ANÁLISE DE INADIMPLÊNCIA CLIENTES USANDO INTELIGÊNCIA DE NEGÓCIO E ANÁLISE DE DADOS

Richard Carlos de Oliveira[1],
Fabio Goldner[2],
Bruno Bastos Stoll[2]

[1] Discente do curso de Engenharia de Computação do Centro Universitário Multivix
[2] Mestres, Docentes do Centro Universitário Multivix Vitória

RESUMO
Dentro das instituições financeiras é um processo comum a avaliação de perfil de clientes afim de ceder crédito. No entanto, existem muitas dificuldades na avaliação desses perfis, pois a análise por vezes é manual e demorada. Existem avanços nas avaliações automatizadas utilizando algoritmos de aprendizado de máquina. O presente artigo compara os resultados estatísticos de modelos preditivos desenvolvidos, com o objetivo de verificar sua performance e aderência utilizando uma base de dados pública e melhorar o desempenho da usabilidade do tratamento dos dados referentes a inadimplência de clientes. Para tanto, foi desenvolvido as etapas de análise de dados e inteligência de negócio, sendo que a apresentação de resultados foi inferida perfis de clientes afim de verificar a sua inadimplência ou não ao modelo.
PALAVRAS-CHAVE
Análise de dados; Análise estatística; Aprendizado de Máquina; Classificação; Inteligência de Negócio; Previsão.

ABSTRACT
Within financial institutions, evaluating customer profiles for the purpose of granting credit is a common process. However, there are many challenges in assessing these profiles, as the analysis is often manual and time-consuming. There have been advances in automated evaluations using machine learning algorithms. This article compares the statistical results of developed predictive models with the aim of verifying their performance and adherence using a public dataset, as well as improving the usability of data processing related to customer delinquency. To this end, data analysis and business intelligence steps were developed, and the results presentation inferred customer profiles to determine whether they were likely to default or not according to the model.
KEYWORDS
Data analysis; Statistical analysis; Machine learning; Classification; Business intelligence; Forecasting.

INTRODUÇÃO

Desde a idade antiga as instituições financeiras eram vistas com um olhar muito formalizado e burocrático, onde só os ricos podiam obter crédito e contas. No império romano essas instituições realizavam a troca de moedas, com o passar dos anos o mercado foi mudando e as funções também, com isso durante o século XV, criou-se o primeiro banco moderno o famoso *Banco di San Giorgio* em Gênova. Mas o banco como conhecemos, aquele que tem a função de troca papel-moeda só veio a aparecer em 1710 segundo (JEHNIFFER, 2021).

Com o passar dos anos essas instituições foram moldando a economia global, em 1929 tivemos a primeira grande crise mundial, onde várias instituições bancárias,

foram a falência, como principal causa, tivemos a especulação financeira e a superprodução. Após um tempo a economia conseguiu superar essa crise e em meados dos anos 1983 tivemos um segundo grande marco, a primeira utilização de serviços eletrônicos, o mercado avançou bastante com essas tecnologias, no entanto em 2008 tivemos outra grande crise que ficou conhecida com a bolha imobiliária, fez com que grandes instituições financeiras novamente fossem a falência, e o mercado teve que se renovar outra vez (GAZETA DO POVO, 2018).

Com isso as empresas do mercado de crédito vêm passado por uma reformulação nessa nova era de ouro dos mercados abertos, deixando de desenvolver agências físicas e passando para o mundo digital, onde o contato pode se dar através de uma simples mensagem de texto ou um simples acesso no sitio da empresa, desenvolvendo uma desformalização do mercado. Dessa maneira, as empresas descobriram um novo nicho econômico, que não era atendido anteriormente, pessoas que sequer tinha a possibilidade de solicitar um crédito, agora podem. Então visando sempre um aumento nos lucros de forma abrupta.

As empresas solicitam os dados de futuros clientes e fazem pesquisas em birôs de créditos[1], afim de extrair o histórico de compras daquele cliente, e partir daí liberar ou não crédito para a pessoa. No entanto, na maioria das vezes as análises são desenvolvidas de forma manual em uma planilha. Dessa forma, o presente trabalho tem por objetivo avançar nos estudos de técnicas automatizadas para classificação de clientes inadimplentes, utilizando algoritmo *Random Forest* em bases de dados desbalanceadas e como método de pré-processamento, foi utilizado o algoritmo de geração de dados sintético *SMOTE*.

Segundo Sharda, Delen e Turban, (2019) o *Business Intelligence* (BI) nasceu com a finalidade de apoiar as tomadas de decisões das empresas, munidos a conceitos computacionais como: Inteligência Artificial (IA), Análise de Dados, Computação em Nuvem e Aprendizado de Máquina se desenvolvem de forma orgânica munidos de técnicas matemáticas avançadas, na finalidade de automatizar essa análise de crédito. Com isso o grande volume de dados que as instituições financeiras têm, estão se tornando informação útil.

> As instituições financeiras têm se tornado repositórios fantásticos de informação. A quantidade de dados gerados pela interação de clientes em seus canais digitais aumenta exponencialmente em volume e em

[1] https://www.migalhas.com.br/depeso/344945/o-que-sao-biros-de-credito

> complexidade e, extrapola a fronteira de serviços financeiros. (GIOVANOLLI, 2017, p. 01).

1. REFERENCIAL TEÓRICO

Esta seção apresenta os principais conceitos relacionados a preparação dos dados, aprendizagem de máquina, modelagem preditiva e avaliação e exibição do modelo.

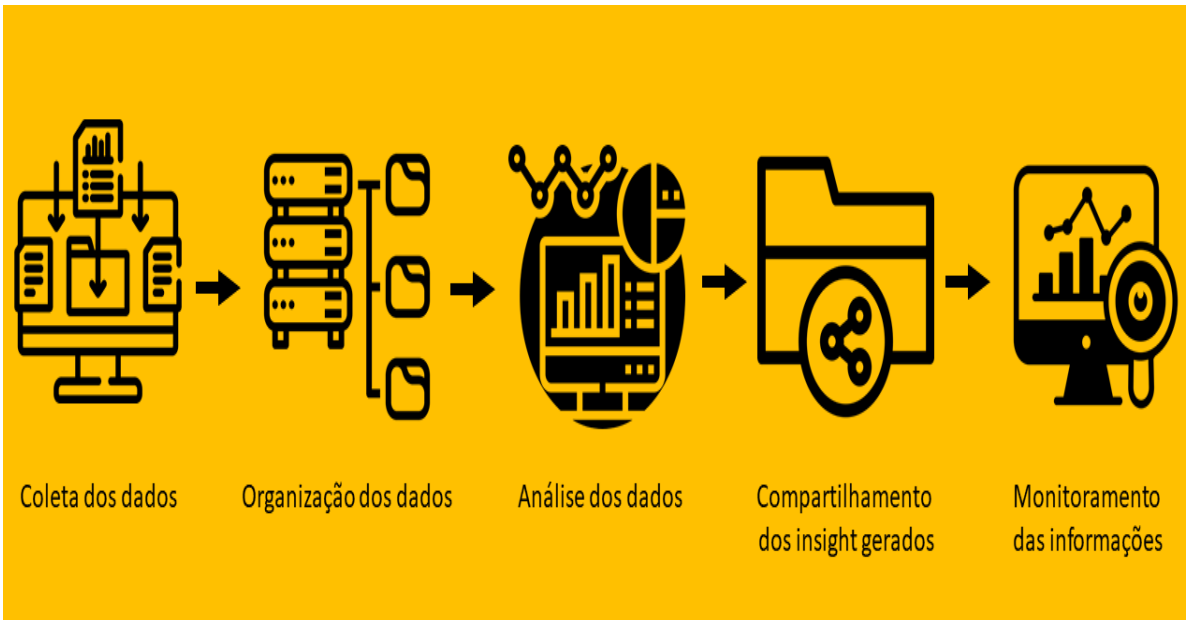
1.1 Business Intelligence

Esse termo pode ser visto como um processo agregado a tecnologia afim de analisar dados e visualizar as informações relevantes ao negócio, ajudando na tomada de decisão e mantendo o negócio sempre bem informado. Há inúmeras bibliografias com definições bem detalhadas sobre o tema, uma delas diz que:

> O conceito de *Business Intelligence* com o entendimento de que é Inteligência de Negócios ou Inteligência Empresarial compõe-se de um conjunto de metodologias de gestão implementadas através de ferramentas de software, cuja função é proporcionar ganhos nos processos decisórios gerenciais e da alta administração nas organizações, baseada na capacidade analítica das ferramentas que integram em um só lugar todas as informações necessárias ao processo decisório (ANGELONI; REIS, 2006, p. 03).

O *Business Intelligence* ou Inteligência de Negócio, pode ser visto como cinco processos separados conforme a imagem a seguir:

Figura 1 - Etapas da análise usando o *Business Intelligence*



Fonte: Produzido pelo autor

1.2 Mineração de dados

A mineração de dados ou *data mining* consiste em um processo de análise crítica exploratória de um volume alto de dados. Nessa análise há uma busca por padrões consistentes ou de relacionamento sistemático entre as variáveis do processo. Gerando assim um padrão detectado de novos subconjuntos de dados (CETAX, 2020).

Faz parte dessa análise exploratória tecnologias de banco de dados estatística e inteligência artificial, segundo (AMO, 2003) a tarefa de mineração de dados é o que desejamos buscar nos dados, ou seja, que tipo de categorias de padrões ou de regularidades podemos retirar. Já as técnicas de mineração são métodos que nos garantem em como descobrir os padrões que nos interessam. Já segundo (ELMASRI e NAVATHE, 2002) as técnicas utilizadas devem ter os seguintes propósitos: previsão, identificação, classificação ou a otimização dos recursos.

1.3 Inteligência artificial

O termo está vinculado a dois campos de estudos um deles tem o propósito de desenvolver e empregar máquinas, afim de realizar as atividades que humanos desenvolvem de maneira automática, é obter uma melhor performance unido a uma eficiência. Um outro lado desse campo e o Aprendizado de Máquina, do inglês *Machine Learning (ML)*, que vem tornando empresas mais analíticas nesse campo da pesquisa. Trazendo à tona uma busca de padrões que pode ser aplicado em diversas tecnologias.

Segundo Barr e Feigenbaum (1981) a IA nasceu como uma vertente da ciência da computação em meados de 1956, com a finalidade de desenvolver sistemas inteligentes de computadores capazes de associar a inteligência humana. Segundo o pai da inteligência artificial e da ciência da computação Alan Turing ele diz que: *“Nós só podemos ver um pouco do futuro, mas o suficiente para perceber que há muito a fazer”.* Com os passar dos anos muitas tecnologias foram desenvolvidas através deste campo de pesquisa, algoritmos foram criados e recriados e guerras foram vencidas, sempre com um olhar para o futuro e vendo que tem muito a se desenvolver ainda.

Dentre as tecnologias que utilizamos hoje algumas delas nos ajudam a agendar um compromisso, ligar para alguém, colocar uma música para tocar, etc. como a Siri, Alexa e Assistente Google. Essas são tecnologias com fundamentos de IA, ou seja, tentam imitar o ser humano. Atrás de todo aquele algoritmo temos base matemática

fundamentada em estatística e álgebra linear, que nos ajudam a estudar pontos em grafos, utilizando matrizes e vetores como fundamento teórico básico e através disso criar padrões que possam ser analisados lexicamente através da linguística computacional podendo utilizar uma modelagem computacional e ontológica para percepção de palavras.

Uma outra parte da desse campo de pesquisa, está mais ligado a mecânica e eletrônica, pois tentam criar agentes fiscos que sejam capazes de inferir no mundo real de maneira autônoma, ou seja, os robôs parecidos com o que podemos ver no filme “Eu, Robô” de 2004 com Will Smith, onde foi possível criar tal máquina.

1.4 Aprendizado de máquina

O Aprendizado de Máquina, do inglês *Machine Learning*, é considerado uma vertente da inteligência artificial que tem a finalidade de criar algoritmos ou utilizar técnicas capazes de criar modelos preditivos, utilizando a estatística pura e aplicada para relatar experiências no mundo real. O processo inicia-se com a observação dos dados, com o principal objetivo de buscar padrões e tomar as melhores decisões baseado em uma lógica, sem inferência de humanos no processo de aprendizado. O aprendizado se dá através de algoritmos pré ou pós programados, eles ditam as regras e o passo a passo para solucionar o problema que foi colocado (SHARDA, DELEN, TURBAN, 2019).

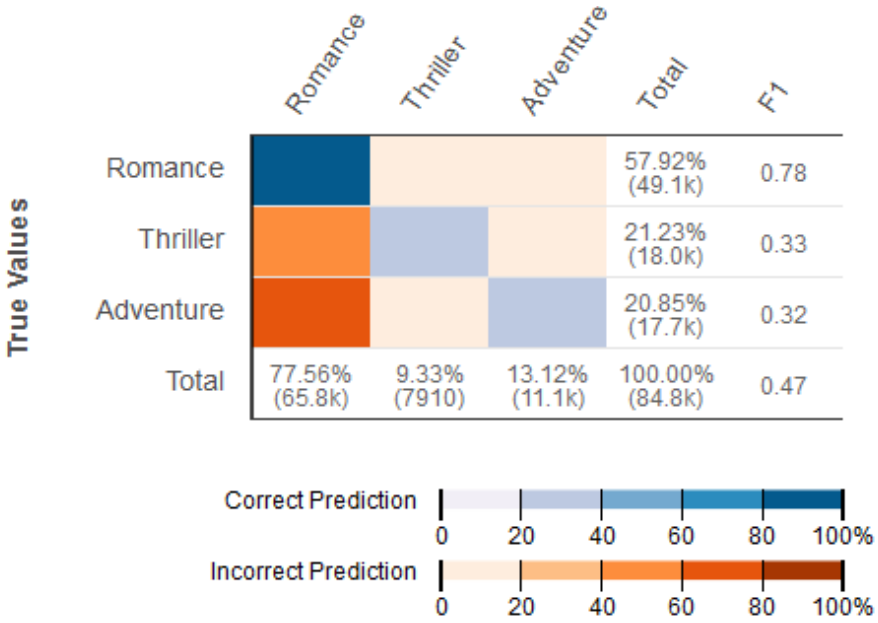
1.4.1 Algoritmos classificadores

Tais algoritmos são utilizados no aprendizado supervisionado, que tem a finalidade de prever uma classe associada dada um conjunto de dados de entrada. Esses algoritmos funcionam com o envio de pré-dados (conjunto de dados setado ou conhecido) para o seu treinamento, pois o algoritmo foi desenvolvido de forma genérica, podendo ser aplicadas em diferentes contextos e bases de dados. Munido de tais informações pode-se inferir que o conjunto terá somente duas ou várias classes, que são conhecidas como classificação binária ou multi-classe respectivamente (SILVA, 2020).

A classificação binária é usada para prever categorias de uma instância ao qual o dado pertence. Reparando que a entrada desse tipo é um conjunto de exemplos rotulados, onde cada rótulo contém um inteiro conhecido como binário (0 ou 1). Já o multi-classe é quando a categoria dos problemas do tipo de classificação é inferida

mais de duas classes. Um bom exemplo para analisar tais situações é quando montamos uma matriz de confusão, onde no exemplo abaixo você não precisa optar por um limite de pontuação para as previsões, tais classes são os próprios rótulos, conforme exemplificado na Figura 2.

Figura 2 - Matriz de confusão para classificadores multi-classes



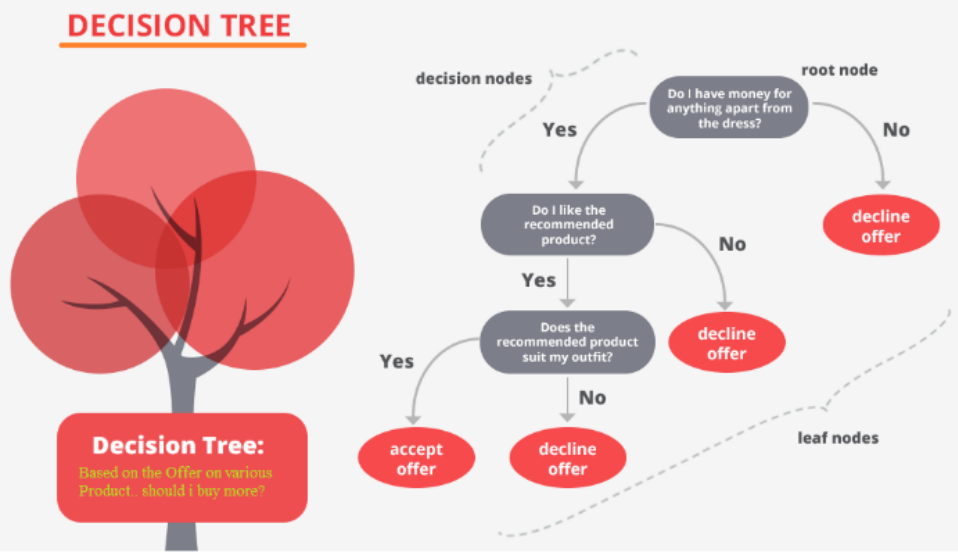
Fonte: (Amazon Web Services, Inc. and/or its., 2021)

1.4.2 Árvore de decisão

Segundo Sacramento (2021) esse termo / conceito é utilizados em vários campos de pesquisa, desde a Pesquisa Operacional que é voltada para a programação de processos até mesmo no aprendizado de máquina, onde é utilizado junto com técnicas de classificação ou até mesmo de regressão afim de chegar em um resultado estatisticamente melhor, analisando probabilisticamente nó à nó da árvore de decisão, qual tem o melhor custo, recurso ou possibilidades de eventos futuros.

Em suma uma árvore de decisão é como o próprio nome já diz uma árvore que pode conter vários ramos ou nós conhecidos como *decision nodes* que se ligam hierarquicamente. O nó principal é conhecido como nó raiz ou *root node* e os resultados são representados pelas folhas ou *leaf nodes.* Tendo como processo de criação do algoritmo de árvore de decisão a indução de dados para criar uma estrutura com várias ramificações partindo dos dados de entrada, pode se estimar os resultados mais prováveis, conforme apresentada na Figura 3:

Figura 3 - Funcionamento da Árvore de Decisão



Fonte: (SACRAMENTO, 2021)

Dentre os algoritmos existente o de árvore de decisão é o mais utilizado, pois é o que se assemelha a um fluxograma, onde pode-se coloca o processo em tal árvore e abrir ele em nós.

1.4.3 Random Forest

Tal algoritmo entra no rol de algoritmos de classificação onde foi desenvolvido com algumas melhorias feitas com base no algoritmo Árvores de Decisão, ele utiliza um método de reamostragem ou *bootstrap* associado a agregação.

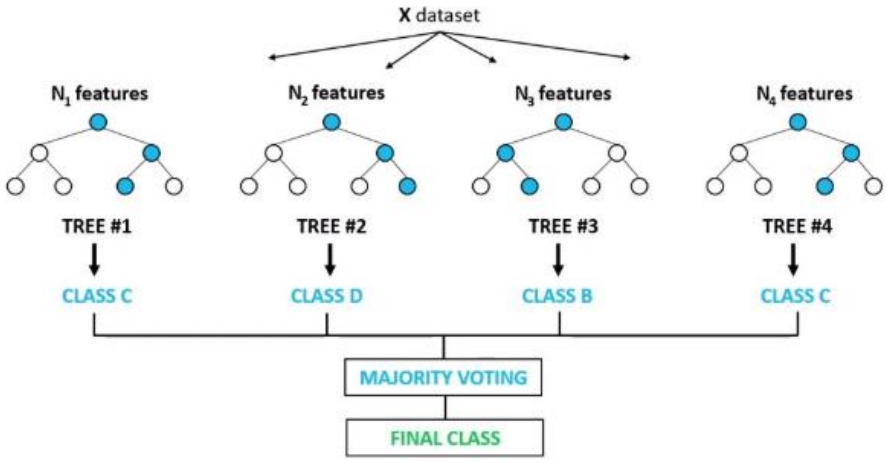
> O método de *bootstrap* se baseia em uma amostra aleatória $y = (y1, y2, \dots, yn)$ cujos valores são realizações de variáveis aleatórias independentes e identicamente distribuídas $Y1, \dots, Yn$, cada uma possuindo função de densidade de probabilidade e função de distribuição denotadas por f e F, respectivamente. A amostra é usada para realizar inferências sobre alguma característica da população, genericamente denotada por θ, através de uma estatística T, cujo valor na amostra é t. (CRIBARI 2010, p. 33).

Essa consociação entre a reamostragem e agregação produz *bootstrap aggregation*, procedimento generalizado utilizado para retratar a variância dos algoritmos.

A primeira versão do algoritmo foi criada por Tin Kam Ho em 95, que utilizou o princípio do subespaço aleatório, onde ele criava uma discriminação estocástica para classificação proposta. Com o passar dos anos várias incrementações foram sendo feitas no algoritmo (OLIVEIRA, 2017).

Com ele e possível criar modelo de predição usando ML para apresentar resultados bem próximos do real. Pois com as várias árvores que ele cria a probabilidade de um ponto de convergência e muito concreta (ANDRADE, 2020), conforme apresentado na Figura 4.

Figura 4 - Random Forest



Fonte: (ANDRADE, 2020)

1.5 Métricas de avaliação

Dentre as inúmeras métricas de avaliação a capacidade de predizer um modelo, via análise estatística e até hoje a mais usual. Diversas ferramentas podem mensurar o poder de predição de um modelo e qual a precisão para a inserção de novos dados.

Tabela 1 - Faixas confiabilidade da acurácia

% de acurácia	Confiabilidade	Explicação	Risco para organização
0% até 30%	Baixa	Há pouca certeza de que os resultados encontrados podem gerar valor para a análise.	Elevado
30% até 80%	Média	Mesmo os níveis de sendo médios os valores ainda podem gerar bons *insight* para a organização, dependendo da sua aplicação.	Moderado
80% até 100%	Alta	No geral os valores demonstram validação para o negócio, podendo ser considerados como prováveis.	Menos sujeitos

Fonte: Produzido pelo autor

Segundo Fávero (2017), uma forma de demonstrar os resultados é através da acurácia que pode variar de 0 a 100%, no geral, quanto mais próximo dos 100% maior é a proximidade do resultado encontrado com o valor real, ela pode-se dividir em três grupos conforme a Tabela 1. Já quando se trata de modelo de classificação binário dividimos os resultados em quatro subcategorias (Matriz de confusão), como na tabela 2:

Tabela 2 - Siglas das métricas de avaliação

Sigla	Descrição
TP	Resultados positivos previstos corretamente como positivos.
TN	Resultados negativos previstos corretamente como negativos.
FN	Resultados positivos previstos erradamente como negativos.
FP	Resultados negativos previstos erradamente como positivos.

Fonte: Produzido pelo autor

Uma outra forma de avaliação é a utilização do coeficiente *kappa* que mostra a taxa de aceitação relativa é a taxa hipotética de aceitação, esse coeficiente varia conforme a tabela abaixo:

Tabela 3 – Valor do *Kappa*

Valor do coeficiente *KAPPA*	Nível de concordância
< 0	Não existe Concordância
0 – 0,20	Concordância Mínima
0,21 – 0,40	Concordância Razoável
0,41 – 0,60	Concordância Moderada
0,61 – 0,80	Concordância Substancial
0,81 – 1,0	Concordância Perfeita

Fonte: Produzido pelo autor

Outros métodos de avaliação de estatística são: precisão, sensibilidade, especificidade e p-valor, onde a precisão é dentre todas as classificações que foram dadas com positivas na matriz de confusão (valores positivos divididos pela soma do valores positivos mais os falsos positivos), onde o grau de porcentagem varia de 0% a 100% entre elas, esse método está estritamente ligado a acurácia, já a sensibilidade e a porcentagem dos valores positivos da matriz de confusão divididos pela soma dos valores positivos mais os falsos negativos, essa variação também vai de 0% a 100%. No que se trata da especificidade ela também é uma variação que vai de 0% a 100%

sendo que a sua avaliação é dada pelos valores positivos divididos pela soma dos valores negativos mais os falsos positivos do conjunto de dados. O p-valor que também é chamado de significância mostra se um teste foi correto ou foi errado, sendo que p menor que 0,05 seu teste é possível, já se ele for maior não se pode inferir resultados (FAVERO, 2017).

1.6 Linguagem R e bibliotecas (*SMOTE*)

A linguagem R é uma dessas novas linguagem de programação que é de código aberto, muito utilizada por cientistas de dados e estatísticos, ela não é limitada a sessões interativas. Os scripts desenvolvidos nela podem ser empacotados em bibliotecas e distribuídos de forma gratuita.

Essa linguagem é extremante simples e existem vários manuais e fóruns pela internet mostrando novas funções e como puxar esses novos pacotes para os scripts, devido a essa facilidade, a linguagem vem sendo muito utilizada na análise de exploratória de dados. Uma das bibliotecas mais utilizadas para essas analises é a *SMOTE* [2] (*Synthetic Minority Oversampling Technique*) que é uma função usada classificação em dados desbalanceados, ou seja, ela usa uma técnica de sobre amostragem minoritária sintética de dados, com isso é possível inflar o conjunto de dados, aumentando a performance e aderência dos resultados em algoritmos preditivos (ZUMEL; MOUNT, 2020).

1.7 Avaliação de inadimplência

Para Sandroni (1999), a inadimplência é “a falta de cumprimento das cláusulas contratuais em determinado prazo” (p. 293), ou seja, quando não há uma quitação de operação no prazo combinado. Tem-se caracterizado como inadimplente, podendo então gerar uma cobrança dessa dívida por vários meios, sejam eles amigáveis e/ou judiciais.

Deve-se ressaltar que a cobrança será feita de forma justa e condizente com os direitos do consumidor, levando em consideração as determinações do Código de Defesa do Consumidor [3].

[2] https://www.rdocumentation.org/packages/DMwR/versions/0.4.1/topics/SMOTE
[3]https://www2.senado.leg.br/bdsf/bitstream/handle/id/533814/cdc_e_normas_correlatas_2ed.pdf

3. METODOLOGIA E MÉTODO DA PESQUISA

Essa sessão contém os procedimentos e técnicas que foram utilizados para a comparação do modelo preditivo, assim como a base de dados, os softwares e pacotes que foram utilizados para manipulação dos dados e o passo a passo para obtenção dos resultados.

3.1 Definição

Para o desenvolvimento do trabalho foi utilizado as etapas do processo de inteligência de negócio, que consiste em: preparar os dados; selecionar as variáveis; construir o modelo; testar o modelo; comparar os resultados do algoritmo e verificar a contribuição das variáveis para o modelo; retirar do modelo as não contribuem; melhorar o algoritmo; testar outra vez, até que se chegue em um modelo conciso e com uma acurácia alta e *Kappa*; após isso podemos inferir perguntas e definir perfis de clientes.

Durante a primeira parte é detalhado os métodos e técnicas da pesquisa, detalhando os programas usados as características da base que foi coletada, mostrando os procedimentos e funções utilizados para preparação dos dados e seleção das variáveis preditoras.

A seguir é detalhado os testes e simulações realizados para obtenção dos resultados utilizando o algoritmo de *Randon Forest*. Após isso será demonstrado os resultados estatísticos deste algoritmo com base no modelo criado. E, por fim, a inferência de resultados no modelo criado para prever se um cliente será inadimplente ou não com a sua conta.

3.2 Preparação dos dados

A base de dados foi obtida de um conglomerado de dados abertos da *University of California Irvine* – Universidade da Califórnia Irvine (UCI), que fica situada nos Estados Unidos da América do Norte, mais especificamente no estado da Califórnia. Essa Universidade coleta diversos dados e os distribui de forma aberta, sempre seguindo as normas da Lei de dados Abertos Americana.

Os dados retirados do site: https://archive.ics.uci.edu/ml/machine-learning-databases/00350/ são de uma pesquisa objetiva de inadimplência de clientes de Taiwan, os dados dessa pesquisa foram liberados em 26 de janeiro de 2016, é a pesquisa foi realizada no ano antecedente. Os dados estão em um formato de .csv

(*Comma-separated values)*, na descrição dos dados foi indicado que a característica dos conjuntos é multi-variável, contendo valores inteiros e reais, com total de 30.000 instâncias e 24 atributos e sem valores ausentes.

Tabela 4 - Descrição do conjunto de dados

Variável	Descrição
X0	Coluna chave
X1	Valor do crédito concedido (dólar NT)
X2	Gênero (1 = masculino; 2 = feminino)
X3	Educação (1 = pós-graduação; 2 = universidade; 3 = ensino médio; 4 = outros).
X4	Estado civil (1 = casado; 2 = solteiro; 3 = outros).
X5	Idade (ano).
X6	Estado de reembolso em setembro de 2005
X7	Situação de amortização em agosto de 2005
X8	Estado de reembolso em julho de 2005
X9	Situação de amortização em junho de 2005
X10	Estado de reembolso em maio de 2005
X11	Situação de amortização em abril de 2005
X12	Valor da fatura em setembro de 2005
X13	Valor da fatura em agosto de 2005
X14	Valor da fatura em julho de 2005
X15	Valor da fatura em junho de 2005
X16	Valor da fatura em maio de 2005
X17	Valor da fatura em abril de 2005
X18	Valor pago em setembro de 2005
X19	Valor pago em agosto de 2005
X20	Valor pago em julho de 2005
X21	Valor pago em junho de 2005
X22	Valor pago em maio de 2005
X23	Valor pago em abril de 2005

Fonte: (UCI Machine Learning Repository, 2009)

Munidos dos dados foi utilizado a ferramenta *Sublime Text 3* para conferir se os dados que foram baixados, realmente constavam dentro do arquivo, após isso foi utilizado outro software conhecido com *R Studio*[4], versão 1.4.1106, para criação do modelo teste do mesmo. Vale ressaltar que essa ferramenta é um ambiente livre de desenvolvimento em linguagem R, muito utilizada para aprendizado de máquina e predição de medidas estatísticas nos campos de pesquisa econômicas e de engenharia.

Para iniciar a criação do modelo foi aberta a ferramenta e definido a pasta de trabalho, após isso foi instado os pacotes para o projeto e foi carregado os pacotes e feito o inserte da fonte de dados, figura a seguir:

[4] https://www.rstudio.com/

Figura 5 – Bibliotecas e pacotes

```
# Definindo a pasta de trabalho
setwd("C:/Users/rcovv/OneDrive/Documentos/Arq")
getwd()

# Instalando os pacotes para o projeto
install.packages("Amelia")
install.packages("caret")
install.packages("ggplot2")
install.packages("dplyr")
install.packages("reshape")
install.packages("randomForest")
install.packages("e1071")

# Carregando os pacotes
library(Amelia)
library(ggplot2)
library(caret)
library(reshape)
library(randomForest)
library(dplyr)
library(e1071)

# Carregando o dataset
# Fonte: https://archive.ics.uci.edu/ml/datasets/default+of+credit+card+clients
dados_clientes <- read.csv("dados/dataset.csv")
```

Fonte: Produzido pelo autor

Após isso foi utilizado a função *view* para visualizar os dados dentro na ferramenta a função *dim* para mostrar quantidade dimensões dentro do conjunto através de um vetor atômico de dados e a função a *str* para visualizar os dados, conforme apresentada a Figura 6.

A função *summary* para mostrar o resumo das variáveis em forma estatística incluindo desvio padrão, média intervalos e percentis, conforme a figura 7. Dessa forma, é possível notar que os valores das variáveis eram todos inteiros.

Figura 6 – Visualização dos dados

```
> dados_clientes <- read.csv("dados/dataset.csv")
> View(dados_clientes)
> dim(dados_clientes)
[1] 30000    25
> str(dados_clientes)
'data.frame':   30000 obs. of  25 variables:
 $ ID                        : int  1 2 3 4 5 6 7 8 9 10 ...
 $ LIMIT_BAL                 : int  20000 120000 90000 50000 50000 500000 100000 140000 20000
 $ SEX                       : int  2 2 2 2 1 1 1 2 2 1 ...
 $ EDUCATION                 : int  2 2 2 2 2 1 1 2 3 3 ...
 $ MARRIAGE                  : int  1 2 2 1 1 2 2 2 1 2 ...
 $ AGE                       : int  24 26 34 37 57 37 29 23 28 35 ...
 $ PAY_0                     : int  2 -1 0 0 -1 0 0 0 0 -2 ...
 $ PAY_2                     : int  2 2 0 0 0 0 0 -1 0 -2 ...
 $ PAY_3                     : int  -1 0 0 0 -1 0 0 -1 2 -2 ...
 $ PAY_4                     : int  -1 0 0 0 0 0 0 0 0 -2 ...
 $ PAY_5                     : int  -2 0 0 0 0 0 0 0 -1 -1 ...
 $ PAY_6                     : int  -2 2 0 0 0 0 0 -1 0 -1 ...
 $ BILL_AMT1                 : int  3913 2682 29239 46990 8617 64400 367965 11876 11285 0 ...
 $ BILL_AMT2                 : int  3102 1725 14027 48233 5670 57069 412023 380 14096 0 ...
 $ BILL_AMT3                 : int  689 2682 13559 49291 35835 57608 445007 601 12108 0 ...
 $ BILL_AMT4                 : int  0 3272 14331 28314 20940 19394 542653 221 12211 0 ...
 $ BILL_AMT5                 : int  0 3455 14948 28959 19146 19619 483003 -159 11793 13007 ...
 $ BILL_AMT6                 : int  0 3261 15549 29547 19131 20024 473944 567 3719 13912 ...
 $ PAY_AMT1                  : int  0 0 1518 2000 2000 2500 55000 380 3329 0 ...
 $ PAY_AMT2                  : int  689 1000 1500 2019 36681 1815 40000 601 0 0 ...
 $ PAY_AMT3                  : int  0 1000 1000 1200 10000 657 38000 0 432 0 ...
 $ PAY_AMT4                  : int  0 1000 1000 1100 9000 1000 20239 581 1000 13007 ...
 $ PAY_AMT5                  : int  0 0 1000 1069 689 1000 13750 1687 1000 1122 ...
 $ PAY_AMT6                  : int  0 2000 5000 1000 679 800 13770 1542 1000 0 ...
 $ default.payment.next.month: int  1 1 0 0 0 0 0 0 0 0 ...
```

Fonte: Produzido pelo autor

Figura 7 - Sumário dos dados

```
> summary(dados_clientes)
       ID          LIMIT_BAL           SEX          EDUCATION        MARRIAGE          AGE
 Min.   :    1   Min.   :  10000   Min.   :1.000   Min.   :0.000   Min.   :0.000   Min.   :21
 1st Qu.: 7501   1st Qu.:  50000   1st Qu.:1.000   1st Qu.:1.000   1st Qu.:1.000   1st Qu.:28
 Median :15000   Median : 140000   Median :2.000   Median :2.000   Median :2.000   Median :34
 Mean   :15000   Mean   : 167484   Mean   :1.604   Mean   :1.853   Mean   :1.552   Mean   :35
 3rd Qu.:22500   3rd Qu.: 240000   3rd Qu.:2.000   3rd Qu.:2.000   3rd Qu.:2.000   3rd Qu.:41
 Max.   :30000   Max.   :1000000   Max.   :2.000   Max.   :6.000   Max.   :3.000   Max.   :79
     PAY_0              PAY_2              PAY_3              PAY_4              PAY_5
 Min.   :-2.0000   Min.   :-2.0000   Min.   :-2.0000   Min.   :-2.0000   Min.   :-2.0000
 1st Qu.:-1.0000   1st Qu.:-1.0000   1st Qu.:-1.0000   1st Qu.:-1.0000   1st Qu.:-1.0000
 Median : 0.0000   Median : 0.0000   Median : 0.0000   Median : 0.0000   Median : 0.0000
 Mean   :-0.0167   Mean   :-0.1338   Mean   :-0.1662   Mean   :-0.2207   Mean   :-0.2662
 3rd Qu.: 0.0000   3rd Qu.: 0.0000   3rd Qu.: 0.0000   3rd Qu.: 0.0000   3rd Qu.: 0.0000
 Max.   : 8.0000   Max.   : 8.0000   Max.   : 8.0000   Max.   : 8.0000   Max.   : 8.0000
     PAY_6            BILL_AMT1          BILL_AMT2          BILL_AMT3          BILL_AMT4
 Min.   :-2.0000   Min.   :-165580   Min.   :-69777   Min.   :-157264   Min.   :-170000
 1st Qu.:-1.0000   1st Qu.:   3559   1st Qu.:  2985   1st Qu.:   2666   1st Qu.:   2327
 Median : 0.0000   Median :  22382   Median : 21200   Median :  20089   Median :  19052
 Mean   :-0.2911   Mean   :  51223   Mean   : 49179   Mean   :  47013   Mean   :  43263
 3rd Qu.: 0.0000   3rd Qu.:  67091   3rd Qu.: 64006   3rd Qu.:  60165   3rd Qu.:  54506
 Max.   : 8.0000   Max.   : 964511   Max.   :983931   Max.   :1664089   Max.   : 891586
   BILL_AMT5          BILL_AMT6          PAY_AMT1           PAY_AMT2           PAY_AMT3
 Min.   :-81334   Min.   :-339603   Min.   :     0   Min.   :     0   Min.   :     0
 1st Qu.:  1763   1st Qu.:   1256   1st Qu.:  1000   1st Qu.:   833   1st Qu.:   390
 Median : 18105   Median :  17071   Median :  2100   Median :  2009   Median :  1800
 Mean   : 40311   Mean   :  38872   Mean   :  5664   Mean   :  5921   Mean   :  5226
 3rd Qu.: 50191   3rd Qu.:  49198   3rd Qu.:  5006   3rd Qu.:  5000   3rd Qu.:  4505
 Max.   :927171   Max.   : 961664   Max.   :873552   Max.   :1684259   Max.   :896040
    PAY_AMT4           PAY_AMT5           PAY_AMT6       default.payment.next.month
 Min.   :     0   Min.   :     0.0   Min.   :     0.0   Min.   :0.0000
 1st Qu.:   296   1st Qu.:   252.5   1st Qu.:   117.8   1st Qu.:0.0000
 Median :  1500   Median :  1500.0   Median :  1500.0   Median :0.0000
 Mean   :  4826   Mean   :  4799.4   Mean   :  5215.5   Mean   :0.2212
 3rd Qu.:  4013   3rd Qu.:  4031.5   3rd Qu.:  4000.0   3rd Qu.:0.0000
 Max.   :621000   Max.   :426529.0   Max.   :528666.0   Max.   :1.0000
```

Fonte: Produzido pelo autor

3.3 Análise, limpeza e transformação dos dados

Agora que os dados foram setados para dentro do programa e foram conferidos a suas quantidades e perfils, pode-se começar a principal etapa do processo, que é a análise, limpeza e transformação dos dados. Para tanto iniciamos com a remoção da primeira coluna chave pois para a análise essa coluna não traz nenhuma informação útil. Após isso foi renomeado a coluna 24 para inadimplente e verificado se possuía valores ausentes, durante essa observação não foi constatado valores ausentes, conforme o código abaixo:

Figura 8 - Análise dos atributos

```
# Removendo a primeira coluna ID
dados_clientes$ID <- NULL
dim(dados_clientes)
View(dados_clientes)

# Renomeando a coluna de classe
colnames(dados_clientes)
colnames(dados_clientes)[24] <- "inadimplente"
colnames(dados_clientes)
View(dados_clientes)

# Verificando valores ausentes e removendo do dataset
sapply(dados_clientes, function(x) sum(is.na(x)))
?missmap
missmap(dados_clientes, main = "Valores Missing Observados")
dados_clientes <- na.omit(dados_clientes)
```

Fonte: Produzido pelo autor

Após isso foi renomeado e convertido os atributos de gênero, escolaridade, estado civil e idade para fatores categóricos, pois estavam com os atributos inconsistentes com formato de inteiros, no geral foi utilizado as funções *str, summary, cut* e *labels* para fazer essas mudanças, conforme as Figuras 9 e 10 e 11.

Figura 9 - Tratamento dos dados (a)

```
# Renomeando colunas categoricas
colnames(dados_clientes)
colnames(dados_clientes)[2] <- "Genero"
colnames(dados_clientes)[3] <- "Escolaridade"
colnames(dados_clientes)[4] <- "Estado_Civil"
colnames(dados_clientes)[5] <- "Idade"
colnames(dados_clientes)
View(dados_clientes)

# Genero
View(dados_clientes$Genero)
str(dados_clientes$Genero)
summary(dados_clientes$Genero)
?cut
dados_clientes$Genero <- cut(dados_clientes$Genero,
                             c(0,1,2),
                             labels = c("Masculino",
                                        "Feminino"))
View(dados_clientes$Genero)
str(dados_clientes$Genero)
summary(dados_clientes$Genero)
```

Fonte: Produzido pelo autor

Figura 10 - Tratamento de dados (b)

```
# Escolaridade
str(dados_clientes$Escolaridade)
summary(dados_clientes$Escolaridade)
dados_clientes$Escolaridade <- cut(dados_clientes$Escolaridade,
                                   c(0,1,2,3,4),
                                   labels = c("Pos Graduado",
                                              "Graduado",
                                              "Ensino Medio",
                                              "Outros"))
View(dados_clientes$Escolaridade)
str(dados_clientes$Escolaridade)
summary(dados_clientes$Escolaridade)

# Estado Civil
str(dados_clientes$Estado_Civil)
summary(dados_clientes$Estado_Civil)
dados_clientes$Estado_Civil <- cut(dados_clientes$Estado_Civil,
                                   c(-1,0,1,2,3),
                                   labels = c("Desconhecido",
                                              "Casado",
                                              "Solteiro",
                                              "Outro"))
View(dados_clientes$Estado_Civil)
str(dados_clientes$Estado_Civil)
summary(dados_clientes$Estado_Civil)
```

Fonte: Produzido pelo autor

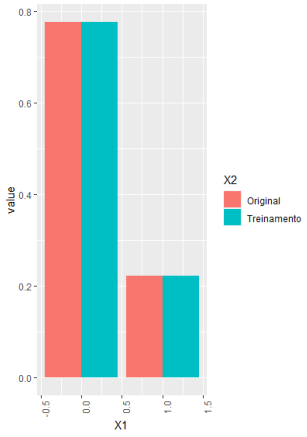
Após essa parte foi convertido as variáveis de pagamento para o tipo fator, as variáveis de endividamento dos meses permaneceram categóricas com os valores reais e a variável de inadimplência continuou como binaria (0 para não inadimplente e 1 para inadimplente). Em seguida foi feito uma amostragem estratificada e definido os dados de treinamento como subconjunto do conjunto de dados original, reparando que a porcentagem de um conjunto era maior que a do outro em 55%. Assim, evitando possíveis problemas, tais entraves expostos através da função *missmap* junto de uma equação para verificar se o novo conjunto tinha valores em branco ou nulos, na finalidade de retirar esses valores. Depois isso foi replicado o conjunto para se ter os dados originais e os dados de treinamento, e subsequente a essa parte foi utilizado a função *melf* para converte colunas em linhas e plotado um gráfico com a distribuição dos dados de treinamento pelos dados originais com as respectivas variáveis de inadimplência (0 e 1), conforme a Figura 12.

Figura 11 - Tratamento de dados (c)

```
# Convertendo a variavel para o tipo fator com faixa etária
str(dados_clientes$Idade)
summary(dados_clientes$Idade)
hist(dados_clientes$Idade)
dados_clientes$Idade <- cut(dados_clientes$Idade,
                            c(0,30,50,100),
                            labels = c("Jovem",
                                       "Adulto",
                                       "Idoso"))
view(dados_clientes$Idade)
str(dados_clientes$Idade)
summary(dados_clientes$Idade)
view(dados_clientes)
```

Fonte: Produzido pelo autor

Figura 12 - Categorização dos dados



Fonte: Produzido pelo autor

3.4 Modelo Aprendizado de Máquina

Foi decidido utilizar o algoritmo de *Random Forest* para a criação do modelo. Como um primeiro modelo foi utilizado os dados desbalanceados, código nas Figuras 13 e 14. Foi feita a matriz de confusão (*confusionMatrix*) e análise estatística dos dados que trouxe a: acurácia, IC, p-valor, taxa sem informação, sensibilidade, prevalência, especificidade, precisão balanceada, *kappa*, valor preditivo negativo e positivo.

Foi desenvolvido um segundo modelo, só que dessa vez com os dados balanceado, para tanto foi utilizando a função *SMOTE* Figura 14 rodado o algoritmo de *Random Forest* e verificando os resultados. Optou-se por fazer um terceiro modelo,

acrescentado a análise de variáveis influenciadoras no processo, onde notou-se que certas variáveis não influenciavam no processo e foi exibido os resultados, salvo o modelo e feito algumas previsões com base no modelo três.

Figura 13 - Código para classificação

```
# Construindo a primeira versão do modelo
?randomForest
View(dados_treino)
modelo_v1 <- randomForest(inadimplente ~ ., data = dados_treino)
modelo_v1

# Avaliando o modelo
plot(modelo_v1)

# Previsão com dados de teste
previsoes_v1 <- predict(modelo_v1, dados_teste)

# Confusion Matrix
?caret::confusionMatrix
cm_v1 <- caret::confusionMatrix(previsoes_v1, dados_teste$inadimplente, positive = "1")
cm_v1

# Calculando Precision, Recall e F1-Score, metricas de avaliação do modelo preditivo
y <- dados_teste$inadimplente
y_pred_v1 <- previsoes_v1

precision <- posPredValue(y_pred_v1, y)
precision

recall <- sensitivity(y_pred_v1, y)
recall

F1 <- (2 * precision * recall) / (precision + recall)
F1
```

Fonte: Produzido pelo autor

Figura 14 - Uso algoritmo SMOTE

```
# Aplicando o SMOTE - SMOTE: Synthetic Minority Over-sampling Technique
# https://arxiv.org/pdf/1106.1813.pdf
table(dados_treino$inadimplente)
prop.table(table(dados_treino$inadimplente))
set.seed(9560)
dados_treino_bal <- SMOTE(inadimplente ~ ., data  = dados_treino)
table(dados_treino_bal$inadimplente)
prop.table(table(dados_treino_bal$inadimplente))
```

Fonte: Produzido pelo autor

4. RESULTADOS E DISCUSSÃO

Os resultados obtidos do processo estão descritos nesta parte, acompanhados de algumas análises. No que concerne a parte de tratamento da base de dados, foi verificado cada tipo de variável envolvida no processo é pareado os seus tipos. Foi trocado de variável booleana para categórico as variáveis de: gênero, escolaridade, estado civil, e idade. Ficando com 11888 registros de gênero masculino e 18112 registros do gênero feminino.

Na parte de escolaridade foi divido em quatro classes (pós-graduados, graduados, ensino médio e outros) com isso a base ficou com 10585, 14030, 4917 e 123 respectivamente. Nessa parte apareceu 345 registros que não pertenciam a nenhuma dessas categorias. Na variável de estado civil também foi divido em quatro (desconhecido, casado, solteiro, outros) ficam com 54, 13659, 15964 e 323 registros respectivamente. A variável de idade, foi escolhido dividir em três classes (jovem de 12-29, adulto de 30-59 e idoso acima de 60) ficam com 11013, 16718, 2269 registros respectivamente. Outras conversões foram feitas como a troca de variáveis para o tipo fator para os meses de pagamento. Foi verificado também a quantidade de registros com clientes inadimplente o que mostrou 23045 (77.71%) não inadimplentes e 6610 (22.28%) inadimplentes.

O primeiro modelo criado demonstra cerca de 0.81 de acurácia sendo que o seu p-valor ficou em 1.01e^-15, ou seja, mesmo com valores em branco e desbalanceados a confiabilidade dos dados era alta. Já para o segundo modelo retirando os valores

em banco apresenta a nova acurácia mostrou um valor de 0.80 e um p-valor de 0.0001 e que mostra um resultado bem melhor que o segundo e analisando o kappa ficou entre regular e boa (0,4 – 0,6) diferente do primeiro modelo que ficou entre ruim e fraca (0,0 – 0,2). Já para o terceiro modelo analisa as variais dependente, onde constata quais variáveis de gênero, estado civil, idade e escolaridade não influencia no modelo. Então retirado essas colunas e rodado o modelo três, que mostrou uma kappa entre fraca e regular (0,3 – 0,4), ou seja, com uma acuraria alta. No entanto, o valor de p-valor aumentou e a sensibilidade ficou próximo de 0,91 diferente de todos os outros modelos que ficaram por volta de 0,7 e 0,8. Assim como a precisão foi de 0,83 para 0,84 no modelo três.

Foi inferido as caraterísticas para o cliente, apresentada no Bloco A -Figura 15. Com isso pode-se notar que o cliente submetido ao modelo preditivo seria inadimplente enquanto os outros 2 não, apresentado no Bloco B – Figura 15.

Figura 15 - Dados de clientes

Bloco A - Código

```
# Dados dos clientes
PAY_0 <- c(0, 0, 0)
PAY_2 <- c(0, 0, 0)
PAY_3 <- c(1, 0, 0)
PAY_AMT1 <- c(1100, 1000, 1200)
PAY_AMT2 <- c(1200, 1300, 1150)
PAY_5 <- c(0, 0, 0)
BILL_AMT1 <- c(350, 420, 280)
```

Bloco B – Resultado

	V1
1	1
2	0
3	0

Fonte: Produzido pelo autor

5. CONSIDERAÇÕES FINAIS

Esta pesquisa estudou sobre o uso de dados na análise preditiva na detecção de clientes inadimplentes. Para isso, foram utilizadas técnicas de mineração de dados, bem como algoritmos classificadores e técnicas de pré-processamento para reamostragem de dados. Cada etapa do método de análise teve a sua relevante importância para a criação do modelo e exibição dos resultados, sendo possível concluir que os objetivos foram atendidos.

O uso advindo das técnicas apresentadas neste trabalho permite que análise preditiva em dados seja melhorada. Bem como a aplicação prática desta pesquisa poderia ser uma interface com o usuário de uma instituição financeira para avaliar o perfil do cliente que está solicitando um crédito e apresentar se ele será inadimplente.

6. REFERÊNCIAS

AFFILIATES., Amazon Web Services, Inc. and/or its. Classificação multiclasse. *In:* AWS. **Amazon Machine Learning Guia do desenvolvedor.** USA, 4 ago. 2016. Disponível em: https://docs.aws.amazon.com/pt_br/machine-learning/latest/dg/multiclass-classification.html. Acesso em: 15 out. 2021.

ANDRADE, António César de. **Como usar algoritmos baseados em árvore para aprendizado de máquina. *In:* Ciber Sistemas**. cibersistemas Games e Tecnologia. *[S.l.].* 6 ago. 2020. Disponível em: https://cibersistemas.pt/tecnologia/como-usar-algoritmos-baseados-em-arvore-para-aprendizado-de-maquina/. Acesso em: 15 out. 2021.

ANGELONI, Maria T.; REIS, Eduardo S. **Business Intelligence como Tecnologia de Suporte a Definição de estratégias para melhoria da qualidade do ensino. In: Encontro da ANPAD 2006**, 2006, Salvador. XXX Encontro Nacional de Pós-Graduação em Administração, 2006. v. 1. p. 16 páginas.

ANONIMO, **TIPOS de análise de dados: Conheça os 4 principais!**, 2021. Disponível em: https://www.fiveacts.com.br/tipos-de-analise-de-dados/. Acesso em: 20/10/2021.

ANTONELLI, Ricardo Adriano. **Conhecendo o Business Intelligence (BI); Uma Ferramenta de Auxílio à Tomada de Decisão**. Revista TECAP, Nº 3, Ano 3, Volume 3, 2009. Disponível em: Acesso em 29 Set. 2021.

AMO, S**. Curso de data mining: programa de mestrado em ciência da computação**. Uberlândia: Universidade Federal de Uberlândia, 2003. Disponível em: www.deamo.prof.ufu.br/CursoDM.html. Acesso em: 01 Set. 2021.

BARR, Avron; FEIGENBAUM, Edward A. **The Handbook of Artificial Intelligence. Stanford - California: HeurisTech Press**, Department of Computer Science, Stanford University, volume 1, p. 20 – 25. 1981.

BECKER, K. ; TUMITAN, D. **Introdução à Mineração de Opiniões: Conceitos, Aplicações e Desafios.** In: Joao Eduardo Ferreira. (Org.). Lectures of the 28th Brazilian Symposium on Databases. 1ed.Pernambuco: CIN - UFPE, 2013, v. , p. 27-52

CARDOSO, Olinda Nogueira Paes e Machado, Rosa Teresa Moreira. **Gestão do conhecimento usando data mining: estudo de caso na Universidade Federal de Lavras.** Revista de Administração Pública [online]. 2008, v. 42, n. 3, Acessado: 20 out 2021, pp. 495-528. Disponível em: <https://doi.org/10.1590/S0034-76122008000300004>. Epub 02 Set 2008. ISSN 1982-3134. https://doi.org/10.1590/S0034-76122008000300004.

CARVALHO FILHO, José Adail. **Mineração de textos: análise de sentimentos utilizando Tweets referentes à Copa do Mundo 2014**. 2014. 44 f. TCC (graduação em Engenharia de Software) - Universidade Federal do Ceará, Campus Quixadá, Quixadá, 2014.

CETAX. Data Mining: O que é, conceito e definição. *In:* CETAX. **DATA ANALYTICS, BIG DATA, DATA SCIENCE.** São Paulo, 25 jul. 2020. Disponível em: https://www.cetax.com.br/blog/data-mining/. Acesso em: 7 set. 2021.

CONCEIÇÃO, Thayná; ROSSI, Rafael. **Desenvolvimento de uma Ferramenta para Análise de Sentimentos de Textos Publicados no Twitter** - Trabalho de Conclusão de Curso - Sistemas de Informação - UFMS/CPTL. 2017.

CRIBARI, Francisco. **Método Boostrap**. 19 de abril de 2010. Disponível em: http://www.ebah.com.br/content/ABAAAAr8cAA/metodo-bootstrap. Acesso em: 18 out. 2021.

ELMASRI, R.; NAVATHE, S. B. **Sistemas de banco de dados: fundamentos e aplicações**. 3. ed. Rio de Janeiro: LTC, 2002.

FÁVERO, Luiz Paulo Lopes; BELFIORE, Patrícia Prado. **Manual de análise de dados: estatística e modelagem multivariada com excel, SPSS e stata. [S.l: s.n.]**, 2017.

GIOVANOLLI, Regina. **O IMPACTO DA INTELIGÊNCIA ARTIFICIAL NA INDÚSTRIA FINANCEIRA. Provider IT & Business Solutions**. Julho, 2017. Disponível em: https://provider-it.com.br/consultoria-de-ti/o-impacto-da-inteligencia-artificial-na-industria-financeira/. Acesso em: 18 out. 2021.

JEHNIFFER, Jaíne. **Origem dos bancos: Surgimento das cédulas e dos empréstimos.** *In:* Investidor Sardinha. **ISardinha.** São Paulo, 12 fev. 2021. Disponível em: https://investidorsardinha.r7.com/aprender/origem-dos-bancos-historia/. Acesso em: 15 out. 2021.

JUNIOR, J. R. C. **DESENVOLVIMENTO DE UMA METODOLOGIA PARA MINERAÇÃO DE TEXTOS**, 2007.

OLIVEIRA, RICARDO DANTAS DE. **OTIMIZAÇÃO DE ALGORITMOS PARA PREDIÇÃO DE DESEMPENHO ACADÊMICO DE ESTUDANTES EM AMBIENTES EDUCACIONAIS.** Orientador: Prof. Dr. Rafael Ferreira Leite de Mello. 2017. 33 f. TCC (Graduação) - Curso de Ciência da Computação, Universidade Federal Rural de Pernambuco, RECIFE, 2017. Disponível em: http://www.bcc.ufrpe.br/sites/ww3.bcc.ufrpe.br/files/Ricardo%20Dantas.pdf. Acesso em: 18 out. 2021.

REDAÇÃO, **5 grandes crises econômicas que abalaram o mundo**. Gazeta do Povo, São Paulo, 14 de set. de 2018. Disponível em: https://www.gazetadopovo.com.br/mundo/5-grandes-crises-economicas-que-abalaram-o-mundo-atheycnpmtjjl1dfe9srhaapl/. Acesso em: 01 set. 2021.

ROMANI, Mateus Flach. **Comparação de algoritmos de aprendizagem de máquina na construção de modelos preditivos para rentabilidade de clientes bancários**. 2017. 50 f., il. Trabalho de Conclusão de Curso (Bacharelado em Engenharia de Produção) - Universidade de Brasília, Brasília, 2017.

SACRAMENTO, Gabriel. **ÁRVORE DE DECISÃO: ENTENDA ESSE ALGORITMO DE MACHINE LEARNING**. *In:* TERA. **SOMOSTERA.** São Paulo, 12 jul. 2021.

Disponível em: https://blog.somostera.com/data-science/arvores-de-decisao. Acesso em: 15 out. 2021.

SANDRONI, P. **Novíssimo dicionário de economia**. São Paulo: Best Seller, 1999.

SHARDA, R., DELEN, D., & TURBAN, E. (2009). **Business Intelligence e Análise de Dados para Gestão do Negócio. 4. Ed**. Bookman Editora.

SILVA, DILANE RIBEIRO DA. Data Mining. **UMA VISÃO GERAL SOBRE MACHINE LEARNING: CLASSIFICAÇÃO**. *In:* OPER DATA. **OPER.** São Paulo, 13 ago. 2020. Disponível em: https://operdata.com.br/blog/uma-visao-geral-sobre-machine-learning/. Acesso em: 7 set. 2021.

YEH, I. C., LIEN, C. H. **comparisons of data mining techniques for the predictive accuracy of probability of default of credit card clients**. 2009. Disponível em: https://archive.ics.uci.edu/ml/datasets/default+of+credit+card+clients. Acesso em: 01 set. 2021.

ZUMEL, N.; MOUNT, J. Unsupervised methods. In: ZUMEL, N.; MOUNT, J. (Ed.) **Practical Data Science with R. New York**, NY: Ed. Manning, 2020.