Análise de Perfil de Cliente para Recuperação de Crédito com Aprendizado de Máquina

Guilherme B. Carvalho[1], **André Constantino da Silva**[1], **Adriano da Silva Ferreira**[1]

[1]Instituto Federal de Ciência e Tecnologia de São Paulo (IFSP)
Avenida Thereza Ana Cecon Breda, N.º 1896,
Vila São Pedro – 13183-250 – Hortolândia – SP – Brasil

gbordotti20@gmail.com, andre.constantino@ifsp.edu.br,

adriano.ferreira@ifsp.edu.br

Abstract. *It's known default in Brasil is represented by 40% of adult population, consequently financial institutions try to recover their credit and this process implements several technologies to achieve the debtor even thought not them all have tendency to pay, and recognize who has more tendency could create opportunities, so a lot of tecniquies, one of them is statistical analisys. This paper uses machine learning and describle how was implemented an neural network for predict a profile between tendency to pay or not as result was 85.1% of precision and 79.6% of accuracy based on 8.000 contracts processed.*

Resumo. *A inadimplência no Brasil é representada por aproximadamente 40% dos brasileiros adultos. Diante disso, instituições financeiras buscam retomar o crédito cedido por processos de recuperação de crédito que utilizam diversas tecnologias para entrar em contato com o devedor. Contudo nem todos os devedores possuem propensão ao pagamento, desta forma identificar os que possuem a propensão ao pagamento pode gerar oportunidades e para isso existem técnicas como aprendizado de máquina a qual este trabalho se propôs a aplicar neste contexto com a implementação de uma rede neural para criação de um modelo capaz de predizer os perfis de clientes entre a propensão a pagamento ou não pagamento, por fim obteve-se uma rede neural com 85,1% de precisão e acurácia de 79,6% no processamento de 8.000 contratos.*

1. Introdução

Um assunto que é comumente falado e conhecido por muitos e envolve aproximadamente 40% da população adulta brasileira é a inadimplência, esta que, de acordo com [Serasa 2018a], tem como as principais causas no Brasil: Desemprego (26%), Redução de renda (14%) e Descontrole Financeiro (11%). Além destas causas um ponto que se torna agravante é a compra por impulso [Cerbasi 2018] que certas vezes são realizadas sem necessidade ou para suprir um "status social"[Alcoforado et al. 2019], em certos casos também pela falta de educação financeira [Claudino et al. 2009].

No Brasil existe um alto volume de inadimplência. Conforme o registro da Serasa Experiean, em agosto de 2018 haviam 61,5 milhões de brasileiros inadimplentes no país [Serasa 2018b] e, comparando julho de 2018 com julho de 2019, é observado ainda um crescimento de 2,7%, elevando para 63,3 milhões de pessoas com contas atrasadas

[Stephanie 2019] e, segundo pesquisa da Confederação Nacional do Comércio de Bens, Serviços e Turismo [Confederação Nacional do Comércio de Bens 2020], em junho de 2019 64,0% das famílias brasileiras possuíam dívidas, contudo em junho de 2020 este número saltou para 67,1% com crescimento de 3,1% e, dos que responderam a pesquisa em 2020, 16,1% se consideram muito endividado.

Há também o estudo [Tiryakia et al. 2017] que buscou relacionar o balanço do PIB com a da inadimplência, no qual foram analisadas pessoas jurídicas e foi possível observar que, em momentos de recessão, a inadimplência não cresce da forma como ocorre em momentos de melhora do meio econômico, dado que é percebido um afrouxamento nas regras e exigências das empresas credoras quando há uma melhora na economia do país, expondo-se assim a maiores riscos, refletindo então em maior concessão de crédito.

Com tal inadimplência as empresas credoras buscam uma forma de reaver seu crédito, neste caso elas mesmas pode entrar em contato com o devedor ou podem contratar uma empresa terceira para realizar tal serviço que é descrito como recuperação de crédito. Existem empresas especializadas para realizar este tipo de trabalho, onde consiste em contatar o devedor e convencê-lo a devolver o crédito concedido, podendo então concretizar a devolução por completo ou não.

O processo de recuperação de crédito tem como princípio o contato com o cliente buscando um acordo de forma amigável para o pagamento da dívida em atraso. Este processo engloba técnicas de persuasão além de conhecimento do cliente do contatado. Para a estimativa do perfil de forma superficial leva-se um tempo médio de 3 minutos analisando os dados disponíveis, tal estimativa pode ser muitas vezes imprecisa de acordo com o nível de experiência do consultor. Análise esta que, ao se multiplicar no volume de contratos a se trabalhar, torna lento o processo além da exaustão.

Contudo não é tão simples concretizar a devolução por completo devido aos tramites legais, sua burocracia, que geram custos neste processo. Então escolher um cliente ou contrato atrelado ao cliente que tenha uma maior probabilidade de trazer o resultado esperado é uma forma de se otimizar o trabalho.

No mercado existem ferramentas que realizam diversas análises estatísticas dos dados buscando classificar estes clientes, para isso, profissionais que realizam este tipo de trabalho se baseiam em um conjunto de dados (do inglês *dataset*) do cliente e do contrato que possam apontar um padrão e assim subdividir estes com rótulos esperados, neste caso os rótulos buscados são clientes que possuem propensão a pagamento e os que não possuem. Dada complexidade de se identificar padrões nestes *datasets* os custos de ferramentas que realizam tais tarefas automatizadas é muito elevado, não sendo viável para a maioria das pequenas e médias empresas.

Estes conjuntos de dados em sua maioria possuem informações como: valor de financiamento, valor da entrada, parcelas, valor das parcelas, garantias, em alguns casos podem ou não possuir informação como sexo, cidade, unidade federativa (estado), idade e profissão. O desafio então é realizar o pré-processamento de forma adequada e identificar um modelo que classifique estes clientes utilizando estes dados e buscando quais atributos melhor descrevem a classe a qual estes pertencem.

Assim o objetivo deste trabalho foi aplicar métodos de classificação com aprendizado de máquina para auxílio da recuperação de crédito buscando classificar o perfil dos

clientes. Para isso foi utilizado o aprendizado de máquina, ideia que vem sendo desenvolvida desde 1960, tornando-se possível aplicar em decorrência do aumento na capacidade de processamento dos computadores na atualidade.

A execução manual do processo de predição de perfil é de alta complexidade e os dados para análise podem se tornarem obsoletos em poucos meses dependendo do cenário sócio-econômico que estão contidos, tornando inviável para execução de forma manual devido a seu alto custo e tempo para desenvolvimento. Assim, neste projeto o objetivo foi desenvolver uma aplicação que utiliza-se de dados reais coletados previamente com seus devidos resultados esperados e treinar uma máquina, validando-a.

Na Seção 2 deste trabalho é apresentado o referencial teórico onde haverá base para contextualização dos assuntos que serão abordados no desenvolvimento deste. Os trabalhos correlatos são apresentados na Seção 3 e buscam resolver o mesmo problema com abordagens parecidas ou outros problemas com a mesma abordagem deste. A seguir, na Seção 4, são apresentados materiais, métodos, ferramentas e tecnologias utilizadas durante o desenvolvimento deste trabalho. O descritivo do trabalho realizado é apresentado na Seção 5 com maiores detalhes do processo realizado. Por fim, mas não menos importante é apresentado a Seção 6 com a conclusão e trabalhos futuros.

2. Referencial Teórico

Desde 1950 vem se desenvolvendo Redes Neurais Artificiais (RNAs) baseadas em neurônios biológicos na busca de classificação e de reconhecimento de padrões, dado que esta tecnologia possibilita reduzir modelos matemáticos com alta densidade de probabilidade para cada atributo [Costa et al. 1999].

Estas redes podem ter arquiteturas de aprendizado supervisionado ou não-supervisionado, onde o primeiro é dado um volume de dados com seus resultados esperados para validação e o segundo a rede classifica os dados sem um direcionamento específico [Kovács 2002]. Este trabalho irá utilizar o modo supervisionado.

O aprendizado de máquina supervisionado é um processo de ensinar ao computador um comportamento dado a apresentação de exemplos com dados de entrada e seus resultados esperados, então com o treinamento de um algoritmo a máquina irá criar um modelo que consiga reproduzir de forma aproximada tais resultados quando lhe for fornecido dados para serem classificados [Domingos 2017].

Um exemplo de aprendizado supervisionado é a identificação de números em escrita manual onde o treinamento do modelo ocorre por uma série de imagens e seus respectivos valores, em seguida é realizada a predição onde, dado um modelo treinado para aqueles valores, é passado uma nova imagem, porém desta vez a máquina buscará identificar o padrão desta e o resultado será um número que possui a melhor aproximação encontrada.

Como forma de identificar a melhor combinação entre atributos e quantidade de linhas do *dataset* a ser utilizado no treinamento a métrica de coeficiente de correlação de Pearson é utilizada[Stephanie 2013] cuja formula é apresentada na Equação 1, onde mostra quão próximo e interligados são os dados. Por meio da aplicação da Equação 1 obtem-se uma variação de -1 até 1 cujo resultado representa 3 estados:

- Para valores mais próximos de -1 indica que os valores são inversamente correla-

cionados, ou seja, quando x é ascendente, y será descendente dentro da proporção apresentada e vice-versa.

- Para o caso do coeficiente apresentar-se próximo a 0 este possui baixa correlação ou nenhuma, demonstrando independência entre si.
- No momento que o resultado apresentar valores próximos a 1 mostra que x e y variam na mesma direção.

Conhecendo tais comportamentos, quando o coeficiente dos dados apresentam-se próximos as extremidades estima-se uma melhor predição.

$$r = \frac{n(\sum xy) - (\sum x)(\sum y)}{\sqrt{[n(\sum x^2) - (\sum x)^2][n(\sum y^2) - (\sum y)^2]}} \tag{1}$$

Afim de melhorar a visualização dos dados e também do diagrama de correlação criado a partir do cruzamento de atributos com a fórmula acima foi utilizada a biblioteca chamada Matplotlib [Tosi 2009]. Tal biblioteca permite a criação de gráficos 2D e 3D em diferentes formatos dando uma ampla gama de configurações opcionais a serem incrementadas.

Para criação de um diagrama de correlação foi-se decidido utilizar a biblioteca Seaborn [Waskom et al. 2015] com o método *heatmap* por apresentar uma maior agilidade na construção da visualização.

Empregou-se neste trabalho outros dois gráficos que são o *scatter plot* e o *boxplot*. No gráfico *scatter plot*, também conhecido como gráfico de dispersão, é usado para analisar a relação entre duas variáveis (uma no eixo x e outra no eixo y), permitindo analisar se a correlação entre as variáveis é forte ou fraca: quando traçado uma linha média entre os pontos e indicar um aumento no eixo x e y com menor dispersão maior a correlação porém quanto maior a dispersão também menor a correlação; também quando os valores no eixo x aumentam e no ponto y reduzem indicam uma correlação negativa porém também ligado a menor dispersão mais forte a corrrelação e vice-versa.

O segundo gráfico, *bloxplot*, apresenta uma caixa demonstrando a distribuição dos dados com a mediana dos valores e a maior recorrência dentro da caixa; a linha acima e a linha abaixo é onde se encontra mais uma parte com menor frequência dos dados e as bolinhas fora das linhas são considerados *outliers*, pontos fora da curva também que não representam o comum do conjunto de dados. Abaixo segue uma representação gráfica.

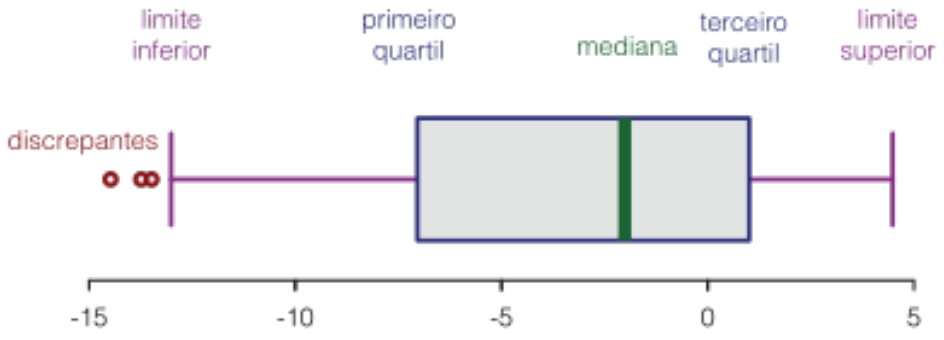

Figura 1. Elementos de um *boxplot*.

Um *heatmap*, também conhecido como mapa de calor, dispõe as variáveis tanto no eixo x quanto no eixo y e, se utiliza cores para facilitar a identificação de informações. Quando combinado com os resultados de um mapa de correlação, as cores indicam se as variáveis possuem correlação forte ou fraca. A forma de leitura do *heatmap*, também conhecido como mapa de calor, em conjunto com um mapa de correlação começa pela identificação de cores mais escuras o que significa que são valores mais altos que os demais. Pela legenda também é possível identificar que os valores de colunas são os mesmos das linhas pois cada valor refere-se a correlação entre os dois atributos, ainda quando dito sobre correlação haverá uma linha diagonal com valores sempre iguais a 1 pois este é o cálculo de correlação entre os mesmo valores o que deve ser desconsiderado para análise, por fim quanto mais próximos os resultados forem de 1 ou -1 indicam uma maior correlação entre os atributos o que para a rede neural se torna mais fácil de classificar tais clientes por haverem padrões mais claros.

Existe também a necessidade de redução da dimensionalidade dos dados quando há grandes volumes de variáveis impossibilitando a visualização dos dados e também podendo retardar o treinamento de uma rede ou dificultar o treinamento e predição podendo ainda ocorrer um *overfitting*, este indica que a predição está tão próxima dos dados treinados que ao inserir novos dados ela pode ter um resultado abaixo do esperado e até muito diferente dos dados de teste.

Foi testado também o *Principal Component Analysis*, ou simplesmente PCA, para redução de espaço. Este essencialmente visa a redução das dimensões das variáveis, podendo ser executado de duas formas diferentes [Jolliffe 2002]: (1) pela *Feature Elimination*, que exclui um conjunto de colunas e mantém somente as desejadas, ação esta que pode não ser vantajosa visto que atributos menos relativos não terão influência nenhuma no treinamento quando em algum momento estes teriam sua relevância no refinamento do treinamento. Algo que pode então suprir tal necessidade é a (2) *Feature Extraction* que busca manter o máximo possível da variação dos dados e gerando então um novo conjunto com variáveis independentes mas estas são criadas a partir da combinação dos atributos originais [Brems 2017].

Outro objetivo na utilização do PCA foi gerar as visualizações dos dados utilizados no trabalho em conjunto com a biblioteca *Matplotlib* nos gráficos 3D de pontos (*scatter plot*).

A Figura 2 exemplifica o fluxo de dados no treinamento e predição dentro de uma rede neural artificial, onde nos círculos a esquerda pintados de verde são os os atributos de entrada, nos laranjas ao meio são os neurônios nas camadas ocultas e em azul ao lado direito as camadas de saída.

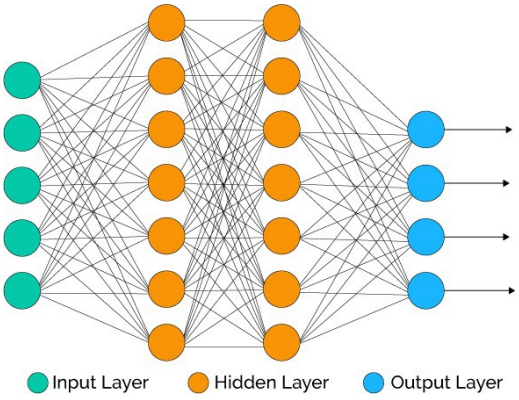

Figura 2. Fluxo de dados em uma rede neural.

Quantificando desta forma o modelo criado, ele pode ter 5 neurônios na camada de entrada representando os atributos, na camada oculta foi utilizado 7 neurônios em 2 camadas, podendo variar a quantidade de neurônios por camada e a quantidade de camadas de acordo com o modelo, e por fim na camada de saída são 4 neurônios representando suas classes de saída.

3. Trabalhos Correlatos

Nesta Seção são apresentados três trabalhos correlatos a este, sendo o primeiro com o objetivo de realizar a modelagem do risco de crédito, o segundo um estudo sobre a aplicação de rede neurais na predição de risco de crédito e o terceiro que investigou a extração de regras de redes neurais para avaliações de risco de crédito.

Para a Modelagem de Risco de Crédito existe uma padronização que diferencia o bom pagador do mau pagador a partir de modelos matemáticos. A partir desses modelos, o estudo desenvolvido por [Neto and Carmona 2002] utilizou métodos de análise discriminante e de regressão logística para predizer o risco e a segmentação de cliente incluídos em um *dataset*, sendo o segundo método de melhor resultado de precisão.

No estudo de [Pacelli and Azzollini 2011] é apresentada a comparação entre regressão logística e Rede Neural Artificial para análise do mesmo *dataset* a fim de se obter melhores resultados e apresentar suas diferenças. As conclusões obtidas foram que esta abordagem pode servir de suporte a modelagem do risco de crédito, contudo não substituir ainda os modelos tradicionais.

O trabalho de [Baesens et al. 2003] utilizou três diferentes algoritmos sendo eles, *Neurorule, Trepan, Nefclass*, a fim de tornar regras usadas na avaliação do risco de crédito claras para a visualização de pessoas leigas. Com os diferentes algoritmos e os resultados de cada análise concluiu-se que os modelos *Neurorule* e *Trepan* obtiveram os melhores resultados.

Com isso, o trabalho desenvolvido por [Neto and Carmona 2002] é o que mais se assemelha a este em seu desenvolvimento e objetivo, por buscar responder as mesmas perguntas e com cálculos estatísticos, podendo este ser considerado como um trabalho futuro ao realizado por Neto e Carmona.

4. Materiais e Métodos

A Figura 3 apresenta o fluxo de dados do projeto, desde a origem dos dados até a apresentação dos gráficos para análise.

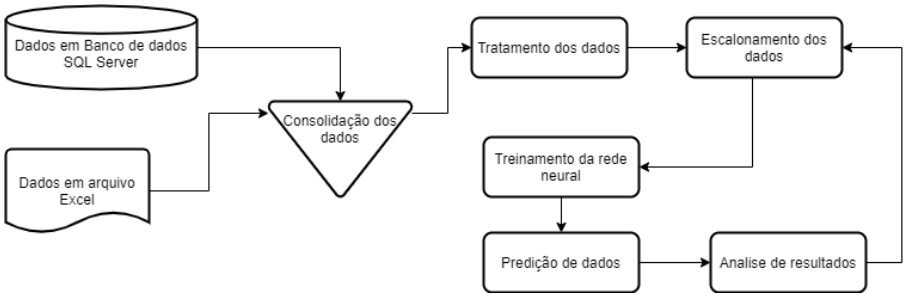

Figura 3. Fluxo de dados do projeto.

Inicialmente foi necessário extrair os dados de um banco de dados SQL Server com base em uma *query SQL* desenvolvida para este projeto contendo somente um dado identificador, que possibilita a conexão entre os dados existentes em uma planilha em formato .xls (formato de planilha eletrônica Excel). Os dados, embora estruturados, são armazenados separados, necessitando assim ser consolidados.

Também foi necessário realizar o tratamento destes dados pois haviam dados faltantes, o que prejudicava o desempenho da rede neural levando a identificação de padrões inexistentes; neste caso os registros foram descartados da base.

Em seguida também foi necessário realizar o escalonamento dos dados pois haviam números em diferentes dimensões, como idade entre 18 e 80 anos e valores de financiamento entre 10.000 e 1.000.000. Para o escalonamento foi aplicado o método *Standard Scaler* do *Sci-kit Learn* que realiza a subtração de uma unidade pela mediana e, em seguida, divide pelo desvio padrão, conforme apresenta a Equação 2, reduzindo a magnitude dos dados sem perder suas devida proporções.

$$r = \frac{(x - u)}{s} \tag{2}$$

Em seguida, foi realizada uma análise exploratória aplicando diferentes algoritmos de classificação a fim de identificar qual obteria melhor aproximação, para que então fosse utilizado nos testes finais buscando classificar o perfil de clientes entre as classes de não pagador e pagador, objetivo deste trabalho.

O fluxo de dados deste projeto foi programado em linguagem Python com as bibliotecas Pandas, Numpy, Sci-Kit Learn, Multithreading e Matplotlib, onde a biblioteca Pandas foi utilizada para carregar, tratar e filtrar os dados. A biblioteca Numpy [McKinney 2012] foi escolhida por ter um ótimo desempenho com números e foi utilizado para cálculos de correlação, mesclagem de forma randômica e cálculos de vetores. Da biblioteca Sci-Kit Learn(SK-Learn) foi utilizado algoritmos de classificação como o MLP Classifier, Random Forest Classifier e KNN Classifier. A biblioteca Multithreading foi utilizada para permitir processamento em paralelo no treinamento de modo a se ob-

ter mais rapidamente os resultados. A biblioteca Matplotlib foi utilizada para gerar os gráficos e visualizações.

Também foi utilizado o Google Colab, uma plataforma no formato de um *jupyter notebook* e possui hardware compartilhado para estudo havendo, assim, disponível uma GPU para processamento com bibliotecas que suportem o processamento por GPU.

Os dados utilizados foram disponibilizados por uma empresa privada no setor de recuperação de crédito para criação do modelo de predição e os algoritmos aplicados foram *Multilayer Perceptron (MLP)*, *K-nearest neighbors (KNN)* e *Random Forest Classifier* onde foi observado um melhor desempenho com a utilização do MLP. O nome da empresa foi mantido em sigilo devido aos dados sensíveis e acordo realizado para a realização deste trabalho.

Foi realizada uma pesquisa de campo com profissionais da área de recuperação de crédito afim de identificar quais atributos eram relevantes para uma análise manual e que poderiam ser utilizado nos algoritmos, assim foi buscado principalmente dados do contrato e do cliente sendo estes: valor de parcela, total de parcelas, percentual pago, valor de financiamento e valor da entrada, também dados qualitativos do cliente como sexo e estado cível.

Com apoio da entrevista foram selecionados os atributos que seriam utilizados no modelo, estes são atributos numéricos e categóricos o que torna necessário realizar o processo de codificação de dados categóricos e também aplicado a normalização ou escalonamento dos dados para facilitar o treinamento da rede, utilizando-se das ferramentas *Metrics*, *Standard Scaler*, *Normalizer*, *Neural Networks* do *Sci-Kit Learn*.

Para a visualização dos dados foi utilizado um algoritmo de redução de dimensões, o *PCA*, reduzindo-se para 3 dimensões assim sendo possível desenhar em um gráfico facilitando a visualização do conjunto dos dados e suas distribuições.

O método de treinamento conta com a estratégia de validação cruzada buscando resultados mais precisos com a variação dos parâmetros de cada algoritmo aplicado, quando utilizado a MLP é incrementado o número de neurônios, camadas e alteração do alpha. Para o *Kneighbors* é incrementado o número de clusters. Ao fim dos testes é escolhido um único algoritmo que será tratado com maior atenção para refinamento do resultado.

A medida de desempenho aplicada foram acurácia, precisão e *recall*. Onde a acurácia é medida pela divisão de resultados corretos pelo volume analisado apresentando a taxa de acertos. A precisão é o resultado entre os Verdadeiros Positivos(VP) dividido pelos Falsos Positivos(FP) mais os VP; tal métrica é importante quando se precisa reduzir os Falsos Positivos. O *Recall* é a divisão dos pontos VP pela divisão da soma dos VP com Falsos Negativos(FN) seu resultado aponta a quantidade de positivos marcados corretamente. Dentre as três medidas apresentadas anteriormente o melhor cenário é 1 para cada uma delas e o pior é 0.

Foram definidos 4 cenários e 8 variações no volume do *dataset*: nos cenários 1 e 3 aplicando-se a proporção 75% para treinamento e 25% para validação e com a amostra em ordem aleatória; e para o cenário 2 e 4 95% para treinamento e 5% para validação e com amostra sequencial por distribuição temporal.

Esta abordagem foi escolhida dado que em um caso de utilização em produção os dados para predição seriam apenas 3 a 5% do volume utilizado para o treinamento da rede, ainda há um possível viés temporal dado a variação econômica do meio que está contida a amostragem, podendo haver um pré-filtro dos contratos antes que seja integrado a base de dados e então colhidos para previsão.

O formato do teste escolhido é o de caixa preta sendo aplicado os dados ao sistema para predição, verificando sua saída sendo considerado o teste como aceito em caso de sua acurácia apresentar taxa superior a 80% e sua precisão acima de 80%.

5. Desenvolvimento

Primeiramente foi necessário desenvolver um query em SQL que realizasse a busca dos dados necessários para o projeto onde consistia na união de atributos de duas tabelas de um banco de dados relacional SQL Server, então os dados foram extraídos para uma tabela Excel. Ainda foi necessário unir o atributo de referência de pagamento para os contratos consultados anteriormente que eram originados de um outro arquivo Excel, com a união destes dados a partir de um identificador único; posteriormente foram eliminados todos os atributos que pudessem identificar diretamente um cliente.

Em posse dos dados consolidados o desafio era identificar quais os melhores atributos a serem utilizados no modelo, com isso primeiro verificou-se quais estavam disponível e buscado a correlação entre eles, embora muitos dos coeficientes de correlação obtidos estivessem próximo a zero o que indica independência dos atributos. Inicialmente o cálculo de correlação por meio da função de correlação foi realizado diretamente no Excel e, em um segundo momento, este cálculo foi realizado com o apoio de uma função do *Numpy* para correlação de atributos.

Em um segundo momento foi realizada uma pesquisa de campo que apontou um conjunto de atributos, que manualmente eram observados, e com base na análise e identificação de padrões em métricas e de expertise era definido uma prioridade caso houvesse um potencial de pagamento elevado.

Visto que o objetivo é a classificação de clientes, era necessário saber quais classes deveriam ser utilizadas para classificação. Havia uma que era mais segmentada pelo tipo de pagamento havendo então 4 classes, enquanto o outro formato era de somente 2 classes mostrando sim ou não para a chance de pagamento. Foi selecionado o formato de 2 classes para seguir nos testes a fim de tornar mais simples o treinamento da rede e visualização.

Durante a análise do conjunto de atributos levantados na entrevista foi identificado dados classificatórios sendo estes transformados em números como representação do item sendo criado uma tabela de conversão para futuras verificações.

Observando os dados foi possível identificar a magnitude das informações o que pode prejudicar o treinamento da rede, onde há informações de idade que variam entre 18 e 80 anos e valores de financiamento que variam de 10.000 a 1.000.000, assim foi aplicado o método *Standard Scaler* para reduzir a magnitude dos dados.

Durante o processo de análise foi desenvolvido um script em Python que auxiliou no processo de manipulação dos dados e também para gerar as visualizações que estão apresentadas a seguir.

Na Figura 4 usando um gráfico *scatter plot*, os dados foram filtrados removendo poucos pontos muito distantes (*outliers*) que seriam casos raros na amostra e é possível identificar que há uma separação visível de dois grupos de clientes mas não é clara a divisão entre os que realizaram o pagamento e os que não.

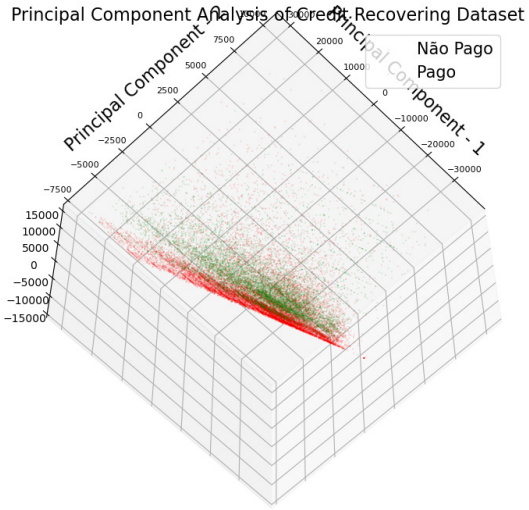

Figura 4. Distribuição de dados sem *Standard Scaler*.

Na Figura 5 os dados foram mais filtrados para mostrar que além da separação vista anterior há uma leve separação vertical de clientes. Ainda assim um grupo menos denso que mostram uma maior frequência de clientes pagantes, mesmo não dividido claramente, e no outro grupo alguns poucos clientes pagantes.

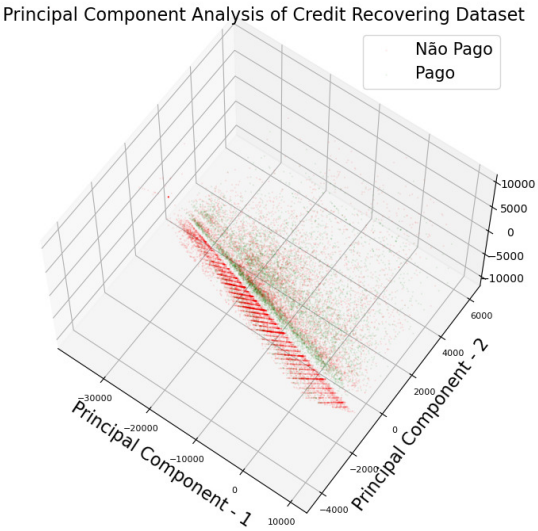

Figura 5. Aplicado um filtro de dados em relação ao gráfico anterior.

Na Figura 6 os dados não foram filtrados para mostrar que, do ângulo da imagem,

um aglomerado de clientes pagantes sobrepõem os não pagantes possuindo uma determinada característica ainda desconhecida.

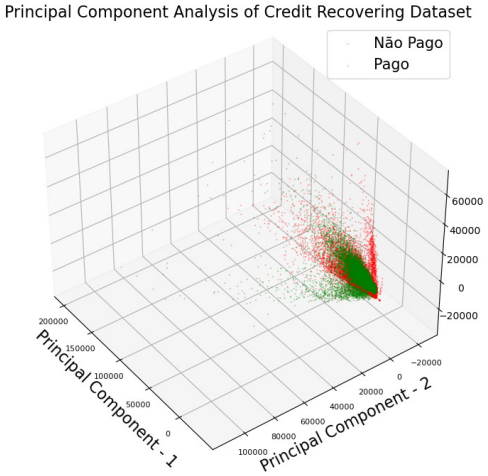

Figura 6. Dados sem filtro e com alteração de angulo.

Finalmente, na Figura 7, os dados não filtrados foram padronizados com a biblioteca *Standard Scaler* para mostrar que alguns grupos possuem uma frequência maior de pagantes enquanto outros não, não apresentando uma clara separação.

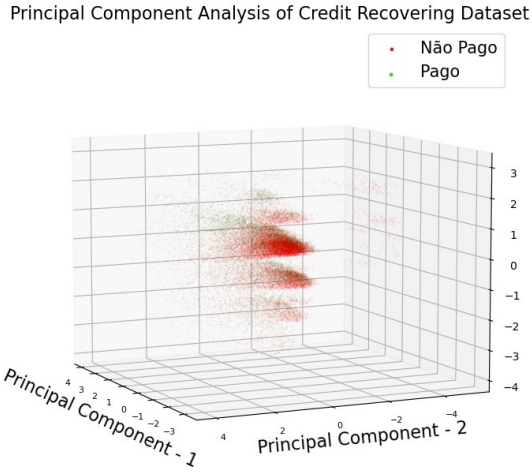

Figura 7. Dados escalonados com *Standard Scaler* apresentando outro angulo.

Para entender a correlação dentre os atributos selecionados para o teste é apresentado um mapa de calor na Figura 8 onde é possível identificar a correlação entre os atributos utilizados. Para isso é considerado uma correlação direta quando se aproximam de 1 ou inversa quando se aproximam de -1; desta forma, quanto mais próximo de 0, indica que menor a correlação entre eles então uma independência entre os atributos.

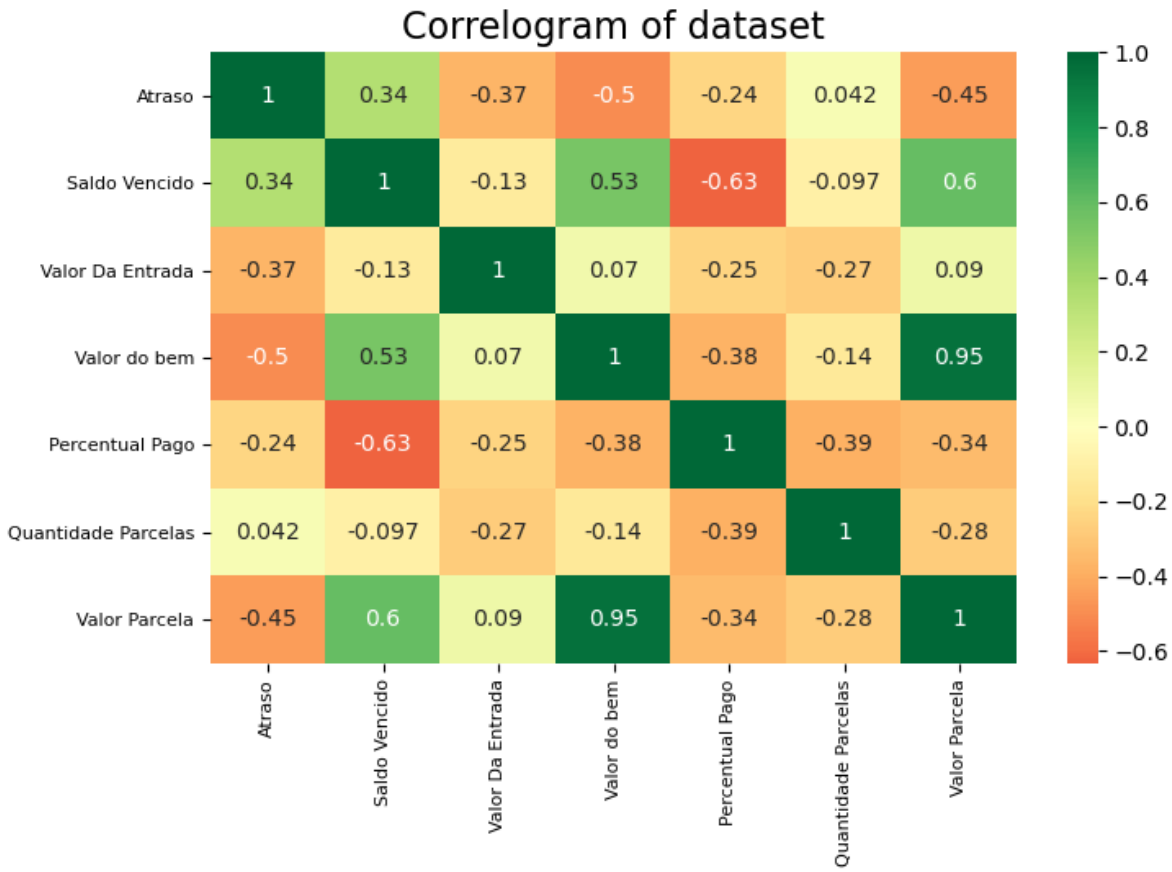

Figura 8. Correlação dos dados usados no treinamento e teste.

Uma informação que pode ser obtida simplesmente pela observação do mapa de calor acima é a correlação entre o saldo vencido e o percentual pago, é apresentado como -0,63, mostrando que, não exclusivamente mas, quanto maior o percentual pago do contrato menor o saldo vencido. Outro exemplo é a ligação entre o valor da entrada com o atraso que também se mostra negativa (-0,37) apresentando que quando maior o valor de entrada menor o atraso, embora seja um valor baixo indica a ocorrência desta condição.

Após a análise da distribuição dos dados ficou claro que não seria uma tarefa fácil realizar o treinamento desta rede neural, decidiu-se então aplicar um arquitetura descrita a seguir.

Para aplicação na MLP foi testado diversos modelos que variavam a quantidade de camadas e de neurônios por camada a fim de buscar o modelo que melhor se adequava aos dados sem causar um *overfitting*. As variações foram de 1 a 4 camadas ocultas contendo de 1 a 15 neurônios cada camada, o que gerou 50.625 combinações, embora um alto número de combinações, o objetivo era em uma análise exploratória em uma grande abrangência e identificar quão diferente eram os resultados com as variações da arquitetura aplicada.

Foram utilizados os volumes de 500, 1.000, 2.500, 3.000, 4.000, 8.000, 10.000, 15.000, 20.000 e 28.000 contratos, respectivamente, durante a primeira bateria de treino e teste nas diferentes arquiteturas da MLP e os resultados obtidos foram variados embora

um menor volume de camadas tenha apresentado melhor desempenho.

Aplicando as combinações de arquitetura aos 10 volumes de *dataset*, o total de execuções de treinamento e validação foram 506.250, cujos resultados foram analisados por meio de filtros, auxílio necessário devido a quantidade de resultados.

Durante o período de treinamento foi identificado um acréscimo gradual no tempo de cada configuração de treino, em alguns momentos para o *dataset* de 20.000 linhas (contratos) cada treinamento apresentava de 2 a 6 minutos e ainda assim era atingido o número máximo de iterações antes de convergir (o algoritmo foi configurado para executar 1.000 iterações).

Ainda não próximo do esperado realizou-se uma segunda bateria de testes com as combinações de neurônios por camada, camadas e alpha que obtiveram melhores resultados em comparação com a bateria anterior, sendo então o alpha em 0,001, um máximo de iterações em 10.000 e 10 combinações de neurônios e camadas, descritas a seguir: (1, 7, 5), (1, 11, 4), (2), (6), (7), (2, 2, 4), (2, 5, 2), (2, 7, 8), (2, 8, 7), (2, 11, 6). Ressaltando que cada conjunto dentro de parênteses representa uma combinação executada, sendo os números representando a quantidade de neurônios e as vírgulas separando as camadas, dado como exemplo esta combinação "(1, 7, 5)"onde temos 3 camadas.

Entretanto antes da segunda bateria foi realizado uma nova análise para remoção de *outliers* vistos anteriormente, uma dispersão mais clara pode ser observado nos gráficos do tipo *boxplot* (Figura 9) com os atributos de valor de parcela e dias em atraso. Onde dias em atraso se mostravam negativos significa que um contrato do cliente estava pago, embora ele houvesse outro contrato em atraso; nestes casos os valores foram convertidos para zero e utilizado um máximo de dias em atraso de até 800 dias. Para as parcelas foram utilizados valores inferiores a R$ 5.000,00. Com a aplicação deste filtro, foram desconsideradas aproximadamente 3,4% das linhas do *dataset*.

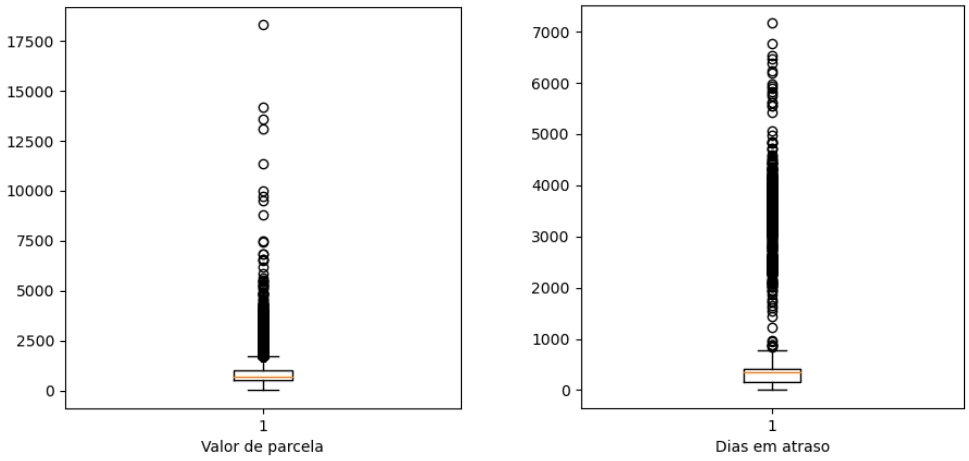

Figura 9. Gráfico boxplot dos atributos de parcela e atraso.

Durante a última bateria de execução foram coletados também o número de iterações realizadas pelo algoritmo até convergir ao resultado, apresentados na Figura 10.

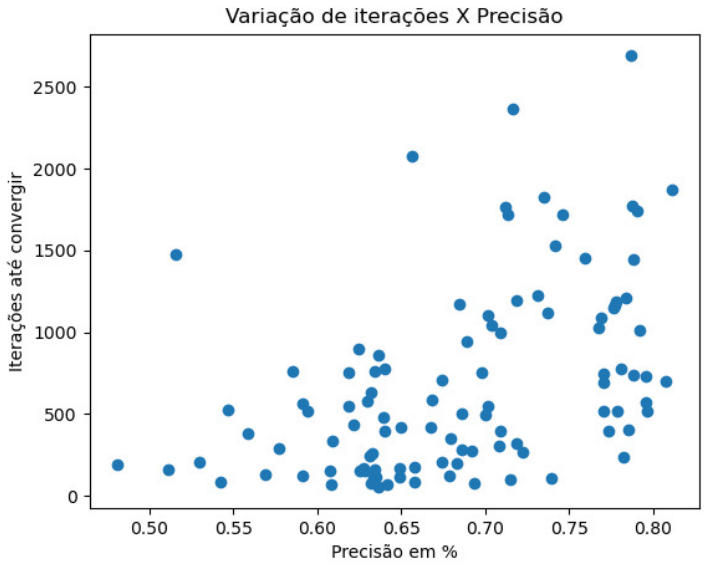

Figura 10. Volume de iterações por execução até convergir.

O modelo escolhido ao final possuía 1 camada oculta de neurônios contendo 7 neurônios sendo o volume de dados em 8.000 contratos com a proporção de 80,0% para treinamento e 20,0% para validação. Nessa configuração, observou-se os resultados de acurácia com 85,11%, 79,57% de precisão, 81,42% de *recall* na validação e a matriz confusão exposta na Figura 11.

		Predição		
		Não Propensão de Pagamento	Propensão de Pagamento	Total corretos
Atual	Não Propensão de Pagamento	870	126	996
	Propensão de Pagamento	112	491	603
	Total da predição	982	617	

Figura 11. Matriz Confusão de resultados.

Com a matriz confusão é possível identificarmos que houveram 1.361 contratos preditos corretamente (870 mais 491) e 238 predições incorretas (112 mais 126). Chega-se a esses números por meio da interpretação da matriz confusão sendo o elemento da primeira linha com a primeira coluna o "Verdadeiro Negativo"(VN) indicando classificação correta para um dado negativo; o elemento da primeira linha com a segunda coluna o "Falso Positivo"(FP) indicando classificação incorreta como positivo; na segunda linha com a primeira coluna, encontra-se o elemento "Falso Negativo"(FN) indicando classificação incorreta como negativo; e na segunda linha com a segunda coluna, o elemento "Verdadeiro Positivo"(VP) indicando classificação correta para um dado positivo. Na Figura 12 é possível visualizar o conceito na própria matriz.

		Predição	
		Negativo	Positivo
Atual	Negativo	VERDADEIRO	FALSO
	Positivo	FALSO	VERDADEIRO

Figura 12. Conceito de matriz confusão.

6. Conclusão

Então, com o objetivo de realizar uma classificação de cliente de forma automatizada utilizando o conceito de aprendizado de máquinas a fim de reduzir o processo manual e entregar com maior rapidez uma diretriz de quais clientes podem ser focados inicialmente. Foi possível concluir que o maior trabalho é durante o pré-processamento, nos passos de colheita, limpeza e junção de dados, escolha correta dos atributos a serem utilizados para enfim chegar ao treinamento das redes neurais.

Também foi visto que a rede possui capacidade para atingir os níveis de aceite contudo com mais atributos acrescidos ao modelo atual e ou substituição de atributos no modelo. Assim apresentando que o dados possui um padrão porém ainda não bem definido no modelo ou algoritmo utilizado. Os resultados obtidos tiveram aproximadamente 85,1% dos contratos classificados corretamente e aproximadamente 79,6% de precisão dos valores treinados.

Uma das dificuldades encontradas neste tipo de dados é que estas são sensíveis ao tempo, onde em um determinado cenário um cliente pode ter uma maior probabilidade de pagamento que outro, entretanto variáveis econômicas, políticas, naturais podem afetar diretamente na atitude do cliente que caso venha novamente se tornar inadimplente neste momento ele pode não ter a mesma condição de pagamento que a anterior. Ainda de acordo com as variações dos últimos anos no cenário político e econômico é possível dizer que um conjunto de dados pode se deteriorar em apenas 3 meses, ainda mais se considerarmos o impacto sócio-econômico gerado por um fenômeno natural recentemente presenciado e não esperado pela pandemia do Covid-19.

Dentre as principais disciplinas que foram importantes para este projeto estão as de Lógica de Programação, Programação Orientada a Objetos e Estrutura de Dados (que auxiliaram no desenvolvimento e estruturação do script), Introdução a Inteligência Artificial (que apoiou fortemente devido ser o método empregado), também as disciplinas de Projeto de Sistemas I e Projeto de Sistemas II (que auxiliaram no refinamento de escrita técnica e documentação), entre outras.

Para trabalhos futuros recomenda-se a utilização desta ideia aplicando um refinamento por *feature engineer*, novos algoritmos, comparação de cenários sócio-econômicos em diferentes tempos e ainda novos modelos e também aplicação de probabilidade de cada contrato para cada classe classificada e também a utilização de *deep learning* que se baseia em um maior volume de camadas gerando diversas interpretações para retro alimentação podendo identificar padrões sutis que podem trazer ótimos resultados.

Referências

Alcoforado, C. F., Alcoforado, L. F., Dutt-Ross, S., and dos Santos Simão, A. (2019). Identificando fatores que influenciam no endividamento do cadete da aeronáutica. *Revista do Seminário Internacional de Estatística com R*, 4(1):15.

Baesens, B., Setiono, R., Mues, C., and Vanthienen, J. (2003). Using neural network rule extraction and decision tables for credit-risk evaluation. *Management science*, 49(3):312–329.

Brems, M. (2017). A one-stop shop for principal component analysis. https://towardsdatascience.com/a-one-stop-shop-for-principal-component-analysis-5582fb7e0a9c. [Online; acesso em 30-Agosto-2018].

Cerbasi, G. (2018). Compras por impulso: Como reverter esse problema? `https://www.gustavocerbasi.com.br/blog/compras-por-impulso/`. [Online; acesso em 20-Novembro-2020].

Claudino, L. P., Nunes, M. B., Oliveira, A. R., and Campos, O. V. (2009). Educação financeira e endividamento: um estudo de caso com servidores de uma instituição pública. In *Anais do Congresso Brasileiro de Custos-ABC*.

Confederação Nacional do Comércio de Bens, S. e. T. C. (2020). Endividamento das famílias alcança novo recorde, e inadimplência acelera em junho. `http://www.cnc.org.br/sites/default/files/2020-06/An%C3%A1lise%20Peic%20-%20junho%20de%202020.pdf`. [PDF Online; acesso em 20-Novembro-2020].

Costa, J. A. F. et al. (1999). *Classificação automática e análise de dados por redes neurais auto-organizáveis*. PhD thesis, Universidade Estadual de Campinas.

Domingos, P. (2017). *O algoritmo mestre: como a busca pelo algoritmo de machine learning definitivo recriará nosso mundo*. Novatec Editora.

Jolliffe, I. (2002). *Principal Component Analysis*. Springer-Verlag New York.

Kovács, Z. L. (2002). *Redes neurais artificiais*. Editora Livraria da Fisica.

McKinney, W. (2012). *Python for data analysis: Data wrangling with Pandas, NumPy, and IPython*. "O'Reilly Media, Inc.".

Neto, A. A. A. and Carmona, C. U. D. M. (2002). Modelagem do risco de crédito: um estudo do segmento de pessoas físicas em um banco de varejo. *Revista Eletrônica de Administração*, 10(4).

Pacelli, V. and Azzollini, M. (2011). An artificial neural network approach for credit risk management. *Journal of Intelligent Learning Systems and Applications*, 3(02):103.

Serasa, E. (2018a). Conheça as 7 principais causas de inadimplência no brasil hoje. [Online; acesso em 30-Agosto-2018].

Serasa, E. (2018b). Inadimplência do consumidor afeta 61,5 milhões no país, revela serasa.

Stephanie (2013). Correlation coefficient. [Online; Acesso em 18-Março-2019].

Stephanie (2019). Crescente inadimplencia. [Online; Acesso em 09-Outubro-2019].

Tiryakia, G., Gavazzab, I., Andradec, C., and Mota, A. (2017). Ciclos crédito, inadimplência e as flutuações econômicas no brasil. *Journal of Contemporary Economics*, 21(01):1–33.

Tosi, S. (2009). *Matplotlib for Python developers*. Packt Publishing Ltd.

Waskom, M., Botvinnik, O., Hobson, P., Warmenhoven, J., Cole, J. B., Halchenko, Y., Vanderplas, J., Hoyer, S., Villalba, S., Quintero, E., et al. (2015). Seaborn: V0. 6.0 (june 2015).

Documento Digitalizado Público

Anexo I - artigo TCC - aluno Guilherme Bordotti de Carvalho - HT1621017

Assunto: Anexo I - artigo TCC - aluno Guilherme Bordotti de Carvalho - HT1621017
Assinado por: Andre Constantino
Tipo do Documento: Outro
Situação: Finalizado
Nível de Acesso: Público
Tipo do Conferência: Documento Original

Documento assinado eletronicamente por:

- **Andre Constantino da Silva**, **PROFESSOR ENS BASICO TECN TECNOLOGICO**, em 15/01/2021 18:24:40.

Este documento foi armazenado no SUAP em 15/01/2021. Para comprovar sua integridade, faça a leitura do QRCode ao lado ou acesse https://suap.ifsp.edu.br/verificar-documento-externo/ e forneça os dados abaixo:

Código Verificador: 591694
Código de Autenticação: 5893711360