Instituto Serzedello Corrêa – ISC
Pós-Graduação em _____

Marcel Azevedo Coutinho de Freitas

Utilização de aprendizagem supervisionada para a detecção de benefício assistencial com maior probabilidade de conter irregularidade

Orientador(a):
Prof. Dr. Remis Balaniuk

Brasília
2020

Marcel Azevedo Coutinho de Freitas

Utilização de aprendizagem supervisionada para a detecção de benefício assistencial com maior probabilidade de conter irregularidade

Orientador:

Prof. Dr. Remis Balaniuk

Monografia submetida ao Instituto Serzedello Corrêa do Tribunal de Contas da União como requisito para a obtenção do grau de especialista.

Brasília
2020

Marcel Azevedo Coutinho de Freitas

Utilização de aprendizagem supervisionada para a detecção de benefício assistencial com maior probabilidade de conter irregularidade

Monografia submetida ao Instituto Serzedello Corrêa do Tribunal de Contas da União como requisito para a obtenção do grau de especialista.

Data de aprovação:

31/03/2020

Banca examinadora:

_____

Prof. Dr. Remis Balaniuk

_____

Prof. Rodrigo Otávio Coelho Hildebrand

Brasília
2020

Agradecimentos

A Deus por tudo que tem me propiciado em minha vida.

Aos meus pais, Marcelino e Selma, por tudo que fizeram por mim, especialmente por aquela parte que superava suas possibilidades, saibam que vocês foram grandes incentivadores e fonte de inspiração nessa jornada.

À Laura Martins por sua compreensão e por me fazer acreditar, do início ao fim, que poderia realizar esse sonho que ora materializo.

Aos professores do curso pela dedicação no planejamento das aulas e das atividades extras.

Aos companheiros de caminhada pela experiência compartilhada e pelo incentivo recíproco.

Resumo

A Previdência Social e a Assistência Social são importantes políticas públicas e nelas são investidos uma quantidade expressiva de recursos públicos. Nesse sentido, a verificação da regularidade dos benefícios deve, necessariamente, levar esse fator em consideração. A tecnologia de mineração de dados (Data Mining) é uma abordagem viável quando se lida com grande volume e tem diversas áreas de aplicação, inclusive na detecção de fraudes. Diante disso, este trabalho foi realizado com objetivo criar um modelo preditivo capaz de identificar benefícios assistenciais com maior probabilidade de conter irregularidades utilizando como insumos as características dos conjuntos de dados disponíveis. Por fim, registra-se a importância de testar o modelo preditivo com dados reais afim de medir o seu real desempenho, visto que ele é treinado e testado com base em um conjunto de dados que reflete o passado. Todavia, o seu objetivo é identificar futuros benefícios com maior probabilidade de conter fraude.

Palavras-chave: detecção de fraude; modelo de classificação; seguridade social

Abstract

Social Security is an essential public policy, and it is invested a significant amount of public resources in it. Therefore, the regularity of benefits must necessarily consider that circumstance. Data mining is a viable approach when dealing with large volumes of data, and it has several applications, including fraud detection. This paper proposes to create a predictive model able to identify pensions with a higher probability of containing irregularities using the characteristics of the available data sets as inputs. The results obtained from the proposal reinforces the importance of testing the predictive model with real data to measure the actual performance because it is trying to predict future data using a dataset that reflects the past.

Keywords: fraud detection; classification models; pensions

Lista de gráficos

Lista de tabelas

Lista de quadros

Sumário

1. Introdução

Em dezembro de 2019, havia 35,4 milhões de benefícios ativos na folha de pagamento do INSS e, no ano de 2019, foi gasto R$ 616 bilhões com o pagamento de benefícios previdenciários e assistenciais. Esses valores demonstram a materialidade dessa política pública no que diz respeito ao orçamento público, bem como quanto ao impacto na vida da população.

O TCU realizou uma auditoria na qual estimou[1] que 11,4% do percentual de benefícios eram pagos indevidamente. Isso representa um montante de R$ 70.2 bilhões, o que equivale a 62% do orçamento da educação no ano de 2019.

Considerando os números expostos acima, percebe-se que, além da grande quantidade de recursos desembolsados anualmente, há também o desafio de lidar com uma enorme quantidade de dados com potencial de gerar substancial economia de recursos públicos.

Todavia, qualquer iniciativa no sentido de identificar benefícios concedidos irregularmente tem de ser planejada para que seja escalável a fim de o custo dessa identificação e posterior revisão de benefícios irregulares não seja superior à economia decorrente da cessação desses benefícios.

Por outro lado, entre janeiro de 2014 e dezembro de 2019, foram efetuados pagamentos para 95 espécies de benefícios previdenciários e assistenciais pelo INSS. Essas espécies de benefícios possuem diferentes objetivos, público-alvo e regras de elegibilidade o que tornou necessário escolher uma espécie de benefício, qual seja Benefícios de Prestação Continuada ao Idoso (BPC idoso) para a execução deste trabalho.

Diante disso, este trabalho se propôs a responder a seguinte pergunta: é possível distinguir os benefícios com maior probabilidade de conter irregularidades dos benefícios regulares com base nas características disponíveis dos dados com razoável margem de acerto?

---

[1] Acordão 1.057/2018-TCU-Plenário

Nesse sentido, o objetivo geral deste trabalho é a criação de um modelo preditivo que auxilie na identificação de benefícios da espécie BPC idoso com a maior probabilidade de conterem irregularidades.

Por fim, as etapas cumpridas neste trabalho estão distribuídas por capítulos seguindo o modelo CRISP-DM, contemplando a compreensão do negócio, a compreensão dos dados e justificativa, a preparação dos dados, a modelagem dos dados, a avaliação, a aplicação e considerações finais.

2. Compreensão do negócio

Conforme registrado no art. 203 da Constituição Federal de 1988 (CF/88), a assistência social será prestada a quem dela necessitar, independentemente de contribuição, e visa proteger a família, a infância, a adolescência e a velhice.

Na regulamentação do art. 203 da CF/88, a Lei nº 8.742, de 7 de dezembro de 1993, (Lei Orgânica da Assistência Social – LOAS) dispôs que:

> Art. 12. Compete à União:
>
> I - Responder pela concessão e manutenção dos benefícios de prestação continuada definidos no art. 203 da Constituição Federal;

Além disso, por intermédio da Lei Orgânica da Assistência Social, foi definido o valor do Benefício de Prestação Continuada (BPC), que visa proteger a pessoa com deficiência e o idoso em situação de vulnerabilidade. A lei definiu os critérios de elegibilidade para a concessão do benefício, nos seguintes termos:

> Art. 20. O **benefício de prestação continuada** é a garantia de um salário-mínimo mensal à pessoa com deficiência e ao **idoso com 65 (sessenta e cinco) anos ou mais** que comprovem **não possuir meios de prover a própria manutenção nem de tê-la provida por sua família**.
>
> § 1º Para os efeitos do disposto no caput, a família é composta pelo requerente, o cônjuge ou companheiro, os pais e, na ausência de um deles, a madrasta ou o padrasto, os irmãos solteiros, os filhos e enteados solteiros e os menores tutelados, desde que vivam sob o mesmo teto.
>
> (...)
>
> § 3º Considera-se incapaz de prover a manutenção da pessoa com deficiência ou idosa a família cuja **renda mensal per capita seja inferior a 1/4 (um quarto) do salário-mínimo**.
>
> § 4° **O benefício** de que trata este artigo **não pode ser acumulado pelo beneficiário com qualquer outro no âmbito da seguridade social ou de outro regime**, salvo os da assistência médica e da pensão especial de natureza indenizatória.

Todavia, cabe ressaltar que há 26 Ações Civis Públicas (ACPs) em vigor no país que alteram o critério de elegibilidade para o Benefício de Prestação Continuada, produzindo efeitos em diferentes regiões do país, conforme é tratado no Anexo A deste trabalho. Além disso, a produção de efeitos dessas ACPs, em sua maioria, fica restrita à jurisdição do juízo que proferiu a decisão.

Um exemplo é a ACP 2007.83.05.000083-0 que estabeleceu para os municípios situados na jurisdição da 23ª Vara Federal de Garanhuns (Pernambuco) critério de renda familiar *per capita* valor igual ou inferior a 1/2 salário mínimo, diferente do valor de1/4 do salário mínimo estabelecido no § 3º do art. 20 da LOAS.

Outro exemplo é a ACP 0004265-82.2016.4.03.6105 que definiu para os municípios localizados na jurisdição da 8ª Vara Federal de Campinas (São Paulo) que se exclua do cálculo da renda familiar benefícios previdenciários e assistenciais no valor de até um salário mínimo recebidos por outro membro do grupo familiar do idoso.

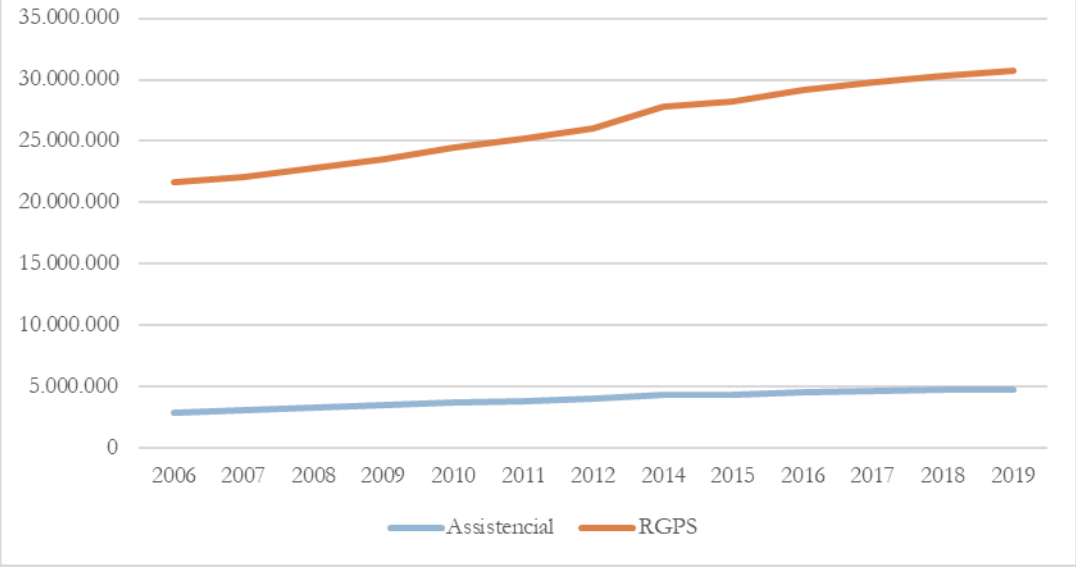
Além das ACPs, há casos em que os benefícios são concedidos em atendimento à decisão judicial. Diferente do que ocorre com as ACPs, essas decisões produzem efeito *inter partes*, ou seja, restrito àqueles que participaram da respectiva ação judicial.

Diante disso, considera-se importante que os efeitos decorrentes de decisões judiciais sejam levados em conta na execução de trabalhos como este em etapas como: seleção de dados, criação de novos atributos, integração de atributos provenientes de outras bases de dados e avaliação dos atributos considerados mais relevantes pelo modelo preditivos.

A operacionalização do pagamento do BPC idoso é realizada pelo Instituto Nacional do Seguro Social (INSS), assim como ocorre com a outra espécie de BPC cujo público alvo são as pessoas com deficiência.

Isso faz com que o INSS seja responsável por operacionalizar a concessão, manutenção e pagamento de 35,4 milhões de benefícios. O gráfico below ilustra a evolução da quantidade de benefícios pagos pelo INSS entre os anos de 2006 e 2019.

Gráfico 1: Quantidade de benefícios previdenciários e assistenciais pagos pelo INSS



Fonte: Boletim Estatístico da Previdência Social (BEPS) de outubro de 2019.

Nota: Entre 2006 e 2018 foi adotado o mês de dezembro como mês de referência e em 2019 utilizou-se os dados do mês de outubro visto que era a informação mais atualizada daquele ano

Inicialmente havia 21,6 milhões de benefícios previdenciários e outros 2,9 milhões benefícios assistenciais em dezembro de 2006 pagos por meio da Maciça, que é a folha de pagamento ordinária do INSS. Esse número aumentou, alcançando 30,7 milhões de benefícios previdenciários, além de 4,7 milhões de benefícios assistenciais em outubro de 2019.

Além disso, é possível identificar uma tendência no aumento da quantidade de benefícios previdenciários e assistenciais, uma vez que entre 2006 e 2019 houve um aumento na quantidade de benefícios previdenciários e assistenciais de 42% e 62%, respectivamente.

Vale ressaltar que, como há a possibilidade de o INSS realizar pagamento de benefícios sem que eles sejam registrados na Maciça, como é o caso do Pagamento Alternativo de Benefício (PAB) ou primeira concessão, pode-se afirmar que estes números não abarcam a totalidade dos benefícios pagos pelo INSS.

Por meio da tabela e do gráfico below é apresentado o pagamento de benefícios sob a ótica da despesa.

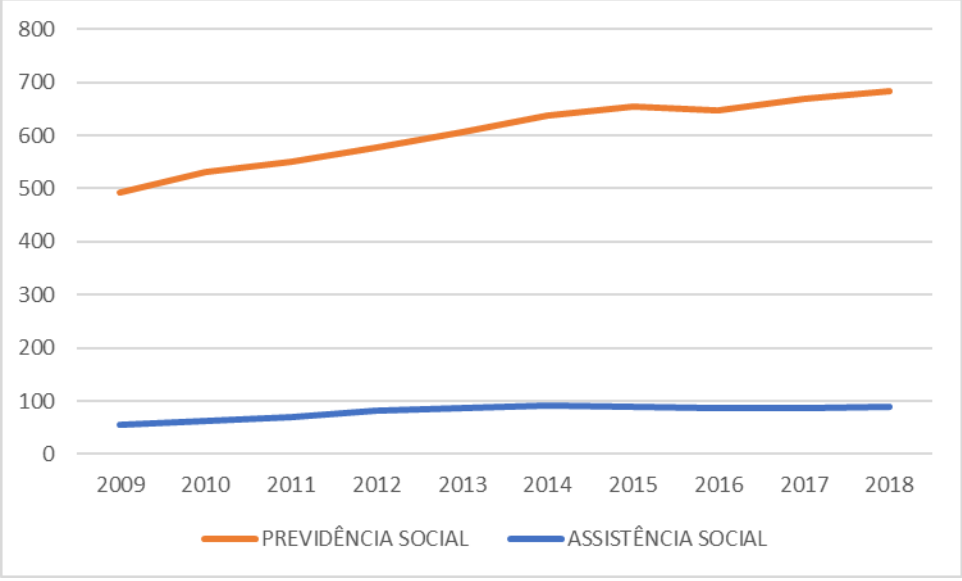
Tabela 1: Despesas Empenhadas por Função de Governo - a valores de 2018 pelo INPC

Função de Governo	2009	2010	2011	2012	2013	2014	2015	2016	2017	2018
Previdência Social	493,39	531,64	550,57	576,73	605,59	636,91	655,14	647,78	669,33	684,26
Assistência Social	56,5	63,68	69,69	81,64	87,75	90,57	88,64	86,76	86,46	88,67
Total	**549,89**	**595,32**	**620,26**	**658,37**	**693,34**	**727,48**	**743,78**	**734,54**	**755,79**	**772,93**

Fonte: Tesouro Gerencial e INPC/IBGE.

Nota: Valores em bilhões de reais.

Gráfico 2: Despesas Empenhadas por Função de Governo - a valores de 2018 pelo INPC



Fonte: Tesouro Gerencial e INPC/IBGE.

Nota: Valores em bilhões de reais.

Primeiramente, ressalta-se que as despesas acima foram atualizadas a valores de 2018 pelo Índice Nacional de Preços ao Consumidor (INPC) de forma a expurgar os efeitos inflacionários. Isto posto, verifica-se que houve aumento real no período analisado (2009-2018) das despesas com previdência e assistência social de 39% e 57%, respectivamente.

Com isso, percebe-se que há uma tendência de aumento tanto do montante dispendido como da quantidade de benefícios com o passar dos anos. Dessa forma, é possível concluir que a avaliação da regularidade desses benefícios precisa ser realizada de maneira escalável, sob o risco de se tornar inviável no médio e longo prazo.

Neste ponto, vale destacar que há previsão legal para a realização de revisão de benefícios, há iniciativas de gestão que podem melhorar a qualidade dos dados utilizados pelo INSS e há fatores relacionados ao TCU que favorecem à execução deste trabalho.

No que diz respeito ao aspecto legal, no caso dos benefícios de prestação continuada há duas leis que autorizam a revisão do benefício. A primeira é a Lei Orgânica da Assistência Social na qual está definido que:

> Art. 21. **O benefício de prestação continuada deve ser revisto a cada 2 (dois) anos** para avaliação da continuidade das condições que lhe deram origem.
>
> § 1º O pagamento do benefício cessa no momento em que forem superadas as condições referidas no caput, ou em caso de morte do beneficiário.
>
> § 2º **O benefício será cancelado quando se constatar irregularidade na sua concessão** ou utilização.

A segunda é a Lei nº 13.846, de 18 de junho de 2019 (Lei de Combate às Fraudes Previdenciárias) que instituiu o Programa Especial para Análise de Benefícios com Indícios de Irregularidade e criou mecanismos que mitigam o risco de que a implantação desse Programa pelo INSS afete o atendimento aos cidadãos que requerem novos benefícios.

No que concerne às iniciativas da gestão do INSS que podem melhorar a qualidade dos dados utilizados, cita-se, como exemplo, o INSS Digital e a criação do Sistema Nacional de Informações de Registro Civil (SIRC).

O INSS Digital consiste na alteração do fluxo de atendimento ao cidadão com objetivo de aumentar a capacidade da autarquia, otimizando o uso de sua força de trabalho. Um dos pilares do INSS Digital é a implantação do processo eletrônico o que permitirá registros mais precisos da data que o benefício foi requerido e a data de sua concessão ou indeferimento.

Vale ressaltar que a diferença entre a data de entrada do requerimento no INSS e a data de concessão do benefício (D2_DDBxD2_DER) foi uma variável considerada relevante pelo modelo de classificação para identificar os benefícios com maior probabilidade de conter irregularidades.

O SIRC, por sua vez, é o sistema por meio do qual os Cartórios de Registro Civil alimentam uma base de dados com informações de registros civis de pessoas naturais do Brasil. Com isso, o INSS terá acesso ao dados relativos a registros de nascimento, de casamento, de óbito e de natimorto, conferindo maior celeridade e completude ao sistema.

Cabe destacar que esses dados são essenciais para a administração do sistema de benefícios previdenciário e assistencial e permitem a otimização de processos de trabalho. Por um lado, reduz a quantidade de documentos que os cidadãos precisam apresentar para requerer um benefício e mitiga erros na internalização das informações constantes nesses documentos para as bases de dados da autarquia. Por outro lado, permite a concessão automática de algumas espécies de benefícios como é o caso do auxílio maternidade.

Diante dos aspectos legais e das iniciativas acima descritos, pode-se afirmar que o ambiente de negócio é propício e, portanto, é viável a revisão de benefícios que venham a ser apontados com maior probabilidade de conter irregularidades pelo INSS.

Com relação aos fatores relativos ao TCU, registra-se que o acesso aos dados disponíveis no LabContas. Com isso, foram utilizados como fontes de dados para este trabalho bases de dados como, por exemplo, Concessão_2[2], Maciça[3], além de tabelas auxiliares que permitem identificar os códigos utilizados para determinados campos.

Outro fator relacionado ao TCU é a existência da Diretoria de Análise de Dados e Tecnologia da Informação (DCAD) que é formada por uma a equipe de auditores os quais são especialistas nas bases de dados utilizadas neste trabalho, haja vista a expertise adquirida na realização dos Ciclos de Fiscalização Contínua de Benefícios (FCB), conforme evidenciado na tabela **Error! Reference source not found.**.

Quadro 1: Ciclos da Fiscalização Contínua de Benefícios sociais

Função Ciclo	Acórdão	Data da sessão	Relator	TC
Previdência Ciclo 1	718/2016-P	30/3/2016	Ministro Vital do Rêgo	010.947/2015-9
Previdência Ciclo 2	1.057/2017-P	24/5/2017	Ministro Vital do Rêgo	016.216/2016-4
Previdência Ciclo 3	1.057/2018-P	9/5/2018	Ministro-Substituto André Luís de Carvalho	017.519/2017-9
Previdência Ciclo 4	1.947/2019-P	18/8/2019	Ministro Raimundo Carreiro	021.408/2018-1
Trabalho Ciclo 1	1.181/2016-P	11/5/2016	Ministro-Substituto Weder de Oliveira	022.036/2015-6
Trabalho Ciclo 2	1.058/2017-P	24/5/2017	Ministro Vital do Rêgo	016.474/2016-3
Trabalho Ciclo 3	1.343/2018-P	13/6/2018	Ministro Benjamin Zymler	020.992/2017-3
Trabalho Ciclo 3	1.947/2019-P	18/8/2019	Ministro Raimundo Carreiro	021.408/2018-1
Assistência Ciclo 1	1.009/2016-P	27/4/2016	Ministro-Substituto Weder de Oliveira	030.760/2015-1
Assistência Ciclo 2	1.344/2017-P	28/6/2017	Ministro-Substituto Weder de Oliveira	012.474/2016-9
Assistência Ciclo 3	12.162/2018-2ª	4/12/2018	Ministro-Substituto André Luís de Carvalho	020.222/2017-3
Assistência Ciclo 4	1.947/2019-P	18/8/2019	Ministro Raimundo Carreiro	021.408/2018-1

[2] Base de dados [BD_BENEFICIOS_HIST].[dbo].[CONCESSAO_2] do LabContas
[3] Base de dados [BD_BENEFICIOS_HIST].[dbo].[MACIÇA] do LabContas

Fonte: Relatório do Acórdão 1.947/2019-P (TC 021.408/2018-1) com adaptações

Além dos trabalhos da FCB, vale destacar que o TCU, por meio do Acordão 1.057/2018-TCU-Plenário (TC 017.519/2017-9), utilizou simulações de Monte Carlo[4] para estimar em 11,4% (limite inferior, utilizando um intervalo de confiança de 90%) o percentual de benefícios pagos indevidamente.

O órgão responsável por operacionalizar as políticas de Previdência Social, Assistência Social e Trabalho no Reino Unido (*Department for Work & Pensions*) apresentou relatório, em maio de 2019, no qual estimou o montante gasto com pagamentos incorretos de benefícios naquele país.

Naquele trabalho, foi estimado que 2,2 % do montante gasto por aquele órgão com o pagamento dos benefícios sociais decorriam de pagamentos acima do valor devido, o que representava £4.1 bilhões (R$ 21,8 bilhões[5]).

No quarto ciclo da Fiscalização Contínua de Benefícios realizado pelo Tribunal de Contas da União em 2018 (TC 021.408/2018-1), foram identificados 60.717 benefícios previdenciários e 56.739 benefícios assistenciais com indícios de irregularidade por meio do uso de tipologias, conforme demonstrado na tabela below:

Tabela 2: Quantidade e valores das tipologias verificadas na FCB 2018

Função de Governo	Tipologias FCB		Maciça Dez/2018		Tipologia FCB como percentual da Maciça	
	Quant.	Valor	Quant.	Valor	Quant.	Valor
Previdência	60.717	73.621.236,49	32.847.306	45.518.152.040,34	0,18%	0,16%
Assistência	56.739	54.129.006,00	2.046.956	1.952.798.534,50	2,77%	2,77%

Fonte: Relatório do Acórdão 1.947/2019-P (TC 021.408/2018-1) com adaptações

Cabe destacar que tipologia consiste no conjunto de testes realizados na base de dados com objetivo de identificar benefícios com indícios de irregularidade em virtude de descumprimento das regras de negócio, dos critérios para elegibilidade e valores pagos em desacordo com legislação em vigor.

[4] Designa-se por método de Monte Carlo (MMC) qualquer método de uma classe de métodos estatísticos que se baseiam em amostragens aleatórias massivas para obter resultados numéricos, isto é, repetindo sucessivas simulações um elevado número de vezes, para calcular probabilidades heuristicamente, tal como se, de fato, se registrassem os resultados reais em jogos de cassino (Wikipedia)

[5] Cálculo realizado por meio do site do Banco Central (https://www.bcb.gov.br/conversao) considerando a taxa de câmbio de 31/12/2019.

As tipologias utilizadas na FCB 2018 que diziam respeito ao Benefícios de Prestação Continuada podem ser divididas em dois grupos, quais sejam renda e cadastro.

A tipologia relacionada ao grupo renda busca identificar os beneficiários com renda formal, por meio de cruzamento com outras bases do governo.

As tipologias que concernem ao cadastro têm o objetivo de identificar benefícios com indícios de que o titular do benefício: faleceu, teve o CPF cancelado ou suspenso, possui informações cadastradas incompatíveis com a base do CPF ou não possui CPF.

Cumpre ressaltar que os indícios de erros cadastrais ou falecimento dos beneficiários ou dos membros da família, mesmo que confirmados, não indicam necessariamente uma fraude aos programas.

Por outro lado, a existência desses erros indica falhas de atualização e manutenção das informações cadastrais que podem ser aproveitadas para o cometimento de fraudes.

Considerando a estimativa supracitada de que 11,4% dos benefícios são pagos indevidamente, observa-se que, em tese, há oportunidade de melhorias que contribuam para a identificação de benefícios previdenciários (0,18%) e assistenciais (2,77%) com indícios de irregularidades.

Por último, cumpre registrar que para considerar o resultado obtido pelo modelo desenvolvido como bem sucedido sob a perspectiva do negócio, é importante levar em consideração o custo da revisão dos benefícios indicados como aqueles em que há a maior probabilidade de conter irregularidades. Ao tratar do custo-benefício da implantação do controle, Dantas *et al* (2010) concluiu que:

> "(..) **uma das razões para as limitações de eficácia do controle interno é a relação custo versus benefício da implementação de determinado mecanismo de controle**. Entre as razões que justificam esse tipo de análise, pode-se destacar: **o custo dos controles internos não deve ser superior aos benefícios que deles se esperam**; as organizações têm recursos limitados e devem priorizar sua utilização nas atividades (incluindo os controles) que agregam mais valor; no caso dos controles, os recursos devem ser investidos para mitigar os riscos mais relevantes; e o excesso de controles pode onerar demasiadamente o processo, tornando-o dispendioso e contraproducente."

Portanto, é importante que o modelo desenvolvido tenha uma razoável assertividade ao propor que o INSS revise benefícios previstos como os de maior probabilidade de conter irregularidade, tendo em vista o custo de revisão de cada benefício.

2.1 Problema e Objetivos da pesquisa

Isto posto, verifica-se um aumento tanto do montante despendido como da quantidade de benefícios previdenciários e assistências. Isso torna necessário pensar no custo-benefício de uma iniciativa que busque identificar benefícios irregulares.

Nesse sentido, verifica-se a existência de iniciativas que tendem a tornar o ambiente de negócio cada vez mais propício para a uso de *machine learning* no auxílio de identificar irregularidades.

Diante do exposto, este trabalho se propõe a responde a seguinte pergunta: é possível distinguir os benefícios com maior probabilidade de conter irregularidades dos benefícios regulares com base nas características disponíveis dos dados com razoável margem de acerto?

Para obter subsídios que permitam responder a essa pergunta, foi definido como objetivo geral deste trabalho propor um modelo preditivo capaz de auxiliar na identificação de benefícios com maior probabilidade de conter irregularidades.

Ademais, também foram definidos os seguintes objetivos específicos:

a) Identificar as bases oficiais que podem ser utilizadas para fornecer insumos para a criação do modelo preditivo
b) Identificar nas bases oficiais os atributos que podem compor um conjunto de dados a ser utilizado por um modelo preditivo;
c) Selecionar a espécie de benefício mais adequada para a criação de um modelo preditivo
d) Criar um conjunto que permita identificar nas bases de dados oficiais elementos que permitam construir o modelo preditivo;
e) Avaliar os resultados obtidos pelo modelo preditivo

3. Modelos Preditivos

Machine learning (aprendizado de máquina) é um campo de estudo que faz parte da inteligência artificial e trata a respeito de como tornar as máquinas aptas a aprender (RÄTSCH, 2004). Grus (2019), por sua vez, conceitua aprendizado de máquina como a criação e uso de modelos cujo aprendizado é obtido a partir dos dados. O autor acrescenta que é possível utilizar o aprendizado de máquina para desenvolver modelos a partir dos os dados existentes de forma a prever os resultados para novos dados, como:

a) Se uma mensagem de e-mail é spam ou não;
b) Se uma transação com cartão de crédito é fraudulenta;
c) Em qual anúncio um comprador provavelmente clicará;
d) Qual time de futebol americano vencerá o *Super Bowl.*

O aprendizado de máquina pode ser dividido em dois subgrupos, quais sejam aprendizagem não-supervisionada e a aprendizagem supervisionada.

Duda *et al* (1973) ensina que na aprendizagem não-supervisionada não são utilizadas as informações das variáveis de saída. Assim, os dados de entradas são analisados e agrupados segundo a proximidade de seus valores, de forma que é utilizado um rótulo para cada grupo ou cluster, o que permite identificar a qual grupo cada registro pertence (DUDA *et al*, 1973).

De acordo com Sathya e Abraham (2013), a aprendizagem supervisionada consiste no treinamento de uma amostra de dados do conjunto de dados na qual a classificação correta já atribuída. Em outras palavras, no conjunto de dados consta os valores das variáveis de saída, ou variáveis-objetivo (*target*), que são as variáveis as quais se deseja predizer a partir dos dados existentes.

Hastie *et al* (2009) ressalta que uma dificuldade que os métodos supervisionados encontram é o *overfitting* (sobreajuste) do modelo ao conjunto de dados. Segundo os autores, isso dificulta a predição de novos exemplos uma vez que o modelo deixa de aprender com os dados e passa a decorá-los. Os autores alertam que isso ocorre com distribuições não balanceadas, em que a grande maioria de casos é de uma classe, como, normalmente, é o caso de detecção de fraudes eventos fraudulentos que são raros.

Em busca do alcance do objetivo geral deste trabalho foram utilizados nove algoritmos diferentes para identificar o algoritmo mais adequado, haja vista as

características do conjunto de dados. Considerando que a literatura sobre o tema está escrita, predominantemente, em inglês, procurou-se manter os nomes dos algoritmos também em língua inglesa para permitir uma melhor identificação dos algoritmos utilizados.

Assim, os nove algoritmos utilizados neste trabalho foram: *Multi-layer Perceptron Classifier*, *KNeighbors Classifier*, *Support Vector Classifier* (SVC), *Gaussian Process Classifier*, *Decision Tree Classifier*, *Random Forest Classifier*, *Ada Boost Classifier*, *Gaussian Naive Bayes* e *Quadratic Discriminant Analysis*.

O *Multi-layer Perceptron* (MLP) é um algoritmo de aprendizado supervisionado que aprende uma função $f(\cdot) : R^m \rightarrow R^o$ treinando em um conjunto de dados, onde m é o número de dimensões para entrada e o é o número de dimensões para saída. Dado um conjunto de recursos $X = x_1, x_2, \ldots, x_m$ e a coluna objetivo (*target*) é o *y*, ele pode aprender um aproximador de função não linear para classificação ou regressão. É diferente da regressão logística, pois entre a camada de entrada e a de saída, pode haver uma ou mais camadas não lineares, chamadas camadas ocultas.

O algoritmo *KNeighbors Classifier*, por sua vez, é um tipo de aprendizado baseado em instância ou não generalizante. Neste algoritmo a classificação é calculada a partir do número de vizinhos (k) mais próximos de cada ponto a ser classificado

No que se refere ao algoritmo *Support Vectors Classifier* (SVC), esse algoritmo se propõe a encontrar o melhor hiperplano para separar as diferentes classes, maximizando a distância entre os pontos de amostra e o hiperplano.

No que concerne ao algoritmo *Gaussian Process Classifier* (processos gaussianos), esse algoritmo é um método genérico de aprendizado supervisionado, projetado para resolver problemas de regressão e classificação probabilística. As vantagens dos processos gaussianos são: A previsão interpola as observações (pelo menos para kernels regulares).

Com relação ao algoritmo *Decision Tree Classifier* (árvore de decisão), esse algoritmo consiste em uma representação simples para classificar exemplos. Ele é dos algoritmos usados em aprendizado de máquina supervisionado e nele os dados são continuamente divididos de acordo com um determinado parâmetro.

No que diz respeito ao algoritmo *Random Forest Classifier* (floresta aleatória), ele é um algoritmo de classificação que consiste em muitas árvores de decisão. Ele cria um conjunto de árvores de decisão a partir do subconjunto selecionado aleatoriamente

do conjunto de treinamento. Em seguida, agrega os votos de diferentes árvores de decisão para decidir a classe final do objeto de teste.

Acerca ao algoritmo *Ada Boost Classifier*, ele combina vários classificadores para aumentar a precisão dos classificadores. O classificador *Ada Boost* cria um classificador forte, combinando vários classificadores de baixo desempenho, para obter um classificador forte de alta precisão. O conceito básico por trás do Ada Boost é definir os pesos dos classificadores e treinar a amostra de dados em cada iteração, de modo a garantir previsões precisas de observações incomuns. Qualquer algoritmo de aprendizado de máquina pode ser usado como classificador base, se ele aceitar pesos no conjunto de treinamento.

No que tange ao algoritmo *Gaussian Naive Bayes*, presume-se que os valores contínuos associados a cada atributo sejam distribuídos de acordo com uma distribuição gaussiana, que consiste em uma distribuição normal. Além disso, esse classificador trabalha com base no teorema de Bayes, que descreve a probabilidade de um evento, com base no conhecimento prévio das condições e estar relacionado às condições do evento.

Por último, o algoritmo *Quadratic Discriminant Analysis* é um classificador quadrático usado na aprendizagem de máquina e na classificação estatística para separar medidas de duas ou mais classes de objetos ou eventos por uma superfície quadrática.

Uma vez apresentados os algoritmos a serem usados para a criação do modelo preditivo, o próximo passo consiste na definição de quais parâmetros seriam utilizados para definir quais eram os melhores modelos.

Mas antes de estabelecer quais parâmetros são os mais adequados para mensurar se o modelo desenvolvido é razoavelmente assertivo, é necessário realizar breves considerações acerca da matriz de confusão e de alguns conceitos que dela decorrem como: *accuracy*, *precision* e *recall*, os quais se optou por traduzir para não perder suas cargas semânticas.

Quadro 2: Matriz de confusão para um problema de classificação com duas classes

		Valor previsto	
		Fraude	**Não Fraude**
Valor real	**Fraude**	Verdadeiros positivos	Falsos negativos

	Não fraude	Falsos positivos	Verdadeiros Negativos

Fonte: VISA *et al* (2011) com adaptações

Como é possível inferir a partir da matriz de confusão constante no quadro above, há quatro respostas possíveis a serem obtidas pelo modelo, quais sejam verdadeiros positivos, falsos positivos, falsos negativos, e verdadeiros negativos.

Segundo Grus (2019), verdadeiros positivos são os casos em que a classe que se busca identificar é prevista corretamente pelo modelo, isto é, o benefício foi cessado em razão de constatação de fraude ou irregularidade e o modelo o identificou corretamente como fraudado ou irregular.

Com relação ao falso positivo, Grus (2019) afirma que é quando o modelo prevê incorretamente a classe que é alvo da identificação. Em outras palavras, é o caso de o benefício não ter sido cessado por motivo de fraude ou irregularidade, mas o modelo o classifica como benefício fraudado ou irregular.

De acordo com Grus (2019), falsos negativos ocorrem quando o modelo prevê incorretamente a classe que não está sendo buscada. Em outros termos, esse é o caso em que o modelo classifica um benefício como não sendo irregular ou fraudado, não obstante esse benefício tenha sido cessado em virtude de identificação de fraude ou irregularidade.

Por último, Grus (2019) define falso verdadeiro como a previsão correta do modelo da classe que não está sendo buscada. De outra forma, é quando o modelo classifica um benefício como não sendo irregular ou fraudado e ele, de fato, não foi cessado por constatação de irregularidade ou fraude.

Quanto à definição de *accuracy*, Ben-David (2007) a descreve como a proporção de previsões acertadas do modelo, além disso o autor acrescenta que esta medida é de longe a mais dominante quando se fala em medir a precisão dos classificadores. Em outros termos, a *accuracy* permite verificar a proporção de previsões corretas de benefícios cessados por irregularidade e benefícios regulares das previsões possíveis, o que pode ser representado pela fórmula abaixo:

$$\textit{Accuracy} = \frac{verdedeiros\ positivos + verdadeiros\ negativos}{\text{verdadeiros positivos} + \text{falsos positivos} + \text{falsos negativos} + \text{verdadeiros negativos}} = =$$

$$\frac{predições\ corretas}{total\ de\ predições}$$

Precision, por sua vez, é definido por Powers (2011) como a proporção de casos previstos como positivos pelo modelo que de fato eram positivos, isto é, por intermédio

desta medida é possível dizer a proporção dos benefícios previstos pelo modelo como irregulares ou fraudados que de fato haviam sido cessados por este motivo. A fórmula para calcular o *precision* é a seguinte:

$$\textit{Precision} = \frac{Verdadeiros\,positivos}{Verdadeiros\,positivos + falsos\,positivos}$$

No que tange ao conceito de *recall*, Powers (2011) o define como a proporção de positivos identificados corretamente. De outra forma, por meio do *recall* é possível aferir o quão bom é o modelo ao prever quais benefícios foram cessados em virtude de identificação de fraude ou irregularidade. A fórmula para calcular o *recall* é a seguinte:

$$\textit{Recall} = \frac{verdadeiro\,positivo}{verdadeiro\,positivo + falso\,negativo}$$

Considerando os conceitos brevemente abordados acima e a necessidade de se verificar que o modelo desenvolvido tenha uma razoável assertividade. Tendo em vista o custo de revisão de cada benefício e que *accuracy* e *recall* são medidas importantes para avaliar um modelo de classificação, verifica-se que o conceito de *precision* mostra-se como o mais relevante para a execução deste trabalho.

Em face do exposto, considerou-se que o alcance de pelo menos 0,9 de *precision* confere desempenho aceitável para o modelo obter na previsão de benefícios com maior probabilidade de conter irregularidade ou fraude de forma a considerá-lo como bem sucedido.

Para cumprir as etapas necessárias para o atendimento dos objetivos específicos deste trabalho, ele foi realizado seguindo a metodologia de mineração de dados *Cross Industry Standard Process for Data Mining* (CRISP-DM). Chapman *et al* (2000) conceituam essa metodologia como um modelo de processo hierárquico, contemplando quatro níveis de abstração (do geral ao específico): fase, tarefa genérica, tarefa especializada e instância do processo.

Além disso, segundo Chapman *et al* (2000) há seis fases no modelo CRISP-DM: compreensão do negócio, compreensão dos dados, preparação dos dados, modelagem, avaliação e distribuição. Por fim, os autores destacam que que não existe uma seqüência rigorosa entre essas fases, sendo quase sempre necessário voltar e seguir em frente entre diferentes fases.

4. Compreensão dos Dados

Neste capítulo, são tratadas as etapas da coleta de dados iniciais, da descrição dos dados, da exploração dos dados, analisando sua descrição e utilizando técnicas de visualização, da verificação da qualidade dos dados e, por último, da apresentação do problema a ser tratado, da motivação para a realização deste trabalho e do escopo utilizado.

No que diz respeito à coleta inicial de dados, para a execução deste trabalho, foram utilizadas duas bases de dados, quais sejam, Concessão_2[6], Maciça[7], além de tabelas auxiliares que permitem identificar os códigos utilizados para determinados campos.

Segundo GRUS (2019), este é o momento em que as perguntas para as quais se busca resposta e os dados disponíveis podem tentar o cientista de dados a começar imediatamente a construir modelos e obter respostas. Todavia, GRUS (2019), reforça que é necessário resistir a esse desejo e iniciar pelo primeiro passo, que é explorar os dados.

Seguindo essas orientações, começamos pela descrição da Maciça, que é a base de dados que coleciona os pagamentos ordinários mensais realizados pelo INSS de benefícios previdenciários e assistenciais de prestação continuada.

Além disso, é importante ressaltar que, quando este trabalho foi executado, os dados da Maciça disponíveis abarcavam o período de janeiro de 2014 até dezembro de 2019.

Avançando para a análise da descrição dos dados da Maciça e para o uso de técnicas de visualização, a base Maciça contém 148 atributos e 2.408.312.164 registros, conforme demonstrado na tabela below:

Tabela 3: Quantidade de benefícios na Maciça agrupados por mês de pagamento

Ano/Mês	Jan	Fev	Mar	Abr	Mai	Jun	Jul	Ago	Set	Out	Nov	Dez	Total
2014	31,0	31,0	31,1	31,3	31,3	30,2	31,2	31,6	31,7	31,9	31,9	32,0	376,0

[6] Base de dados [BD_BENEFICIOS_HIST].[dbo].[CONCESSAO_2] do LabContas
[7] Base de dados [BD_BENEFICIOS_HIST].[dbo].[MACIÇA] do LabContas

2015	32,0	32,0	32,1	32,2	32,3	32,4	32,5	32,5	32,4	32,3	32,4	32,5	387,6
2016	32,6	32,7	32,9	33,0	33,1	33,2	33,3	33,4	34,9	35,0	33,5	33,6	401,2
2017	33,6	33,6	33,7	33,7	33,8	33,9	33,9	34,0	34,1	34,1	34,3	34,3	407,1
2018	34,3	34,4	34,4	34,5	34,5	34,6	34,6	34,6	34,7	34,8	34,9	34,9	415,1
2019	34,9	34,8	34,9	35,0	35,0	35,0	35,1	35,1	35,2	35,3	35,4	35,5	421,2
Total													2.408,3

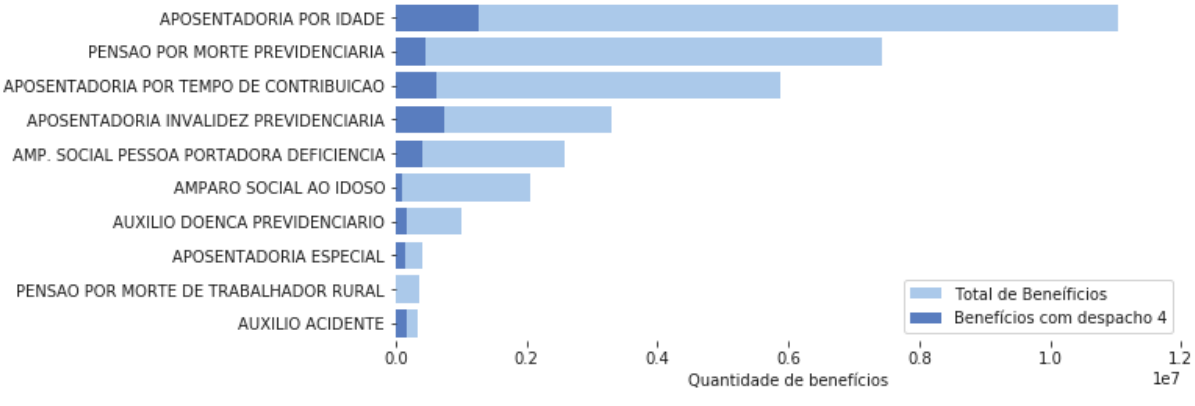
Fonte: Elaboração pelo autor com dados da Maciça

Nota: Quantidade em milhões de registro

Com relação aos atributos, por meio meio deles é possível identificar informações relativas a cada benefício, como por exemplo sua espécie, a Agência da Previdência Social (APS) responsável pela concessão do benefício, a matrícula do servidor responsável pela concessão, as datas de requerimento do benefício perante o INSS e de concessão do benefício, além de outras informações conforme descrição dos demais atributos registrados no Apêndice A.

Para selecionar a espécie de benefício a ser utilizada neste trabalho, foi analisada a distribuição dos 35,5 milhões de benefícios ativos em dezembro de 2019. Nessa análise, foi identificado que eles estavam divididos em 63 espécies distintas. As dez espécies mais representativas foram demonstradas por meio gráfico below:

Gráfico 3: Distribuição dos benefícios da Maciça de dezembro de 2019 por espécie



Fonte: Elaboração pelo autor com dados da Maciça

No que concerne à distribuição dos benefícios presentes na Maciça de dezembro de 2019, observa-se que 31% eram da espécie aposentadoria por idade, isto é, 11.025.829 de registros, dos quais 1.272.924 (11,5%) foram benefícios concedidos por força de decisão judicial.

A segunda espécie com mais registros é a pensão por morte previdenciária com 7.420.592 benefícios, sendo 463.800 (6,3%) concedidos em cumprimento à decisão judicial (despacho 4).

Por último, vale destacar a espécie identificada na Maciça como amparo social ao idoso, a qual é chamada de BPC idoso neste trabalho. Embora essa espécie seja apenas a sexta com maior quantidade de benefícios, com 2.057.255 registros, ela foi a selecionada para a execução deste trabalho uma vez que ela tem maior quantidade registros cessados por motivo de identificação de fraude.

Passando para a descrição dos dados da base Concessão_2, diferente do que se pode inferir pelo seu nome, ela contém o registro dos benefícios previdenciário e assistenciais cessados. Por isso, a fim de permitir maior fluidez na descrição das etapas cumpridas neste trabalho e evitar incompreensões decorrentes do nome desta base no LabContas, passamos a partir deste ponto a chamá-la de Cessados.

Antes de seguir para a análise da descrição da base Cessados e para o uso de técnicas de visualização, vale ressaltar que, quando este trabalho foi executado, os dados disponíveis abarcavam os benefícios cessado até dezembro de 2019.

Feitas as considerações acima, na base de benefícios cessados, são colecionados os benefícios previdenciários e assistenciais cujo pagamento foi descontinuado. Ela contém 9 atributos e 5,9 trilhões de registros, considerando o período selecionado para execução deste trabalho, qual seja, de janeiro de 2014 até dezembro de 2019, conforme demonstrado na tabela below.

Tabela 4: Quantidade de benefícios na base Cessados agrupados por mês de pagamento

Ano/Mês	Jan	Fev	Mar	Abr	Mai	Jun	Jul	Ago	Set	Out	Nov	Dez	Total
2014	67,9	68,3	68,7	69,1	69,6	70,0	70,4	70,9	71,4	71,9	26,8	72,8	797,8
2015	73,2	73,5	73,9	74,4	74,8	75,2	75,7	76,0	76,2	76,5	76,8	77,2	903,5
2016	77,5	77,9	78,3	78,9	79,4	79,9	80,4	80,8	81,2	81,7	82,0	82,5	960,4
2017	82,8	83,2	83,6	84,1	84,4	84,9	85,4	85,8	86,3	86,7	87,2	87,6	1.022,0
2018	88,0	88,4	88,7	89,2	89,7	90,2	90,6	91,0	91,5	91,9	92,4	92,8	1.084,4
2019	93,2	93,5	93,9	94,3	94,7	95,2	95,6	96,1	96,6	97,0	97,6	98,0	1.145,5
Total													5.913,7

Fonte: Elaboração pelo autor com dados da base Cessados
Nota: Quantidade em milhões de registro

No que concerne aos 5,9 trilhões de registros da base Cessados, é importante registrar que esse valor não representa adequadamente a quantidade de benefícios cessados até dezembro de 2019.

Diferente da Maciça, na atualização mensal da base de cessados, são informados todos os benefícios cessados até aquele mês. Portanto, no mês de fevereiro de 2014, foram cessados 380.271 benefícios e não os 68,3 milhões de registros atribuídos na base Cessados.

Diante do exposto, a quantidade de benefícios cessados até dezembro de 2019 é de 98,0 milhões de registros.

No que diz respeito aos 9 atributos da base Cessados, há quatro deles (DT_ATUALIZACAO_ETL, DS_ERRO, NM_ARQUIVO e ANO_MES_REF) que registram informações que dizem respeito à atualização da base e, portanto, não foram considerados na execução deste trabalho.

Com isso, por meio do quadro below é realizada a descrição dos cinco atributos da base de dados Cessados utilizados na execução deste trabalho:

Quadro 3: Atributos da base Cessados

Atributo	Descrição
NU_NB	Número do benefício
D2_DCB	Data da cessação do benefício
CS_SITUACAO_BENEF	Código que identifica situação do seu benefício quanto a continuidade do pagamento
CS_MOTIVO	Código que identifica o motivo para a cessação do benefício
ID_OL_MANUTENCAO	Identificação do APS mantenedora do benefício

Fonte: Elaboração pelo autor

Ao analisar a base Cessados, constatou-se a distribuição dos 98,0 milhões de registros de benefícios previdenciários e assistenciais cessados em 99 motivos diferentes para cessação. Entre os motivos para cessação há a não comprovação de fé de vida, a cessação em virtude de concessão de benefício de outra espécie e a alta médica que é aplicável ao auxílio doença.

Porém, como este trabalho se propõe a identificar benefícios previdenciários ou assistenciais com maior probabilidade de conter irregularidade, entre aqueles 99 motivos para cessação foram selecionados os sete motivos relacionados à identificação de fraudes ou irregularidades, conforme disposto no Quadro 4. Dessa forma, a quantidade de benefícios a serem utilizados neste trabalho reduziu de 97.9 milhões para 132.583 registros.

Quadro 4: Motivos de cessação de benefício relacionados à fraude ou irregularidade

CS_MOTIVO	Descrição
30	Constatação de fraude
31	Constatação de irregularidade ou erro administrativo
53	Fraude informada pela auditoria
55	Irregularidade médico-pericial quando na revisão médico-pericial realizada for constatada a ausência de elementos médicos-periciais que confirmem a concessão dos benefícios

74	Cancelamento por fraude identificada pela auditoria
75	Cancelamento por erro administrativo identificado pela auditoria
77	Manutenção irregular identificado pela perícia médica ou auditoria

Fonte: [BD_BENEFICIOS_HIST].[dbo].[COD_MOTIVO_CONCESSAO2] do Labcontas com adaptações

Considerando que havia 95 espécies de benefícios previdenciários e assistenciais diferentes na Maciça entre janeiro de 2014 e dezembro de 2019 e que essas espécies possuem diferentes objetivos, público-alvo e regras de elegibilidade, foi necessário escolher a espécie de benefício previdenciário ou assistencial com a maior quantidade de cessações pelos motivos listados no Quadro 4.

Para isso, utilizou-se o campo NU_NB como chave-primária para incorporar os demais campos constantes na Maciça aos 132.583 benefícios selecionados da base Cessados.

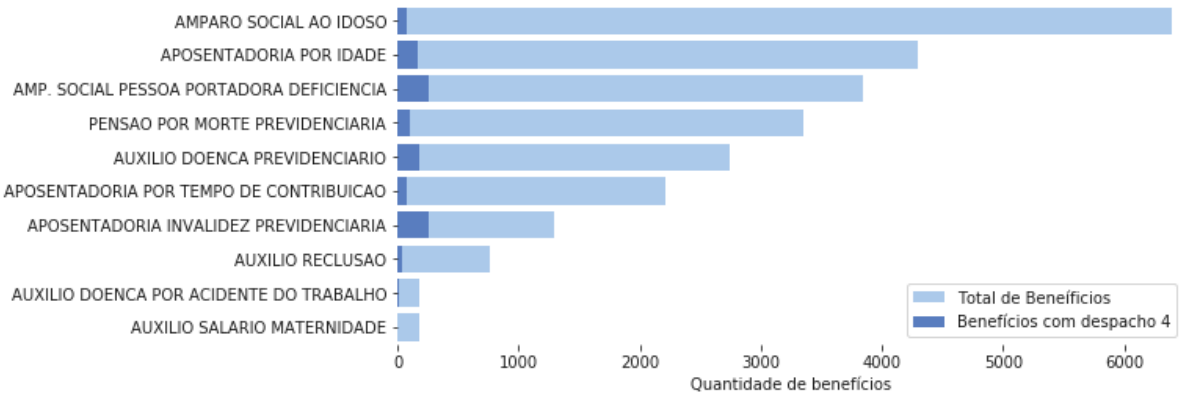
Todavia, dos 132.583 da base Cessados apenas 25.753 foram identificados na Maciça. Isso já era esperado visto que havia 113.179 benefícios na base de Cessados cuja a data de cessação do benefício (D2_DCB) era anterior a janeiro de 2014.

Cabe registrar que a data de cessação, assim como a data de início de benefício (D2_DIB) estão relacionadas ao pagamento do benefício sob o regime de competência e não o de caixa. Em outras palavras, as datas de início de benefício (D2_DIB) e de cessação do benefício (D2_DCB) dizem respeito, respectivamente, à data de início e à data de fim do período no qual o cidadão teve direito ao benefício.

Por outro lado, os campos de data de despacho (D2_DDB) e data de início de pagamento (D2_DIP) dizem respeito, respectivamente, à data na qual o benefício foi concedido pelo INSS, isto é, a data de deferimento do benefício; e à data na qual o benefício foi incluído na folha de pagamento.

Por meio do gráfico below, são apresentadas as 10 espécies de benefícios que concentraram as maiores quantidades daqueles 25.753 benefícios que foram identificados na Maciça.

Gráfico 4: Distribuição dos benefícios cessados por espécie



Fonte: Elaboração pelo autor

Diante do resultado demonstrado por intermédio do gráfico above, foi selecionado o benefício assistencial Amparo Social ao Idoso, uma vez que é a espécie que possui a maior quantidade de benefícios cessados (6.399 registros) em decorrência de identificação de fraude ou irregularidade.

Os benefícios com despacho 4 são aqueles cuja concessão foi realizada em cumprimento à decisão judicial. Como as decisões judiciais podem alterar os critérios para elegibilidade aos benefícios para o caso concreto, conforme consta no Relatório do Levantamento da Previdência Social em 2019, elaborado pelo TCU (TC 009.811/2019-2), e como tratava-se de 78 registros, optou-se por desconsiderá-los.

Cabe destacar que o INSS não identifica os benefícios concedidos em decorrência da vigência das ACPs com despacho 4, tampouco a Autarquia criou uma identificação que permita distinguir esses benefícios. Desse modo, não foi possível retirar da base os benefícios concedidos em virtude da alteração de regras de elegibilidade promovida pelas ACPs.

Portanto, ao final dos procedimentos acima citados, a base de benefícios cessados em razão de identificação de fraude colecionava 6.320 registros.

Feita a escolha da espécie de benefício a ser utilizada neste trabalho, Amparo Social ao Idoso (BPC idoso), e definido que seriam descartados os benefícios concedidos em cumprimento de decisão judicial (despacho 4), revisitou-se a base Maciça para selecionar um conjunto de benefícios com essas características, para os quais não tivesse sido constatada fraude ou irregularidade.

Em que pese o ato de concessão do benefício gozar da presunção de legitimidade, uma vez que se trata de um ato administrativo, essa presunção é relativa, isto é, admite prova em contrário (DI PIETRO, 2019).

Em face do exposto, para fins deste trabalho, não é possível presumir que os benefícios selecionados por mio de uma amostra aleatória da Maciça estejam livres de fraudes ou irregularidades que no futuro possam ser identificadas.

Como forma de mitigar o relatado acima, foram considerados como benefícios não fraudados aqueles que foram concedidos antes do ano de 2015 (ano da DDB menor que 2015) e que estavam ativos até dezembro de 2019. Assim, seriam selecionados benefícios que passaram por quatro Ciclos de Fiscalização Contínua de Benefícios (**Error! Reference source not found.**) sem que tivessem sido identificados como fraudados ou irregulares.

5. Preparação dos Dados

O sucesso de um projeto de aprendizado de máquina para a solução de uma tarefa é significativamente influenciado pela qualidade dos dados selecionados e sua demonstração (SALEEM, 2014).

De acordo com Beniwal e Arora (2012), após a seleção dos dados, é necessário levar em conta que eles podem estar em diferentes formatos, pois são obtidos por meio de diferentes fontes, além de conterem problemas como ruídos, atributos irrelevantes e dados ausentes o que, de acordo com Saleem (2011), pode levar a resultados incorretos.

Diante disso, Saleem (2014) destaca que o pré-processamento de todos os conjuntos de dados que serão utilizados é um fator fundamental na aplicação da mineração de dados para a solução de problemas.

Chapman (2000) afirma que a fase de preparação de dados abrange todas as atividades necessárias para construir o conjunto de dados final, que são os dados a serem utilizado como insumo do modelo a partir dos dados brutos iniciais. Além disso, segundo Chapman (2000), é provável que as tarefas de preparação de dados sejam executadas várias vezes e sem uma ordem pré-determinada.

Diante disso, é possível compreender o motivo de Saleem *et al* (2014) afirmar que o saneamento dos dados, o que inclui sua preparação e filtragem, pode consumir até 80% do tempo no pré-processamento de dados em qualquer projeto de mineração de dados do mundo real.

Chapman *et al* (2000) ensina que a preparação dos dados é composta por tarefas como selecionar, limpar, construir, integrar e formatar os dados.

Com relação à seleção de dados, Chapman *et al* (2000) registra que, nesta etapa, são tomadas decisões a respeito dos dados a serem usados para análise. Os autores acrescentam que os critérios a serem utilizados devem levar em consideração a relevância para as metas da mineração de dados, a qualidade, e as restrições técnicas, como limites de volume ou tipos de dados.

Por último, Chapman *et al* (2000) destaca que a seleção de dados abrange a seleção de atributos (colunas), bem como a seleção de registros (linhas) em uma tabela. Além disso, Beniwal e Arora (2012) ressaltam a importância da seleção de atributos, uma vez que nem todos são relevantes. Diante disso, Beniwal e Arora (2012)

acrescentam que é essencial eleger um subconjunto de atributos relevantes para a mineração entre todos os atributos originais.

Após superada a etapa de compreensão dos dados descrita no capítulo above, a seleção de dados deste trabalho teve início com um conjunto de dados inicial que colecionava 149 atributos e 12.640 registros.

Cumpre registrar que, entre aqueles 149 atributos, havia um atributo, chamado ‘target’, que foi criado para distinguir os benefícios fraudados (target = 1) dos não fraudados (target = 0). Por isso, havia mais do que os 148 atributos da base da Maciça, cuja descrição consta no apêndice A deste trabalho.

No que tange aos 12.640 registros que compunham o conjunto de dados inicial, 6.320 deles se tratavam de benefícios cessados em razão de identificação de fraude e os outros 6.320 correspondiam a benefícios não fraudados selecionados conforme descrito no capítulo above.

Ao analisar a qualidade dos dados com relação à quantidade de registro nulos, foi identificado que os 32 atributos constantes na tabela below possuíam mais de 90 % dos registros nulos.

Tabela 5: Atributos em que mais de 90% dos registros são nulos

Atributos	Benefícios irregulares			Benefícios regulares			Tipo de atributo
	Quant. de registros nulos	% Nulos	Quant. total de registros	Quant. de registros nulos	% Nulos	Quant. total de registros	
D2_INI_INCAPAC	6320	100,0	6320	6320	100,0	6320	A1
D2_INICIO_DOENCA	6320	100,0	6320	6320	100,0	6320	A1
D2_OBITO_RECLUSAO	6320	100,0	6320	6320	100,0	6320	A1
CS_DIAGNOSTICO_N	6320	100,0	6320	6319	100,0	6320	A1
CS_DIAGNOSTICO_1	6320	100,0	6320	6319	100,0	6320	A1
D2_OBITO_T	6320	100,0	6320	6320	100,0	6320	A5
NM_INSTITUIDOR_I	6320	100,0	6320	6320	100,0	6320	A2
NM_MAE_I	6320	100,0	6320	6320	100,0	6320	A2
DT_NASCIMENTO_I	6320	100,0	6320	6320	100,0	6320	A2
CTPS_UF_I	6320	100,0	6320	6320	100,0	6320	A2
NU_IDENTIDADE_I	6320	100,0	6320	6320	100,0	6320	A2
IDENTIDADE_UF_I	6320	100,0	6320	6320	100,0	6320	A2
D2_OBITO_I	6320	100,0	6320	6320	100,0	6320	A2
NM_PROCURADOR_P	6254	99,0	6320	6206	98,2	6320	A3
NM_MAE_P	6254	99,0	6320	6206	98,2	6320	A3
DT_NASCIMENTO_P	6254	99,0	6320	6206	98,2	6320	A3
CTPS_UF_P	6281	99,4	6320	6246	98,8	6320	A3
NU_IDENTIDADE_P	6254	99,0	6320	6206	98,2	6320	A3
IDENTIDADE_UF_P	6254	99,0	6320	6206	98,2	6320	A3

NM_BAIRRO_P	6254	99,0	6320	6207	98,2	6320	A3
NM_MUNICIPIO_P	6255	99,0	6320	6206	98,2	6320	A3
NM_UF_MUNICIPIO_P	6254	99,0	6320	6206	98,2	6320	A3
NM_REPRESENTANTE_R	6314	99,9	6320	6285	99,4	6320	A4
NM_MAE_R	6314	99,9	6320	6285	99,4	6320	A4
DT_NASCIMENTO_R	6314	99,9	6320	6285	99,4	6320	A4
CTPS_UF_R	6316	99,9	6320	6299	99,7	6320	A4
NU_IDENTIDADE_R	6314	99,9	6320	6285	99,4	6320	A4
IDENTIDADE_UF_R	6315	99,9	6320	6288	99,5	6320	A4
DT_ULTIMA_ALTER	6320	100,0	6320	6320	100,0	6320	A1
D2_LIMITE	6320	100,0	6320	6320	100,0	6320	A1
DS_ERRO	6317	100,0	6320	6317	100,0	6320	A6
DT_ULTIMA_PERICIA	6320	100,0	6320	6319	100,0	6320	A1

Fonte: Elaboração pelo autor

Com relação ao fato de 21% dos atributos (32 de 149) terem sido listados na tabela above, isso decorre do fato de o conjunto de dados inicial conter todos os atributos da Maciça. Vale lembrar que a Maciça contempla os pagamentos de outras espécies de benefícios e, por isso, nem todos os atributos são utilizados para registrar informações relacionadas ao pagamento do BPC idoso.

Os atributos listados na tabela above são utilizados para registrar informações relativas, por exemplo, a outras espécies de benefícios, ao instituidor da pensão, ao procurador e ao representante legal. A relação dos códigos da coluna Tipo de atributo e seu uso é apresentada por meio do quadro below.

Quadro 5: Códigos utilizados na coluna Tipo de atributo da Tabela 5

Tipo de atributo	Uso do atributo
A1	Registro de informações relativas à outras espécies de benefícios
A2	Registro de informações relativas ao instituidor da pensão
A3	Registro de informações relativas ao procurador do beneficiário
A4	Registro de informações relativas ao representante legal do beneficiário
A5	Data de óbito do titular do benefício
A6	Registro de informação relativa à carga da base no LabContas

Fonte: Elaboração pelo autor

Conforme abordado no capítulo que trata da compreensão do negócio, o BPC idoso não é um benefício relacionado à incapacidade laboral. Além disso, esse benefício não gera pensão, tampouco é necessário constituir procurador ou representante legal para recebê-lo.

Ademais, segundo exposto no capítulo que trata da compreensão do negócio, o INSS utiliza o SIRC para obter informações relativas à data de óbito dos beneficiários e

tomar as medidas cabíveis. Portanto, não faz sentido manter o atributo D2_OBITO_T no conjunto de dados.

Em face do exposto, entendeu-se que era aplicável a utilização de técnicas para atribuir valores aos atributos listados na Tabela 5. Portanto, aqueles 32 atributos foram excluídos do conjunto de dados que passou a conter 117 atributos.

Após analisar os registros nulos, foi verificado se o conjunto de dados continha atributos cujo valor não variava e, nesta etapa, foram identificados os 17 atributos da tabela below:

Tabela 6: Atributos com todos os registros iguais

Atributos	Benefícios irregulares		Benefícios regulares	
	Quant. de registros distintos	Quant. total de registros	Quant. de registros distintos	Quant. total de registros
CS_TRATAMENTO	1	6320	1	6320
CS_ESPECIE	1	6320	1	6320
NU_NB_ANT	1	6320	1	6320
NU_MATR_MP2	1	6320	1	6320
NU_CPF_I	1	6320	1	6320
ID_NIT_I	1	6320	1	6320
CTPS_I	1	6320	1	6320
CTPS_SERIE_I	1	6320	1	6320
NU_TIT_ELEITOR_I	1	6320	1	6320
CS_VAL_CNIS_I	1	6320	1	6320
CS_SEXO_I	1	6320	1	6320
QT_DEP_VAL_NB	1	6320	1	6320
QT_DEP_CADASTRO	1	6320	1	6320
CS_RUBRICA_1	1	6320	1	6320
CS_RUBRICA_10	1	6320	1	6320
VL_RUBRICA_10	1	6320	1	6320
FASE_ULTIMA_PERICIA	1	6320	1	6320

Fonte: Elaboração pelo autor

Primeiramente, cumpre destacar que os atributos CS_RUBRICA_1, CS_RUBRICA_10 e VL_RUBRICA_10, listados na tabela above, foram mantidos no conjunto de dados por estarem relacionados às rubricas e aos valores que fazem parte do contracheque do beneficiário. Diante disso, esses atributos foram utilizados para criar atributos conforme descrito no próximo capítulo que aborda a elaboração do modelo preditivo.

Já os atributos CS_TRATAMENTO e CS_ESPECIE possuem todos os valores iguais em razão de todos os benefícios serem da mesma espécie (BPC Idoso).

Com relação ao atributo FASE_ULTIMA_PERICIA, embora conste na tabela above 125 registros com valores iguais, todos os demais 6196 registros são NaN (*Not a Number*).

Quanto aos demais atributos listados na tabela above, os demais valores estão preenchidos com o valor 0, assim como consta na Maciça, em razão de essa não ser a regra para eles quando seus valores são nulos.

Considerando que o objetivo do modelo desenvolvido neste trabalho é identificar os benefícios da espécie BPC idoso com maior probabilidade de conter irregularidade, foi identificada a necessidade de excluir o atributo CS_SITUACAO_BENEF.

Isso se deve ao fato de, além desse atributo identificar quais benefícios estão ativos e quais foram cessados, ele não ser passível de utilização em uma etapa de validação ou até mesmo em produção. Por meio da tabela below é apresentada a distribuição de valores entre benefícios regulares e irregulares.

Tabela 7: Valores do atributo CS_SITUACAO_BENEF no conjunto de dados

Valor	Quant. de registro benefícios irregulares	Quant. de registro benefícios regulares
Ativo	0	6.320
Cessado	6.316	0
Cessado pela Auditoria	4	0
Total	**6.320**	**6.320**

Fonte: Elaboração pelo autor

Portanto, com exceção dos campos CS_RUBRICA_1, CS_RUBRICA_10 e VL_RUBRICA_10, todos os outros demais 14 atributos foram excluídos do conjunto de dados, que passou a conter 103 atributos.

A etapa seguinte no modelo CRISP-DM é a de limpeza dos dados e Chapman *et al* (2000) ensina que nesta etapa deve-se buscar aumentar a qualidade dos dados para o nível exigido pelas técnicas de análise selecionadas. Beniwal e Arora (2012), por sua vez, afirmam que a limpeza de dados abrange, por exemplo, a detecção e correção de erros nos dados, além do preenchimento de valores faltantes.

Isto posto, a limpeza dos dados neste trabalho teve início com a identificação dos atributos restantes que possuíam registros nulos. O resultado obtido é apresentado na tabela below:

Tabela 8: Atributos em com pelo menos 1 registo nulo

Atributos	Benefícios irregulares	Benefícios regulares

	Quant. de registros nulos	% Nulos	Quant. total de registros	Quant. de registros nulos	% Nulos	Quant. total de registros
D2_DCB	0	0,0	6320	6320	100,0	6320
NU_CONTA_CORRENTE	3668	58,0	6320	0	0,0	6320
D2_FORMAT_CONC	3016	47,7	6320	3727	59,0	6320
CTPS_UF_T	3465	54,8	6320	2395	37,9	6320
NU_IDENTIDADE_T	473	7,5	6320	187	3,0	6320
IDENTIDADE_UF_T	500	7,9	6320	261	4,1	6320
NM_BAIRRO_T	30	0,5	6320	27	0,4	6320
NU_DDD_T	4685	74,1	6320	4468	70,7	6320
NU_TELEFONE_T	4560	72,2	6320	4339	68,7	6320
NM_MUNICIPIO_T	1	0,0	6320	3	0,0	6320
NM_UF_MUNICIPIO_T	0	0,0	6320	2	0,0	6320

Fonte: Elaboração pelo autor

O campo D2_DCB significa data de cessação de benefício. Essa data é fixada sob a ótica do regime de competência da contabilidade, portanto, esse atributo registra a data até a qual o beneficiário tem direito ao benefício.

No caso do benefício BPC idoso, não há previsão legal para fixação dessa data no momento de sua concessão. Por isso, verifica-se que não há registros nulos para o atributo D2_DCB quando o benefício foi cessado por irregularidade. Por outro lado, todos os registros desse atributo são nulos quando se trata de benefício regular. Com isso, optou-se pela exclusão do atributo D2_DCB.

O atributo NU_CONTA_CORRENTE, por sua vez, indica o número da conta corrente na qual é depositado o valor do benefício. Ao explorar este atributo verificou-se que ele segue a distribuição apresentada por meio da tabela below:

Tabela 9: Distribuição de valores para o atributo NU_CONTA_CORRENTE

Valor	Benefícios irregulares		Benefícios regulares	
	Pagamentos por cartão magnético	Depósitos em conta corrente	Pagamentos por cartão magnético	Depósitos em conta corrente
Nulo	3.668	0	0	0
0000000000	618	0	3.459	0
Outro valor	0	2.034	0	2.861
Subtotal	**4.286**	**2.034**	**3.459**	**2.861**
Total	**6.320**		**6.320**	

Fonte: Elaboração pelo autor

Diante da distribuição dos dados observada, optou-se por transformar a variável NU_CONTA_CORRENTE em categórica com as 3 categorias que constam na tabela above, quais sejam ‘Nulo’, '0000000000' e ‘Outro valor’.

Quanto ao atributo D2_FORMAT_CONC, não foi possível atribuir um valor com base nas outras variáveis do tipo data. Assim, a partir desse atributo foi criada uma variável binária que informa se o campo atributo D2_FORMAT_CONC foi preenchido.

No que concerne aos outros atributos listados na Tabela 8, eles registram informações de cadastro do beneficiário, conforme pode ser verificado realizando cotejamento com as informações tratadas no Apêndice A. Dessa forma, foi criada uma variável binária para identificar quando cada um desses atributos foi preenchido.

Avançando para a etapa de construção de dados, a qual, segundo Chapman *et al* (2000), inclui operações construtivas de preparação de dados, como a criação de atributos derivados da combinação de um ou mais atributos existentes.

Nesse sentido, foram criadas variáveis que exploram a interação dos atributos que se referem às rubricas constantes na folha de pagamento dos beneficiários, ao valor bruto (‘VL_BRUTO’), ao valor líquido (‘VL_LIQUIDO’) e ao total de descontos ('TOT_DESCONTOS') dos benefícios, bem como às datas relacionadas às etapas do processo de concessão do benefício[8].

Cada um dos campos 10 atributos 'CS_RUBRICA'[9] e seus respectivos valores[10] correspondem, respectivamente, à cada linha constante no contracheque do beneficiário.

No caso, por exemplo, de um beneficiário que contrair dois empréstimos consignados, cada um deles fica registrado em uma linha diferente do contracheque. Diante disso, optou-se por consolidar o somatório de todos os empréstimos consignados do beneficiário em um único atributo.

Dessa forma, foram criados atributos para registrar o somatório dos valores pagos e descontados em cada rubrica constante no contracheque do beneficiário.

---

[8] Os atributos que registram as datas relacionadas às etapas do processo de concessão do benefício são: ‘D2_DER’, ‘D2_DRD’, ’D2_DDB’, ‘D2_DIB’ e ’D2_DIP’

[9] Os atributos que contém os códigos das rubricas da folha de pagamento são: ‘CS_RUBRICA_1', ‘CS_RUBRICA_2', ‘CS_RUBRICA_3', ‘CS_RUBRICA_4', ‘CS_RUBRICA_5', ‘CS_RUBRICA_6', ‘CS_RUBRICA_7', ‘CS_RUBRICA_8', ‘CS_RUBRICA_9' e ‘CS_RUBRICA_10'.

[10] Os atributos que contém os valores das rubricas são: 'VL_RUBRICA_1', 'VL_RUBRICA_2', 'VL_RUBRICA_3', 'VL_RUBRICA_4', 'VL_RUBRICA_5', 'VL_RUBRICA_6', 'VL_RUBRICA_7', 'VL_RUBRICA_8', 'VL_RUBRICA_9' e 'VL_RUBRICA_10'.

Isso resultou na criação de 25 atributos, de forma que cada um deles registrava o somatório dos valores pagos em cada rubrica naquele mês. Como exemplo deste processo, vamos tratar da rubrica 203 que corresponde às consignações do beneficiário, como, por exemplo, o empréstimo consignado.

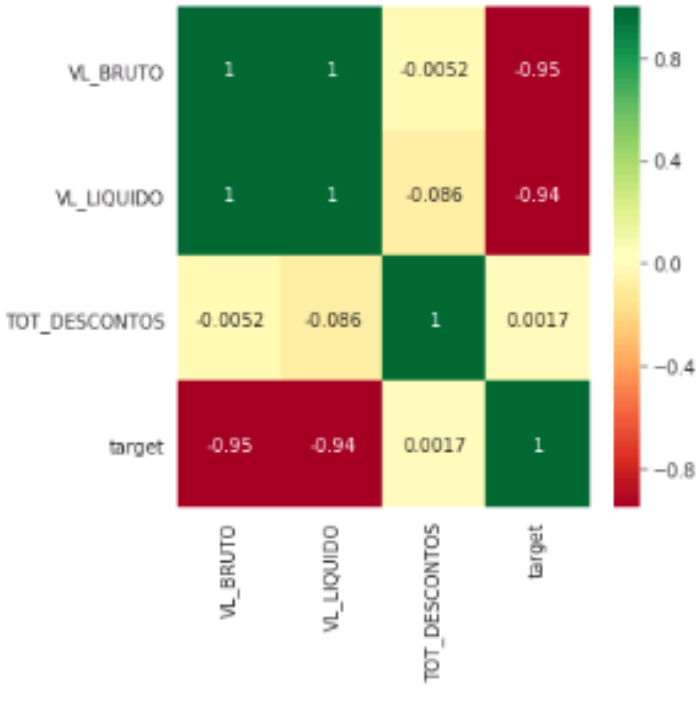
Cada um dos empréstimos consignados do beneficiário é apresentado na Maciça como um atributo “CS_RUBRICA” e seu respectivo “VL_RUBRICA”. Para agrupá-las foi criado um atributo “SUM_VL_RUBRICA_203” para consolidar o seu valor em um único atributo. Esse processo foi repetido para as outras 24 rubricas encontradas.

No que diz respeito à criação de variável que demonstrasse a interação entre os atributos valores brutos, valores líquidos e o total de descontos, isso foi feito calculando a proporção do benefício que era efetivamente recebido pelo beneficiário. Esse cálculo foi feito por meio da fórmula abaixo:

$$'PER\_VL\_LIQUIDO\_VL\_BRUTO' = \frac{'\text{VL\_LIQUIDO}'}{'\text{VL\_BRUTO}'}$$

Neste ponto, é importante destacar que os campos valor bruto (‘VL_BRUTO’) e valor líquido (‘VL_LIQUIDO’) possuem alta correlação negativa com o atributo que identifica os benefícios fraudados e não fraudados (*target*), como demonstram o Gráfico 5 e o Gráfico 6.

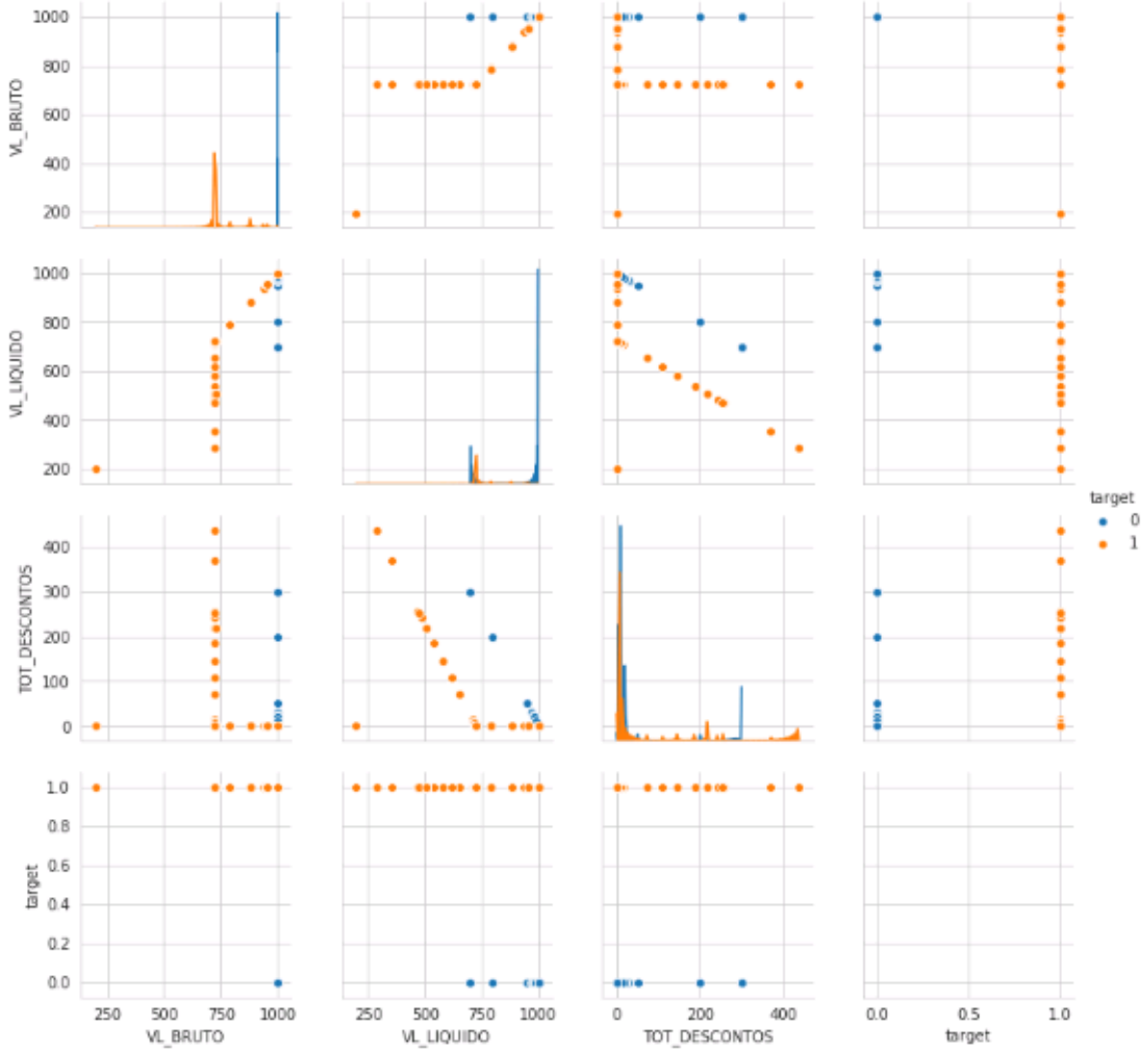
Gráfico 5: Mapa de calor da correlação entre os atributos valor bruto, valor líquido, o total de descontos e ‘target’



Fonte: Elaboração pelo autor

Além disso, foi constado que os atributos valor bruto (‘VL_BRUTO’), valor líquido (‘VL_LIQUIDO’) e o total de descontos ('TOT_DESCONTOS') indicavam indiretamente quais eram os benefícios fraudados e os não fraudados, como demonstrado por meio da Gráfico 6.

Gráfico 6: Relações entre os atributos valor bruto, valor líquido e o total de descontos fazendo a distinguindo os benefícios fraudados (*target* = 1) dos não fraudados (*target* = 0)



Fonte: Elaboração pelo autor

Considerando a relação entre valores dos atributos VL_BRUTO, VL_LIQUIDO, TOT_DESCONTOS, optou-se por realizar uma breve exposição tratando apenas sob à ótica do atributo VL_BRUTO para elucidar o motivo desses três atributos identificarem indiretamente quais eram os benefícios fraudados e os não fraudados.

O campo VL_BRUTO é o valor do benefício sem considerar os descontos. Conforme definido pelo do art. 20 da LOAS, o valor do BPC idoso é igual a um salário

mínimo, que, por sua vez, foi alterado entre janeiro de 2014 e dezembro de 2019 como apresentado por meio do quadro abaixo.

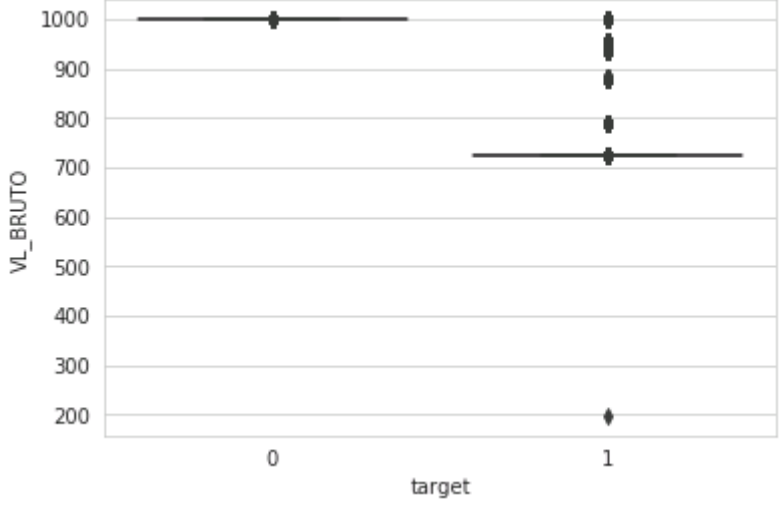
Quadro 6: Valor do salário mínimo de janeiro de 2014 até dezembro de 2019

Início da vigência	Dispositivo legal	Valor (em R$)
01/01/2014	Decreto nº 8.166, de 2013	724,00
01/01/2015	Decreto nº 8.381, de 2014	788,00
01/01/2016	Decreto nº 8.618, de 2015	880,00
01/01/2017	Decreto nº 8.948, de 2016	937,00
01/01/2018	Decreto nº 9.255, de 2017	954,00
01/01/2019	Decreto nº 9.661, de 2019	998,00

Fonte: Elaboração pelo autor

Segundo Capela e Capela (2011), o gráfico *boxplot* permite avaliar a simetria dos dados, a dispersão e a existência ou não de *outliers*, além disso, ele pode ser utilizado para a comparação de dois ou mais grupos. Com isso, foi elaborado o gráfico *boxplot* below para apresentar como é a distribuição dos valores no campo VL_BRUTO, no caso dos benefícios fraudados (target = 1) e o dos não fraudados (target = 0).

Gráfico 7: Distribuição dos valores brutos dos benefícios



Fonte: Elaboração pelo autor

Analisando o gráfico *boxplot* above, verifica-se que todos os valores do campo VL_BRUTO para os benefícios não fraudados (target = 0) são aproximadamente R$ 1.000,00. Quanto aos benefícios fraudados (target = 1), verifica-se que os valores se concentram um pouco acima de R$ 700,00.

No que concerne à distribuição dos valores dos benefícios não fraudados (target = 0), ela é resultado desses benefícios terem sido selecionados na Maciça de dezembro

de 2019 e, portanto, o valor do benefício é igual a R$ 998,00, conforme conta no Quadro 6.

A distribuição de valores dos benefícios fraudados (target = 1), por sua vez, se concentra no valor de R$ 724,00. Isso se deve ao fato de a incorporação dos atributos da Maciça na base Cessados, conforme tratado no capítulo 0, ter sido realizada selecionando a Maciça mais antiga que continha o registro do benefício fraudado e a Maciça mais antiga disponível é a de janeiro de 2014.

Considerando que a obtenção do atributo ‘VL_BRUTO’ se deu por meio da Maciça, conforme tratado no capítulo 0, a concentração dos valores dos benefícios fraudados (target = 1), em R$ 724,00, é consequência da maior parte dos benefícios cessados por fraude ter ocorrido anteriormente ao ano de 2015 e a Maciça mais antiga disponível é a de janeiro de 2014.

Cumpre destacar que a variação do atributo ‘VL_BRUTO’, acima descrita, gera efeitos semelhantes nos campos ‘VL_LIQUIDO’ e 'TOT_DESCONTOS'. Isso, na verdade, é resultado da correlação artificial entre esses atributos e a classe, causada pelo viés involuntário nos dados derivado da escolha de benefícios de diferentes períodos.

Diante dessa correlação artificial entre os atributos e a classe, foi necessária a exclusão dos atributos ‘VL_BRUTO’, ‘VL_LIQUIDO’ e 'TOT_DESCONTOS' do conjunto de dados.

Quanto à criação de variáveis capazes de captar a interação entre as datas das diferentes etapas do processo de concessão do benefício, ela era necessária para permitir a utilização desses atributos no modelo. Com isso, essas variáveis foram criadas partindo de duas perspectivas.

Mas, antes de discorrer sobre aquelas duas perspectivas, é importante realizar uma breve descrição a respeito dos atributos que registram as datas das diferentes etapas do processo de concessão do benefício.

O atributo ‘D2_DER’ registra a data em que o beneficiário realizou o requerimento junto ao INSS com o objetivo de obter a concessão do benefício.

Quanto ao atributo ‘D2_DRD’, nele está consignada a data em que foi regularizada alguma pendência de documentação não entregue pelo beneficiário quando da realização do requerimento.

Enquanto o atributo ’D2_DDB’ registra a data em que ocorreu o despacho do servidor concedendo o benefício, no atributo ‘D2_DIB’, por sua vez, está lançada a data em que teve início o direito ao benefício por parte do beneficiário

Por último, no atributo ‘D2_DIP’, está consignada a data em que o INSS efetuou o primeiro pagamento do benefício, o que não isenta a autarquia de pagar ao beneficiário o valor retroativo à data em que teve início o direito ao benefício (‘D2_DIB’).

Isso posto, na primeira abordagem utilizada para criação de variáveis, foram criados 10 campos para registrar a diferença entre cada um dos atributos que dizem respeito às datas das diferentes etapas do processo de concessão do benefício acima descrito, conforme as fórmulas abaixo.

$$DIF\_\mathrm{D2\_DIP\_D2\_DIB} = D2\_DIP\ - D2\ \_DIB$$

$$DIF\_\mathrm{D2\_DIP\_D2\_DDB} = D2\_DIP\ - D2\ \_DDB$$

$$DIF\_\mathrm{D2\_DIP\_D2\_DRD} = D2\_DIP\ - D2\ \_DRD$$

$$DIF\_\mathrm{D2\_DIP\_D2\_DER} = D2\_DIP\ - D2\ \_DER$$

$$DIF\_\mathrm{D2\_DIB\_D2\_DDB} = D2\_DIB\ - D2\ \_DDB$$

$$DIF\_\mathrm{D2\_DIB\_D2\_DRD} = D2\_DIB\ - D2\ \_DRD$$

$$DIF\_\mathrm{D2\_DIB\_D2\_DER} = D2\_DIB\ - D2\ \_DER$$

$$DIF\_\mathrm{D2\_DDB\_D2\_DRD} = D2\_DDB\ - D2\ \_DRD$$

$$DIF\_\mathrm{D2\_DDB\_D2\_DER} = D2\_DDB\ - D2\ \_DER$$

$$DIF\_\mathrm{D2\_DRD\_D2\_DER} = D2\_DRD\ - D2\ \_DER$$

No que diz respeito à criação das variáveis a partir das datas das diferentes etapas do processo de concessão do benefício sob a segunda perspectiva, foram criadas variáveis que identificassem se essas datas correspondem a um dia de semana ou a um final de semana.

Portanto, a partir dos atributos que registram as datas das diferentes etapas do processo de concessão do benefício, foram criadas 10 variáveis decorrentes da primeira perspectiva mais 5, como consequência da segunda.

6. Elaboração do Modelo Preditivo

De acordo com Chapman *et al* (2000), antes de criarmos um modelo preditivo, é necessário gerar um procedimento ou mecanismo para testar sua qualidade e validade. Os autores citam como exemplo em projetos que utilizam a aprendizagem supervisionadas de mineração de dados, como a classificação, é comum uso taxas de erro como medidas de qualidade para os modelos de mineração.

Diante disso, Chapman *et al* (2000) afirma que normalmente o conjunto de dados é separado em conjunto de treino e de teste. Uma vez feita esta separação, os autores sugerem que o treinamento do modelo ocorra no conjunto de treino e estimativa de sua qualidade seja realizada no conjunto de teste.

Nesse sentido, o conjunto de dados, o qual colecionava de 12640 registros, foi dividido de forma em que 67% (8.468 registros) foi utilizado como conjunto de dados para treino e o restante (4.172) foi utilizado como teste.

Vale ressaltar que tanto no conjunto de dados de treino como no de teste foi mantido a mesma distribuição de classes do conjunto de dados original. Em outras palavras, metade dos registros correspondia a benefícios fraudados (6320 registros) e o restante a benefícios não fraudados (target = 0).

Após se avançou para a identificação do algoritmo mais adequado para responder à pergunta proposta neste trabalho, qual seja, é possível distinguir os benefícios com maior probabilidade de conter irregularidades dos benefícios regulares com base nas características disponíveis dos dados com razoável margem de acerto?

Para obter subsídio para responder a esta pergunta foram realizadas três versões do modelo de classificação utilizando 9 algoritmos diferentes em cada delas com objetivo identificar o algoritmo mais adequado, haja vista as características do conjunto de dados.

O que diferenciou cada versão do modelo preditivo era a preparação dos dados realizada.

Segundo Wirth e Hipp (2000), a avaliação é a fase do projeto na qual se tem um ou mais modelos que parecem ter alta qualidade, sob a perspectiva da análise de dados. Os autores acrescentam que, no final desta fase, uma decisão sobre o uso dos resultados da mineração de dados deve ser alcançada.

Neste trabalho, a avaliação dos resultados obtidos em cada versão do modelo era utilizada de insumo para preparação do conjunto de dados antes do processamento da versão seguinte até se chegar na versão final.

Vale ressaltar que, na primeira versão do modelo preditivo foram identificados atributos que possuíam um viés involuntário derivado da escolha de benefícios de diferentes períodos e, por isso necessitavam ser excluídos.

Todavia, antes de excluí-los eram avaliadas alternativas que permitissem criar outros atributos a partir deles, como por exemplo, a criação de variável que identificasse se cada etapa do processo de concessão de benefícios ocorreu em dia de sema ou final de semana, conforme descrito no capítulo anterior.

O objetivo de registrar e apresentar os resultados obtidos nessas versões é verificar as diferenças de resultados alcançados em cada uma delas. De maneira sistemática as versões dos modelos foram processadas da seguinte forma:

a) Versão 1: Execução dos algoritmos após a limpeza dos dados, mas sem atributos adicionais.
b) Versão 2: Execução dos algoritmos após a limpeza dos dados, sem os atributos com viés decorrente da seleção dos dados e sem atributos adicionais.
c) Versão 3: Execução dos algoritmos após a limpeza dos dados, sem os atributos com viés decorrente da seleção dos dados e com atributos adicionais.

Nessas três versões do modelo foram utilizados nove algoritmos: *Multi-layer Perceptron Classifier*, *KNeighbors Classifier*, *Support Vector Classifier* (SVC), *Gaussian Process Classifier*, *Decision Tree Classifier*, *Random Forest Classifier*, *Ada Boost Classifier*, *Gaussian Naive Bayes* e *Quadratic Discriminant Analysis*. Os parâmetros utilizados para esses algoritmos constam no quadro below:

Quadro 7: Parâmetros utilizados nos algoritmos em cada versão de criação de modelo

Algoritmo	Parâmetros
Decision Tree Classifier	*{'class_weight': None, 'criterion': 'gini', 'max_depth': None, 'max_features': None, 'max_leaf_nodes': None, 'min_impurity_decrease': 0.0, 'min_impurity_split': None, 'min_samples_leaf': 1, 'min_samples_split': 2, 'min_weight_fraction_leaf': 0.0, 'presort': False, 'random_state': None, 'splitter': 'best'}*
Random Forest Classifier	*{'bootstrap': True, 'class_weight': None, 'criterion': 'gini', 'max_depth': None, 'max_features': 'auto', 'max_leaf_nodes': None, 'min_impurity_decrease': 0.0, 'min_impurity_split': None, 'min_samples_leaf': 1, 'min_samples_split': 2, 'min_weight_fraction_leaf': 0.0, 'n_estimators': 10, 'n_jobs': 1, 'oob_score': False, 'random_state': None, 'verbose': 0, 'warm_start': False}*

Ada Boost Classifier	*{'algorithm': 'SAMME.R', 'base_estimator': None, 'learning_rate': 1.0, 'n_estimators': 50, 'random_state': None}*
Quadratic Discriminant Analysis	*{'priors': None, 'reg_param': 0.0, 'store_covariance': False, 'store_covariances': None, 'tol': 0.0001}*
Multi-layer Perceptron Classifier	*{'activation': 'relu', 'alpha': 0.0001, 'batch_size': 'auto', 'beta_1': 0.9, 'beta_2': 0.999, 'early_stopping': False, 'epsilon': 1e-08, 'hidden_layer_sizes': (100,), 'learning_rate': 'constant', 'learning_rate_init': 0.001, 'max_iter': 200, 'momentum': 0.9, 'nesterovs_momentum': True, 'power_t': 0.5, 'random_state': None, 'shuffle': True, 'solver': 'adam', 'tol': 0.0001, 'validation_fraction': 0.1, 'verbose': False, 'warm_start': False}*
Gaussian Naive Bayes	*{'priors': None}*
KNeighbors Classifier	*{'algorithm': 'auto', 'leaf_size': 30, 'metric': 'minkowski', 'metric_params': None, 'n_jobs': 1, 'n_neighbors': 5, 'p': 2, 'weights': 'uniform'}*
Support Vectors Classifier	*{'C': 1.0, 'cache_size': 200, 'class_weight': None, 'coef0': 0.0, 'decision_function_shape': 'ovr', 'degree': 3, 'gamma': 'auto', 'kernel': 'rbf', 'max_iter': -1, 'probability': False, 'random_state': None, 'shrinking': True, 'tol': 0.001, 'verbose': False}*
Gaussian Process Classifier	*{'copy_X_train': True, 'kernel': None, 'max_iter_predict': 100, 'multi_class': 'one_vs_rest', 'n_jobs': 1, 'n_restarts_optimizer': 0, 'optimizer': 'fmin_l_bfgs_b', 'random_state': None, 'warm_start': False}*

Fonte: Elaboração pelo autor

6.1 Primeira versão do modelo de classificação

Antes de dar início à criação da primeira versão do modelo, foram realizadas as etapas de preparação do conjunto de dados a seguir:

- Exclusão dos atributos que possuíam mais de 90% dos dados faltantes;

- Exclusão das colunas em que não há alteração do dado;

- Conversão das variáveis do tipo data em variáveis numéricas de forma que representassem a quantidades de dias que antecediam a data de 31/12/2019 (último dia da Maciça mais recente utilizada).

- Substituição dos valores que representavam menos que 10% do total de ocorrência pelo valor '*other*' em cada atributo de variável categórica

- Criação de variáveis binárias (*dummy*) para representar os valores de cada atributo categórico. Por exemplo, para o atributo CS_SEXO_T, que corresponde ao sexo do titular do benefício, foi criado um atributo que quando assumia o valor 1 quando o sexo do titular do benefício era masculino e 0 quando era feminino.

Após a realização destas etapas de preparação dos dados, foram realizados o treinamento e o teste da primeira versão do modelo, considerando os parâmetros

descritos no Quadro 7 e obteve-se os resultados apresentados por intermédio da tabela below.

Tabela 10: Resultados obtidos pelos algoritmos na 1ª versão do modelo

Algoritmo	***Accuracy***	***Recall***	***Precision***
Decision Tree Classifier	1	1	1
Random Forest Classifier	1	1	1
Ada Boost Classifier	1	1	1
Quadratic Discriminant Analysis	1	1	1
Multi-layer Perceptron Classifier	0,999041	1	0,998063
Gaussian Naive Bayes Classifier	0,997843	1	0,995652
KNeighbors Classifier	0,993768	0,989811	0,997555
Support Vectors Classifier	0,585331	1	0,543656
Gaussian Process Classifier	0,571668	0,132945	1

Fonte: Elaboração pelo autor

Conforme exposto no capítulo que trata da compreensão do negócio, foi estabelecido meta 0,9 ao aferir a *precision* do modelo.

Diante disso, considerando os dados constantes na Tabela 10, a avaliação dos resultados alcançados pelos algoritmos, em um primeiro momento, pode levar à conclusão de um resultado extraordinário.

Todavia não é razoável pensar que 4 algoritmos[11] seriam capazes de acertar 100% das previsões realizadas, além de outros 3[12] terem acertado mais de 99% das previsões realizadas, conforme demonstra a Tabela 10.

Com isso, passou-se a investigar quais variáveis eram consideradas como mais relevantes para os modelos por meio, por exemplo, seleção univariada de atributos (*univariate feature selection*) e o uso de florestas de árvores para avaliar a importância dos atributos (*feature importances with forests of trees*).

[11] Os algoritmos que acertaram todas as previsões foram: *Decision Tree Classifier, Random Forest Classifier, Ada Boost Classifier e Quadratic Discriminant Analysis*.

[12] Os algoritmos que acertaram mais de 99% das previsões foram*: Multi-layer Perceptron Classifier, KNeighbors Classifier* e *Gaussian Naive Bayes*.

De acordo com PAES (2010), na análise univariada, investiga-se isoladamente a relação entre cada variável explicativa e a variável resposta, sem levar em conta as demais. A autora acrescenta que a análise univariada também pode ser entendida como uma análise bivariada, pois investiga a associação entre uma variável explicativa e uma resposta. Por último, a autora registra que existe muitas variáveis explicativas, portanto, a análise univariada pode servir como critério de seleção das variáveis que entrarão em um modelo final.

Segundo Ni (2012) o Chi-quadrado é uma das técnicas de análise univariada que se aplica quando se está diante de um problema de duas classes, que é o caso deste trabalho.

Na seleção univariada de atributos, são selecionados os melhores atributos com base em testes estatísticos univariados. Por meio dessa seleção, buscou-se identificar os 10 atributos cuja pontuação era a mais alta e obteve-se os resultados apresentados na tabela below.

Tabela 11: Os dez atributos mais relevantes segundo a análise univariada (1ª versão do modelo)

Atributo	**X²**
D2_DCB_	19.079.828,0
D2_DIB_	277.438,0
D2_DDB_	274.579,3
D2_DRD_	268.901,1
D2_DIP_	268.745,9
D2_DER_	263.584,9
VL_RUBRICA_2	246.829,9
VL_LIQUIDO	230.401,9
VL_RUBRICA_1	230.209,7
VL_BRUTO	230.203,5

Fonte: Elaboração pelo autor

Nessa análise, constatou-se que a estratégia de criar atributos calculando a diferença no dia 31/12/2019 e cada um dos campos do tipo data não era adequada. Era necessário criar atributos a partir dos campos que diziam respeito ao contracheque dos beneficiários. Isso se deve ao fato de esses atributos estarem distinguindo os benefícios fraudados dos não fraudados.

Com isso, foi calculada a diferença entre as datas das diferentes etapas do processo de concessão do benefício e a criação de atributos que indicavam quais daquelas datas correspondiam a final de semana, conforme descrito no capítulo que trata da Preparação dos Dados.

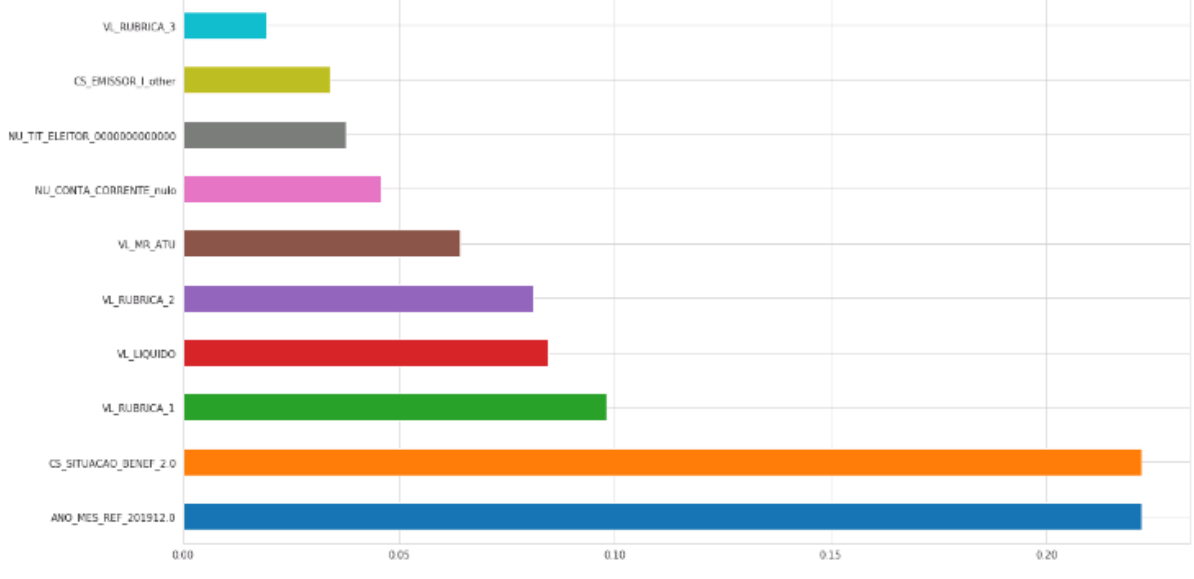
Além disso, foram criados atributos a partir dos quais se calculava a proporção do valor líquido do benefício em relação ao valor total, e a soma de cada uma das rubricas por meio das quais se realizam os pagamentos dos benefícios.

No que concerne à avaliação da importância dos atributos utilizando o método das florestas de árvores (*Feature importances with forests of trees*), Strobl (2008) afirma que as árvores de decisão funcionam dividindo os dados em subconjuntos que pertencem mais fortemente a uma classe. A autora destaca que a nesse método a árvore continuará construindo subconjuntos diferentes até entender e representar o relacionamento das variáveis com a variável alvo.

Os métodos de árvores calculam suas divisões, determinando matematicamente qual divisão de atributos permite distinguir com mais eficiência as classes.

Ao avaliar a importância dos atributos utilizando o método das florestas de árvores (*Feature importances with forests of trees*), foram identificados os 10 atributos[13] mais relevantes, conforme apresentado por meio da gráfico below:

Gráfico 8: Seleção dos 10 atributos mais relevantes segundo o método de florestas de árvores (1ª versão do modelo)



[13] Os 10 atributos mais relevantes de acordo com o método das florestas de árvores (Feature importances with forests of trees) foram: 'ANO_MES_REF_201912.0', 'CS_SITUACAO_BENEF_2.0', 'VL_RUBRICA_1', 'VL_LIQUIDO', 'VL_RUBRICA_2', 'VL_MR_ATU', 'NU_CONTA_CORRENTE_nulo', 'NU_TIT_ELEITOR_0000000000000', 'CS_EMISSOR_I_other' e 'VL_RUBRICA_3'.

Fonte: Elaboração pelo autor

Nesse método, implementou-se a meta estimadores que define um número aleatório de árvores de decisão para extrair os atributos mais relevantes (PETKOVIĆ, 2017). Desconsiderando os atributos já abordados na análise anterior, nessa análise, identificou-se a necessidade de excluir os atributos ‘CS_SITUAÇÃO_BENEF’ e ‘ANO_MES_REF’.

Os motivos para exclusão do atributo ‘CS_SITUAÇÃO_BENEF’ foram expostos no capítulo que trata da preparação dos dados. Já a exclusão do atributo ‘ANO_MES_REF’ é necessária visto que esse atributo é igual para todos benefícios regulares, conforme é descrito no capítulo que trata da compreensão dos dados e justificativa.

6.2 Segunda Versão do modelo de classificação

A segunda versão do modelo foi processada após a exclusão do conjunto de dados dos atributos que possuíam um viés involuntário derivado da escolha de benefícios de diferentes períodos. Essa opção decorre da necessidade de ter uma referência para avaliar o ganho proporcionado pelos novos atributos.

Considerando as conclusões descritas acima, antes de construir a segunda versão do modelo de classificação, foram realizados os seguintes ajustes no conjunto de dados:

- Exclusão dos atributos relacionados à data [14]

- Exclusão dos atributos relacionados ao valor do benefício[15]

-Exclusão do atributo Relacionadas à Situação do Benefício ('CS_SITUACAO_BENEF')

[14] 'D2_DCB_', 'D2_DDB_', 'D2_DRD_', 'D2_DIP_', 'D2_DER_', 'D2_DIB_', 'D2_FORMAT_CONC_'

[15] 'VL_LIQUIDO', 'VL_BRUTO', 'VL_MR_ATU', 'VL_RMI', 'TOT_DESCONTOS', 'VL_RUBRICA_1', 'VL_RUBRICA_2', 'VL_RUBRICA_3', 'VL_RUBRICA_3', 'VL_RUBRICA_4', 'VL_RUBRICA_5', 'VL_RUBRICA_6', 'VL_RUBRICA_7', 'VL_RUBRICA_8', 'VL_RUBRICA_9', 'VL_RUBRICA_10'

- Exclusão do atributo relacionado ao mês de pagamento da Maciça ('ANO_MES_REF')

Após a realização desses ajustes no conjunto de dados, foram realizados o treinamento e o teste da segunda versão dos modelos preditivos, considerando os parâmetros descritos no Quadro 7. Os resultados obtidos são apresentados por intermédio da tabela below.

Tabela 12: Resultados obtidos pelos algoritmos na 2ª versão do modelo

Algoritmo	***Accuracy***	***Recall***	***Precision***
Ada Boost Classifier	0,97	0,94	0,99
Random Forest Classifier	0,97	0,95	0,99
Quadratic Discriminant Analysis	0,97	0,93	1,00
Decision Tree Classifier	0,96	0,96	0,96
Gaussian Naive Bayes	0,91	0,88	0,93
Gaussian Process Classifier	0,64	0,58	0,65
KNeighbors Classifier	0,61	0,53	0,62
Support Vector Classifier	0,60	0,54	0,61
Multi-layer Perceptron Classifier	0,50	1,00	0,49

Fonte: Elaboração pelo autor

Comparando os dados constantes na Tabela 10 e na Tabela 12, verifica-se que houve uma piora no desempenho dos algoritmos. Esse resultado já era esperado tendo em vista a exclusão dos atributos supracitados.

Sem embargo, verifica-se que 4 algoritmos permanecem com um índice de acerto das previsões acima de 90% nas três medidas de desempenho utilizadas. Dessa forma, avançou-se para identificação dos atributos que poderiam ser considerados mais relevante sob o prisma da análise univariada.

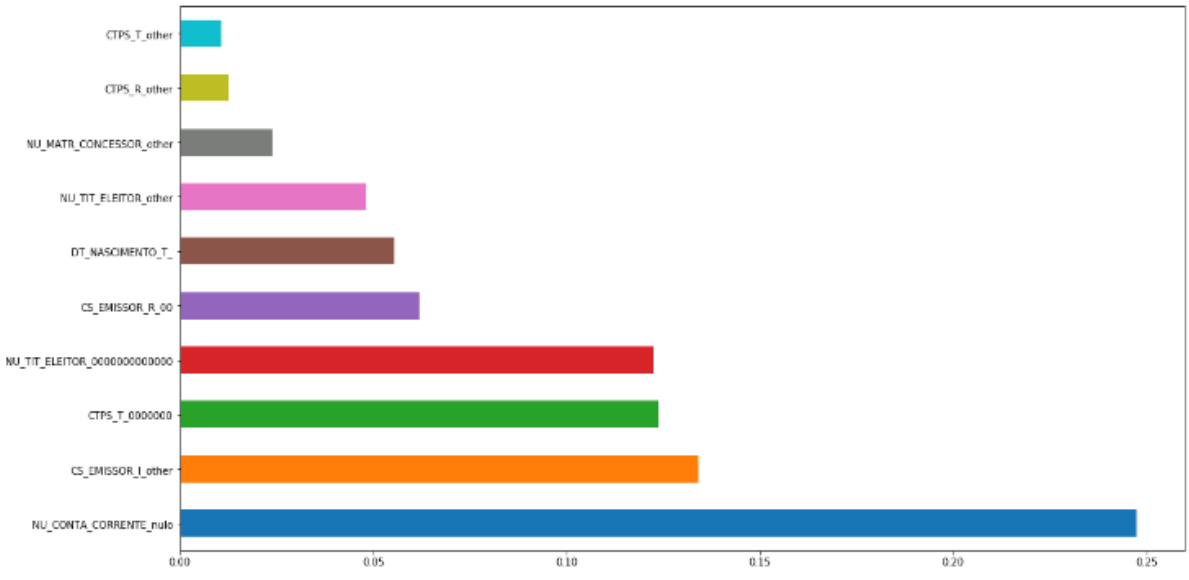
Tabela 13: Os dez atributos mais relevantes segundo a análise univariada (2º versão do modelo)

Atributo	**X^2**
DT_NASCIMENTO_T_	9.493,7
NU_CONTA_CORRENTE_nulo	3.668,0
CS_EMISSOR_I_other	2.425,0
CTPS_T_0000000	1.993,0
NU_TIT_ELEITOR_0000000000000	1.100,5
CS_EMISSOR_R_00	423,3
CS_RUBRICA_5_902.0	250,7
CS_RUBRICA_4_316.0	247,4
CTPS_T_other	171,3
CTPS_SERIE_T_other	170,7

Fonte: Elaboração pelo autor

Analisando os dados constante na Tabela 13, verifica-se que já não constam atributos que permitam distinguir os benefícios fraudados dos não fraudados. Tampouco há atributos que necessitem ser excluídos quando se analisa os 10 atributos[16] mais relevantes identificados por meio do método das florestas de árvores, conforme demonstrado no gráfico below.

Gráfico 9: Seleção dos 10 atributos mais relevantes segundo o método de florestas de árvores (2ª versão do modelo)



Fonte: Elaboração pelo autor

6.3 Terceira versão do modelo de classificação

Considerando os resultados obtidos pela 2ª versão do modelo como referência, foram incluídos os atributos cuja criação foi tratada no capítulo de preparação dos dados. Além disso, foram mantidos os parâmetros descritos no Quadro 7 e obteve-se os resultados apresentados por meio da tabela below:

[16] Os 10 atributos mais relevantes de acordo com o método das florestas de árvores (Feature importances with forests of trees) foram: 'NU_CONTA_CORRENTE_nulo', 'CS_EMISSOR_I_other', 'CTPS_T_0000000', 'NU_TIT_ELEITOR_0000000000000', 'CS_EMISSOR_R_00', 'DT_NASCIMENTO_T_', 'NU_TIT_ELEITOR_other', 'NU_MATR_CONCESSOR_other', 'CTPS_R_other', 'CTPS_T_other'

Tabela 14: Resultados obtidos pelos algoritmos na 3ª versão do modelo

Algoritmo	***Accuracy***	***Recall***	***Precision***
Ada Boost Classifier	0,97	0,94	1,00
Quadratic Discriminant Analysis	0,97	0,93	1,00
Random Forest Classifier	0,96	0,94	0,98
Decision Tree Classifier	0,96	0,95	0,96
Gaussian Naive Bayes	0,89	0,89	0,89
Gaussian Process Classifier	0,57	0,50	0,57
KNeighbors Classifier	0,55	0,53	0,55
Support Vector Classifier	0,54	0,63	0,52
Multi-layer Perceptron Classifier	0,51	0,00	0,40

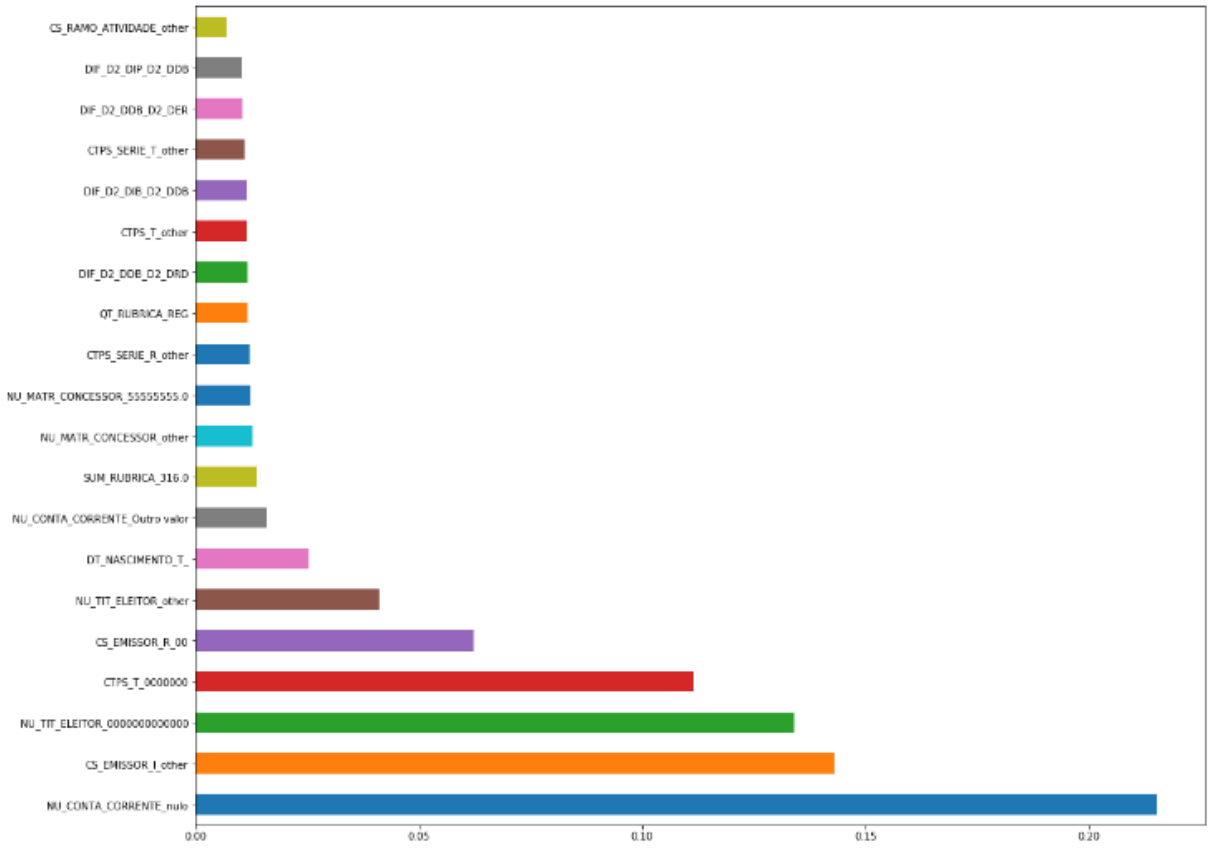
Fonte: Elaboração pelo autor

Embora não haja uma diferença significativa entre os quatro algoritmos de melhor desempenho, os dados da tabela above demonstram que o melhor desempenho foi obtido utilizando o algoritmo *Ada Boost Classifier*.

Ao comparar os resultados constantes na Tabela 12 e Tabela 14, verifica-se que, a princípio, os ajustes realizados no conjunto de dados resultaram em uma melhora na precisão (*precision)* do modelo. Além disso, também foi observado melhoras nas outras medidas, mas na quarta casa decimal.

Para obter evidência de que essa melhoria era decorrente da inclusão de novos atributos, foi utilizado o método das florestas e árvores para identificar os 20 atributos mais relevantes, os quais são apresentados por meio do gráfico below

Gráfico 10: Seleção dos 20 atributos mais relevantes segundo o método de florestas de árvores (3ª versão do modelo)



Fonte: Elaboração pelo autor

Ao analisar os 20 atributos mais relevantes constantes no gráfico above, verifica-se que nenhum dos novos atributos estão entre os 10 mais relevantes[17]. Todavia, há cinco deles[18] entre os 20 mais relevantes.

Segundo Chapman *et al* (2000), na etapa de avaliação, busca-se verificar o grau de atendimento do modelo aos objetivos de negócios e procura-se determinar se há razão para que o modelo seja considerado deficiente sob a perspectiva do negócio. Os autores apresentam como alternativa testar o(s) modelo(s) em casos reais se as restrições de tempo e orçamento permitirem.

---

[17] Os 10 atributos mais relevantes de acordo com o método das florestas de árvores na 3ª versão do modelo foram: 'NU_CONTA_CORRENTE_nulo', 'NU_TIT_ELEITOR_0000000000000', 'CS_EMISSOR_I_other', 'CTPS_T_0000000', 'CS_EMISSOR_R_00', 'NU_TIT_ELEITOR_other', 'DT_NASCIMENTO_T_', 'NU_MATR_CONCESSOR_55555555.0', 'NU_CONTA_CORRENTE_Outro valor', 'NU_MATR_CONCESSOR_other'.

[18] Os cinco atributos que foram criados e constam entre os 20 mais relevantes de acordo com o método das florestas de árvores na 3ª versão do modelo são: 'DIF_D2_DDB_D2_DER', 'DIF_D2_DIB_D2_DDB', 'DIF_D2_DIP_D2_DDB', 'DIF_D2_DDB_D2_DRD', 'SUM_RUBRICA_316.0'

Diante do exposto e considerando os resultados obtidos, entende-se que é necessário avançar para a etapa de validação do modelo de forma a aferir seu resultado com dados reais.

No entanto não é possível colher os resultados desta nova etapa a tempo de apresentá-la neste trabalho. Vários órgãos como, por exemplo, o Tribunal de Conta da União, Controladoria-Geral da União (CGU), Polícia Federal e Ministério Público Federal propõe que o INSS revise benefícios por conterem indícios de irregularidade. Isso faz com que o INSS contenha um estoque de benefícios a serem revisados o que inviabiliza a revisão tempestiva dos benefícios identificados neste trabalho como aqueles com maior probabilidade de conter irregularidade.

7. Considerações Finais

No que diz respeito à aplicação, Chapman *et al* (2000) afirma que nesta etapa é importante resumir a experiência adquirida durante o projeto, citando, por exemplo armadilhas e abordagens enganosas utilizadas.

Dessa forma, é importante registrar algumas das experiências adquiridas no processo que culminou na construção deste modelo.

Os benefícios assistenciais e previdenciários possuem diferentes objetivos e, consequentemente, têm público-alvo diferentes. Portanto, não é razoável pensar em construir um modelo de classificação que seja capaz de identificar entre os benefícios previdenciários e assistenciais aqueles com maior probabilidade de conter irregularidades.

Assim, a melhor estratégia a ser adotada para realizar um trabalho como este é definir a espécie de benefício assistencial com a qual vai se trabalhar.

Com relação a definir a espécie de benefício, entende-se que não é viável realizar esta escolha antes de analisar a base de registros cessados por irregularidade. No pré-projeto deste trabalho, por exemplo, pretendia-se criar um modelo para identificar benefícios da espécie aposentadoria por idade visto que essa espécie representava 31% de todos os benefícios da Maciça de dezembro de 2019.

Em princípio, essa seria uma estratégia razoável, uma vez que o TCU estimou em 11,4% o percentual de benefícios pagos indevidamente, conforme exposto no capítulo que trata da compreensão dos dados e justificativa. Assim, a escolha da espécie de benefício mais representativa seria uma forma de aumentar o benefício potencial a ser gerado pelo modelo.

No entanto, não havia registros suficientes de aposentadoria por idade cessadas em virtude de irregularidade. Com isso, optou-se pela seleção da espécie de benefício que colecionava a maior quantidade de benefícios cessados em decorrência de constatação de irregularidade, qual seja, o BPC idoso.

Outro ponto que vale destacar é o que diz respeito aos atributos que integram o conjunto de dados. É importante que a avaliação do desempenho do modelo não se restrinja a medidas como *accuracy*, *precision* e *recall* ou análise de matrizes de confusão.

De acordo com a experiência adquirida neste trabalho é importante utilizar um método que permita identificar quais são os atributos mais relevantes para o modelo. Dessa forma, é possível identificar atributos que estejam distinguindo indiretamente os benefícios irregulares do regulares.

Vale destacar que, por mais que se conheça o negócio e as bases de dados que serão utilizadas, a identificação desses atributos não é óbvia. A construção do conjunto de dados para treinamento e teste para um modelo de classificação demanda uma série de decisões e manipulações dos dados. Isso pode fazer com que os valores dos atributos passem a distinguir quais são os benefícios irregulares.

Neste trabalho, por exemplo, isso ocorria com o atributo VL_BRUTO. O seu valor indicava quais benefícios eram regulares e irregulares, visto que os benefícios regulares foram selecionados na Maciça de dezembro de 2019[19].

Embora o desempenho alcançado na terceira versão do modelo atenda ao critério de sucesso definido neste trabalho, entende-se que é necessário testar o modelo com dados reais para aferir adequadamente o desempenho do modelo.

Essa conclusão decorre da impossibilidade de descartar que este resultado decorra de um sobreajuste do modelo em razão de não haver dados suficientes para serem aprendidos. Nesse caso, o modelo estaria memorizando os dados reconhecidos e, com isso, não conseguiria fazer a generalização para identificar novos casos.

Outro fator a ser considerado são as mudanças das regras de negócio que ocorrem com o passar do tempo os seus feitos no conjunto de dados.

Um exemplo dessas alterações identificadas neste trabalho são os aumentos do benefício ocorridos entre 2014 e 2019, como consequência do aumento do salário mínimo[20].

Outro exemplo são as ACPs que alteram o critério de apuração da renda *per capita* para elegibilidade ao benefício. Aliás, com relação ao critério renda, cumpre registrar que no dia 11/03/2020 o Congresso Nacional derrubou o veto presidencial ao

---

[19] Os motivos desta seleção constam no capítulo que trata da compreensão dos dados

[20] No Quadro 6 estão registrados os valores do salário mínimo no período de janeiro de 2014 até dezembro de 2019.

Projeto de Lei do Senado Federal n° 55/1996 o que resulta no aumento do critério renda de ¼ para ½ salário mínimo.

Por fim, destaca-se a importância das considerações realizadas neste trabalho, uma vez que o modelo de classificação é treinado e testado com base em um conjunto de dados que reflete o passado. Mas, o objetivo desse modelo é identificar futuros benefícios com maior probabilidade de conter fraude.

Referências bibliográficas

BARBETTA, Pedro. **Estatística Aplicada às Ciências Sociais.** 5ª edição. Florianópolis: Editora da UFCS, 2002.

BEN-DAVID, Arie. **A lot of randomness is hiding in accuracy.** Engineering Applications of Artificial Intelligence, v. 20, n. 7, p. 875-885, 2007.

BENIWAL, Sunita; ARORA, Jitender. **Classification and feature selection techniques in data mining**. International journal of engineering research & technology (ijert), v. 1, n. 6, p. 1-6, 2012.

BRASIL. **Constituição da República Federativa do Brasil**. 1988.

BRASIL. **Lei nº 13.846, de 18 de junho de 2019.** Institui o Programa Especial para Análise de Benefícios com Indícios de Irregularidade e dá outras providências. 2019.

BRASIL. **Lei nº 8.742, de 7 de dezembro de 1993**. Dispõe sobre a organização da Assistência Social e dá outras providências. 1993

CAPELA, Marisa Veiga; CAPELA, Jorge MV. **Elaboração de gráficos box-plot em planilhas de cálculo**. In: CONGRESSO DE MATEMÁTICA APLICADA E COMPUTACIONAL DA REGIÃO SUDESTE–CNMAC Sudeste. 2011.

CHAPMAN, Pete et al. **CRISP-DM 1.0: Step-by-step data mining guide**. SPSS inc, v. 16, 2000.

DANTAS, José Alves *et al*. **Custo-benefício do controle: proposta de um método para avaliação com base no COSO**. 2010.

DEPARTMENT FOR WORK & PENSIONS. **Fraud and Error in the Benefit System: 2018/19 Estimates**. Great Britain. 2019. Disponível em: https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/801594/fraud-and-error-stats-release-2018-2019-estimates.pdf. Acessado em: 10 de fevereiro de 2020.

DI PIETRO, Maria Sylvia Zanella. **Direito administrativo**. São Paulo: Forense, 2019.

DUDA, Richard O.; HART, Peter E.; STORK, David G. **Pattern classification and scene analysis**. New York: Wiley, 1973.

GRUS, Joel. **Data science from scratch: first principles with python**. O'Reilly Media, 2019.

HASTIE, Trevor; TIBSHIRANI, Robert; FRIEDMAN, Jerome. **The elements of statistical learning: data mining, inference, and prediction.** Springer Science & Business Media, 2009.

LANGLEY, Pat et al. **Selection of relevant features in machine learning.** In: Proceedings of the AAAI Fall symposium on relevance. 1994. p. 245-271.

NI, Weizeng. **A review and comparative study on univariate feature selection techniques.** 2012. Tese de Doutorado. University of Cincinnati.

PAES, Ângela Tavares. **Análise univariada e multivariada.** Revista de Educação Continuada em Saúde, v. 8, parte 2, p. 1 e 2, 2010.

PETKOVIĆ, Matej; DŽEROSKI, Sašo; KOCEV, Dragi. **Feature ranking for multi-target regression with tree ensemble methods**. In: International Conference on Discovery Science. Springer, Cham, 2017. p. 171-185.

POWERS, David Martin. **Evaluation: from precision, recall and F-measure to ROC, informedness, markedness and correlation.** 2011.

RÄTSCH, Gunnar. **A brief introduction into machine learning.** Friedrich Miescher Laboratory of the Max Planck Society, 2004.

SALEEM, Asma et al. **Pre-processing methods of data mining.** In: 2014 IEEE/ACM 7th International Conference on Utility and Cloud Computing. IEEE, 2014. p. 451-456.

SATHYA, R.; ABRAHAM, Annamma. **Comparison of supervised and unsupervised learning algorithms for pattern classification.** International Journal of Advanced Research in Artificial Intelligence, v. 2, n. 2, p. 34-38, 2013.

STROBL, Carolin et al. **Conditional variable importance for random forests.** BMC bioinformatics, v. 9, n. 1, p. 307, 2008.

VISA, Sofia et al. **Confusion Matrix-based Feature Selection.** MAICS, v. 710, p. 120-127, 2011.

WIRTH, Rüdiger; HIPP, Jochen**. CRISP-DM: Towards a standard process model for data mining**. In: Proceedings of the 4th international conference on the practical applications of knowledge discovery and data mining. London, UK: Springer-Verlag, 2000. p. 29-39.

ANEXO A – Ações Civis Públicas cujo objeto são os Benefícios de Prestação Continuada

Foi elaborado o quadro below para registrar as Ações Civis Públicas que promovem alterações nos critérios de elegibilidade para acessar o benefício de prestação continuada ao idoso (BPC idoso) cuja idade é igual ou superior a 65 anos.

Quadro 8: Ações Civis Públicas que tratam dos benefícios de prestação continuada

Ação Civil Pública	Objeto	Localidade de abrangência da decisão
2009.38.00.005945-2	Desconsideração de outro BPC e de benefício previdenciário de valor mínimo na análise da renda per capita familiar	Estado de Minas Gerais
5000339-37.2011.4.04.7210	Desconsideração de outro BPC e de benefício previdenciário de valor mínimo na análise da renda per capita familiar	**Municípios:** Anchieta, Bandeirante, Barra Bonita, Belmonte, Bom Jesus do Oeste, Caibi, Campo Erê, Cunha Porã, Cunhataí, Descanso, Dionísio Cerqueira, Flor do Sertão, Guaraciaba, Guarujá do Sul, Iporã do Oeste, Iraceminha, Itapiranga, Maravilha, Mondaí, Palma Sola, Paraíso, Princesa, Riqueza, Romelândia, Saltinho, Santa Helena, Santa Terezinha do Progresso, São Bernardino, São João do Oeste, São José do Cedro, São Miguel da Boa Vista, Tigrinhos, Tunápolis e São Miguel do Oeste
2005.71.00045257-0	Desconsideração de outro BPC e de benefício previdenciário de valor mínimo na análise da renda per capita familiar	**Municípios**: Alvorada, Arambaré, Arroio do Sal, Arroio dos Ratos, Balneário Pinhal, Barão do Triunfo, Barra do Ribeiro, Brochier do Marata, Butiá Cachoeirinha, Capão da Canoa, Capela de Santana, Capivari do Sul, Caraá, Cerro Grande do Sul, Charqueadas, Cidreira, Dom Pedro de Alcântara, Eldorado do Sul, Fazenda Vilanova, General Câmara, Glorinha, Gravataí, Guaíba, Imbé, Itati, Mampituba, Maquine, Marata, Mariana Pimentel, Minas do Leão, Montenegro, Morrinhos do Sul, Mostardas, Osório, Palmares do Sul, Pareci Novo, Paverama, Porto Alegre, Santo Antônio da Patrulha, São Jerônimo, Sentinela do Sul, Sertão Santana, Tabaí, Tapes, Taquari, Tavares, Terra de Areia, Torres, Tramandaí, Três Cachoeiras, Três Forquilhas, Triunfo, Viamão e Xangri-lá

2006.71.17.001095-3	Desconsideração de outro BPC e de benefício previdenciário de valor mínimo na análise da renda per capita familiar	**Municípios:** Água Santa, Camargo, Capão Bonito do Sul, Casca, Caseiros, Ciríaco, Coxilha, David Canabarro, Ernestina, Gentil, Ibiaçá, Ibiraiaras, Lagoa Vermelha, Marau, Mato Castelhano, Montauri, Muliterno, Nicolau Vergueiro, Nova Alvorada, Passo Fundo, Pontão, Santa Cecília do Sul, Santo Antônio do Palma, São Domingos do Sul, Sertão, Tapejara, Tupanci do Sul, União da Serra, Vanini, Vila Lângaro, Vila Maria
2002.71.04.000395-5	Desconsideração de outro BPC e de benefício previdenciário de valor mínimo na análise da renda per capita familiar	**Municípios:** Ajuricaba, Augusto Pestana, Bozano, Catuípe, Chiapeta, Coronel Barros, Ijuí, Inhacorá, Jóia, Nova Ramada, Santo Augusto, São Valério do Sul
0000003-61.2010.4.04.7111	Desconsideração de outro BPC e de benefício previdenciário de valor mínimo na análise da renda per capita familiar	**Municípios:** Gramado Xavier, Herveiras, Santa Cruz do Sul, Sinimbu, Vale do Sol, Vale Verde e Vera Cruz
5000852-57.2015.4.04.7212	Desconsideração do valor decorrente de qualquer benefício assistencial ou previdenciário de renda mínima percebido por idoso ou por pessoa com deficiência membro do grupo familiar do requerente de BPC	**Municípios:** Alto Bela Vista, Arabutã, Corcórdia, Faxinal do Guedes, Ipira, Ipumirim, Irani, Itá, Jaborá, Lindóia do Sul, Passos Maia, Peritiba, Piratuba, Ponte Serrada, Presidente Castelo Branco, Seara, Vargeão e Xavantina, todos do Estado de Santa Catarina.
0006972-83.2012.4.01.3400	Determinação para que o INSS que se abstenha de indeferir pedidos de benefícios assistenciais exclusivamente em razão da nacionalidade dos requerentes, a fim de garantir, em todo território nacional, aos estrangeiros residentes no País em situação regular, idosos ou com deficiência, o direito ao benefício assistencial previsto no inc.V do art. 203 da Constituição Federal.	**Nacional**
0000083-10.2007.4.05.8305	Concessão de Benefício de Prestação Continuada-BPC considerando ½ salário-mínimo como critério objetivo de apuração de miserabilidade. Exclusão de benefício previdenciário no valor de salário mínimo recebido por idoso	**Municípios:** Angelim, Águas Belas, Brejão, Bom Conselho, Caetés, Capoeiras, Canhotinho, Correntes, Calçado, Garanhuns, Iati, Ibirajuba, Jucati, Jupi, Lajedo, Lagoa do Ouro, Palmeirina, Paranatama, Quipapá, São João, São Bento do Una, Saloá e Terezinha, todos do Estado de Pernambuco
0004265-82.2016.4.03.6105	Exclusão, no cálculo da renda familiar, de benefícios previdenciários e assistenciais no valor de até um	**Municípios:** Amparo, Campinas, Capivari, Elias Fausto, Holambra, Hortolândia, Indaiatuba, Itatiba, Jaguariúna, Jarinu, Mombuca, Monte Mor, Morungaba, Paulínia,

	salário mínimo recebidos por outro membro do grupo idoso ou deficiente.	Pedreira, Rafard, Santo Antônio de Posse, Sumaré, Valinhos e Vinhedo, todos do Estado de São Paulo
5002350-92.2013.404.7202	Desconsideração, na análise dos requerimentos de BPC, dos valores decorrentes de qualquer benefício assistencial ou previdenciário de renda mínima, percebido por membro do grupo, idoso ou pessoa com deficiência.	**Municípios:** Abelardo Luz, Águas de Chapecó, Águas Frias, Arvoredo, Bom Jesus, Caxambu do Sul, Chapecó, Cordilheira Alta, Coronel Freitas, Coronel Martins, Entre Rios, Formosa do Sul, Galvão, Guatambu, Ipuaçu, Irati, Jardinópolis, Jupiá, Lageado Grande, Marema, Modelo, Nova Erechim, Nova Itaberaba, Novo Horizonte, Ouro Verde, Paial, Palmitos, Pinhalzinho, Planalto Alegre, Quilombo, Santiago do Sul, São Carlos, São Domingos, São Lourenço do Oeste, Saudades, Serra Alta, Sul Brasil, União do Oeste, Xanxerê e Xaxim, todos do estado de Santa Catarina.
2005.72.05.001947-1	Desconsideração de benefício assistencial percebido por familiar idoso ou deficiente e benefício previdenciário de valor mínimo percebido por familiar idoso, na análise da renda per capita familiar a que se refere o art. 20 da Lei nº 8.742/93	**Municípios:** Apiúna, Ascurra, Benedito Novo, Blumenau, Botuverá, Brusque, Doutor Pedrinho, Gaspar, Guabiruba, Ilhota, Indaial, Luiz Alves, Pomerode, Rio dos Cedros, Rodeio e Timbó, todos do estado de Santa Catarina.
526-61.2017.4.01.3603	Excluir do cálculo da renda per capita familiar os benefícios previdenciários de um salário mínimo recebidos por outro idoso da família, bem como excluir do referido cálculo o benefício assistencial recebido por outro membro da família idoso ou com necessidades especiais; e adotar como critério objetivo de presunção de miserabilidade a renda per capita familiar de ½ salário mínimo	**Municípios:** Sinop, Alta Floresta, Apiacás, Carlinda, Cláudia, Colíder, Feliz Natal, Guarantã do Norte, Ipiranga do Norte, Itaúba, Lucas do Rio Verde, Marcelândia, Matupá, Nova Bandeirantes, Nova Canaã do Norte, Nova Guarita, Nova Monte Verde, Nova Santa Helena, Novo Mundo, Paranaíta, Peixoto de Azevedo, Santa Carmem, Sorriso, Terra Nova do Norte, União do Sul, Vera, Juara, Juína, Colniza, Nova Mutum, Novo Horizonte do Norte, Porto dos Gaúchos, Tabaporã, Brasnorte, Castanheira, Juruena, Itanhangá, Nova Ubiratã, Santa Rita do Trivelato, Tapurah, Aripuana e Cotriguaçu, todos do estado de Mato Grosso.
0001038-69.2007.4.03.6115	Exclusão no cálculo da renda familiar de benefícios previdenciários e assistenciais no valor de até um salário mínimo recebidos por outro membro do grupo idoso ou deficiente, independentemente de renúncia de benefícios, em âmbito territorial da Subseção Judiciária de São Carlos/SP	**Municípios:** Brotas, Descalvado, Dourado, Ibaté, Pirassununga, Porto Ferreira, Ribeirão Bonito, Santa Cruz da Conceição, Santa Cruz das Palmeiras, Santa Rita do Passa Quatro, São Carlos e Tambaú, todos do estado de São Paulo.
0011259-41.2007.403.6106	Desconsideração do valor decorrente de qualquer benefício assistencial ou previdenciário de renda	**Municípios:** Adolfo, Altair, Álvares Florence, Américo de Campos, Bady Bassit, Bálsamo, Cardoso, Cedral, Cosmorama, Floreal, Guapiaçu, Guaraci, Icém, Ipiguá, Irapuã, Jaci, José

	mínima percebido por idoso com mais de 65 anos membro do grupo familiar do requerente de BPC	Bonifácio, Macaubal, Magda, Mendonça, Mirassol, Mirassolândia, Monte Aprazível, Neves Paulista, Nhandeara, Nipoã, Nova Aliança, Nova Granada, Novo Horizonte, Olímpia, Onda Verde, Orindiúva, Palestina, Parisi, Paulo de Faria, Planalto, Poloni, Pontes Gestal, Potirendaba, Riolândia, Sales, São José do Rio Preto, Sebastianópolis do Sul, Severínia, Tanabi, Ubarana, Uchoa, União Paulista, Urupês, Valentim Gentil e Votuporanga, todos do Estado de São Paulo.
500339-37.2011.404.7210	Desconsideração de outro BPC e de benefício previdenciário de valor mínimo na análise da renda per capita familiar	**Municípios:** Anchieta, Bandeirante, Barra Bonita, Belmonte, Bom Jesus do Oeste, Caibi, Campo Erê, Cunha Porã, Cunhataí, Descanso, Dionísio Cerqueira, Flor do Sertão, Guaraciaba, Guarujá do Sul, Iporã do Oeste, Iraceminha, Itapiranga, Maravilha, Mondaí, Palma Sola, Paraíso, Princesa, Riqueza, Romelândia, Saltinho, Santa Helena, Santa Terezinha do Progresso, São Bernardino, São João do Oeste, São José do Cedro, São Miguel da Boa Vista, Tigrinhos, Tunápolis e São Miguel do Oeste, todos do Estado de Santa Catarina
2007.71.02.000569-5	Todos os Benefícios assistenciais, oriundos das cidades citadas, deverá ser desconsiderado no cálculo da renda per capita familiar, o valor de qualquer outro benefício assistencial percebido por outro membro do grupo familiar	**Municípios:** Agudo, Dilermando de Aguiar, Dona Francisca, Faxinal do Soturno, Formigueiro, Itaara, Ivorá, Jari, Julio de Castilhos, Mata, Nova Palma, Pinhal Grande, Quevedos, Restinga Seca, Santa Margarida do Sul, Santa Maria, São João do Polesini, São Martinho da Serra, São Pedro do Sul, São Sepé, Silveira Martins, Toropi, Vila Nova do Sul
2007.71.19.000090-8	Deverá desconsiderar no cálculo da renda per capta o valor de outro benefício assistencial (B-87 ou 88), percebido por outro membro do grupo familiar.	**Municípios:** Arroio do Tigre, Cachoeira do Sul, Caçapava do Sul, Cerro Branco, Encruzilhada do Sul, Ibarama, Lagoa Bonita do Sul, Novo Cabrais, Paraíso do Sul, Passa Sete, Segredo e Sobradinho
2007.71.20.000785-2	Deverá desconsiderar no cálculo da renda per capta o valor de outro benefício assistencial (B-87 ou 88), percebido por outro membro do grupo familiar.	**Municípios:** Bossoroca, Capão do Cipó, Itacurubi, Jaguari, Nova Esperança do Sul, Santiago, São Vicente do sul e Unistalda
0002356-52.2002.404.7209	Nos requerimentos de benefício assistencial formulados por idosos o INSS deverá deduzir os gastos comprovados e relacionados diretamente ao próprio idoso, representados por medicamentos, alimentação especial, fraldas descartáveis, plano de saúde, tratamento médico, psicológico e fisioterápico e transporte especial	**Municípios:** Corupá, Guaramirim, Jaraguá do Sul, Massaranduba e Schroeder, todos do Estado de Santa Catarina

2003.72.00.001108-0	Decisão judicial determinou ao INSS que: a) na análise dos requerimentos de BPC exclua as despesas do requerente relacionadas diretamente com a deficiência, incapacidade ou idade avançada, em especial, despesas com medicamentos, alimentação especial, fraldas descartáveis, tratamento médico, psicológico e fisioterápico; b) exclua do cômputo da renda per capita familiar o valor do benefício assistencial concedido à pessoa idosa (B/88), membro do grupo familiar	**Municípios:** Águas Mornas, Alfredo Wagner, Angelina, Anitápolis, Antônio Carlos, Biguaçu, Canelinha, Florianópolis, Governador Celso Ramos, Palhoça, Paulo Lopes, Rancho Queimado, Santo Amaro da Imperatriz, São Bonifácio, São João Batista, São José, São Pedro de Alcântara, Tijucas, todos do Estado de Santa Catarina
2007.83.05.000083-0	Concessão de Benefício de Prestação Continuada-BPC considerando ½ salário-mínimo como critério objetivo de apuração de miserabilidade. Exclusão de benefício previdenciário no valor de salário mínimo recebido por idoso	**Municípios:** Angelim, Águas Belas, Brejão, Bom Conselho, Caetés, Capoeiras, Canhotinho, Correntes, Calçado, Garanhuns, Iati, Ibirajuba, Jucati, Jupi, Lajedo, Lagoa do Ouro, Palmeirina, Paranatama, Quipapá, São João, São Bento do Una, Saloá e Terezinha, todos do Estado de Pernambuco
2001.72.05.007738-6	Determinou ao INSS que deixe de aplicar o critério objetivo de avaliação da renda per capita do grupo familiar para a concessão dos Benefícios de Prestação Continuada da Assistência Social-BPC à pessoa com deficiência, conforme dispõe o § 3º do art. 20 da Lei nº 8.742, de 7 de dezembro de 1993	**Municípios:** Agrolândia, Agronômica, Apiúna, Ascurra Atalanta, Aurora, Benedito Novo, Blumenau, Botuverá, Braço do Trombudo, Brusque, Chapadão do Lageado, Dona Emma, Doutor Pedrinho, Gaspar, Guabiruba, Ibirama, Ilhota, Imbuia, Indaial, Ituporanga, José Boiteux, Laurentino, Lontras, Luiz Alves, Mirim Doce, Petrolândia, Pomerode, Pouso Redondo, Presidente Getúlio, Presidente Nereu, Rio do Campo, Rio do Oeste, Rio do Sul, Rio dos Cedros, Rodeio, Salete, Santa Terezinha, Taió, Timbó, Trombudo Central, Vidal Ramos, Vitor Meireles e Witmarsum.
2007.72.15.000170-9	Modificar a forma objetiva de cálculo da renda per capita do grupo familiar para acesso ao BPC, requerido por pessoa com deficiência e a pessoa idosa.	**Município:** Nova Trento/SC
2001.72.03.001315-9	Na análise de requerimentos de BPC excluir da renda per capita familiar as despesas relacionadas diretamente à doença do requerente.	**Município:** Água doce, Alto Bela Vista, Arabutã, Arroio Trinta, Brunópolis, Caçador, Calmon, Campos Novos, Capinzal, Catanduvas, Concórdia, Erval Velho, Fraiburgo, Herval D'Oeste, Ibiam, Ibicaré, Iomerê, Ipira, Ipumirim, Irani, Irineópolis, Jaborá, Joaçaba, Lacerdópolis, Lindóia do Sul, Luzerna, Macieira, Matos Costa, Monte Carlo, Ouro, Passos Maia, Peritiba, Pinheiro Preto, Piratuba, Ponte Serrada, Porto União, Presidente Castelo

		Branco, Rio das Antas, Salto Veloso, Tangará, Treze Tílias, Vargeão, Vargem, Vargem Bonita, Videira e Zortéa
5044874-22.2013.404.7100	Exclusão do cálculo da renda per capita familiar das despesas do requerente de benefício assistencial que decorram diretamente da deficiência, incapacidade ou idade avançada, com medicamentos, alimentação especial, fraldas descartáveis e consultas na área de saúde, requeridas e negados pelo Estado	Nacional
0012938-20.1997.4.04.7005	Exclusão no cálculo da renda familiar de benefícios previdenciários e assistenciais de renda mínima recebidos por idoso com 65 (sessenta e cinco) anos ou mais, ou em razão de deficiência, independentemente de idade	**Municípios:** Ampére, Anahy, Barracão, Boa Esperança do Iguaçu, Boa Vista da Aparecida, Braganey, Cafelândia, Campina da Lagoa, Campo Bonito, Capanema, Capitão Leônidas Marques, Cascavel, Catanduvas, Corbélia, Cruzeiro do Sul, Dois Vizinhos, Enéas Marques, Flor da Serra do Sul, Francisco Beltrão, Guaraniaçu, Ibema, Iguatu, Lindoeste, Maripá, Marmeleiro, Nova Esperança do Sudoeste, Nova Prata do Iguaçu, Ouro Verde do Oeste, Pérola D´Oeste, Pinhal do São Bento, Planalto, Pranchita, Quedas do Iguaçu, Realeza, Renascença, Salgado Filho, Salto do Lontra, Santa Izabel do Oeste, Santa Lúcia, Santa Tereza do Oeste, Santo Antônio do Sudoeste, São João, São Jorge D´Oeste, São Pedro do Iguaçu, Toledo, Três Barras do Paraná, Tupãssi, Ubiratã, Verê e Vitorino, todos do Estado do Paraná

Fonte: Relação das Ações Civis Públicas[21] constante no site do INSS, com adaptações

[21] Acessado em 22 de fevereiro de 2020 por meio do link: https://www.inss.gov.br/wp-content/uploads/2019/09/Rela%C3%A7%C3%A3o-de-ACPs-de-BPC.pdf

APÊNDICE A – Descrição dos atributos da Maciça

A Maciça contém 148 atributos cuja descrição consta no quadro abaixo

Quadro 9: Descrição dos atributos da Maciça

Atributo	Formato	Descrição
NU_NB	9(10)	Número do benefício
ID_OL_CONCESSAO	9(08)	Identificador do órgão concessor - conforme tabela corporativa Uos TB0700
ID_OL_MANUTENCAO	9(08)	Identificador do órgão mantenedor - conforme tabela corporativa Uos TB0700
ID_OL_MANUT_ANT	9(08)	Identificador do órgão mantenedor anterior - conforme tabela corporativa Uos TB0700
CS_PA	9(01)	Classificador Pensão Alimentícia: 0 = Sem pensão, 1 = Com pensão (tit. ben. Próprio), 3 = Com pensão (tit. Pensão)
VL_MR_ATU	9(12)V99	Valor da mensalidade reajustada atual
VL_RMI	9(12)V99	Valor da renda mensal inicial
CS_TRATAMENTO	9(02)	Tratamento do benefício
CS_ESPECIE	9(02)	Espécie do benefício
CS_RAMO_ATIVIDADE	9(01)	Ramo de atividade
CS_FORMA_FILIACAO	9(01)	Forma de filiação
CS_DOC_EMPREGADOR	9(01)	Tipo de documento do empregador: 1 = CNPJ/CGC, 3 = CEI, 5 = CPF, 7 = NIT
NU_DOC_EMPREGADOR	9(14)	Número do documento do empregador
NU_NB_ANT	9(10)	Número do benefício anterior
D2_DER	9(08)	Data de entrada do requerimento
D2_DIB	9(08)	Data de início do benefício
D2_DDB	9(08)	Data do despacho do benefício
D2_DCB	9(08)	Data de cessação do benefício
D2_DIP	9(08)	Data do início do pagamento
D2_INI_INCAPAC	9(08)	Data de início da incapacidade
D2_INICIO_DOENCA	9(08)	Data de início da doença
D2_DRD	9(08)	Data da regularização da documentação
D2_OBITO_RECLUSAO	9(08)	Data do óbito ou reclusão
CS_CLIENTELA	X(01)	Classificador de urbano e rural: U = Urbano, R = Rural
NU_MATR_CONCESSOR	9(08)	Número de matrícula do concessor
NU_MATR_HABILITADOR	9(08)	Número de matrícula do habilitador
CS_SITUACAO_BENEF	9(02)	Classificador da situação do benefício
ID_BANCO	9(03)	Cbc do banco
ID_ORGAO_PAGADOR	9(06)	Sinônimo bancário (cod. Agência Prev.)
CS_MEIO_PAGTO	9(02)	Classificador de meio de pagamento
NU_AGENCIA_PAG	9(06)	Número da agência pagadora
NU_CONTA_CORRENTE	X(10)	Número da conta-corrente
CS_DIAGNOSTICO_N	X(06)	Código da doença
CS_DIAGNOSTICO_1	X(06)	Código da doença
NU_MATR_MP1	9(07)	Número de matrícula do médico perito
NU_MATR_MP2	9(07)	Número de matrícula do médico perito
D2_FORMAT_CONC	9(08)	Data da formatação da concessão
CS_DESPACHO	9(02)	Classificador de despacho
DT_DIA_UTIL_PAGTO	9(02)	Dia útil da efetivação do pagamento
NM_RECEBEDOR	X(28)	Nome do recebedor
DN_RECEBEDOR	9(08)	Data de nascimento do recebedor

NU_CPF_R	9(11)	Número de CPF do recebedor
CS_SEXO_R	9(01)	Sexo do recebedor: 1 = Masculino, 2 = Feminino, 9 = Não informado
NM_TITULAR_BENEF_T	X(40)	Nome do titular do benefício
NM_MAE_T	X(32)	Nome da mãe do titular do benefício
NU_CPF_T	9(11)	Número de CPF do titular do benefício
ID_NIT_T	9(11)	Número de NIT do titular do benefício
DT_NASCIMENTO_T	9(08)	Data de nascimento do titular do benefício
CTPS_T	9(07)	Número da CTPS do titular do benefício
CTPS_SERIE_T	9(05)	Número série CTPS do titular do benefício
CTPS_UF_T	X(02)	UF da CTPS do titular do benefício
NU_IDENTIDADE_T	X(14)	Número de identidade do titular do benefício
IDENTIDADE_UF_T	X(02)	UF da identidade do titular do benefício
CS_EMISSOR_T	9(02)	Órgão emissor da identidade do titular do benefício
NU_TIT_ELEITOR	9(13)	Número do título de eleitor do titular do benefício
CS_VAL_CNIS	9(02)	Classificador de validade do NIT do titular no CNIS (
CS_SEXO_T	9(01)	Sexo do titular do benefício: 1 = Masculino, 2 = Feminino, 9 = Não informado
TE_ENDERECO_T	X(40)	Endereço do titular (logradouro e número)
NM_BAIRRO_T	X(17)	Nome do bairro do titular do benefício
NU_CEP_T	9(08)	CEP do titular do benefício
NU_DDD_T	X(04)	DDD do titular do benefício
NU_TELEFONE_T	X(08)	Número de telefone do titular do benefício
ID_MUN_SINPAS_T	9(05)	Município de nascimento do titular
ID_MUN_IBGE_T	9(06)	Município de nascimento do titular
NM_MUNICIPIO_T	X(24)	Nome município do titular do benefício
NM_UF_MUNICIPIO_T	X(02)	UF do município do Titular do benefício
D2_OBITO_T	9(08)	Data de óbito do titular
NM_INSTITUIDOR_I	X(40)	Nome do Instituidor do Benefício
NM_MAE_I	X(32)	Nome da mãe do instituidor do benefício
NU_CPF_I	9(11)	CPF do instituidor do benefício
ID_NIT_I	9(11)	NIT do instituidor do benefício
DT_NASCIMENTO_I	9(08)	Data de nascimento do instituidor do benefício
CTPS_I	9(07)	Número da CTPS do instituidor do benefício
CTPS_SERIE_I	9(05)	Número de série da CTPS do instituidor do benefício
CTPS_UF_I	X(02)	UF da CTPS do instituidor do benefício
NU_IDENTIDADE_I	X(14)	Número de identidade do instituidor do benefício
IDENTIDADE_UF_I	X(02)	UF da identidade do instituidor do benefício
CS_EMISSOR_I	9(02)	Órgão emissor da identidade do instituidor do benefício
NU_TIT_ELEITOR_I	9(13)	Número do título de eleitor do instituidor
CS_VAL_CNIS_I	9(02)	Classificador de validade do Nit do instituidor no CNIS
CS_SEXO_I	9(01)	Sexo do segurado instituidor: 1 = Masculino, 2 = Feminino, 9 = Não informado
D2_OBITO_I	9(08)	Data de óbito do segurado instituidor
NM_PROCURADOR_P	X(70)	Nome do procurador
NM_MAE_P	X(70)	Nome da mãe do procurador
NU_CPF_P	9(11)	CPF do procurador
ID_NIT_P	9(11)	Nit do procurador
DT_NASCIMENTO_P	9(08)	Data de nascimento do procurador
CTPS_P	9(07)	Número da CTPS do procurador
CTPS_SERIE_P	9(05)	Número de série da CTPS do procurador
CTPS_UF_P	X(02)	UF da CTPS do procurador

NU_IDENTIDADE_P	X(14)	Número da identidade do procurador
IDENTIDADE_UF_P	X(02)	UF da identidade do procurador
CS_EMISSOR_P	9(02)	Órgão emissor da identidade do procurador
CS_SEXO_P	9(01)	Sexo do procurador: 1 = Masculino, 2 = Feminino, 9 = Não informado
NM_BAIRRO_P	X(17)	Nome do bairro do procurador
NU_CEP_P	9(08)	Número do CEP do procurador
TE_ENDERECO_P	X(40)	Endereço do procurador
NM_MUNICIPIO_P	X(40)	Nome do município do procurador
NM_UF_MUNICIPIO_P	X(02)	Sigla da UF do município
MUNICIP_NASC_P	9(06)	Município de Nascimento do Procurador
NM_REPRESENTANTE_R	X(40)	Nome do representante legal do benefício
NM_MAE_R	X(32)	Nome da mãe do representante legal
ID_NIT_R	9(11)	Nit do representante legal
DT_NASCIMENTO_R	9(08)	Data de nascimento do representante legal
CTPS_R	9(07)	Número da CTPS do representante legal
CTPS_SERIE_R	9(05)	Série da CTPS do representante legal
CTPS_UF_R	X(02)	UF da CTPS do representante legal
NU_IDENTIDADE_R	X(14)	Número de identidade. do representante legal
IDENTIDADE_UF_R	X(02)	UF da identidade do representante legal
CS_EMISSOR_R	9(02)	Órgão emissor identidade do repres. legal
CS_TIPO_R	9(01)	Classificador do tipo de representante
QT_DEP_IR	9(02)	Quantidade de dependentes no I.R.
QT_DEP_VAL_NB	9(02)	Quantidade de dependentes válidos no benefício
QT_DEP_CADASTRO	9(02)	Quantidade de dependentes cadastrados no benefício
QT_RUBRICA_REG	9(03)	Quantidade de rubricas registradas
CS_RUBRICA_1	9(03)	Descrição da rubrica 1
CS_RUBRICA_2	9(03)	Descrição da rubrica 2
CS_RUBRICA_3	9(03)	Descrição da rubrica 3
CS_RUBRICA_4	9(03)	Descrição da rubrica 4
CS_RUBRICA_5	9(03)	Descrição da rubrica 5
CS_RUBRICA_6	9(03)	Descrição da rubrica 6
CS_RUBRICA_7	9(03)	Descrição da rubrica 7
CS_RUBRICA_8	9(03)	Descrição da rubrica 8
CS_RUBRICA_9	9(03)	Descrição da rubrica 9
CS_RUBRICA_10	9(03)	Descrição da rubrica 10
VL_RUBRICA_1	9(12)V99	Valor da rubrica 1
VL_RUBRICA_2	9(12)V99	Valor da rubrica 2
VL_RUBRICA_3	9(12)V99	Valor da rubrica 3
VL_RUBRICA_4	9(12)V99	Valor da rubrica 4
VL_RUBRICA_5	9(12)V99	Valor da rubrica 5
VL_RUBRICA_6	9(12)V99	Valor da rubrica 6
VL_RUBRICA_7	9(12)V99	Valor da rubrica 7
VL_RUBRICA_8	9(12)V99	Valor da rubrica 8
VL_RUBRICA_9	9(12)V99	Valor da rubrica 9
VL_RUBRICA_10	9(12)V99	Valor da rubrica 10
VL_BRUTO	9(12)V99	Valor da rubrica 11
TOT_DESCONTOS	9(12)V99	Valor da rubrica 12
VL_LIQUIDO	9(12)V99	Valor da rubrica 13

NU_CPF	9(09)	CPF do representante legal
NU_CPF_DV	9(02)	*inserir acima
CS_SEXO	9(01)	Sexo representante legal: 1 = Masculino, 2 = Feminino, 9 = Não informado
DT_ULTIMA_ALTER	9(08)	Data da última alteração do arquivo utilizado para fazer a carga dos dados
D2_LIMITE	9(08)	
ANO_MES_REF	9(06)	Ano e mês do pagamento da folha da Maciça

Observação		Completar com "0"s à esquerda