PROJETO DE GRADUAÇÃO

COMPARAÇÃO DE ALGORITMOS DE APRENDIZAGEM DE MÁQUINA NA CONSTRUÇÃO DE MODELOS PREDITIVOS PARA RENTABILIDADE DE CLIENTES BANCÁRIOS

Por,
Mateus Flach Romani

Brasília, 13 de dezembro de 2017

UNIVERSIDADE DE BRASILIA

FACULDADE DE TECNOLOGIA
DEPARTAMENTO DE ENGENHARIA DE PRODUÇÃO

UNIVERSIDADE DE BRASILIA
Faculdade de Tecnologia
Departamento de Engenharia de Produção

PROJETO DE GRADUAÇÃO

COMPARAÇÃO DE ALGORITMOS DE APRENDIZAGEM DE MÁQUINA NA CONSTRUÇÃO DE MODELOS PREDITIVOS PARA RENTABILIDADE DE CLIENTES BANCÁRIOS

Por,

Mateus Flach Romani

Relatório submetido como requisito parcial para obtenção

do grau de Engenheiro de Produção

Banca Examinadora

Prof. Ph.D. Reinaldo Crispiniano Garcia, UnB/EPR (Orientador) _____

Prof. Dr. Annibal Affonso Neto, UnB/EPR _____

Brasília, 13 de dezembro de 2017

RESUMO

O presente trabalho compara os resultados de precisão de algoritmos de aprendizagem de máquina, com o intuito de verificar sua performance e adequação em relação a uma base de dados obtida em uma Instituição Financeira brasileira, construindo modelos de previsão sobre a rentabilidade de clientes bancários, e assim, sugerir uma aplicação de Inteligência Artificial em *Business Intelligence*. Três algoritmos de classificação, são utilizados e, são apresentadas as etapas que antecedem a obtenção dos resultados, como *data wrangling* e seleção de variáveis.

Palavras chave: Algoritmos, Aprendizagem de máquina, *Business Intelligence*, Classificação, Construção, Modelos, Precisão, Previsão, Rentabilidade.

ABSTRACT

The present work compares the accuracy of machine learning algorithms with the purpose of verifying their performance and adequacy using a database obtained in a Brazilian Financial Institution, building prediction models over the clients profitability, and then, suggest an Artificial Intelligence application in Business Intelligence. Three different classification algorithms were used and, each step that precede the results, like data wrangling and feature selection.

Keywords*: Accuracy, Algorithms, Build, Business Intelligence, Classification, Machine Learning, Models, Prediction, Profitability.*

SUMÁRIO

LISTA DE FIGURAS

LISTA DE TABELAS

LISTA DE SÍMBOLOS

Abreviaturas

AM	Aprendizado de Máquina
AD	Armazém de Dados
BI	*Business Intelligence* (Inteligência de Negócios)
EIS	*Executive Information Systems* (Sistemas de Informação Executivo)
FN	*False Negative* (Falso Negativo)
FP	*False Positive* (Falso Positivo)
IA	Inteligência Artificial
IF	Instituições Financeiras
MD	Mineração de Dados
MLG	Modelos Lineares Generalizados
NB	Naive Bayes
RF	*Random Forest* (Floresta Aleatória)
TI	Tecnologia da Informação
TN	*True Negative* (Verdadeiro Negativo)
TP	*True Positive* (Verdadeiro Positivo)

1. INTRODUÇÃO

A estratégia corporativa é um dos meios pelos quais as empresas podem obter vantagens competitivas para disputar mercados com seus concorrentes. Porter (2004), diz que a estratégia competitiva faz um exame do modo como uma empresa pode competir com maior eficácia para fortalecer sua posição no mercado. Atualmente, novas ferramentas tecnológicas têm sido utilizadas para aprimorar a estratégia das organizações e, é de seu grande interesse, utilizar a tecnologia da informação para auxiliar no processo de tomada de decisão. De acordo com Primak (2008), o mundo empresarial começou a se comportar de modo mais complexo e a tecnologia da informação (TI) progrediu para oferecer informações precisas para alinhar ações com foco na melhoria do desempenho no mundo dos negócios.

O desenvolvimento de TI trouxe ao setor corporativo o *Executive Information Systems* (EIS). Uma tecnologia cujo objetivo era fornecer informações empresariais a partir de uma base de dados. Surgiu então na década de 80 o termo *Business Intelligence* (BI), uma evolução do EIS, o que despertou grande interesse das empresas pois trazia informações de forma rápida e apropriada para tomada de decisão. (PRIMAK 2008).

O contínuo desenvolvimento em TI fez surgir novas tecnologias que estão sendo usadas como ferramentas no ambiente corporativo e, podem causar grande impacto nos negócios e na economia, como é o caso da inteligência artificial (IA), *big data* e *blockchain*. Os estudos referentes a IA remontam ao período subsequente à Segunda Guerra Mundial e, de acordo com Russel e Norvig, (2013), é um dos campos mais recentes da ciência e engenharia que tentam, não apenas compreender, mas também construir entidades inteligentes.

A empresa de consultoria Accenture, analisou 12 economias desenvolvidas e constatou que a IA é "o novo fator de produção", com o potencial de dobrar o crescimento econômico dos países estudados. O jornal *The New York Times*, cita a ameaça da IA para o mercado de trabalho também tem sido citada, pois espera-se que haja decréscimo na quantidade de postos de trabalho devido a automação de atividades. (PURDY e DAUGHERTY 2016). (LEE 2017).

Inserido na área de IA encontra-se o aprendizado de máquina (AM), cujo estudo visa o desenvolvimento de técnicas computacionais sobre o aprendizado além da construção de sistemas capazes de adquirir conhecimento de forma automática. O aprendizado de máquina permite que se obtenha *insights* a partir dos dados da empresa e com isso, gerar valor para o negócio. São

citados diversos exemplos de sucesso na aplicação de AM em seu ambiente de negócios como, Airbnb, Zynga e GE. (BARANAUSKAS 2007). (TAURION 2015).

Instituições financeiras (IF), também estão interessadas nos benefícios de aplicação de AM, respostas mais rápidas às solicitações, interações assertivas e consequente rentabilização dos clientes, fazem parte do objetivo que os bancos têm em relação ao desenvolvimento de IA no setor financeiro. Novamente a empresa Accenture, traz dados recentes sobre a aplicação de IA nesse setor. De acordo com seu relatório *Banking Technology Vision*, 79% dos bancos entrevistados concordam que a IA revolucionará o modo como obtêm informação e interagem com os clientes. (ACCENTURE 2017).

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

Apesar do campo de estudo em aprendizagem de máquina ser algo relativamente recente, o uso de IA e AM na indústria financeira ainda está direcionado à detecção de fraudes, análise de crédito bancário e, recentemente, para sugestão de investimentos. Porém, há potencial para serem utilizados na obtenção de informações para o nível estratégico, como por exemplo, prever o custo variável com cada cliente, ou na área de marketing, na verificação de quais ofertas de produtos são mais assertivas de acordo com cada perfil de consumo, ou de acordo com dados obtidos em fontes externas, e até mesmo, prever qual a rentabilidade de cada cliente, para assim, planejar ações que possam antecipar receitas.

O Capítulo 2 trará conceitos relevantes para o entendimento dos modelos estatísticos que estão inseridos nos algoritmos utilizados, além de fornecer um contexto histórico e corporativo sobre o desenvolvimento da inteligência artificial nas empresas. O Capítulo 3 fornecerá os procedimentos e ferramentas utilizados na obtenção dos resultados, além de alguns detalhes de implementação específicos do software utilizado. No Capítulo 4 serão disponibilizados os resultados respectivas análises. Por fim, o Capítulo 5 fornecerá as conclusões a respeito dos objetivos específicos e sugestões de trabalhos futuros.

Este trabalho pode ser classificado como uma pesquisa quantitativa e qualitativa de natureza aplicada e caráter exploratório, uma vez que o estudo tem mais familiaridade com problema de aplicação prática já que a fonte de informação advém do campo corporativo.

1.1 Objetivos

1.1.1 Problema

O problema ao qual esse trabalho deseja tratar é, prever a rentabilidade a partir de outros dados dos clientes.

1.1.2 Objetivo Geral

Comparar o desempenho de três algoritmos classificadores de aprendizagem de máquina, utilizando uma base de dados obtida em uma instituição financeira brasileira, na construção de modelos de previsão, sobre a rentabilidade de clientes pessoa física.

1.1.3 Objetivos Específicos

Serão considerados os seguintes pontos para a análise comparativa proposta:

- O aumento no número de variáveis aumenta a precisão do modelo.
- O aumento da quantidade de dados aumenta a precisão do modelo.
- As variáveis que apresentarem maior correlação, em relação a variável alvo, serão as que darão maior contribuição à precisão do modelo.

1.1.4 Hipóteses

As hipóteses foram levantadas a partir do método da pesquisa a fim de responder os objetivos específicos:

- A precisão dos modelos será maior, quanto maior for a quantidade de dados de treino.
- A precisão dos modelos será menor se a quantidade de dados de teste for maior que a quantidade de dados de treino, em relação aos testes em que a quantidade de dados nos conjuntos de treino e teste forem iguais.
- A precisão dos modelos terá incrementos percentuais proporcionais à medida que as variáveis forem adicionadas.

2. REFERENCIAL TEÓRICO

Este capítulo apresenta os principais conceitos relacionados a preparação dos dados, aprendizagem de máquina, modelagem preditiva e avaliação de modelos.

2.1 *Business Intelligence*

As corporações necessitam de auxilio nas mais diferentes situações para a tomada de decisão. A inteligência de negócios (*Business Intelligence*) tem sido amplamente demandada pelas empresas pois consegue auxiliar a tomada de decisão de forma rápida e adequada.

De acordo com Primak (2008), o termo *Business Intelligence* surgiu nos anos 80 no Gartner Group e faz referência ao processo inteligente de coleta, organização, análise, compartilhamento e monitoração de dados contidos em *Data Warehouse* / *Data Mart*, gerando informações para o suporte a tomada de decisões no ambiente de negócios.

> "O entendimento de aspectos básicos de inteligência de negócios envolve a definição de alguns termos. Os dados são observações, percepções e fatos. Informação é o resultado do processamento dos dados que estejam em um contexto e tenham relevância e propósito. Conhecimento é a crença justificada sobre um relacionamento o qual é relevante para tomada de decisão". (SABHERWAL e BECERRA-FERNANDEZ 2011).

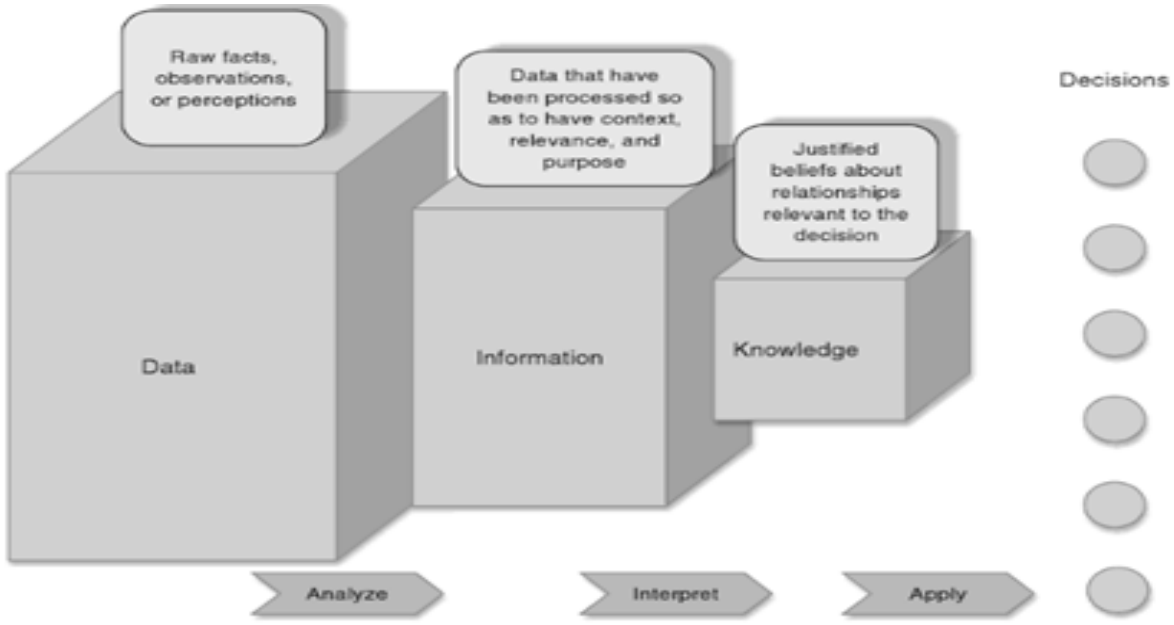

Figura 2.1 – Dados, Informação e Conhecimento.

Fonte: Saberwhal e Becerra-Fernandez, 2011.

O BI funciona captando dados de armazém de dados (AD), ou *data warehouses*. Com esses dados aplicam-se técnicas como mineração de dados (MD), ou *data mining*, que buscam obter padrões e correlações entre os dados, fornecedo informações importantes para o negócio.

De acordo com Dutra (2005), os sistemas de BI têm como principais características:

a) Extrair e integrar dados de múltiplas fontes;
b) Fazer uso da experiência e conhecimento adquirido por seus usuários;
c) Analisar dados dentro de uma cadeia de processos negócios;
d) Trabalhar com múltiplas hipóteses e simulações;
e) Extrair padrões de comportamento e classifica-los em categorias.

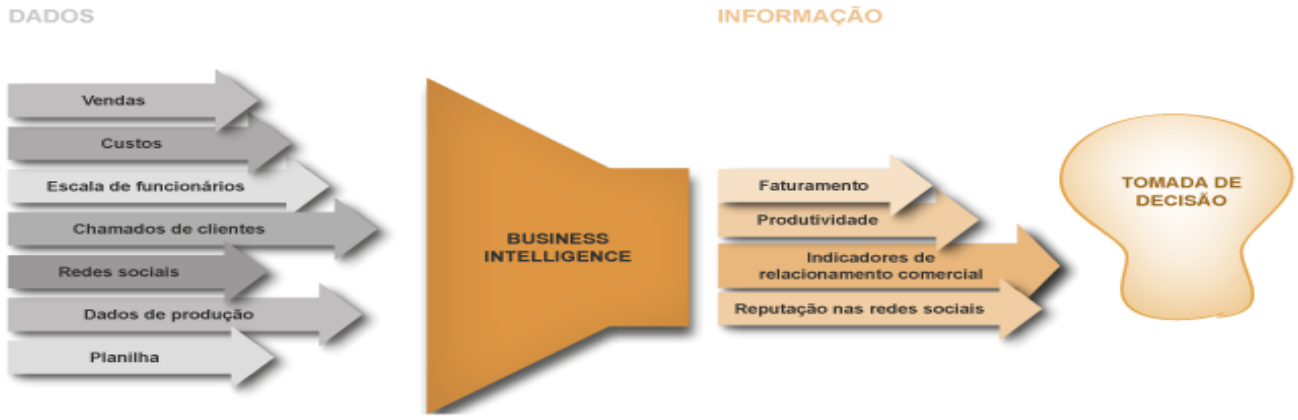

Figura 2.2 – Como Funciona o BI.

Fonte: Know Solutions.

2.2 Inteligência Artificial

A inteligência artificial está sendo utilizada para aprimorar os processos de BI. Análises reativas e relatórios estáticos estão dando lugar a análises proativas e relatórios em tempo real. Essa mudança está alterando a forma como organizações realizam seu BI, exigindo maiores conhecimentos do que análise de dados.

A inteligência artificial é parte da ciência da computação preocupada em desenvolver sistemas de computadores inteligentes, ou seja, sistemas que apresentem características que associamos com a inteligência humana. (BARR e FEIGENBAUM 1981).

Russel e Norvig (2013), listam as áreas de conhecimento que contribuíram com ideas e pontos de vista para a IA, dentre elas estão a filosofia, com contribuições desde Aristóteles que desenvolveu um sistema informal de silogismos que permitia gerar conclusões mecanicamente, até a teoria da confirmação de Carnap e Hempel que tentava compreender a aquisição de conhecimento por meio da experiência. Na matemática houve contribuições das áreas de lógica, computação e

probabilidade, das quais tem-se os respectivos cientístas George Boole, Alan Turing e Thomas Bayes. A economia contribuiu com os campos de conhecimento da teoria dos jogos, processos de decisção de Markov e a pesquisa operacional, os quais ajudam a compreender como as pessoas fazem escolhas. A Fig. (2.3) ilustra a historia da IA em alguns de seus momentos importantes.

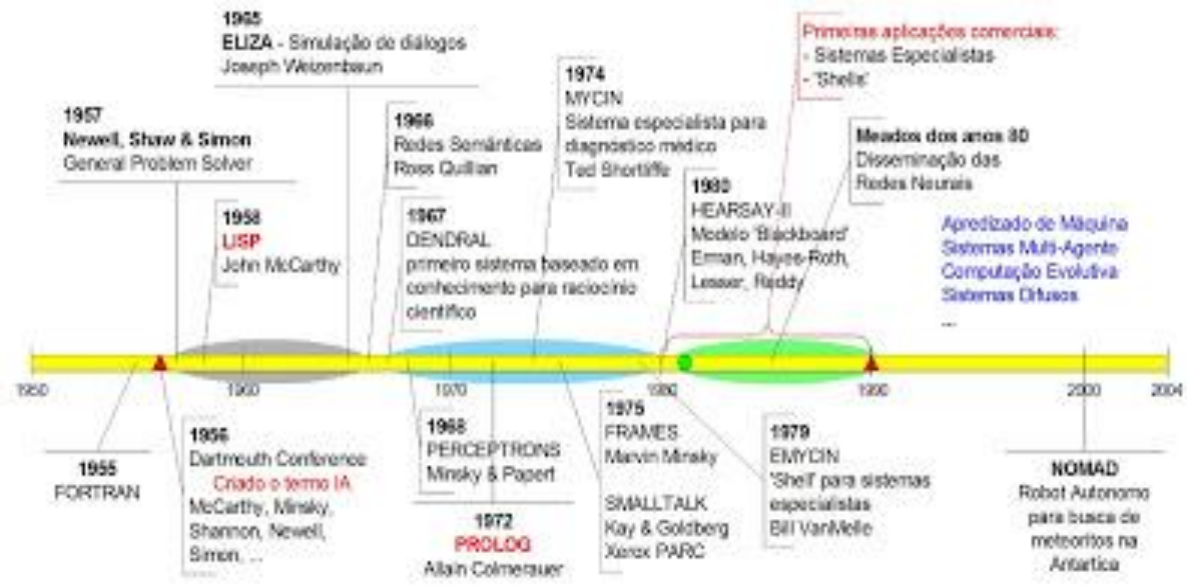

Figura 2.3 – História da IA.

Fonte: PUC Minas.

A Tab. (2.1) resume algumas das subáreas da IA e seus respecitvos campos de estudo.

Tabela 2.1 – Subáreas da IA e seus Campos de Estudo.

Subárea	Campo de estudo
Sistemas Baseados em Agentes	Criar sistemas que permitam o estudo dos agentes inteligentes estudando seu comportamento e desempenho.
Busca (*Pathfinding*)	Desenvolvimento de algoritmos mais eficientes baseados em grafos.
Planejamento Automatizado	Estudar como maximizar o desempenho de um agente inteligente na realização de suas tarefas.
Aprendizado de Máquina	Criar algoritmos e técnicas que permitam a criação de modelos matemáticos baseados na experiência.
Robótica	Criar agentes físicos que por meios técnicos se tornem capazes de interferir no mundo real.

Fonte: Elaborado pelo autor.

2.3 Algoritmos Utilizados

A capacidade de aprendizado é considerada essencial para um comportamento inteligente. Em AM, computadores são programados para aprender com a experiência passada. Para tal, empregam um princípio de inferência denominado indução. (FACELLI 2011).

O aprendizado indutivo segue dois caminhos dependendo das tarefas de aprendizado, divididas em tarefas preditivas e descritivas. Tarefas preditivas geram saídas a partir de valores de entrada e requerem um supervisor externo que conheça a saída, portanto o aprendizado é supervisionado. As tarefas descritivas têm como objetivo a exploração e agrupamento de um conjunto de dados sem a necessidade de gerar uma saída, nesse caso o aprendizado é não supervisionado. A Fig. (2.4) mostra a divisão por tipo de tarefas.

Figura 2.4 - Tipos de Aprendizado de Máquina.

Fonte: Facelli et al. 2011.

A principal diferença entre algoritmos de classificação e regressão é que o primeiro é usado para valores discretos e o segundo para valores contínuos. Além disso o primeiro não necessita de uma série histórica para análise, apenas os dados em um ponto no tempo ou de um período curto são suficientes para realizar as análises, ao contrário do segundo.

2.3.1 *Random Forest*

O algoritmo *Random Forest* (RF) é um algoritmo de classificação construido sobre algumas melhorias feitas em algoritmos de árvores de decisão, utilizando um método de reamostragem (*bootstrap*) associado a agregação.

> "O método de *bootstrap* se baseia em uma amostra aleatória $y = (y_1, y_2, \dots, y_n)$ cujos valores são realizações de variáveis aleatórias independentes e identicamente distribuídas $Y_1, \dots, Y_n$, 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).

A associação entre reamostragem e agregação produz o *baggin* (*bootstrap aggregation*), um procedimento generalizado, utilizado para reduzir a variância de alguns algoritmos, como por exemplo, o algoritmo de árvore de decisão. O processo de construção utiliza a indução para criar uma estrutura com ramificações a partir dos dados de entrada, com o intuito de estimar os resultados mais prováveis.

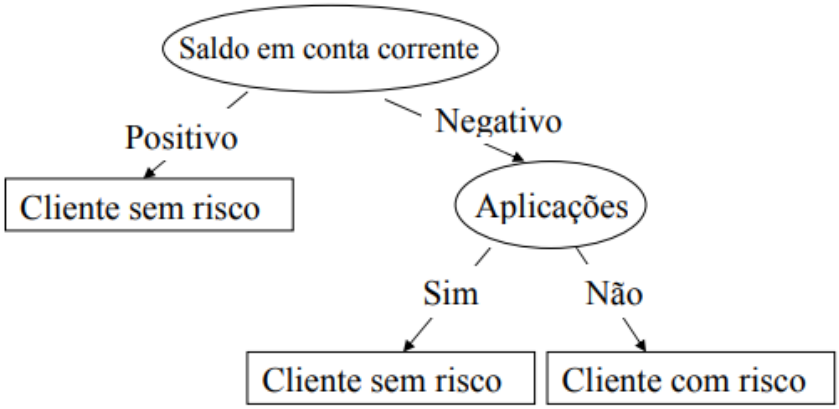

Figura 2.5 – Exemplo Fictício de Árvore de Decisão com Clientes de Alguma IF.

Fonte: Zuben e Attux.

Hastie, Tibshirani e Friedman (2009), colocam o RF como um algoritmo muito popular na áreas de pesquisa e de negócios, dada sua performace satisfatória. Seu método reduz a variância de um grande número de modelos complexos, ao contrário dos modelos de *boosting* que reduzem o viés dos modelos, além de ser fácil de treinar e modificar seus parâmetros. Os autores definem as etapas do algoritmo com a seguinte sequência de passos:

1. Para $b = 1$ até B:
    a. Desenhe a reamostragem $\mathbb{Z}^*$ de tamanho N dos dados de treino.
    b. Construa uma árvore da floresta aleatória T_b para os dados da reamostragem, por recursividade repetindo os seguintes passos para cada nó da árvore, até que o nó de menor tamanho n_{min} seja alcançado.
        i. Selecione m variáveis aleatórias dentre p variáveis
        ii. Escolha a melhor variável/ponto de divisão entre m.

iii. Divida o nó em dois nós filhos.

2. Gere a saída com o conjunto de árvores $\{T_b\}_1^B$.

Para fazer uma previsão em um novo ponto x:

Regressão: $\hat{f}_{rf}^B(x) = \frac{1}{B}\sum_{b=1}^{B} T_B(x)$

Classificação: Seja $\hat{C}_b(x)$ a classe preditora da b-ésima árvore de floresta aleatória. Faça a função classe, $\hat{C}_{rf}^B(x) = maioria\ dos\ votos\ \{\hat{C}_b(x)\}_1^B$. (HASTIE, TIBSHIRANI e FRIEDMAN 2009).

A principal vantagem do RF em relação aos algoritmos de *boosting* e árvore de decisão é sua resistência ao *overfitting*, ou sobreajuste. Vários algoritmos complexos sofrem com esse problema, quando uma grande quantidade de dados de treino é passada ao algoritmo e esse se adapta muito bem ao conjunto fornecido, mas se mostra ineficaz para prever novos conjuntos de dados.

2.3.2 Naive Bayes

O classificador Naive Bayes é uma rede bayesiana com uma estrutura fixa onde cada variável preditora é independente das outras variáveis, dada a variável alvo. As conexões saem da variável alvo em direção a todas as variáveis preditoras. (OLIVEIRA 2016).

O algoritmo Naive Bayes (NB) usa o teorema de Bayes em seu código, para prever o resultado de variáveis numéricas e categóricas. Esse teorema fornece uma forma de calcular a probabilidade posterior $P(B|A)$, a partir de $P(B)$, $P(A)$ e $P(A|B)$. "Seja $(\Omega, \mathcal{A}, P)$ um espaço de probabilidade. Se $B \in \mathcal{A}$ e $P(B) > 0$, a probabilidade condicional de $A \in \mathcal{A}$ dado B é definida por:"

$$P(A|B) = \frac{P(A \cap B)}{P(B)}. \qquad (2.1)$$

(NETO 2010). Outra forma de escrever o teorema é:

$$P(B = k|A = x) \propto \pi_k\, f_k(x). \qquad (2.2)$$

Em que π_k é o quão provável o caso k é, antes que se saibam os dados x e, $f_k(x)$ é a probabilidade dos dados x dada a probabilidade B em k.

De acordo com Carvalho (2017), considerando A um vetor grande $A = (A_1, A_2, \dots, A_p)$, e fazendo

$$f_k(x) = P\big(A_1 = x_1, A_2 = x_2, \dots, A_p = x_p \big| B = k\big), \qquad (2.3)$$

para um p grande. Como as variáveis são linearmente independentes dado B, então:

$$f_k(x) = f_k^1(x_1) f_k^2(x_2),\ f_k^i(x_i) = P(A_i = x_i | B = k). \quad (2.4)$$

Assim, para um valor geral de p tem-se:

$$f_k(x) = \prod_{i=1}^{p} f_k^i(x_i). \quad (2.5)$$

O fato de que o algoritmo assume a independência entre as variáveis o torna muito rápido em seus cálculos, mesmo utilizando grandes quantidades de dados, pois não se utiliza processamento para calcular a correlação entre elas, isso pressupõe uma economia de tempo, o que, em situações práticas, faz com que seja muito utilizado por grupos de pesquisa quando deseja-se obter estimativas ou previsões rápidas.

2.3.3 Regressão Logística

A regressão logística é uma técnica estatística que permite a previsão de variáveis categóricas ou, variáveis discretas, na qual se tenham apenas duas respostas possíveis. Para um conjunto de dados que apresentem mais de duas variáveis categóricas possíveis, utilizam-se os modelos multicategóricos.

O princípio da regressão logística tem base na técnica de regressão linear onde a variável quantitativa y é definida como uma relação linear das variáveis preditoras x de acordo com a seguinte fórmula: $h_x = w_o + w_1 x_1 + \cdots + w_k x_k$. Nesse caso, w_k representa os pesos dados a cada varíavel preditora. (OLIVEIRA 2016).

Esse algoritmo se encontra dentro de uma família denominada de modelos lineares generalizados (MLG), os quais se diferenciam pela função de ligação. Os modelos analisados será o modelo multicategórico Logit.

Os modelos Logit multicategóricos são análogos aos modelos de regressão logística tradicionais, exceto pelo fato de que a distribuição de probabilidades de resposta é multicategórica ao invés de binomial e, tem-se $J - 1$ equações ao invés de uma. Esse modelo assume também, que o log das razões de probabilidade para cada categoria de resposta seguem a seguinte equação:

$$\eta_{ij} = \log \frac{\pi_{ij}}{\pi_{iJ}} = \alpha_j + x_i' \beta_j, \quad (2.6)$$

onde π_{ij} é a probabilidade que a i-ésima observação caia na j-ésima categoria de resposta em um total de J categorias possíveis, α_j é uma constante, β_j é o vetor de regressão dos coeficientes para $j = 1, \ldots, J - 1$.

Quando a função de ligação segue a distribuição logística, o modelo Logit multicategórico, segue a seguinte expressão.

$$P_1 = \frac{exp(D_1)}{\sum_{i=1}^{3} exp(D_i)}. \quad (2.7)$$

2.4 Métricas de Avaliação

A capacidade preditiva de um modelo pode ser avaliada por diversas ferramentas estatísticas. As ferramentas mais utilizadas procuram mensurar o quanto o modelo consegue explicar os dados fornecidos e, qual sua precisão ao tentar prever novos dados.

Para modelos binários de classificação, os resultados são divididos em quatro categorias. Os resultados positivos que foram previstos corretamente como positivos (TP), os resultados negativos que foram previstos corretamente como negativos (TN), os resultados positivos que foram previstos como negativos (FN), e os resultados negativos previstos como positivos (FP).

Os modelos multicategóricos tem maior quantidade de resultados possíveis do que os modelos binários, dado que cada categoria existente se relaciona com a outra, assim para um número n de categorias tem-se n^2 resultado possíveis. Dentre os resultados possíveis existem n resultados previstos corretamente. Uma das métricas utilizadas é a precisão (ACC) ou taxa de acerto, que é calculada somando todos os resultados previstos corretamente e os dividindo pela soma de todos os resultados, tanto os previstos corretamente quanto os previstos de forma incorreta. Para o caso binário a fórmula seria:

$$ACC = \frac{\sum TP + \sum TN}{\sum Total\ de\ resultados}. \quad (2.8)$$

Os modelos construídos devem ter como resultado de precisão mínimo, o valor de 70% para poder ser considerado satisfatório. Esse valor é uma convenção estabelecida pelos analistas e estatísticos que atuam na área modelos preditivos e será utilizado nesse trabalho como parâmetro comparativo entre os resultados dos modelos.

Outra métrica que auxilia a precisão é o intervalo de confiança (IC). Essa estimativa de parâmetro populacional calcula, a partir das observações, qual a frequência em que o parâmetro de interesse, no caso a precisão, contém o parâmetro de interesse quando o experimento é repetido várias vezes. De acordo com Paternalli (2009) o IC é calculado como

$$IC = \bar{X}\ \frac{t_\alpha}{2} \pm \frac{s}{\sqrt{n}}. \quad (2.9)$$

2.5 Rentabilidade

Existem diversos conceitos e diferentes formas de se medir a rentabilidade de clientes. Uma das formas de se mensurar tal indicador é pela análise de contribuição marginal. Essa análise depende dos custos variáveis e dos custos fixos. Na ausência de informações sobre custos fixos o melhor é trabalhar com a margem de contribuição, que é dada pela diferença entre o preço de venda e o custo variável de cada cliente. (COBRA 2009).

Para o caso de clientes bancários, é difícil precificar o custo fixo que cada cliente tem, pois, a demanda pelos serviços prestados pode variar ao longo do tempo para um mesmo cliente, por isso, trabalha-se com o indicador de margem de contribuição.

Alguns autores consideram o uso da margem de contribuição inadequado para tomada de decisão, devido às desvantagens do custeio variável. A existência de custos mistos (custos com parcela fixa e outra variável), torna difícil de separar o valor de cada uma, mesmo utilizando técnicas estatísticas, essa divisão, muitas vezes, se torna arbitrária. (BRUNI e FAMÁ 2009).

Complementando essa visão, argumenta-se que a análise de rentabilidade dos clientes pode ser aprimorada com a alocação dos custos e despesas.

> "[...] utilizando exclusivamente a margem de contribuição, os dispêndios de recursos com atendimento aos clientes são considerados como despesas do exercício e, portanto, lançados na demonstração de resultados, [...]. Destarte, tais procedimentos acabariam distorcendo o valor a ser considerado para atendimento aos clientes [...]" (WERNKE e LEMBECK 2004).

Entretanto, muitas empresas ainda utilizam esse indicador em seus sistemas de informação, ainda que não seja o indicador ideal para mensurar resultado e otimizar recursos disponíveis, a informação disponibilizada pela margem de contribuição é útil e em boa medida, permite que as organizações tomem decisões. Uma possível análise de margem de contribuição então, é cruzar as informações de clientes e produtos, permitindo a eliminação daqueles com margem seja negativa, melhorando o resultado geral. (MÜLLER e KRIGER 2002).

3. METODOLOGIA

Este capítulo descreve os procedimentos e técnicas utilizados para a comparação dos modelos preditivos, bem como a base de dados obtida, o softwares utilizados, alguns pacotes para manipulação dos dados e, o processo para obtenção dos resultados.

A metodologia utilizada seguiu as etapas de preparação dos dados, seleção de variáveis, construção dos modelos e, sequências de testes, assim, será possível comparar a precisão de cada algoritmo e verificar quais variáveis podem dar maior contribuição para a predição dos modelos construídos.

A primeira etapa será de explanação dos métodos e técnicas da pesquisa, sendo detalhados os programas usados, características da base de dados coletada e detalhamento de informações dessa, procedimentos e funções utilizados para uniformizar e preparar os dados, seleção das variáveis preditoras e respectivas considerações. Portanto, pretende-se fazer uma contextualização inicial da pesquisa a fim de que se possa depreender maior entendimento quanto a situação do trabalho.

Em seguida pretende-se explicar os testes e simulações realizados para obtenção dos resultados dos três algoritmos. Na sequência, será feito o detalhamento dos parâmetros utilizados em cada algoritmo e possíveis especificidades de cada um. A figura abaixo simplifica a visualização do método deste trabalho.

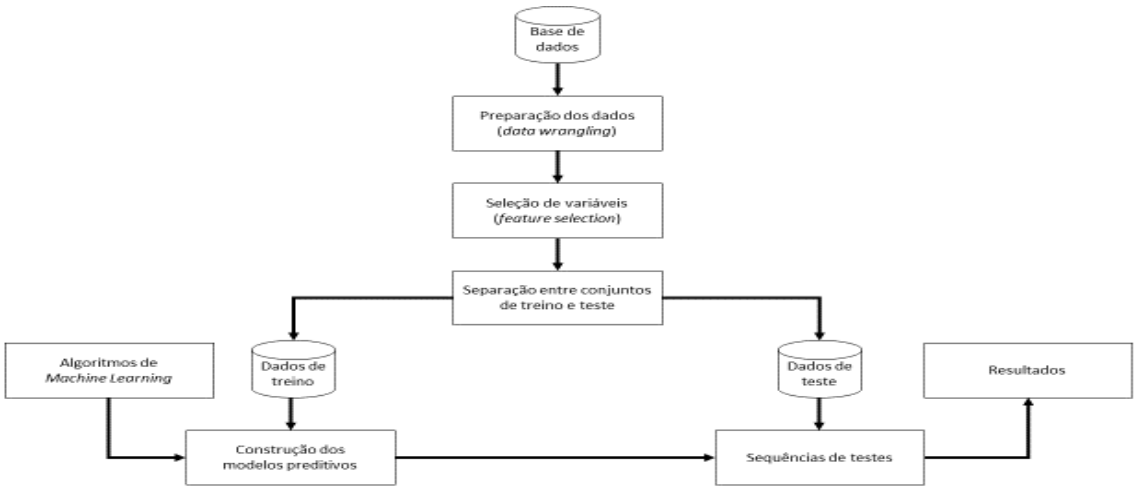

Figura 3.1 – Método usado no trabalho.

Fonte: Elaborado pelo autor.

3.1 Preparação dos Dados

A base de dados foi obtida em uma grande Instituição Financeira Brasileira, que atua em todo o território nacional e em diversos segmentos de mercado. A coleta dos dados foi realizada no mês de abril de 2017 e, esses referem-se ao mês de março do mesmo ano. Após a coleta, os dados foram consolidados em um único arquivo de extensão ".txt". Não foi feito qualquer tipo de seleção específica para os registros, exceto um filtro para selecionar clientes pessoa física de forma aleatória.

Essa base apresentada diversas informações de clientes denominadas variáveis, como faixa de renda, ocupação, nível de escolaridade, margem de contribuição, endividamento dentro da Instituição Financeira, endividamento no Sistema Financeiro Nacional, quantidade de produtos, tarifa mensal, segmento ao qual o cliente pertence, uso de cheque especial, uso de cartão de crédito, entre outros. A variável que se deseja prever, margem de contribuição, possui vinte e duas categorias, cada uma indicando uma faixa de valores de rentabilidade, distribuídas da conforme Tab. (4.1). Para calcular essa variável, o sistema da Instituição Financeira considera principalmente os juros pagos ao mês pelo cliente, a tarifa mensal da conta corrente cobrada, a taxa de administração dos investimentos do cliente e, demais produtos que o cliente venha a ter como previdência privada e seguros de vida.

Os softwares utilizados para a seleção de variáveis e construção dos modelos foram o *Microsoft Azure Machine Learning Studio* e o *R Studio* versão 1.0.143. O primeiro é uma plataforma online de computação na nuvem que possibilita a implementação de modelos de ML de forma rápida e fácil, sendo um ambiente com aparência amigável e operação intuitiva. O segundo é um ambiente livre de desenvolvimento integrado em linguagem R, muito utilizada para cálculos estatísticos e aprendizagem de máquina.

Os dados foram primeiramente importados no *R Studio* utilizando o método "*read.fwf*", pertencente ao pacote "*utils*", nativo do *R Studio*, o qual lê informações cujos caracteres tenham comprimento fixo definido pelo usuário. Após esse processo, os dados estavam organizados na forma de um *data frame*, com as respectivas colunas nomeadas porém, se encontravam em formato numérico, assim, utilizou-se o método "*lapply*" e a função de retorno "*as.factor*", em todo o *data frame* para transformar os dados de formato numérico para categórico.

Ao fim, foi utilizada a função "*replace*" do pacote básico do *R*, para uniformizar a sequência de categorias de todo o *data frame*, dado que a categoria que representa o estado "sem informação" é o número "70", esse foi substituído pelo número "0". Esse procedimento foi efetuado devido a

necessidade de ordenação das categorias de cada variável, de forma a permitir que os algoritmos ordinais de regressão pudessem estabelecer a ordem das categorias de cada variável e assim, calcular a precisão dos modelos.

Após a análise da base de dados, verificou-se que algumas variáveis estavam com grande quantidade de dados faltantes e outras não forneciam informação relevante quando associadas às demais variáveis, portanto, decidiu-se excluí-las para economizar tempo e processamento. Para isso foi associado o atributo "*NULL*" à cada variável que não era de interesse do estudo, finalizando assim o processo de preparação dos dados o qual, teve como resultado final uma base de dados com 76 variáveis e 2.782.560 registros. As variáveis estão detalhadas nos Anexos deste trabalho.

3.2 Seleção de Variáveis

A seleção de variáveis ou, *feture selection*, é o processo que seleciona um subconjunto das variáveis mais relevantes para a construção de modelos preditivos. Esse processo é diferente de uma simples redução de dimensionalidade, que procura criar novas variáveis a partir da associação de variáveis existentes. Ambos os métodos objetivam reduzir o número de atributos em uma base de dados, porém a seleção de variáveis apenas inclui ou exclui atributos sem alterá-los.

O método utilizado para selcionar as variáveis foi analisar uma matriz de correlação de Pearson e verificar quais das variáveis possuem maior correlação com a variável margem de contribuição. A correlação de Pearson mede o grau de correlação linear entre duas variáveis quantitativas. É um índice adimensional com valores situados entre os números -1 e 1, e reflete a intensidade de uma relação linear entre dois conjuntos de dados. (VARGAS 2012).

Construiu-se então uma matriz de correlação para detectar quais variáveis poderiam dar maior contribuição à construção dos modelos preditivos. O programa utilizado para esse cálculo foi o *Microsoft Azure Machine Learning Studio*, ferramenta online de computação na nuvem, disponibilizada pela *Microsoft* e, gratuíta para uma base de dados de até dez *gigabytes* de tamanho. Essa ferramenta foi escolhida pois possui uma interface gráfica amigável, o processamento é feito na nuvem de computadores, é facil de se programar e, é compatível com os demais produtos da *Microsoft*, como *Excel* e *Word*. Abaixo segue figura do experimento realizado.

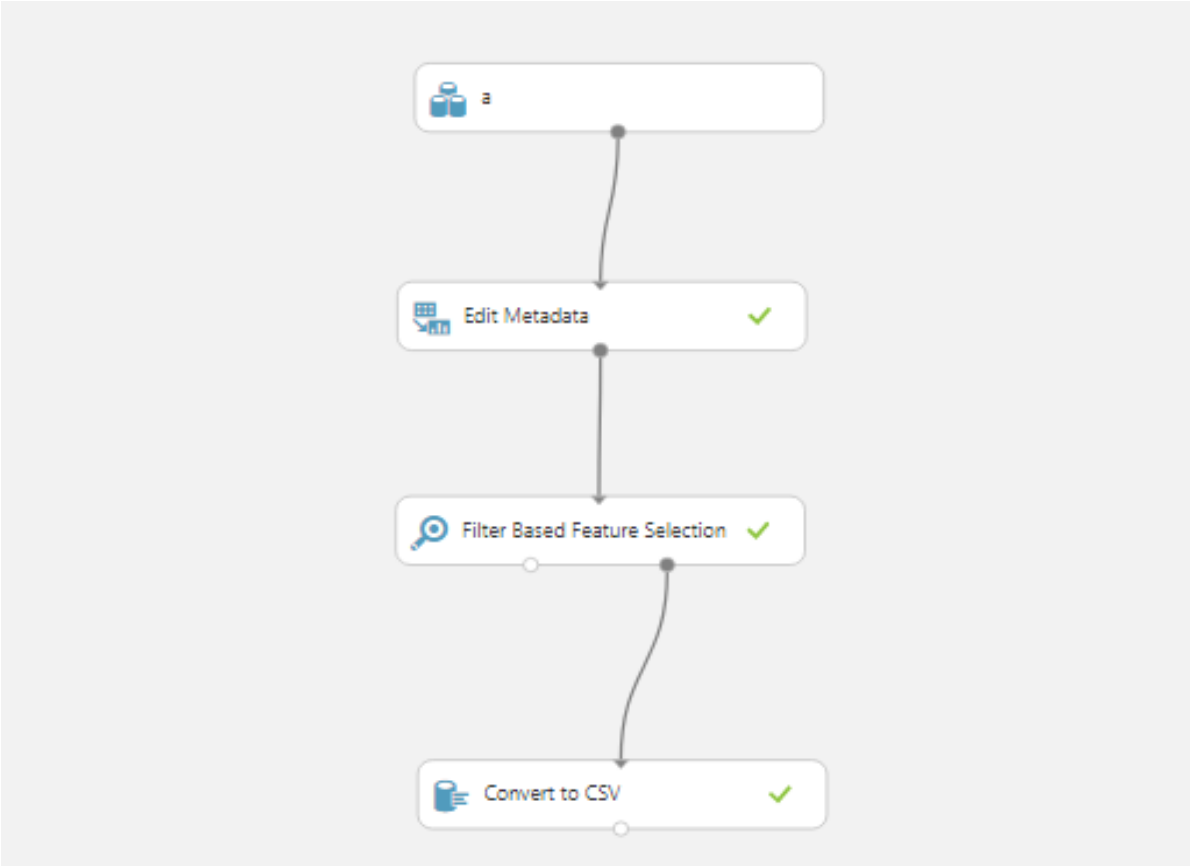

Figura 3.2 - Seleção de Variáveis no *Microsoft Azure Machine Learning Studio*.

Fonte: Elaborado pelo autor.

O arquivo "a.csv", contém a base de dados tratada no *R Studio*, a qual foi exportada para o ambiente da *Microsoft*. O bloco "*Edit Metadata*" transforma as variáveis de numéricas para categóricas. Na sequência, o bloco "*Filter Based Feature Selection*" possibilita a escolha de critérios para selecionar as variáveis, no caso, foi escolhida a correlação de Pearson. Por fim, o bloco "*Convert to CSV*" converte os resultados em um arquivo de extensão ".csv" contendo as correlações entre as variáveis. A Tab. (4.1) ilustra a matriz de correlações obtida nessa etapa.

A Tab. (3.1) apresenta as variáveis resultantes desta etapa e sua descrição.

Tabela 3.1 – Índice, Abreviatura, Legenda e Descrição das Variáveis Selecionadas.

Índice	Abreviatura	Legenda	Descrição
0	mc	Margem de Contribuição	Estima o quanto cada cliente dá de rentabilidade à empresa.
1	assi	Possui Empréstimo	Indica se o cliente possui contrato de empréstimo assinado.
2	end_if	Endividamento na IF	Classifica a faixa de volume de endividamento do cliente dentro da Instituição Financeira.
3	prod	Produtos	Quantidade de produtos que o cliente possui.
4	chq_ac	Uso Cheque Especial Acima de 50%	Quantidade de vezes que o cliente usou o cheque especial acima de 50% nos últimos três meses.
5	chq_ab	Uso Cheque Especial acima 10% Abaixo de 50%	Quantidade de vezes que o cliente usou o cheque especial acima de 10% e abaixo de 50% nos últimos três meses.
6	sms	Serviço de SMS	Indica se o cliente possui ou não serviço de mensagens por SMS.
7	pgt_car	Pagamento da Fatura do Cartão de Crédito	Faixa de valor pago da fatura do cartão de crédito nos últimos 60 dias.
8	dbt	Débito Automático	Indica quantos débitos automáticos ativos o cliente possui em sua conta corrente.
9	salario	Recebimento de Salário	Informa se o cliente recebe ou não salário pela conta da IF.
10	car_adq	Pontos do Programa de Relacionamento Adquiridos no Mês.	Quantidade de pontos que o cliente adquiriu ao pagar a fatura do cartão de crédito.
11	p_cdc	Prestação Disponível para Empréstimo Direto ao Consumidor	Indica qual a faixa de prestação disponível para contratação de empréstimo direto ao consumidor que o cliente possui.
12	pac	Tarifa Mensal da Conta Corrente	Classifica por faixa de valores a tarifa mensal que o cliente paga em sua conta corrente.
13	end_sfn	Endividamento no Sistema Financeiro Nacional	Informa a faixa de volume de endividamento que o cliente possui no Sistema Financeiro Nacional.
14	p_con	Prestação Disponível para Empréstimo Consignado	Indica qual a faixa de prestação disponível para contratação de empréstimo consignado que o cliente possui.
15	pts	Pontos do Programa de Relacionamento	Quantidade de pontos que o cliente possui no programa de relacionamento.
16	inv	Investimentos	Faixa de valor investido que o cliente possui na Instituição Financeira.
17	lim	Limite Disponível no Cartão de Crédito	Valor que o cliente possui disponível para compras no cartão de crédito.

Fonte: Elaborado pelo autor.

Inicialmente, foram selecionadas como variáveis preditoras aquelas que possuiam maior correlação com a variável margem de contribuição, considerando como valor de corte a correlação em torno de 0,50, totalizando 17 variáveis preditoras. Em seguida foi definida uma abreviação e legenda para as variáveis selecionadas de modo a facilitar a operação e manipulação dos algoritmos durante a construção dos modelos preditivos. A Tab (3.1) resume a abreviatura, legenda e descrição das variáveis preditoras selecionadas e da variável a ser prevista.

3.3 Construção dos Modelos

Selecionadas as variáveis preditoras, fez-se a construção dos modelos preditivos a partir dos métodos que cada algoritmo possui no programa *R Studio*. A construção desses modelos poderia ser feita também no *Microsoft Azure Machine Learning Studio*, entretanto, considerando que a linguagem *R* é mais usada no ambiente de AM e, seus métodos são mais consolidados e robustos, decidiu-se que os modelos seriam criados e testados no programa *R*.

Ao construir o modelo RF, foi utilizado o método "*randomForest*", do pacote de mesmo nome, utilizando como argumento, além das variáveis de interesse, os argumentos "*nTree*" igual a "10", "*nodesize*" igual a "10" e, "*importance*" igual a "*TRUE*". O primeiro estabelece o número de árvores que crescerão durante a construção do modelo, já que caso não fosse preenchido, o algoritmo tomaria como padrão o número 7. O segundo argumento toma o número 10 como tamanho mínimo dos nós da árvore gerada. O argumento "*importance*" elenca quais variáveis o modelo considerou as mais importantes na sua construção, tendo sido usado apenas para obter parâmetros de comparação.

O algoritmo baseado no método Naive Bayes, foi mais bem implementado na linguagem *R* pelo pacote "*e1071*", cujo método se chama "*naiveBayes*". Há outros pacotes que implementam o mesmo algoritmo, entretanto, o pacote utilizado apresenta melhor documentação de sua implementação e consegue atingir resultados mais precisos. Dentre todos os algoritmos esse foi o que consumiu o menor tempo de processamento na construção dos modelos.

O modelo Logit foi construído a partir do método "*multinom*", do pacote "*nnet*". A execução do algoritmo exigiu, o ajuste do argumento "*MaxNWts*", que define o maior número de pesos permitidos, para 1.000.000, já que a configuração padrão do algoritmo define um número máximo de pesos relativamente baixo, de forma a evitar que a precisão do obtida seja baixa e mais processamento seja consumido.

3.4 Sequências de Testes

Antes de iniciar a sequência de testes, foi necessário separar a base de dados em dois grandes conjuntos, um conjunto de treinamento e um conjunto de teste. Para ambos os conjuntos, foram atribuídos dois subconjuntos de dados, um contendo 500.000 e o outro 1.000.000 de registros, sendo que os registros do conjunto de treinamento diferem dos registros do conjunto de teste, dado que para o conjunto de treino, foram tomados os dados do iníciais da base e, para o conjunto de teste foram tomados os registros finais. Utilizou-se a atribuição "dados_de_treino ← dados [-x:-2282560,]" para separar o conjunto de treino e teste, sendo que "x" representa o limite inferior da seleção dos registros, podendo representar o número 500.000 ou 1.000.000.

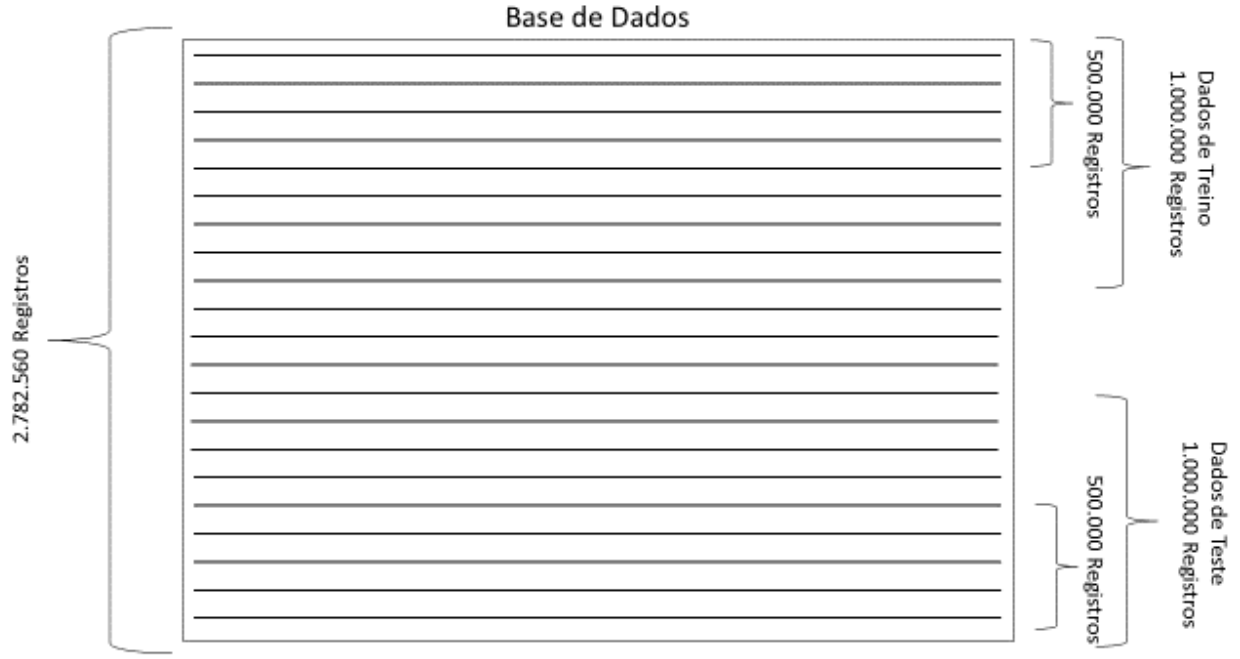

Figura 3.3 - Ilustração dos Conjuntos de Treino e Teste.

Fonte: Elaborado pelo autor.

Os modelos foram então construídos para cada algoritmo citado na seção 2.3. Incialmente foram utilizados 500.000 dados de treino e teste e, 11 variáveis, sendo desde as variáveis de índice 1 até 11 listadas na Tab. (4.2). Com o mesmo conjunto de dados, esse procedimento foi repetido para 14 e 17 variáveis sendo, desdes os índices 1 a 14 e, 1 a 17, respectivamente, finalizando a primeira sequência de testes.

A segunda sequência de testes utilizou 500.000 dados de treino e 1.000.000 de dados de teste, passando pelas mesmas quantidades de variáveis utilizadas na primeira sequência. Por fim, a

terceira sequência de teste utilizou 1.000.000 de dados de treino e 1.000.000 de dados de teste, passando também pelas mesmas variáveis que as sequencias anteriores. A Tab. (3.2), organiza todas as três sequências de testes de acordo com os parâmetros de cada uma e os separa conforme um indice, em que cada sequência possui testes A, B e C e, o índice das variáveis segue a Tab. (3.1).

Tabela 3.2 – Sequência de Testes.

Primeira sequência de testes			
Índice do teste	**Índice das variáveis**	**Dados de Treino**	**Dados de Teste**
A	1 até 11	500.000	500.000
B	1 até 14	500.000	500.000
C	1 até 17	500.000	500.000
Segunda sequência de testes			
Índice do teste	**Índice das variáveis**	**Dados de Treino**	**Dados de Teste**
A	1 até 11	500.000	1.000.000
B	1 até 14	500.000	1.000.000
C	1 até 17	500.000	1.000.000
Terceira sequência de testes			
Índice do teste	**Índice das variáveis**	**Dados de Treino**	**Dados de Teste**
A	1 até 11	1.000.000	1.000.000
B	1 até 14	1.000.000	1.000.000
C	1 até 17	1.000.000	1.000.000

Fonte: Elaborado pelo autor.

Em cada uma das sequências de experimentos foram registrados os valores de precisão, com o intuítuo de comparar a performace de cada algoritmo. Esses valores foram obtidos utilizando o pacote "*caret*" do *R Studio*, por meio do método "*confusionMatrix*", que retorna o valor de precisão que o modelo construído teve em relação aos dados de teste, usando como argumentos o objeto do modelo construído e o a coluna referente à margem de contribuição dos dados de teste.

Ao final da sequência de testes obtiveram-se vinte e sete resultados, dado que, para cada sequência de testes foram utilizados três algoritmos . Esses resultados estão dispostos no Capítulo 4, em uma série de tabelas.

4. RESULTADOS

Os resultados obtidos de todas as etapas do processo serão aqui descritos, acompanhados de alguma análise ou discussão.

4.1 Resultado da Seleção de Variáveis

A seleção das variáveis resultou em um conjunto particular, conforme visto na Tab. (3.1), pelos cálculos feitos a partir do *Microsoft Azure Machine Learning Studio* demonstrados na Fig. (3.1).

Os cálculos de correlação estão disponíveis na Tab (4.1), na qual estão relacionadas todas as variáveis a partir das abreviaturas dadas a cada uma. Como se trata de uma matriz de correlação, os dados se repetem acima e abaixo da diagonal principal, portanto, somente os dados da parte superior da diagonal principal foram disponibilizados no intuito de facilitar sua compreensão.

Conforme mencionado no capítulo anterior, foram tomadas como as candidatas mais adequadas a serem as variáveis preditoras, aquelas que possuem correlação superior ou, aproximadamente igual a 0,50 em relação à variável margem de contribuição (mc). Assim, as variáveis selecionadas encontram-se ordenadas da esquerda para a direita, de forma decrescente, na Tab. (4.1), de acordo com seu grau de correlação em relação à variável alvo.

Os valores de correlação inferiores a 0,50, foram considerados para testar o terceiro objetivo específico levantado na Introdução do trabalho, desse modo, foi possível comparar a contribuição à precisão do modelo tanto das variáveis que possuem maior correlação, quanto daquelas que apresentam menor correlação em relação à variável alvo.

É possível notar que há a ocorrência de multicolinearidade entre as variáveis preditoras, cujas correlações foram grafadas em negrito. Essa constatação poderia indicar que essas variáveis não seriam as escolhas mais adequadas para a construção dos modelos, dado que, uma elevada multicolinearidade pode causar grande variância do modelo e assim, torná-lo suscetível a variações bruscas se houver alteração nas variáveis preditoras altamente correlacionadas.

Foi investigada a natureza das informações que as variáveis em questão apresentam e, verificou-se que em alguns casos, a elevada correlação é apenas uma coincidência da amostra utilizada, como, por exemplo, as variáveis Contrato de Empréstimo Assinado (assi) com as variáveis Produtos (prod) e Tarifa Mensal (pac) dado que, cada uma dessas variáveis exprime informações diferentes entre si. A variável “assi” indica apenas se o cliente possui contratos de empréstimo, o

que a princípio, não tem relação com a quantidade de produtos que o cliente possui ou, sua tarifa mensal.

Constatou-se também que algumas variáveis possuem correlação elevada pois suas informações podem estar inseridas em outras variáveis sem, no entanto, representar o conjunto de todos os dados coletados, como as variáveis Cheque Especial Acime de 50% (chq_ac), Cheque Especial Abaixo de 50% (chq_ab), Endividamento na Instituição Financeira (end_if), Endividamento no Sistema Financeiro Nacional (end_sfn), Prestação Disponível Empréstimo (p_cdc) e, Prestação Disponível Consignado (p_con). Para o primeiro par de variáveis, é possível notar que boa parte dos clientes que utilizaram o cheque especial em uma proporção maior que 50%, também o fizeram na proporção inferior a 50%, contudo, isso não exclui o fato de que há clientes que somente utilizaram o cheque especial em uma proporção inferior a 50%.

O mesmo ocorre para o segundo e terceiro par de variáveis, na qual, quem possui endividamento no Sistema Financeiro Nacional, não necessariamente possui dívidas junto a IF e, que possuem margem de prestação para contratar empréstimo, não necessariamente possui linha de consignado disponível. Essas considerações levaram a conclusão de que a exclusão dessas variáveis poderia trazer mais prejuízo à construção dos modelos do que benefícios, levando à decisão de mantê-las da forma como estão.

Por fim, o último par que apresenta correlação elevada são as variáveis Pagamento de Cartão (pgt_car) e Pontos Adquiridos no Programa de Relacionamento (car_adq). A primeira variável indica o nível de pagamento da fatura do cartão de crédito do cliente, conforme tabela em Anexos. Essa variável está mais associada à aspectos de inadimplência e análise de crédito, pois o cliente pode efetuar o pagamento de sua fatura em diversas faixas de valor ou, não fazer qualquer pagamento. Naturalmente, quanto maior for o pagamento da fatura do cartão de crédito, maior será a quantidade de pontos que o cliente recebera no programa de relacionamento, contudo, existem diversas modalidade de cartão de crédito que a IF disponibiliza e, há modalidades de cartões que dão mais pontos do que outras modalidades. Portanto, para um mesmo valor de fatura, o cliente poderá receber mais pontos dependendo da modalidade de cartão que possua.

Geralmente, as modalidades de cartões que possuem maior pontuação são oferecidas aos clientes que possuem renda mais elevada e tenham um segmento diferenciado em relação ao atendimento levando ao encarecimento das anuidades dessas modalidades de cartões, o que eleva significativamente a margem de contribuição dos clientes. Portanto, apesar de haver uma razão

para que haja correlação dessas variáveis, ambas fornecem diferentes informações relevantes para a construção dos modelos, fazendo com que fossem mantidas para a etapa de testes.

Ao longo da matriz de correlações é possível perceber também, que há outras variáveis preditoras, além das já citadas, que apresentam correlação superior a 0,65 entre si. Em estatística, é considerada uma correlação linear positiva forte para valores a partir de 0,70, entretanto, considerando que os cálculos podem conter erros, foi considerado o valor de 0,65 a fim de abranger a maior quantidade de variáveis preditoras que poderiam fornecer estimativas com alto desvio padrão aos modelos. (RUMSEY 2016).

Tabela 4.1 – Matriz de Correlação das Variáveis Preditoras Selecionadas.

	mc	assi	end_if	prod	chq_ac	chq_ab	sms	pgt_car	dbt	salario	car_adq	p_cdc	pac	end_sfn	p_con	pts	inv	lim
mc	1	0,70	0,68	0,64	0,60	0,59	0,59	0,57	0,55	0,52	0,51	0,51	0,49	0,48	0,46	0,45	0,42	0,41
assi		1	0,49	**0,68**	0,54	0,53	0,58	0,45	0,40	0,51	0,37	0,44	**0,66**	0,38	0,38	0,34	0,27	0,39
end_if			1	0,47	0,43	0,43	0,50	0,49	0,44	0,36	0,48	0,35	0,40	0,68	0,30	0,23	0,18	0,30
prod				1	0,55	0,54	0,54	0,48	0,46	0,42	0,44	0,36	0,50	0,34	0,42	0,27	0,31	0,33
chq_ac					1	**0,91**	0,46	0,53	0,50	0,44	0,49	0,51	0,33	0,30	0,47	0,28	0,37	0,33
chq_ab						1	0,45	0,50	0,48	0,43	0,46	0,49	0,34	0,29	0,45	0,28	0,37	0,32
sms							1	0,50	0,42	0,40	0,43	0,36	0,44	0,39	0,31	0,30	0,27	0,32
pgt_car								1	0,45	0,36	**0,75**	0,39	0,29	0,35	0,34	0,24	0,28	0,35
dbt									1	0,34	0,50	0,37	0,27	0,31	0,33	0,23	0,43	0,28
salario										1	0,33	0,44	0,37	0,27	0,45	0,21	0,22	0,29
car_adq											1	0,36	0,26	0,31	0,33	0,19	0,32	0,28
p_cdc												1	0,30	0,26	**0,66**	0,25	0,30	0,56
pac													1	0,31	0,24	0,24	0,17	0,29
end_sfn														1	0,21	0,18	0,10	0,25
p_con															1	0,20	0,27	0,39
pts																1	0,46	0,23
inv																	1	0,24
lim																		1

Fonte: Elaborado pelo autor.

4.2 *Random Forest*

A partir dos primeiros resultados, observa-se que o aumento da quantidade de dados tende a aumentar a precisão do modelo, entretanto, como mostrado na Tab. (4.2), o aumento é pouco significativo considerando onze e quatorze variáveis preditoras, o que responde ao primeiro ponto dos objetivos específicos. Outra constatação importante, é a verificação de que o aumento do número de variáveis não necessariamente aumenta a precisão dos modelos, como pode ser percebido pelos resultados quando utilizadas onze e quatorze variáveis. Por fim, as variáveis que possuem a menor correlação com a variável alvo, nas três sequências de testes (teste C que contém todas as variáveis), apresentaram resultados de precisão significativamente superior em relação a onze e quatorze variáveis.

Tabela 4.2 – Resultados *Random Forest*.

Primeira Sequência de Testes (500.000 dados de treino e teste)		
Índice do Teste	**Precisão**	**Intervalo de Confiança 95%**
A (11 variáveis)	0,6429	0,6416 a 0,6442
B (14 variáveis)	0,6311	0,6298 a 0,6325
C (17 variáveis)	0,6846	0,6833 a 0,6859
Segunda Sequência de Testes (500.000 dados de treino e 1.000.000 dados de teste)		
Índice do Teste	**Precisão**	**Intervalo de Confiança 95%**
A (11 variáveis)	0,6416	0,6406 a 0,6425
B (14 variáveis)	0,6313	0,6303 a 0,6322
C (17 variáveis)	0,6783	0,6774 a 0,6792
Terceira Sequência de Testes (1.000.000 dados de treino e teste)		
Índice do Teste	**Precisão**	**Intervalo de Confiança 95%**
A (11 variáveis)	0,6438	0,6429 a 0,6447
B (14 variáveis)	0,6328	0,6319 a 0,6338
C (17 variáveis)	0,7036	0,7027 a 0,7045

Fonte: Elaborado pelo autor.

4.3 Naive Bayes

Assim como o algoritmo anterior, as hipóteses levantadas são refutadas. A precisão do algoritmo Naive Bayes mostrou-se, em termos gerais, superior à precisão do *Random Forest*, entretanto, o intervalo de confiança no NB é mais amplo do que o RF, indicando que o erro associado ao NB é maior.

Novamente, o aumento no número de variáveis não necessariamente mostrou-se eficaz para o aumento de precisão e, as variáveis que possuem menor correlação (teste tipo C contendo as variáveis que possuem menor correlação), mostraram-se importantes para atingir maiores resultados de precisão. Para esse algoritmo, o aumento na quantidade de dados de treino diminuiu a precisão dos modelos, indicando que nem sempre a quantidade de dados tornará os modelos mais precisos, por conta do sobreajuste.

Tabela 4.3 – Resultados Naive Bayes.

Primeira Sequência de Testes (500.000 dados de treino e teste)		
Índice do Teste	**Precisão**	**Intervalo de Confiança 95%**
A (11 variáveis)	0,6348	0,6335 a 0,6362
B (14 variáveis)	0,6385	0,6371 a 0,6398
C (17 variáveis)	0,7067	0,7054 a 0,7080
Segunda Sequência de Testes (500.000 dados de treino e 1.000.000 dados de teste)		
Índice do Teste	**Precisão**	**Intervalo de Confiança 95%**
A (11 variáveis)	0,6348	0,6339 a 0,6358
B (14 variáveis)	0,6445	0,6435 a 0,6454
C (17 variáveis)	0,7162	0,7154 a 0,7171
Terceira Sequência de Testes (1.000.000 dados de treino e teste)		
Índice do Teste	**Precisão**	**Intervalo de Confiança 95%**
A (11 variáveis)	0,6335	0,6325 a 0,6334
B (14 variáveis)	0,6428	0,6419 a 0,6438
C (17 variáveis)	0,7136	0,7127 a 0,7145

Fonte: Elaborado pelo autor.

4.4 Logit

O modelo Logit mostrou o primeiro conjunto de resultados que confirmam o ponto de que o aumento no número de variáveis tende a aumentar a precisão do modelo. Apenas os testes B e C da terceira sequência tiveram resultados inferiores em relação aos mesmos testes das sequências anteriores, indicando que esse modelo pode ter sofrido sobreajuste com a quantidade de dados de treino utilizada na última sequência.

Os testes C da primeira e segunda sequência, sugerem o mesmo resultado obtido com os algoritmos anteriores. As variáveis preditoras com maior correlação fornecem valores de precisão significativos, mas, as que possuem os menores valores de correlação incrementam os resultados consideravelmente.

Tabela 4.4 – Resultados Logit.

Primeira Sequência de Testes (500.000 dados de treino e teste)		
Índice do Teste	**Precisão**	**Intervalo de Confiança 95%**
A (11 variáveis)	0,6375	0,6362 a 0,6388
B (14 variáveis)	0,6483	0,6470 a 0,6496
C (17 variáveis)	0,7228	0,7215 a 0,7240
Segunda Sequência de Testes (500.000 dados de treino e 1.000.000 dados de teste)		
Índice do Teste	**Precisão**	**Intervalo de Confiança 95%**
A (11 variáveis)	0,6333	0,6324 a 0,6342
B (14 variáveis)	0,6550	0,6540 a 0,6559
C (17 variáveis)	0,7291	0,7282 a 0,7300
Terceira Sequência de Testes (1.000.000 dados de treino e teste)		
Índice do Teste	**Precisão**	**Intervalo de Confiança 95%**
A (11 variáveis)	0,6338	0,6328 a 0,6347
B (14 variáveis)	0,6523	0,6514 a 0,6533
C (17 variáveis)	0,7192	0,7183 a 0,7200

Fonte: Elaborado pelo autor.

4.5 Análise dos Resultados

O resultado geral que obteve maior precisão foi o teste C da segunda sequência de testes do modelo linear generalizado Logit, com 72,91% de precisão. Coincidentemente, o algoritmo Naive Bayes teve nesse teste a sua maior precisão e ficando em segundo lugar em resultado geral com 71,62%, considerando o teste C da segunda sequência de testes. Nesse ponto ambos os modelos apresentaram precisão superior a 70%, valor mínimo considerado para fins de resultados satisfatórios, porém, deve-se ressaltar que o NB, por considerar todas as variáveis independentes, consumiu um tempo menor de processamento em relação ao Logit. A partir do ponto de vista da eficiência de tempo e processamento, o NB é um algoritmo bastante interessante para análises preliminares em que se tem pouca quantidade de dados.

Outro comportamento interessante do Logit em relação aos demais algoritmos pode ser observado ao incrementarmos a quantidade de variáveis para construção do modelo. O Logit manteve um padrão crescente de resultados em todos os testes. Nesse sentido, pode se inferir que para esse algoritmo, o incremento na quantidade de variáveis oferece maior contribuição à precisão do algoritmo do que o aumento na quantidade de dados de treino.

O algoritmo *Random Forest* obteve, em sua maioria, resultados abaixo de 70%, exceto no teste C da terceira sequência de testes. Esperava-se pela quantidade de dados, resultados mais promissores para o referido algoritmo, entretanto, como esse apresenta bastante robustez a sobreajuste, entende-se que a quantidade de dados de treino pode ter insuficiente para a construção de modelos mais precisos.

Um ponto importante para todos os algoritmos ocorre nos testes de índice C, em todas as sequências de testes. Obteve-se resultados de precisão mais elevados em relação a variação do teste A para o teste B. Para o RF, os resultados da primeira sequência de testes tiveram variação percentual média de 4,76% entre os testes A para C e B para C. A segunda sequência de testes teve variação média percentual 4,14% e, a terceira sequência teve variação de 6,38% para os mesmos testes. Já o NB obteve variações de 7,00%, 7,65% e, 7,55% respectivamente para a primeira, segunda e terceira sequencias de testes, entre os testes de índice A e B em relação ao teste C. Por fim, o Logit teve as variações percentuais médias de 7,99%, 8,50% e, 7,57% para as mesmas sequências de testes.

Esses resultados indicam que as variáveis Pontos do Programa de Relacionamento (pts), Investimentos (inv) e, Limite Disponível no Cartão de Crédito (lim), apesar de possuírem as menores correlações com a variável alvo, fornecem grande contribuição para a construção de

modelos preditivos mais precisos. Um dos possíveis motivos para esse fenômeno é que essas variáveis quase não apresentam multicolinearidade com as demais variáveis, ocasionando baixo erro associado a elas, assim a contribuição que essas variáveis fornecem se torna superior seu erro.

A hipótese de que a quantidade de dados utilizada para construir os modelos influência na sua precisão foi parcialmente respondida, entretanto, analisando os resultados de todos os modelos, percebe-se que tal hipótese nem sempre é verdadeira. Comparando a segunda e terceira sequência de testes dos algoritmos Naive Bayes, Logit, há uma pequena queda de precisão, não ocorrendo o mesmo comportamento para o *Random Forest*. Isso indica que a quantidade de dados de treino deve ser adequada para cada algoritmo a fim de obter resultados mais precisos de acordo com cada método estatístico associado.

Outro ponto de interesse é a variação na quantidade de dados de teste entre a primeira e segunda sequências. Esperava-se que haveria uma queda de precisão em todos os modelos dado que a quantidade de dados de teste era o dobro da quantidade de dados de treino. Entretanto, observa-se que esse fenômeno não ocorreu para o modelo Logit e Naive Bayes, e para um caso no modelo *Random Forest*. Todos os testes da segunda sequência foram mais precisos do que os testes da primeira sequência para os dois primeiros algoritmos e, para o RF, o teste B da segunda sequência foi mais preciso do que o mesmo teste da primeira sequência. Desse modo, não há indício de sobreajuste na primeira e segunda sequências de testes nos referidos modelos, dado que os modelos conseguem manter os mesmos padrões de precisão em diferentes conjuntos de testes.

A quantidade de variáveis também nem sempre fornecerá resultados melhores à medida que forem sendo adicionadas aos modelos, conforme já mencionado, entretanto, deve-se ressaltar que os testes de índice C tiveram resultados mais precisos para três dos quatros modelos, indicando que para a maioria dos casos, o aumento na quantidade de variáveis trará um resultado de precisão mais elevado.

5. CONCLUSÃO

O objetivo principal deste trabalho foi comparar os modelos preditivos a partir de seus resultados e características de seus algoritmos. Foram discutidos e apresentados os principais conteúdos teóricos e práticos dos algoritmos utilizados bem como, os principais conceitos para entendimento do tema e contextualização e aplicações do campo de estudo da Inteligência Artificial.

O processo utilizado no trabalho foi documentado a fim de permitir que os objetivos específicos levantados fossem respondidos com satisfatório embasamento dos resultados, assim, cada etapa do método utilizado teve sua importância para que os resultados fossem atingidos.

Os resultados obtidos colocam o modelo linear generalizado Logit como o melhor modelo para a previsão da base de dados utilizada, dado que se trata de um modelo geral, e por isso, dentre os modelos utilizados, é o que mais se adequa aos dados fornecidos pois constrói relações lineares entre as variáveis. Por um lado, a construção de relações lineares é um método simples e que garante resultados de previsão acurados, entretanto, podem existir outros modelos não-lineares que forneçam resultados mais precisos, porém, para um conjunto específico de dados tratados de forma específica, e não de forma genérica como ocorrido neste trabalho.

O algoritmo Naive Bayes foi o segundo melhor em termos de precisão, apresentando três resultados com precisão acima de 0,7. Verifica-se nos modelos Logit e Naive Bayes há um indício de sobreajuste da segunda para a terceira sequência de testes, assim, a quantidade de dados adequada para treinar esses modelos é inferior a 1.000.000 de dados de treino. Portanto, para se ter resultados mais precisos nesses modelos, seria necessário utilizar mais de 17 variáveis a fim de evitar que os modelos preditivos fiquem enviesados.

A comparação dos modelos preditivos propostas foi finalizada atendendo ao objetivo geral deste trabalho, dessa forma foi possível comparar os algoritmos de aprendizagem de máquina a partir dos parâmetros propostos.

O primeiro objetivo específico foi demonstrado de forma geral para todos os algoritmos, ou seja, o aumento no número de variáveis aumenta a precisão do modelo, ainda que haja modelos mais sensíveis a esse aumento do que outros. Em seguida o segundo objetivo específico também se mostrou verdadeiro para todos os algoritmos, o aumento da quantidade de dados aumenta a precisão dos modelos. Por fim, o último objetivo específico não foi verificado em nenhum dos modelos, assim, as variáveis com maior correlação em relação à variável alvo, não necessariamente fornecerão a maior contribuição em termos de precisão para os modelos.

As hipóteses consideradas puderam ser verificadas e, apenas uma pode ser constatada como verdadeira, ou seja, a precisão dos modelos será maior, quanto maior for a quantidade de dados de treino, dado que para todos os algoritmos, houve melhora nos resultados de precisão utilizando 1.000.000 de dados de treino. As demais hipóteses não se mostraram necessariamente verdadeiras, assim foram refutadas, considerando que, não há incremento proporcional de precisão à medida que as variáveis são adicionadas e, a precisão dos modelos não diminuiu em dois dos três algoritmos quando a quantidade de dados de teste foi superior a quantidade de dados de treino.

Dentre as limitações encontradas nesse trabalho, há a carência de uma análise estatística preliminar mais profunda das variáveis e, a utilização de outros métodos mais modernos de seleção de variáveis como o *cross validation* e *k nearest number*. A linguagem utilizada também foi outro fator delimitador, dado que seria necessário utilizar um software livre em decorrência do custo envolvido com softwares mais consolidados no mercado com o SAS, SPSS e STATA. O *R Studio* é um excelente software para análise estatística, construção de gráficos e para aprendizagem de máquina, entretanto, sua configuração armazena todos os dados analisados em memória, isso exige um elevado armazenamento de memória RAM e processamento, sendo necessário utilizar um computador bastante robusto para que os testes fossem executados.

Como sugestão de trabalhos futuros, pode se analisar qual seria o melhor método de seleção de variáveis para um determinado algoritmo, ou mesmo analisar a quantidade de dados que cada algoritmo suporta até que ocorra o sobreajuste. Também há a possibilidade de estudar outros métodos de tratamento de dados que possibilitariam obter resultados melhores com os algoritmos utilizados e, comparar os mesmos algoritmos com outras métricas de avaliação como *F-Score*, ou curva AUC.

REFERÊNCIAS BIBLIOGRÁFICAS

ABREU, Mery; SIQUEIRA, Arminda; CAIAFFA, Waleska. **Regressão logística ordinal em estudos epidemiológicos**. Belo Horizonte: Revista Saúde Pública, nº 43, p. 184. 2009.

ACCENTURE. **BANK TECHNOLOGY VISION 2017**. Accenture, p. 1-25. 2017. Disponível em: <https://www.accenture.com/t20170322T205838Z__w__/us-en/_acnmedia/PDF-47/Accenture-Banking-Technology-Vision2017.pdf#zoom=50>. Acesso em: 03/09/2017.

BARANAUSKAS, José Augusto. **Aprendizado de Máquina Conceitos e Definições**. Ribeirão Preto: Departamento de Física e Matemática – FFCLRP – USP, 2007. Disponível em: <dcm.ffclrp.usp.br/~augusto/teaching/ami/AM-I-Conceitos-Definicoes.pdf>. Acesso em: 26/08/2017.

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.

BLANCO, Sylvie; LESCA, Humbert; CARSON-FASAN, Marie-Laurence. **Developing Capabilities to Create Collective Intelligence within Organizations**. *Journal of Competitive Intelligence and Management*, *volume* 1, *number* 1, p. 81 – 84. 2003.

BRUNI, Adriano Leal; FAMÁ, Rubens. **Gestão de Custos e Formação de Preços**: Com Aplicações na Calculadora HP 12C. São Paulo: Atlas, ed. 5ª, v. 2, p. 177. 2009.

CARVALHO, Carlos. ***Classification****: Logistic Regression and Naive Bayes*. *Austin*: *The University of Texas McCombs School of Business*, p. 1 – 71. 2017. Disponível em: <https://faculty.mccombs.utexas.edu/carlos.carvalho/teaching/Sec3_Classification.pdf>. Acesso em: 24/10/2017.

COBRA, Marcos. **Administração de Marketing no Brasil**. Rio de Janeiro: Elsevier, 3ª ed, p. 200 – 210. 2009.

CRIBARI, Francisco. **Método Boostrap**. 19 de abril de 2010. Disponível em: <http://www.ebah.com.br/content/ABAAAAr8cAA/metodo-bootstrap>. Acesso em: 23/10/2017.

DUTRA, Rogério Garcia. **Aplicação de Métodos de Inteligência Artificial em Inteligência de Negócios**. Porto Alegre: Enegep 2005, p. 4956 – 4960, 29 de outubro de 2005.

FACELLI, Katti. et al. **Inteligência Artificial**: Uma Abordagem de Aprendizado de Máquina. Rio de Janeiro: LTC, 1ª ed, p. 2 – 32. 2011.

GIOVANOLLI, Regina. **O IMPACTO DA INTELIGÊNCIA ARTIFICIAL NA INDÚSTRIA FINANCEIRA**. Provider IT & Business Solutions. Julho, 2017. Disponível em:

<http://provider-it.com.br/consultoria-de-ti/o-impacto-da-inteligencia-artificial-na-industria-financeira/>. Acesso em: 26/08/2017.

HASTIE, Trevor; TIBSHIRANI, Robert; FRIEDMAN, Jerome. **The Elements of Statistical Learning**. *New York – NY*: *Springer*, p. 587 – 588. 2009.

HILLIER, Frederick; LIEBERMAN Gerald. **Introdução à Pesquisa Operacional**. Traduzido por Ariovaldo Griesi. Porto Alegre: AMGH, 9ª ed, p. 657 – 659. 2013.

KNOW SOLUTIONS. **O que é Business Intelligence (BI)**. Disponível em: < http://knowsolution.com.br/o-que-e-business-intelligence-bi/>. Acesso em: 20/10/2017.

LEE, Kai-Fu. ***The Real Threat of Artificial Intelligence***. The New York Times: Pequim, 24 de junho de 2017. Disponível em: <https://www.nytimes.com/2017/06/24/opinion/sunday/artificial-intelligence-economic-inequality.html>. Acesso em: 26/08/2017.

LESCA, Humbert. ***Veille stratégique pour le management stratégique: Etat de la question et axes de recherche***. *In Economies et Societé, Série Sciences de Gestion, SG*, nº. 20:5, p. 31-35. 1994.

MÜLLER, Cláudio José; KRIGER, Joel Szmelsztayn. **Gestão de Custos em Empresas de Distribuição**. São Paulo: Revista da FAE – USP, p. 28 – 31. 2002.

NETO, Joaquim. **Inferência Bayesiana**. Juiz de Fora: Departamento de Estatística – ICE, p. 22. Disponível em: < http://www.ufjf.br/joaquim_neto /files/2009/09/IB-Slides-v1.1.pdf>. Acesso em: 24/10/2010.

OLIVEIRA, André Rodrigues. **Comparação de algoritmos de aprendizagem de máquina para construção de modelos preditivos de diabetes não diagnosticado**. Porto Alegre: UFRGS, Instituto de Informática, p. 15 – 70. 2016.

PORTER, Michael. **Estratégia competitiva: técnica para análise de indústrias e da concorrência**. Tradução Elizabeth Maria de Pinho Braga. Rio de Janeiro: Elsevier, 2ª ed, p. xii – xiii. 2004.

PRIMAK, Fábio Vinícius. **DECISÕES COM B.I. (BUSINESS INTELLIGENCE)**. Ciência Moderna, 1ª ed., p. 1 – 5. 2008.

PUC – MG. **Inteligência Artificial**: Uma breve introdução. 12 de abril de 2012. Disponível em: <http://icpucminas.blogspot.com.br/2012/04/inteligenciaartificial-uma-breve.html>. Acesso em: 21/10/2017.

PURDY, Mark; DAUGHERTY, Paul. ***WHY ARTIFICIAL INTELLIGENCE IS THE FUTURE OF GROWTH***. Accenture, p. 3 – 10. 2016. Disponível em: <https://www.accenture.com/lv-en/_acnmedia/PDF-33/Accenture-Why-AI-is-the-Future-of-Growth.pdf>. Acesso em: 26/08/2017.

PETERNELLI, Luiz Alexandre. **CAPÍTULO 7 – Intervalos de confiança**. Viçosa: Departamento de Informática – UFV, INF 162, p. 88. 2009.

RODRÍGUEZ, German. ***Generalized Linear Models****: The Multinomial Logit Model*. *Princeton*: *Princeton University*. 2017. Disponível em: <http://data.princeton.edu/wws509/notes/c6s2.html>. Acesso em: 28/10/2017.

RUMSEY, Deborah. **HOW TO INTERPRET A CORRELATION COEFFICIENT *R***. Junho de 2016. Disponível em: <http://www.dummies.com/education/math/statistics/how-to-interpret-a-correlation-coefficient-r/>. Acesso em: 01/12/2017.

RUSSEL, Stuart; NORVIG, Peter. **Inteligência Artificial**. Tradução Regina Célia Simille. Rio de Janeiro: Elsevier, 3ª ed, p. 28 – 29. 2013.

SABERWHAL, Rajiv; BECERRA-FERNANDEZ, Irma. ***Business Intelligence: Practices, Technologies, and Management***. Estados Unidos: Wiley, 2ª ed, p. 5 – 13. 2011.

TAURION, Cezar. **Aprendizado de máquina começa a entrar no radar do mundo corporativo**. CIO COMPUTERWORLD. Disponível em: <http://cio.com.br/opiniao/2015/07/27/aprendizado-de-maquina-comeca-a-entrar-no-radar-do-mundo-corporativo/>. Acesso em: 03/09/2017.

VARGAS, Vera do Carmo Comparsi de. **CORRELAÇÃO**. Santa Catarina: Centro Tecnológico da Universidade Federal de Santa Catarina – UFSC. 2012. Disponível em: <http://www.inf.ufsc.br/~vera.carmo/Correlacao/Correlacao_Pearson_Spearman_Kendall.pdf>. Acesso em: 04/11/2017.

WEEKS, Melvin. ***THE MULTINOMIAL PROBIT MODEL REVISITED:*** *A DISCUSSION OF PARAMETER ESTIMABILITY, IDENTIFICATION AND SPECIFICATION TESTING*. *Journal of Economic Surveys*, p. 300-303. 1997.

WERNKE, Rodney; LEMBECK, Marluce. **Análise de rentabilidade dos segmentos de mercado de empresa distribuidora de mercadorias**. São Paulo: Revista Contabilidade & Finanças, vol. 15, n. 35, p. 69 – 79. 2004.

ZUBEN, Fernando Von; ATTUX, Romis. **Árvore de Decisão**. Unicamp. Disponível em: ftp://ftp.dca.fee.unicamp.br/pub/docs/vonzuben/ia004_1s10/notas_de_aula/topico7_IA004_1s10.pdf>. Acesso em: 23/10/2017.

ANEXOS

Tabela I - Margem de Contribuição.

Categoria	Faixa de valores (R$)	Quantidade de registros
0	Sem informação	1.725.442
1	Valor negativo	72.231
2	Valor igual a zero	192.777
3	0,01 a 4,99	290.636
4	5,00 a 9,99	44.446
5	10,00 a 15,99	47.673
6	15,00 a 19,99	38.994
7	20,00 a 27,99	43.637
8	28,00 a 44,99	66.871
9	45,00 a 49,99	15.702
10	50,00 a 54,99	13.164
11	55,00 a 59,99	11.703
12	60,00 a 67,99	16.298
13	68,00 a 99,99	45.049
14	100,00 a 130,99	27.571
15	131,00 a 149,99	12.740
16	150,00 a 179,99	16.439
17	180,00 a 199,99	8.731
18	200,00 a 221,99	8.183
19	222,00 a 299,99	20.386
20	300,00 a 399,99	15.933
21	Maior ou igual a 400,00	48.604

Fonte: Elaborado pelo autor.

Tabela II – Possui Empréstimo.

Categoria	Correspondência	Quantidade de registros
1	Possui	593.814
2	Não possui	2.188.746

Fonte: Elaborado pelo autor.

Tabela III – Endividamento IF.

Categoria	Faixa de valores (R$)	Quantidade de registros
0	Sem informação	2.279.336
1	Não possui	140.974
2	0,01 a 500,00	70.877
3	500,01 a 1.000,00	47.964
4	1.000,01 a 3.000,00	79.053
5	3.000,01 a 5.000,00	32.027
6	5.000,01 a 10.000,00	38.901
7	10.000,01 a 25.000,00	38.435
8	25.000,01 a 50.000,00	21.253
9	50.000,01 a 75.000,00	10.563
10	75.000,01 a 100.000,00	7.981
11	100.000,01 a 150.000,00	7.075
12	150.000,01 a 200.000,00	2.677
13	Acima de 200.000,01	5.444

Fonte: Elaborado pelo autor.

Tabela IV – Produtos.

Categoria	Correspondência	Quantidade de registros
0	Sem informação	2.229.590
1	1 Produto	73.632
2	2 Produtos	84.530
3	3 Produtos	67.052
4	4 Produtos	51.766
5	5 Produtos	38.199
6	6 Produtos	27.925
7	7 Produtos	19.503
8	8 Produtos	13.457
9	9 Produtos	9.203
10	Mais de 9 produtos	18.016
11	0 Produto	149.687

Fonte: Elaborado pelo autor.

Tabela V – Uso Cheque Especial Acima de 50%.

Categoria	Correspondência	Quantidade de registros
0	Sem informação	2.545.684
1	3 Últimos meses	17.567
2	2 Últimos meses	5.644
3	Último mês	7.832
4	Penúltimo mês	7.948
5	Antepenúltimo mês	5.450
6	Não usou	192.435

Fonte: Elaborado pelo autor.

Tabela VI – Uso Cheque Especial Abaixo de 50%.

Categoria	Correspondência	Quantidade de registros
0	Sem informação	2.545.684
1	3 Últimos meses	17.110
2	2 Últimos meses	7.611
3	Último mês	11.734
4	Penúltimo mês	13.299
5	Antepenúltimo mês	8.496
6	Não usou	178.626

Fonte: Elaborado pelo autor.

Tabela VII – Possui Serviço de SMS.

Categoria	Correspondência	Quantidade de registros
1	Não possui	2.398.230
2	Possui	384.330

Fonte: Elaborado pelo autor.

Tabela VIII – Pagamento Fatura do Cartão de Crédito.

Categoria	Faixa de valores	Quantidade de registros
0	Sem informação	2.594.951
1	Pagou abaixo do valor mínimo	6.122
2	Pagou valor mínimo	1.688
3	Pagou acima do valor mínimo e abaixo do valor total	15.148
4	Pagou valor total da fatura	164.651

Fonte: Elaborado pelo autor.

Tabela IX – Débito Automático.

Categoria	Faixa de valores	Quantidade de registros
0	Sem informação	2.499.900
1	Possui até 2	175.717
2	Possui de 3 a 5	66.875
3	Possui de 6 a 8	23.697
4	Possui de 9 a 11	9.926
5	Possui de 12 a 15	4.754
6	Possui mais de 15	1.691

Fonte: Elaborado pelo autor.

Tabela X – Salário.

Categoria	Correspondência	Quantidade de registros
1	Sim	248.877
2	Não	2.533.683

Fonte: Elaborado pelo autor.

Tabela XI – Pontos Adquiridos no Programa de Relacionamento.

Categoria	Faixa de valores	Quantidade de registros
0	Sem informação	2.626.407
1	1 a 499	111.608
2	500 a 999	21.557
3	1.000 a 2.499	17.466
4	2.500 a 4.999	3.189
5	5.000 a 9.999	1.867
6	10.000 a 19.999	403
7	20.000 a 49.999	55
8	50.000 a 99.999	7
9	Acima de 100.000	1

Fonte: Elaborado pelo autor.

Tabela XII – Prestação Disponível para Empréstimo Direto ao Consumidor.

Categoria	Faixa de valores (R$)	Quantidade de registros
0	Sem informação	1.885.734
1	Igual a 0,00	336.143
2	0,01 a 10,00	10.215
3	10,01 a 30,00	13.646
4	30,01 a 50,00	23.691
5	50,01 a 100,00	141.571
6	100,01 a 150,00	70.245
7	150,01 a 300,00	75.454
8	300,01 a 500,00	58.612
9	500,01 a 1000,00	47.181
10	Maior que 1000,00	93.375
11	Negativo	26.693

Fonte: Elaborado pelo autor.

Tabela XIII – Tarifa Mensal.

Categoria	Correspondência	Quantidade de registros
0	Sem informação	1.934.472
1	Modalidade 10	53.179
2	Modalidade 20	49.613
3	Modalidade 30	319
4	Modalidade 40	30.581
5	Simples	24.344
6	Modalidade 50	8.233
8	Pacote PF Resolução 3402	26.239
11	Modalidade 18	53
12	Padronizado	115.407
13	Conta jovem	7.940
14	Conta universitária	38.190
16	Bônus 10	391
17	Bônus 15	2.496
18	Bônus 25	1.582
19	Eletrônica	145
20	Simplificada	110
21	Fácil	541
23	Cesta 1	69
24	Cesta 3	3
26	On line	2.725
27	Plus	3.111
28	Classic	3.493
29	Personalizado	205
30	Livre opção bancária	1
31	Econômico	24.574
32	Modalidade 3	11.144
34	Resolução 3104	23.386
35	Especial	17.630
38	Outros pacotes	7.017
39	Sem pacote	247.314
44	Completo	32.961
45	Universtário (banco incorporado I)	424
46	Econômico (banco incorporado I)	3.760
47	Especial (banco incorporado I)	2.161
48	Completo (banco incorporado I)	1.828
49	Pleno (banco incorporado I)	741
50	Conta Digital	3.745
51	Postal	2.803

52	Banco postal	13.998
53	Mais econômico	5.797
55	Total	8.538
61	INSS (banco incorporado II)	230
62	Econômico (banco incorporado II)	881
63	Especial (banco incorporado II)	31.421
64	Completo (banco incorporado II)	28.880
65	Pleno (banco incorporado II)	9.885

Fonte: Elaborado pelo autor.

Tabela XIV – Endividamento SFN.

Categoria	Faixa de valores (R$)	Quantidade de registros
0	Sem informação	2.279.336
1	0,00	197.661
2	0,01 a 500,00	24.789
3	500,01 a 1.000,00	30.532
4	1.000,01 a 3.000,00	64.719
5	3.000,01 a 5.000,00	31.554
6	5.000,01 a 10.000,00	42.266
7	10.000,01 a 25.000,00	49.897
8	25.000,01 a 50.000,00	25.770
9	50.000,01 a 75.000,00	11.551
10	75.000,01 a 100.000,00	7.460
11	100.000,01 a 150.000,00	7.547
12	150.000,01 a 200.000,00	3.326
13	Acima de 200.000,01	6.162

Fonte: Elaborado pelo autor.

Tabela XV – Prestação Disponível para Empréstimo Consignado.

Categoria	Faixa de valores (R$)	Quantidade de registros
0	Sem informação	2.406.213
1	Igual a 0,00	84.546
2	0,01 a 10,00	1.462

3	10,01 a 30,00	1.741
4	30,01 a 50,00	2.296
5	50,01 a 100,00	5.895
6	100,01 a 150,00	6.398
7	150,01 a 300,00	33.500
8	300,01 a 500,00	79.637
9	500,01 a 1000,00	58.156
10	Maior que 1000,00	93.339
11	Negativo	9.377

Fonte: Elaborado pelo autor.

Tabela XVI – Pontos no Programa de Relacionamento.

Categoria	Faixa de valores	Quantidade de registros
0	Sem informação	2.132.522
1	1 a 40 pontos	13.707
2	41 a 50 pontos	1.032
3	51 a 60 pontos	749
4	61 a 70 pontos	636
5	71 a 80 pontos	515
6	81 a 90 pontos	402
7	91 a 100 pontos	318
8	101 a 150 pontos	1.013
9	151 a 200 pontos	512
10	201 a 250 pontos	308
11	251 a 300 pontos	200
12	301 a 350 pontos	133
13	351 a 400 pontos	102
14	401 a 450 pontos	63
15	451 a 500 pontos	48
16	Acima de 500	320
17	Restrição para participação no programa	629.980

Fonte: Elaborado pelo autor.

Tabela XVII – Investimento.

Categoria	Faixa de valores (R$)	Quantidade de registros
1	0,00	2.114.364
3	0,01 a 5.000,00	525.100
5	5.000,01 a 10.000,00	34.079
7	10.000,01 a 15.000,00	20.155
9	15.000,01 a 20.000,00	13.269
11	20.000,01 a 30.000,00	17.507
13	30.000,01 a 40.000,00	11.027
15	40.000,01 a 60.000,00	13.804
17	60.000,01 a 80.000,00	7.792
19	80.000,01 a 100.000,00	4.915
21	100.000,01 a 150.000,00	7.203
23	150.000,01 a 200.000,00	3.588
25	Acima de 200.000,00	9.757

Fonte: Elaborado pelo autor.

Tabela XVIII – Limite Disponível no Cartão de Crédito.

Categoria	Faixa de valores (R$)	Quantidade de registros
0	Sem informação	1.863.251
1	Igual a 0,00	240.722
2	0,01 a 50,00	14.228
3	50,01 a 200,00	35.288
4	200,01 a 600,00	111.733
5	600,01 a 800,00	59.539
6	800,01 a 1000,00	45.562
7	1000,01 a 2000,00	99.506
8	2000,01 a 3000,00	45.562
9	3000,01 a 5000,00	38.036
10	5000,01 a 7000,00	17.943
11	7000,01 a 10000,00	13.665
12	Maior que 10000,00	17.530
13	Negativo	173.520

Fonte: Elaborado pelo autor.