Instituto Serzedello Corrêa – ISC
Pós-Graduação em Análise de dados para o Controle _____

Leandro Rangel Santos

# Aprendizado de máquina para identificação de fraudes em pensão no INSS

Orientador(a):
Prof. Msc. Gustavo Cordeiro  Galvão Van Erven

**Brasília**
**2019**

Leandro Rangel Santos

# Aprendizado de máquina para identificação de fraudes em pensão no INSS

Orientador(a):

Prof. Msc. Gustavo Cordeiro  Galvão Van Erven

Trabalho de conclusão de curso submetido ao Instituto Serzedello Corrêa do Tribunal de Contas da União como requisito para a obtenção do grau de especialista.

**Brasília**
**2019**

Leandro Rangel Santos

# Aprendizado de máquina para identificação de fraudes em pensão no INSS

Trabalho de conclusão de curso submetido ao Instituto Serzedello Corrêa do Tribunal de Contas da União como requisito para a obtenção do grau de especialista.

**Data de aprovação:**

__ / __ / ____

**Banca examinadora:**

_____

Prof. Dr.

_____

Prof. Dr.

**Brasília**
**2019**

# Agradecimentos

Agradeço a minha família.

# Resumo

Em resposta aos novos desafios impostos pela sociedade, a Controladoria Geral da União (CGU) tem investido no aperfeiçoamento dos processos e métodos de trabalho visando o controle mais efetivo da aplicação dos recursos públicos, para atender esses desafios as atividades de auditoria contínua vêm crescendo na CGU. Essas atividades de auditoria contínua nos últimos anos foram focadas na identificação de tipologias que apresentam indícios de irregularidades, porém é notada a carência de trabalhos de identificação de fraudes. Atrelado a isso, existe também a carência de auditores e técnicas para efetivamente fazer o combate a fraudes em âmbito nacional, principalmente nas fraudes de benefícios previdenciários, os quais representam mensalmente 35 milhões de benefícios, com uma materialidade mensal superior a R$ 49 bilhões. Com o quantitativo reduzido de auditores qualificados e inúmeras responsabilidades, a auditoria contínua precisa de métodos automatizados aplicáveis a grandes volumes de dados buscando assim a priorização e atuação eficaz. Esse trabalho apresenta a aplicação de aprendizado de máquina para gerar um modelo preditivo de fraude em pensões por morte no INSS. O modelo construído ao final desse trabalho obteve resultados satisfatórios de aproximadamente 92% de recall, 94% de precisão, 93% de acurácia e 92% de F-Score.

**Palavras-chave:** Fraude; Aprendizado de máquina; Auditoria contínua; Previdência Social.

# Abstract

In response to the new challenges imposed by society, the Brazilian Office of the Comptroller General (CGU) have invested in the improvement of processes and work methods aimed at more effective control of the application of public resources, to meet these challenges continuous audit activities have been growing in CGU. These continuous audit activities in recent years has focused on the identification of typologies that show signs of irregularities, but there is a lack of fraud identification work. Linked to this, there is also the lack of auditors and techniques to effectively combat fraud at a national level, mainly in frauds of social security benefits, which represents 35 million benefits per month, witch represents more than R$ 49 billions per month. With the reduced number of qualified auditors and numerous responsibilities, continuous auditing requires automated methods that fits to large volumes of data, thus seeking to prioritize and act effectively. This paper presents the application of machine learning to generate a predictive model of fraud in death pensions in the Brazilian social security. The model constructed at the end of this paper obtained satisfactory results of approximately 92% por recall, 94% of precision, 93% of accuracy and 92% F-Score.



**Keywords**: Fraud; Machine Learning; Continuous auditing; Social Security

# Lista de figuras

**Figura 1:** Fases do CRISP-DM .....18
**Figura 2:** Matriz de confusão .....20
**Figura 3:** Evolução do Mean test score com o acréscimo de variáveis no modelo .....30

# Lista de tabelas

**Tabela 1:** Exemplo de colunas da base da Maciça.....25
**Tabela 2:** Resultado do modelo gerado .....27
**Tabela 3:** Importância das variáveis da Baseline 1.....27
**Tabela 4:** Resultado do modelo gerado .....31
**Tabela 5:** Performance dos distintos modelos de classificação .....31
**Tabela 6:** Performance dos modelos antes e depois dos ajustes dos hiper parâmetros .....32
**Tabela 7:** Resultado do modelo gerado .....33

**SUMÁRIO**

**1. INTRODUÇÃO ..... 10**

**2. PROBLEMA E JUSTIFICATIVA ..... 12**

**3. OBJETIVOS ..... 14**

3.1. OBJETIVO GERAL ..... 14

3.2. OBJETIVOS ESPECÍFICOS ..... 14

**4. METODOLOGIA ..... 15**

**5. FUNDAMENTAÇÃO TEÓRICA ..... 17**

**6. DESENVOLVIMENTO ..... 22**

**6.1. ENTENDIMENTO DO NEGÓCIO ..... 22**

**6.2. ENTENDIMENTO DOS DADOS ..... 24**

**6.3. PREPARAÇÃO DOS DADOS ..... 25**

**6.4. MODELAGEM ..... 26**

**6.5. AVALIAÇÃO ..... 33**

**7. CONSIDERAÇÕES FINAIS ..... 34**

# 1. Introdução

A Previdência Social é um dos ramos da Seguridade Social do país e tem por objetivo amparar os trabalhadores de certos riscos sociais conforme definidos na Constituição Federal. O Instituto Nacional do Seguro Social (INSS), operacionalizador das políticas de seguridade social do pais é uma autarquia federal vinculada atualmente ao Ministério da Economia, e foi instituído pelo Decreto nº 99;350/1990, oriundo da fusão do Instituto de Administração Financeira da Previdência e Assistência Social (IAPAS) com o Instituto Nacional da Previdência Social (INPS) (BRASIL,1990a ;1990b)

A pensão por morte é um dos benefícios concedidos pelo INSS, e visa assegurar os dependentes do segurado que veio a óbito, de forma que os mesmos não fiquem desassistidos após o falecimento do segurado, ela é devida aos dependentes do segurado, aposentado ou não, que falece.

Praticamente qualquer organização, como órgão, autarquia, empresa pública, sociedades de economia mista, parcerias público-privadas, fundações, organizações sociais, fundos de pensão etc. está sob risco de fraude e corrupção, bastando para tal a existência de recursos públicos disponíveis para atrair a cobiça dessas máfias (BRASIL, 2016).

Conforme assinala o Tribunal de Contas da União (BRASIL, 2018), as fraudes dentro do INSS correspondem a pelo menos 11.41% de tudo que é pago pelo instituto, isso representou, em 2018, um prejuízo potencial superior a R$ 64 bilhões.

Apesar dos números serem preocupantes, há aspectos positivos a serem observados. Face o número de benefícios a serem analisados, e as modernas técnicas que são criadas, os órgãos de controle passaram a recorrer mais a ferramentas de tecnologia da informação para responder a altura os desafios que lhe são impostos. Tais iniciativas permitiram aumentar o espetro de atuação dos órgãos, e vêm trazendo resultados promissores. Antigamente seria inviável imaginar uma análise de todos os benefícios do INSS, mas atualmente muitas questões de auditoria podem ser realizadas por processos de trabalho que fazem uso de tecnologia da informação. Como isso, os órgãos de controle vêm adotando mais tais práticas.

Outro fator que corrobora com o uso de tecnologia da informação nas ações de controle é a maior facilidade de acesso as informações. A adoção dos dispositivos

constantes da Lei nº 12.527, de 18 de novembro de 2011, lei de acesso a informação, fomentou na administração pública federal a transparência ativa, onde os órgãos e entidades são obrigados a divulgar na internet, um conjunto mínimo de dados referentes a sua gestão.

# 2. Problema e justificativa

A Controladoria-Geral da União (CGU) está estruturada em cinco unidades finalísticas, que atuam de forma articulada, em ações organizadas entre si: Secretaria de Transparência e Prevenção da Corrupção (STPC), Secretaria Federal de Controle Interno (SFC), Corregedoria-Geral da União (CRG), Secretaria de Combate à Corrupção (SCC) e Ouvidoria-Geral da União (OGU).

A Secretaria Federal de Controle Interno (SFC) exerce as atividades de órgão central do sistema de controle interno do Poder Executivo Federal. Nesta condição, fiscaliza e avalia a execução de programas de governo, inclusive ações descentralizadas a entes públicos e privados realizadas com recursos oriundos dos orçamentos da União; realiza auditorias e avalia os resultados da gestão dos administradores públicos federais; apura denúncias e representações; exerce o controle das operações de crédito; e, também, executa atividades de apoio ao controle externo.

Para subsidiar as auditorias relacionadas ao tema previdenciário, a SFC conta com a Diretoria de Auditoria de Previdência e Benefícios, criada pelo Decreto nº 9.681/2019, unidade responsável pela auditoria de toda a área previdenciária, bem como todo e qualquer benefício social pago no âmbito da União.

A DPB é dividida em seis divisões, uma delas é a divisão de auditoria contínua, a qual é responsável por levantar informações sobre quaisquer dados previdenciários e de outros benefícios com o intuito de realizar auditoria contínua, o que nas palavras de Alles et al. (2006) pode ser definida como:

> A auditoria contínua estende os métodos analíticos de auditoria tradicional por meio do exame de fluxos contínuos de dados, com modelos de comportamento dos sistemas usados para estabelecer expectativas para o conteúdo dos dados.

Os trabalhos da auditoria contínua nos últimos anos foram focados na identificação de tipologias que apresentam indícios de irregularidades, porém é notada a carência de trabalhos de identificação de fraudes. Atrelado a isso, existe também a carência de servidores e recursos para efetivamente fazer o combate a fraudes em âmbito nacional, principalmente nas fraudes de benefícios previdenciários, os quais representam mensalmente 35 milhões de benefícios, com uma materialidade mensal superior a R$ 49 bilhões.

A avaliação de fraudes em pensão por morte no INSS é uma atividade complexa. Atualmente o papel é desempenhado principalmente pela Coordenação-Geral de Inteligência Previdenciária (COINP), ligada a Secretaria Especial de Previdência e Trabalho, juntamente com a Polícia Federal. Uma das técnicas utilizadas para a identificação de fraudes é a análise de fraudes documentais, ou seja, documentos feitos com a intenção de ludibriar o agente público, com objetivo de obter um benefício indevido por meio da apresentação de documentos fabricados ou adulterados.

Infelizmente, a análise documental é um processo custoso, que demanda tempo e servidores especializados. Um dos meios pelos quais se pode agilizar o processo de identificação de concessões fraudulentas é através da visão computacional aliada à aprendizagem de máquina, duas áreas que entraram em plena convergência nos últimos anos. Com base nisso, e considerando a evolução das técnicas de aprendizado de máquina, bem como a disponibilidade de dados para uso dos órgãos de controle, identificou-se o seguinte problema de pesquisa: **É possível identificar fraudes em pensão por morte no INSS através de aprendizado de máquina?**

# 3. Objetivos

Para responder o problema de pesquisa definido acima, definem-se os objetivos, classificando-os em:

## 3.1. Objetivo geral

Criar um modelo preditivo para identificação de concessões de pensão por morte no INSS que possuem alta probabilidade de serem uma concessão fraudulenta.

## 3.2. Objetivos específicos

- Identificar e selecionar atributos relevantes para a identificação da fraude
- Avaliar modelos gerados a partir dos diversos algoritmos de classificação

# 4. Metodologia

O trabalho foi executado utilizando o processo CRISP-DM (CHAPMAN,2000), permeando, de forma resumida as seguintes etapas:

- Entendimento do negócio: Buscando criar um produto que fosse atender as necessidades da CGU, a ideia do pré-projeto que visava criar um modelo preditivo de identificação de fraudes em pensão no INSS foi submetido para a área responsável na casa, e o pré-projeto foi aprovado como tema correlato as atividades da casa.

- Entendimento dos dados: Os dados previdenciários já eram coletados pela CGU e recebiam o devido tratamento, portanto essa etapa foi resumida. Os dados coletados são armazenados e acessados usando a ferramenta SQL Server Management Studio[1] .

- Preparação dos dados: Uma vez de posse dos dados, foram realizadas algumas atividades de eliminação de outliers, eliminação de dados nulos, criação de novos dados derivados, com o intuito de criar um conjunto de dados aptos a serem passados para o modelo. Essas atividades foram realizadas usando a linguagem Python com uso da biblioteca Pandas[2] no editor Jupyter Notebook[3].

- Modelagem: Para a criação desse trabalho foram testados 6 modelos distintos de classificação, todos da biblioteca scikit-learn[4].

- Avaliação: A avaliação dos modelos consistiu no uso da matriz de confusão.

- Aplicação: Com o modelo construído, os casos apontados como alta probabilidade de serem fraudes serão discutidos com especialistas para a geração de um modelo que possa ser implantado na identificação real de casos de fraude.

---

[1] SQL Server Management Studio. Link: https://www.microsoft.com/pt-br/download/details.aspx?id=29062

[2] Disponível em https://pandas.pydata.org/index.html

[3] Disponível em http://jupyter.org/

[4] Disponível em https://scikit-learn.org

- Implantação: Esta fase busca colocar em prática o conhecimento adquirido, e ela será executada através da apresentação dos resultados as partes interessadas, tais como a auditoria geral do INSS, a COINPE e a Polícia Federal.

# 5. Fundamentação teórica

Nesse item são apresentados os principais referenciais teóricos e os trabalhos acadêmicos relacionados.

## 5.1. Mineração de Dados

A mineração de dados é o processo de descobrir padrões perspicazes, interessantes e novos, bem como modelos descritivos, compreensíveis e preditivos de dados em grande escala (ZAKI,MEIRA 2013).

A identificação de um problema utilizando mineração de dados é implementada em duas etapas, o pré-processamento e o pós-processamento. O pré-processamento é a fase responsável por transformar os dados brutos em um formato mais adequado a mineração de dados. O pós-processamento é a fase responsável pela avaliação do modelo de mineração de dados criado, com vistas a interpretar os modelos, padrões, e confirmar as hipóteses.

Em linhas gerais, a mineração de dados comporta as técnicas de análise exploratória de dados, descoberta de padrões frequentes, agrupamento e modelos de classificação.

A análise exploratória de dados visa explorar os atributos numéricos e categóricos de dados, individualizados ou em conjunto, para extrair características chave da amostra de dados (ZAKI,MEIRA 2013). A descoberta de padrões frequentes se refere a tarefa de extrair padrões informativos e úteis em conjuntos de dados massivos e complexos. Agrupamentos, ou clustering, é a tarefa de particionar os pontos em grupos naturais, chamados clusters, de tal forma que os pontos dentro de um grupo são muito semelhantes, enquanto pontos entre os clusters são tão diferentes quanto possível. A tarefa de classificação é prever o rótulo ou classe para um produto não classificado (ZAKI,MEIRA 2013). Existem inúmeros algoritmos que podem ser usados para classificação, como por exemplo o naive bayes, o classificador de vizinhos mais próximos e árvores de decisão.

## 5.2. CRISP-DM

O CRISP-DM (Cross Industry Standard Process for Data Mining) é um modelo de referência para processos de mineração de dados, consistindo de seis fases, apresentadas abaixo:

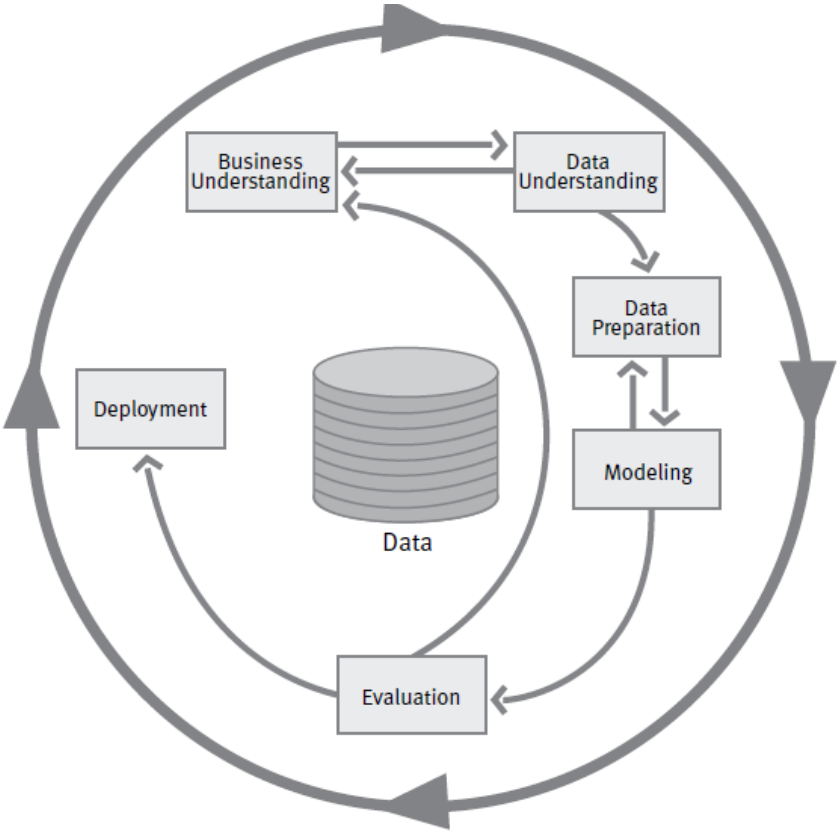


**Figura 1:** Fases do CRISP-DM

Fonte: (CHAPMAN, 2000)

**Entendimento do negócio:**

Esta fase inicial se concentra na compreensão dos objetivos e requisitos do projeto de uma perspectiva de negócios, em seguida, converter esse conhecimento em uma definição de problema de mineração de dados e um plano preliminar projetado para alcançar os objetivos.

**Entendimento dos dados:**

A fase de compreensão dos dados começa com a coleta de dados inicial e prossegue com atividades que permitem familiarizar-se com os dados, identificar problemas de qualidade de dados, descobrir os primeiros insights sobre os dados e/ou detectar subconjuntos interessantes para formar hipóteses sobre informações ocultas.

**Preparação dos dados:**

A fase de preparação de dados abrange todas as atividades necessárias para construir o dataset inicial (conjunto de dados que serão introduzidos na ferramenta de modelagem). É provável que a fase de preparação dos dados seja executada várias vezes e não em qualquer ordem prescrita. As tarefas incluem seleção de dados, bem como a transformação e limpeza dos dados para a ferramenta de modelagem.

**Modelagem:**

Nesta fase, várias técnicas de modelagem são selecionadas e aplicadas e seus parâmetros são calibrados. Normalmente, existem várias técnicas para o mesmo tipo de problema de mineração de dados. Algumas técnicas possuem requisitos sobre a forma de dados. Portanto, voltar para a fase de preparação de dados é muitas vezes necessário.

**Avaliação:**

Neste estágio do projeto, um modelo de alta qualidade de análise de dados já foi construído. Antes de prosseguir para a implantação final do modelo, é importante avaliá-lo completamente e revisar as etapas que foram executadas para cria-lo, para ter certeza de que o modelo atinge adequadamente os objetivos de negócio.

**Implantação:**

A criação do modelo geralmente não é o final do projeto. Mesmo que o propósito do modelo seja aumentar o conhecimento dos dados, o conhecimento adquirido deverá ser organizado e apresentado de forma que o cliente possa utilizá-lo. A fase de implantação pode ser tão simples quanto gerar um relatório ou tão completa quanto implementar uma mineração de dados repetida e contínua por toda a empresa.

## 5.3. Métricas de validação

A matriz de confusão é uma maneira conveniente de avaliar um modelo de classificação. Ela consiste de uma tabela que contém a quantidade de elementos que foram classificados de forma correta e de forma errada. A imagem abaixo exemplifica uma matriz de confusão:

| | **True Class** | |
|---|---|---|
| **Predicted Class** | Positive ($c_1$) | Negative ($c_2$) |
| Positive ($c_1$) | True Positive ($TP$) | False Positive ($FP$) |
| Negative ($c_2$) | False Negative ($FN$) | True Negative ($TN$) |

**Figura 2:** Matriz de confusão

Fonte: (ZAKI,MEIRA 2013)

Conforme ZAKI e MEIRA, quando o modelo de classificação possui apenas duas classes, a matriz de confusão é dita binária, e os campos recebem nomes especiais, conforme definidos abaixo:

Verdadeiro Positivos (TP): Total de elementos que o modelo classificou corretamente como positivo.

Falso Positivo (FP): Total de elementos que o modelo previu que seriam positivos, mas de fato pertencem a classe negativa.

Falso Negativo (FN): Total de elementos que o modelo previu que seriam negativos, mas de fato pertencem a classe positiva.

Verdadeiro Negativo (TN): Total de elementos que o modelo classificou corretamente como negativo.

n: Total de elementos classificados.

Com base na matriz de confusão, foram derivadas algumas métricas de avaliação, as quais definimos abaixo:

Acurácia é a fração das predições corretas.

$$Acurácia = \frac{TP + TN}{n}$$

Precisão é o total de positivos previstos pelo modelo

$$Precisão = \frac{TP}{TP + FP}$$

Sensibilidade, também chamada de recall, é a fração correta do total de positivos corretamente previstos como positivos

$$recall = \frac{TP}{TP + FN}$$

Em adição as métricas acima, será utilizado também o F-Score para avaliar a qualidade do modelo. O F-Score é calculado da seguinte forma:

$$F - SCORE = \frac{2 * (recall * precisão)}{(recall + precisão)}$$

# 6. Desenvolvimento

O desenvolvimento desse trabalho foi executado seguindo as etapas do CRISP-DM, e foram realizadas duas iterações.

### Primeira Baseline

## 6.1. Entendimento do negócio

Como órgão de controle interno, a CGU possui como missão:

”Promover o aperfeiçoamento e a transparência da gestão pública, a prevenção e o combate à corrupção, com participação social, por meio da avaliação e controle das políticas públicas e da qualidade do gasto.”

Corroborando com a missão institucional da casa, a Diretoria de Auditoria da Área de Previdência e Benefícios - DPB tem por atribuição realizar o controle interno das áreas de previdência, em especial o Instituto Nacional do Seguro Social - INSS.

A DPB é a área técnica da CGU responsável pelo acompanhamento do Sistema Previdenciário Brasileiro, que pode ser classificado da seguinte forma:

**Regime Geral de Previdência Social (RGPS):** operacionalizado pelo INSS, possui gestão pública e corresponde à previdência social geral de todos os trabalhadores de filiação obrigatória e dos segurados facultativos.

**Regime de Previdência Privada Complementar (RPPC):** possui gestão privada e a adesão é facultativa com natureza contributiva. O objetivo desse Regime é incrementar a renda do participante no momento da aposentadoria. É um complemento à renda da previdência social básica.

**Regime Próprio do Servidor Público (RPPS):** é a previdência dos servidores públicos dos entes (União, Estados e Municípios). Também tem gestão pública e a filiação nesse regime é obrigatória para os servidores.

O Instituto Nacional do Seguro Social (INSS), autarquia federal vinculada atualmente ao Ministério da Economia, foi instituído pelo Decreto nº 99.350/1990, por

meio da fusão do Instituto de Administração Financeira da Previdência e Assistência Social (IAPAS) com o Instituto Nacional de Previdência Social (INPS), fundamentado no disposto no Art. 17 da Lei nº 8.029/1990. Atualmente possui cerca de 34.000 servidores ativos.

O INSS [5]tem como objetivo prestar um bom atendimento aos cidadãos que buscam por serviços previdenciários, assistenciais ou alguns benefícios trabalhistas. Os demandantes dos serviços do INSS são oriundos de todas as faixas sociais, etárias e com os mais variados graus de escolaridade. O citado instituto possui como finalidade primordial a concessão e manutenção de benefícios previdenciários. Esta autarquia mantém o pagamento de 34 milhões de benefícios ao mês, com uma folha de pagamento em 2017 com valor superior a R$ 560 bilhões anuais, sendo, portanto, atividade de relevante materialidade às contas públicas.

Apesar da enorme missão atribuída a DPB, a força de trabalho atual está nitidamente reduzida. Sabendo da escassez de auditores para realizar a missão institucional, a DPB está investindo cada vez mais em auditorias utilizando recursos tecnológicos a fim de maximizar a abrangência da fiscalização, bem como tentar entregar mais resultado com menos recursos.

Face ao corpo escasso de auditores, e a elevada informatização dos processos do INSS, os trabalhos de auditoria de dados tem por objetivo permear as espécies de benefícios pagos, a afim de identificar, através de análise de bases de dados, benefícios que são pagos irregularmente.

O critério de sucesso para o negócio é que uma auditoria de dados consiga, com poucos recursos envolvidos, apontar pagamentos indevidos, para que o INSS possa ser notificado e providencie a cessação dos mesmos.

Face a existência de inúmeros tipos de benefícios diferentes, esse primeiro exercício tem por objetivos identificar, através de análise de dados, pensões que são concedidas com base em matrimônios fraudulentos, que são apresentados ao INSS

[5] www.inss.gov.br

apenas com a finalidade de gerar um benefício a um instituidor que não deixou dependentes.

Para identificar as pensões concedidas com base em matrimônios fraudulentos, estão planejados os seguintes passos:

1 – Seleção dos benefícios de pensão por morte

2 – Obter junto a órgãos parceiros (Policia Federal, Auditoria Interna do INSS, Inteligência previdenciária) os benefícios dos últimos 5 anos que foram cancelados por fraude documental.

3 – Criar um modelo preditivo, treinando esse modelo com base nos casos dos benefícios que foram cancelados por fraude em certidão de nascimento

4 – Com base em alguns casos apontados pelo modelo, selecionar uma amostra, e solicitar os processos físicos dessa amostra

5 – Proceder com a perícia das certidões de casamento a fim de testar a assertividade do modelo.

As etapas 4 e 5 citadas acima serão realizadas fora do âmbito deste trabalho, e serão realizadas em ações de controle da CGU.

## 6.2. Entendimento dos dados

Como fonte primária dos dados, o trabalho utilizou a base da Maciça, que representa a folha de pagamento de benefícios previdenciários.

Os dados referentes as pensões fraudulentas foram disponibilizadas pela Coordenação Geral de Inteligência Previdenciária (Coinp), área da Secretaria de Previdência responsável por combater crimes contra o sistema previdenciário. Foram disponibilizados 1.839 números de benefícios que foram cessados por fraude documental.

Com os números de benefícios fornecidos, buscamos no histórico da Maciça os dados cadastrais desses benefícios de foram cancelados.

Somado a esses benefícios fraudados, foram adicionados 1.839 benefícios de pensão não fraudadas, constantes da Maciça, para gerar os dados do modelo.

A Maciça possui 145 colunas, e aproximadamente 45 milhões de registros por mês, que representam o pagamento de toda a folha previdenciária brasileira.

A tabela a seguir ilustra algumas colunas contidas na base da Maciça:

**Tabela 1:** Exemplo de colunas da base da Maciça

| Coluna | Tipo | Comentário |
|---|---|---|
| G_OL_Concessao | String | Agência da previdência social responsável pela concessão do benefício |
| G_Clientela | String | Classificador entre urbano e rural |
| G_Despacho | Int | Código do despacho da concessão. Ex. concessão normal, concessão judicial |
| G_Especie | Int | Espécie do benefício previdenciário. Ex. aposentadoria por idade, aposentadoria por invalidez |
| G_Forma_Filiacao | Int | Classificador da forma de filiação com a atividade. Ex. Empregado, empresário, avulso. |
| G_Meio_Pagamento | Int | Classificador do meio de pagamento. Ex. Conta corrente, cartão de pagamento |
| G_VL_MRI | Numeric | Valor da remuneração inicial do benefício |
| G_VL_MR_ATU | Numeric | Valor da remuneração atual do benefício |
| G_Ramo_Atividade | Int | Classificador do ramo de atividade. Ex. Servidor, empregado, empresário |
| G_DIB | Date | Data de início do benefício |
| G_DER | Date | Data de entrada do requerimento |
| G_DDB | Date | Data de despacho do benefício |

Fonte: Base de dados da MACICA disponibilizada pelo INSS.

## 6.3. Preparação dos dados

Nessa etapa foram feitas diversas ações para prepara e tratar os dados. Primeiramente foram excluídos registros com informações sem preenchimento ou

inválidas (Datas 1900-01-01, CPF 00000000000). Muitos campos da maciça são categóricos, ou seja, eles possuem apenas valores simbólicos, e muitas das operações aritméticas não podem ser realizadas diretamente nestes valores simbólicos (ZAKI,MEIRA 2013). Para esses campos, exemplo o G_FORMA_FILIACAO, foi utilizada a técnica de codificar os valores inteiros dos campos em uma nova coluna, representando as informações como uma coluna binária, representando a existência ou não do referido valor.

Para as colunas de identificação pessoal ( NIT , CPF , RG ) foi utilizada a técnica de transformar essa coluna em uma variável binária, representando a existência ou não do campo, criando assim variáveis como possuiCPF , possuiNIT.

Para as variáveis de Data, elas foram substituídas por variáveis de diferença de tempo, o que demonstra o tempo de processamento. A diferença entre a G_DER e a G_DDB representa o quão rápido foi conferido o despacho para o benefício. A diferença entre a G_DIB e a G_DER representa quanto tempo o segurado demorou para dar entrada no benefício, após ter direito ao mesmo.

Os valores de remuneração onde a remuneração que constava da base de dados era maior de que o teto remuneratório do INSS foram considerados como outliers, e foram excluídos.

Para as colunas que representavam valores numéricos, como o G_VL_MR_ATU, foram criadas outras colunas representando as relações quadradas , cúbicas e logarítmicas entre esses valores. Nem todo dado possui uma relação linear e em algumas situações nosso modelo precisa se adaptar a uma relação mais complexa com as colunas, desta forma, foram criadas novas colunas representando o G_VL_MR_ATU ao quadrado e ao cubo, de forma prover mais variáveis para nosso modelo, e caso a relação entre essas variáveis e o target não seja linear, nosso modelo também poderá utilizar essas variáveis.

As colunas que não possuem nenhuma correlação com os benefícios de pensão, como por exemplo o número de telefone do procurador, foram excluídas.

Por fim, criamos uma coluna chamada target, que recebeu o valor 0 para os benefícios não fraudados, e valor 1 para os benefícios fraudados.

Feita a limpeza e a conversão, o modelo ficou com 182 colunas.

## 6.4. Modelagem

Nessa etapa, uma vez que conhecemos antecipadamente os benefícios que são fraudados e os que não são, optamos por utilizar modelos de classificação. Para identificarmos uma primeira modelagem, passamos os dados tratados (todas as 182 colunas) para um modelo de Random Forrest, e separamos 20% dos casos para teste, realizando o treino do modelo com 80% dos casos.

Um classificador de árvore de decisão é uma árvore recursiva baseada em predição que prevê a classe y para cada ponto de x. Random Forrest por sua vez é um algoritmo do tipo ensemble (ou composto) que se utiliza de vários outros para gerar melhores resultados (BREIMAN,2001).

O modelo de classificação do Random Forrest, o qual chamamos de Baseline 1, apresentou a seguinte performance:

**Tabela 2:** Resultado do modelo gerado

| Baseline | Recall | Precisão | Acurácia | F-Score |
|---|---|---|---|---|
| Baseline 1 | 90,27% | 95,42% | 92,90% | 91.15% |

O modelo acima utilizou todas as variáveis existentes para criar sua predição, entretanto muitas variáveis não tiveram poder preditivo algum para o modelo. Observamos abaixo as primeiras 40 variáveis utilizadas e a sua importância para o modelo, conforme o método feature_importances_[6] do scikit learn.

**Tabela 3:** Importância das variáveis da Baseline 1

| Variável | Importância |
|---|---|
| G_VL_MR_ATU | 0.18588639384410707 |
| G_VL_MR_ATU-log | 0.14186731819928217 |
| G_VL_MR_ATU-2 | 0.11554398927315837 |
| difDIBDER | 0.10054032857068587 |
| G_VL_MR_ATU-Sq | 0.08057904734091254 |
| PAG_Total_Descontos | 0.07625450530186212 |
| G_VL_MR_ATU-3 | 0.06779325473718516 |

[6] https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestClassifier.html#sklearn.ensemble.RandomForestClassifier.feature_importances_

| Variável | Importância |
|---|---|
| PossuiNBAnterior | 0.03477115054177427 |
| difDERDDB | 0.018783395594881947 |
| PAG_Qtde_Rubricas_Registrada-2 | 0.016916932266604588 |
| PAG_Qtde_Rubricas_Registrada-Sq | 0.014392742289676677 |
| G_Ramo_Atividade_8 | 0.010652334897980535 |
| PAG_Qtde_Rubricas_Registrada-3 | 0.010543162263763766 |
| G_Tratamento_1 | 0.01018980347217147 |
| D_Qtde_Validos_Benef-2 | 0.010182602578310666 |
| G_Clientela_U | 0.00969970360956726 |
| D_Qtde_Cadastrado_Benef-2 | 0.009038294860671144 |
| G_Tipo_Doc_Empregador_0 | 0.008802954198454246 |
| D_Qtde_Validos_Benef-Sq | 0.008382320502294533 |
| G_Forma_Filiacao_7 | 0.007256544785832865 |
| G_Ramo_Atividade_2 | 0.006805722750522479 |
| D_Qtde_Cadastrado_Benef-3 | 0.006700077890443671 |
| D_Qtde_Validos_Benef-3 | 0.006167346310491742 |
| G_Forma_Filiacao_8 | 0.006099190949960252 |
| G_Meio_Pagto_2 | 0.005647851510737002 |
| titularMasculino | 0.0055641941457016 |
| PossuiDocEmpregador | 0.005372692675266964 |
| G_Tipo_Doc_Empregador_1 | 0.0051420490471608985 |
| recebedorMasculino | 0.004716634782861525 |
| G_Forma_Filiacao_1 | 0.004689339621741612 |
| D_Qtde_Validos_Benef_0 | 0.004677300750382111 |
| G_Banco_341 | 0.004592354894963104 |
| PAG_Qtde_Rubricas_Registrada_3 | 0.0045503190329540255 |
| D_Qtde_Cadastrado_Benef-Sq | 0.004209586003266023 |
| G_Tratamento_81 | 0.004080623120057131 |
| RL_Tipo_0 | 0.004057203036231037 |
| I_Validade_NIT_CNIS_1 | 0.003706254643008324 |
| instituidorMasculino | 0.00334498559265664 |
| D_Qtde_Cadastrado_Benef_1 | 0.003308964631775798 |
| I_Validade_NIT_CNIS_0 | 0.0032906686282643156 |
| G_Forma_Filiacao_0 | 0.0032135027541412946 |

Observando as variáveis da tabela acima, vemos que muitas delas são correlacionadas, como por exemplo o G_VL_MR_ATU, que aparece repetidas vezes sendo representado em seu valor original, seu valor ao quadrado e seu valor ao cubo.

Removendo essas duplicidades (pegando apenas a primeira ocorrência com maior importância), temos a seguinte relação de variáveis utilizadas pela Baseline 1:

- G_VL_MR_ATU
- difDIBDER

- PAG_Total_Descontos
- PossuiNBAnterior
- difDERDDB
- PAG_Qtde_Rubricas_Registrada-2
- G_Ramo_Atividade_8
- G_Tratamento_1
- D_Qtde_Validos_Benef-2
- G_Clientela_U
- D_Qtde_Cadastrado_Benef-2
- G_Tipo_Doc_Empregador_0
- G_Forma_Filiacao_7
- G_Ramo_Atividade_2
- G_Forma_Filiacao_8
- G_Meio_Pagto_2
- titularMasculino
- PossuiDocEmpregador
- G_Tipo_Doc_Empregador_1
- recebedorMasculino
- G_Forma_Filiacao_1
- D_Qtde_Validos_Benef_0
- G_Banco_341
- PAG_Qtde_Rubricas_Registrada_3
- G_Tratamento_81
- RL_Tipo_0
- I_Validade_NIT_CNIS_1
- instituidorMasculino
- D_Qtde_Cadastrado_Benef_1
- I_Validade_NIT_CNIS_0
- G_Forma_Filiacao_0.

Com o intuito de incrementar os percentuais obtidos na Baseline 1, foi feita uma análise da quantidade de variáveis que devem ser passadas ao modelo, para isso foi utilizado a funcionalidade SelectKBest[7] do scikit learn, tal funcionalidade apresenta o

[7] https://scikit-learn.org/stable/modules/generated/sklearn.feature_selection.SelectKBest.html

resultado do modelo utilizando K variáveis, e mostra o progresso do modelo a cada acréscimo de K. A análise de variáveis do modelo é demonstrada no gráfico abaixo:

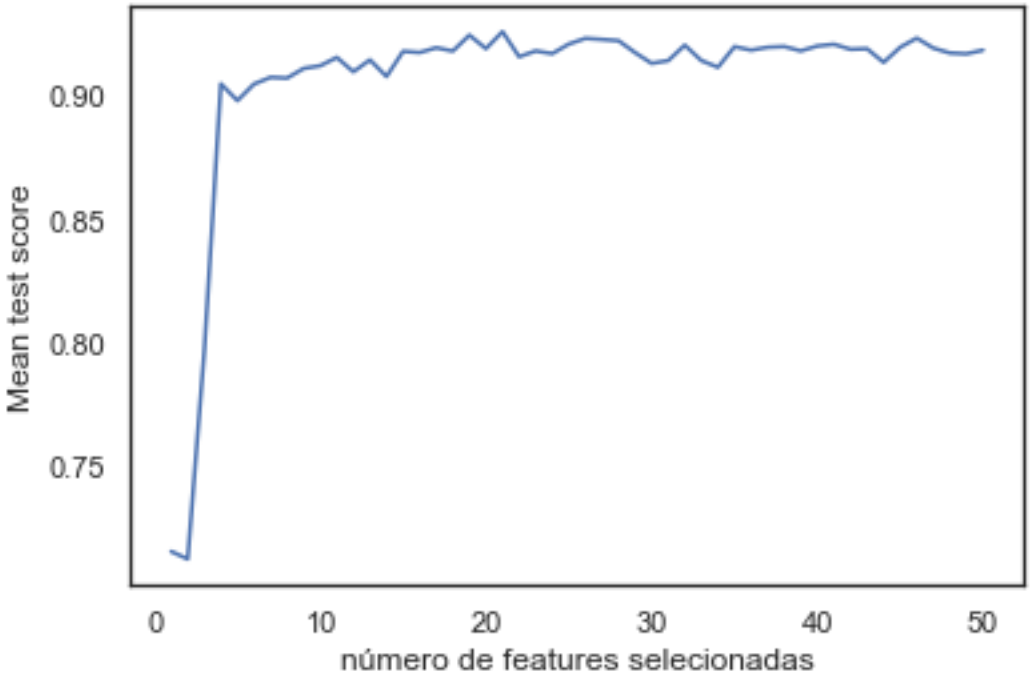


**Figura 3:** Evolução do Mean test score com o acréscimo de variáveis no modelo

Observamos pela imagem acima que a partir de 20 variáveis o modelo não ganha performance alguma, até perdendo performance em alguns casos. Portanto, vamos gerar uma nova baseline utilizando 20 variáveis, ao invés das 182 utilizadas na Baseline 1.

## Segunda Baseline

Utilizando as 20 principais variáveis:

- G_VL_MR_ATU
- difDIBDER
- PAG_Total_Descontos
- PossuiNBAnterior
- difDERDDB
- PAG_Qtde_Rubricas_Registrada-2
- G_Ramo_Atividade_8
- G_Tratamento_1
- D_Qtde_Validos_Benef-2

- G_Clientela_U
- D_Qtde_Cadastrado_Benef-2
- G_Tipo_Doc_Empregador_0
- G_Forma_Filiacao_7
- G_Ramo_Atividade_2
- G_Forma_Filiacao_8
- G_Meio_Pagto_2
- titularMasculino
- PossuiDocEmpregador
- G_Tipo_Doc_Empregador_1
- recebedorMasculino

Com isso, foi gerada uma nova Baseline, a qual demonstramos o resultado abaixo:

**Tabela 4:** Resultado do modelo gerado

| Baseline | Recall | Precisão | Acurácia | F-SCORE |
|---|---|---|---|---|
| Baseline 1 | 90,27% | 95,42% | 92,90% | 91.15% |
| Baseline 2 | 91.62% | 95.22% | 93.36% | 92.26% |

## Terceira Baseline

Uma vez que a biblioteca scikit learn oferece uma série de modelos de classificação, a terceira baseline buscou utilizar o mesmo dataset gerado na segunda baseline, mas tentando utilizar distintos modelos de classificação para ver qual apresentava a melhor performance sem ajudar ainda os hiper parâmetros dos modelos. Para avaliação da performance de cada modelo foi utilizado o valor do recall, e os modelos apresentaram o seguinte desempenho:

**Tabela 5:** Performance dos distintos modelos de classificação

| Modelo | Recall |
|---|---|
| KNeighborsClassifier | 85,94% |
| GaussianProcessClassifier | 20% |
| DecisionTreeClassifier | 92,16% |

| Modelo | Recall |
|---|---|
| RandomForestClassifier | 92.70% |
| AdaBoostClassifier | 93.24% |
| GaussianNB | 70.54% |
| QuadraticDiscriminantAnalysis | 74.59% |
| MLPClassifier | 84,32% |

Observando os resultados acima, e com objetivo de melhorar a performance do modelo, foram escolhidos os seis melhores algoritmos para fazermos uma segunda rodada com ajustes nos hiper parâmetros, utilizando a ferramenta GridSearch[8] Uma vez que a biblioteca scikit learn oferece uma série de modelos de classificação. Os seis modelos escolhidos foram:

- KNeighborsClassifier
- DecisionTreeClassifier
- RandomForestClassifier
- AdaBoostClassifier
- QuadraticDiscriminantAnalysis
- MLPClassifier

Os ajustes dos hiper parâmetros apresentaram a seguinte variação nos modelos:

**Tabela 6:** Performance dos modelos antes e depois dos ajustes dos hiper parâmetros

| Modelo | Recall sem ajustes dos hiper parâmetros | Recall com ajuste dos hiper parâmetros | Variação do Recall |
|---|---|---|---|
| KNeighborsClassifier | 85,94% | 85,94% | 0% |
| DecisionTreeClassifier | 92,16% | 92,16% | 0% |
| RandomForestClassifier | 92,70% | 91.62% | -1,08% |
| AdaBoostClassifier | 93,24% | 93,24% | 0% |
| QuadraticDiscriminantAnalysis | 74,59% | 74,59% | 0% |
| MLPClassifier | 84,32% | 81,08% | -3,24% |

[8] https://scikit-learn.org/stable/modules/generated/sklearn.model_selection.GridSearchCV.html

Dentre os resultados apresentados, escolhemos o modelo AdaBoostClassifier por ter apresentado melhor performance. Chamamos essa primeira versão de Baseline 3. Ao verificar os outros indicadores de performance com essa baseline, e comparando-os com as baselines anteriores, temos a seguinte relação:

**Tabela 7:** Resultado do modelo gerado

| Baseline | Recall | Precisão | Acurácia | F-SCORE |
|---|---|---|---|---|
| Baseline 1 | 90,27% | 95,42% | 92,90% | 91.15% |
| Baseline 2 | 91.62% | 95.22% | 93.36% | 92.26% |
| Baseline 3 | 92.70% | 94.49% | 93.49% | 92.43% |

## 6.5. Avaliação

Para avaliar nosso modelo optamos por usar uma análise conjunta de acurácia, recall e precisão. A acurácia em resumo mostra o quão frequente o nosso classificador é correto. A precisão por sua vez busca medir dos valores classificados como correto, quantos efetivamente eram corretos. O recall é a frequência em que o classificador encontra exemplos de uma classe.

Uma vez que o nosso classificador busca identificar benefícios fraudados, o recall torna-se uma avaliação muito importante, e o baixo valor do recall na primeira baseline motivou a tentativa de incrementar esse valor.

A primeira baseline do modelo apresentou uma boa precisão, mas com recall baixo. Com os ajustes descritos no modelo, conseguimos implementar um classificador com 93% de acurácia, 94% de precisão e 92% de recall.

A seguir, em complemento, serão delineadas as principais inferências obtidas pelas features e seus respectivos efeitos na fraude.

**Aumentam** o risco de fraude:

- Tempo de processamento entre a data de início do benefício e a data de entrada do requerimento. Quanto maior esse tempo, maior a correlação com a fraude.

- Total de descontos na folha de pagamento. Quanto maior o total de descontos na folha de pagamento, maior a correlação com a fraude.

- Valor do benefício pago. Quanto maior o valor do benefício, maior a correlação com a fraude.

**Diminuem** o risco de fraude:

- Possuir benefício anterior. Se o titular da pensão já possui um número de benefício anterior, isso diminui sensivelmente a correlação com a fraude.

- Instituidor ser masculino. Se o instituidor da pensão for masculino, isso diminui a correlação com a fraude.

- Possuir documento identificador do empregador. Se o titular da pensão possui uma documentação identificadora do empregador, isso diminui a correlação com a fraude.

Com base nas informações acima, será buscado a interlocução com os órgãos responsáveis pela identificação e tratamento das fraudes no âmbito da previdência social para identificação e tratamento das potenciais fraudes.

# 7. Considerações finais

O presente trabalho realizou o estudo e a aplicação de técnicas de aprendizado de máquina para criação de modelos preditivos para classificar uma nova concessão de benefício de pensão por morte no INSS como fraudulenta ou não.

Após algumas iterações e aperfeiçoamentos no modelo, uma versão final foi consolidada obtendo resultados de 94% de precisão, 92% de recall, 93% de acurácia e 92% de F-Score.

Como planos de trabalhos futuros, levando em conta a necessidade da continuidade deste trabalho, foram levantadas as seguintes:

- Averiguação, junto a órgãos de polícia e inteligência, dos casos onde o modelo aponta alta probabilidade de fraude, se a análise documental realizada por esses órgãos permite aferir se o benefício foi de fato fraudado

- Implantação do modelo desenvolvido nas unidades e órgãos que vão fazer o enfrentamento das fraudes, como a Coinpe, Policia Federal e Auditoria Interna do INSS.

# Referências bibliográficas


Alles, M.A., Brennan, G., Kogan, A., and Vasarhelyi, M.A. Continuous Monitoring of Business Process Controls: **A Pilot Implementation of a Continuous Auditing System at Siemens**. International Journal of Accounting Information Systems, Vol 7. 137-161. 2006

BRASIL. Tribunal de Contas da União. **Referencial de Combate à Fraude e à Corrupção**. 1. ed. Brasília, DF: TCU, p. 164, 2016. Disponível em: < https://portal.tcu.gov.br/lumis/portal/file/fileDownload.jsp?fileId=8A8182A15A235CCB015 A29ACF7D11830>. Acesso em: 26 jun. 2019

BRASIL: **Decreto nº 99.350**, junho de 1990a. http://www.planalto.gov.br/ccivil_03/ decreto/Antigos/D99350.htm.

BRASIL: **Lei nº 8029**, abril de 1990b. http://www.planalto.gov.br/ccivil_03/ leis/L8029cons.htm.

BRASIL. Tribunal de Contas da União. **Acórdão no 1057/2018-TCU-Plenário**

CHAPMAN, Pete, CLINTON, Julian ,KERBER, Randy , KHABAZA , Thomas, REINARZ Thomas e WIRTH , Rüdiger: **CRISP-DM 1.0: Step-by-step data mining guide**. SPSS, 2000.

BREIMAN, L , **Random forests**. 2001. Disponível em http://www.springerlink.com/index/U0P06167N6173512.pdf

Zaki, Mohammed J and Meira, Wagner. **Data mining and analysis: fundamental concepts and algorithms**, 2013.