CENTRO UNIVERSITÁRIO CHRISTUS
SISTEMAS DE INFORMAÇÃO

FLÁVIO WILKER CHAVES PINTO

CLASSIFICAÇÃO DE RECLAMAÇÕES DO PORTAL CONSUMIDOR.GOV.BR UTILIZANDO APRENDIZAGEM AUTOMÁTICA

FORTALEZA

2021

FLÁVIO WILKER CHAVES PINTO

CLASSIFICAÇÃO DE RECLAMAÇÕES DO PORTAL CONSUMIDOR.GOV.BR UTILIZANDO APRENDIZAGEM AUTOMÁTICA

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)

P659c Pinto, Flávio Wilker Chaves Pinto.
Classificação de reclamações do portal Consumidor.gov.br utilizando aprendizagem automática / Flávio Wilker Chaves Pinto Pinto. - 2021.
39 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. Consumidor.gov.br. 2. Aprendizado de Máquina. 3. Classificação. I. Título.

CDD 005

FLÁVIO WILKER CHAVES PINTO

CLASSIFICAÇÃO DE RECLAMAÇÕES DO PORTAL CONSUMIDOR.GOV.BR UTILIZANDO APRENDIZAGEM AUTOMÁTICA

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

Aprovado em: ____ / ____ / _____

BANCA EXAMINADORA

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

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

Prof. MSc. Nicksson Ckayo Arrais de Freitas
Centro Universitário Christus – Unichristus

AGRADECIMENTOS

Agradeço aos meus pais e minha irmã, que sempre me ajudaram, acreditando em mim e me incentivando nos estudos, não permitindo que eu desistisse.

À minha esposa pelo apoio nas minhas escolhas, seu companheirismo e encorajamento nos momentos difíceis.

Aos meus filhos por me darem tanta motivação para sempre seguir adiante.

Ao meu orientador Prof. Felipe Timbó Brito por toda ajuda, compreensão e confiança na conclusão deste trabalho.

Por fim, agradeço aos professores que tive durante minha jornada. Todos tiveram sua parcela de contribuição na formação do profissional que me tornei.

RESUMO

Devido à crescente complexidade dos problemas nas últimas décadas, surgem cada vez mais ferramentas computacionais autônomas, que reduzem a intervenção humana por meio da utilização de Inteligência Artificial (IA). Um subgrupo da IA, denominado Aprendizado de Máquina, lida com algoritmos computacionais, que podem predizer certos comportamentos por meio da utilização de dados históricos. No Brasil existe uma plataforma denominada Consumidor.gov.br, a qual visa solucionar conflitos entre consumidores e empresas. Os consumidores, usuários da plataforma Consumidor.gov.br por muitas vezes têm suas reclamações não atendidas ou mesmo recusadas. Por esse motivo, seria importante para o consumidor ter uma análise prévia, de maneira automática, se suas reclamações serão atendidas ou recusadas pelas empresas, anteriormente a abertura de uma reclamação na plataforma. Dessa forma, consumidores podem procurar diretamente o Programa de Proteção e Defesa do Consumidor (PROCON) de sua cidade antes de buscar um acordo prévio com a empresa por meio do Consumidor.gov.br. Este trabalho visa utilizar Aprendizado de Máquina para classificar automaticamente reclamações não atendidas, ou mesmo recusadas, contidas na plataforma Consumidor.gov.br. Resultados mostram que é possível classificar tais reclamações com acurácia acima de 85% para os diversos contextos analisados.

Palavras-chave: Consumidor.gov.br. Aprendizado de Máquina. Classificação.

ABSTRACT

Due to the growing complexity of the problems in recent decades, there are more and more autonomous computational tools that reduce human intervention through the use of Artificial Intelligence (AI). A subset of AI, called Machine Learning, deals with computational algorithms, which can predict certain behaviors through the use of historical data. In Brazil there is a platform called Consumidor.gov.br which aims to solve conflicts between consumers and companies. Consumers, users of the Consumidor.gov.br platform, often have their complaints unanswered or even refused. For this reason, it would be important for consumers to have a prior analysis, automatically, whether their complaints will be answered or rejected by companies, prior to opening a complaint on the platform. In this way, consumers can look for PROCON in their city directly before seeking a prior agreement with the company through Consumidor.gov.br. This work aims to use Machine Learning to automatically classify unanswered or even refused complaints contained in the Consumidor.gov.br platform. Results show that it is possible to classify such complaints with accuracy above 85% for the different contexts analyzed.

Keywords: Consumidor.gov.br. Machine Learning. Classification.

LISTA DE FIGURAS

LISTA DE TABELAS

LISTA DE ABREVIATURAS E SIGLAS

CRISP-DM	*Cross Industry Standard Process for Data Mining*
CSV	*Comma Separated Values*
IA	Inteligência Artificial
KNN	*K-nearest neighbors*
PROCON	Programa de Proteção e Defesa do Consumidor
Senacon/MJ	Secretaria Nacional do Consumidor do Ministério da Justiça

SUMÁRIO

1 INTRODUÇÃO

A Inteligência Artificial teve início após a Segunda Guerra Mundial, sendo uma das ciências mais recentes e, atualmente é aplicada em muitos campos e contextos da sociedade, como jogos, carros automáticos, diagnósticos de doenças e muitos outros. A Inteligência Artificial auxilia diversas atividades relacionadas ao pensamento humano, tomada de decisões e resolução de problemas. Sendo assim, torna-se muito relevante para qualquer área da atividade intelectual humana (RUSSELL; NORVIG, 2004).

Devido à crescente complexidade dos problemas nas últimas décadas, surgem cada vez mais ferramentas computacionais autônomas, que reduzem a intervenção humana. Essas técnicas têm a capacidade de criar por si mesmas uma hipótese ou função, a partir de uma experiência passada, que seja capaz de resolver o problema desejado. Um exemplo é a descoberta de uma hipótese na forma de regra, ou conjunto de regras, para saber quais clientes de um supermercado irão receber uma determinada propaganda, utilizando-se do histórico do cliente na base de dados do próprio supermercado (CARVALHO *et al.*, 2011).

Dentro da Inteligência Artificial surgiu o aprendizado de máquina, que lida com algoritmos computacionais que podem predizer certos comportamentos, por meio da utilização de dados históricos sem programação explícita. Apesar de ser considerada uma ferramenta poderosa para geração de conhecimento, no aprendizado automático não existe um único algoritmo capaz de apresentar melhor desempenho para todos os tipos de problemas. Dessa forma, é importante compreender os pontos fortes e as principais limitações dos diversos tipos de algoritmos. (MONARD; BARANAUSKAS, 2003).

1.1 Contextualização e delimitação do tema

No Brasil existe uma plataforma chamada Consumidor.gov.br, lançada pelo Ministério da Justiça, que funciona como um serviço de defesa do consumidor. Através dessa plataforma, os consumidores podem fazer reclamações de empresas, por sua vez, essas empresas podem se manifestar para que possam explicar o problema, ou dar uma solução aos problemas relatados nas reclamações (FERNANDES; FILHO, 2015).

O Consumidor.gov.br é um serviço público para solucionar conflitos de consumo

através da internet, permitindo a interlocução entre os consumidores e empresas. Esse serviço também permite a resolução desses conflitos de forma rápida e sem burocracia. A participação das empresas no serviço deve ser solicitada diretamente pela empresa, com o comprometimento de solucionar os problemas apresentados a elas. Por outro lado, os consumidores devem cadastrar os seus dados adequadamente e apresentar todos os dados necessários ao cadastrar uma reclamação contra uma empresa (CONSUMIDOR.GOV.BR, 2021b).

1.2 Problematização

Os consumidores, que utilizam a plataforma Consumidor.gov.br, muitas vezes têm suas reclamações não atendidas. Algumas vezes, essas reclamações podem ser recusadas pela empresa, isto é, quando a empresa comunica ao gestor da plataforma que a reclamação não é válida. Em alguns casos, o gestor acata a decisão da empresa e a recusa é, portanto, classificada como procedente. Caso contrário, a recusa é considerada improcedente e a empresa deverá dar prosseguimento ao tratamento da reclamação aberta pelo consumidor. Com base no fluxo descrito acima, seria importante para o consumidor ter uma análise prévia, de maneira automática, se suas reclamações serão atendidas ou mesmo recusadas pelas empresas, anteriormente a abertura de uma reclamação na plataforma.

1.3 Pressuposto

Assume-se que a utilização de algoritmos de aprendizagem de máquina pode auxiliar na classificação das reclamações contidas na plataforma Consumidor.gov.br. Primeiramente, a plataforma possui um vasto conjunto de dados que pode ser utilizado para treinamento de modelos computacionais. Além disso, um estudo recente (SILVA, 2021) mostrou que, para dados semelhantes aos deste trabalho, vários algoritmos de aprendizado de máquina conseguem classificar reclamações atendidas e não atendidas. Por fim, uma ferramenta capaz de classificar não só as reclamações atendidas ou não, mas também reclamações recusadas, auxiliaria o consumidor a procurar diretamente o PROCON de sua cidade, antes de buscar um acordo prévio com a empresa por meio do Consumidor.gov.br.

1.4 Objetivos

A seguir, são apresentados o objetivo geral e os objetivos específicos deste trabalho.

1.4.1 Objetivo geral

Este trabalho visa classificar, de maneira automática, as reclamações da plataforma Consumidor.gov.br, utilizando aprendizado de máquina.

1.4.2 Objetivos específicos

Como objetivos específicos, pretende-se classificar os seguintes cenários:

a) Se uma reclamação do consumidor, contra uma determinada empresa, será atendida.
b) Se a empresa irá recusar a reclamação do consumidor antes do atendimento.
c) Caso a empresa recuse uma reclamação aberta na plataforma, se a recusa é procedente ou improcedente.

1.5 Estrutura do trabalho

Os próximos capítulos são estruturados da seguinte forma: No capítulo 2 é apresentado a fundamentação teórica deste trabalho. No capítulo 3 é mostrado a metodologia utilizada e todo o passo a passo utilizado para construção dos modelos computacionais. No capítulo 4 é mostrado os resultados obtidos em cada contexto apresentado nos objetivos. Por fim, o trabalho é concluído e direções futuras são apontadas no capítulo 5.

2 FUNDAMENTAÇÃO TEÓRICA

Nesta seção será apresentada a fundamentação teórica deste trabalho, abordando os principais conceitos que compõem a abordagem proposta. Inicialmente são detalhados conceitos sobre a plataforma Consumidor.org.br. Em seguida, são especificadas noções sobre aprendizagem de máquina, bem como formas de dividir conjuntos de dados e estratégias para *Encoding*. Por fim, são descritas técnicas de correlação entre atributos e métricas de avaliação de modelos.

2.1 Consumidor.gov.br

Na plataforma Consumidor.gov.br, o consumidor pode se comunicar com qualquer empresa participante, a qual se compromete em receber, analisar e responder qualquer reclamação do consumidor no prazo de 10 dias. Primeiramente, deve ser verificado se a empresa está cadastrada no serviço. Depois, com o cadastro da reclamação feito, a empresa tem até 10 dias para analisar e responder. Após a resposta da empresa o consumidor tem 20 dias para avaliar a resposta, informando se foi "Resolvida" ou "Não Resolvida". Por fim, ele pode informar qual foi o seu nível de satisfação com o atendimento (CONSUMIDOR.GOV.BR, 2021a).

Existem algumas condutas, inclusas nos termos de uso que são avaliadas pelas próprias empresas quando recebem uma reclamação. Dessa forma, caso o consumidor esteja infringindo alguma delas, a empresa pode recusar a reclamação (CONSUMIDOR.GOV.BR, 2021c). Essa recusa é analisada por um órgão gestor responsável, que tem um prazo de 15 dias para avaliar se a recusa é procedente ou improcedente. Se a recusa for procedente, a reclamação é cancelada. Se a recusa for improcedente, a reclamação volta para a empresa e o tempo de resposta é retomado (SENACON/MJ, 2020). Caso a reclamação do consumidor não seja atendida, ele deve buscar o serviço prestado pelos órgãos de defesa do consumidor, como por exemplo, os Procons de cada Unidade Federativa (CONSUMIDOR.GOV.BR, 2021a).

O Consumidor.gov.br não substitui os serviços dos órgãos de defesa do consumidor, também não constitui um procedimento administrativo. A ideia dessa plataforma é possibilitar um contato direto entre consumidores e empresas, disponibilizando um ambiente público e transparente, e dispensando a intervenção do poder público na reclamação. Isso permite a conversa direta entre o consumidor e a empresa para solução de conflitos de consumo através da

internet (CONSUMIDOR.GOV.BR, 2021a).

Todos os dados das reclamações registradas no Consumidor.gov.br são registrados em uma base de dados pública. Essa base de dados tem informações como: empresas que têm os melhores índices de solução e satisfação, as que responderam nos menores prazos, e muitos outros indicadores (CONSUMIDOR.GOV.BR, 2021a).

O conteúdo das reclamações também é acessível. Campos específicos como segmento de mercado, área, assunto, problema, classificação (resolvida / não resolvida/ não avaliada), nota de satisfação, e muitos outros estão abertos ao público para consulta. Isso permite que qualquer indivíduo tenha acesso aos dados e faça diversas análises (CONSUMIDOR.GOV.BR, 2021a).

O Consumidor.gov.br é monitorado por diversos órgãos como: a Secretaria Nacional do Consumidor do Ministério da Justiça (Senacon/MJ), Procons, Defensorias Públicas, Ministérios Públicos, Agências Reguladoras (CONSUMIDOR.GOV.BR, 2021a).

2.2 Aprendizado de Máquina

Há alguns anos a área de Inteligência artificial tinha pouco valor prático, sendo utilizada mais na área teórica. Na década de 1970 tivemos uma maior popularização de técnicas computacionais com base em Inteligência Artificial na solução de problemas reais. Antes esses problemas eram tratados computacionalmente com conhecimento inicial de especialistas em uma determinada área, como medicina, por exemplo. Esse conhecimento era codificado em programas computacionais (CARVALHO *et al.*, 2011).

Para um comportamento inteligente a capacidade de aprendizado é essencial. Memorizar, observar e explorar situações para aprendizado, melhorar habilidades através da prática e organizar o conhecimento são atividades relacionadas ao aprendizado. Em aprendizado de máquina, os computadores aprendem programaticamente de acordo com a experiência passada (CARVALHO *et al.*, 2011).

Em todos os casos não triviais, tirar algum conhecimento dos dados não é óbvio, por exemplo, detectar se um e-mail é ou não spam. Será necessário procurar por certas palavras usadas juntas, combinado com o tamanho do e-mail, utilizar de estatística e outros fatores, assim

transformando dados em informação (HARRINGTON, 2012).

Por outro lado, temos outros problemas que podem ser não determinísticos, por exemplo, a motivação humana. Em ciências sociais, acertar 60% das vezes é considerado um sucesso. Por exemplo, será que não poderíamos prever o resultado de eventos humanos com base na suposição que humanos agem para aumentar a sua felicidade? Isso seria possível, mas é difícil definir o que faz alguém feliz, isso pode ser diferente para cada pessoa (HARRINGTON, 2012).

Uma das formas de aprendizado de máquina que temos, é a aprendizagem não supervisionada, nesse caso o agente em questão aprende os padrões de entrada, sem nenhum *feedback* passado diretamente. Usando a detecção de grupos de exemplos de entradas que sejam potencialmente úteis. Um exemplo seria um agente de táxi desenvolver gradualmente o conceito de “dia de tráfego bom” e “dia de tráfego ruim” sem a necessidade de ter passado exemplos para ele por um humano (RUSSELL; NORVIG, 2004).

Outra forma de aprendizado de máquina é a aprendizagem supervisionada. Nesse caso o agente observa alguns exemplos de entrada e saída, depois aprende uma função para fazer o mapeamento de entrada e saída. Tendo como exemplo anterior do táxi, as entradas podem ser as percepções do agente ao conceito de "dia"e a saída é fornecida por um humano que vai informar o momento de frear e virar a esquerda, por exemplo, assim o agente vai aprendendo com a ajuda de uma base inicial (RUSSELL; NORVIG, 2004).

2.2.1 Como dividir o conjunto de dados?

Podemos dividir os dados em dois conjuntos diferentes para o aprendizado e avaliação da predição dos dados. A divisão em treino e teste é uma técnica que visa avaliar o desempenho de um algoritmo de aprendizado de máquina (BROWNLEE, 2021).

A divisão em treino e teste é uma técnica em que seus resultados permitem comparar o desempenho de algoritmos de aprendizado de máquina para um problema de modelagem desejado. Embora seja simples de usar e interpretar, existem casos em que não devemos usá-lo. Quando você tem um pequeno conjunto de dados e situações em que uma configuração adicional é necessária, ou quando é usado para classificação e o conjunto de dados não está balanceado (BROWNLEE, 2021).

Essa técnica é mais adequada quando existe um conjunto de dados muito grande, um modelo muito custoso para treinar, ou precisa de uma boa estimativa de desempenho do modelo rapidamente. Ela pode ser usada para problemas de classificação ou regressão e pode ser usada também para qualquer algoritmo de aprendizado supervisionado (BROWNLEE, 2021).

A técnica de divisão em treino e teste funciona da seguinte forma: pega-se um conjunto de dados e divide em dois subconjuntos. O primeiro subconjunto é usado para treinar o modelo, é o conjunto de dados de treinamento. No segundo subconjunto o elemento de entrada do conjunto de dados é fornecido ao modelo e, em seguida, as previsões são feitas e comparadas aos valores esperados. O objetivo é fazer uma estimativa da performance do modelo de aprendizado de máquina em novos dados que não foram usados para treinamento do modelo (BROWNLEE, 2021).

2.2.2 Como lidar com os tipos de dados?

Nos modelos de Aprendizado de Máquina temos casos em que os dados de entrada são dados de texto e que precisam ser convertidos em uma representação numérica para se encaixar nos modelos de aprendizado de máquina. Existem duas formas mais comuns para pre-processamento de dados categóricos, uma seria usar *Label Encoder* e outra seria usar *One Hot Encoder*. Ambos possuem vantagens e desvantagens, o que irá depender das características do conjunto de dados para decidir qual deles deve ser usado (JIANG; LIN; RAGHAVAN, 2020).

O *Label Encoder* tem a vantagem de que ele não aumenta o tamanho dos dados de entrada. Ele converte cada valor da variável categórica em um valor numérico. Essa técnica cria apenas uma única variável numérica correspondente a uma variável categórica, em que cada valor corresponde a um número. Normalmente, o primeiro recebe valor 1, o segundo recebe valor 2 e assim por diante (THOMAS; JUDITH, 2020). A Tabela 1 mostra um exemplo de *Label Encoder*:

Tabela 1 – Exemplo de *Label Encoder*.

Grade	Grade_label_encoded
average	1
excellent	2
average	1
good	3
excellent	2
average	1
excellent	2

Fonte: (THOMAS; JUDITH, 2020).

Uma das desvantagens de usar o *Label Encoder* é que ele introduz uma comparação numérica entre os diferentes valores. Isso pode trazer má interpretação, afetando assim o desempenho do modelo. Não se deve haver diferença de peso ou ordem para os valores, e eles devem ser tratados igualmente durante o treinamento do modelo (JIANG; LIN; RAGHAVAN, 2020). Para evitar isso, utiliza-se o *One Hot Encoder*.

A técnica de o *One Hot Encoder* cria novos parâmetros para representar diferentes valores categóricos, depois atribui 0 ou 1 para indicar se pertence à uma determinada categoria. Todos as colunas processadas têm um relacionamento independente (JIANG; LIN; RAGHAVAN, 2020). A Tabela 2 mostra um exemplo de utilização do *One Hot Encoder*:

Tabela 2 – Exemplo de *One Hot Encoder*.

Grade	Grade_average	Grade_excellent	Grade_good
average	1	0	0
excellent	0	1	0
average	1	0	0
good	0	0	1
excellent	0	1	0
average	1	0	0
excellent	0	1	0

Fonte: (THOMAS; JUDITH, 2020).

No entanto, a desvantagem dessa técnica é que se os valores únicos para a variável categórica forem muitos, o *One Hot Encoder* cria diversas novas colunas iguais ao número desses

valores. Isso pode acarretar em problemas de processamento e memória (THOMAS; JUDITH, 2020).

2.3 Seleção e Correlação entre Atributos

A etapa de seleção de atributos busca remover atributos redundantes ou mesmo não-informativos do modelo proposto. Testes de correlação são uma forma promissora de medir quais atributos devem ser selecionados (HALL, 1999). Para dados numéricos, os coeficiente de *Pearson* e *Spearman* são os mais conhecidos. Para dados categóricos, testes de Qui-quadrado e V de Cramér funcionam bem.

Muitas vezes é necessário calcular a correlação entre variáveis categóricas e numéricas. Para isso, existe o teste Qui-quadrado, o qual mede a relação de dependência entre duas variáveis categóricas, verificando como os valores esperados diferem dos valores observados. Para determinar a estatística do teste de Qui-quadrado, determina-se inicialmente as frequências que seriam esperadas caso as variáveis fossem completamente independentes. Estas frequências são obtidas multiplicando as margens da tabela e dividindo pelo número total de observações. Em seguida, são calculadas as diferenças entre as frequências observadas e as esperadas e somam-se por meio da expressão (AKOGLU, 2018):

$$X^2 = \sum \frac{(O_i - E_i)^2}{E_i}$$

Na expressão acima, E_i representa a frequência esperada e O_i a observada. O teste Qui-quadrado já é um bom teste para calcular a correlação entre variáveis categóricas, contudo, esta medida não está limitada ao intervalo [0, 1] e o seu valor máximo depende do número total de observações (AKOGLU, 2018).

Como forma de contornar as limitações do teste Qui-quadrado existe o teste V de Cramér, ou coeficiente de Cramér. Esse coeficiente mostra uma medida de associação entre duas variáveis categóricas, e pode ser aplicado em situações onde a informação se encontra distribuída por categorias nominais e não ordenáveis. Uma vantagem em relação ao teste Qui-quadrado é que o valor retornado está limitado ao intervalo [0, 1], e ele é calculado da seguinte forma (AKOGLU, 2018):

$$C = \sqrt{\frac{X^2}{n(k-1)}}$$

Na expressão acima, X^2 representa o resultado obtido pelo teste Qui-quadrado, n representa o número total de observações e k o menor número de categorias de cada variável. Este trabalho adota o teste V de Cramér para analisar a correlação entre as variáveis, independente de serem categóricas ou numéricas, e assim selecionar um conjunto de atributos os quais mais representam o problema (AKOGLU, 2018).

2.4 Métricas para classificação

Métricas são importantes para avaliar um modelo, obtendo o classificador ideal durante o treinamento de classificação. O modelo pode fornecer melhores resultados quando avaliado usando uma métrica adequada. Algumas métricas são mais adequadas que outras para determinados casos, cada caso deve ser analisado de acordo com a necessidade (HOSSIN; SULAIMAN, 2015).

2.4.1 Acurácia

A Acurácia é a métrica de avaliação mais usada na prática para problemas de classificação binários ou multi-classe. Através dela, a qualidade da solução produzida é avaliada com base na porcentagem de previsões corretas sobre o total de instâncias. Essa métrica tem algumas vantagens, como ser fácil de calcular com menos complexidade, aplicável para problemas de multi-classe e fácil de entender por humanos. Entretanto a simplicidade dessa métrica pode levar a soluções sub-ótimas, especialmente quando se trata de distribuição de classes desequilibrada (HOSSIN; SULAIMAN, 2015).

A Acurácia é medida através do número de predições corretas dividido pelo número total de objetos (SAMMUT; WEBB, 2017), cuja expressão é dada por:

$$\text{acurácia} = \frac{\text{número de predições corretas}}{\text{número total de objetos}}$$

2.4.2 Matriz de Confusão

A Matriz de Confusão resume o desempenho de classificação de um classificador com relação a alguns dados de teste. Um caso em a Matriz de Confusão é frequentemente usada é com duas classes, uma usada como classe positiva e a outra como classe negativa. Dessa forma, as quatro células da matriz são formadas como verdadeiros positivos (VP), verdadeiros negativos (VN), falsos positivos (FP) e falsos negativos (FN) (SAMMUT; WEBB, 2017), conforme a Tabela 3:

Tabela 3 – Matriz de Confusão

	Positivo	**Negativo**
Positivo	VP	FN
Negativo	FP	VN

Fonte: (SAMMUT; WEBB, 2017).

Explicando a tabela 3 de acordo com os termos:

- Verdadeiros Positivos (VP): são os exemplos positivos que são classificados corretamente por um modelo de classificação.
- Verdadeiros Negativos (VN): são os exemplos negativos que são classificados corretamente por um modelo de classificação.
- Falsos Positivos (FP): é um exemplo de uma classe negativa que foi incorretamente classificado como positivo.
- Falsos Negativos (FN): é um exemplo de classe positiva que foi incorretamente classificada como negativa.

A acurácia total da matriz é medida através do número de verdadeiros positivos mais o número de verdadeiros negativos dividido pelo total de elementos da matriz (CARVALHO *et al.*, 2011), conforme expressão a seguir:

$$\text{acurácia total} = \frac{\text{número de verdadeiros positivos} + \text{número de verdadeiros negativos}}{\text{total de elementos da matriz}}$$

2.4.3 Precision

A métrica *Precision* é usada para medir os padrões positivos que são previstos corretamente a partir do total de padrões previstos em uma classe positiva. Ela é definida como a razão de verdadeiros positivos e o número total de positivos previstos por um modelo (HOSSIN; SULAIMAN, 2015), conforme abaixo:

$$\text{Precision} = \frac{\text{verdadeiros positivos}}{\text{verdadeiros positivos} + \text{falsos positivos}}$$

2.4.4 Recall

De acordo com (HOSSIN; SULAIMAN, 2015), a métrica *Recall* é usada para medir a fração de padrões positivos que são classificados corretamente. Ela é definida como a razão de verdadeiros positivos e o número de exemplos que são de fato positivos, conforme encontra-se a seguir:

$$\text{Recall} = \frac{\text{verdadeiros positivos}}{\text{verdadeiros positivos} + \text{falsos negativos}}$$

2.4.5 F1 Score

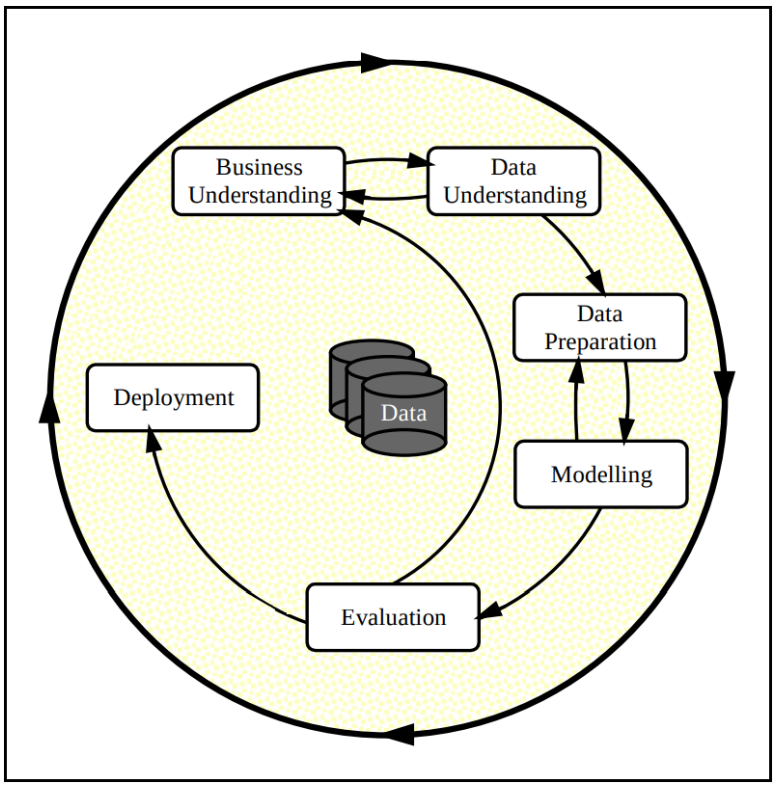
A métrica *F1 Score* é a média harmônica entre os valores das métricas *Precision* e *Recall*. Seu valor pode variar no intervalo [0,1], onde o mínimo é atingido para Verdadeiros Positivos = 0, ou seja, quando todas as amostras positivas são classificadas incorretamente, e o máximo para Falsos Negativos = Falsos Positivos = 0, ou seja, para classificação perfeita. Quanto maior for o *F1 Score*, melhor é o desempenho do modelo, segundo (CHICCO; JURMAN, 2020), cuja expressão é dada por:

$$\text{F1 Score} = 2 * \frac{precision * recall}{precision + recall}$$

3 METODOLOGIA

Este trabalho emprega a metodologia *Cross Industry Standard Process for Data Mining (CRISP-DM)*, por ser amplamente utilizada em projetos de Mineração e Ciência de Dados. Ela é dividida em seis etapas, vide Figura 1: Compreensão do problema, Interpretação dos dados, Preparação dos dados, Modelagem, Avaliação e Implantação/Publicação (WIRTH; HIPP, 2000). Os detalhes de cada uma dessas etapas, aplicadas neste trabalho, são descritos na Figura 1.

Figura 1 – Fases da metodologia de processo CRISP-DM



Fonte: (WIRTH; HIPP, 2000).

3.1 Compreensão do Problema

Inicialmente foram identificados os objetivos centrais deste projeto. Para isso, foram realizadas uma série de análises sobre os dados do portal Consumidor.gov.br. Observou-se que era possível predizer, utilizando algoritmos de Aprendizagem de Máquina, informações sobre reclamações do portal em questão. Particularmente, este trabalho busca predizer os seguintes cenários:

- **C1.** Se uma reclamação do consumidor, contra uma determinada empresa, será atendida.
- **C2.** Se a empresa irá recusar a reclamação do consumidor antes do atendimento.
- **C3.** Caso a empresa recuse, se a recusa é procedente ou improcedente.

3.2 Interpretação dos dados

Todos os dados utilizados nesta pesquisa foram retirados do site dados.gov.br. O site disponibiliza vários arquivos .csv com os dados das reclamações ao longo dos anos, desde 2014. Foram usadas as reclamações de 2014 a setembro de 2021 (SENACON/MJ, 2021). O Apêndice A mostra cada atributo do conjunto de dados, a descrição e o tipo de cada atributo.

3.3 Preparação dos dados

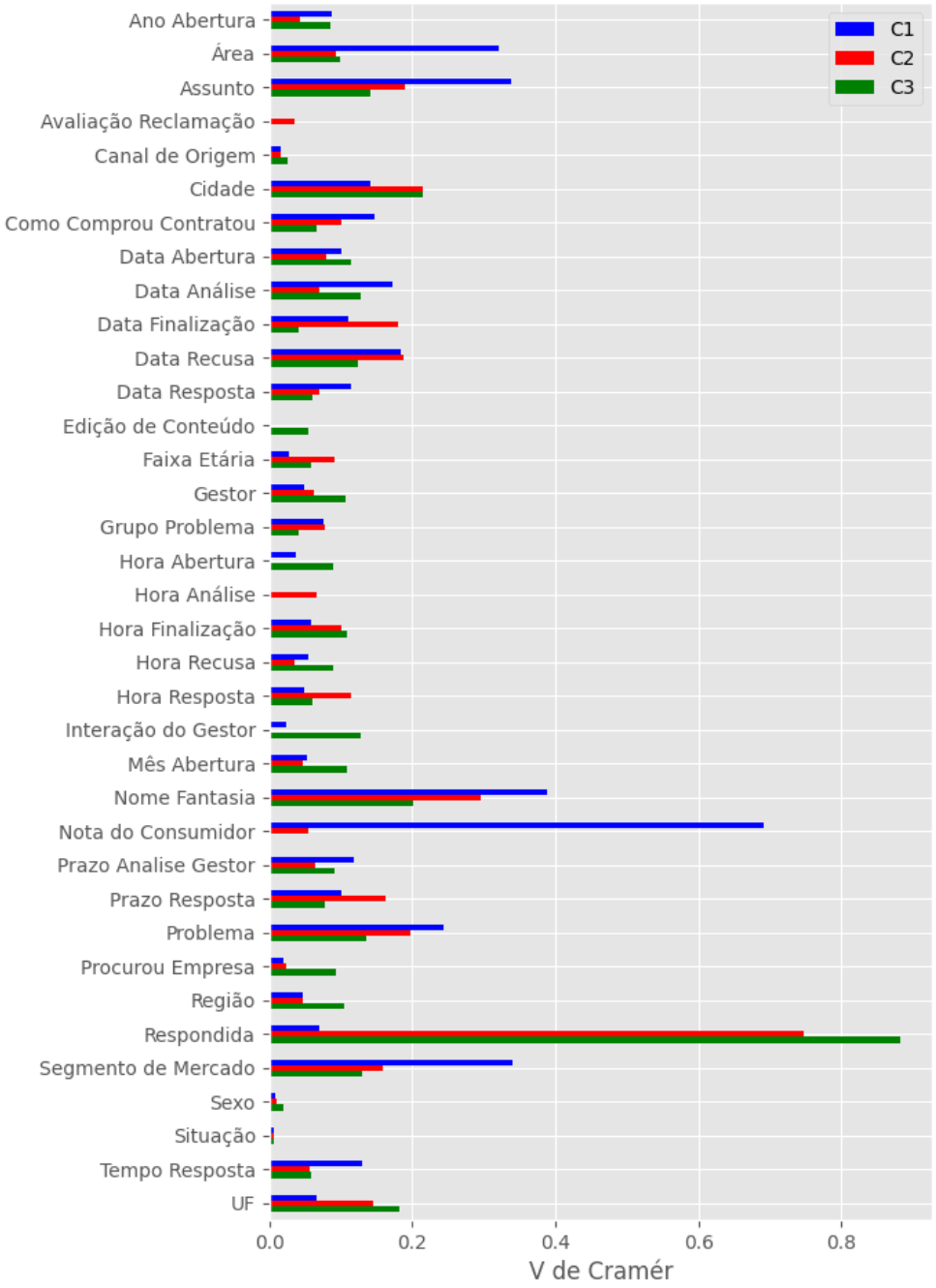
Inicialmente todos os dados coletados foram concatenados em um só arquivo de extensão *Comma Separated Values* (CSV), preservando suas respectivas colunas correspondentes. Para isso, a codificação de caracteres de alguns arquivos foi tratada. Os aquivos de 2014 até setembro de 2020 possuíam codificação de caracteres ISO-8859-1, em sua maioria, com exceção dos meses de abril, maio, junho, julho e setembro de 2019. Esses últimos, além dos arquivos entre outubro de 2020 a setembro de 2021, possuíam codificação UTF-8.

Posteriormente, alguns atributos do tipo texto foram convertidos em atributos numéricos. Os atributos convertidos foram: Avaliação Reclamação, Análise da Recusa, Procurou Empresa, Respondida e Avaliação Reclamação. Dessa forma, para os atributos que apresentavam valor "S"(SIM), foi atribuído o valor 1, e, para os atributos que apresentavam valor "N"(NÃO), foi atribuído o valor 0. Para o campo Avaliação Reclamação, em particular, foi atribuído o valor 1 para as reclamações com valor "Resolvida", e valor 0 para reclamações com valor "Não Resolvida".

Adicionalmente, foi criado um campo denominado Recusa, para identificar as reclamações recusadas pelas empresas. Dessa forma, caso uma reclamação tenha sido recusada, o valor 1 é atribuído a esse campo, e 0, caso contrário. Já para o campo "Análise da Recusa", foi atribuído o valor 0 para recusas improcedentes e 1 para recusas procedentes, isto é, quando um gestor do portal Consumidor.gov.br acata que a empresa irá recusar uma reclamação do consumidor.

Em seguida, para cada cenário deste trabalho, foi calculada a correlação V de Cramér em relação a variável na qual se desejava predizer. A Figura 2 mostra o grau de correlação entre as diversas variáveis do conjunto de dados em relação às variáveis "Avaliação Reclamação"(Problema referente a C1), "Recusa"(Problema referente a C2) e "Análise da Recusa"(Problema referente a C3), respectivamente.

Figura 2 – Grau de correlação V de Cramér entre as diversas variáveis do conjunto de dados em relação a C1, C2 e C3.



Fonte: Autoria própria.

Com as devidas correlações calculadas, foram selecionados os atributos que fariam parte do treinamento dos modelos, para cada cenário de predição. Para a predição de atendimento de reclamação do consumidor, foram selecionados os cinco atributos de maior correlação com o campo "Avaliação Reclamação", resultando nos seguintes atributos: "Nota do Consumidor", "Nome Fantasia", "Segmento de Mercado", "Assunto"e "Área", além do próprio atributo "Avaliação Reclamação".

Para a predição de Recusa, isto é, se a empresa irá recusar a reclamação do consumidor antes do atendimento, foram selecionados os seguintes atributos: "Respondida", "Cidade", "Nome Fantasia", "Assunto"e "Problema", além do próprio atributo "Recusa".

Já para a predição da recusa ser procedente ou improcedente, também foram selecionados os mesmos atributos do problema anterior: "Respondida", "Cidade", "Nome Fantasia", "Assunto"e "Problema", além do próprio atributo "Análise da Recusa", ou seja, os cinco atributos com maior correlação segundo a métrica V de Crámer. Vale ressaltar que, apesar do campo "UF"possuir uma correlação entre as cinco maiores com a variável "Análise da Recusa", com valor 0,1818, ele foi descartado, pois o campo "Cidade"já está sendo utilizado, isto é, uma generalização do campo "UF". Amostras dos conjuntos de dados para cada contexto, sendo as cinco primeiras variáveis, variáveis de entrada, e a última variável sendo a variável de saída (a ser predita), são exibidas nas Figuras 3, 4, 5, respectivamente para C1, C2 e C3.

Figura 3 – Amostra de dados utilizados para predição de atendimento (C1).

	Nota do Consumidor	Nome Fantasia	Segmento de Mercado	Assunto	Área	Avaliação Reclamação
0	1	Magazineluiza.com	Comércio Eletrônico	Aparelho celular	Produtos de Telefonia e Informática	0
1	5	Banco Santander Cartões	Bancos, Financeiras e Administradoras de Cartão	Cartão de Crédito / Cartão de Débito / Cartão ...	Serviços Financeiros	1
2	5	Hipercard	Bancos, Financeiras e Administradoras de Cartão	Cartão de Crédito / Cartão de Débito / Cartão ...	Serviços Financeiros	1
3	3	Philips TV e Monitores	Fabricantes - Eletroeletrônicos, Produtos de ...	Acessórios e periféricos (monitor, impressora,...	Produtos de Telefonia e Informática	1
4	1	Claro Fixo - Embratel	Operadoras de Telecomunicações (Telefonia, Int...	Telefonia Fixa	Telecomunicações	0
...	...	...	...	...	...	...
2561890	5	Globoplay	Provedores de Conteúdo e Outros Serviços na In...	Serviços na internet (provedor, hospedagem, ap...	Demais Serviços	1
2561891	5	Azul Linhas Aéreas	Transporte Aéreo	Aéreo	Transportes	1
2561892	5	Azul Linhas Aéreas	Transporte Aéreo	Aéreo	Transportes	1
2561893	3	LG Electronics	Fabricantes - Eletroeletrônicos, Produtos de ...	Televisão	Produtos Eletrodomésticos e Eletrônicos	0
2561894	1	Neoenergia Coelba	Energia Elétrica	Energia Elétrica	Água, Energia, Gás	0

2561895 rows × 6 columns

Fonte: Autoria própria.

Figura 4 – Amostra de dados utilizados para predição de recusa pela empresa (C2).

	Respondida	Nome Fantasia	Cidade	Problema	Assunto	Recusada
0	1	Hipercard	Itaboraí	Cobrança de compra/saque não reconhecido	Cartão de Crédito / Cartão de Débito / Cartão ...	0
1	1	Magazineluiza.com	Serra	Não entrega / demora na entrega do produto	Aparelho celular	0
2	1	Banco Santander Cartões	Volta Redonda	Dificuldade / atraso na devolução de valores p...	Cartão de Crédito / Cartão de Débito / Cartão ...	0
3	0	Hipercard	São Luís	Dados pessoais ou financeiros consultados, col...	Cartão de Crédito / Cartão de Débito / Cartão ...	0
4	1	American Express - Amex	Recife	Cobrança por serviço/produto não contratado / ...	Cartão de Crédito / Cartão de Débito / Cartão ...	0
...	...	...	...	...	...	...
4336621	0	Sanepar	Curitiba	Cobrança de tarifas, taxas, valores não previs...	Água e Esgoto	1
4336622	1	Faculdade FAMA	Florianópolis	Má qualidade no atendimento (descortesia / des...	Superior (Graduação e Pós)	1
4336623	1	Facebook / Instagram	Uberlândia	Funcionamento inadequado do serviço	Serviços na internet (provedor, hospedagem, ap...	1
4336624	0	Sanepar	Curitiba	Cobrança de tarifas, taxas, valores não previs...	Água e Esgoto	1
4336625	1	Neoenergia Coelba	Salvador	Cobrança por Irregularidade ou defeito na medição	Energia Elétrica	0

4336626 rows × 6 columns

Fonte: Autoria própria.

Figura 5 – Amostra de dados utilizados para predição da procedência da recusa (C3).

	Respondida	Cidade	Nome Fantasia	Assunto	Problema	Análise da Recusa
0	1	São Paulo	Magazineluiza.com	Ar condicionado e aquecedor	Produto entregue incompleto / diferente do pe...	0
1	1	Santa Rita de Cássia	Banco BMG	Crédito Consignado (Empréstimo descontado em f...	Dificuldade para obter boleto de quitação ou i...	0
2	1	Curitiba	Magazineluiza.com	Vestuário e Artigos de Uso Pessoal (roupa, cal...	Não entrega / demora na entrega do produto	0
3	1	Belo Horizonte	Magazineluiza.com	Acessórios e periféricos (monitor, impressora,...	Oferta não cumprida / serviço não fornecido/ v...	0
4	1	Guarapuava	Itaú Consórcio	Consórcios (exceto imóveis)	Cobrança de tarifas, taxas, valores não previs...	0
...	...	...	...	...	...	...
439436	0	Curitiba	Itaucard	Cartão de Crédito / Cartão de Débito / Cartão ...	Cobrança de tarifas, taxas, valores não previs...	1
439437	0	Curitiba	Sanepar	Água e Esgoto	Cobrança de tarifas, taxas, valores não previs...	1
439438	0	Florianópolis	Faculdade FAMA	Superior (Graduação e Pós)	Má qualidade no atendimento (descortesia / des...	1
439439	0	Uberlândia	Facebook / Instagram	Serviços na internet (provedor, hospedagem, ap...	Funcionamento inadequado do serviço	1
439440	0	Curitiba	Sanepar	Água e Esgoto	Cobrança de tarifas, taxas, valores não previs...	1

439441 rows × 6 columns

Fonte: Autoria própria.

Após seleção de atributos, registros com valores incompletos foram removidos dos seus respectivos conjuntos de dados. Essa etapa só foi realizada posteriormente a seleção de atributos para não haver perda de informação desnecessária, isto é, remoção de registros incompletos contendo campos que poderiam ser utilizados na criação dos modelos. Por exemplo, "Segmento de Mercado"é um campo que contém vários valores nulos. Ele é utilizado apenas

para a predição de C1. Consequentemente, registros contendo valores nulos de segmentos de mercado foram removidos apenas do conjunto de dados referente a C1. Assim, o número de registros utilizados para cada um dos contextos, bem como a porcentagem de registros de cada classe alvo, são mostrados na Tabela 4.

Tabela 4 – Número de registros e porcentagem de cada classe alvo por contexto.

Contexto do Problema	Número de Registros	% de Registros da Classe 0	% de Registros da Classe 1
C1	2.209.145	37,29%	62,97%
C2	4.336.626	91,62%	8,38%
C3	439.441	16,3%	83,7%

Fonte: Autoria própria.

Dessa forma, para C1, a classe 0 representa reclamações "Não Atendidas"e a classe 1 representa reclamações "Atendidas". Já para C2, a classe 0 reflete reclamações "Não Recusadas", enquanto a classe 1 reflete as reclamações "Recusadas". Por fim, para C3, a classe 0 significa que a recusa por parte da empresa é improcedente, enquanto a classe 1 quer dizer procedente.

Anteriormente à modelagem do problema, foi aplicada a técnica de *One Hot Encoding* para transformar os dados categóricos, selecionados na etapa anterior, em dados numéricos. A escolha desse método de codificação se deu devido a maior parte dos atributos selecionados para cada contexto ser do tipo categórico, por exemplo: "Nome Fantasia", "Segmento de Mercado", "Assunto", "Área", entre outros. Além disso, essas categorias não possuem um relacionamento ordinal, para se aplicar, por exemplo, uma codificação do tipo *Label Encoding*. Consequentemente, ao utilizar *One Hot Encoding*, espera-se evitar um baixo desempenho do algoritmo de Aprendizado de Máquina aplicado. Em contrapartida, pode-se aumentar consideravelmente o número de colunas geradas.

3.4 Modelagem

A etapa de modelagem consistiu na aplicação do algoritmo de aprendizagem *Random Forest*, por ser amplamente utilizado para tratar problemas de classificação em várias áreas distintas, como bancos, mercado de ações, medicina, comércio eletrônico e muitas outras. Além disso, um amplo estudo mostrou que não são necessários vários classificadores para

resolver problemas do mundo do mundo real (FERNÁNDEZ-DELGADO *et al.*, 2014), pois algoritmos baseados em florestas, como o *Random Forest*, classificam muito bem dados como nos contextos citados acima. Ainda assim, um estudo recente (SILVA, 2021) mostrou que, para dados semelhantes aos deste trabalho, o algoritmo *Random Forest* performou muito bem quando comparado aos algoritmos *Gradient Boosting*, *K-nearest neighbors* (KNN), *Regressão Logística* e *Naive Bayes*.

Por fim, para cada contexto deste trabalho, os conjuntos de dados foram divididos em treino e teste, utilizando a proporção 90% e 10%, respectivamente. A Tabela 5 sumariza o número de registros de treino e teste para cada contexto analisado.

Tabela 5 – Número de registros de treino e teste por contexto.

Contexto do Problema	**Número de Registros para Treino (90%)**	**Número de Registros para Teste (10%)**
C1	1.988.231	220.914
C2	3.902.963	433.663
C3	395.496	43.945

Fonte: Autoria própria.

A fase de avaliação do modelo CRISP-DM será apresentada no capítulo seguinte.

4 RESULTADOS

Este trabalho foi desenvolvido utilizando a linguagem de programação Python, sistema operacional Linux Ubuntu versão 20.04, 64 bits, 8 cores de CPU, memória RAM de 64GB e armazenamento de dados em SSD. Devido ao grande volume de dados, foi necessário utilizar uma máquina com mais memória RAM.

Os resultados deste trabalho, para cada contexto avaliado, em termos de Acurácia, *Precision*, *Recall* e *F1 Score*, são mostrados na Tabela 6.

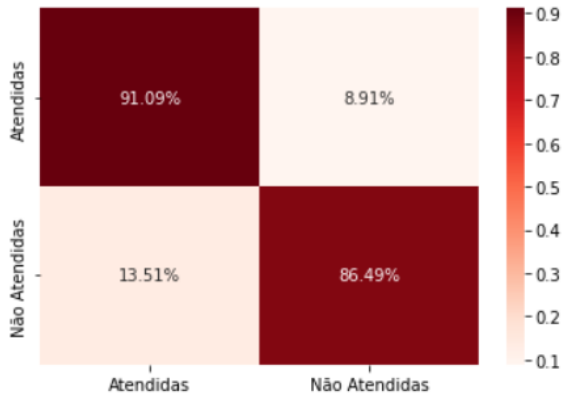
Tabela 6 – Resultados referentes a Acurácia, *Precision*, *Recall* e *F1 Score* para cada contexto.

Contexto do Problema	Acurácia	Precision	Recall	F1 Score
C1	86,39%	89,22%	87,14%	88,72%
C2	94,87%	87,55%	79,21%	84,48%
C3	97,31%	98,72%	98,96%	98,89%

Fonte: Autoria própria.

Em particular, para o contexto 1, no qual se buscava predizer se uma determinada reclamação do consumidor, contra uma determinada empresa, seria atendida, o modelo atingiu uma acurácia de 86,39%. Isto é, o modelo proposto acerta mais de 85% das vezes se uma reclamação será atendida ou não. A Figura 6 mostra a matriz de confusão para esse contexto.

Figura 6 – Matriz de confusão para predição de atendimento (C1).



Fonte: Autoria própria.

Observa-se que, para as reclamações "Atendidas", por se tratar de 62,97% das reclamações, o modelo acerta 91,09% das vezes. Por outro lado, para as reclamações "Não Atendidas", o modelo acerta 86,49% das vezes.

Já para o contexto 2, no qual se buscava predizer se a empresa iria recusar a reclamação do consumidor antes do atendimento, o modelo atingiu uma acurácia de 94,87%. Valores de *Precision*, *Recall* e *F1 Score* para esse contexto foram respectivamente 87,55%, 79,21% e 84,48%. A Figura 7 mostra a matriz de confusão para esse contexto.

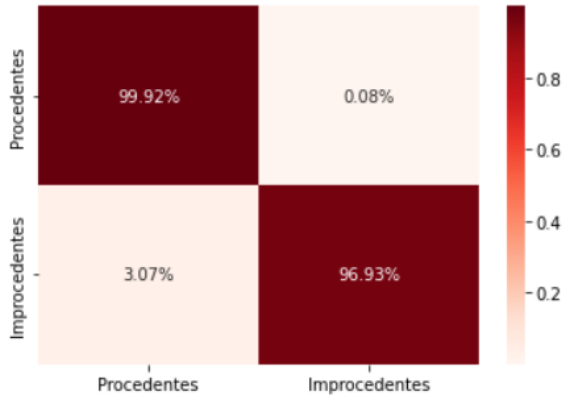
Figura 7 – Matriz de confusão para predição de recusa pela empresa (C2).

Fonte: Autoria própria.

Observa-se que, para as reclamações "Não Recusadas"pelas empresas, por representar mais de 90% das reclamações, o modelo acerta quase sempre, isto é, em 99,29% das vezes. Já para as reclamações "Não Atendidas", o modelo acerta bem menos, em 82,45% das vezes.

Por fim, para o contexto 3, no qual se buscava predizer se as recusas pela empresa eram procedentes ou improcedentes, o modelo atingiu uma acurácia ainda maior, de 97,31%. Valores de *Precision*, *Recall* e *F1 Score* para esse contexto foram respectivamente 98,72%, 98,96% e 98,89%. A Figura 8 mostra a matriz de confusão para o contexto em questão.

Figura 8 – Matriz de confusão para predição da procedência da recusa (C3).



Fonte: Autoria própria.

Pela matriz de C3 é possível ver que em mais de 95% das vezes, o modelo proposto acerta se as recusas pelas empresas foram procedentes ou improcedentes. Em particular, o modelo alcança 99,92% de acerto para as recusas procedentes, e 96,93% para as recusas improcedentes.

5 CONCLUSÃO E TRABALHOS FUTUROS

Neste trabalho foi apresentada uma abordagem para classificar automaticamente reclamações abertas na plataforma Consumidor.gov.br, um serviço do Ministério da Justiça que auxilia na defesa do consumidor. Reclamações entre consumidor e empresa são cadastradas na plataforma para que possam ser solucionados rapidamente e da melhor forma possível. Essas reclamações podem gerar alguns casos indesejados, como a reclamação não ser atendia, ou a reclamação ser recusada pela empresa, que pode ser ratificada como procedente ou improcedente. Essas questões foram analisadas utilizando aprendizado de máquina para auxiliar o consumidor a ter uma previsão do que pode acontecer com sua reclamação mesmo sem abri-la diretamente na plataforma Consumidor.gov.br.

Após diversas análises, três contextos foram escolhidos buscando prevê-los : C1: Se uma reclamação do consumidor, contra uma determinada empresa, será atendida; C2: Se a empresa irá recusar a reclamação do consumidor antes do atendimento; C3: Caso a empresa recuse, se a recusa é procedente ou improcedente.

Para o contexto 1 o modelo atingiu uma Acurácia de 86,39%, Precision de 89,22%, Recall de 87,14%, e F1 Score de 88,72% para prever se uma reclamação será ou não atendida. No contexto 2 o modelo atingiu um Acurácia de 94,87%, Precision de 87,55%, Recall de 79,21%, e F1 Score de 84,48% para prever se a empresa irá recusar a reclamação. Já no contexto 3 o modelo atingiu uma Acurácia de 97,31%, Precision de 98,72%, Recall de 98,96%, e F1 Score de 98,89% para prever se a a reclamação que foi recusada será classificada como procedente ou improcedente. Tais valores são considerados muito bons na literatura, e, em termos de acurácia, todos atingiram mais de 85% de acerto.

Como trabalhos futuros, pretende-se desenvolver uma ferramenta para que o consumidor possa cadastrar sua futura reclamação, e já verificar as predições analisadas, possibilitando ele tomar a decisão de prosseguir ao cadastro no Consumidor.gov.br ou buscar outro serviço prestado pelos órgãos de defesa do consumidor, como o PROCON.

Adicionalmente, planeja-se analisar e tratar os casos de desbalanceamento de registros, principalmente nos contextos C2 e C3. Esses contextos tiveram muitos registros a mais em uma das classes analisadas e resultados bastante significativos. Contudo, na prática, isso

pode levar a uma predição não assertiva, visto que classificar ingenuamente um registro com sua classe de maior ocorrência também trás resultados significativos, mas não garante um bom classificador.

Por fim,espera-se analisar a troca de mensagens propriamente dita nas reclamações, levando a uma eventual análise de sentimento por parte do consumidor, e assim inferir seu grau de satisfação e a resolução da reclamação também de maneira automática.

REFERÊNCIAS

AKOGLU, H. User's guide to correlation coefficients. **Turkish journal of emergency medicine**, Elsevier, v. 18, n. 3, p. 91–93, 2018.

BROWNLEE, J. **Train-Test Split for Evaluating Machine Learning Algorithms**. 2021. Disponível em: https://machinelearningmastery.com/train-test-split-for-evaluating-machine-learning-algorithms/. Acesso em: 25 de novembro 2021.

CARVALHO, A.; FACELI, K.; LORENA, A.; GAMA, J. **Inteligência Artificial–uma abordagem de aprendizado de máquina**. [S.l.]: LTC, 2011.

CHICCO, D.; JURMAN, G. The advantages of the matthews correlation coefficient (mcc) over f1 score and accuracy in binary classification evaluation. **BMC genomics**, Springer, v. 21, n. 1, p. 1–13, 2020.

CONSUMIDOR.GOV.BR. **Conheça o Consumidor.gov.br**. 2021. Disponível em: https://www.consumidor.gov.br/pages/conteudo/publico/1. Acesso em: 02 de novembro 2021.

CONSUMIDOR.GOV.BR. **Sobre o Serviço**. 2021. Disponível em: https://consumidor.gov.br/pages/conteudo/sobre-servico. Acesso em: 02 de novembro 2021.

CONSUMIDOR.GOV.BR. **Termos de Uso Consumidor.gov.br**. 2021. Disponível em: https://www.consumidor.gov.br/pages/conteudo/publico/7. Acesso em: 13 de novembro 2021.

FERNANDES, C. M.; FILHO, A. S. A proteção do consumidor na sociedade da informação: uma análise da plataforma consumidor. gov. br. In: **Anais do Congresso Brasileiro de Processo Coletivo e Cidadania**. [S.l.: s.n.], 2015. p. 467–474.

FERNÁNDEZ-DELGADO, M.; CERNADAS, E.; BARRO, S.; AMORIM, D. Do we need hundreds of classifiers to solve real world classification problems? **The journal of machine learning research**, JMLR. org, v. 15, n. 1, p. 3133–3181, 2014.

HALL, M. A. Correlation-based feature selection for machine learning. University of Waikato Hamilton, 1999.

HARRINGTON, P. **Machine learning in action**. [S.l.]: Simon and Schuster, 2012.

HOSSIN, M.; SULAIMAN, M. A review on evaluation metrics for data classification evaluations. **International Journal of Data Mining & Knowledge Management Process**, Academy & Industry Research Collaboration Center (AIRCC), v. 5, n. 2, p. 1, 2015.

JIANG, D.; LIN, W.; RAGHAVAN, N. A novel framework for semiconductor manufacturing final test yield classification using machine learning techniques. **IEEE Access**, IEEE, v. 8, p. 197885–197895, 2020.

MONARD, M. C.; BARANAUSKAS, J. A. Conceitos sobre aprendizado de máquina. **Sistemas inteligentes-Fundamentos e aplicações**, Manole Ltda, v. 1, n. 1, p. 32, 2003.

RUSSELL, S. J.; NORVIG, P. **Inteligência artificial**. [S.l.]: Elsevier, 2004.

SAMMUT, C.; WEBB, G. I. **Encyclopedia of machine learning and data mining**. [S.l.]: Springer Publishing Company, Incorporated, 2017.

SENACON/MJ. **Dicionário de Dados Consumidor.gov.br**. 2020. Disponível em: http://dados.mj.gov.br/dataset/0182f1bf-e73d-42b1-ae8c-fa94d9ce9451/resource/90aedbfe-3c91-4c18-86a5-f408d07e7210/download/dicionario-de-dados—consumidorgovbr-v3.pdf. Acesso em: 13 de novembro 2021.

SENACON/MJ. **Dados Consumidor.gov.br**. 2021. Disponível em: https://dados.gov.br/dataset/reclamacoes-do-consumidor-gov-br1. Acesso em: 15 de novembro 2021.

SILVA, T. Um estudo comparativo entre algoritmos de aprendizagem de máquina supervisionados para predição de solução de reclamações no procon. Centro Universitário Christus-Unichristus, 2021.

THOMAS, R.; JUDITH, J. A novel ensemble method for detecting outliers in categorical data. **International Journal**, v. 9, n. 4, 2020.

WIRTH, R.; HIPP, J. Crisp-dm: Towards a standard process model for data mining. In: SPRINGER-VERLAG LONDON, UK. **Proceedings of the 4th international conference on the practical applications of knowledge discovery and data mining**. [S.l.], 2000. v. 1.

APÊNDICE A – CONJUNTO DE DADOS DO SERVIÇO CONSUMIDOR.GOV.BR

Tabela 7 – Conjunto de dados do serviço Consumidor.gov.br

Atributo	Descrição	Tipo
Gestor	Nome da entidade de defesa do consumidor, responsável pela gestão da reclamação do consumidor	string
Canal de Origem	Origem do registro no consumidor.gov.br	string
Região	Sigla da região geográfica do consumidor reclamante	string
UF	Sigla do estado do consumidor reclamante	string
Cidade	Município do consumidor reclamante	string
Sexo	Sigla do sexo do consumidor reclamante	string
Faixa Etária	Faixa etária do consumidor	string
Ano Abertura	Ano de abertura da reclamação pelo consumidor	inteiro
Mês Abertura	Número do mês de abertura da reclamação pelo consumidor	inteiro
Data Abertura	Data de abertura da reclamação pelo consumidor	string (formato dd/mm/aaaa)
Data Resposta	Data de resposta da reclamação pela empresa	string (formato dd/mm/aaaa)
Data Análise	Data da análise da recusa da empresa pelo Gestor da reclamação	string (formato dd/mm/aaaa)
Data Recusa	Data de registro da recusa pela empresa e envio da reclamação para análise do Gestor	string (formato dd/mm/aaaa)
Data Finalização	Data de finalização da reclamação	string (formato dd/mm/aaaa)
Prazo Resposta	Data limite para resposta da empresa. Caso a reclamação tenha sido recusada pela empresa e encaminhada para análise do Gestor, o prazo se altera, considerando o tempo que a reclamação tenha ficado em análise pelo gestor	string (formato dd/mm/aaaa)

Fonte: (SENACON/MJ, 2020).

Tabela 8 – Conjunto de dados do serviço Consumidor.gov.br (continuação)

Atributo	Descrição	Tipo
Prazo Análise Gestor	Número de dias levados para a análise da recusa da empresa pelo Órgão Gestor da reclamação. Diferença, em dias, entre a Data Análise e a Data Recusa	string (mas somente números inteiros)
Tempo Resposta	Número de dias para a resposta da reclamação, entre a Data de Resposta e a Data de Abertura, desconsiderado o tempo que a reclamação tenha ficado em análise pelo Gestor (se for o caso)	string (mas somente números inteiros)
Nome Fantasia	Nome pelo qual a empresa reclamada é conhecida no mercado	string
Segmento de Mercado	Principal segmento de mercado da empresa participante	string
Área	Área à qual pertence o assunto objeto da reclamação	string
Assunto	Assunto objeto da reclamação	string
Grupo Problema	Agrupamento do qual faz parte o problema classificado na reclamação	string
Problema	Descrição do problema objeto da reclamação	string
Como Comprou Contratou	Descrição do meio utilizado para contratação/aquisição do produto ou serviço reclamado	string
Procurou Empresa	Sigla da resposta do consumidor à pergunta: "Procurou a empresa para solucionar o problema?"	string
Respondida	Sigla que indica se a empresa respondeu à reclamação ou não	string
Situação	Situação atual da reclamação no sistema	string
Avaliação Reclamação	Classificação atribuída pelo consumidor sobre o desfecho da reclamação	string
Nota do Consumidor	Número da nota de 1 a 5 atribuída pelo consumidor ao atendimento da empresa	string (mas somente números inteiros)
Análise da Recusa	Resultado da Análise da Recusa pelo Gestor, se procedente (aceita) ou improcedente (indeferida ou não aceita)	string

Fonte: (SENACON/MJ, 2020).