

Machine Learning Aplicado ao Resultado de Pedido de Concessão de Benefícios do INSS – Análise Ampliada

Machine Learning Applied to the INSS Benefit Request - Extended Analysis

Ney Barchilon¹, Tatiana Escovedo², Marcos Kalinowski³

¹Departamento de Informática, Pontifícia Universidade Católica (PUC-Rio)
Rio de Janeiro, RJ, Brazil

²Departamento de Informática, Pontifícia Universidade Católica (PUC-Rio)
Rio de Janeiro, RJ, Brazil

³Departamento de Informática, Pontifícia Universidade Católica (PUC-Rio)
Rio de Janeiro, RJ, Brazil

ney@barchilon.com, tatiana@inf.puc-rio.br, kalinowski@inf.puc-rio.br

Abstract. *The materialization of social protection, foreseen in the Brazilian Constitution's Social Security chapter, specifically in the scope of Welfare, occurs through the granting and maintenance of benefits to all Brazilians who need this protection. This right generates a huge demand of millions of requests for annual benefits to the INSS (National Institute of Social Security), which is the operator of these services. Receiving and analyzing benefit requests, in a timely manner and with assertiveness, is complex and challenging. The volume of millions of applications for benefits annually, the diversity of benefits available, different criteria for granting and the urgency that the nature of these applications requires for the maintenance of life for applicants, express this complex and challenging environment. Within this context, the present study aims to develop some models, using machine learning techniques, and select the best one, which can predict whether a certain benefit request will be granted or rejected. This prediction would help in the analysis of new benefit requests, making room for the dynamics of the analysis process to be directed more quickly and assertively. The data source for the construction of the models in this work was obtained from the INSS Open Data Portal, which are included in the INSS Open Data Plan. This dataset is composed of monthly files of Decided Benefits (Granted and Dismissed) from December 2018 to June 2020. As a scope of analysis, algorithms such as KNN, SVC, Decision Trees, Logistic Regression, etc. were addressed. Models were also built using the Ensemble Bagging and Boosting techniques, reaching a set of seventeen analyzed algorithms. The algorithm that obtained the best performance, using the F1 metric as the determinant, was the eXtreme Gradient Boosting (XGB) Classifier with 80%. With this, the model performs the prediction with approximately 84%*

Accuracy, 76% Sensitivity and 81% for AUC (Area Under the Curve). As a result of the study, a model capable of predicting whether a given benefit requirement would be granted or denied was obtained, based on the requirement data, with a performance within the expectations established in the objectives. Therefore, this article analyzed seventeen machine learning algorithms with the aim of building a model to predict whether a given benefit request to the INSS will be granted or rejected. As a contribution of this article to the IS, we highlight the initiative to apply a machine learning model in a domain not yet explored by other research, using, as a data source, open data made available by the federal government in the cloud. It is hoped that the result of this study will open space for new scientific research in the field of machine learning to be developed in this domain, with the aim of helping with real problems in this sensitive part of the lives of Brazilians and of public administration.

Keywords. *Algorithms; Machine Learning; Models; Prediction; Welfare.*

Resumo. *A materialização da proteção social, prevista no capítulo da Seguridade Social da Constituição do Brasil, especificamente no âmbito da Previdência, se dá através da concessão e manutenção de benefícios a todos os brasileiros que necessitem dessa proteção. Esse direito gera uma demanda enorme de milhões de requisições de benefícios anuais ao INSS (Instituto Nacional de Seguridade Social), que é o operador desses serviços. Receber e analisar os pedidos de benefícios, em tempo hábil e com assertividade, é algo complexo e desafiador. O volume de milhões de pedidos de benefícios anuais, a diversidade de benefícios disponíveis, diferentes critérios de concessão e a urgência que a natureza desses pedidos exige para a manutenção da vida dos requerentes, expressam esse ambiente complexo e desafiador. Dentro deste contexto, o presente estudo vem no sentido de desenvolver alguns modelos, utilizando técnicas de aprendizado de máquina, e selecionar o melhor deles, que possa prever se determinada requisição de benefício será concedida ou indeferida. Essa predição ajudaria na análise de novos pedidos de benefício, abrindo espaço para que a dinâmica do processo de análise possa ser direcionada de forma mais ágil e assertiva. A fonte de dados para a construção dos modelos, neste trabalho, foi obtida no Portal de Dados Abertos do INSS, que constam do Plano de Dados Abertos do INSS. Esse conjunto de dados é composto de arquivos mensais de Benefícios Decididos (Concedidos e Indeferidos) no período de dezembro de 2018 a junho de 2020. Como escopo de análise, foram abordados algoritmos como KNN, SVC, Árvores de Decisão, Regressão Logística etc. Também foram construídos modelos que utilizam as técnicas de Ensemble Bagging e Boosting, chegando a um conjunto de dezessete algoritmos analisados. O algoritmo que obteve o melhor desempenho, utilizando a métrica F1 como determinante, foi o Classificador eXtreme Gradient Boosting (XGB) com 80%. Com este, o modelo realiza a predição com aproximadamente 84% de Precisão, 76% de Sensibilidade e 81% para AUC (Area Under the Curve). Como resultado do estudo, obteve-se um modelo capaz de efetuar a predição se determinado requerimento de benefício seria concedido ou indeferido, com base nos dados de requerimento, com uma*

performance dentro das expectativas estabelecidas nos objetivos. Portanto, este artigo analisou dezessete algoritmos de aprendizado de máquina com o objetivo de construir um modelo para prever se determinada requisição de benefício ao INSS será concedida ou indeferida. Destacamos como contribuição desse artigo ao SI, a iniciativa da aplicação de um modelo de aprendizado de máquina em um domínio ainda não explorado por outras pesquisas, utilizando, como fonte de dados, os dados abertos disponibilizados pelo governo federal na nuvem. Espera-se que o resultado desse estudo abra espaço para que novas pesquisas científicas, no campo da aprendizagem de máquina, sejam desenvolvidas nesse domínio com o objetivo de ajudar nos problemas reais dessa parte sensível da vida dos brasileiros e da administração pública.

Palavras-Chave. Algoritmos; Aprendizado de Máquina; Modelos; Predição; Previdência

1. INTRODUÇÃO

A Previdência Social no Brasil, segundo o relatório do Tesouro Nacional em dezembro de 2019 [Tesouro 2019], representa mais de 48% de todos os gastos do Governo Federal. São concedidos mais de 5 milhões de novos pedidos de benefícios por ano em todo o país, segundo o relatório da Secretaria de Previdência [INSS 2019]. O desafio de receber, analisar e apresentar o resultado desses novos pedidos, com agilidade e em tempo hábil, se dá em várias dimensões da administração, seja na quantidade de pessoal qualificado, na necessidade da revisão de processos ou na infraestrutura física e tecnológica de suporte. O prazo legal máximo de 60 dias (lei 9.784/99) não está sendo cumprido: segundo notícia do jornal Estadão [Estadão 2020], o problema já foi judicializado, uma vez que o Ministério Público Federal entrou com uma ação popular, cobrando o cumprimento desses prazos sob pena de multa [MP-DF 2019]. Portanto, ferramentas que ajudem na agilização do processo de análise desses pedidos de benefício são bem-vindas nesse momento.

O Portal de Dados Abertos do Governo Federal disponibiliza alguns conjuntos de dados sobre os benefícios concedidos e indeferidos na Previdência Social, que contêm características sobre os pedidos de benefícios, tais como tipo de benefício, localidade, sexo etc. Neste contexto, o objetivo deste trabalho é investigar se as características contidas nesses conjuntos de dados são suficientes para viabilizar a construção de um modelo que consiga prever se determinado novo pedido de benefício será concedido ou indeferido com um nível de precisão aceitável (acima de 80%).

Para alcançar o objetivo supracitado, efetuamos uma exploração e tratamento dos dados abertos e, em seguida, construímos modelos que ofereçam respostas a este problema com um nível de confiança alto para as predições, escolhendo aquele que ofereceu a melhor resposta. A seção 2 apresenta conceitos básicos sobre Previdência Social, INSS e seu Processo de Análise de Concessão de Benefícios, bem como conceitos de Aprendizado de Máquina. A seção 3, por sua vez, detalha o problema estudado e descreve a solução proposta. A seguir, a seção 4 apresenta os experimentos realizados com os 17 algoritmos analisados e os resultados obtidos. Finalmente, a seção 5 conclui este trabalho.

Como resultado, o algoritmo que obteve o melhor desempenho, utilizando a métrica F1 como determinante, foi o Classificador eXtreme Gradient Boosting (XGB)

com 80%. Com este, o modelo realiza a predição com aproximadamente 84% de Precisão, 76% de Sensibilidade e 81% para AUC (Area Under the Curve).

Não encontramos estudos relacionados na aplicação de modelos de aprendizagem para pedidos de benefícios previdenciários ou assistenciais. Os que existem, mesmo assim estrangeiros, são voltados para detecção de fraudes, o que implica em uma modelagem diferente com um espaço de características e target diferentes, pois, segundo Pacheco Junior [Pacheco Junior, 2019], uma das características marcantes das bases de dados que possuem informações de fraude é o desbalanceamento, o que não é o caso em estudo, sendo que esta particularidade causa um efeito negativo no processo de modelagem do evento, se fazendo necessária a utilização de abordagens específicas para este tipo de base de dados. Talvez existam trabalhos similares, mas que, por questões de segurança, as empresas não divulguem por receio de que alguém, de má-fé, possa desenvolver procedimentos para burlar os mesmos.

1.1. Evolução em Relação ao Artigo Original

No artigo original, apresentado e publicado na SBSI 2021 [Barchilon e Escovedo 2021], investigamos se era possível desenvolver um modelo de predição que fosse capaz de prever se determinado requerimento de benefício ao INSS seria ou não concedido no início do processo e se os dados de benefícios decididos, disponíveis no Portal de Dados Abertos do Governo Federal, fonte de dados para os modelos, seriam suficientes para construção deles com uma boa performance. Para tanto, utilizamos os dados da unidade federativa de Tocantins no treinamento/validação/teste dos modelos e concluímos que o melhor modelo alcançou boa performance nas suas predições.

Neste artigo, estendemos nosso trabalho anterior da seguinte forma: (a) ampliamos a descrição do processo de entrada de um requerimento no INSS, cujo entendimento é a base do problema em estudo; (b) incluímos algumas imagens e gráficos que ajudam no entendimento do texto; (c) incluímos uma análise SHAP (SHapley Additive exPlanations) para importância de características, tornando o modelo mais explicável, isto é, dando explicabilidade às predições, informando quais variáveis mais influenciaram/impactaram nas predições; (d) apresentamos os parâmetros utilizados no melhor modelo dando espaço para reprodutibilidade deste; (e) incluímos uma análise da aplicação do melhor modelo em mais sete unidades federativas (Região Sudeste: SP, RJ e MG, Região Sul: RS, Região Centro-Oeste: GO, Região Norte: RO e Região Nordeste: PE) e seus indicadores para avaliar se o modelo está generalizando bem com dados não vistos de outras unidades federativas; e (f) incluímos um link para um protótipo em que é possível efetuar uma simulação de utilização do melhor modelo em diversas situações para que o leitor possa avaliar on-line o resultado das predições do melhor modelo.

2. FUNDAMENTAÇÃO TEÓRICA

Nesta seção serão abordados conceitos básicos relevantes, utilizados nesse artigo, sobre a Previdência Social, o INSS e seu Processo de Análise de Concessão de Benefícios. Também será abordado o conceito de Aprendizado de Máquina.

2.1. Previdência Social no Brasil e o INSS

A Previdência Social tem por fim assegurar aos seus beneficiários meios indispensáveis de manutenção, por motivo de incapacidade, idade avançada, tempo de serviço, desemprego involuntário, encargos de família e reclusão ou morte daqueles de quem dependiam economicamente [Senado Federal 2008]. Dentro desse escopo, o sistema previdenciário brasileiro tem como principais objetivos ou funções evitar a queda brusca de renda na velhice, garantir renda ao segurado no caso de alguma contingência (acidente, gravidez, doença, velhice, entre outros) e atuar na prevenção à pobreza [Rangel e Caetano 2015]. Compete ao INSS a operacionalização do reconhecimento dos direitos dos segurados do Regime Geral de Previdência Social (RGPS) que abrange mais de 50 milhões de segurados e aproximadamente 33 milhões de beneficiários em 2017 [INSS 2017].

O desafio de atender a todos esses objetivos com uma população do tamanho do Brasil se revela em números. A Previdência Social brasileira – regime geral e diversos regimes próprios – tem como segurados aproximadamente 58% da população economicamente ativa (PEA) em 2008 [IPEA 2009]. Dentro desse contexto da universalização dos serviços prestados, abrangência de benefícios e ordem de grandeza dos números movimentados pela Previdência Social no Brasil, fica evidente que o desafio de gestão é enorme e que ferramentas que venham a ajudar na agilidade de seus processos serão bem-vindas.

2.2. O Processo de Análise de Benefício

O INSS é a autarquia federal operadora das ações para prestação dos serviços e pagamentos devidos aos segurados da Previdência Social. Sendo assim, quando o contribuinte cumpre os requisitos para receber um benefício, este pode se dirigir ao INSS e iniciar um processo administrativo se habilitando para receber esse benefício.

Um processo administrativo é iniciado para qualquer requerimento ao INSS pelo segurado. Atualmente, são cerca de 44 Espécies de benefícios [INSS Boletim 2020] e nove tipos de Forma de Filiação.

O INSS disponibiliza, através do seu site, um grande conjunto de informações acerca dos benefícios e do acesso para habilitação/acompanhamento desses. Apesar disso, não disponibiliza um documento formal contendo todas as etapas, tarefas e atividades que compõe o Processo Administrativo no INSS. No entanto, como o foco desse estudo é somente no resultado da análise de concessão (ou indeferimento) do benefício, então, através das informações disponibilizadas, podemos deduzir como o macroprocesso funciona até chegar ao resultado da análise (concedido ou indeferido). As etapas constantes desse macroprocesso podem ser reunidas em três processos, a saber: Pedido de Agendamento Atendimento, Recebimento do Requerimento e Análise do Requerimento.

Tomando como base a Instrução Normativa (IN) n. 77/2015 [DOU 2015], o processo administrativo previdenciário possui quatro fases: inicial; instrutória; decisória; e recursal [CNJ 2020]. Conforme ilustrado na Figura 1, na fase inicial, onde acontece o processo de Pedido de Agendamento Atendimento, é quando se efetua o requerimento da prestação previdenciária postulada, através dos diversos canais de atendimento do INSS, seja por meio de suas agências ou por meio do telefone 135 ou pelo portal "Meu INSS" na Internet [CNJ 2020]. Em seguida, na fase instrutória, onde acontece o processo de

Recebimento do Requerimento, conforme a IN n. 77/2015 [DOU 2015], são desenvolvidas as atividades de instrução destinadas a averiguar e comprovar os requisitos legais para o reconhecimento de direito aos benefícios e serviços da Previdência Social que serão realizadas pelo INSS.

Superada a fase de instrução, chega-se a fase decisória, onde acontece o processo de Análise do Requerimento, em que são desenvolvidas as atividades de análise e a decisão fundamentada da concessão ou indeferimento do requerimento. A fase recursal é aquela onde são desenvolvidas as atividades para revisão da decisão proferida na fase decisória. Segundo o art. 126, caput, da Lei n. 8.213/1991, regulamentado pelo art. 305 e seguintes do Decreto n. 3.048/1999, prevê que das decisões proferidas pelo INSS poderão os interessados, quando não conformados, interpor recurso ordinário ao Conselho de Recursos da Previdência Social [CNJ 2020].

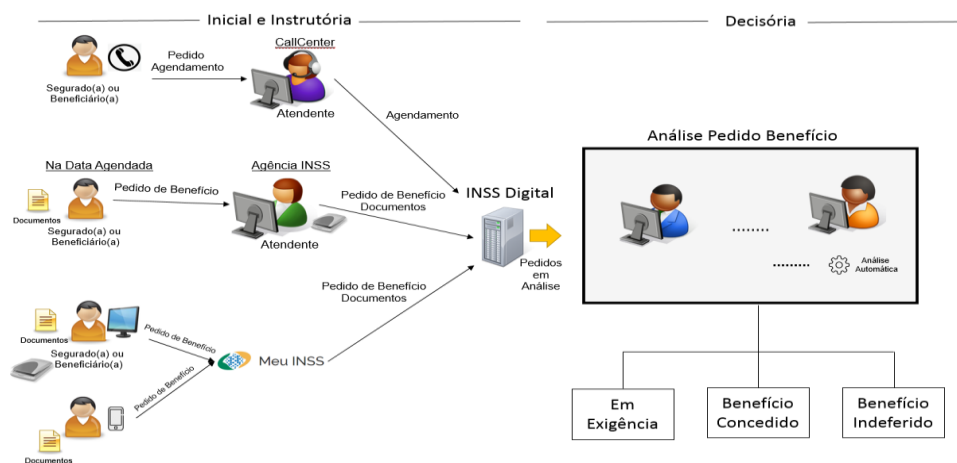


Figura 1. Processo Administrativo de Requerimento de Benefício INSS

Existe uma modalidade, lançada em 09/2017 e implementada gradualmente desde então, que é a de análise automática de requerimentos. Efetuado o requerimento pela Central 135 ou pela Internet e constatado que todas as informações nas bases cadastrais do INSS estão corretas, o benefício é concedido de forma automática [INSS 2018]. No momento do desenvolvimento deste artigo, em função da pandemia, o INSS suspendeu os serviços presenciais [INSS Atendimento 2020].

2.3. Aprendizado de Máquina

Aprendizado de Máquina (Machine Learning) é a ciência (e a arte) da programação de computadores para que eles possam aprender com os dados. Como exemplo o filtro de spam é um programa de Aprendizado de Máquina que pode aprender e assinalar e-mails como spam e como regulares [Geron 2019]. Mitchel [Mitchel, Tom 1997] se refere ao aprendizado de máquina como o campo da Inteligência Artificial responsável pelo desenvolvimento de modelos (hipóteses) gerados a partir de dados, e que automaticamente aperfeiçoam-se com a experiência.

No aprendizado Supervisionado, os dados de treinamento fornecidos ao algoritmo incluem as soluções que se desejam prever, chamadas rótulos [Geron 2019]. Escovedo e Koshiyama [Escovedo e Koshiyama 2020] ainda completam que o Aprendizado

Supervisionado é geralmente utilizado em problemas de Classificação e Regressão. O filtro de spam é um bom exemplo de um problema de Classificação, pois ele é treinado com muitos exemplos de e-mails junto às classes (spam ou não spam) e deve aprender a classificar novos e-mails [Geron 2019]. Já um exemplo de problema de Regressão é a predição do valor estimado das vendas em uma nova filial de uma determinada cadeia de lojas [Escovedo e Koshiyama 2020].

Alguns dos algoritmos mais conhecidos de aprendizado supervisionado são: K-Nearest Neighbours, Regressão Linear, Regressão Logística, Máquinas de Vetores de Suporte (SVM), Árvores de Decisão, Florestas Aleatórias e Redes Neurais [Geron 2019]. Com exceção das Redes Neurais, todos os algoritmos listados anteriormente, e muitos outros, como aqueles que utilizam as técnicas de Ensemble Bagging e Boosting, serão objeto desse estudo na construção de modelos capazes de uma boa generalização e que alcancem uma boa performance nas suas predições.

3. SOLUÇÃO PROPOSTA

O processo de análise de pedido de concessão de benefício, como mencionado anteriormente, tem sido parte do foco do INSS para melhorias na prestação de serviços ao cidadão. Diariamente, o processo é operado por cerca de 6 mil servidores dedicados [INSS Resolução 2019], recebendo, mensalmente, uma média de 833 mil novos requerimentos [INSS Boletim 2020]. Entre benefícios concedidos e indeferidos, chega-se a uma média mensal de 783 mil requerimentos [INSS Boletim 2020], ficando a diferença mensal entre aqueles da fila virtual nas situações em análise, em exigência ou foram cancelados. Em janeiro de 2019, cerca de 1,3 milhão de pedidos aguardavam, por mais de 45 dias, uma conclusão no requerimento [EBC 2020].

O principal problema no processo de análise de pedidos de benefícios, do ponto de vista do requerente, é o tempo de conclusão dos requerimentos. Para a agilidade desse processo, e a consequente diminuição do seu tempo, o INSS vem investindo em aumento de pessoal, melhorias em políticas de pessoal, treinamento de equipes, melhorias de processos, melhorias no suporte tecnológico, análise automática de requerimentos etc. [INSS Resolução 2019]. Dentro dessa perspectiva, um modelo capaz de predizer, logo no início do requerimento, e instantaneamente, que esse requerimento tem alta chance de ser concedido ao final do processo, poderia abrir uma janela de melhorias nesse processo

Conforme apresentado na Figura 2, o modelo poderia ser utilizado, já na fase de instrução, indicando, a todo momento, a chance (probabilidade) de que aquele requerimento seja, ao final do processo, concedido ou indeferido, abrindo espaço para que o processo de análise possa ser direcionado de forma mais ágil.

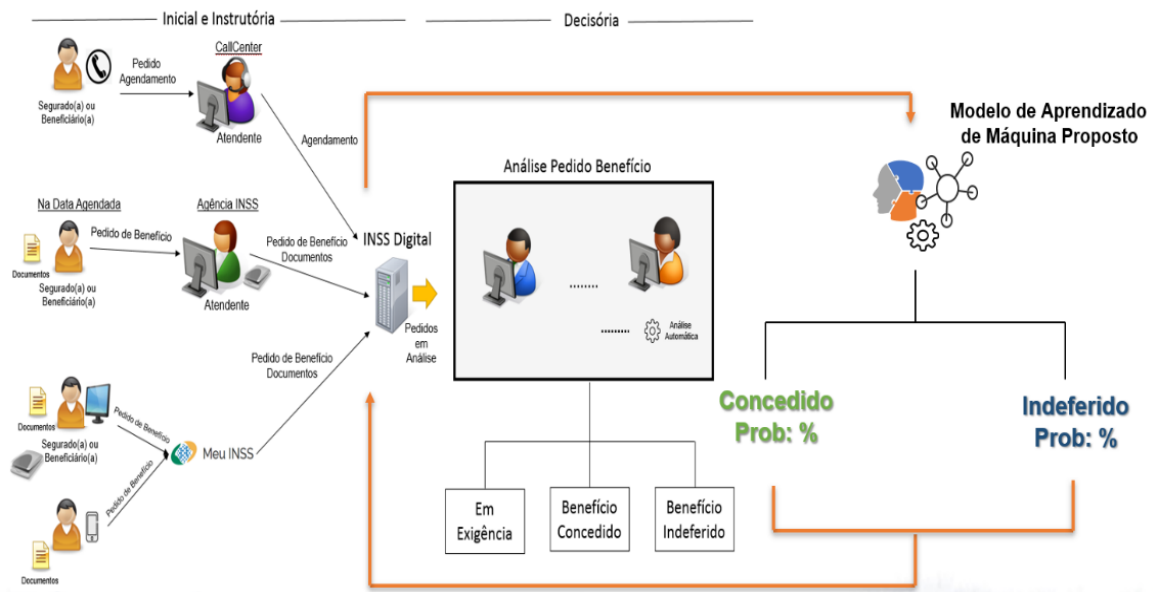


Figura 2. Processo Administrativo de Requerimento de Benefício INSS com um Modelo de Aprendizado de Máquina

Esse modelo seria uma ferramenta que se integraria ao conjunto de iniciativas hoje já em andamento, como a análise automática de requerimentos, abrindo espaço para que o processo de análise possa ser direcionado de forma mais ágil. A ideia é que já na abertura do pedido fosse indicada a possibilidade desse pedido ser concedido ou indeferido com a respectiva probabilidade dessa predição. Dessa forma, o processo poderia ser direcionado com base nessa nova informação, oferecendo, assim, uma contribuição para a diminuição desse tempo.

Nesse artigo, construímos modelos baseados em algoritmos de aprendizado Supervisionado, capazes de maximizar a generalização, mas com a precaução de evitar um sobreajuste, e, assim, efetuar boas predições. Construídos os modelos, esses foram avaliados nos indicadores produzidos e selecionado o modelo que apresentou a melhor capacidade de efetuar uma predição assertiva.

Um dos desafios com relação a construção desse modelo foi a escolha dos dados utilizados para seu treinamento, de modo que ele pudesse fornecer predições confiáveis. Para tal, utilizamos dois dos conjuntos de dados, relacionados no Plano de Dados Abertos do INSS [INSS Plano 2020], que contém os chamados Benefícios Decididos contendo os conjuntos de dados de Benefícios Concedidos e Benefícios Indeferidos.

Portanto, o objetivo desse estudo é responder a três perguntas: 1) Os conjuntos de dados a serem utilizados estão em condições de serem utilizados para construção dos modelos de aprendizado? 2) É possível construir um modelo de aprendizado de máquina, treinado com os conjuntos de dados abertos selecionados, que consiga prever, se um determinado pedido de concessão de benefício terá como resultado se será concedido ao final? 3) O modelo construído será capaz de fornecer uma predição com boa precisão (acima de 80%)?

Outra possível aplicação do modelo seria como mais uma ferramenta para suportar o novo programa de revisão administrativa de benefícios concedidos, chamado Programa Especial para Análise de Benefícios com Índícios de Irregularidade [INSS Fraudes 2019]. Segundo notícia recente veiculada no próprio site do INSS, informa que em todo o país, mais de 1 milhão de beneficiários de todas as espécies de benefício serão notificados por meio de carta de cumprimento de exigência [INSS Notificação 2020]. Nesse caso, o modelo poderia, junto com as ferramentas já desenvolvidas pelo órgão, ajudar apontando naqueles benefícios concedidos quais deles teriam menor probabilidade de serem concedidos para um tratamento em separado.

4. EXPERIMENTOS REALIZADOS

Para construção do modelo de aprendizado de máquina que efetuará a predição do resultado da concessão de benefício, vamos seguir alguns passos que nos levarão ao melhor modelo. Existem algumas possibilidades de um conjunto de passos a serem seguidos para a construção deste modelo, como a sugerida por Escovedo e Koshiyama [Escovedo e Koshiyama 2020] que define um Esquema Básico de um projeto de Data Science em sete etapas: Problema, Necessidade ou Ideia; Coleta e análise de dados; Pré-processamento; Modelagem e inferência; Pós-Processamento; Apresentação de Resultados e Implantação do modelo e geração de valor. Neste artigo vamos seguir a abordagem sugerida por Escovedo e Koshiyama, com algumas adaptações de procedimentos de acordo com a natureza do problema desse estudo. Basicamente, a diferença foi que o melhor modelo não foi implantado para utilização em um ambiente de produção real e sim em um ambiente de demonstração com um protótipo que foi desenvolvido especialmente para esse artigo.

4.1. Coleta e Análise dos Conjuntos de Dados

Os conjuntos de dados de Benefícios Concedidos e Indeferidos, identificados como Benefícios Decididos, utilizados como fonte primária desse estudo, foram obtidos do Portal de Dados Abertos do INSS [INSS Dados 2020] e que constam no Plano de Dados Abertos do INSS [INSS Plano 2020].

O conjunto de dados de Benefícios Concedidos, conforme apresentado na Figura 3, é, na realidade, composto de vários arquivos mensais, em um total de, no momento deste artigo, dezoito arquivos abrangendo os benefícios concedidos entre os meses de Dezembro de 2018 a Junho de 2020 (Janeiro de 2020 não constava). Cada instância contida nos arquivos contém um pedido de benefício, que passou pelo processo de análise, e cujo resultado foi concedido no mês de referência. Provavelmente por questões de privacidade, não existe qualquer identificação do beneficiário nesses arquivos.

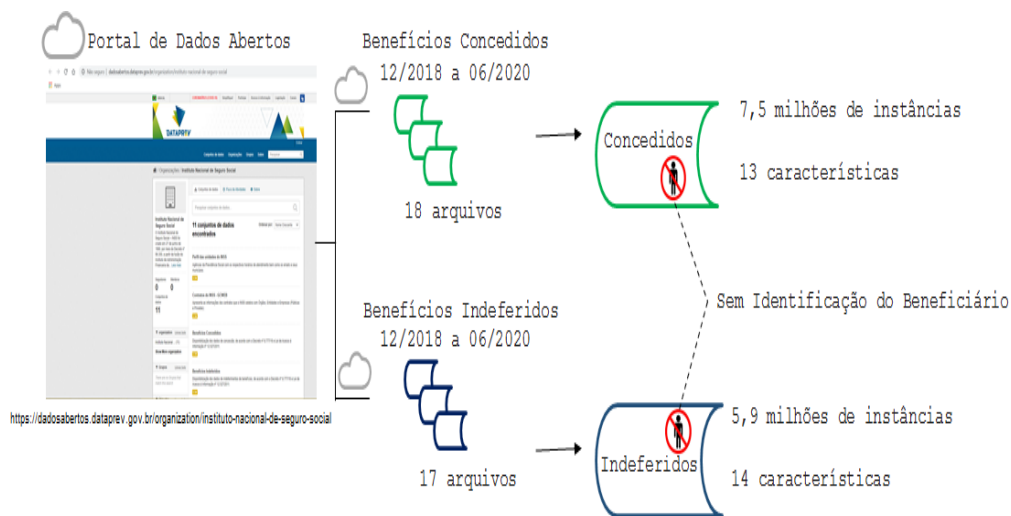


Figura 3. Coleta de Dados

O total de instâncias de todos estes arquivos juntos é de 7,5 milhões de instâncias de benefícios concedidos, com uma média mensal de 417 mil instâncias. Os arquivos possuem o mesmo layout com treze características, a saber: Competência Concessão, Espécie (Tipo de Benefício), CID (Classificação Estatística Internacional de Doenças e Problemas Relacionados com a Saúde), CID.1 (Descrição do CID), Despacho (Justificativa da Concessão), Dt Nascimento, Sexo, Clientela (Rural ou Urbano), Mun Resid, Vínculo dependentes (Filho, Cônjuge etc), Forma Filiação (Autônomo, Empregado, etc), UF e Qt SM RMI (Quantidade de Salários Mínimos na Renda Mensal Individual). A característica Clientela só consta até abril de 2019.

O conjunto de dados de Benefícios Indeferidos é, assim como o de Benefícios Concedidos, composto de vários arquivos mensais, em um total de, no momento deste artigo, dezessete arquivos abrangendo os benefícios indeferidos (não concedidos) entre os meses de Dezembro de 2018 a Junho de 2020 (sendo que Novembro de 2019 está inválido e Janeiro de 2020 não consta). Cada instância contida nos arquivos contém um pedido de benefício, que passou pelo processo de análise, e cujo resultado do processo de análise foi indeferido (não concedido) no mês de referência. Da mesma forma que os arquivos de Benefícios Concedidos, não existe qualquer identificação do beneficiário. O total de instâncias de todos estes arquivos juntos é de 5,9 milhões de instâncias de benefícios indeferidos, com uma média mensal de 347 mil instâncias. Os arquivos possuem o mesmo layout, com características comuns e outras diferentes do layout dos arquivos de Benefícios Concedidos, com nove características, a saber: Competência indeferimento, Espécie (Aposentadoria, Invalidez, Pensão etc), Motivo Indeferimento, Dt Nascimento, Sexo, Clientela (Rural ou Urbano), Forma Filiação (Autônomo, Empregado etc), UF, Dt Indeferimento.

Até o momento, não existe disponível uma documentação formal com a semântica de cada característica contida nesses arquivos. No entanto, as principais características, de ambos os conjuntos, podem ser identificadas confrontando as mesmas com o documento, produzido pela equipe do INSS, de Treinamento em Serviço em Reconhecimento Inicial de Direito e Cadastro para Concessores [INSS Treinamento 2019], disponível no site do INSS.

4.2. Preparação dos Conjuntos de Dados

Como vimos anteriormente, os conjuntos de dados, à primeira vista, já possuem algumas imperfeições, como a falta de instâncias em alguns meses, e precisamos tratá-las para que não influenciem negativamente na construção dos modelos. A primeira providência foi harmonizar os layouts desses conjuntos, de modo a torná-lo um conjunto único, acrescentando uma variável dependente (target) binária indicando se o benefício foi concedido (valor 1) ou indeferido (valor 0). Essa variável dependente (target) no conjunto único foi preenchida com 1 (concedido) se a instância é oriunda do conjunto de dados de Benefícios Concedidos e 0 (indeferido) se oriunda do conjunto de dados de Benefícios Indeferidos. As características não coincidentes nos conjuntos de dados foram descartadas. Com esse procedimento o conjunto de dados único ficou com as seguintes características: Competência, Espécie, Dt Nascimento, Sexo, Clientela, Forma Filiação, UF e Concedido (target), isto é, sete características independentes e uma dependente (target).

Seguindo no tratamento do conjunto único, entre todas as instâncias, verificamos problemas de características com valores faltantes (missing values) e que, no caso da característica Clientela, chegou a 80% do total de instâncias. Para tratamento destes, primeiro adotamos o procedimento de excluir a característica Clientela do conjunto de dados único devido ao grande percentual de valores faltantes. Para as outras características com menor percentual de valores faltantes, removemos as instâncias com valores faltantes (cerca de 152 mil instâncias) uma vez que não tínhamos informação suficiente para o preenchimento das mesmas e que representavam cerca de 1% do total de instâncias.

Avançando no tratamento dos dados no conjunto único, diversos ajustes para limpeza, correção e padronização dos dados foram efetuados. Entre eles, o ajuste em características retirando caracteres especiais, criação de atributos derivados dos originais que facilitem a análise dos modelos, ajuste de valores discrepantes etc.

Superada a atividade de tratamento dos dados, o conjunto único ficou com as características Competência (AAAAMM), Espécie, Idade, Sexo, FMFiliacao, UF e a variável dependente binária Concedido (target). Portanto, o dataset final contém seis características independentes e a variável dependente binária (Concedido), informando se aquela instância produziu um resultado de pedido de análise de benefício concedido (1) ou indeferido (0), com um total de 12,8 milhões de instâncias, sendo 54% concedido (1) e 46% indeferido (0) indicando um balanceamento entre as classes. Em que pese o pequeno número de características, trata-se de base rica para desenvolvimento do nosso estudo.

Com os dados preparados, foi possível efetuar uma série de inferências através de técnicas de análise exploratória. Várias informações sobressaem com essa análise, dentre as quais, destacamos: (i) a diferença de perfil das quantidades da Forma de Filiação dos segurados entre as unidades federativas e ao longo do tempo; (ii) a influência na variação do volume de benefícios decididos no período da tramitação da reforma da previdência e da pandemia; (iii) a diferença de perfil das quantidades da Espécie de Benefício entre as unidades federativas e ao longo do tempo; e (iv) a diferença de perfil da quantidade no Sexo do requerente entre as unidades federativas e ao longo do tempo. As visualizações e apontamentos dessa análise exploratória estão disponíveis em <https://zenodo.org/record/5517490>.

4.3. Conjunto de Dados para o Modelo

Vimos anteriormente que os volumes de dados disponíveis têm uma ordem de grandeza na casa de dezena de milhões de instâncias, o que é ótimo para análise de dados, extração de conhecimento e para a construção dos modelos. No entanto, trabalhar o treinamento de modelos nessa ordem de grandeza demanda uma disponibilidade de processamento altíssima. Isto posto, atingir os resultados com limitações de tempo e de recursos exige a capacidade de focar exatamente naqueles dados que podem fornecer as respostas as nossas questões sem comprometer a performance dos indicadores gerados.

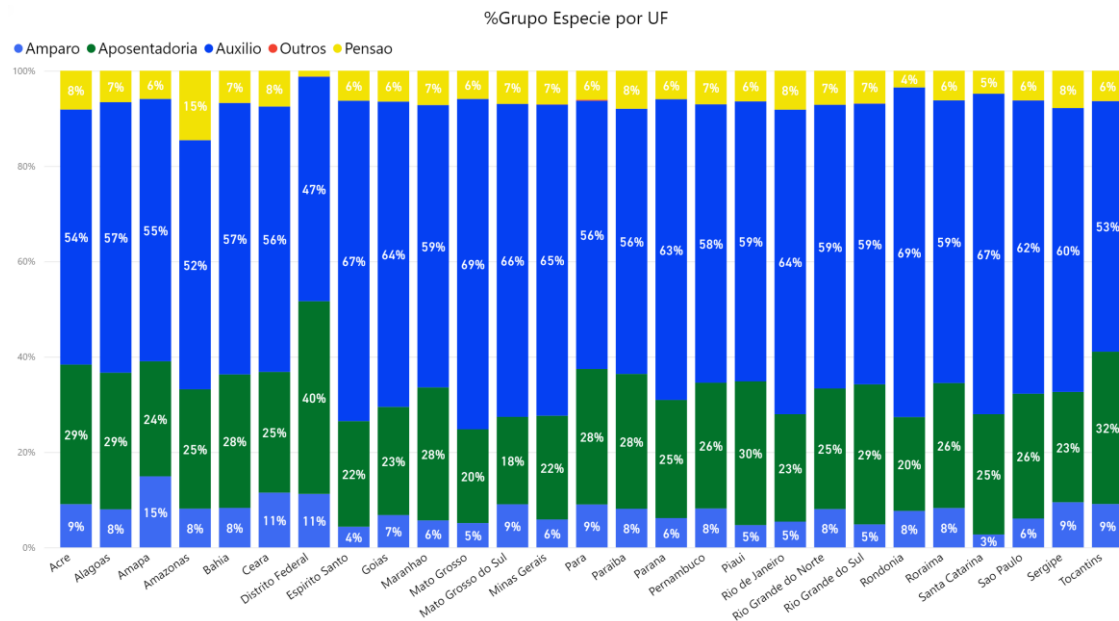


Figura 4. Percentual do Número de Benefícios Decididos entre 12/2018 e 06/2020 – por Grupo Espécie e Unidade Federativa

Na análise exploratória dos dados, verificamos que existe uma indicação de alta variação nas principais características do conjunto de dados entre as diferentes unidades da federação (Figura 4 e 5) e, também, em diferentes intervalos de tempo.

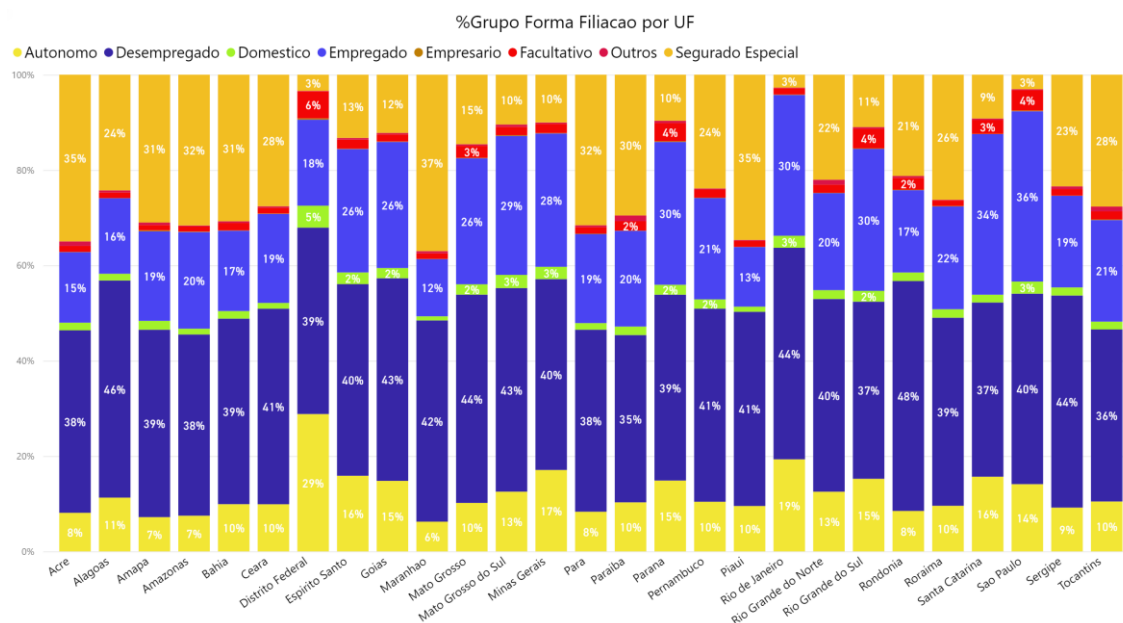


Figura 5. Percentual do Número de Benefícios Decididos entre 12/2018 e 06/2020 – por Grupo Forma de Filiação e Unidade Federativa

Também verificamos que algumas Unidades Federativas guardam dimensões e proporções semelhantes nas características trabalhadas e, portanto, podemos utilizar os dados de uma delas como base para uma estimativa do comportamento global. (Vide fls 8, 12, 20 e 22 da análise exploratória). Embora essas variações e padrões possam não influenciar negativamente a performance dos modelos desenvolvidos, essa informação foi importante para que pudéssemos avaliar uma boa estratégia para diminuir o escopo de dados submetidos ao treinamento dos modelos. Por exemplo, utilizando dados referentes a uma unidade federativa e por um período menor. Ainda pela análise exploratória, verificamos que os períodos de 03/2019 a 10/2019 foram altamente influenciados pela Reforma da Previdência (Figura 6), voltando ao padrão a partir de 12/2020, e pela Pandemia entre 03/2020 a 06/2020. No período da Pandemia não foi possível efetuar uma análise maior, pois foram coletados apenas os quatro meses iniciais (03/2020 a 06/2020), embora já fosse possível perceber uma variação anormal no comportamento da curva de benefícios decididos nesse período. Essa variação (vide análise exploratória fls 22) se deu pela queda no início da pandemia de todos as espécies de benefícios com uma vigorosa retomada dos benefícios de Auxílios a partir de maio de 2021.

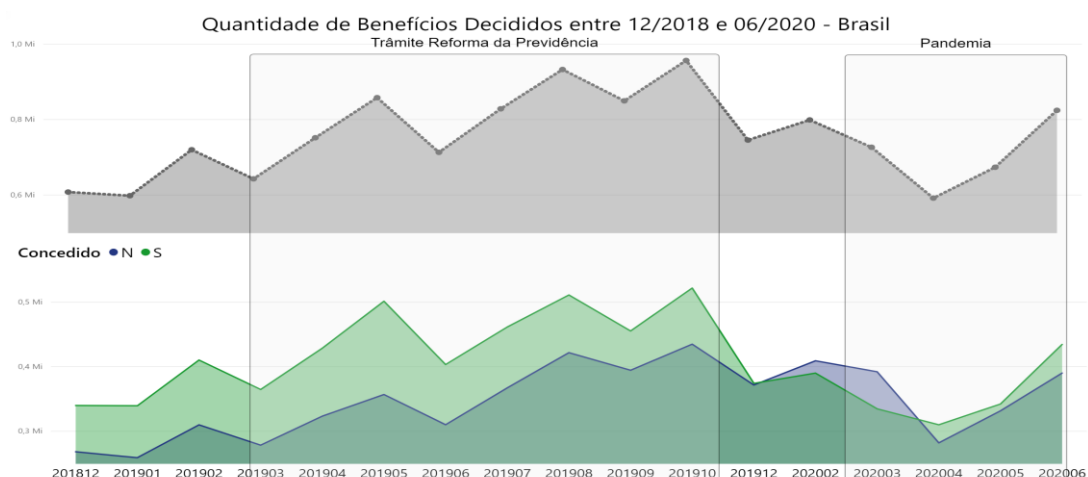


Figura 6. Número de Benefícios Decididos entre 12/2018 e 06/2020 – Brasil

Ainda investigando a movimentação desses períodos extraordinários, podemos verificar, pela Figura 7 (Gráfico Animado), que o período durante o trâmite da reforma da previdência as movimentações em algumas UF's, em especial o DF, foram bem fora de seus padrões normais. Não existem, nos conjuntos de dados coletados, informações que permitam saber ao certo o motivo dessa variação, especial com DF. No entanto, uma das possibilidades é que seja devido aos segurados que adiantaram os seus pedidos devido a perspectivas anunciadas de diminuição de direitos com a reforma da previdência.

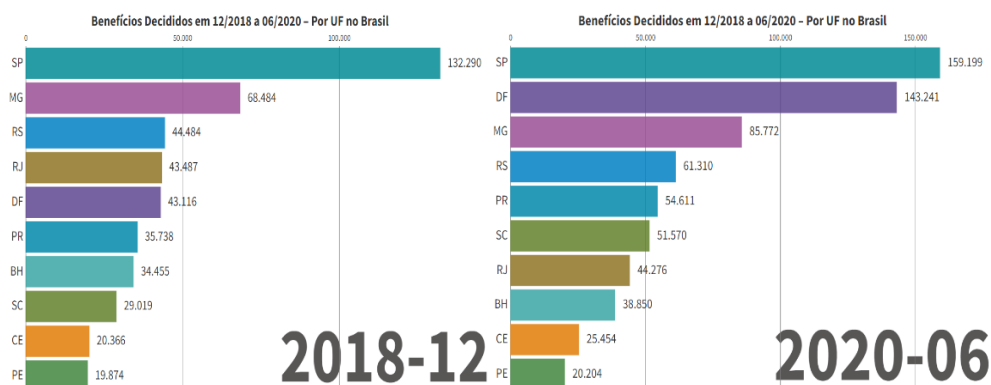


Figura 7. Número de Benefícios Decididos em 12/2018 e 06/2020 – Por UF no Brasil

Para visualizar esse gráfico animado (RaceBar) acesse: [Visualizar Gráfico](#)

Ou acesse e faça o download: [Gráfico em GIF para Download](#)

Portanto, evitamos a maior parte desses períodos na construção dos modelos desse estudo. Isto posto, definimos dois cortes no conjunto de dados único, adequando o mesmo à nossa capacidade de processamento, sem perder de vista a capacidade de resposta as nossas perguntas. Dessa forma, primeiro selecionamos, do conjunto único, todas as instâncias da unidade federativa de Tocantins (TO), com 78.560 instâncias, por estar no grupo padrão citado e por possuir um volume de dados adequado a nossa capacidade.

Dessas, utilizaremos o período de 12/2018 a 06/2019, que tem menor influência daqueles dois movimentos extraordinários que citamos, para serem trabalhadas pelos modelos de aprendizagem, totalizando cerca de 26 mil instâncias a serem trabalhadas pelos modelos de aprendizagem em treinamento, validação e teste.

4.4. Separação dos Dados

Antes de iniciar a construção dos modelos de aprendizagem, precisamos definir como será a distribuição dos dados que serão utilizados por esses modelos. Grus [Grus 2016] sugere que devemos dividir os dados em três partes: um conjunto de treinamento para construir modelos, um conjunto de validação para escolher entre os modelos treinados e um conjunto de teste para avaliar o modelo final.

Vamos adotar as três divisões sugeridas acima, assumindo que o modelo a ser utilizado em um mês foi avaliado com os dados dos seis meses anteriores, separamos para treinamento e validação dos modelos, os meses de 2018-12 a 2019-05 (22.682 instâncias) e o conjunto de teste para avaliação do modelo final o mês de 2019-06 (3.443 Instâncias). O conjunto de treinamento e validação (22.682 instâncias), utilizados para treino e avaliação da predição dos modelos, foi dividido em 80% para treinamento e 20% para validação, ficando, então, o conjunto de dados dividido em três partes: treinamento (18.145 instâncias sendo 80% do conjunto separado para treino/validação), validação (4.537 instâncias sendo 20% do conjunto separado para treino/validação) e, finalmente, teste (3.443 Instâncias) para a avaliação da predição do modelo final.

Na parte de treino utilizamos o método de validação cruzada para a construção do preditor e a avaliação do seu erro. A ideia por trás da validação cruzada é que a construção do preditor, e a avaliação do seu erro, sejam efetuadas em subconjuntos diferentes do conjunto de dados total e, para isso, propõe a divisão do conjunto de dados total em amostra de treino e validação, sendo os métodos mais usados de validação cruzada: Hold-Out, K-fold e Leave-one-out [Cunha 2019]. Nesse estudo, utilizaremos o método de validação cruzada K-fold que divide o conjunto de dados em K subconjuntos aleatórios e complementares, com tamanhos parecidos, sendo o modelo treinado em combinações diferentes desses subconjuntos e validado em relação às partes restantes [Geron 2019].

Para manter nos subconjuntos a proporção de instâncias em cada uma das classes, utilizamos a técnica de amostra estratificada (Stratified K-Folds) em cada um dos subconjuntos. Portanto, neste experimento, dividimos aleatoriamente o conjunto de dados separado para treino com 18.145 instâncias em 10 subconjuntos disjuntos (folds) com aproximadamente o mesmo número de elementos (18.145 instâncias / 10) e mantendo a proporção de instâncias em cada uma das classes. Cada um dos 10 subconjuntos é usado como conjunto de treino e os restantes são reunidos em um conjunto de validação deste treino. O processo é repetido 10 vezes iterativamente, sendo os modelos gerados e avaliados nestes. O Scikit Learn implementa essa técnica através da classe StratifiedKFold.

4.5. Algoritmos Utilizados

Para construção dos modelos de predição, utilizamos dezessete algoritmos, a saber: Naive Bayes (NBAY), K vizinhos mais próximos (KNN), Máquinas de Vetores de Suporte Linear (LinearSVC), Regressão Logística (RLOG), Árvore de Decisão (CART), Máquinas de Vetores de Suporte (SVC), Gradiente Descendente Estocástico (SGD),

Floresta Aleatória (RFOR), Árvores Extra (ETREE), Adaptive Boosting com Árvore de Decisão (ADAB_CART), Bagging com K vizinhos mais próximos (BAGG_KNN), Bagging com Regressão Logística (BAGG_RLOG), Bagging com Árvore de Decisão (BAGG_CART), Stochastic Gradient Descent (SGD), Gradiente Boosting (GB), eXtreme Gradient Boosting (XGB) e Light Gradient Boosting (LGB).

Adicionalmente, também construímos um modelo DUMMY que é útil como uma linha de base simples para comparar com outros classificadores (reais) priorizando sempre a classe mais frequente. Quase todos os modelos foram construídos com a utilização da biblioteca Scikit-Learn release 0.23 para Python (3.6.9), com exceção dos modelos XGB e LGB que possuem bibliotecas próprias para Python. Para o desenvolvimento e processamento dos modelos utilizamos o ambiente do Google Colab sobre o navegador Chrome versão 90.0. Para análise exploratória utilizamos o Power BI Desktop versão 2.91 64bits. O hardware utilizado foi um PC Intel Core I7 2.20GHz, com 8GB de RAM, SSD 128M e S.O. Windows 8.1.

4.6. Métricas de Avaliação dos Resultados

Para avaliar o desempenho dos modelos de classificação de aprendizado, precisamos definir quais são as métricas que desejamos adotar para efetuar a medição de desempenho (ou performance) de modo a avaliar o resultado da predição de um modelo. Geron [Geron 2019] sugere que uma boa maneira de avaliar o desempenho de um classificador é olhar para a matriz de confusão. Segundo Escovedo e Koshiyama [Escovedo e Koshiyama 2020], a matriz de confusão oferece um detalhamento do desempenho do modelo de Classificação, mostrando, para cada classe, o número de classificações corretas em relação ao número de classificações indicadas pelo modelo.

Ainda segundo Escovedo e Koshiyama [Escovedo e Koshiyama 2020], a partir da matriz de confusão, é possível verificar o número de Falsos Positivos, (FP - também conhecido como Alarme Falso; ou seja, quando o resultado esperado é negativo, mas o modelo resulta em positivo) e Falsos Negativos (FN - também conhecido como Alarme Defeituoso; ou seja, quando o resultado esperado é positivo, mas o modelo resulta em negativo), bem como os números de Verdadeiros Positivos, (VP, quando o resultado esperado é positivo e o modelo resulta em positivo) e Verdadeiros Negativos (VN, quando o resultado esperado é negativo e o modelo resulta em negativo). No nosso caso, o Positivo representa o Benefício Concedido (1) e o Negativo representa o Benefício Indeferido (0). A Tabela 1 ilustra uma matriz de confusão.

Tabela 1. Representação da Matriz de Confusão

Matriz de Confusão		Predição	
Real		Negativo	Positivo
	Negativo	TN _{ou} VN	FP
	Positivo	FN	TP _{ou} VP

Partindo da observação da matriz de confusão, diversas métricas podem ser reconhecidas. Escovedo e Koshiyama [Escovedo e Koshiyama 2020] relacionam as

métricas de Sensibilidade, que representa a capacidade de se identificar corretamente os indivíduos que apresentam a característica de interesse, a Especificidade, que mede a quantidade de negativos verdadeiros corretamente classificados, e representa a capacidade em identificar corretamente os indivíduos que não apresentam a condição de interesse. Geron [Geron 2019] acrescenta as métricas Precisão, que mede a acurácia das predições positivas, e a pontuação F1, que combina a Precisão e Sensibilidade através da sua média harmônica.

Em complemento, adotamos uma métrica que não deriva da matriz de confusão, mas que ajudou a avaliar melhor o conjunto de resultados que é a Curva ROC. Escovedo e Koshiyama [Escovedo e Koshiyama 2020] sugerem a curva ROC (Receiver Operating Characteristic) como sendo uma métrica de avaliação de desempenho de problemas de Classificação muito utilizada, que contrasta os benefícios de uma classificação correta e o custo de uma classificação incorreta, e definem a curva ROC como representação da "Sensibilidade" no eixo y e "1-Especificidade" (1 menos Especificidade) no eixo x. Quanto mais a curva estiver próxima do canto superior esquerdo, melhor o classificador. Geron [Geron 2019] observa a curva ROC como uma ferramenta comum utilizada com classificadores binários (que é o caso nesse estudo), e sugere que uma maneira de comparar classificadores é medir a área abaixo da curva (AUC, do inglês). Complementa ainda que um classificador perfeito terá um AUC igual a 1, enquanto um classificador puramente aleatório terá um AUC igual a 0,5. O Scikit-Learn fornece uma ferramenta para calcular o AUC. A Tabela 2 resume as fórmulas de cálculo das métricas que vamos observar e que são derivadas da matriz de confusão.

Tabela 2. Medidas de Desempenho derivadas da Matriz de Confusão

Fórmulas	
Medida	Fórmula
Acurácia	$\frac{(TN + TP)}{(TP + FP + TN + FN)}$
Precisão	$\frac{TP}{(TP + FP)}$
Sensibilidade (Recall ou TPR)	$\frac{TP}{(TP + FN)}$
F1	$\frac{2TP}{(2TP + FP + FN)}$
NPV	$\frac{TN}{(TN + FN)}$
Especificidade (TNR)	$\frac{TN}{(TN + FP)}$

Nesse estudo, estamos interessados em prever se um determinado processo de análise de benefício terá como resultado concedido, abrindo espaço para a possibilidade de que esse processo possa seguir um caminho mais ágil na análise. Isto posto, estamos mais preocupados com a performance dos chamados Verdadeiros Positivos (VP ou TP), que são aqueles cuja predição foi verdadeira e que, na realidade, são verdadeiros e, minimizando os Falsos Positivos (FP), que são aqueles cuja predição foi verdadeira e que, na realidade, eram falsos. Portanto, precisamos dar maior atenção as métricas Precisão, Sensibilidade, F1 e AUC. De acordo com especialistas consultados, os valores mínimos aceitáveis para esse tipo de experimento para Precisão, Sensibilidade e F1 seriam, respectivamente, 80%, 75% e 79% e mínimo de 80% para AUC. Foi considerado o melhor modelo aquele que alcançou esses valores mínimos e apresentou a melhor performance na Precisão seguido do melhor F1. Adicionalmente, o tempo de execução do modelo foi considerado dentro dos critérios de escolha, pois espera-se que o modelo possa ser utilizado em processos com respostas online e, portanto, o tempo de execução é importante.

4.7. Construção dos Modelos e Predição

Para construção dos modelos, efetuamos primeiro o treinamento destes buscando a melhor combinação de hiperparâmetros que gerem um modelo que alcance a melhor performance nas métricas selecionadas. Esse procedimento é chamado por Geron [Geron 2019] de ajuste do modelo. Como esse seria um trabalho muito tedioso e que demanda tempo para explorar as diversas combinações desejadas, Geron [Geron 2019] sugere a utilização da função GridSearchCV do Sckit-Learn para efetuar a busca nas combinações desejadas por meio de validação cruzada. Desta forma, seguimos com o treinamento dos modelos com a busca de 1.952 combinações de hiperparâmetros nos diversos modelos.

Os resultados do treinamento e validação em cada modelo, da melhor combinação de hiperparâmetros, podem ser verificados na Tabela 3, dispondo das métricas em cada melhor modelo por ordem descendente da métrica Precisão na validação cruzada. Aplicando os critérios descritos anteriormente sobre os resultados, verificamos que sete deles (SVC, XGB, GB, LGB, RFOR, BAGG_CART e BAGG_KNN) conseguiram alcançar os valores mínimos estabelecidos para avaliação da performance dos modelos.

Tabela 3. Resultados de Treinamento dos 17 Modelos

Ordenação: Média da Precisão de Treino com Validação Cruzada (DESCENDENTE)							
Modelo	Validação Cruzada						
	Acc	Prec	Rec	F1	Npv	Esp	AUC
01.NBAY	57.8	93.1	18.6	31.0	53.8	98.6	58.6
02.XGB	80.4	84.6	75.2	79.6	76.9	85.8	80.5
03.GB	80.4	84.6	75.2	79.6	76.9	85.8	80.5
04.LGB	80.4	84.5	75.3	79.6	76.9	85.7	80.5
05.KNN	79.1	84.3	72.5	78.0	75.1	86.0	79.2
06.CART	78.9	84.2	72.1	77.6	74.7	85.9	79.0
07.BAGG_KNN	80.1	84.0	75.1	79.3	76.7	85.2	80.2
08.BAGG_CART	80.0	83.9	75.3	79.4	76.8	84.9	80.1
09.RFOR	80.1	83.8	75.6	79.5	77.0	84.9	80.2
10.ETREE	79.5	83.5	74.6	78.8	76.3	84.6	79.6
11.SVC	80.3	83.3	76.7	79.8	77.6	84.0	80.3
12.SGD	74.0	82.8	61.9	70.8	68.6	86.6	74.3
13.ADAB_CART	79.1	82.8	74.4	78.4	75.9	84.0	79.2
14.LinearSVC	77.5	82.2	71.2	76.3	73.7	84.0	77.6
15.BAGG_RLOG	73.7	81.0	63.1	70.9	68.8	84.6	73.9
16.RLOG	77.3	80.8	72.8	76.6	74.4	82.0	77.4
17.DUMMY	50.0	50.9	50.7	50.8	49.0	49.2	50.0

Os resultados da aplicação desses melhores modelos treinados sobre o conjunto de validação para avaliar as métricas das suas predições constam na Tabela 4. Aplicando os critérios descritos anteriormente, sobre as predições de validação, verificamos que seis deles (XGB, SVC, GB, LGB, BAGG_KNN e RFOR) superaram os valores mínimos estabelecidos para avaliação da performance dos modelos, sendo que o melhor deles foi o modelo XGB.

No treinamento do modelo XGB foi utilizado o método de booster baseado em árvore (gbtree). Para o tuning do modelo foram selecionados os seguintes parâmetros: gamma, max_depth e n_estimator. Segundo a documentação do XGBoost [Developers 2020], o parâmetro gamma especifica o valor mínimo de redução necessário para que um nó se divida. O parâmetro max_depth define a profundidade máxima de uma árvore e o n_estimator define o número de árvores sequenciais a serem modeladas. A combinação que alcançou a melhor performance no treinamento foi gamma: 2.5, max_depth: 10 e n_estimators: 50.

Portanto, o modelo que apresentou a melhor performance, segundo os critérios estabelecidos, foi o XGB alcançando uma Precisão de 83.8%, Sensibilidade 76% e F1 79.7%. Esses resultados significam que as predições geradas pelo modelo têm a expectativa de alcançar um nível de acerto se o benefício será concedido em 83.8% e que essa predição deve alcançar, sobre todo o universo de concedidos, um acerto de 76%. O tempo de treinamento/validação deste modelo foi de 00:43:14.

Tabela 4. Resultados da Validação do Melhor Modelo de Treino nos 17 Modelos

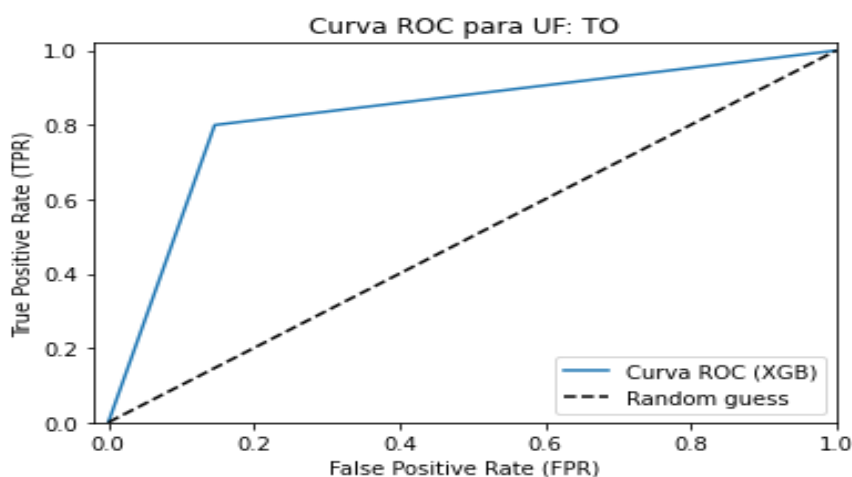
Modelo	Predição de Validação						
	Acc	Prec	Rec	F1	Npv	Esp	AUC
01.NBAY	59.3	94.4	19.3	32.0	55.3	98.9	59.1
02.XGB	80.7	83.8	76.0	79.7	78.2	85.4	80.7
03.GB	80.6	83.8	75.5	79.5	77.9	85.5	80.5
04.LGB	80.5	83.5	75.8	79.4	78.0	85.1	80.4
05.RFOR	80.3	83.4	75.5	79.2	77.8	85.1	80.3
06.CART	79.8	83.3	74.3	78.5	77.0	85.2	79.7
07.KNN	79.4	83.1	73.6	78.1	76.5	85.2	79.4
08.BAGG_KNN	80.3	82.9	76.2	79.4	78.2	84.4	80.3
09.ETREE	79.8	82.7	75.1	78.8	77.4	84.5	79.8
10.ADAB_CART	79.6	82.7	74.5	78.4	77.0	84.6	79.5
11.SVC	80.4	82.3	77.2	79.7	78.7	83.6	80.4
12.BAGG_CART	80.0	82.2	76.2	79.1	78.0	83.7	79.9
13.LinearSVC	78.0	82.1	71.3	76.3	74.9	84.6	77.9
14.RLOG	78.1	80.8	73.3	76.9	75.8	82.8	78.0
15.BAGG_RLOG	73.9	80.5	62.7	70.5	69.7	85.0	73.8
16.SGD	74.3	72.5	77.9	75.1	76.4	70.7	74.3
17.DUMMY	50.1	49.9	51.9	50.8	50.4	48.3	50.1

Para validar essa performance, efetuamos uma simulação do ambiente real submetendo o modelo XGB a um conjunto de teste do mês seguinte ao período que foi submetido em treino/validação. Esse teste tem como objetivo avaliar se a expectativa das métricas serão correspondidas em um novo período não visto. Para tanto, o modelo foi novamente treinado com todo o conjunto de treino/validação (22.682 instâncias), salvo e submetido ao conjunto de teste do mês de competência 06/2019 da mesma unidade federativa (TO) com 3.443 instâncias. O resultado dessa avaliação consta na Tabela 5.

Tabela 5. Resultados de Teste do Melhor Modelo (XGB)

Modelo	Predição de Teste						
	Acc	Prec	Rec	F1	Npv	Esp	AUC
XGB	82.7	83.3	79.7	81.4	82.3	85.5	82.6

A curva ROC do modelo XGB, para o conjunto de teste, pode ser observada na Figura 8. Verificamos que o modelo XGB, quando submetido ao conjunto de teste, alcançou boas medidas de desempenho em todos os indicadores, confirmando as expectativas de treino e validação. Este modelo respondeu bem às métricas dentro das expectativas em todos os conjuntos de treino, validação e teste, não só quanto as predições positivas (benefício será concedido), mas, também, apresentou uma boa performance nas predições negativas (benefício será indeferido), se habilitando a ser a melhor escolha, para o momento, dentro do escopo desse estudo. O tempo de teste deste modelo foi de 00:00:02.453413.

**Figura 8. Curva ROC do Modelo XGB para o Conjunto de Teste**

Encontrado o melhor modelo, iniciou-se a investigação se ele generalizaria bem em dados de outras unidades federativas, uma vez que, como citado anteriormente, efetuamos o treinamento e validação dos modelos com os dados da unidade federativa de TO. Para verificar se o melhor modelo XGB generalizaria bem em outras unidades federativas, submetemos o mesmo, separadamente, a sete conjuntos de teste contendo todos os dados das unidades federativas para o mesmo mês do conjunto de teste de TO, isto é, junho de 2019: Região Sudeste: SP (148.209 registros), RJ (47.683 registros) e MG (78.410 registros); Região Sul: RS (49.468 registros); Região Norte: RO (6.255 registros); Região Centro-Oeste: GO (19.111); Região Nordeste: PE (20.910). Essas unidades federativas foram selecionadas por apresentarem volumes consideráveis de pedidos e por representarem diferentes regiões do país com cerca de 52% do total deste período.

A performance do melhor modelo XGB nesses novos conjuntos de teste das unidades federativas, constam da Tabela 6. Verificamos que o modelo manteve boa performance (acima dos padrões mínimos desejados) nessas unidades federativas.

Tabela 6. Resultados de Teste do Melhor Modelo (XGB) em 7 UF's

	Predição de Teste em 7 UF's						
UF	Acc	Prec	Rec	F1	Npv	Esp	AUC
SP	80.3	86.8	76.0	81.0	74.1	85.6	80.8
RJ	81.5	84.0	81.5	82.8	78.6	81.4	81.5
MG	83.0	88.7	82.3	85.3	75.7	84.0	83.1
RS	80.4	89.8	75.3	81.9	71.3	87.8	81.5
RO	83.5	88.1	76.8	82.1	80.0	89.9	83.4
GO	82.3	90.6	79.0	84.4	73.0	87.4	83.2
PE	80.0	85.8	75.1	80.1	74.8	85.6	80.3

A diferença na performance das diferentes métricas em cada unidade federativa, embora acima dos padrões estabelecidos, pode sugerir que as diferenças regionais nas características do conjunto de dados, apontadas na análise exploratória deste trabalho, podem afetar a performance da predição do modelo, o que poderia indicar a necessidade de um treinamento direcionado para esses casos. A percepção dessa possível queda de performance por unidade federativa representa um avanço em relação ao artigo original, uma vez que aponta a necessidade de um estudo mais aprofundado em cada uma das unidades para verificação da performance do modelo.

Superada a avaliação da performance do melhor modelo XGB, finalizamos esse trabalho investigando quais características, do conjunto de dados que utilizamos para o treino do nosso melhor modelo, que mais influenciaram na construção de suas predições. Essa investigação tem como objetivo oferecer uma maior explicabilidade para o comportamento do modelo. O procedimento de avaliar a importância de cada característica para o desempenho de classificação do modelo é chamado de feature importance [Casalicchio G et al 2019], sendo essa uma das técnicas mais comuns na busca pela explicabilidade do modelo [Bhatt U et al 2020].

Na busca pelas características que mais influenciaram no melhor modelo XGB, utilizamos o método SHAP (SHapley Additive exPlanations) [Lundberg e Lee 2017]. Segundo Molnar [Molnar, C 2019], o objetivo do SHAP é explicar a previsão de uma instância x calculando a contribuição de cada característica para a previsão. O método de explicação SHAP calcula os valores de Shapley a partir da teoria dos jogos de coalizão. Ainda segundo Molnar [Molnar, C 2019] a ideia geral é que uma previsão pode ser explicada assumindo que cada valor de característica da instância é um "jogador" em um jogo em que a previsão é o pagamento. Os valores de Shapley nos dizem como distribuir de forma justa o "pagamento" entre as características. De forma simplificada, a forma de

cálculo do valor de Shapley é a média das contribuições marginais em todas as permutações com as características de cada instância.

Molnar [Molnar, C 2019] explica que a ideia por trás da importância de uma característica (feature importance) SHAP é que características com grandes valores Shapley absolutos são importantes. Como queremos a importância global, somamos os valores Shapley absolutos por característica nos dados.

O resultado desse método, utilizando o nosso melhor modelo no conjunto de dados em que foi treinado, pode ser visualizado na Figura 9. É possível observar que as características mais importantes para as previsões foram Forma de Filiação, Espécie e Idade, seguidas pela característica Sexo em menor grau. Portanto, Forma de Filiação, Espécie, Idade e Sexo foram as características mais importantes para explicar o comportamento do modelo na geração de suas previsões.

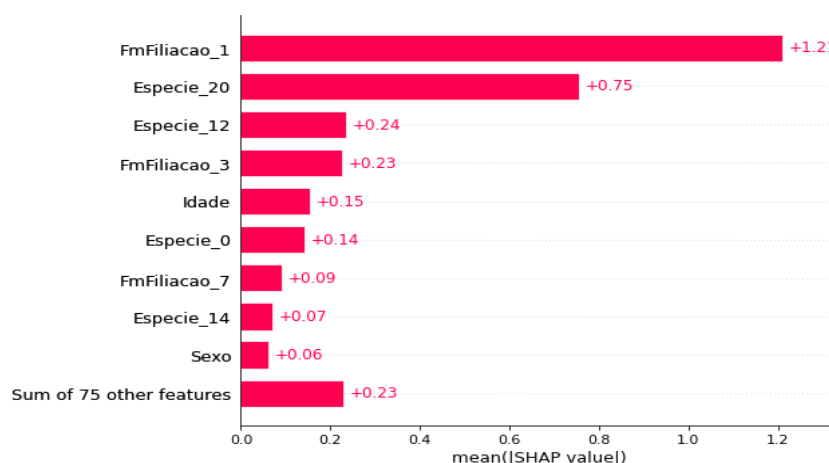


Figura 9. Importância das Características pelo valor SHAP

O modelo construído neste trabalho está disponível para utilização através de um protótipo de aplicação web em <https://streamlit-cbenef.herokuapp.com/>. Cabe ressaltar que, em função da limitação de espaço no github, só foi possível carregar um pequeno volume de dados para utilização no mesmo.

5. CONCLUSÃO

Neste artigo, avaliamos dezessete modelos de classificação de aprendizado de máquina, a fonte de dados dos diversos modelos tem origem em dados abertos do INSS no portal de dados abertos do governo federal, nos conjuntos de benefícios decididos (Concedidos e Indeferidos). Para não onerar demais o tempo de treinamento, o conjunto de dados, utilizado no treinamento/validação/teste dos modelos, ficou restrito a um período de sete meses (Dezembro de 2018 a Junho de 2019) e a uma unidade federativa Tocantins (TO).

Os modelos desenvolvidos foram avaliados em seus desempenhos por várias métricas: precisão, sensibilidade, F1, especificidade e AUC, para classificar a concessão de benefício, ou seu indeferimento. Os modelos de classificação de aprendizado de máquina foram treinados usando a técnica de validação cruzada estratificada dez vezes, em um conjunto de treinamento com 18.145 instâncias de processos (80% das instâncias

separadas para treinamento) e 4.537 instâncias de processos para validação da predição dos modelos treinados (20% das instâncias separadas para treinamento) e escolha do melhor modelo. O ajuste dos hiperparâmetros dos modelos alcançou um total de 1.916 combinações até que chegássemos aos melhores ajustes. Como avaliação final, o melhor modelo nas etapas de treino/validação foi submetido ao conjunto de dados de teste (3.443 instâncias) do mês seguinte (Junho de 2019) ao período selecionado para treinamento e validação, de modo a simular uma utilização real e avaliar se o melhor modelo selecionado confirmaria a performance de predição alcançada no treinamento e validação.

Neste experimento, o melhor modelo ajustado na fase de treino/validação foi o XGB, que apresentou uma Sensibilidade de 76% no resultado para análise de benefício para concessão do benefício, Especificidade de 85% no resultado para análise de benefício para indeferimento do benefício, Precisão de 84% e pontuação F1 de 80%. Complementando esses números, obteve ainda uma área estimada sobre a curva AUC de 81%. Adicionalmente, este melhor modelo ajustado foi submetido a um conjunto de dados de teste do mês seguinte (Junho de 2019) ao período de treino e validação (Dezembro de 2018 a Maio de 2019), simulando uma utilização real. Nesse novo experimento, as métricas de predição alcançadas foram, uma Sensibilidade de 80% no resultado para análise de benefício para concessão do benefício, Especificidade de 86% no resultado para análise de benefício para indeferimento do benefício, Precisão de 83%, uma pontuação F1 de 81% e uma área estimada sobre a curva AUC de 83%, confirmando, assim, os bons números de treino/validação.

Portanto, constatamos que o modelo de aprendizado de máquina desenvolvido, eXtreme Gradient Boosting (XGB), utilizando os conjuntos de dados abertos do INSS de Benefícios Decididos, pode fornecer uma predição precisa e consistente do resultado da análise de concessão de benefícios ao INSS. Isto posto, esse modelo pode oferecer uma oportunidade de que possa ser utilizado como mais uma ferramenta para ajudar na análise de novos pedidos de benefício. A utilização da predição deste modelo pode abrir espaço para que o processo de análise (e/ou análise automática) possa ser direcionado de forma mais ágil e assertiva. Desta forma, esse pedido, que já se saberia, pelo modelo, tem alta possibilidade de ser concedido ou indeferido, poderia ter o seu trâmite no processo agilizado, oferecendo, assim, uma contribuição para a diminuição desse tempo.

Neste artigo estendido, o melhor modelo foi, ainda, submetido aos dados das unidades federativas de SP, RJ, MG, RS, RO, GO e PE para o mesmo mês de junho de 2019, alcançando boa performance nas métricas de predição. No entanto, a diferença na performance das métricas de predição, entre as unidades federativas, pode sugerir que as diferenças regionais nas características do conjunto de dados, apontadas na análise exploratória deste trabalho, podem afetar a performance da predição do modelo. A percepção dessa possível queda de performance por unidade federativa representa um avanço em relação ao artigo original, uma vez que aponta a necessidade de um estudo mais aprofundado em cada uma das unidades para verificação da performance do modelo. Também como parte dessa extensão do artigo, efetuamos uma análise SHAP sobre as características do conjunto de dados e verificamos que Forma de Filiação, Espécie, Idade e Sexo foram as características mais importantes para explicar o comportamento do melhor modelo na geração de suas predições.

Como trabalhos futuros, propomos investigar um conjunto de novos procedimentos e técnicas que entendemos podem melhorar ainda mais a performance e

assertividade das respostas dos modelos, diminuindo também a possibilidade de viés na predição, são elas: (a) avaliar a construção de modelos em separado por unidade federativa ou um modelo único para todo o Brasil; (b) verificar a possibilidade da ampliação do espaço de características disponíveis para treinamento dos modelos incorporando novas informações dos sistemas do próprio INSS e de outras fontes; (c) avaliar o melhor período para novo treinamento do modelo e com base em que janela de tempo anterior esse treinamento seria efetuado; (d) aprofundar novas combinações de hiperparâmetros, principalmente nos modelos Ensemble, de modo a obter melhor performance nas métricas de suas predições; (e) propor um conjunto de novas técnicas de Ensemble, não utilizadas nesse artigo, que combinem resultados da predição de diferentes modelos individuais ponderando seus resultados de diferentes formas, alcançando resultados de predição ainda melhores que os modelos individuais; (f) propor a utilização de modelos de redes neurais profundas para gerar novas predições, observando as condições de disponibilidade de dados e limitações tecnológicas para seu desenvolvimento.

REFERÊNCIAS

- Bhatt, U., Xiang, A., Sharma, S., Weller, A., Taly, A., Jia, Y., Ghosh, J., Puri, R., Moura, J.M., Eckersley, P. (2020) “Explainable machine learning in deployment”. In: Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency, pp 648–657, Disponível em: <https://arxiv.org/pdf/1909.06342.pdf>, Acessado em: 08/08/2021.
- Barchilon, N. e Escovedo, T. (2021) “Machine Learning Applied to the INSS Benefit Request”. In: XVII Brazilian Symposium on Information Systems (SBSI).
- Casalicchio G., Molnar C., Bischl, B. (2019) “Visualizing the Feature Importance for Black Box Models”. Lect Notes Comput Sci 11051:655–670, Disponível em: https://www.researchgate.net/publication/324587839_Visualizing_the_Feature_Importance_for_Black_Box_Models, Acessado em: 08/08/2021.
- Cunha, J.P.Z. (2019). “Um estudo comparativo das técnicas de validação cruzada aplicadas a modelos mistos”. Dissertação de Mestrado. Instituto de Matemática e Estatística, USP. São Paulo, 2019. Disponível em <https://www.teses.usp.br/teses/disponiveis/45/45133/tde-26082019-220647/en.php>. Acessado em: 10/09/2020.
- CNJ (2020), “A judicialização de benefícios previdenciários e assistenciais / Instituto de Ensino e Pesquisa (INSPER)”, Relatório Final de Pesquisa, Conselho Nacional de Justiça - CNJ, Disponível em https://bibliotecadigital.cnj.jus.br/jspui/bitstream/123456789/530/1/Suma%cc%81rio-Executivo-Previde%cc%82ncia-Insper-CNJ_2020-12-01.pdf. Acessado em 08/05/2020
- Developers (2020) “XGBoost Documentation”, Disponível em: <https://xgboost.readthedocs.io/en/latest/python/index.html>, Acessado em: 08/08/2021.
- DOU (2015), “INSTRUÇÃO NORMATIVA Nº 77, DE 21 DE JANEIRO DE 2015”, Diário Oficial da União, 21/01/2015, Disponível em https://www.in.gov.br/materia/-/asset_publisher/Kujrw0TZC2Mb/content/id/32120879/do1-2015-01-22-instrucao-normativa-n-77-de-21-de-janeiro-de-2015-32120750. Acessado em 08/05/2020

- EBC (2020), “INSS realiza força-tarefa para agilizar concessão de benefícios”, EBC – Empresa Brasileira de Notícias - Agência Brasil, 11/01/2020, 2020, Disponível em <https://agenciabrasil.ebc.com.br/economia/noticia/2020-01/inss-realiza-forca-tarefa-para-agilizar-concessao-de-beneficios>. Acessado em 08/05/2020
- Escovedo, T. e Koshiyama, A. (2020), “Introdução a Data Science – Algoritmos de Machine Learning e métodos de análise”, Casa do Código.
- Estadão, Jornal 2020, “Tribunal manda INSS cumprir prazo na análise de pedidos de benefícios”, Estadão, 12/02/2020, Disponível em: <https://www.istoedinheiro.com.br/tribunal-manda-inss-cumprir-prazo-na-analise-de-pedidos-de-beneficios/>. Acessado em: 08/05/2020.
- Geron, A. (2019), “Mãos à Obra: Aprendizado de Máquina com Scikit-Learn & TensorFlow”, Editora Alta Books; Edição: 1.
- Grus, J. (2016), “Data Science do Zero”, Editora Alta Books.
- INSS (2018), “89,7 mil benefícios foram concedidos de forma automática pelo INSS”, Ministério da Economia - INSS, Disponível em <https://www.gov.br/inss/pt-br/assuntos/noticias/813-mil-beneficios-foram-concedidos-de-forma-automatica-pelo-inss>. Acessado em 08/05/2020.
- INSS Dados (2020), “Conjunto de Dados Abertos”, Ministério da Economia - INSS, Disponível em <https://dadosabertos.dataprev.gov.br/organization/instituto-nacional-de-seguro-social>. Acessado em 20/05/2020
- INSS Atendimento (2020), “INSS suspende atendimento presencial nas suas agências em todo o país”, Ministério da Economia - INSS, Disponível em <https://www.gov.br/inss/pt-br/assuntos/noticias/inss-suspende-atendimento-presencial-nas-suas-agencias-em-todo-o-pais>. Acessado em 08/05/2020.
- INSS (2017), “Institucional”, Ministério da Economia - INSS, Disponível em: <https://dadosabertos.dataprev.gov.br/organization/about/instituto-nacional-de-seguro-social>. Acessado em: 10/05/2020
- INSS 2019, “Resultado do Regime Geral de Previdência Social – RGPS”, Ministério da Economia - INSS - Secretaria de Previdência - Subsecretaria do Regime Geral de Previdência Social, Disponível em: <https://www.gov.br/trabalho-e-previdencia/pt-br/aceso-a-informacao/dados-abertos/dados-abertos-previdencia/previdencia-social-regime-geral-inss/ResultadodoRGPS201912urbanorural.pdf>. Acessado em: 08/05/2020.
- INSS Notificação (2020), “INSS inicia notificação de beneficiários após revisão administrativa”, Ministério da Economia - INSS, 2020, Disponível em <https://www.gov.br/inss/pt-br/assuntos/noticias/inss-inicia-notificacao-de-beneficiarios-apos-revisao-administrativa>. Acessado em 05/09/2020.
- INSS Plano (2020), “Plano de Dados Abertos - Julho/2016-Julho/2018”, Ministério da Economia -INSS, Disponível em <https://www.gov.br/inss/pt-br/centrais-de-conteudo/publicacoes/relatorios/plano-de-dados-abertos-do-inss-pdf/view>. Acessado em 08/05/2020.
- INSS Resolução (2019), “Resolução publicada pelo INSS vai acelerar concessão de benefícios”, 12/08/2019, Disponível em <https://www.gov.br/inss/pt-br>

[br/assuntos/noticias/resolucao-publicada-pelo-inss-vai-acelerar-concessao-de-beneficios](https://www.gov.br/assuntos/noticias/resolucao-publicada-pelo-inss-vai-acelerar-concessao-de-beneficios). Acessado em 08/05/2020

INSS Fraudes (2019), “Medida provisória combate fraudes e melhora a qualidade dos gastos na Previdência”, Ministério da Economia - INSS, Disponível em <https://www.gov.br/fazenda/pt-br/assuntos/noticias/2019/janeiro/medida-provisoria-combate-fraudes-e-melhora-a-qualidade-dos-gastos-na-previdencia>. Acessado em: 10/05/2020.

INSS Boletim (2020), “Boletim Estatístico da Previdência Social - Volume 25 Nr 4”, Ministério da Economia - Secretaria de Políticas de Previdência Social, Abril 2020, Disponível em https://www.gov.br/trabalho-e-previdencia/pt-br/aceso-a-informacao/dados-abertos/dados-abertos-previdencia/previdencia-social-regime-geral-inss/arquivos/Beps042020_trab_Final_sem_TMC.pdf. Acessado em: 10/05/2020.

INSS Treinamento (2019), “Instituto Nacional do Seguro Social, Treinamento em Serviço em Reconhecimento Inicial de Direito e Cadastro para Concessionários”, Ministério da Economia - INSS, Disponível em <https://tinyurl.com/432mwyf5> Acessado em 20/05/2020.

IPEA 2009, “Acompanhamento e análise - Vinte Anos da Constituição Federal - Volume 1”, IPEA - Diretoria de Estudos e Políticas Sociais, Políticas Sociais, Disponível em: https://www.ipea.gov.br/portal/index.php?option=com_content&view=article&id=5609&catid=305. Acessado em: 10/05/2020.

Lundberg, S. e Lee, S.I. (2017). “A Unified Approach to Interpreting Model Predictions”, Disponível em: https://www.researchgate.net/publication/317062430_A_Unified_Approach_to_Interpreting_Model_Predictions, Acessado em: 08/08/2021.

Mitchel, T. (1997), “Machine Learning”. New York; McGraw-Hill Science/Engineering/Math; (March 1, 1997)

Molnar, C. (2019) "Interpretable machine learning. A Guide for Making Black Box Models Explainable", Disponível em: <https://christophm.github.io/interpretable-ml-book>, Acessado em: 08/05/2020.

MP-DF 2019, “Ação Civil Pública”, Ministério Público Federal, Procuradoria da República no DF, 31/07/2019, Disponível em: <http://www.mpf.mp.br/df/sala-de-imprensa/docs/inicial-acp-serv-inss.pdf>. Acessado em: 02/05/2020.

Pacheco Junior, João Carlos. Modelos para detecção de fraudes utilizando técnicas de aprendizado de máquina / João Carlos Pacheco Junior. - 2019. Disponível em: https://bibliotecadigital.fgv.br/dspace/bitstream/handle/10438/27166/Dissertacao_Joao_Carlos_Pacheco_VFinal_2.pdf. Acessado em: 02/05/2020.

Rangel, L. A. e Caetano, M.A.R. (2015), “Previdência Social, Políticas Sociais: Acompanhamento e Análise”, Artigos - Nr 2, IPEA, Disponível em: <http://repositorio.ipea.gov.br/handle/11058/4225>. Acessado em: 10/05/2020

Senado Federal (2008), “Direitos da Seguridade Social”, Coleção Direitos Sociais - Volume II, Lei Orgânica da Seguridade Social - Lei nº8212 de 24/07/1991 / PL - Poder Legislativo Federal (D.O.U. 25/07/1991), Disponível em:

<https://www2.senado.leg.br/bdsf/bitstream/handle/id/496324/000977786.pdf?sequence=1&isAllowed=y>. Acessado em: 10/05/2020.

Tesouro (2019), “Boletim Resultado do Tesouro Nacional (RTN)”, Ministério da Economia – Secretaria Especial de Fazenda - Secretaria do Tesouro Nacional, Disponível em:

https://sisweb.tesouro.gov.br/apex/f?p=2501:9::::9:P9_ID_PUBLICACAO:31547.

Acessado em: 08/05/2020.