CENTRO UNIVERSITÁRIO CHRISTUS
SISTEMAS DE INFORMAÇÃO

THALYS MELICIO DA COSTA SILVA

UM ESTUDO COMPARATIVO ENTRE ALGORITMOS DE APRENDIZAGEM DE MÁQUINA SUPERVISIONADOS PARA PREDIÇÃO DE SOLUÇÃO DE RECLAMAÇÕES NO PROCON

FORTALEZA

2021

THALYS MELICIO DA COSTA SILVA

UM ESTUDO COMPARATIVO ENTRE ALGORITMOS DE APRENDIZAGEM DE MÁQUINA SUPERVISIONADOS PARA PREDIÇÃO DE SOLUÇÃO DE RECLAMAÇÕES NO PROCON

Trabalho de Conclusão de Curso (TCC) apresentado ao Curso de Sistemas de Informação do Centro Universitário Christus, como requisito parcial para obtenção do grau de bacharel em Sistemas de Informação.

Orientador: Prof. MSc. Felipe Timbó Brito

FORTALEZA

2021

Dados Internacionais de Catalogação na Publicação
Centro Universitário Christus - Unichristus
Gerada automaticamente pelo Sistema de Elaboração de Ficha Catalográfica do
Centro Universitário Christus - Unichristus, com dados fornecidos pelo(a) autor(a)

S586e Silva, Thalys Melicio da Costa.
Um Estudo Comparativo Entre Algoritmos de Aprendizagem de Máquina Supervisionados para Predição de Solução de Reclamações no Procon / Thalys Melicio da Costa Silva. - 2021.
49 f. : il. color.

Trabalho de Conclusão de Curso (Graduação) - Centro Universitário Christus - Unichristus, Curso de Sistemas de Informação, Fortaleza, 2021.
Orientação: Prof. Me. Felipe Timbó Brito.

1. Aprendizado de Máquina. 2. Aprendizagem Supervisionada. 3. Predição de Solução. 4. PROCON. I. Título.

CDD 005

THALYS MELICIO DA COSTA SILVA

UM ESTUDO COMPARATIVO ENTRE ALGORITMOS DE APRENDIZAGEM DE MÁQUINA SUPERVISIONADOS PARA PREDIÇÃO DE SOLUÇÃO DE RECLAMAÇÕES NO PROCON

Trabalho de Conclusão de Curso (TCC) apresentado ao curso de Sistemas de Informação do Centro Universitário Christus, como requisito parcial para obtenção do grau de bacharel em Sistemas de Informação.

Aprovada em:_____/_____/ _____

BANCA EXAMINADORA

_____

Prof. Ms. Felipe Timbó Brito (Orientador)
Centro Universitário Christus – Unichristus

_____

Prof. Dr. Daniel Nascimento Teixeira
Centro Universitário Christus – Unichristus

_____

Prof. Ms. David Kenned Ferreira Andrade Viana
Centro Universitário Christus – Unichristus

AGRADECIMENTOS

Primeiramente, agradeço a Deus por ter permitido a conquista de um sonho tão importante para mim e minha família.

À minha futura esposa pelo seu companheirismo e apoio em todos os momentos difíceis dessa jornada, o que sempre me motivou a não desistir mesmo achando impossível.

Ao Prof. Me Felipe Timbó Brito pela confiança, orientação objetiva, parceria e incentivo na realização deste trabalho. Ao Prof. Dr. Daniel Nascimento Teixeira, por me iniciar na pesquisa de iniciação científica. Obrigado pelos ensinamentos!

Aos meus pais e familiares, que sempre me incentivaram a trilhar o caminho dos estudos de forma humilde e perseverante.

Por fim, agradeço profundamente a todos os professores que nunca desistiram de mim como aluno, sempre me motivando a aprender cada vez mais e me desafiando a ser um profissional e cidadão melhor a cada dia.

RESUMO

A utilização de algoritmos de aprendizado de máquina já é uma realidade nas empresas e indústrias. Sua utilização visa melhorar os processos, otimizar suas estratégias e, consequentemente, maximizar os lucros. Essa realidade já é comum em várias empresas ao redor do mundo, como Netflix, Facebook, Amazon, Walmart, entre outras. No Brasil, vários órgãos, empresas e instituições ainda carecem da utilização de algoritmos de aprendizado de máquina para alavancarem seus resultados, como é o caso do PROCON - Programa de Proteção e Defesa do Consumidor. Este órgão está presente em todos os estados brasileiros e tem como finalidade tratar conflitos entre consumidores e empresas que vendem produtos ou prestam serviços. A fim de prever para novas entradas de reclamações, se elas serão solucionadas ou não, este trabalho tem como objetivo realizar um estudo comparativo entre seis algoritmos de aprendizado de máquina sobre dados de reclamações, realizadas por meio do PROCON. Foi utilizada uma metodologia exploratória e descritiva, conjuntamente à metodologia CRISP-DM, com o intuito de auxiliar na compreensão do problema, interpretação e preparação dos dados, modelagem, avaliação dos resultados e disponibilização da solução. Resultados demonstram que, para algumas técnicas, é possível obter mais de 70% de acurácia na predição de atendimento das reclamações, isto é, se a reclamação foi atendida ou não. Dentre as técnicas aplicadas, o algoritmo *Random Forest* foi o que obteve os melhores resultados, chegando a 73,34% de acurácia.

Palavras-chave: Aprendizado de Máquina. Aprendizagem Supervisionada. Predição de Solução. PROCON.

ABSTRACT

The use of machine learning algorithms is already a reality in companies and industries. Its use aims to improve processes, optimize its strategies and, consequently, maximize profits. This reality is already common in several companies around the world, such as Netflix, Facebook, Amazon, Walmart, among others. In Brazil, several agencies, companies and institutions still lack the use of machine learning algorithms to leverage their results, as PROCON - Programa de Proteção e Defesa do Consumidor. This agency is present in all Brazilian states and aims to deal with conflicts between consumers and companies that sell products or provide services. In order to predict for new entries, if they will be solved or not, this work aims to conduct a comparative study between six machine learning algorithms on data of complaints made through PROCON. An exploratory and descriptive methodology was used, together with the CRISP-DM methodology, in order to assist in understanding the problem, interpreting and preparing the data, modeling, evaluating the results and publishing the solution. Results demonstrate that, for some techniques, it is possible to obtain more than 70 % accuracy in the prediction of complaints' attendance, that is, if the complaint was answered or not. Among the applied techniques, the textit Random Forest algorithm was the one that obtained the best results, reaching 73.34 % of accuracy.

Keywords: Machine Learning. Supervised Learning. Solution Prediction. PROCON.

LISTA DE FIGURAS

LISTA DE TABELAS

LISTA DE ABREVIATURAS E SIGLAS

API	*Application Programming Interface*
CDC	*Código de Defesa do Consumidor*
CRISP-DM	*Cross Industry Standard Process for Data Mining*
CSV	*Comma Separated Values*
EQM	*Erro Quadrático Médio*
GBDT	*Gradient Boosted Decision Trees*
IDE	*Integrated Development Environment*
KNN	*K-nearest neighbors*
OCR	*Reconhecimento óptico de caracteres*
PROCON	*Programa de Proteção e Defesa do Consumidor*
REQM	*Raiz Quadrada do Erro Quadrático Médio*
SAC	*Serviço de Atendimento ao Cliente*
SINDEC	*Sistema Nacional de Informações de Defesa do Consumidor*
SNDC	*Sistema Nacional de Defesa do Consumidor*
SPC	*Proteção ao Crédito*

SUMÁRIO

1 INTRODUÇÃO

Não é de agora que os temas envolvendo Inteligência Artificial são frutos de discussão sobre o futuro da humanidade. Essa área da computação vem constantemente evoluindo com a criação de novos algoritmos e múltiplas aplicações, desde sistemas financeiros, cujo impacto pode ser global, a uma simples aplicação de entretenimento para seus usuários.

Uma das áreas da Inteligência Artificial é o Aprendizado de Máquina. Essa área consiste em um método que permite que o computador tome decisões através de software, seguindo a análise de dados baseada em determinadas regras e algoritmos. Em outras palavras, o Aprendizado de Máquina é uma forma automatizada dos softwares encontrarem informações relevantes em meio a uma grande quantidade de dados. Conforme são alimentados com mais dados, esses programas são capazes de aprender a reconhecer padrões e fornecer *insights* relevantes para as organizações. É nesse ponto que está a importância do Aprendizado de Máquina para as empresas (CENTRALSERVER, 2020).

À medida que as ferramentas de Aprendizado de Máquina coletam dados de dentro e de fora de uma organização, elas conseguem adquirir conhecimentos estratégicos para o negócio. A análise do funcionamento de processos internos, de acordo com o desempenho dos colaboradores, ou a verificação da satisfação dos clientes, segundo comentários nas redes sociais, são alguns exemplos (BATISTA, 2017).

Um dos conceitos utilizados no Aprendizado de Máquina é a aprendizagem supervisionada. O programa recebe um conjunto de dados para ser treinado e consegue entender como eles se relacionam. Assim, quando o software precisar analisar novos dados, ele saberá o tipo de problema que precisará resolver e será capaz de chegar a conclusões por conta própria. Trabalhar com Aprendizado de Máquina só é possível se existirem dados suficientes para que os algoritmos sejam capazes de gerar um classificador de entradas, de tal forma que para novos valores, o algoritmo consiga classificar estas entradas com base nas informações antigas (CENTRALSERVER, 2020).

1.1 Contextualização e delimitação do tema

Diversas empresas utilizam do Aprendizado de Máquina para maximizarem as experiências dos usuários com os seus serviços ofertados e, assim, aumentarem os seus lucros com propagandas e serviços personalizados, exemplo disso são empresas como Netflix, o Facebook, os bancos digitais como a Digio e o Nubank, a Amazon, o Walmart entre outros.

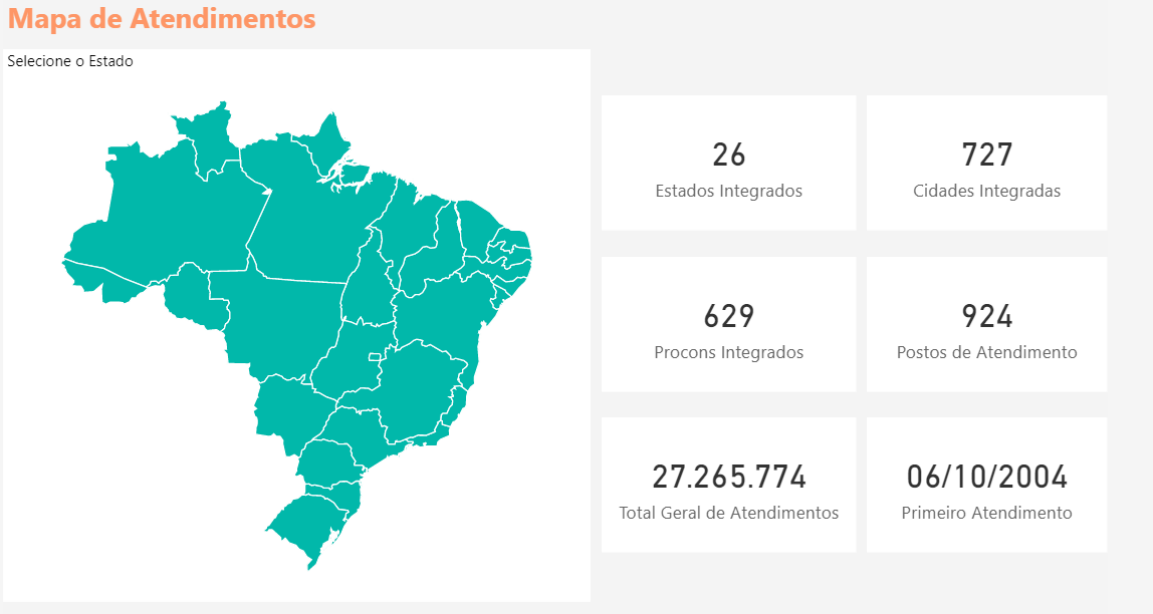
No Brasil, existe uma fundação que está presente em todos os estados e que trata de conflitos entre consumidores e empresas, que vendem produtos ou prestam serviços. Esta fundação é o *Programa de Proteção e Defesa do Consumidor (Procon)*, que tenta solucionar o conflito através de um acordo entre as partes envolvidas. Caso um acordo prévio não seja possível, o conflito será encaminhado para um Juizado Especial Cível. Sempre que um consumidor se sentir lesado por um serviço prestado, ele deve recorrer ao Procon e, através dele, se dará a defesa. Infelizmente, nem sempre é possível conseguir um acordo e o processo acaba indo à justiça.

Graças a era da informação, os consumidores podem abrir reclamações e estas reclamações entram no *Sistema Nacional de Informações de Defesa do Consumidor (Sindec)* , que é uma política pública e, por meio de um conjunto de soluções tecnológicas, representa um eixo fundamental de integração do *Sistema Nacional de Defesa do Consumidor (SNDC)* e de fortalecimento da ação coordenada e harmônica entre seus órgãos. O Sindec permite o registro dos atendimentos individuais a consumidores, a instrução dos procedimentos de atendimento e dos processos de reclamação, além da gestão das políticas de atendimento e fluxos internos dos Procons integrados e a elaboração de Cadastros Estaduais e Nacional de Reclamações Fundamentadas. Todo esse trabalho, harmônico e articulado entre os Procons, gera informações que são consolidadas nos bancos de dados estaduais e replicados na base de dados nacional do Sindec no âmbito do Ministério da Justiça e Segurança Pública. Atualmente, este sistema integra 26 Procons estaduais e 351 Procons municipais espalhados por todo o Brasil. Ao todo, o sistema opera em 629 unidades espalhadas por 448 cidades brasileiras, onde atendem em uma média mensal 261 mil consumidores. Tais informações se configuram em amostra bastante qualificada das diversas demandas e reclamações de consumidores levadas, diariamente, aos órgãos de defesa do consumidor (SINDEC, 2020).

Os dados de todas as reclamações, realizadas entre 2015 e 2019, estão em um

conjunto de dados do governo federal no Portal Brasileiro de Dados Abertos [1]. Esses dados somam quase 100 Mega de texto em formato de tabela, e cada linha desta tabela descreve uma reclamação feita por um consumidor. É possível notar, conforme Figura 1, o imenso número de atendimentos realizados pelo Procon (SINDEC, 2020).

Figura 1 – Gráficos com mapa de atendimentos dos Procons no Brasil



Fonte: (SINDEC, 2020)

Usando os conceitos de Aprendizado de Máquina e o conjunto de dados das reclamações feitas no Procon, há diversas informações sobre as empresas, Estados, municípios das ocorrências. Além disso, a possibilidade da criação de um classificador para as reclamações, predizendo a resolubilidade de uma nova ocorrência registrada no Procon.

1.2 Problematização

Atualmente, quando ocorre alguma problemática relacionada a um serviço ou produto de uma determinada empresa, o consumidor fica ávido pela conclusão do problema o mais rápido possível sem ter a mínima noção se sua reivindicação será atendida ou não pela empresa, apesar da legislação (o decreto federal 6.523 de 2008, conhecido como a Lei do SAC [2]) determinar que a solução ocorra em cinco dias úteis após o registro da queixa.

[1] <http://dados.gov.br/>
[2] <http://www.planalto.gov.br/ccivil_03/_ato2007-2010/2008/decreto/d6523.htm>

Muitas vezes são abertos protocolos no *Serviço de Atendimento ao Cliente (SAC)* e sites de reclamações como Reclame Aqui[3] diante da ineficiência das centrais de atendimento e, com isso, as reclamações dos consumidores aos Procons crescem a cada ano.

Para agilizar o processo de conclusão da contestação do consumidor, mesmo desfrutando dos recursos citados, como o cliente saberá se a demanda será atendida? Esse trabalho tem como intuito produzir a solução e responder a este questionamento.

1.3 Pressuposto

As grandes organizações já começaram a implementar a inteligência artificial de forma mais cotidiana. Em alguns casos, os fornecedores de software incorporaram o Aprendizado de Máquina em ferramentas usadas para um propósito específico, obtendo resultados mais rápidos, precisos e relevantes. Se comparado este mesmo processo sendo realizado de forma humana, seria considerado praticamente inviável a obtenção dos mesmos resultados, devido a enormes quantidades de dados a serem analisados em curto período de tempo.

Com isso muitos órgãos, como o próprio Procon, verificaram a importância de implementação dessa tecnologia para oferecer aos consumidores uma resposta assertiva. Diante do aumento dos números de reclamações nos últimos anos, atrelado à situação atual de pandemia, o distanciamento se tornou algo emergencial e, com isso, a utilização da tecnologia se tornou uma ferramenta essencial e necessária na adaptação desse processo.

1.4 Objetivos

O objetivo geral e os objetivos específicos deste trabalho, são descritos a seguir:

1.4.1 Objetivo geral

O objetivo geral deste trabalho é fazer um estudo sobre as reclamações registradas no Procon utilizando algoritmos de Aprendizado de Máquina, de modo que a partir de novos registros seja possível prever se a determinada reclamação será atendida ou não pela empresa acusada.

[3] <https://www.reclameaqui.com.br>

1.4.2 Objetivos específicos

Como objetivos específicos, destacam-se:

a) Compreender e modelar o problema por meio da escolha de algoritmos de aprendizagem de máquina supervisionados e de técnicas de validação;
b) Interpretar e preparar os dados para utilização, a fim de otimizar os resultados do trabalho;
c) Realizar experimentos com o intuito de obter resultados quantitativos em termos de acurácia dos algoritmos propostos;
d) Disponibilizar a aplicação para o público geral e assim auxiliá-los a entender se uma determinada reclamação será atendida ou não.

1.5 Justificativa

No contexto empresarial a informação é um recurso fundamental para as organizações, uma vez ela pode ser utilizada como vantagem competitiva. Cada vez mais as empresas estão num cenário de crescente competitividade, e, sendo assim, existe a necessidade de tomar decisões rápidas e ter respostas mais assertivas sobre seus negócios. Neste aspecto, a aplicação do Aprendizado de Máquina se torna primordial quando se trabalha com uma grande quantidade de dados, e precisam apresentar resultado preciso de maneira eficiente (SANTOS, 2013).

Tendo em vista a atual situação que se enfrenta por conta da pandemia de Coronavirus [4], é cada vez mais necessário a realçar a importância da transparência nas relações de consumidores e fornecedores. Desta forma, este trabalho torna-se extremamente relevante, pois vai contribuir grandemente como um recurso a mais a ser utilizado pelos consumidores e, por ventura, pelas empresas, a fim de priorizar uma resposta assertiva e mais rápida sobre ocorrências atendidas pelo *Procon*.

1.6 Estrutura do trabalho

Os próximos capítulos são estruturados da seguinte forma:

- No Capítulo 2, são apresentados os conceitos necessários para o entendimento dos capítulos

[4] https://coronavirus.saude.gov.br/sobre-a-doenca

seguintes.

- No Capítulo 3, é apresentada a técnica proposta pelo trabalho, explicando como foi o processo de desenvolvimento dos modelos de aprendizado para atingir os objetivos do trabalho.
- No Capítulo 4, são apresentados os resultados obtidos dos algoritmos estudados, análises da sua utilização e algumas métricas de desempenho.
- Por fim, no Capítulo 5, são apresentadas as conclusões sobre o trabalho e são propostos direcionamentos para trabalhos futuros.

2 REFERENCIAL TEÓRICO

Nesta seção serão apresentados os conceitos e definições que serviram como base para a construção do trabalho.

2.1 Procon

O Procon é um serviço público, mantido pelo governo do estado, que tem como finalidade proteger, amparar e defender o consumidor de práticas comerciais enganosas ou que lhe tragam danos ou prejuízos. Todos os problemas relacionados à compra de produtos e prestações de serviços podem ser encaminhados ao Procon. Alguns destes problemas são:

- Alimentação: qualidade e quantidade, higiene dos estabelecimentos, prazo de validade vencido, etc;
- Assuntos financeiros: cobranças indevidas, multas mal calculadas, envio de cartão sem solicitação, nome do consumidor enviado indevidamente ao Serviço de *Proteção ao Crédito(SPC)* [1], falhas em transações eletrônicas, etc.
- Habitação: problemas na prestação de serviços essenciais (água, esgoto, energia elétrica, gás, telefone), como cobranças indevidas (ligações telefônicas não reconhecidas, elevação injustificada de consumo, serviço não solicitado), interrupção do serviço sem justificativa; aumento abusivo de prestação, problemas com aluguel, condomínio, etc.
- Produtos: defeito ou mau funcionamento não causado por uso indevido do produto, não cumprimento do prazo de entrega prometido ou a entrega de um produto que não corresponde ao que foi comprado, produto não corresponde ao que foi anunciado ou não cumpre o que foi dito em sua propaganda, etc.
- Saúde: problemas relacionados a hospitais, clínicas, laboratórios, medicamentos, planos de saúde, além de serviços veterinários.
- Serviços: problemas relacionados à prestação de serviços contratados com empresas telefônicas, escolas particulares, planos de saúde, consórcios, cartões de crédito, assistência técnica e serviços autônomos em geral, etc.

[1] <https://www.spcbrasil.org.br/>

2.1.1 Como utilizar os serviços do Procon

Algumas informações podem ser fornecidas pelo telefone mas, para encaminhar reclamações e denúncias, é preciso comparecer ao Procon. Para isso, é fundamental que o consumidor junte cópias de toda documentação que puder (nota fiscal, recibos, contratos, certificado de garantia, cartões de cobrança, carnês e comprovantes de pagamento em geral), para que fique bastante caracterizado o prejuízo causado facilitando, assim, a solução ou o encaminhamento do problema.

2.1.2 Prazos para reclamação

De acordo com o artigo *Código de Defesa do Consumidor(CDC)* [2], quando o defeito é aparente, o prazo para reclamação é de 30 dias para produtos não duráveis e 90 dias para os duráveis, contados a partir da data da compra. Se o problema for oculto, os prazos são os mesmos, mas começam a valer no momento em que o defeito é detectado pelo consumidor. Além disso, de acordo com o artigo 18 do *Código de Defesa do Consumidor(CDC)*[3], no caso de o produto ter defeito, o consumidor pode reclamar tanto ao fabricante quanto à loja onde comprou a mercadoria (PROCONSP, 2020).

Abaixo é possível observar alguns exemplos de produtos ou serviços acompanhados de seus respectivos prazos:

- 30 dias: para produtos ou serviços não duráveis, por exemplo: alimentos, serviços de lavagem de roupas numa lavanderia, etc.
- 90 dias: para produtos ou serviços duráveis, por exemplo: eletrodomésticos, reforma de uma casa, pintura do carro, etc.

2.2 Aprendizado de Máquina

O uso do Aprendizado de Máquina pode ajudar as empresas a processarem grandes quantidades de dados complexos para melhorar a análise, a acurácia preditiva e a tomada de decisões e, com isso, elevar a produtividade da companhia, além de facilitar a substituição de serviços prestados por soluções mais práticas (DOMINGOS, 2017).

[2] <https://www.jusbrasil.com.br/topicos/10604414/artigo-26-da-lei-n-8078-de-11-de-setembro-de-1990>
[3] <https://www.jusbrasil.com.br/topicos/10605675/artigo-18-da-lei-n-8078-de-11-de-setembro-de-1990>

Tornando as empresas que utilizam mais competitivas e referências em suas respectivas áreas de atuação, o Aprendizado de Máquina tem a capacidade ampla de compreender as preferências do público através de interfaces como *chatbots* e aplicativos autônomos, vindo a oferecer, em um segundo momento, soluções para problemas futuros (SANTOS, 2013).

2.2.1 Como funciona o Aprendizado de Máquina na prática

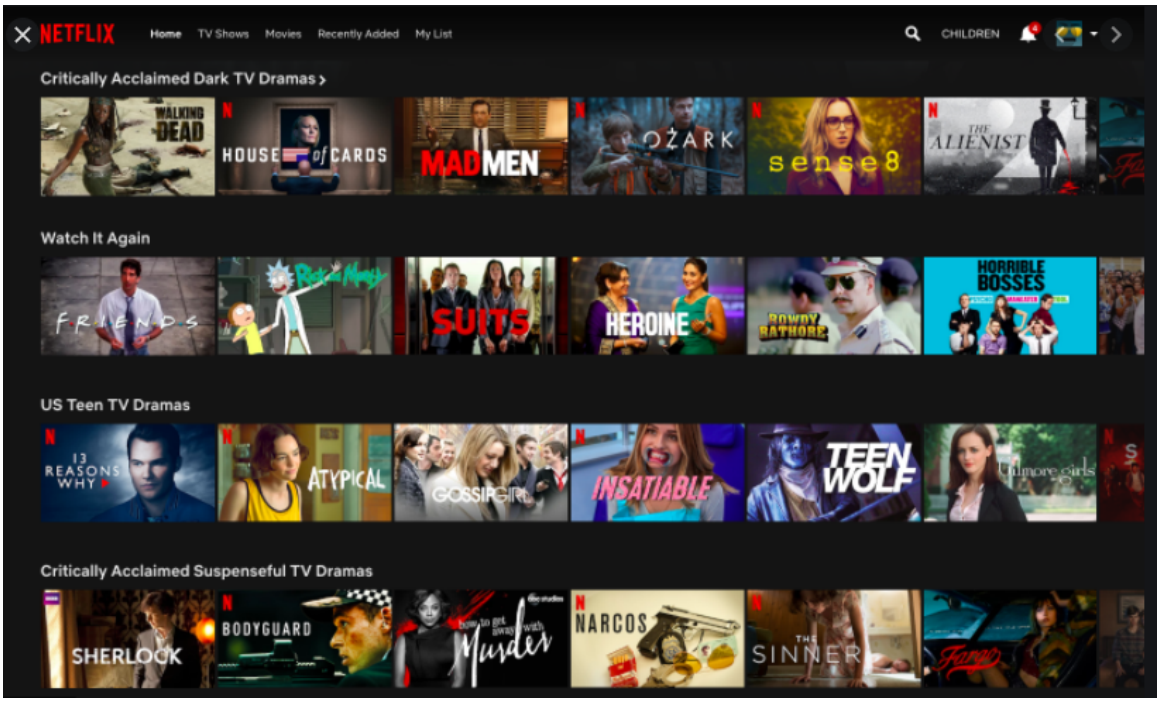
Os algoritmos de aprendizagem, geralmente são baseados em diversos cálculos estatísticos, no qual se utilizam de uma grande massa de dados para realizar os treinos desses algoritmos, com intuito de processar novos dados e capaz de demonstrar uma conclusão esperada. Essa tecnologia funciona porque a essência se faz através de previsão: ela prevê o que se quer, os resultados de nossas ações, como atingir nossos objetivos (DOMINGOS, 2017).

Nem todos os algoritmos de aprendizado funcionam da mesma forma, e as diferenças têm consequências, como por exemplo, os algoritmos de recomendação da Amazon e Netflix. Se numa hipótese um deles estivesse guiado por uma livraria física, tentando determinar o que é interessante, neste caso, provavelmente a Amazona o levaria às prateleiras que você consultou anteriormente conforme exemplo da Figura 2; o da Netflix, o levaria à seção da loja desconhecida, mais com itens que você provavelmente adoraria, conforme Figura 3.

Figura 2 – Imagem ilustrativa de um anúncio direcionado a um perfil de usuário.

Fonte: (AMAZON, 2020)

Figura 3 – Imagem ilustrativa da página inicial do site Netflix com a lista de filmes, séries recomendadas.

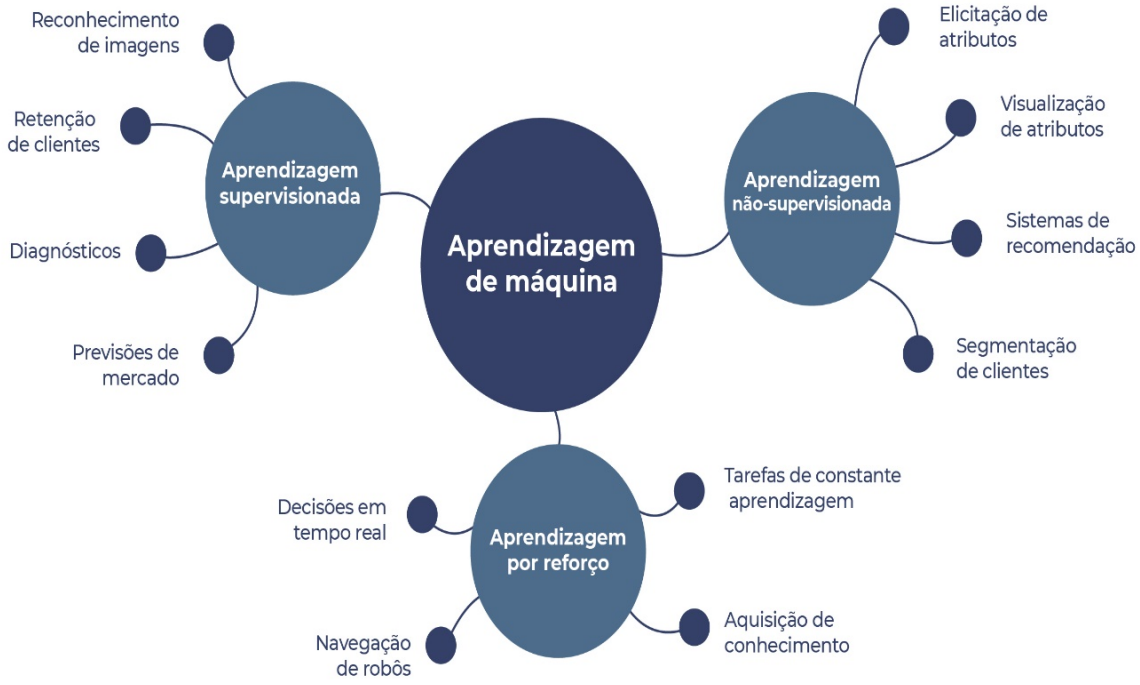


Fonte: (NETFLIX, 2020)

É possível usar os diversos tipos de algoritmos, de acordo com a real necessidade. O modelo de algoritmo utilizado depende do direcionamento da demanda, e saber o que a empresa precisa, é o ponto chave para o sucesso da aplicação do aprendizado de máquina. Na Figura 4 observa-se um mapa mental do fluxo de Aprendizagem de Máquina.

Análise preditiva consiste na aplicação de algoritmos para compreender a estrutura dos dados existentes e gerar regras de predição. Esses algoritmos podem ser utilizados em um cenário não supervisionado, nos quais apenas preditores (covariáveis) estão disponíveis no conjunto de dados, ou em problemas supervisionados, quando, além dos preditores, está disponível também uma resposta de interesse, responsável por guiar a análise (SANTOS, 2013). Por outro lado existe a aprendizagem não supervisionada, que nos permite abordar problemas com pouca ou nenhuma ideia do que nossos resultados devem aparentar (CAMPBELL *et al.*, 2019).

Figura 4 – Imagem ilustrativa do fluxo básico do processo de aprendizagem de máquina



Fonte:(ISI-TICS, 2018)

2.2.2 Formas de Aprendizado de Máquina

Para uma aprendizagem efetiva não basta apenas treinar, mas treinar da maneira mais correta possível e com foco no objetivo principal. No Aprendizado de Máquina não é muito diferente (RUSSELL; NORVIG, 2013). Existem três tipos de *feedback* que determinam os três principais tipos de aprendizagem segundo (RUSSELL; NORVIG, 2013):

a) **Aprendizagem não supervisionada:** O agente aprende padrões na entrada, embora não seja fornecido nenhum *feedback* explícito. A tarefa mais comum de aprendizagem não supervisionada é o agrupamento: a detecção de grupos de exemplos de entrada potencialmente úteis. Por exemplo, um agente de táxi pode desenvolver gradualmente um conceito de "dia de tráfego bom" e "dia de tráfego ruim" sem nunca ter sido rotulados exemplos de cada um deles por um professor.

b) **Aprendizagem por reforço:** O agente aprende a partir de uma série de reforços, recompensas ou punições. Por exemplo, a falta de gorjeta ao final de uma corrida

dá ao agente do táxi a indicação de que algo saiu errado. Os dois pontos de vitória no final de um jogo de xadrez informam ao agente que fez a coisa certa. Cabe ao agente decidir qual das ações anteriores ao reforço foram as maiores responsáveis por isso.

c) **Aprendizagem supervisionada:** O agente observa alguns exemplos de pares de entrada e saída, e aprende uma função que faz o mapeamento da entrada para a saída. No componente 1 dos parágrafos anteriores, as entradas são percepções e a saída é fornecida por um instrutor que diz “Freie!” ou “Vire à esquerda”. No componente 2, as entradas são imagens da câmera, e as saídas vêm de um instrutor que diz “isso é ônibus”. Em 3, a teoria da frenagem é uma função de estados e ações de frenagem até à distância de parada. Nesse caso, o valor da saída está disponível diretamente da percepção do agente (após o fato); o ambiente é o instrutor.

2.2.3 Missões de um algoritmo que aprende

Um algoritmo que aprende necessita executar alguma tarefa segundo (RUSSELL; NORVIG, 2013). Existem alguns tipos de tarefas que são mais presentes nesse tipo de algoritmo:

a) **Classificação:** o algoritmo possui o objetivo de classificar dados. Portanto, consegue classificar um tipo de dado com base em rótulos. Por exemplo, imagine que seu algoritmo conhece os modelos “animais, plantas e montanhas”. Caso você apresente um gato para o algoritmo, ele saberá classificá-lo como um animal com base no seu treinamento prévio..

b) **Regressão:** o algoritmo possui a tarefa de prever valores e comportamentos a partir da análise de dados. Por exemplo, supõe-se que você vai investir seu dinheiro em um imóvel. Não seria interessante saber qual casa valorizará mais no futuro, ou qual o bairro da cidade que terá o maior crescimento urbano? Desenvolver um algoritmo de regressão pode ser a maneira de obter as respostas para essas perguntas.

c) **Agrupamento:** o algoritmo possui a tarefa de separar os dados em grupos, segmentando por características similares. não seria interessante a Netflix saber o que seus principais assinantes possuem em comum? Por exemplo, um algo-

ritmo de agrupamento poderia dividir todos os usuários em grupos conforme as preferências. Uma possível situação seria agrupar conforme o tempo em que determinados usuários assistem um gênero de filme.

2.2.4 Aplicações e problemas

Algoritmos de aprendizado foram implantados com sucesso em uma variedade de aplicações.

- **Classificação de texto ou documento:**
  Exemplos: detecção de spam, previsão de reclamações de um determinado órgão entre outros.
- **Processamento de linguagem natural:**
  Exemplos: análise morfológica, marcação de parte da fala, análise estatística, reconhecimento de entidade nomeada; Reconhecimento de fala, síntese de fala, verificação de falante; *Reconhecimento óptico de caracteres (OCR)* .
- **Aplicações de biologia computacional:**
  Exemplos: função proteica ou predição estruturada.
- **Aplicações na visão computacional:**
  Exemplos: reconhecimento de imagem, detecção de faces; Detecção de fraudes (cartão de crédito, telefone) e intrusão de rede; Jogos, por exemplo, xadrez, gamão.
- **Aplicações gerais:**
  Exemplos:(robôs, navegação), Diagnóstico médico; Sistemas de recomendação, motores de busca, sistemas de extração de informação.

2.2.5 Importância do tratamento dos dados

Segundo (SANTOS; CONTE; PAOLO, 2018) o tratamento prévio dos dados se faz necessário antes de sua utilização no estudo pretendido, qualquer que seja, para que se possa averiguar a natureza dos dados, sua distribuição e possíveis anomalias. Ainda que os dados recebidos já tenham sido utilizados em outra pesquisa, isso não garante que estejam prontos para uso imediato. De acordo com o foco do estudo faz-se necessário prepará-los o que, por vezes implica uma limpeza, agrupamento informações e/ou transformações de parte das informações

e, talvez, selecionar parte de seus descritores principalmente quando estes se mostram bastante extensos.

Dessa maneira os autores concluíram a grande importância da organização e classificação dos dados para conseguir um melhor resultado e acurácia na tomada de decisão a partir dos resultados encontrados que, no caso deste trabalho, é a identificação se uma demanda de reclamação será atendida ou não.

2.2.6 Técnicas de validação cruzada no treinamento dos dados

A validação cruzada é uma técnica para classificar a aptidão de generalização de um modelo, desde um conjunto de dados. Essa técnica é abundantemente empregada em problemas onde o objetivo da modelagem é a predição. Procura estimar a relevância do modelo na prática, ou seja, o seu desempenho para um novo conjunto de dados.

2.2.6.1 Validação cruzada *Hold-Out*

Na validação cruzada *Hold-Out*, é definida uma porcentagem dos dados de forma aleatória para realizar o treino do modelo e outra parte voltada, especificamente, para realizar os testes do modelo. Neste trabalho, está sendo utilizado 75% dos dados para treino e 25% para teste, observa-se na Figura 5 a explanação da fórmula utilizada no *Hold-Out*.

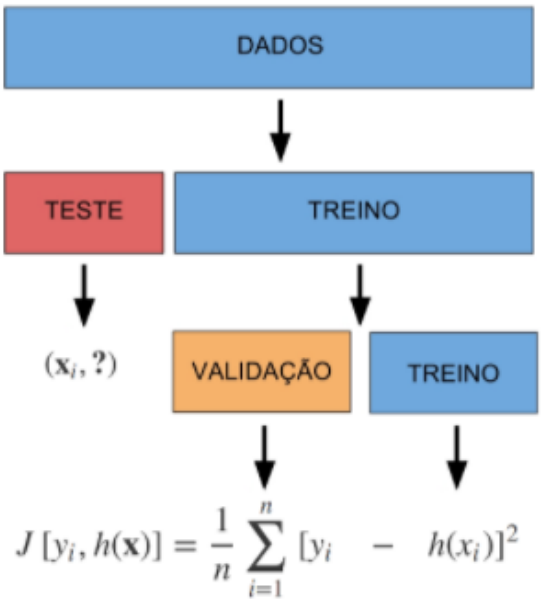
A equação a seguir representa o modelo de validação cruzada *Hold-out*:

$$J[y_i, h(\mathbf{x})] = \frac{1}{n}\sum_{i=1}^{n}[y_i - h(x_i)]^2.$$

Também conhecida como erro quadrático médio *EQM*, trata da soma do quadrado da diferença entre o valor observado: y_i, e o estimado: $h(x)$. Essa métrica é mais útil quando erros grandes são particularmente indesejáveis. Quando se quer uma métrica na mesma unidade da variável de interesse, toma-se a raiz quadrada do erro quadrático médio *REQM* (GEOINFORMAÇÃO, 2016).

Diante da equação $\frac{1}{n}\sum_{i=1}^{n}[y_i - h(x_i)]^2$, os dados de validação são representados por $\frac{1}{n}\sum_{i=1}^{n}\cdot$

Figura 5 – Ilustração da validação cruzada Hold-Out



Fonte: (GEOINFORMAÇÃO, 2016)

Já os dados de treinamento são representados por $[y_i - h(x_i)]^2$,os dados de validação juntamento com os dados de treino que é o valor esperado da equação.

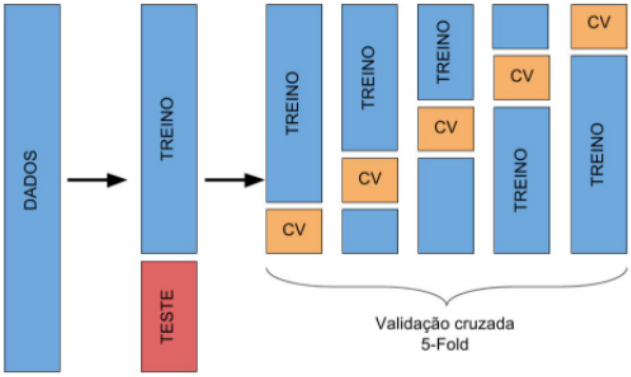
2.2.6.2 Validação cruzada *K-fold*

Na validação cruzada *K-fold*, os dados são divididos em *K-fold* conforme Figura 6, no qual é exequível realizar o treinamento do modelo em k-1 partes com uma parte retida para teste. Esse processo é iterado para certificar que cada parte do conjunto de dados tenha a chance de ser o conjunto retido. Assim que o processamento é concluído, é possível sintetizar a métrica de avaliação usando a média e o desvio padrão.

2.2.7 Algoritmos de aprendizagem

A seguir são detalhados os algoritmos de aprendizagem supervisionados utilizados neste trabalho:

Figura 6 – Imagem ilustrativa de validação cruzada utilizando K-Fold.



Fonte: (GEOINFORMAÇÃO, 2016)

2.2.7.1 Logistic Regression

A regressão logística é o método que utiliza conceitos de estatística e probabilidade. É um algoritmo que trabalha com indagações e impasses de classificação, investigando aspectos divergentes ou variáveis de um objeto para posteriormente definir uma classe na qual ele se encaixa de forma mais favorável possível. Na Figura 8, percebe-se um gráfico básico de regressão logística.

A regressão logística tornou-se uma das técnicas de classificação mais populares para problemas de medicina, marketing e análise de dados, pontuação de crédito, saúde pública e outras aplicações (RUSSELL; NORVIG, 2013).

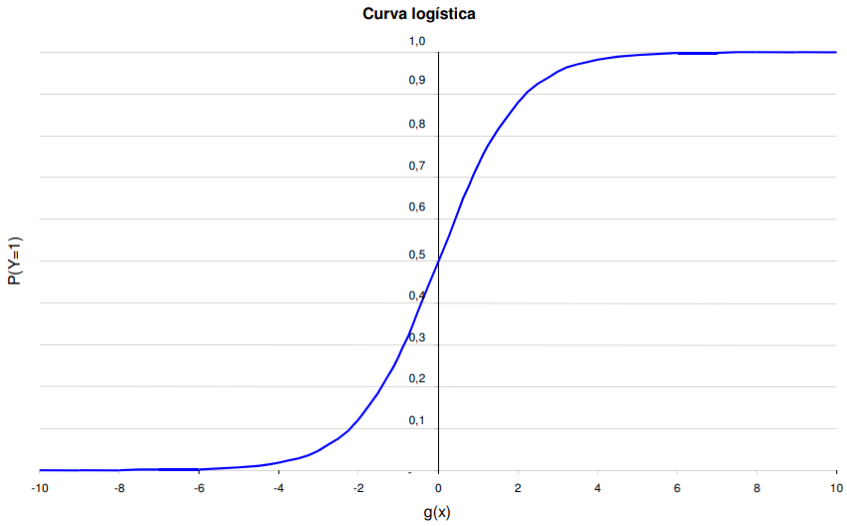
A base da regressão logística é a função logística, também chamada de função sigmoide conforme ilustrada na Figura 7, que obtém qualquer número de valor real e o mapeia para um valor entre 0 e 1 (SONER, 2020).

Figura 7 – Imagem ilustrativa uma função sigmoide

$$Sigmoid\ Function\colon\ y = \frac{1}{1 + e^{-x}}$$

Fonte: (SONER, 2020)

Figura 8 – Imagem ilustrativa do gráfico da regressão logística



Fonte: (EDISCIPLINAS, 2020)

2.2.7.2 Naive bayes

Naive Bayes é um algoritmo de aprendizado supervisionado usado para tarefas de classificação. Por isso, também é chamado de Classificador Naive Bayes. Na Figura 9, é possível observar o teorema de Naive Bayes. Esse algoritmo admite que os recursos são independentes e que não há correspondência entre eles . Contudo, este não é o caso na vida real. Essa suposição ingênua de que os recursos não estão correlacionados é a razão pela qual esse algoritmo é chamado de “ingênuo” (SONER, 2020).

Para alguns tipos de modelos de probabilidade, os classificadores *Naive Bayes* podem ser treinados de forma muito efetiva em um ambiente de aprendizagem supervisionada.

Figura 9 – Imagem ilustrativa do Teorema de Bayes

$$p(A|B) = \frac{p(A).p(B|A)}{p(B)} \qquad (Bayes' Theorem)$$

Fonte: (SONER, 2020)

- p (A | B): Probabilidade de evento A dado evento B já ter ocorrido
- p (B | A): Probabilidade de evento B dado o evento A já ter ocorrido

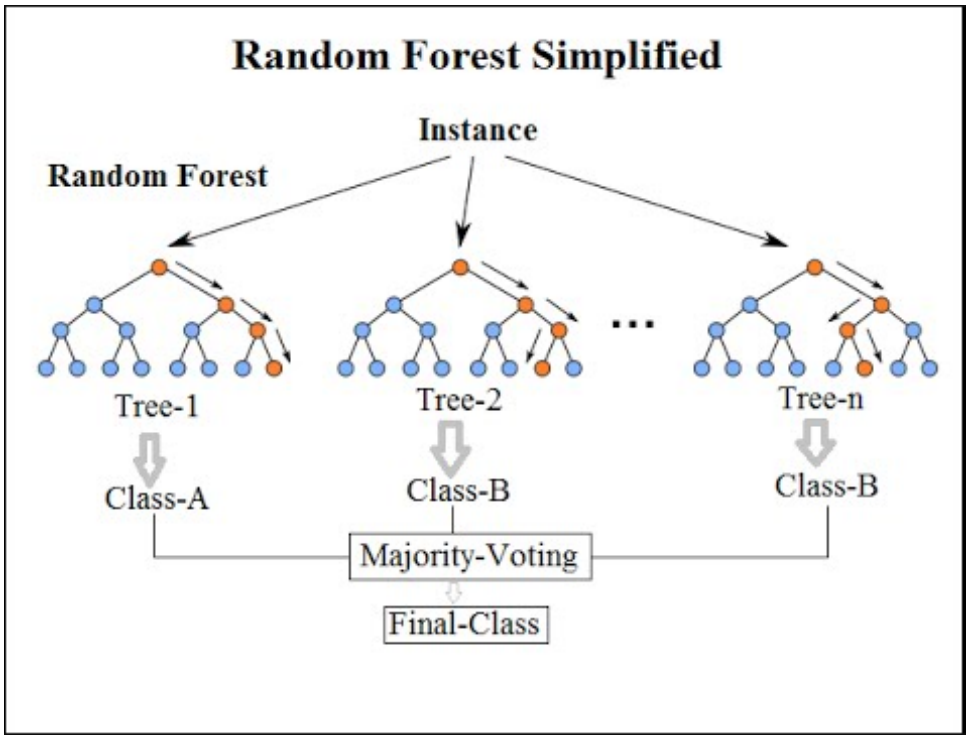
- p (A): Probabilidade do evento A
- p (B): Probabilidade do evento B

A suposição de que todos os recursos são independentes torna o algoritmo *Naive Bayes* muito rápido em comparação com algoritmos mais complexos. Em alguns casos, a velocidade é preferível a maior acurácia.

2.2.7.3 Random Forest

O algoritmo *Random Forest*, traduzido como floresta aleatória, pode ser identificado como uma coleção de árvores de decisão, conforme apresentado na Figura 10, no qual cada árvore estima uma classificação e isso pode ser apelidado como "voto". Nestes termos, considera-se cada voto de cada árvore e se escolhe a classificação mais votada (FURQUIM, 2020).

Figura 10 – Diagrama Random Forest



Fonte: (KDNUGGETS, 2020)

As florestas aleatórias reduzem o risco de readaptação e a acurácia é muito maior do que uma única árvore de decisão. Além disso, as árvores de decisão em uma floresta aleatória são executadas em paralelo e dessa maneira evita que o tempo de execução possa vir a se tornar um problema.

O sucesso de uma floresta aleatória varia muito do uso de árvores de decisão não correlacionadas. Se usarmos árvores iguais ou muito semelhantes, o resultado geral não será muito diferente do resultado de uma única árvore de decisão. O Algoritmo de *Random Forest* conseguem ter árvores de decisão não correlacionadas por ***Bootstrapping*** e ***Feature randomness***, conforme ilustrados na Figura 11 e Figura 12.

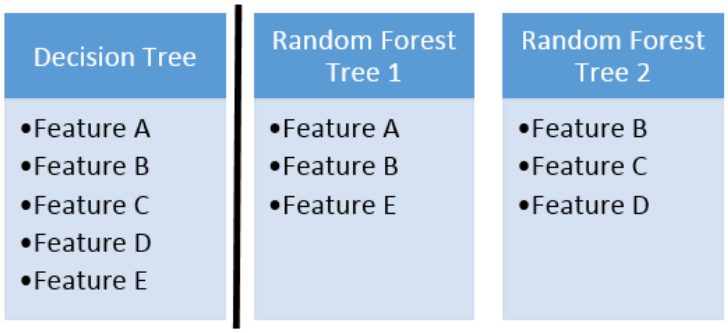
Bootstrapping é a seleção aleatória de amostras de dados de treinamento com substituição. Eles são chamados de amostras de bootstrap.

Figura 11 – Bootsrapping

Fonte: (RESEARCHGATE, 2020)

A ***Feature randomness*** é obtida selecionando recursos aleatoriamente para cada árvore de decisão. O número de recursos usados para cada árvore em uma floresta aleatória pode ser controlado com o parâmetro max_features (SONER, 2020).

Figura 12 – Aleatoriedade de recursos



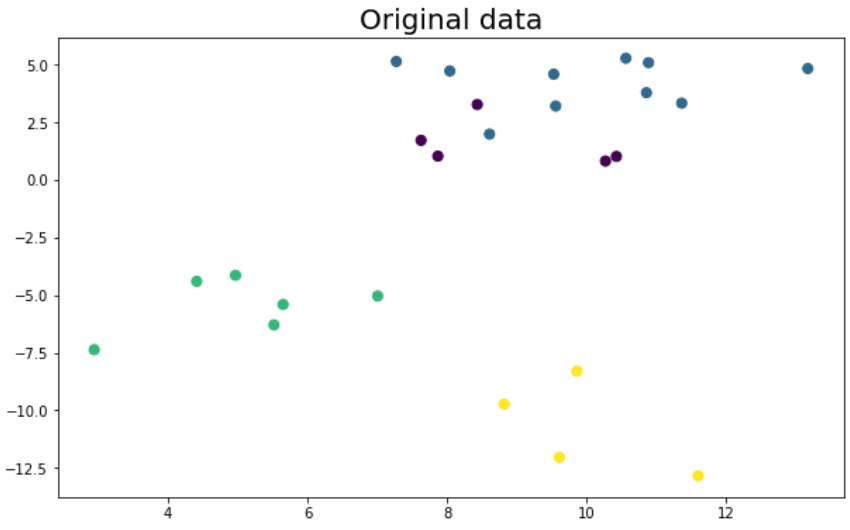
Fonte: (SONER, 2020)

2.2.7.4 KNN (K-Nearest Neighbors)

O algoritmo *K-nearest neighbors (kNN)* é um algoritmo de aprendizado supervisionado que pode ser empregado para solucionar atividades de classificação e regressão. A concepção fundamental por trás do *KNN* é que o valor ou classe de um ponto de dados é definido pelos pontos de dados ao seu redor.

O classificador *KNN* determina a classe de um ponto de dados pelo princípio de votação por maioria. Exemplificando, se k é definido como 5, as classes dos 5 pontos mais próximos são averiguadas. A predição é feita em concordância com a classe predominante. Da mesma forma, a regressão *KNN* assume o valor médio de 5 pontos mais próximos. Segundo a Figura 13, considerando que os seguintes pontos de dados que pertencem a 4 classes diferentes:

Figura 13 – Gráfico exemplo original



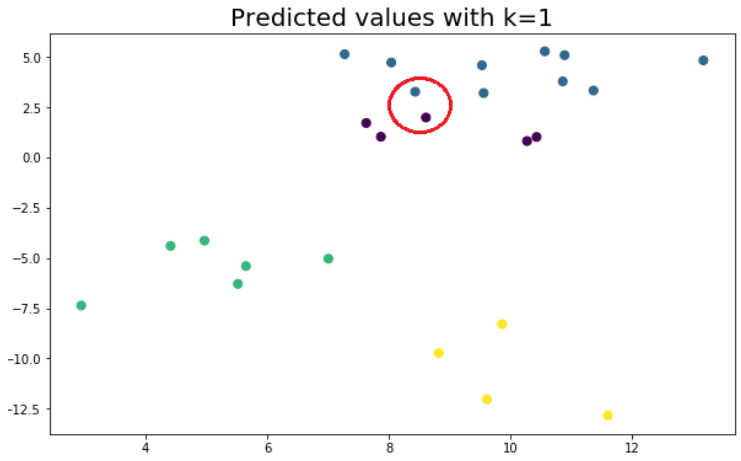
Fonte: (SONER, 2020)

Observa-se como as classes previstas mudam de acordo com o valor k: primeiramente é importante determinar um valor k ótimo. Se k for muito baixo, o modelo é muito específico e não foi bem generalizado. O modelo atinge uma alta acurácia no conjunto de treino, mas será um mau preditor em pontos de dados novos. Logo, é provável que se obtenha um modelo com *overfit*. Por outro lado, se k for muito grande, o modelo é muito generalizado e não é um bom preditor nos conjuntos de treino e teste. Esta situação é conhecida como *underfitting*.

O *kNN* é transparente e fácil de interpretar. Não faz nenhuma teoria, portanto pode

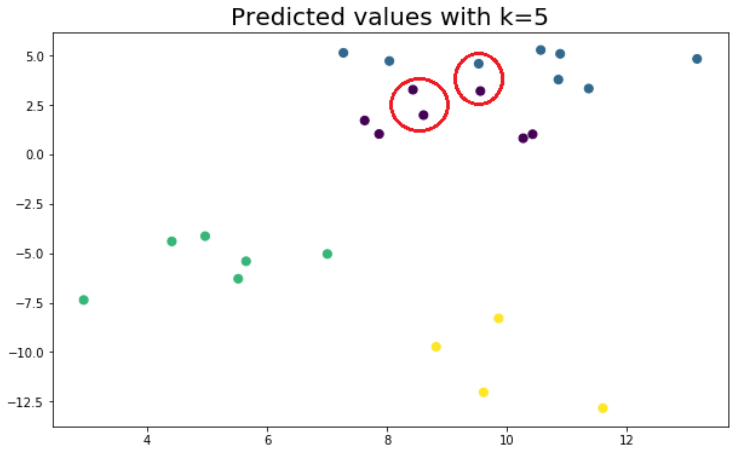
ser implementado em tarefas não lineares. *kNN* faz-se lento à medida que o número de pontos de dados amplia porque o modelo precisa manter todos os pontos de dados. Consequentemente, não é eficiente em termos de memória. Outra desvantagem do *kNN* é que ele é sensível a valores discrepantes. As Figuras 14 e 15 mostram a comparação entre os valores de K, no qual o melhor valor nesse exemplo foi k=5, devido o maior número de agrupamento de pontos semelhantes.

Figura 14 – Gráfico exemplo predição com k = 1



Fonte: (SONER, 2020)

Figura 15 – Gráfico exemplo predição com k = 5



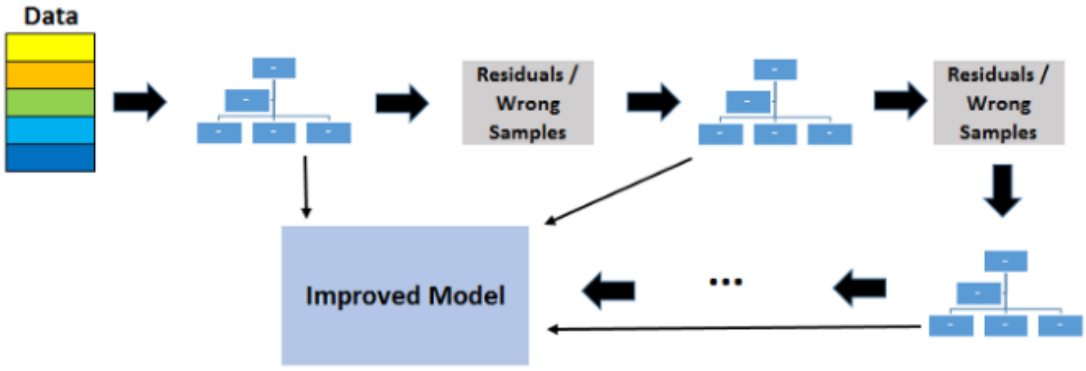
Fonte: (SONER, 2020)

2.2.7.5 Gradient Boosted Decision Trees (GBDT)

O *Gradient Boosted Decision Trees (GBDT)* é um algoritmo de conjunto que usa o método de reforço para combinar árvores de decisões individuais. *Boosting* significa juntar um algoritmo de aprendizagem em série para obter um resultado forte de muitos resultados fracos interligados sequencialmente. No caso do *GBDT*, os resultados fracos são árvores de decisão. Cada árvore tenta minimizar os erros da árvore anterior.

As árvores no *boost* são os resultados fracos, mas a adição de muitas árvores em série e cada uma focando nos erros do anterior torna o *boost* um modelo altamente eficiente e preciso. Cada vez que uma nova árvore é adicionada, ela se adapta a uma versão modificada do conjunto de dados inicial, conforme Figura 16. Como as árvores são adicionadas sequencialmente, os algoritmos de impulsão aprendem lentamente. Na aprendizagem estatística, os modelos que aprendem lentamente têm um desempenho melhor (SONER, 2020).

Figura 16 – Gráfico exemplo Gradient Boosted Decision Trees



Fonte: (SONER, 2020)

O algoritmo *GBDT* é tão dominante que foi desenvolvido diversas versões atualizadas dele, como *XGBOOST*, *LightGBM*, *CatBoost* entre outras.

2.2.7.6 Multinomial Naive Bayes

O Algoritmo *Multinomial Naive Bayes* implementa o algoritmo *Naive Bayes* para dados distribuídos multinomialmente e é uma das duas variantes do *Naive Bayes*, clássico onde

são utilizados na classificação de texto, que por sua vez os dados são normalmente representados como contagens de vetores de palavras, abaixo observa-se melhor a formula e sua definição (SCKITLEARN, 2020).

$$\theta_{c_i} = \frac{N_{c_i} + \alpha}{N_c + \alpha n}$$

Ny é o número total de recursos do evento y (número total de palavras em todas as mensagens de spam), Nyi - contagem de cada recurso (número resumido de repetições de uma palavra em todas as mensagens de spam), n - o número de recursos (número de palavras no vocabulário) e α é um parâmetro de suavização de *Laplace* para descartar a influência de palavras ausentes no vocabulário (HORBONOS, 2020).

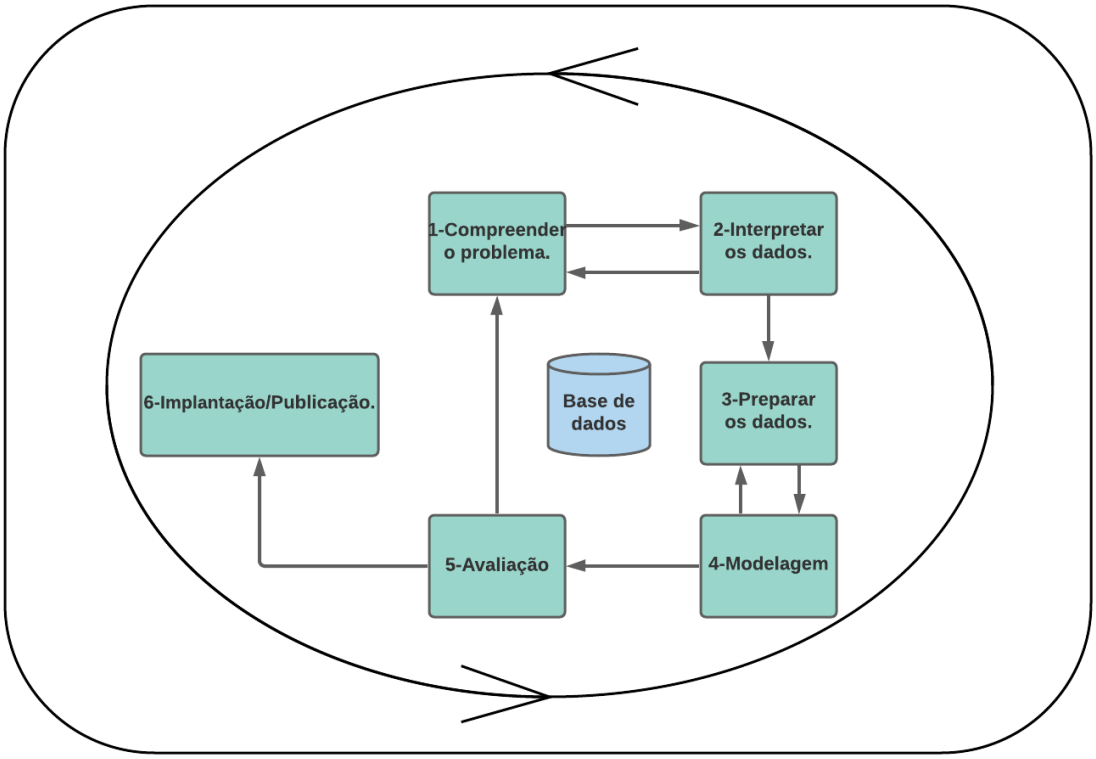
3 METODOLOGIA

Foi utilizado neste trabalho uma classificação de pesquisa exploratória, devido a análise de assuntos relacionados à aplicação de aprendizagem de máquina em dados públicos. Foi aplicada, ainda, a pesquisa descritiva de vários algoritmos para análise e comparação de melhores resultados. Adicionalmente, foi realizado o processo de limpeza e classificação dos dados dispostos no conjunto de dados original, removendo campos indevidos e filtrando colunas específicas para obtenção de melhores resultados. Dessa forma, a metodologia utilizada neste trabalho foi a *Cross Industry Standard Process for Data Mining (CRISP-DM)*. Sua tradução significa Processo Padrão Inter-Indústrias para Mineração de Dados. Essa técnica é amplamente utilizada em trabalhos envolvendo aprendizagem de máquina e tem como principal vantagem a possibilidade de aplicação em diversos tipos de negócios, sem interdependência de uma ferramenta para sua execução. As etapas dessa metodologia são descritas a seguir:

1. **Compreensão do problema:** Nessa fase é preciso ter uma percepção do negócio, aplicando os objetivos do projeto ao negócio.
2. **Interpretação dos dados:** Nessa etapa é onde ocorre a coleta dos dados, tratamento dos dados, exploração e análise dos dados a fim de garantir a qualidade do mesmo.
3. **Preparação dos dados:** Modificar os dados iniciais da fase anterior em possíveis dados finais, geralmente essa fase ocorre diversas vezes em companhia com a Interpretação dos dados.
4. **Modelagem:** Nessa etapa é realizada a construção propriamente dita do seu modelo. Ela consiste na escolha do algoritmo e otimização de parâmetros.
5. **Avaliação:** Nessa fase ocorre a avaliação do modelo, no qual é verificado se o modelo proposto atingiu os objetivos do negócio definidos na primeira fase, caso contrário será necessário retornar a fase inicial para uma nova análise.
6. **Implantação/Publicação:** Nessa fase é realizada a publicação do modelo final alcançado durante todo ciclo.

A Figura 17 mostra o diagrama de etapas do CRISP-DM. O detalhamento de cada uma dessas etapas é descrito nas seções subsequentes.

Figura 17 – Diagrama CRISP-DM



Fonte: Autoria própria

3.1 Compreensão do Problema

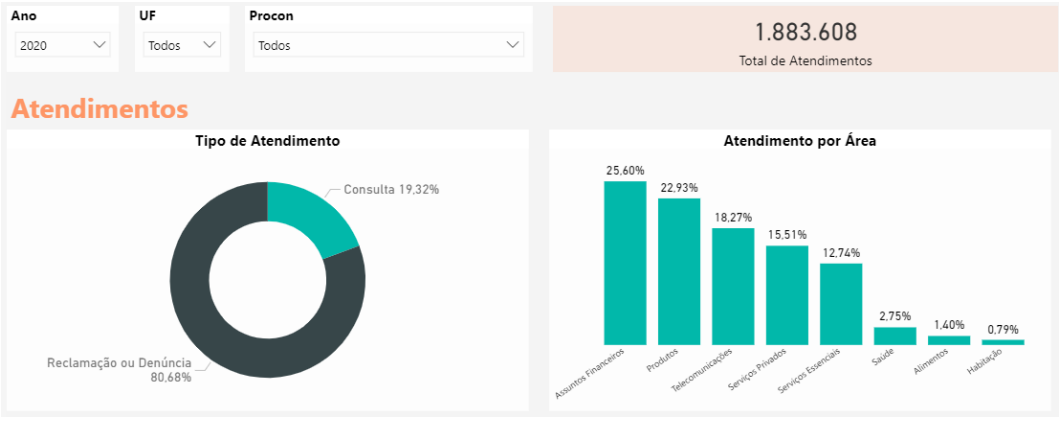
Conforme análise dos dados acessíveis no portal de transparência e coletados pelo *SINDEC* [1], foi possível perceber a quantidade massiva de registros de atendimentos, sendo mais de 80% reclamações ou denúncias, conforme Figuras 18 e 19, que serão apresentadas logo a seguir. Dessa forma, uma aplicação que informa ao consumidor se a sua reclamação será atendida ou não, seria de grande relevância. Logo, poderia ser utilizada como um filtro a mais para o registro de atendimentos realmente pertinentes.

3.2 Interpretação e Preparação dos Dados

Nessa etapa foi analisada a qualidade dos dados, onde foi observado que muitos valores necessitavam de um pré-processamento devido a campos nulos e campos que deveriam ser numéricos, porém estavam como valores não numéricos. Assim, foi realizado o tratamento

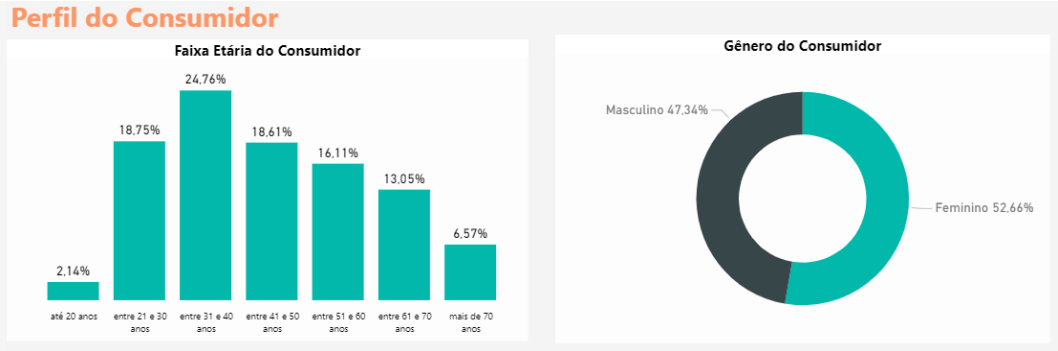
[1] <https://sindecnacional.mj.gov.br/sobre>

Figura 18 – Gráficos com total de atendimentos, tipos de atendimento e atendimentos por área



Fonte: (ATENDIMENTOS, 2020)

Figura 19 – Gráficos com perfil e gênero dos consumidores



Fonte: (ATENDIMENTOS, 2020)

desses valores indevidos. Consequentemente, foi gerado um conjunto de dados unificados para o treino dos modelos de Aprendizagem de Máquina.

Os conjuntos de dados contêm, originalmente, 23 atributos. É possivel verificar cada atributo e suas informações conforme Tabelas 1 e 2.

Neste trabalho foram selecionados 10 atributos, sendo nove provenientes do conjunto de dados original e um atributo criado manualmente. Os atributos utilizados neste trabalho foram: UF, Tipo, CodigoAssunto, CodigoProblema, NumeroCNPJ, CNAEPrincipal, FaixaEtaria-Consumidor, SexoConsumidor, DiasProcessamento e Atendida. O atributo **DiasProcessamento** foi criado a partir de outros dois campos: DataArquivamento e DataAbertura. Já o atributo **Atendida** foi utilizado para validação do objetivo do trabalho, que no caso armazena o valor se foi ou não atendida uma determinada reclamação. Dessa forma, com exceção do atributo **Atendida**, todos os demais campos foram utilizados para treinamento dos algoritmos.

Tabela 1 – Tabela de descrição campos do conjunto de dados

Campo	Tipo	Obrigatório	Descrição
AnoCalendario	Inteiro	Sim	Ano calendário de publicação do cadastro de reclamações fundamentadas.
DataArquivamento	Data	Sim	Data de arquivamento das reclamações (Formato/exemplo: 2011-08-13 12:20:39.000)
DataAbertura	Data	Sim	Data de abertura das reclamações (Formato/exemplo: 2011-08-13 12:20:39.000)
CodigoRegiao	Texto	Sim	Código identificador da região do Procon: 01 - Norte, 02 - Nordeste, 03 - Sudeste, 04- Sul e 05 - Centro-Oeste
Regiao	Texto	Sim	Região do Procon
UF	Texto	Sim	Unidade da Federação do Procon
strRazaoSocial	Texto	Sim	Razão social do fornecedor (empresa) na base de dados do Sindec
strNomeFantasia	Texto	Não	Nome fantasia do fornecedor na base de dados do Sindec (nome comercial / popular / fachada)
Tipo	Inteiro	Sim	Código identificador do tipo da pessoa:1 –Pessoa Jurídica (CNPJ)0 – Pessoa Física (CPF)
NumeroCNPJ	Texto	Não	Número do CNPJ - Cadastro Nacional de Pessoa Jurídica ou CPF - Cadastro de Pessoa Física
RadicalCNPJ	Texto	Não	Aplica-se para pessoa jurídica e serve para agrupar as informações de um mesmo fornecedor (matriz e filiais), sendo os oito primeiros dígitos do número do CNPJ - Exemplo: a matriz (central) do banco e suas filiais (agências)
RazaoSocialRFB	Texto	Não	Razão social do fornecedor na base de dados da RFB – Receita Federal do Brasil.Obs.: somente para os CNPJs (NumeroCNPJ) válidos na base da RFB
NomeFantasiaRFB	Texto	Não	Nome fantasia do fornecedor na base de dados da RFB – Receita Federal do Brasil.Obs.: somente para os CNPJs (NumeroCNPJ) válidos na base da RFB
CNAEPrincipal	Texto	Não	Código identificador da Classificação Nacional de Atividades Econômicas principal do fornecedor. Obs.: somente para os CNPJs (NumeroCNPJ) válidos na base da RFB

Fonte: (SINDEC, 2020)

Tabela 2 – Tabela de descrição campos do conjunto de dados (Continuação)

Campo	Tipo	Obrigatório	Descrição
DescCNAEPrincipal	Texto	Não	Descrição da Classificação Nacional de Atividades Econômicas principal do fornecedor. Obs.: somente para os CNPJs(NumeroCNPJ) válidos na base da RFB
Atendida	Texto	Sim	Código identificador da reclamação fundamentada atendida ou não pela empresa/fornecedor: S – Atendida N – NÃO Atendida
CodigoAssunto	Inteiro	Sim	Código identificador do assunto
DescricaoAssunto	Texto	Sim	Descrição dos assuntos do Sindec (produto ou serviço objeto da reclamação)
CodigoProblema	Inteiro	Sim	Código identificador do problema
DescricaoProblema	Texto	Sim	Descrição dos problemas do Sindec (especificação da lesão sofrida pelo consumidor)
SexoConsumidor	Texto	Sim	Código identificador do sexo do consumidor: M – Masculino F – Feminino N – Não se aplica (são as reclamações (de ofício) em que o Procon é o reclamante)
FaixaEtariaConsumidor	Texto	Não	Faixa etária do consumidor distribuída da seguinte forma: - Até 20 anos - Entre 21 e 30 anos - Entre 31 e 40 anos - Entre 41 e 50 anos - Entre 51 e 60 anos - Entre 61 e 70 anos - Mais de 70 anos - Não Informada (data de nascimento não informada no cadastro do consumidor) - Não se aplica (são as reclamações (de ofício) em que o Procon é o reclamante)
CEPConsumidor	Texto	Não	Código identificador do CEP do consumidor (Código de Endereçamento Postal). Obs.: Não se aplica (são as reclamações (de ofício) em que o Procon é o reclamante)

Fonte: (SINDEC, 2020)

O campo UF é do tipo caractere e descreve a unidade da federação do Procon e, para esse campo, foi criado uma enumeração para representar cada federação.

O campo FaixaEtariaConsumidor é do tipo caractere e é utilizado para identificar a faixa etária do consumidor, e também foi aplicada a técnica de enumeração no qual sua distribuição ficou da seguinte forma:

- Até 20 anos = 15
- Entre 21 e 30 anos = 25
- Entre 31 e 40 anos = 35
- Entre 41 e 50 anos = 45
- Entre 51 e 60 anos = 55
- Entre 61 e 70 anos = 65
- Mais de 70 anos = 75
- NaoInformada = 0

A lógica da enumeração desse campo foi atribuir um valor numérico aproximado da média entre a quantidade de anos de cada faixa etária.

Os dados que foram trabalhados nesse trabalho estão disponíveis no portal de transparência, e foram obtidos através de download, no qual foi selecionado os conjuntos de dados de extensão *CSV* dos anos de 2015 a 2019.

A Figura 20 apresenta uma amostra do conjunto dados do Procon e seus respectivos atributos utilizados neste trabalho.

Figura 20 – Conjunto de dados com as colunas que foram usadas no trabalho.

DiasProces.	UF	Tipo	NºCNPJ	CNAEPrincipal	Cod.Assu	Cod.Prob	Sexo Cons.	Faix.Eta.Con	Atendida
29	PE	1	8279191000184	6512000	101	105	F	entre 31 a 40 anos	S
539	BA	1	8279191000184	6512000	92	105	F	entre 31 a 40 anos	S
65	MS	1	10948651000161	2824102	88	105	F	entre 31 a 40 anos	S
73	CE	1	10882174000189	7911200	233	193	F	entre 21 a 30 anos	N
204	SP	1	5480302000128	2621300	102	102	M	entre 31 a 40 anos	N
15	RJ	1	12069667000120	6319400	141	107	M	entre 41 a 50 anos	N
99	ES	1	12536852000500	4511101	130	105	M	entre 21 a 30 anos	N
33	AL	1	12272084000100	3514000	185	134	F	entre 21 a 30 anos	S
103	BA	1	13607272000104	4752100	101	105	M	entre 21 a 30 anos	S
2922	AP	1	7560958000429	7490104	101	102	F	entre 41 a 50 anos	N
170	BA	1	22770366000182	3240099	102	102	M	entre 41 a 50 anos	S
109	SP	1	19891149000136	NULL	111	116	M	mais de 70 anos	N
35	PE	1	33000118000179	6110801	186	123	M	entre 31 a 40 anos	S
646	DF	1	10310483000184	4619200	96	107	F	entre 31 a 40 anos	N

Fonte: Autoria própria

3.3 Modelagem

Neste trabalho foram utilizados 6 algoritmos de aprendizagem de máquina para verificar a melhor Acurácia dos modelos. Os algoritmos utilizados foram:

1. *Logistic Regression*
2. *Gaussian Naive Bayes (GaussianNB)*
3. *The Random Forest Classifier*
4. *GradientBoostingClassifier*
5. *KneighborsClassifier*
6. *MultinomialNB*

Por fim, foram aplicadas as técnicas de validação *Hold-Out* e *K-fold*, descritas no referencial teórico.

3.4 Avaliação e Implantação

As fases de avaliação do modelo e publicação da solução são descritas no Capítulo 4.

4 RESULTADOS

Esta seção apresenta as especificações de hardware e software utilizadas no trabalho, bem como a análise dos resultados obtidos através da implementação da solução e, por fim, a disponibilização do modelo por meio de um sistema *web*.

4.1 Especificação de hardware e software utilizados

O ambiente utilizado no desenvolvimento do trabalho possui as seguintes especificações: Sistema operacional Windows 10 versão PRO 64 bits, processador AMD FX(tm)-6300 Six-Core Processor 3.50GHz, memória RAM de 10GB e armazenamento de dados em SSD.

Para o desenvolvimento deste trabalho, também foi utilizada a linguagem de programação Python[1], com apoio da IDE *PyCharm Community Edition*[2]. As bibliotecas utilizadas foram:

- Pandas[3], utilizada para leitura e manipulação dos conjuntos de dados.
- Scikit Learn[4], usada para aplicação dos algoritmos de Aprendizado de Máquina.
- zipfile[5], usada para descompactar os conjuntos de dados zipados devido ao tamanho ocupado em disco.
- itertools[6], foi usada para aplicar a técnica de *power set* nos dados dos conjuntos de dados.
- time[7], para cronometrar o tempo de execução de cada algoritmo.
- Flask [8], para disponibilizar a *API* com o modelo de machine learning.
- joblib [9], para exportar o modelo treinado para um arquivo que é consumido pela aplicação.

4.2 Análise dos resultados

Os dados apresentados a seguir mostram os resultados obtidos com a análise de cada algoritmo, no qual foi determinado a acurácia de acerto dos algoritmos e seu tempo de

1 <https://www.python.org/>
2 <https://www.jetbrains.com/pt-br/>
3 <https://pandas.pydata.org/>
4 <https://scikit-learn.org/stable/>
5 <https://docs.python.org/3/library/zipfile.html>
6 <https://docs.python.org/3/library/itertools.html>
7 <https://docs.python.org/3/library/time.html>
8 <https://pypi.org/project/Flask/>
9 <https://pypi.org/project/joblib/>

execução. A acurácia indica uma performance geral do algoritmo utilizado, isto é, dentre todas as classificações (atendida ou não), quantas o modelo classificou corretamente. A acurácia é calculada da seguinte forma:

$$Acuracia = \frac{\text{VP + VN}}{\text{VP + VN + FP + FN}}$$

Onde:

- VP (Verdadeiros Positivos): classificação correta da classe "Atendida";
- VN (Verdadeiros Negativos): classificação correta da classe "Não Atendida".
- FN (Falsos Negativos): erro em que o algoritmo previu a classe "Não Atendida"quando o valor real deveria ser "Atendida";
- FP (Falsos Positivos): erro em que o algoritmo previu a classe "Atendida"quando o valor real deveria ser "Não Atendida";

Adicionalmente, o tempo em que o algoritmo realiza o processamento dos dados de todos os conjuntos de dados desde 2015 à 2019 foram coletados. Os resultados, aplicando as técnicas de validação *Hold-Out* e *K-fold*, são apresentados nas Tabelas 3 e 4, respectivamente.

Tabela 3 – Tabela de resultados com validação Hold-Out

Utilizando Técnica Hold-Out		
Nome Algoritmo	**Precisão Alcançada(%)**	**Tempo de Execução(s)**
Gradient boosting	69.21	80.015
KNN	67.20	24.52
Logistic Regression	61.55	2.523
Multinomial Naive Bayes	48.14	0.752
Naive bayes	61.60	0.907
Random forest	**73.34**	**109.889**

Fonte: Autoria própria

4.3 Aplicação web

Finalmente, após apresentar melhores resultados em termos de acurácia, o algoritmo *Random Forest* foi selecionado para compor a aplicação disponibilizada para o público geral e assim auxiliá-los a entender se uma determinada reclamação será atendida ou não pelo PROCON.

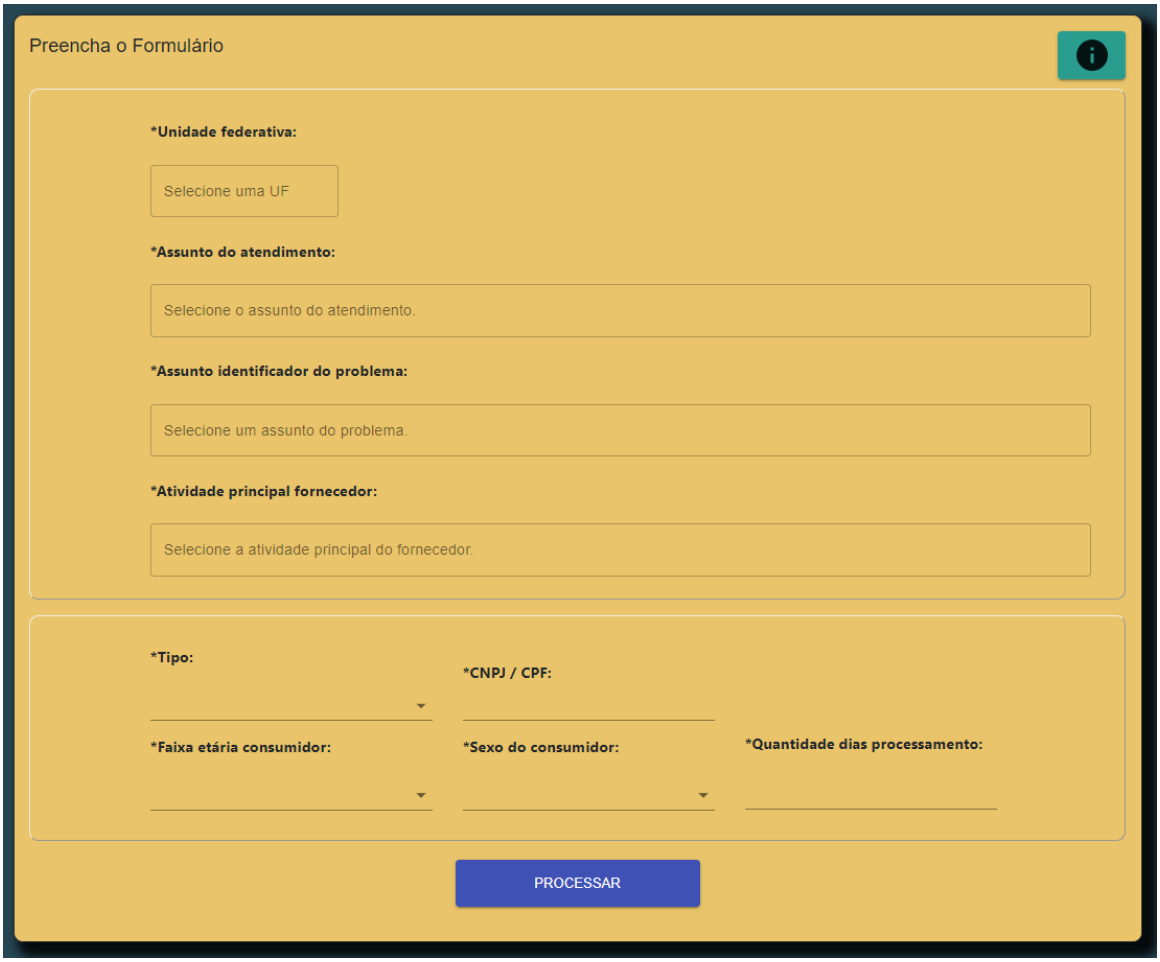
Tabela 4 – Tabela de resultados com validação Kfold

Utilizando Técnica KFold		
Nome Algoritmo	**Acurácia Alcançada(%)**	**Tempo de Execução(s)**
Gradient boosting	68.99	948.173
KNN	67.44	163.219
Logistic Regression	61.64	30.342
Multinomial Naive Bayes	48.06	8.126
Naive bayes	61.66	10.274
Random forest	**73.22**	**1254.376**

Fonte: Autoria própria

Para isso, é necessário que o usuário preencha um formulário com os nove atributos previamente treinados pelo modelo. A Figura 21 apresenta a aplicação web implementada.

Figura 21 – Página inicial da aplicação



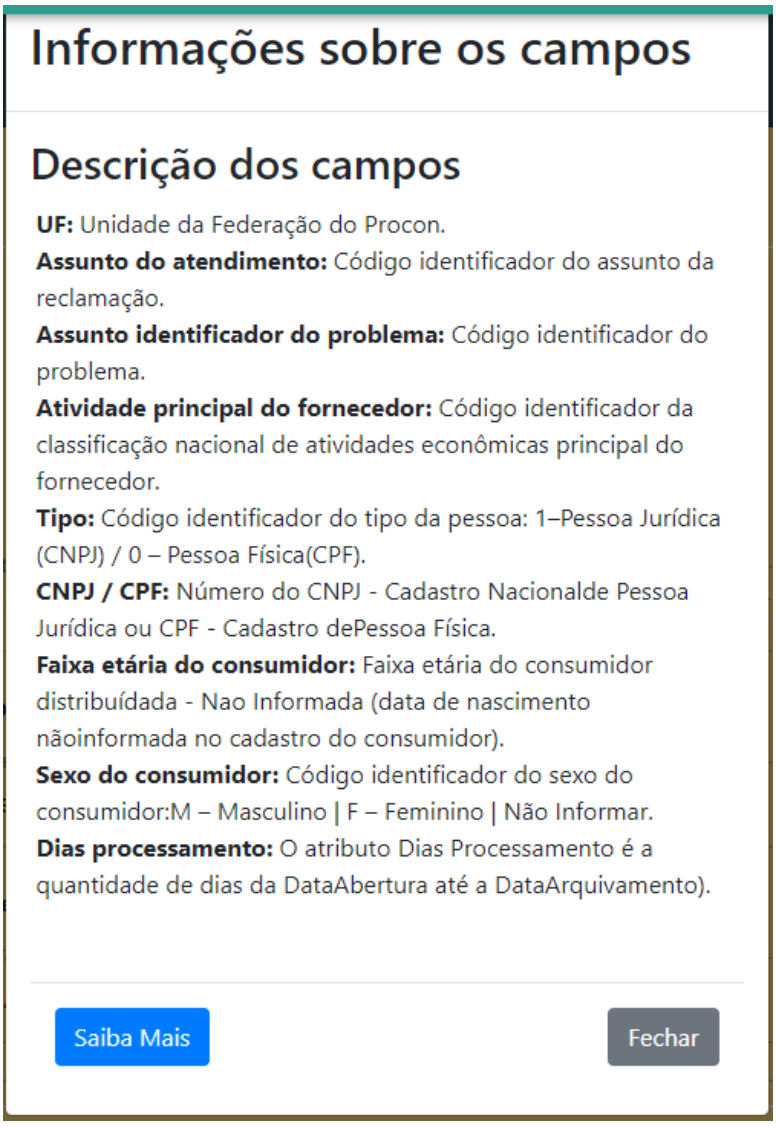
Fonte: Autoria própria

A Figura 22 mostra a descrição de cada campo do formulário para auxiliar o usuário

em seu preenchimento, caso desejado.

Após preenchido todo formulário, a aplicação processa os dados juntamente ao modelo, podendo apresentar dois resultados: "Deverá ser atendida", conforme Figura 23, ou "Não deverá ser atendida", conforme ilustrado na Figura 24. A aplicação foi desenvolvida e executada localmente.

Figura 22 – Modal de ajuda

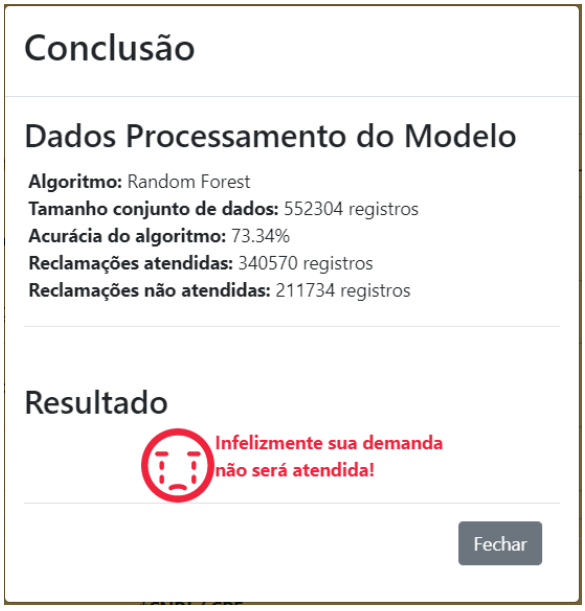


Fonte: Autoria própria

Figura 23 – Resultado da reclamação atendida

Fonte: Autoria própria

Figura 24 – Resultado da reclamação não atendida



Fonte: Autoria própria

5 CONCLUSÃO E TRABALHOS FUTUROS

Este trabalho teve como motivação a incerteza existente sobre a resolução das demandas registradas no PROCON. Todos os dias, inúmeras reclamações são registradas neste órgão, e estas levam uma grande quantidade de tempo para realizar todo processo burocrático de contato e providências entre as partes. Por vezes, o retorno que é dado ao consumidor não é satisfatório e, com isso, sua demanda não é atendida. Este trabalho foi concebido com intuito de orientar a otimização do tempo dos consumidores e minimizar a quantidade de demandas não atendidas no PROCON, uma vez que o consumidor teria uma ferramenta para consultar se sua reclamação eventualmente será atendida ou não. Assim, este trabalho teve como objetivo realizar um estudo sobre as reclamações registradas no PROCON.

Inicialmente, foi realizada a análise e tratamento dos conjuntos de dados disponibilizados pelo portal de transparência do governo, SINDEC. Em seguida foi utilizado o tipo de aprendizado supervisionado juntamente com as técnicas de validação *Hold-Out* e *K-fold* em seis algoritmos: *Logistic Regression*, Naive bayes, Random forest, Gradient boosting, Multinomial Naive Bayes e KNN. Dentre todas as técnicas aplicadas, o algoritmo de *Random forest* foi o que obteve o melhor resultado, tanto aplicando a técnica de validação *Hold-Out* como *K-fold*, em termos de acurácia. Por fim, foi implementada uma aplicação web simples e intuitiva para auxiliar os consumidores de todo Brasil neste processo muitas vezes demorado e burocrático.

Como trabalho futuro, pode ser analisado outros atributos do conjunto de dados e tratá-los para serem úteis aos algoritmos de aprendizado de máquina. Foram utilizados apenas 10 atributos de um total de 23. Adicionalmente, é possível aplicar outras técnicas de Aprendizagem de Máquina, como redes neurais profundas, a fim de obter melhores resultados, de modo a aumentar a acurácia apresentada neste trabalho. Por fim, é plausível utilizar também esses resultados para desenvolver novos trabalhos que apresentem medidas preventivas e corretivas em determinadas regiões, de modo a melhorar não só o atendimento, como também vários outros processos burocráticos.

REFERÊNCIAS

AMAZON. **Pagina Inicial Amazon**. 2020. <https://www.amazon.com.br/>. Acesso em: 20 novembro 2020.

ATENDIMENTOS, S. **Atendimentos MJSP/Senacon - Secretaria Nacional do Consumidor.** 2020. <https://sindecnacional.mj.gov.br/report/Atendimentos>. Acesso em: 02 dezembro 2020.

BATISTA, E. d. O. **Sistemas de informação**. [S.l.]: Saraiva Educação SA, 2017.

CAMPBELL, P. M. d. M.; FRANCELINO, M. R.; FILHO, E. I. F.; ROCHA, P. d. A.; AZEVEDO, B. C. d. Digital mapping of soil attributes using machine learning. **Revista Ciência Agronômica**, SciELO Brasil, v. 50, n. 4, p. 519–528, 2019.

CENTRALSERVER. **Machine learning**. 2020. <https://blog.centralserver.com.br/machine-learning-qual-a-sua-importancia-atual>. Acesso em: 18 janeiro 2021.

DOMINGOS, P. **A revolução do algoritmo mestre**. [S.l.]: Editorial Presença, 2017.

EDISCIPLINAS. **Edisciplinas**. 2020. <https://edisciplinas.usp.br/pluginfile.php/3769787/mod_resource/content/1/09_RegressaoLogistica.pdf>. Acesso em: 21 dezembro 2020.

FURQUIM, C. **10 Algoritmos de Aprendizagem de Máquinas**. 2020. <https://medium.com/@cristianofurquim/10-algoritmos-de-aprendizagem-de-m%C3%A1quinas-machine-learning-que-voc%C3%AA-precisa-saber-c49f9eefe319>. Acesso em: 21 dezembro 2020.

GEOINFORMAÇÃO, L. de Estatística e. **Laboratório de Estatística e Geoinformação - LEG/UFPR. Métodos de reamostragem**. 2016. <http://cursos.leg.ufpr.br/ML4all/apoio/reamostragem.html>. Acesso em: 05 dezembro de 2020.

HORBONOS, P. **Multinomial Naive Bayes Definição**. 2020. <https://towardsdatascience.com/comparing-a-variety-of-naive-bayes-classification-algorithms-fc5fa298379e>. Acesso em: 10 dezembro 2020.

ISI-TICS. **PRINCIPAIS SUB DIVISÕES E APLICABILIDADE DA APRENDIZAGEM DE MÁQUINA**. 2018. <https://isitics.com/2018/05/10/principais-sub-divisoes-e-aplicabilidade-da-aprendizagem-de-maquina/>. Acesso em: 28 novembro de 2020.

KDNUGGETS. **Diagrama Random Forest**. 2020. <https://www.kdnuggets.com/wp-content/uploads/rand-forest-1.jpg>. Acesso em: 21 dezembro 2020.

NETFLIX. **Pagina Inicial NETFLIX**. 2020. <https://www.netflix.com/br/>. Acesso em: 5 dezembro 2020.

PROCONSP. **Atendimento ao Consumidor Procon**. 2020. <https://www.procon.sp.gov.br/>. Acesso em: 14 novembro de 2020.

RESEARCHGATE. **Random Forest Bootstrap**. 2020. <https://www.researchgate.net/figure/An-example-of-bootstrap-sampling-Since-objects-are-subsampled-with-replacement-some_fig2_322179244>. Acesso em: 21 dezembro 2020.

RUSSELL, S. J.; NORVIG, P. Inteligência artificial. Elsevier, 2013.

SANTOS, E. F. D. Descoberta de conhecimento em base de dados do procon utilizando algoritmos e técnicas de inteligência artificial. 2013.

SANTOS, W. J. C.; CONTE, T. N. M.; PAOLO, I. F. D. Avaliação do aprendizado de máquina com base em perspectivas de acurácia como parâmetro principal utilizando cross validation, nltk e knn. In: **CSBC 2018 XXXVIII Congresso da Sociedade Brasileira de Computação**. [S.l.: s.n.], 2018. p. 14.

SCKITLEARN. **Multinomial Naive Bayes**. 2020. <https://scikit-learn.org/stable/modules/naive_bayes.html#multinomial-naive-bayes>. Acesso em: 22 dezembro 2020.

SINDEC. **Dicionário dados sindec**. 2020. <http://dados.mj.gov.br/dataset/8ff7032a-d6db-452b-89f1-d860eb6965ff/resource/d87543d6-cf9d-4752-8f3c-1b0aa075dc45/download/dicionariodadossindec3-0.pdf>. Acesso em: 23 dezembro 2020.

SINDEC, M. P. **Mapa de atendimentos dos Procons.** 2020. <https://sindecnacional.mj.gov.br/report/Mapa>. Acesso em: 02 dezembro 2020.

SONER. **11 Most Common Machine Learning Algorithms Explained in a Nutshell**. 2020. <https://towardsdatascience.com/11-most-common-machine-learning-algorithms-explained-in-a-nutshell-cc6e98df93be>. Acesso em: 21 dezembro 2020.