Predição de fidelização de doadores de sangue utilizando algoritmos de classificação

Fernando Wagner B. H. Filho[1], Bruno de C. Leal[2], Ana G. de Almeida Fernandes[3], Poliana L. dos Santos Campelo[3], Tiago da S. Vinuto[4], Nathália L. Pedrosa[3]

[1]Instituto Federal de Brasília (IFB) - Brasília, DF - Brasil

[2]Instituto Federal do Piauí (IFPI) - Floriano, PI - Brasil

[3]Fundação Hemocentro de Brasília (FHB) - Brasília, DF - Brasil

[4]Universidade Federal do Ceará (UFC) - Fortaleza, CE - Brasil

[1]fernando.filho@ifb.edu.br, [2]brunoleal@ifpi.edu.br, [3]{poliana.campelo, nathalia.pedrosa, ana.fernandes}@fhb.df.gov.br, [4]tiagovinuto@gmail.com

Abstract. *Blood donation is an altruism act capable of save many lives. Lack of blood is a society recurrent problem, and seems to have gotten bigger with COVID-19 pandemic. Health institutions usually promote campaigns that target to grow donors recruitment. Actions like this is important to maintain levels in line with populational demand. This work presents the use of machine learning techniques to predict loyalty blood donor. It´s a work in progress with promising preliminary results and it´s possible that will help in a blood donor recruitment and loyalty.*

Resumo. *Doar sangue é um ato de solidariedade e capaz de salvar inúmeras vidas. A falta de sangue é um problema recorrente na sociedade, que parece ter sido agravada com o advento da pandemia da COVID-19. Instituições de saúde costumam promover campanhas periodicamente no intuito de aumentar a captação de doadores. Ações como essa são importantes para manter os estoques condizentes com a demanda populacional. Este trabalho apresenta a utilização de técnicas de machine learning para predição de fidelização de doadores de sangue. Trata-se de um trabalho em andamento com resultados preliminares promissores e que possibilitará auxiliar em políticas de captação e fidelização de doadores.*

1. Introdução

A doação de sangue representa um ato solidário e de amor ao próximo, possibilitando a recuperação de pessoas acometidas de acidentes, cirurgias e doenças. Em casos mais extremos, a disponibilidade de sangue pode fazer a diferença entre a vida e a morte de um determinado paciente. As transfusões de sangue podem ser feitas de maneira total (ex: em casos de hemorragia), ou apenas de alguns de seus componentes, como eritrócitos, plaquetas ou plasma sanguíneo (ex: tratamento de queimaduras e anemias). Em ambos os casos, é extremamente desejável que tais insumos estejam disponíveis nos bancos de sangue.

Segundo [Ministério da Saúde 2020], aproximadamente 1,6% da população brasileira é doadora de sangue. Apesar de estar acima da recomendação da OMS (que pelo menos 1% da população seja doadora), o Brasil possui dificuldades em manter os

estoques condizentes com a demanda existente. Com o advento do isolamento social, fruto das medidas contra a pandemia de COVID-19, os níveis de doação sofreram uma queda. Estima-se que no início de 2021 houve uma redução da doação de sangue no nosso país entre 15% e 20% em comparação com o ano que antecedeu a pandemia [Silva et al 2021]. Não há dúvidas de que é importante que haja um esforço para difundir a correta informação e conscientização sobre o processo e a importância de se doar sangue perante a população. Como resposta a essa situação, governos costumam promover de tempos em tempos as campanhas de doação de sangue, com o objetivo de intensificar a coleta de sangue, dando ênfase também à questão educacional. Neste contexto, é de extrema importância aumentar a captação de novos doadores e, principalmente, estimular a fidelização e o crescimento dos chamados doadores de repetição, de forma a resultar em um estoque mais adequado para as necessidades dos hospitais e hemocentros. Em paralelo a isso, a área de aprendizagem de máquina vem cada vez mais ganhando notoriedade por oferecer soluções computacionais, apoiadas na inteligência artificial e na estatística, com potencial de promover uma melhora em tomadas de decisão, agregando valor à instituições das mais diversas naturezas.

O objetivo deste trabalho é utilizar algoritmos de aprendizagem de máquina (*machine learning*), para tentar predizer doadores de sangue com tendência a fidelização. Entende-se por fidelizado aquele indivíduo que doa sangue pelo menos duas vezes em um intervalo de 12 meses. Trata-se um problema de classificação binária (doador fidelizado e não-fidelizado), em que será utilizada aprendizagem supervisionada. Para realização deste trabalho, está sendo utilizado um conjunto de dados (*dataset*) com registros de doadores de um hemocentro público localizado no Distrito Federal. É importante salientar que este estudo foi apreciado e aprovado por um comitê de ética em pesquisa (CAEE 40370820.5.0000.5553) e obedece às normas de pesquisa envolvendo seres humanos contidas na Resolução Nº 466/2012. O restante deste trabalho está dividido como se segue: a Seção 2 aborda a metodologia que vem sendo utilizada, a Seção 3 contém os resultados preliminares e discussões, e a Seção 4 apresenta as considerações finais e trabalhos futuros.

2. Metodologia

Esta seção apresenta a metodologia que vem sendo aplicada ao longo do desenvolvimento deste trabalho, consistindo de quatro etapas e ilustrada na Figura 1.

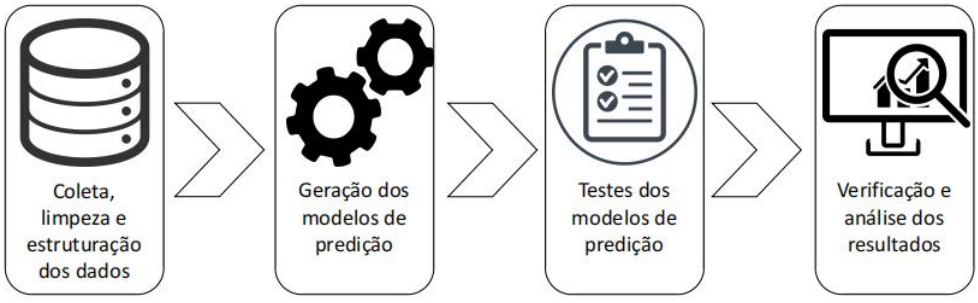

Figura 1. Visão geral da metodologia utilizada.

Na primeira etapa do trabalho (Coleta, limpeza e estruturação dos dados), foram coletados dados 153 registros de doadores fidelizados e 133 de doadores não fidelizados (primeira vez ou frequência esporádica), totalizando 286 registros. Isto representa uma amostra com 95% de nível de confiança. A seleção dos doadores foi realizada através de

sorteio (amostra aleatória simples), havendo dupla checagem dos dados de doação para correta classificação. Após coletados os dados, estes foram estruturados em uma planilha eletrônica tabular, onde cada coluna representa uma variável (*feature*) e as linhas contém os dados dos registros. As variáveis preditoras utilizadas na coleta foram: sexo, raça/cor, escolaridade, faixa de renda familiar, se reside fora da cidade, tipagem sanguínea, faixa etária de idade, se houve motivação por ajudar desconhecidos, se houve motivação por ajudar conhecido, se houve motivação por obter isenção em inscrição de concurso público, se houve motivação por realizar exames de sangue, como avalia o atendimento do hemocentro, e o grau de confiança no serviço do hemocentro. Como variável de desfecho, há apenas se o doador é fidelizado ou não. É importante salientar que todas estas variáveis são categóricas e foram devidamente normalizadas de forma a garantir o correto funcionamento dos algoritmos de geração dos modelos.

A segunda etapa consistiu da escolha e execução dos algoritmos de classificação para a geração dos modelos preditivos. Nesta etapa, utilizou-se a estratégia de *K-Fold* com *Cross-Validation* [Refaeilzadeh and Tang and Liu 2009]. Esta estratégia consiste de dividir o conjunto de dados em k partes (ex: 5 partes) mutuamente exclusivos e de tamanho similar. Após isto, deve-se realizar um número *k* de execuções, sendo que em cada um delas, os dados de treino e teste utilizados serão diferentes, sendo um parte para teste e as demais usadas para treino dos modelos. Ao final das *k* execuções, calcula-se a média aritmética e o desvio padrão dos resultados obtidos. A Figura 2 ilustra essa etapa para um $k = 5$.

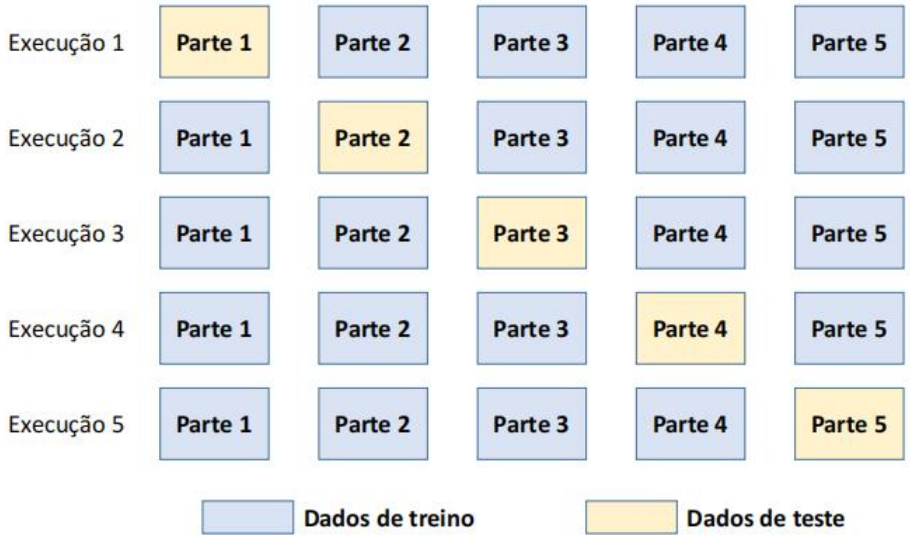

Figura 2. Estratégia *K-Fold* com *Cross-Validation*

Para a implementação do ambiente, foi utilizada a linguagem *Python*[1], em conjunto com as bibliotecas *Pandas*[2], para manipulação dos dados da planilha, e *scikit-learn*[3], para geração dos modelos através dos classificadores. Também está sendo considerada a operação de balanceamento dos dados, que consiste de aproximar o número de registros de doadores fidelizados e não-fidelizados nos dados de treino.

[1] https://www.python.org/
[2] https://pandas.pydata.org/
[3] https://scikit-learn.org/

Utilizando-se das mesmas bibliotecas, realizou-se a terceira etapa, que consistiu dos testes dos modelos de predição. Nesta etapa foram auferidos os resultados utilizando todas as variáveis preditoras, bem como um subconjunto destas, fruto da utilização de estratégias de redução/seleção de variáveis (*feature selection*).

Na quarta última etapa, foram utilizadas algumas métricas frequentemente utilizadas em problemas de classificação para avaliação dos resultados dos testes [Bruce and Bruce 2019]. Dentre estas métricas, citam-se a acurácia (*A*) precisão (*P*), revocação/sensibilidade (*R*), medida-F (*F*). Tais métricas são calculadas a partir da chamada matriz de confusão, que consiste de uma matriz de dimensões 2x2 nas quais armazenam os verdadeiros positivos e verdadeiros negativos (doadores fidelizados e não-fidelizados preditos corretamente), além dos falsos positivos e falsos negativos (doadores fidelizados e não-fidelizados preditos incorretamente), conforme ilustrada na Tabela 1. Em uma matriz de confusão, quanto maiores os valores da diagonal principal, mais assertivos são os modelos.

Tabela 1. Matriz de confusão

	Não-fidelizado	**Fidelizado**
Não-fidelizado	Verdadeiros negativos (V_n)	Falsos positivos (F_p)
Fidelizado	Falsos Negativos (F_n)	Verdadeiros positivos (V_p)

A partir da matriz de confusão, são calculadas as métricas anteriormente citadas, conforme respectivas fórmulas:

$$A = \frac{V_p + V_n}{V_n + F_n + V_p + F_p} \quad (1) \qquad P = \frac{V_p}{V_p + F_p} \quad (2) \qquad R = \frac{V_p}{V_p + F_n} \quad (3)$$

$$E = \frac{V_n}{V_n + F_n} \quad (4) \qquad F = \frac{2 \times (P \times R)}{(P + R)} \quad (5)$$

A acurácia (Eq. 1) representa o total de acertos do modelo; a precisão (Eq. 2) consiste do número de doadores corretamente classificados como fidelizados; a revocação mede a quantidade de doadores fidelizados corretamente identificados em relação aos demais da amostra; a especificidade (Eq. 4) aufere a quantidade de doadores não-fidelizados corretamente classificados; e, por fim, a medida-F (Eq. 5), que representa uma média harmônica entre precisão e revocação.

3. Resultados preliminares e discussões

Para a geração dos resultados, foi considerada a estratégia de *K-Fold* com *cross-validation*, sendo $k = 5$, percentual de registros para treino 80%, e para teste 20%. Além disso, foram utilizados os seguintes algoritmos de classificação: *Decision Tree (*DT*), K-Nearest Neighbors (*KNN*), GradientBoosting (*GB*), Support Vector Machine (*SVM*), Nayve Bayes (*NB*), Logistic Regression (*LR*), Random Forest (*RF*)*, e a rede neural do tipo *Multilayer Perceptron (*MLP*)*. Os valores de média (μ) e desvio-padrão (σ) para as métricas de acurácia, precisão, revocação, especificidade e medida-F, oriundas das cinco execuções encontram-se na Tabela 2.

Conforme pode-se constatar, os níveis obtidos para as métricas são maiores que 0,5, para a maioria dos casos, o que significa que a utilização das soluções de

classificação tendem a ser melhores que a escolha aleatória de candidatos a fidelização. No entanto, percebe-se menores níveis de especificidade nos modelos, denotando certa dificuldade em classificar corretamente doadores não-fidelizados. Posto que é bastante importante descobrir doadores que tendem a ser fidelizados, pode-se dizer que bons níveis de revocação são desejáveis, uma vez que diminui a probabilidade de se perder potenciais doadores fidelizados. O maior nível de acurácia se deu pelo classificador *Random Forest*; o de revocação foi obtido pelo *K-Nearest Neighbors* e; os de precisão, especificidade e medida-F foram auferidas pelo modelo referente ao *GradientBoosting*, tornando-o este o mais interessante classificador (em média) para as previsões.

Tabela 2. Resultados dos testes para cada classificador.

Classificadores	Acurácia	Precisão	Revocação	Especificidade	Medida-F
DT	μ = 0,57 σ = 0,07	μ = 0,59 σ = 0,05	μ = 0,59 σ = 0,09	μ = 0,51 σ = 0,05	μ = 0,61 σ = 0,06
KNN	μ = 0,58 σ = 0,04	μ = 0,57 σ = 0,04	**μ = 0,70** σ = 0,08	μ = 0,41 σ = 0,09	μ = 0,62 σ = 0,03
GB	μ = 0,60 σ = 0.02	**μ = 0,62** σ = 0,01	μ = 0,60 σ = 0,08	**μ = 0,53** σ = 0,03	**μ = 0,66** σ = 0,04
SVM	μ = 0,58 σ = 0,07	μ = 0,59 σ = 0,02	μ = 0,69 σ = 0,11	μ = 0,44 σ = 0,05	μ = 0,65 σ = 0,08
NB	μ = 0,58 σ = 0,03	μ = 0,58 σ = 0,04	μ = 0,69 σ = 0,06	μ = 0,49 σ = 0,13	μ = 0,62 σ = 0,06
LR	μ = 0,56 σ = 0,08	μ = 0,61 σ = 0,01	μ = 0,68 σ = 0,07	μ = 0,44 σ = 0,14	μ = 0,64 σ = 0,07
RF	**μ = 0,61** σ = 0,02	μ = 0,58 σ = 0,03	μ = 0,64 σ = 0,09	μ = 0,44 σ = 0,07	μ = 0,65 σ = 0,06
MLP	μ = 0,57 σ = 0,03	μ = 0,60 σ = 0,05	μ = 0,69 σ = 0,03	μ = 0,40 σ = 0,09	μ = 0,65 σ = 0,03

Com relação às estratégias de seleção de variáveis, foi utilizada o cálculo do grau de impureza de *Gini* [Bruce and Bruce 2019], presente na biblioteca scikit-learn (*feature_importances_*), a partir do modelo considerado o melhor em média (GB). Foram destacadas os 5 variáveis de maior importância (grau de importância ≥ 10%), a saber: faixa de idade, faixa de renda, tipo sanguíneo, grau de confiança no serviço do hemocentro, e se houve motivação por ajudar conhecido. A Tabela 3 mostra os resultados para o modelo gerado a partir destas variáveis. Observa-se uma leve melhora nos índices de acurácia, precisão e revocação (em destaque).

Tabela 3. Resultados para o *GradientBoosting* considerando seleção de variáveis.

Classificador	Acurácia	Precisão	Revocação	Especificidade	Medida-F
GB (5 variáveis)	**μ = 0,63** σ = 0.04	**μ = 0,64** σ = 0,02	**μ = 0,72** σ = 0,07	μ = 0,53 σ = 0,13	μ = 0,66 σ = 0,03

É possível encontrar alguns trabalhos com utilização de aprendizagem de máquina voltada para predição de doação de sangue ([Alajrami et al. 2019], [Silva 2018]), tendo a maioria alcançado níveis de acurácia, precisão, revocação e especificidade entre 50% e 80%, evidenciando a dificuldade de se conseguir resultados excepcionais para este tipo de problema. Além disso, estes trabalhos versam sobre a

probabilidade de receber doações futuras, podendo estas serem esporádicas, e não necessariamente fidelizadas. Portanto tais trabalhos não tratam a questão da fidelização de maneira específica. Considera-se então neste cenário, que o presente trabalho possui relevância, bem como potencial para alcançar melhores resultados.

4. Considerações Finais e trabalhos futuros

O presente trabalho apresentou uma abordagem para a construção de modelos de predição para fidelização de doadores de sangue. O problema da descoberta de potenciais doadores fidelizados foi definido como um problema de classificação binária, e os modelos de predição foram construídos a partir dos dados de um hemocentro público localizado no Distrito Federal.

Para os experimentos realizados, foram utilizados alguns dos principais algoritmos de classificação, bem como uma estratégia para a seleção de variáveis. Foi averiguado que, em média, os modelos construídos se comportam melhor que a escolha aleatória, e que a seleção de variáveis gerou um modelo com melhora nos resultados. Ressalta-se que este é um trabalho em andamento e que há bastante potencial para a melhora dos níveis até aqui obtidos para as métricas. Como trabalhos futuros, está programado a coleta de mais dados de doadores (fidelizados e não-fidelizados), testes com hiperparametrização junto aos classificadores até aqui utilizados, e a inclusão de novos classificadores, como o *CatBoost* [Prokhorenkova et al 2018].

Referências

Alajrami, E. et al. (2019). “Blood donation prediction using artificial neural network”. In: International Journal of Academic Engineering Research (IJAER), v. 3, n. 10, p. 1-7.

Bruce, P. and Bruce, A. (2019) “Estatística Prática Para Cientistas de Dados - 50 Conceitos Essenciais”. Alta Books.

Ministério da Saúde. (2020) “Apresentação - Seja Solidário, Doe Sangue”, Disponível em: https://www.gov.br/saude/pt-br/centrais-de-conteudo/apresentacoes/2020/12-06-2020-campanha-20doao-20de-20sangue-aprovado-final-pdf/@@download/file/12-06-2020_campanha-20doacao-20de.pdf. Acesso em 02 mar de 2022.

Prokhorenkova, L. et al (2018) “CatBoost: unbiased boosting with categorical features”, In: Advances in Neural Information Processing Systems, Edited by S. Bengio and H. Wallach and H. Larochelle and K. Grauman and N. Cesa-Bianchi and R. Garnett, v. 31, Curran Associates, Inc.

Refaeilzadeh P. and Tang L. and Liu H. (2009) Cross-Validation. In: LIU L., ÖZSU M.T. (eds) Encyclopedia of Database Systems. Springer, Boston, MA. https://doi.org/10.1007/978-0-387-39940-9_565.

Silva, D.H. et al. (2021) “Perfil dos doadores novos de sangue do hemocentro de Ribeirão Preto/SP”, In: Hematology, Transfusion and Cell Therapy, v. 43, n. 1, p. S352, Editora Elsevier.

Silva, F. H. da (2018), “Estudo e desenvolvimento de métodos para predição de doadores de sangue”, Universidade Federal de Goiás.