 70,00 | R$ 70,00 | R$ 70,00 | R$ 70,00 | R$ 70,00 | R$ 70,00 | R$ 69,00 | R$ 60,00 | R$ 50,00 | R$ 50,00 | R$ 50,00 | R$ 50,00 |
| | | | 10% | 10% | 10% | 10% | 11% | 13% | 14% | 15% | 17% | 16% | 14% | 17% | 19% | 21% |
| *BRASIL SEM MISERIA | | MINIMO PER CAPITA (BENEFICIOS + RENDIMENTOS) | R$ 85,00 | R$ 77,00 | R$ 70,00 | R$ 70,00 | R$ 70,00 | | | | | | | | | |

| | | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | De acordo com o **Decreto nº 7.494 de 02/06/2011**, a família terá direito a receber **até 05 benefícios variáveis**, conforme o seu perfil. | DECRETO Nº 8.794, DE 29 DE JUNHO DE 2016 | DECRETO Nº 8.232, DE 30 DE ABRIL DE 2014 | | LEI Nº 12.817, DE 5 DE JUNHO DE 2013. | LEI Nº 12.722, DE 3 DE OUTUBRO DE 2012. Medida Provisória nº 590, de 29 de Novembro de 2012 | Lei nº 12.512, de 14/10/2011 | | Decreto no. 6.917 de 30 de julho de 2009 | DECRETO Nº 6.491, DE 26 DE JUNHO DE 2008 | MEDIDA PROVISÓRIA Nº 411, DE 28 DE DEZEMBRO DE 2007. | | LEI No 10.836, DE 9 DE JANEIRO DE 2004. | MEDIDA PROVISÓRIA Nº 132, DE 20 DE OUTUBRO 2003. |
| | | SALÁRIO MINIMO | R$ 880,00 | R$ 788,00 | R$ 724,00 | R$ 678,00 | R$ 622,00 | R$ 545,00 | R$ 510,00 | R$ 465,00 | R$ 415,00 | R$ 380,00 | R$ 350,00 | R$ 300,00 | R$ 260,00 | R$ 240,00 |
| | | BASICO (% DO MINIMO) | 9,66% | 9,77% | 9,67% | 10,32% | 10,93% | 12,48% | 13,33% | 14,62% | 14,94% | 15,26% | 14,29% | 16,67% | 19,23% | 20,83% |
| | | MAXIMO (% DO MINIMO) | 42,27% | 42,64% | 42,27% | 45,13% | 39,23% | 44,77% | 39,22% | 43,01% | 43,86% | 29,47% | 27,14% | 31,67% | 36,54% | 39,58% |

**percentual sempre referente ao salário mínimo vigente**

**BENEFICIO VARIÁVEL DE CARATER EXTRAORDINARIO** parcela do valor dos benefícios em manutenção das famílias beneficiárias dos Programas Bolsa Escola, Bolsa Alimentação, PNAA e Auxílio-Gás que, na data de ingresso dessas famílias no Programa Bolsa Família, exceda o limite máximo fixado para o Programa Bolsa Família.

O benefício variável de caráter extraordinário de que trata o inciso IV terá seu montante arrendondado para o valor inteiro imediatamente superior, sempre que necessário." (NR)

Esse benefício será mantido até a cessação das condições de elegibilidade de cada um dos benefícios que lhe deram origem.

# APÊNDICE B - Divisão do questionário utilizado na aquisiçao dos dados

O questionário da base de dados utilizada neste estudo foi elaborado pelo Consórcio entre o Instituto Internacional de Pesquisa sobre Políticas Alimentares (IFPRI) e Datamétrica Consultoria, Pesquisa e Telemarketing Ltda., instituições responsáveis pelo levantamento de 2009 encomendado pelo MDS. Este questionário está divido nas seções abaixo listadas e encontra-se disponível no site do MDS através do link [http://aplicacoes.mds.gov.br/sagi/PainelPEI/Publicacoes/relatoriosAIBF2.rar]

Quadro B.1: Seções do questionário aplicado para levantamento dos dados

| **SEÇÃO** | | **DESCRIÇÃO** |
|---|---|---|
| | | identificação do entrevistador e do domicílio |
| | | informações para contato posterior |
| **01** | | CARACTERÍSTICAS DO DOMICÍLIO |
| **02** | | CARACTERISTICAS DOS MORADORES, MIGRAÇÃO E ANTROPOMETRIA |
| | a | caracteristicas dos moradores |
| | b | migração (já morava, mudou-se |
| | c | medidas antropométricas para todos os moradores |
| **03** | | EDUCAÇÃO |
| | a | dados gerais (alfabetizado?) |
| | b | gastos com educação |
| **04** | | SAÚDE |
| | a | dados gerais |
| | b | agente de saude |
| | c | gastos com saúde |
| | d | mulheres entre 10 e 49 anos |
| | e | saúde da criança (até 6 anos) - práticas |
| | f | saúde da criança (até 6 anos) |
| **05** | | TRABALHO E TRABALHO INFANTIL |
| | a | informações gerais |
| | b | sobre moradores maiores de 5 anos (últimos 12 meses) |
| | c | moradores de 10 a 29 anos |

Quadro B.1: Seções do questionário aplicado para levantamento dos dados (continuação)

| **SEÇÃO** | | **DESCRIÇÃO** |
|---|---|---|
| **06** | | RENDIMENTOS |
| | a | moradores maiores de 10 anos |
| **07** | | GASTOS INDIVIDUAIS |
| | a | gastos com transporte e comunicação |
| | b | gastos com alimentação fora de casa |
| **08** | | GASTOS COLETIVOS DO DOMICÍLIO |
| | a | habitação/reparos/mobiliario/utensilios/artigos do lar/vestuario/ serviços domésticos/recreação e cultura/higiene pessoal e da casa/outros |
| **09** | | ALIMENTOS E BEBIDAS (65 ITENS) |
| **10** | | BENS DURÁVEIS |
| | a | itenspresentes no domicílio |
| | b | animais e implementos agrícolas |
| | c | imóveis |
| **11** | | CONDIÇÕES DE VIDA |
| | a | convivio social/opinião |
| | b | mulherco- responsável pela casa |
| | c | alocação de tempo do adulto |
| **12** | | ACESSO A CRÉDITO, INCLUSÃO BANCÁRIA E EDUCAÇÃO FINANCEIRA |
| **13** | | PERCEPÇÕES SOBRE POBREZA, BEM-ESTAR E CONFIANÇA |
| **14** | | CHOQUES E MECANISMOS DE LONGO PRAZO (DESASTRES E CALAMIDADES) |
| **15** | | BENEFÍCIOS |
| | a | bolsa família - cadastro/outros benefícios |
| | b | percepções do entrevistado |
| | c | titular do cartão BF - apanhado 12 meses recebidos de Bolsa Família |

*Fonte: elaboração própria baseado na referência [31]*

# APÊNDICE C - Métricas correspondentes a melhor especificidade de cada configuração RNA

De acordo com a melhor especificidade alcançada em cada método de aprendizagem RNA dentre as 50 rodadas, os quadros abaixo apresentam as outras métricas correspondentes.

Tabela C.1: sem intervenção na amostra

| | **trainlm** | **trainbr** | **trainbfg** | **trainrp** | **trainscg** | **traincgf** | **traincgb** | **traincgp** | **trainoss** | **traingdx** | **traingdm** |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Sensibilidade | 0,90646 | 0,93039 | 0,93039 | 0,90797 | 0,91713 | 0,93232 | 0,92127 | 0,94855 | 0,92038 | 0,93039 | 0,94913 |
| **Especificidade** | **0,52000** | **0,48069** | **0,48069** | **0,44444** | **0,44269** | **0,44269** | **0,47111** | **0,44492** | **0,42000** | **0,41333** | **0,08197** |
| Suporte | 0,67451 | 0,71136 | 0,71136 | 0,67656 | 0,67963 | 0,69089 | 0,68270 | 0,73593 | 0,69806 | 0,71136 | 0,72569 |
| Cobertura | 0,82293 | 0,85466 | 0,85466 | 0,83930 | 0,82395 | 0,83521 | 0,82600 | 0,87410 | 0,84545 | 0,85977 | 0,94268 |
| Conf. Pos. (VPP) | 0,81965 | 0,83234 | 0,83234 | 0,80610 | 0,82484 | 0,82721 | 0,82652 | 0,84192 | 0,82567 | 0,82738 | 0,76982 |
| Conf. Neg.(VPN) | 0,60694 | 0,63380 | 0,63380 | 0,57325 | 0,65116 | 0,69565 | 0,66471 | 0,68293 | 0,60927 | 0,62044 | 0,32143 |
| Acurácia | 0,78199 | 0,80348 | 0,80348 | 0,76868 | 0,79427 | 0,80553 | 0,79836 | 0,82190 | 0,79222 | 0,79836 | 0,74411 |
| F-measure | 0,81965 | 0,83234 | 0,83234 | 0,80610 | 0,82484 | 0,82721 | 0,82652 | 0,84192 | 0,82567 | 0,82738 | 0,76982 |
| Área sob a curva | 0,75067 | 0,78924 | 0,78924 | 0,73863 | 0,77349 | 0,80249 | 0,78646 | 0,76988 | 0,77995 | 0,78415 | 0,50632 |
| Eficiência | 0,66323 | 0,66085 | 0,66085 | 0,63471 | 0,67991 | 0,68750 | 0,68396 | 0,66606 | 0,65510 | 0,64998 | 0,51370 |
| G-mean | 0,59338 | 0,59782 | 0,59782 | 0,62185 | 0,59500 | 0,61226 | 0,57802 | 0,62520 | 0,61126 | 0,57955 | 0,19678 |
| MCC | 0,37318 | 0,38724 | 0,38724 | 0,31969 | 0,41385 | 0,44280 | 0,42512 | 0,41750 | 0,36732 | 0,36650 | 0,04999 |
| Índice de Youden | 0,27466 | 0,32416 | 0,32416 | 0,34289 | 0,30828 | 0,33381 | 0,29194 | 0,35266 | 0,33939 | 0,29759 | -0,01052 |
| Kappa | 0,29520 | 0,28779 | 0,28779 | 0,23919 | 0,32361 | 0,33273 | 0,33007 | 0,29478 | 0,27829 | 0,26674 | 0,02222 |
| Taxa de erro positiva | 0,48000 | 0,51931 | 0,51931 | 0,55556 | 0,55731 | 0,55731 | 0,52889 | 0,55508 | 0,58000 | 0,58667 | 0,91803 |
| Taxa de erro negativa | 0,09354 | 0,06961 | 0,06961 | 0,09203 | 0,08287 | 0,06768 | 0,07873 | 0,05145 | 0,07962 | 0,06961 | 0,05087 |
| Taxa de erro global | 0,21801 | 0,19652 | 0,19652 | 0,23132 | 0,20573 | 0,19447 | 0,20164 | 0,17810 | 0,20778 | 0,20164 | 0,25589 |

*Fonte: elaboração própria*

Tabela C.2: Sub-amostragem em beneficiários

| | **trainlm** | **trainbr** | **trainbfg** | **trainrp** | **trainscg** | **traincgf** | **traincgb** | **traincgp** | **trainoss** | **traingdx** | **traingdm** |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Sensibilidade | 0,51626 | 0,73571 | 0,73571 | 0,70714 | 0,74643 | 0,72857 | 0,73050 | 0,71986 | 0,70922 | 0,68929 | 0,32707 |
| **Especificidade** | **0,75403** | **0,75806** | **0,75806** | **0,73878** | **0,72727** | **0,73279** | **0,72152** | **0,72656** | **0,73387** | **0,72374** | **0,75732** |
| Suporte | 0,25149 | 0,40792 | 0,40792 | 0,39208 | 0,41386 | 0,40396 | 0,40792 | 0,40198 | 0,39604 | 0,38218 | 0,17228 |
| Cobertura | 0,38218 | 0,57030 | 0,57030 | 0,56238 | 0,57426 | 0,55644 | 0,56238 | 0,57228 | 0,54257 | 0,53663 | 0,28713 |
| Conf. Pos. (VPP) | 0,65803 | 0,71528 | 0,71528 | 0,69718 | 0,72069 | 0,72598 | 0,72535 | 0,70242 | 0,72993 | 0,71218 | 0,60000 |
| Conf. Neg.(VPN) | 0,61859 | 0,65899 | 0,65899 | 0,62896 | 0,66977 | 0,66071 | 0,65611 | 0,63426 | 0,64502 | 0,62821 | 0,50278 |
| Acurácia | 0,63366 | 0,69109 | 0,69109 | 0,66733 | 0,69901 | 0,69703 | 0,69505 | 0,67327 | 0,69109 | 0,67327 | 0,53069 |
| F-measure | 0,65803 | 0,71528 | 0,71528 | 0,69718 | 0,72069 | 0,72598 | 0,72535 | 0,70242 | 0,72993 | 0,71218 | 0,60000 |
| Área sob a curva | 0,72395 | 0,75279 | 0,75279 | 0,72721 | 0,75344 | 0,74456 | 0,73735 | 0,73423 | 0,73681 | 0,73824 | 0,56375 |
| Eficiência | 0,63072 | 0,68563 | 0,68563 | 0,66246 | 0,69321 | 0,69317 | 0,69036 | 0,66710 | 0,68869 | 0,67131 | 0,54219 |
| G-mean | 0,66932 | 0,73674 | 0,73674 | 0,69219 | 0,73237 | 0,70842 | 0,74340 | 0,72676 | 0,75593 | 0,69994 | 0,43924 |
| MCC | 0,26892 | 0,37276 | 0,37276 | 0,32553 | 0,38844 | 0,38652 | 0,38109 | 0,33544 | 0,37616 | 0,34150 | 0,09313 |
| Índice de Youden | 0,35963 | 0,48304 | 0,48304 | 0,39175 | 0,47083 | 0,41764 | 0,49424 | 0,46448 | 0,51429 | 0,40049 | -0,03168 |
| Kappa | 0,33323 | 0,36096 | 0,36096 | 0,32034 | 0,37310 | 0,38497 | 0,37804 | 0,32611 | 0,38840 | 0,35400 | 0,15481 |
| Taxa de erro positiva | 0,24597 | 0,24194 | 0,24194 | 0,26122 | 0,27273 | 0,26721 | 0,27848 | 0,27344 | 0,26613 | 0,27626 | 0,24268 |
| Taxa de erro negativa | 0,48374 | 0,26429 | 0,26429 | 0,29286 | 0,25357 | 0,27143 | 0,2695 | 0,28014 | 0,29078 | 0,31071 | 0,67293 |
| Taxa de erro global | 0,36634 | 0,30891 | 0,30891 | 0,33267 | 0,30099 | 0,30297 | 0,30495 | 0,32673 | 0,30891 | 0,32673 | 0,46931 |

*Fonte: elaboração própria*

Tabela C.3: Super-amostragem em não-beneficiários

| | **Trainlm** | **trainbr** | **trainbfg** | **trainrp** | **trainscg** | **traincgf** | **traincgb** | **traincgp** | **trainoss** | **traingdx** | **traingdm** |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Sensibilidade | 0,82759 | 0,84803 | 0,84803 | 0,83355 | 0,83875 | 0,83221 | 0,85927 | 0,83576 | 0,86207 | 0,81946 | 0,77989 |
| **Especificidade** | **0,73461** | **0,69835** | **0,69835** | **0,69182** | **0,66878** | **0,65033** | **0,65924** | **0,64718** | **0,65289** | **0,64450** | **0,35759** |
| Suporte | 0,49302 | 0,51356 | 0,51356 | 0,53492 | 0,50863 | 0,50534 | 0,52177 | 0,51849 | 0,51356 | 0,52588 | 0,47165 |
| Cobertura | 0,64503 | 0,64174 | 0,64174 | 0,64914 | 0,65489 | 0,65078 | 0,67132 | 0,65900 | 0,66475 | 0,65325 | 0,72555 |
| Conf. Pos. (VPP) | 0,76433 | 0,80026 | 0,80026 | 0,82405 | 0,77666 | 0,77652 | 0,77723 | 0,78678 | 0,77256 | 0,80503 | 0,65006 |
| Conf. Neg.(VPN) | 0,71065 | 0,74312 | 0,74312 | 0,69555 | 0,71667 | 0,70824 | 0,74000 | 0,70120 | 0,75490 | 0,66588 | 0,51497 |
| Acurácia | 0,74528 | 0,77979 | 0,77979 | 0,77896 | 0,75596 | 0,75267 | 0,76500 | 0,75760 | 0,76664 | 0,75678 | 0,61298 |
| F-measure | 0,76433 | 0,80026 | 0,80026 | 0,82405 | 0,77666 | 0,77652 | 0,77723 | 0,78678 | 0,77256 | 0,80503 | 0,65006 |
| Área sob a curva | 0,81324 | 0,82294 | 0,82294 | 0,82745 | 0,81158 | 0,79769 | 0,80525 | 0,80504 | 0,80627 | 0,78700 | 0,59741 |
| Eficiência | 0,72578 | 0,76152 | 0,76152 | 0,75737 | 0,73357 | 0,73096 | 0,73926 | 0,73282 | 0,74404 | 0,73198 | 0,56874 |
| G-mean | 0,74568 | 0,80305 | 0,80305 | 0,73732 | 0,73671 | 0,71961 | 0,71760 | 0,71165 | 0,72826 | 0,70300 | 0,50149 |
| MCC | 0,46313 | 0,53311 | 0,53311 | 0,51716 | 0,48006 | 0,47319 | 0,49750 | 0,47668 | 0,50739 | 0,46742 | 0,15062 |
| Índice de Youden | 0,49879 | 0,61122 | 0,61122 | 0,49207 | 0,49115 | 0,45541 | 0,45628 | 0,45140 | 0,47456 | 0,42409 | 0,08507 |
| Kappa | 0,41706 | 0,49357 | 0,49357 | 0,50888 | 0,43256 | 0,43100 | 0,43283 | 0,43834 | 0,43741 | 0,45579 | 0,11459 |
| Taxa de erro positiva | 0,26539 | 0,30165 | 0,30165 | 0,30818 | 0,33122 | 0,34967 | 0,34076 | 0,35282 | 0,34711 | 0,3555 | 0,64241 |
| Taxa de erro negativa | 0,17241 | 0,15197 | 0,15197 | 0,16645 | 0,16125 | 0,16779 | 0,14073 | 0,16424 | 0,13793 | 0,18054 | 0,22011 |
| Taxa de erro global | 0,25472 | 0,22021 | 0,22021 | 0,22104 | 0,24404 | 0,24733 | 0,235 | 0,2424 | 0,23336 | 0,24322 | 0,38702 |

*Fonte: elaboração própria*

Tabela C.4: intervenção em ambos os grupos da amostra

| | **trainlm** | **trainbr** | **trainbfg** | **trainrp** | **trainscg** | **traincgf** | **traincgb** | **traincgp** | **trainoss** | **traingdx** | **traingdm** |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Sensibilidade | 0,71310 | 0,69388 | 0,69388 | 0,73805 | 0,70231 | 0,70576 | 0,73805 | 0,73770 | 0,71800 | 0,65199 | 0,49251 |
| **Especificidade** | **0,85591** | **0,79877** | **0,79877** | **0,76891** | **0,76082** | **0,75269** | **0,75926** | **0,75983** | **0,75316** | **0,76000** | **0,66102** |
| Suporte | 0,35655 | 0,35343 | 0,35343 | 0,36902 | 0,34823 | 0,34407 | 0,36902 | 0,37422 | 0,37318 | 0,32328 | 0,23909 |
| Cobertura | 0,48649 | 0,46362 | 0,46362 | 0,49480 | 0,46881 | 0,47817 | 0,52391 | 0,52287 | 0,50624 | 0,46258 | 0,41476 |
| Conf. Pos. (VPP) | 0,73291 | 0,76233 | 0,76233 | 0,74580 | 0,74279 | 0,71957 | 0,70437 | 0,71571 | 0,73717 | 0,69888 | 0,57644 |
| Conf. Neg.(VPN) | 0,72065 | 0,70930 | 0,70930 | 0,74074 | 0,72211 | 0,72510 | 0,72489 | 0,72113 | 0,70316 | 0,67892 | 0,57904 |
| Acurácia | 0,72661 | 0,73389 | 0,73389 | 0,74324 | 0,73181 | 0,72245 | 0,71414 | 0,71830 | 0,72037 | 0,68815 | 0,57796 |
| F-measure | 0,73291 | 0,76233 | 0,76233 | 0,74580 | 0,74279 | 0,71957 | 0,70437 | 0,71571 | 0,73717 | 0,69888 | 0,57644 |
| Área sob a curva | 0,78909 | 0,81350 | 0,81350 | 0,81606 | 0,79552 | 0,79015 | 0,78996 | 0,77662 | 0,79663 | 0,76720 | 0,61578 |
| Eficiência | 0,72661 | 0,73465 | 0,73465 | 0,74324 | 0,73157 | 0,72205 | 0,71414 | 0,71801 | 0,72047 | 0,68785 | 0,57555 |
| G-mean | 0,75261 | 0,78418 | 0,78418 | 0,73701 | 0,73884 | 0,74890 | 0,75131 | 0,75547 | 0,75603 | 0,71915 | 0,55651 |
| MCC | 0,45339 | 0,47047 | 0,47047 | 0,48651 | 0,46402 | 0,44438 | 0,42876 | 0,43643 | 0,44063 | 0,37675 | 0,15327 |
| Índice de Youden | 0,50614 | 0,56840 | 0,56840 | 0,47428 | 0,47924 | 0,49790 | 0,50450 | 0,51543 | 0,51222 | 0,43831 | 0,12291 |
| Kappa | 0,46581 | 0,51560 | 0,51560 | 0,49160 | 0,48983 | 0,45278 | 0,40873 | 0,42301 | 0,45271 | 0,40272 | 0,17684 |
| Taxa de erro positiva | 0,14409 | 0,20123 | 0,20123 | 0,23109 | 0,23918 | 0,24731 | 0,24074 | 0,24017 | 0,24684 | 0,24000 | 0,33898 |
| Taxa de erro negativa | 0,28690 | 0,30612 | 0,30612 | 0,26195 | 0,29769 | 0,29424 | 0,26195 | 0,26230 | 0,28200 | 0,34801 | 0,50749 |
| Taxa de erro global | 0,27339 | 0,26611 | 0,26611 | 0,25676 | 0,26819 | 0,27755 | 0,28586 | 0,28170 | 0,27963 | 0,31185 | 0,42204 |

*Fonte: elaboração própria*

# APÊNDICE D - Variáveis utilizadas no estudo

As váriavéis utilizadas neste estudo encontram-se listadas nos quadros D.1 e D.2 e variáveis derivadas de algumas delas, no quadro D.3

Quadro D.1: Variáveis originais utilizadas no estudo com valor pré-definido

| **variável** | **descrição** | **valor pré-definido** |
|---|---|---|
| a02_est | 2005: a02-estrato do screening [tratamento, controle1, controle2] | 1 domicilio com BF<br>2 domicílio com outro benefício ou cadastrado<br>3 domicílio sem benefício e não cadastrado |
| sitdom | 2005: sitdom-situação do domicílio [urbano ou rural] | 1 rural<br>2 urbano |
| s01x2 | tipo de domicilio | 1 Casa<br>2 Apartamento<br>3 Quarto ou comodo |
| s01x3 | localização do domicilio | 1 Condominio de casas, apartamentos ou casas de vila<br>2 Favelas ou areas invadidas ou ocupadas<br>3 Casa de comodos ou corticos<br>4 Construcao isolada |
| s01x5 | tipo de rua onde se localiza o domicilio | 1 Asfaltada<br>2 Paralelepipedos<br>3 Terra batida ou sem pavimentacao<br>4 Outro tipo |

Quadro D.1: Variáveis originais utilizadas no estudo com valor pré-definido (continuação)

| variável | descrição | valor pré-definido |
|---|---|---|
| s01x6 | condição de ocupação do domicilio | 1 Alugado<br>2 Proprio em aquisicao<br>3 Proprio ja pago<br>4 Cedido por empregador<br>5 Cedido de outra forma<br>6 Outra condição |
| s01x7 | material predominante nas paredes externas | 1 Alvenaria<br>2 Madeira aparelhada<br>3 Tijolo sem revestimento<br>4 Taipa map revestida<br>5 Madeira aproveitada<br>6 Outro material |
| s01x8 | material predominante no piso | 1 Madeira aparelhada<br>2 Carpete<br>3 Ceramica, lajota, ardosia<br>4 Cimento<br>5 Madeira aproveitada<br>6 Terra<br>7 Outro material |
| s01x9 | material predominante no telhado (cobertura externa) | 1 Telha<br>2 Laje de concreto<br>3 Madeira aparelhada<br>4 Zinco ou amianto<br>5 Madeira aproveitada<br>6 Palha<br>7 Outro material |

Quadro D.1: Variáveis originais utilizadas no estudo com valor pré-definido (continuação)

| variável | descrição | valor pré-definido |
|---|---|---|
| s01x15 | tipo de escoadouro do banheiro ou sanitario | 1 Rede coletora de esgoto<br>2 Fossa septica<br>3 Fossa rudimentar<br>4 Vala<br>5 Outro Tipo<br>6 Nao tem |
| s01x17 | principal fonte de abastecimento de agua | 1 Rede geral<br>2 Poco ou nascente na propriedade<br>3 Poco ou nascente fora da propriedade<br>4 Bica Publica<br>5 Carro Pipa<br>6 Cisterna (agua de chuva) construida com recursos proprios<br>7 Cisterna (agua de chuva) construida com recursos do governo<br>8 Outra forma |
| s01x21 | principal tipo de agua usada para beber | 1 Filtrada<br>2 Fervida<br>3 Filtrada e fervida<br>4 Mineral<br>5 Natural<br>6 Directo da rede<br>7 Coada<br>8 Clorada |

Quadro D.1: Variáveis originais utilizadas no estudo com valor pré-definido (continuação)

| variável | descrição | valor pré-definido |
|---|---|---|
| s01x22 | principal tipo de iluminacao | 1 Electrica (rede geral)<br>2 Gerador (domiciliar)<br>3 Lampiao<br>4 Vela ou lamparina<br>5 Outro tipo |
| s01x24 | principal destino do lixo domiciliar | 1 Coletado diretamente por servico de limpeza<br>2 Coletado indirectamente<br>3 Queimado ou enterrado<br>4 Jogado em terreno baldio ou logradouro<br>5 Jogado em rio, lago ou no mar<br>6 Outro destino |
| s02a5 | qual a relacao de convivencia que (nome) tem com o responsavel | 1 Pessoa responsavel<br>2 Conjuge, companheiro (a)<br>3 Filho (a), enteado (a)<br>4 Pai, mae, sogro (a)<br>5 Neto (a), bisneto (a)<br>6 Irmao, irma<br>7 Nora, genro<br>8 Outro parente<br>9 Agregado<br>10 Pensionista<br>11 Empregado domestico<br>12 Parente de empregado domestico |

Quadro D.1: Variáveis originais utilizadas no estudo com valor pré-definido (continuação)

| variável | descrição | valor pré-definido |
|---|---|---|
| s02ad | sexo | 1 Masculino<br>2 Feminino |
| s03a1 | (nome) sabe ler e escrever um bilhete simples no idioma que conhece | 1 Sim<br>2 Não |
| s03a8 | qual o curso mais elevado que (nome) frequentou, no qual conclui ao menos 1 série? | 1 Creche<br>2 Pre-escolar<br>3 Classe de alfabetizacao<br>4 Alfabetizacao de adultos /alfabetizacao de Jovens e Adultos<br>5 Ensino fundamental ou 1 grau- regular seriado<br>6 Ensino fundamental ou 1 grau- regular nao-seriado<br>7 Supletivo / Educacao de Jovens e Adultos<br>8 Ensino medio ou 2 grau- regular seriado<br>9 Ensino medio ou 2 grau- regular nao-seriado<br>10 Supletivo / Educacao de Jovens e Adultos (medio ou 2 grau)<br>11 Pre-vestibular<br>12 Superior- graduacao<br>13 Pos graduacao em geral (Especializacao, Mestrado, ou doutorado) |
| s03a16 | qual o principal meio de transporte habitualmente utilizado | 1 Onibus publico<br>2 Trem/metro<br>3 Metro/onibus<br>4 Trem/metro/onibus<br>5 Transporte escolar (van, kombi, onibus escolar)<br>6 Carro ou moto particular<br>7 Outro veiculo proprio motorizado (lancha, trator)<br>8 Transporte proprio nao-motorizado (bicicleta, cavalo, canoa, etc.)<br>9 Outro tipo<br>10 Nao utiliza (vai a pe) |

Quadro D.1: Variáveis originais utilizadas no estudo com valor pré-definido (continuação)

| variável | descrição | valor pré-definido |
|---|---|---|
| s10a3028 | s10a04-possui ou aluga [automóvel] ? | 1 Sim<br>2 Não |
| s10a5028 | s10a06-ha quanto tempo possui [automóvel] ? | 1 Menos de 1 ano<br>2 De 1 a menos de 2 anos<br>3 2 anos o mais |
| s10a6028 | s10a07-forma de obtencao [automóvel] | 1 Compra a vista<br>2 Compra a prazo<br>3 Aluguel<br>4 Doacao<br>5 Troca<br>6 Recebimento em bens |
| s10a7028 | s10a08-pag. últimos 30 dias (compras prazo - últimos 24 meses)? [automóvel] | 1 Sim<br>2 Não |

*Fonte: elaboração própria*

Quadro D.2: Variáveis originais utilizadas no estudo com livre preenchimento

| **variável** | **descrição** |
|---|---|
| cod_dtm | 2009: identificador do domicilio |
| npes | número de ordem da pessoa |
| s01x13 | quantos banheiros existem neste domicilio (considere apenas os que tem chuveiro ou banheira ou aparelho sanitario) |
| s03b3 | quanto gastou em transporte escolar de (nome) nos ultimos 30 |
| s03b4 | quanto gastou com a merenda escolar de (nome) nos ultimos 30 |
| s03b5 | quanto gastou de material escolar de (nome) em 2009? |
| s03b6 | quanto gastou com a matricula de (nome) em 2009 |
| s03b8 | quanto gastou com (nome) a titulo de outras despesas com educacao nos ultimos 30 dias |
| s04c301 | s04d03-valor gasto pessoas com ate 14 anos [CONSULTAS] |
| s04c302 | s04d03-valor gasto pessoas com ate 14 anos [EXAMES] |
| s04c303 | s04d03-valor gasto pessoas com ate 14 anos [REMEDIO-CONTINUO] |
| s04c304 | s04d03-valor gasto pessoas com ate 14 anos [REMEDIO-OCASIONAL] |
| s04c305 | s04d03-valor gasto pessoas com ate 14 anos [PLANO DE SAUDE] |
| s04c306 | s04d03-valor gasto pessoas com ate 14 anos [INTERNAÇÃO] |
| s04c601 | s04d06-valor gasto pessoas com 15 anos ou mais [CONSULTAS] |
| s04c602 | s04d06-valor gasto pessoas com 15 anos ou mais [EXAMES] |
| s04c603 | s04d06-valor gasto pessoas com 15 anos ou mais [REMEDIO-CONTINUO] |
| s04c604 | s04d06-valor gasto pessoas com 15 anos ou mais [REMEDIO-OCASIONAL] |
| s04c605 | s04d06-valor gasto pessoas com 15 anos ou mais [PLANO DE SAUDE] |
| s04c606 | s04d06-valor gasto pessoas com 15 anos ou mais [INTERNAÇÃO] |
| s05b10 | s05b12 [ocup.1] renda mensal |
| s05b10_2 | s05b12 [ocup.2] renda mensal |
| s05b10_3 | s05b12 [ocup.3] renda mensal |

Quadro D.2: Variáveis originais utilizadas no estudo com livre preenchimento (continuação)

| **variável** | **descrição** |
|---|---|
| s06x2 | quanto [nome] recebeu [aposentadoria/previdencia]? |
| s06x4 | quanto [nome] recebeu [seguro-desemprego]? |
| s06x6 | quanto [nome] recebeu [pensao alimenticia]? |
| s06x8 | qual o valor estimado [alimentos, roupas ou outras mercadorias]? |
| s06x10 | quanto recebeu? [poupança+aluguel+venda+doações+FGTS+outros] |
| s07a2 | quanto [nome] gastou com transporte publico nos ultimos 7 dia |
| s07a4 | quanto [nome] gastou com combustivel, manutencao, estacionam |
| s07a6 | quanto [nome] gastou com comunicacoes nos ultimos 30 dias? |
| s07b2 | quanto [nome] gastou com alimentos e bebidas fora de casa no |
| s07b4 | quanto [nome] gastou nos ultimos 7 dias? |
| s10a4028 | s10a05-quantos possui [automóvel] ? |
| s10a8028 | s10a09-quantas prestacoes faltam [automóvel] |
| s10a9028 | s10a10-valor da prestacao [automóvel] |
| s10a10028 | s10a11-preco de compra (a vista - últimos 24 meses) [automóvel] |

*Fonte: elaboração própria*

Quadro D.3: variaveis criadas durante a preparaçao da base de dados para o estudo

| **variável** | **descrição** |
|---|---|
| a02_est_b | Identificador modificado do domicílio (a02_est) |
| s04c_y14 | Somatório dos gastos com saúde para pessoas menores de 14 anos [s04c301..s04c306] |
| s04c_y15 | Somatório dos gastos com saúde para pessoas maiores de 14 anos [s04c601..s04c606] |
| S05b10_tot | somatorio de prestação de serv. ou venda de produto - seção 5 // [s0cb10 + s0cb10_2 + s0cb10_3] |
| s06x_tot | quanto recebeu? [aposentadoria/previdencia] + [seguro-desemprego] + [pensao alimenticia] + + [alimentos, roupas ou outras mercadorias] + ]poupança+aluguel+venda+doações+FGTS+outros] |
| renda | somatorio de todo dinheiro adquirido (seções 6 e 5) por individuo |
| renda_dom | somatorio de todo dinheiro adquirido (seções 6 e 5) por domicílio |
| qtd_pes | numero pessoas p/domicílio registradas na base de dados |
| perkpta | rendimento percapita (seções 6 e 5): rend_dom/qtd_pes |
| s07a2* | quanto se gastou com transporte publico nos ultimos 7 dia |
| s07a4* | quanto se gastou com combustivel, manutencao, estacionam |
| s07a6* | quanto se gastou com comunicacoes nos ultimos 30 dias? |
| s07b2* | quanto se gastou com alimentos e bebidas fora de casa no |
| s07b4* | quanto se gastou nos ultimos 7 dias? |
| s08x5_tot | somatório dos valores desembolsados [GASTOS COLETIVOS](p/família) |
| s08x6_tot | somatório de valores estimados de produtos adquiridos [GASTOS COLETIVOS](p/família) |
| s08_tot | somatório de todos os valores desembolsados e estimados [GASTOS COLETIVOS](p/família) |
| s15c_tot1 | total, em 12 meses, do BF - Bolsa Família (BF) |
| bf_perkpta | Bolsa Familia (BF) - valor per capita |
| transp | principal meio de transporte habitualmente utilizado no domicílio |

**modificadas para comportar informações a cerca do domicílio, não mais do indivíduo*

*Fonte: elaboração própria*

# APÊNDICE E - Desempenho das métricas utilizadas frente as intervenções na amostra

O quadro E.1 apresenta o desempenho das métricas segundo as intervenções sofridas na amostra enquanto o quadro E.2 organiza esse desempenho segundo o classificador. Os valores alcançados que resultaram nos percentuais abaixo provem da tabela 5.1

Quadro E.1: Desempenho dos classificadores segundo as intervenções na amostra

| | **SUB-AMOSTRAGEM** | | | **SOBRE-AMOSTRAGEM** | | | **AMBOS** | | |
|---|---|---|---|---|---|---|---|---|---|
| | **ABD** | **RL** | **RNA** | **ABD** | **RL** | **RNA** | **ABD** | **RL** | **RNA** |
| Sensibilidade | -13,636% | -0,643% | -23,721% | -6,353% | -1,825% | -10,335% | -18,052% | -1,444% | -26,316% |
| Especificidade | 39,903% | -2,119% | 132,629% | 27,539% | 2,288% | 119,323% | 49,781% | 0,989% | 168,606% |
| Acurácia | -12,014% | -2,873% | 42,750% | -7,378% | -1,310% | 45,152% | -52,891% | -52,145% | -28,206% |
| *F-measure* | -13,636% | -0,643% | 22,882% | -6,353% | -1,825% | 63,107% | -25,996% | -6,156% | 54,396% |
| Área sob a curva | -0,587% | -0,587% | 52,667% | 0,874% | 0,874% | 94,474% | -63,026% | -66,671% | 77,367% |
| Eficiência | 3,490% | -1,347% | 71,182% | 4,491% | 0,150% | 119,014% | -18,569% | -31,841% | 82,412% |
| *G-mean* | 9,820% | -1,458% | -47,065% | 9,330% | 0,204% | -27,436% | -44,734% | -62,507% | -3,352% |
| MCC | 6,576% | 6,576% | -35,841% | 9,569% | 9,569% | -22,808% | -3,912% | 9,757% | -16,683% |
| Índice de Youden | 13,142% | -4,237% | -16,849% | 16,964% | 0,490% | -5,154% | 2,958% | -4,066% | -13,522% |
| Kappa | 9,060% | -20,602% | 20,962% | 12,956% | -6,715% | 22,224% | -24,923% | -53,120% | 26,847% |
| Taxa de erro global | 48,785% | 8,542% | -44,446% | 29,960% | 3,894% | -39,987% | 136,184% | 85,419% | -8,732% |
| Taxa de erro Positiva | -30,635% | 140,606% | 374,958% | -21,142% | 70,347% | 163,361% | 24,412% | 552,627% | 1061,008% |
| Taxa de erro Negativa | 165,789% | -45,409% | 49,957% | 77,237% | -30,458% | 17,808% | 822,632% | 45,152% | 59,662% |

*fonte: elaboração própria*

Quadro E.2: Desempenho de cada classificador frente às intervenções na amostra

| | **ÁRVORE** | | | **REGRESSÃO LOGÍSTICA** | | | **REDE NEURAL** | | |
|---|---|---|---|---|---|---|---|---|---|
| | **Under** | **Over** | **Both** | **Under** | **Over** | **Both** | **Under** | **Over** | **Both** |
| Sensibilidade | -13,636% | -6,353% | -18,052% | -0,643% | -1,825% | -1,444% | -23,721% | -10,335% | -26,316% |
| Especificidade | 39,903% | 27,539% | 49,781% | -2,119% | 2,288% | 0,989% | 132,629% | 119,323% | 168,606% |
| Acurácia | -12,014% | -7,378% | -52,891% | -2,873% | -1,310% | -52,145% | 42,750% | 45,152% | -28,206% |
| *F-measure* | -13,636% | -6,353% | -25,996% | -0,643% | -1,825% | -6,156% | 22,882% | 63,107% | 54,396% |
| Área sob a curva | -0,587% | 0,874% | -63,026% | -0,587% | 0,874% | -66,671% | 52,667% | 94,474% | 77,367% |
| Eficiência | 3,490% | 4,491% | -18,569% | -1,347% | 0,150% | -31,841% | 71,182% | 119,014% | 82,412% |
| *G-mean* | 9,820% | 9,330% | -44,734% | -1,458% | 0,204% | -62,507% | -47,065% | -27,436% | -3,352% |
| MCC | 6,576% | 9,569% | -3,912% | 6,576% | 9,569% | 9,757% | -35,841% | -22,808% | -16,683% |
| Índice de Youden | 13,142% | 16,964% | 2,958% | -4,237% | 0,490% | -4,066% | -16,849% | -5,154% | -13,522% |
| Kappa | 9,060% | 12,956% | -24,923% | -20,602% | -6,715% | -53,120% | 20,962% | 22,224% | 26,847% |
| Taxa de erro global | 48,785% | 29,960% | 136,184% | 8,542% | 3,894% | 85,419% | -44,446% | -39,987% | -8,732% |
| Taxa de erro Positiva | -30,635% | -21,142% | 24,412% | 140,606% | 70,347% | 552,627% | 374,958% | 163,361% | 1061,008% |
| Taxa de erro Negativa | 165,789% | 77,237% | 822,632% | -45,409% | -30,458% | 45,152% | 49,957% | 17,808% | 59,662% |

*nota: **under**sampling (sub-amostragem) **over**sampling (sobre-amostragem) **both** ⇒ sobre-amostragem e sub-amosragem

*fonte: elaboração própria*