UNIVERSIDADE ESTADUAL DA PARAÍBA
CAMPUS I - CAMPINA GRANDE
CENTRO DE CIÊNCIAS E TECNOLOGIA
DEPARTAMENTO DE COMPUTAÇÃO
CURSO DE GRADUAÇÃO EM CIÊNCIA DA COMPUTAÇÃO

KENNEDY JOHNSON DE SOUSA DANTAS

IDENTIFICAÇÃO DE PAUTA FISCAL PARA PRODUTOS DE MADEIRA UTILIZANDO EXPRESSÕES REGULARES

CAMPINA GRANDE
2023

KENNEDY JOHNSON DE SOUSA DANTAS

IDENTIFICAÇÃO DE PAUTA FISCAL PARA PRODUTOS DE MADEIRA UTILIZANDO EXPRESSÕES REGULARES

Trabalho de Conclusão de Curso de Graduação em Ciência da Computação da Universidade Estadual da Paraíba, como requisito à obtenção do título de Bacharel em Ciência da Computação.

Área de concentração: Inteligência Artificial.

Orientadora: Profª. Drª. Kézia de Vasconcelos Oliveira Dantas.

CAMPINA GRANDE
2023

D192i Dantas, Kennedy Johnson de Sousa.
Identificação de pauta fiscal para produtos de madeira utilizando expressões regulares [manuscrito] / Kennedy Johnson de Sousa Dantas. - 2023.
29 p. : il. colorido.

Digitado.
Trabalho de Conclusão de Curso (Graduação em Computação) - Universidade Estadual da Paraíba, Centro de Ciências e Tecnologia, 2023.
"Orientação : Profa. Dra. Kézia de Vasconcelos Oliveira Dantas, Coordenação do Curso de Computação - CCT. "

1. Ciência de dados. 2. Expressões regulares. 3. Pauta fiscal. I. Título

21. ed. CDD 629.895

Elaborada por Talita M. A. Tavares - CRB - CRB 15/971 BC/UEPB

KENNEDY JOHNSON DE SOUSA DANTAS

IDENTIFICAÇÃO DE PAUTA FISCAL PARA PRODUTOS DE MADEIRA UTILIZANDO EXPRESSÕES REGULARES

Trabalho de Conclusão de Curso de Graduação em Ciência da Computação da Universidade Estadual da Paraíba, como requisito à obtenção do título de Bacharel em Ciência da Computação.

Área de concentração: Inteligência Artificial.

Aprovada em 03 de Julho de 2023.

BANCA EXAMINADORA

Kézia de Vasconcelos Oliveira Dantas

Profa. Dra. Kézia de Vasconcelos O. Dantas (CCT/UEPB)
Orientador (a)

Paulo Barbosa

Prof. Dr. Paulo Eduardo e Silva Barbosa (CCT/UEPB)
Examinador (a)

Sabrina de F. Souto.

Profa. Dra. Sabrina de Figueiredo Souto (CCT/UEPB)
Examinador (a)

Dedico este TCC aos meus pais, meus amigos e minha namorada, pois sempre me apoiaram para que eu pudesse dar o meu melhor.

AGRADECIMENTOS

Agradeço a Deus, por ter me guiado para escolher o curso de Computação, no qual, encontrei a minha vocação, assim como, por me conceder força nos momentos difíceis, sempre ao meu lado em meio às adversidades. Logo, obrigado por tudo e por tanto, Senhor.

Sou grato a minha família, em especial, aos meus pais, Maria do Carmo e Lindonjonso, e à minha irmã Kédynna Carmem, que sempre me deram apoio e condições para que eu pudesse dar o meu melhor. Por isso, agradeço de coração a toda minha família, a base da minha vida.

Agradeço a minha namorada Andressa Lucena, principalmente, por ser tão presente durante o período acadêmico, me dando sustentação e incentivo nos momentos mais árduos e dividindo as alegrias cotidianas, sendo assim, um dos pilares da minha vida.

Grato aos amigos que fiz durante a graduação: Ângelo Gabriel, Jefferson Gomes, Klayton Marcos, Renan Rey, Roberto Pereira, Rodolfo Pereira, Mariana Ramos e Natalia Maria, com quem pude dividir tantos momentos especiais ao longo do curso, como também aos amigos de longa data: Phelipe Lacerda e Susana Arruda, pois estão comigo por quase toda minha vida.

Por fim, agradeço à minha orientadora Kézia Dantas, pela disponibilidade e orientação ao longo do TCC e do projeto do NUTES, como também, aos professores Paulo Eduardo e Sabrina Souto pela oportunidade de trabalhar no NUTES, onde pude aprender e me tornar um profissional mais completo, e, por último, a todos os colegas de trabalho e universidade.

RESUMO

As Secretarias de Estado da Fazenda (SEFAZ) controlam a maior parte dos tributos estaduais do Brasil. Na Paraíba, a SEFAZ enfrenta diversos desafios na cobrança de notas fiscais para arrecadar tributos estaduais, destacando-se, o grande volume de dados a serem classificados manualmente, foco deste trabalho. O artigo discute a criação e produção de um classificador, através do NUTES em parceria com a SEFAZ-PB, que utiliza ciência de dados e expressões regulares para analisar as declarações de produtos e identificar sua pauta fiscal. Nesse contexto, o atual trabalho aplica-se após a utilização de um primeiro classificador, que determina a categoria do produto, de forma que, o sistema ao qual se refere este projeto diz respeito a uma segunda etapa de classificação, quando é definida a pauta fiscal do produto. Sendo assim, o objetivo deste documento é demonstrar o processo de automatização, por meio de um classificador, das cobranças fiscais dos itens de madeira comercializados na Paraíba. Esse processo de classificação foi feito apesar de descrições informais realizadas por contribuintes do estado, de forma a solucionar os problemas enfrentados na arrecadação de impostos pelos auditores fiscais da SEFAZ da Paraíba. Com isso, concluiu-se que o classificador apresentou uma acurácia de cerca de 98%, sendo capaz de classificar corretamente a grande maioria das descrições analisadas. O projeto citado já está em uso na Secretaria da Fazenda da Paraíba e culminou em uma redução da carga de trabalho dos auditores fiscais do estado.

Palavras-chave: ciência de dados; expressões regulares; classificador; pauta fiscal.

ABSTRACT

The State Departments of Finance (SEFAZ) control the majority of state taxes in Brazil. In Paraíba, SEFAZ faces several challenges in collecting invoices to collect state taxes, with a notable challenge being the large volume of data that needs to be manually classified, which is the focus of this work. The article discusses the creation and production of a classifier by NUTES in partnership with SEFAZ-PB, which utilizes data science and regular expressions to analyze product declarations and identify their tax classification. In this context, the current work applies after the utilization of an initial classifier that determines the product category. Therefore, the system referred to in this project pertains to a second stage of classification, where the tax classification of the product is determined. As such, the objective of this document is to demonstrate the process of automation through a classifier for the taxation of wood items traded in Paraíba. This classification process was conducted despite informal descriptions provided by taxpayers in the state in order to address the challenges faced in tax collection by tax auditors at SEFAZ Paraíba. It was concluded that the classifier achieved an accuracy of approximately 98%, correctly classifying the vast majority of the analyzed descriptions. The mentioned project is already in use at the Department of Finance of Paraíba and has resulted in a reduction in the workload of the state's tax auditors.

Keywords: data science; regular expressions; classifier; fiscal agenda.

LISTA DE ILUSTRAÇÕES

LISTA DE TABELAS

LISTA DE ABREVIATURAS E SIGLAS

NFE	Nota Fiscal Eletrônica
SEFAZ	Secretaria de Estado da Fazenda
REGEX	Expressões Regulares
NUTES	Núcleo de Tecnologias Estratégicas em Saúde

SUMÁRIO

1 INTRODUÇÃO

As Secretarias de Estado da Fazenda (SEFAZ) controlam a maior parte dos tributos estaduais do Brasil e fiscalizam as finanças dos estados para garantir o cumprimento das obrigações fiscais pelas empresas. Os auditores fiscais são responsáveis por supervisionar o sistema tributário e evitar a sonegação de impostos. Na Paraíba, a SEFAZ enfrenta diversos desafios na cobrança de notas fiscais para arrecadar tributos estaduais, destacando-se, o grande volume de dados a serem classificados manualmente, foco deste trabalho.

A ocorrência de informações erradas e confusas nas descrições de produtos é comum e muitas vezes determina um déficit na arrecadação tributária. Isso se deve ao fato de que a falta de padronização na escrita dessas declarações prejudica a identificação de todos os produtos comercializados e a arrecadação adequada de suas cobranças. Nesse cenário, os auditores fiscais da SEFAZ-PB encaram grandes barreiras para garantir a cobrança eficiente dos tributos, já que a tributação eficiente precisa ocorrer conforme uma pauta fiscal.

As pautas fiscais são tabelas de preços fiscais que determinam os valores presumidos dos itens em cada operação, a fim de aplicar a alíquota (porcentagem cobrada em cima dos rendimentos) e chegar a quantidade do tributo devido (FONTES, 2012).

Com base no que foi apresentado, é factível empregar a tecnologia e suas múltiplas abordagens para desenvolver e aplicar algoritmos que auxiliem neste processo. Pensando nisso, o presente trabalho visa demonstrar o processo de automatização, por meio de um classificador, das cobranças fiscais dos itens de madeira comercializados na Paraíba, mais especificamente com a criação e aplicação de expressões regulares, padronização de sentenças e identificação da pauta fiscal.

Neste estudo, os arquivos de pauta fiscal de madeira (PARAÍBA, 2019) serão usados como modelo para identificar itens comercializados e descritos informalmente, utilizando um algoritmo para adequá-los às descrições formais. Dessa forma, esta pesquisa demonstra a aplicação de expressões regulares, que se tratam de um método formal de especificar um padrão de texto, ou seja, uma composição de símbolos que agrupadas entre si formam uma expressão, interpretada como uma regra que indicará sucesso de uma entrada de dados

(JARGAS, 2016), e a aplicação da ciência de dados, que estuda o dado em todo seu ciclo de vida, da produção ao descarte (AMARAL, 2016).

Apesar de existirem outras técnicas mais avançadas para o tratamento de textos, como a similaridade, as expressões regulares foram escolhidas para este projeto. Isso porque as descrições dos fornecedores possuíam padrões mais simples de serem identificados e tratados, como abreviações, de tal forma que o uso das REGEX seria adequado e pouparia tempo, se mostrando a opção mais eficiente.

Na literatura, diversos trabalhos são direcionados à aplicação de Machine Learning em um contexto de avaliação de tributos e impostos, como na predição de irregularidade fiscal de contribuintes (SOARES e CUNHA 2020), assim como a utilização de Inteligência Artificial (IA) na aplicação de contabilidade nas atividades fiscais e tributárias (SILVA et al. 2022) . Todavia, este trabalho foca na abordagem da pauta fiscal feita através de um classificador utilizando expressões regulares, destacando-se sua importância mediante a massiva e crescente quantidade de notas fiscais emitidas a cada ano[1], que inviabilizam um processamento de dados manual eficaz e evidenciam a necessidade de um processo automatizado.

Vale ressaltar que o atual trabalho aplica-se após a utilização de um primeiro classificador, que determina a categoria do produto, de forma que, o sistema ao qual se refere este projeto diz respeito a uma segunda etapa de classificação, quando é definida a pauta fiscal do produto.

Na sequência, serão apresentados os temas de fundamentação teórica, com conceitos associados às ferramentas utilizadas no desenvolvimento do projeto, baseados em referências científicas, bem como a metodologia empregada e os resultados obtidos por meio da tecnologia desenvolvida.

[1] Estatística da quantidade de notas fiscais emitidas desde 2006, disponível em: <https://www.nfe.fazenda.gov.br/portal/infoEstatisticas.aspx?AspxAutoDetectCookieSupport=1>. Acesso em: 19 de jun. 2023.

2 CONCEITOS E TEORIA

No presente capítulo serão explicados os conceitos fundamentais para o entendimento de Ciência de Dados e bibliotecas de Python (Pandas e re), além do conceito de Pauta fiscal, usadas na aplicação do projeto.

2.1 Ciência de Dados

Segundo Amaral (2016), ciência de dados são os processos, modelos e tecnologias que estudam os dados durante todo o seu ciclo de vida, da produção ao descarte. Já Grus (2016) define como a ciência direcionada para extração de conhecimento a partir de dados desorganizados.

Uma das principais aplicações da ciência de dados é no âmbito financeiro, em que diversas análises podem ser feitas. Pode-se citar: impactos no mercado de ações, uma melhor experiência para os clientes ou até mesmo na prevenção de fraudes. No contexto das fraudes fiscais, é através da coleta de dados que se torna viável criar soluções especializadas, tornando-se assim, uma ciência fundamental na busca de respostas e soluções para problemas reais e cotidianos.

Atualmente, a quantidade de dados é muito grande e está crescendo cada vez mais, principalmente, com o uso da internet. Nesse contexto, o conceito de Big Data[2] torna-se fundamental para a ciência de dados moderna, visto que se refere ao grande volume de dados que são produzidos constantemente, acarretando assim um papel de destaque em meio às técnicas e tecnologias utilizadas para o tratamento de dados.

Nesta pesquisa, a ciência de dados teve papel crucial na identificação da pauta fiscal e no desenvolvimento do classificador, visto que possibilitou o processo de visualização e tratamento dos dados, com identificação dos fatores relevantes e limpeza de elementos dispensáveis, culminando na criação de soluções viáveis.

2.2 Bibliotecas Python

Python[3] é a linguagem mais usada no cenário da ciência de dados, tendo em vista sua velocidade de desenvolvimento e facilidade de uso e leitura. Sendo assim, esta foi a linguagem escolhida para o desenvolvimento de todo o projeto, contando

[2] Disponível em: <https://www.oracle.com/br/big-data/what-is-big-data/>. Acesso em: 25 de jun. 2023.
[3] Disponível em: <https://www.python.org/>. Acesso em: 17 de mar. 2023.

com diversas bibliotecas que auxiliam o processo de análise, tratamento e identificação de textos.

Os próximos tópicos irão tratar justamente das bibliotecas que foram utilizadas no artigo, Pandas e re, mostrando seus conceitos e importância na análise de dados e tratamento de textos, respectivamente.

2.2.1 Pandas

Pandas[4] é uma biblioteca Python bastante utilizada para análise e manipulação de grande quantidade de dados de forma descomplicada. Em virtude disso, foi uma das bibliotecas escolhidas para o projeto, já que possui rapidez e potência para as tarefas como análise exploratória de dados, incluindo identificação, limpeza e manipulação. Além disso, a capacidade de criação de Data Frames, que se trata de um tipo de estruturação de dados, semelhante às tabelas usadas para organização das pautas fiscais[5], é ponto fundamental na estruturação do classificador.

A biblioteca pandas teve papel crucial na manipulação, leitura e visualização da massiva quantidade de dados brutos dos produtos de madeira, sendo assim, uma escolha justificável para a finalidade do projeto.

2.2.2 re

A biblioteca re[6], biblioteca Python que trata das expressões regulares, tem como principais recursos procurar, quebrar e modificar strings, sendo fundamental para grande parte do desenvolvimento deste trabalho.

Na computação, o uso das expressões regulares proporcionam uma maneira maleável de identificar padrões de caracteres ou palavras, de acordo com a criação e estruturação das expressões. A biblioteca foi importante para achar padrões de caracteres e palavras em diversas descrições de produtos, proporcionando, assim, que as pautas dos respectivos produtos fossem escolhidas da melhor forma possível.

Uma situação que pode exemplificar bem o uso de expressões regulares é a identificação de um número de CPF (Cadastro de Pessoa Física). A estrutura de um CPF válido é: 000.000.000-00 ou 00000000000, no exemplo a seguir é mostrado

[4] Disponível em: <https://pandas.pydata.org>. Acesso em: 21 de mar. 2023.
[5] Pauta fiscal e preços sugeridos. Disponível em: <https://www.sefaz.pb.gov.br/legislacao/259-portarias/portarias-2019/7935-portaria-n-00158-2019-sefaz?tmpl=component&format=pdf>. Acesso em: 25 de jun. 2023.
[6] Disponível em: < https://docs.python.org/3/library/re.html>. Acesso em: 1 de abr. 2023.

uma REGEX que é capaz de identificar um cpf válido: /^[0-9]{3}.?[0-9]{3}.?[0-9]{3}-?[0-9]{2}/.

Diante do exposto, a biblioteca re se justifica uma excelente escolha para manipular e identificar as descrições dos produtos de madeira das notas fiscais.

2.3 Pauta fiscal

A pauta fiscal é um documento elaborado pelo legislador estadual, no qual se organiza uma tabela de preços por meio de arbitramento, ou seja, para cada produto que circule no mercado, cria-se uma tabela com determinação de um valor que o referido produto adquirirá em cada etapa da produção (FONTES, 2012). Assim sendo, a partir da descrição do item comercializado é possível adequá-lo a uma descrição pré-determinada pelo estado, através da SEFAZ, e atribuir-lhe o valor de pauta, aplicando a tributação.

Um exemplo deste processo seria a classificação de um produto descrito pelo fornecedor como “MADEIRA CEDRO - RIPA”: ao ser classificado, esse produto seria relacionado a uma categoria, como no caso da pauta de madeira, em que este é determinado no grupo “CEDRO - CAIBRO, RIPA E TÁBUA”, cujo metro cúbico corresponde ao valor de R$ 2.000,00, quantia sobre a qual será aplicado a tributação[5].

O arquivo supracitado contém várias informações sobre os produtos passíveis de cobrança, como fabricante, descrição, valor da pauta, entre outros, e são fundamentais para identificação correta e tributação adequada de todos os produtos comercializados, visando a prevenção de déficit na arrecadação de impostos.

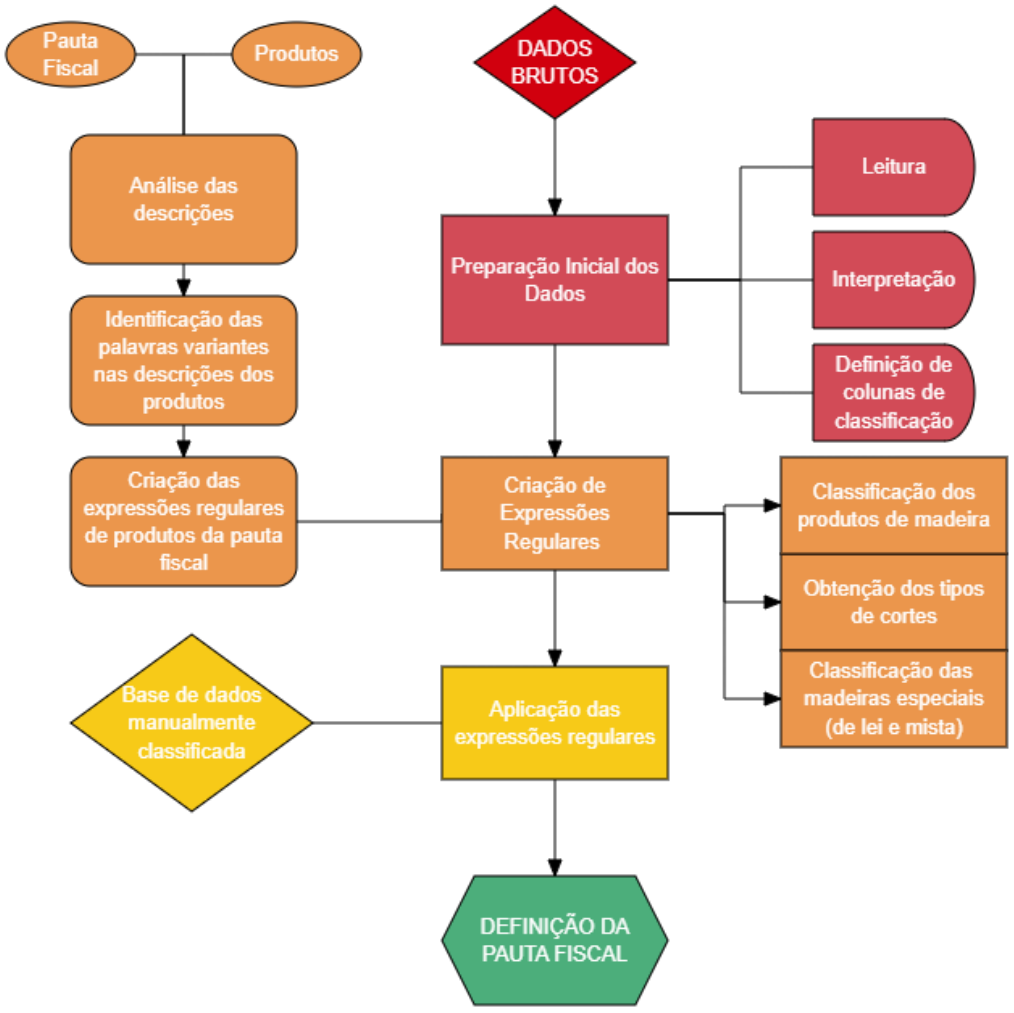
[5] Pauta fiscal e preços sugeridos. Disponível em: <https://www.sefaz.pb.gov.br/legislacao/259-portarias/portarias-2019/7935-portaria-n-00158-2019-sefaz?tmpl=component&format=pdf>. Acesso em: 25 de jun. 2023.

3 METODOLOGIA

Inicialmente, com a identificação do problema das irregularidades fiscais, foram geradas ideias com o propósito de aprimorar o processo de cobrança e qualidade do trabalho realizado pelos profissionais envolvidos. Nesse sentido, o foco principal seria a criação de uma solução que automatizasse, de modo eficaz e ágil, uma parte do processo de cobrança adotado na SEFAZ-PB.

Pensando nisso, definiu-se que o processo seria realizado em cinco etapas principais para se conseguir identificar a pauta fiscal, utilizando ciência de dados e expressões regulares. O processo de classificação é demonstrado na Figura 1 e envolve preparação inicial dos dados, a criação das expressões regulares, a aplicação das expressões regulares e, por último, a definição da pauta fiscal. Nas subseções a seguir serão apresentadas estas etapas serão detalhadas.

Figura 1 - Processo de classificação da pauta fiscal

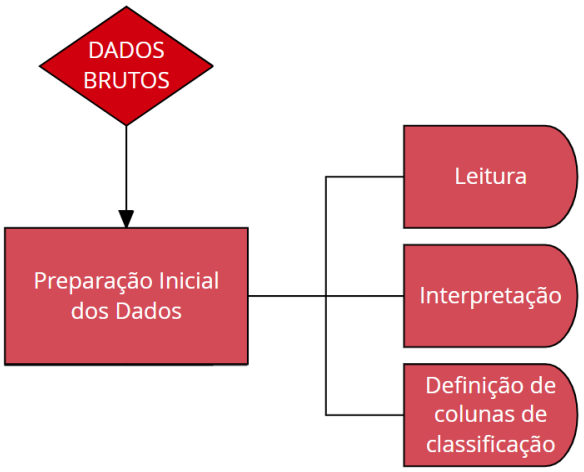


Fonte: Elaborado pelo autor, 2023.

3.1 Preparação inicial dos dados

Para estruturar as informações essenciais necessárias para a criação do classificador, é necessário realizar o pré-processamento das bases de dados, conforme apresentado no fluxograma da Figura 2. Assim, as bases de dados originais passarão por todos os passos mencionados, a fim de gerar bases de dados pré-processadas, que abordam os problemas iniciais e possíveis problemas futuros.

Figura 2 - Preparação inicial dos dados



Fonte: Elaborado pelo autor, 2023.

3.1.1 Processamento de dados brutos (leitura e interpretação)

A ciência de dados é uma disciplina crucial quando é necessário lidar com grandes volumes de dados, permitindo visualizar e organizar as informações. Durante o processamento dos dados, foi possível limpar informações irrelevantes e extrair informações valiosas com o auxílio da biblioteca Pandas, desta forma a leitura dos dados foi facilitada, tornando viável manipulá-los de maneira mais eficiente e clara. Além disso, a interpretação busca detectar potenciais padrões anômalos, como descrições irregulares de produtos.

Com isso em mente, foi possível avaliar as descrições das notas fiscais de madeira, avaliando comportamentos de escrita mais profundamente e assim permitindo um melhor tratamento dos dados, a mencionar nos casos em que as descrições eram feitas com diferentes formatações de letras maiúsculas/minúsculas e foi possível padronizá-las a um único formato.

3.1.2 Definição das colunas importantes e de classificação

Dentre os dados analisados, apresentavam-se diversas informações, organizadas em colunas, sobre declarações de produtos comercializados no estado, assim como as pautas fiscais utilizadas para as cobranças.

No entanto, nem todas as informações eram pertinentes para a criação da ferramenta de classificação de textos declarados, como, por exemplo, a data de extração do produto, que não interferiram no valor da tributação. Por essa razão, algumas colunas que não contribuíam para a precisão do classificador não foram selecionadas, restando somente as colunas relevantes para a tarefa de classificação.

Além disso, para que ao final do classificador, possa ser medida sua acurácia, é preciso que exista uma coluna com as classificações feitas pelo algoritmo. Logo, foi fundamental a criação de uma coluna de classificação durante o desenvolvimento do projeto.

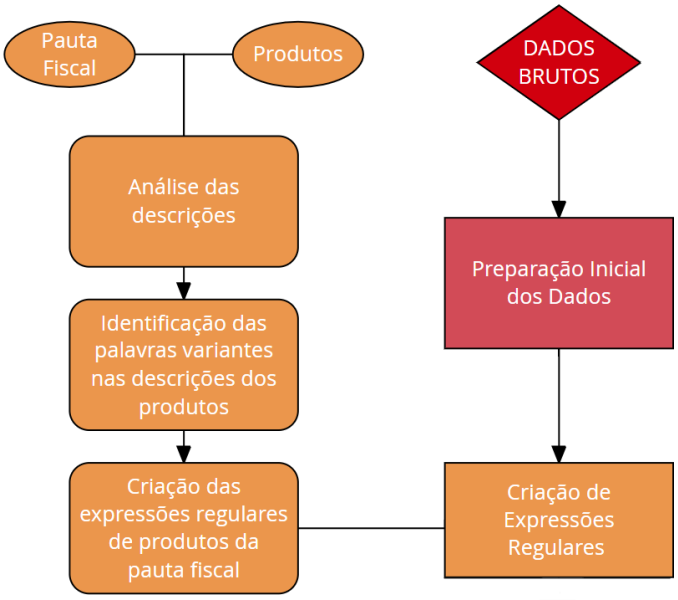
Nesse cenário, após a organização da coluna que armazenaria as classificações, foi então definido que toda a coluna começaria com valor zero, pois até que se prove o contrário, os produtos não têm pauta e somente após a execução do classificador seria possível definir ou não uma pauta válida.

3.2 Criação das expressões regulares

Em meio às notas fiscais analisadas, muitas das descrições dos produtos vêm de forma diferente das descrições da pauta fiscal, por isso a utilização de expressões regulares foi fundamental no sentido de conseguir identificar corretamente a pauta, mesmo que a descrição não esteja escrita como deveria, ou seja, igual a descrição da pauta fiscal.

Sendo assim, a utilização da pauta fiscal e dos produtos das notas fiscais para identificar padrões de linguagem das descrições e criar as expressões regulares foi um ponto crucial no desenvolvimento do algoritmo, processo este que é demonstrado na Figura 3.

Figura 3 - Criação das expressões regulares



Fonte: Elaborado pelo autor, 2023.

3.2.1 Análise das descrições da pauta fiscal e dos produtos

A observação das descrições da pauta foi de extrema importância, já que essa descrição é a forma correta como os produtos deveriam vir dos fornecedores e, portanto, o meio final onde as expressões regulares teriam que ser baseadas.

Já no que tange às análises das descrições dos produtos, sua avaliação também se mostrou fundamental, visto que, dessa maneira, tornou-se possível saber como os fornecedores declaram os produtos das notas fiscais, podendo ter descrições parecidas com as da pauta ou muito diferentes dela.

Sendo assim, durante a análise das descrições, padrões foram estabelecidos e encontrados mais facilmente, ao passo que as descrições incorretas foram encontradas e corrigidas, mostrando-se uma etapa indispensável para a criação das expressões regulares.

3.2.2 Identificação das palavras variantes

Como já mencionado, um dos problemas enfrentados durante a avaliação das notas fiscais é a utilização de termos que fogem ao padrão da pauta, principalmente porque os fornecedores têm liberdade para descrever seus produtos conforme suas próprias competências, apesar de nem sempre essas descrições serem claras e completas. A exemplificar, o uso de termos como “MACARADUBA” ou “MASSAR”, ao invés de “MASSARANDUBA”, termo presente na pauta.

Nesse sentido, um importante processo de identificação das palavras que são escritas de forma diferente e/ou incompletas mostrou-se necessário, tendo em vista a possibilidade de ocorrência de problemas durante a definição da pauta. Sabendo disso, tornou-se imprescindível que as palavras variantes das descrições fossem identificadas e tratadas.

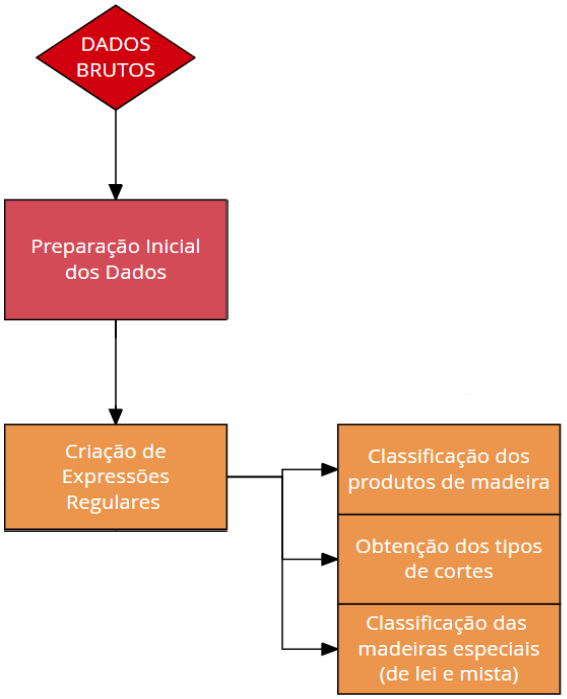
3.2.3 Criação das expressões regulares de produtos da pauta

Após a realização das etapas anteriores, o próximo passo foi a criação das expressões regulares, utilizando como base as descrições dos produtos, da pauta e das palavras variantes.

A criação das expressões regulares foi feita de maneira bem simplificada: cada madeira presente na pauta teve sua expressão regular estabelecida, sendo tratadas todas as possibilidades em que ela estivesse escrita, como por exemplo: “tabua”, “tábua”, “tabuá”, “taboa” e “ta bua”. Em seguida, atribuiu-se o determinado resultado de todas essas descrições para sua variável, que foi usada posteriormente na aplicação das expressões regulares.

Por fim, o processo de criação das expressões regulares foi dividido em três grandes etapas, a saber: classificação dos produtos comuns/usuais, obtenção dos tipos de corte de madeira e classificação dos produtos especiais/não especificados (definidos como de lei e mista). Esse processo é explanado na Figura 4.

Figura 4 - Etapas de criação das expressões regulares



Fonte: Elaborado pelo autor, 2023.

3.2.3.1 Classificação dos produtos de madeira

Inicialmente, foram identificados os produtos mais usuais, que são descritos na pauta fiscal, com organização de acordo com cada tipo de madeira, como mostrado na Figura 5.

Figura 5 - Classificação dos produtos de madeira mais usuais

MAÇARANDUBA - CAIBRO, RIPA E TÁBUA	12851
MAÇARANDUBA - LINHA, BARROTE E PRANCHA	12852
ANDIROBA - CAIBRO, RIPA, TÁBUA, LINHA, BARROTE, E PRANCHA	12826
BÁLSAMO - CAIBRO, RIPA, TÁBUA, LINHA, BARROTE, E PRANCHA	12828
JACARANDÁ - CAIBRO, RIPA, TÁBUA, LINHA, BARROTE, E PRANCHA	12847
PINUS - CAIBRO, RIPA, TÁBUA, LINHA, BARROTE, E PRANCHA	12854
MOGNO - CAIBRO, RIPA, TÁBUA, LINHA, BARROTE, E PRANCHA	12853
SUCUPIRA - CAIBRO, RIPA, TÁBUA, LINHA, BARROTE, E PRANCHA	12855
ANGELIM - CAIBRO, RIPA, TÁBUA, LINHA, BARROTE, E PRANCHA	12827
OUTRAS MADEIRAS DE LEI NÃO ESPECIFICADAS CAIBRO, RIPA, TÁBUA, LINHA, BARROTE, E PRANCHA	12856
OUTRAS MADEIRAS MISTAS NÃO ESPECIFICADAS CAIBRO, RIPA, TÁBUA, LINHA, BARROTE, E PRANCHA	12857

Fonte: Elaborado pelo autor, 2023.

Em seguida, tendo sido identificados, foi criada uma expressão regular para cada tipo de produto, associada a variável estabelecida para cada tipo de madeira. Dessa forma, as primeiras expressões foram geradas, mantendo o padrão da pauta fiscal.

3.2.3.2 Obtenção dos tipos de cortes

Além dos diferentes tipos de madeiras, diferentes cortes, como exposto na Figura 6, foram identificados e separados dentre os produtos da pauta fiscal. Se fez necessário uma boa observação de todos os tipos, já que foi construída uma lógica para separação dos mesmos.

Figura 6 - Classificação dos tipos de corte

MAÇARANDUBA - CAIBRO, RIPA E TÁBUA	12851
MAÇARANDUBA - LINHA, BARROTE E PRANCHA	12852
ANDIROBA - CAIBRO, RIPA, TÁBUA, LINHA, BARROTE, E PRANCHA	12826
BÁLSAMO - CAIBRO, RIPA, TÁBUA, LINHA, BARROTE, E PRANCHA	12828
JACARANDÁ - CAIBRO, RIPA, TÁBUA, LINHA, BARROTE, E PRANCHA	12847
PINUS - CAIBRO, RIPA, TÁBUA, LINHA, BARROTE, E PRANCHA	12854
MOGNO - CAIBRO, RIPA, TÁBUA, LINHA, BARROTE, E PRANCHA	12853
SUCUPIRA - CAIBRO, RIPA, TÁBUA, LINHA, BARROTE, E PRANCHA	12855
ANGELIM - CAIBRO, RIPA, TÁBUA, LINHA, BARROTE, E PRANCHA	12827
OUTRAS MADEIRAS DE LEI NÃO ESPECIFICADAS CAIBRO, RIPA, TÁBUA, LINHA, BARROTE, E PRANCHA	12856
OUTRAS MADEIRAS MISTAS NÃO ESPECIFICADAS CAIBRO, RIPA, TÁBUA, LINHA, BARROTE, E PRANCHA	12857

Fonte: Elaborado pelo autor, 2023.

Observando os cortes, foi decidido que a separação seria definida por dois tipos de cortes: tipo 1 e tipo 2. O primeiro era composto pelos cortes: Caibro, Ripa, Tábua, Viga e Sarrafo, e o segundo pelos cortes: Prancha, Barrote e Linha.

Por fim, fez-se necessário a criação das expressões regulares dos tipos de cortes que foram definidos anteriormente, sendo criado as expressões para o tipo 1 e o tipo 2, e estas foram armazenadas em variáveis para serem usadas posteriormente na identificação da pauta.

3.2.3.3 Classificação das madeiras especiais

Dentre os produtos, percebeu-se que existiam madeiras não especificadas que precisavam primeiramente serem definidas, para que, só então, fossem criadas suas expressões regulares. Esta categoria inclui madeiras de lei e madeiras mista, como destacado na Figura 7.

Figura 7 - Classificação das madeiras especiais

MAÇARANDUBA - CAIBRO, RIPA E TÁBUA	12851
MAÇARANDUBA - LINHA, BARROTE E PRANCHA	12852
ANDIROBA - CAIBRO, RIPA, TÁBUA, LINHA, BARROTE, E PRANCHA	12826
BÁLSAMO - CAIBRO, RIPA, TÁBUA, LINHA, BARROTE, E PRANCHA	12828
JACARANDÁ - CAIBRO, RIPA, TÁBUA, LINHA, BARROTE, E PRANCHA	12847
PINUS - CAIBRO, RIPA, TÁBUA, LINHA, BARROTE, E PRANCHA	12854
MOGNO - CAIBRO, RIPA, TÁBUA, LINHA, BARROTE, E PRANCHA	12853
SUCUPIRA - CAIBRO, RIPA, TÁBUA, LINHA, BARROTE, E PRANCHA	12855
ANGELIM - CAIBRO, RIPA, TÁBUA, LINHA, BARROTE, E PRANCHA	12827
OUTRAS MADEIRAS DE LEI NÃO ESPECIFICADAS CAIBRO, RIPA, TÁBUA, LINHA, BARROTE, E PRANCHA	12856
OUTRAS MADEIRAS MISTAS NÃO ESPECIFICADAS CAIBRO, RIPA, TÁBUA, LINHA, BARROTE, E PRANCHA	12857

Fonte: Elaborado pelo autor, 2023.

Com isso em mente, após breve pesquisa e deliberações feitas em equipe, ficou-se definido que as madeiras de lei seriam as madeiras contidas na Figura 8 e que não tivessem sido especificadas no classificador, já que algumas dessa lista já haviam sido tratadas na pauta. Enquanto as madeiras mistas seriam as demais, ou seja, as pertencentes à categoria de madeira e que não fossem especificadas em nenhuma pauta anterior.

Feito isso, seguiu-se com a criação das expressões regulares para as madeiras de lei e para as madeiras mistas.

Figura 8 - Madeiras de lei

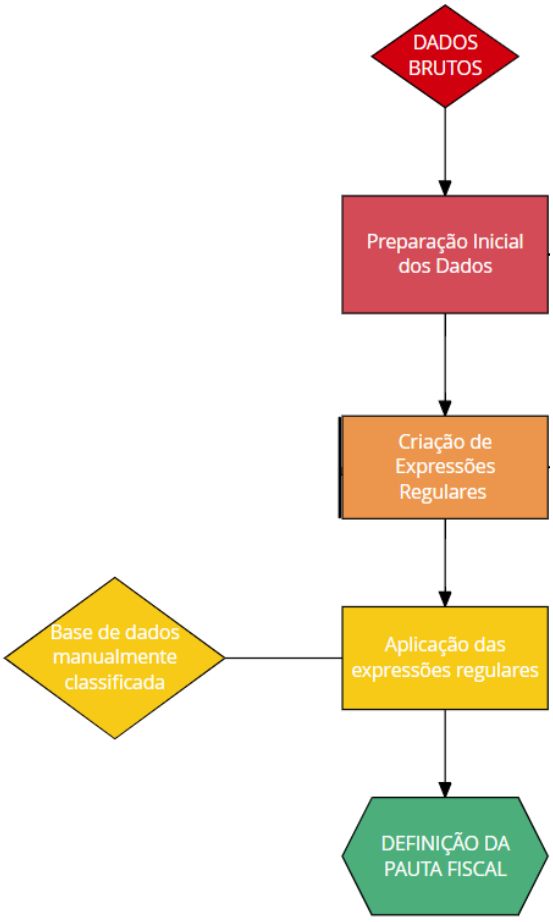
- Acaiacá, também conhecida como Cedro-rosa
- Andiroba
- Angelim-Vermelho, conhecido também como Fevero-Ferro
- Angico
- Araribá
- Imbuia
- Ipê
- Ipê-Felpudo
- Jacarandá
- Jacarandá-da-bahia
- Jacareúba, também conhecida como Guanandi
- Jatobá
- Mogno
- Pau-Brasil
- Pau-Ferro
- Pau-Pereira
- Peroba-Rosa

Fonte: Viva Decora, 2022.

3.3 Definição da pauta fiscal

A última etapa para a identificação da pauta fiscal foi a aplicação de todas as expressões regulares feitas durante o processo de desenvolvimento do classificador, esse processo é mostrado na Figura 9.

Figura 9 - Identificação da pauta fiscal



Fonte: Elaborado pelo autor, 2023.

3.3.1 Aplicação das expressões regulares

Para essa etapa foi importante a utilização da base de dados com os produtos classificados manualmente, que é a base de dados pré-processada resultante da classificação de todos os produtos sem automatização. Esses dados seriam o gabarito para a comparação com a classificação do classificador e foi crucial para o cálculo de acurácia.

Com o carregamento da base manual, a aplicação das expressões regulares foi feita da seguinte forma: execução das expressões regulares de cada tipo de madeira, seguida da execução das expressões regulares dos tipos de corte, a fim de verificar em qual tipo a madeira se encaixava.

Esse processo foi aplicado em cada uma das descrições dos produtos da base de dados classificada manualmente, armazenando, por fim, o resultado encontrado na coluna criada para o resultado do classificador.

Após a aplicação das expressões regulares, os produtos de madeira foram classificados e a acurácia do classificador pode ser verificada na próxima seção.

4 RESULTADOS

A necessidade de criar um classificador surgiu em meio ao cenário de obstáculos enfrentados pela Secretaria da Fazenda da Paraíba (SEFAZ-PB) na cobrança de impostos estaduais, visando simplificar esse processo. Para isso, desenvolveu-se uma solução que identifica a pauta fiscal dos produtos da categoria de madeira, por meio das descrições dos produtos das notas fiscais dos fornecedores, utilizando expressões regulares.

Foram utilizados 8.953 produtos de uma base de dados com produtos da categoria madeira, sendo utilizada uma base de dados com 39 linhas de produtos da pauta fiscal de madeira (PARAÍBA, 2019). Para isso, as bases de dados foram usadas, como mostrado anteriormente, e foi-se obtido um cálculo de acurácia, após a execução do classificador, de 98,77%, alcançando uma taxa de erro muito baixa e relevância considerável da aplicação do classificador.

Nas tabelas a seguir serão mostrados alguns exemplos dos resultados encontrados.

Tabela 1 - Exemplo da pauta fiscal da categoria de madeira

Descrição	sq_pauta
OUTRAS MADEIRAS DE LEI NÃO ESPECIFICADAS CAIBRO, RIPA, TÁBUA, LINHA, BARROTE E PRANCHA	12856
OUTRAS MADEIRAS MISTAS NÃO ESPECIFICADAS CAIBRO, RIPA, TÁBUA, LINHA, BARROTE E PRANCHA	12857
CUMARU - LINHA, BARROTE E PRANCHA	12838
JATOBÁ - CAIBRO, RIPA E TÁBUA	12849
ANGELIM - CAIBRO, RIPA, TÁBUA, LINHA, BARROTE E PRANCHA	12827

Fonte: Elaborada pelo autor, 2023.

As informações presentes na tabela são recortes da base de dados da pauta fiscal de madeira que foi usada como guia para criação do classificador. Nela, a coluna de “Descrição” relata os produtos da pauta fiscal, produtos esses que deverão ser encontrados na base de dados das notas fiscais e assim classificados com seu determinado "sq_pauta".

O “sq_pauta” é um número singular que identifica cada produto da pauta fiscal. Sabendo disso, a base de dados da pauta fiscal presente na Tabela 1 será usada a título de comparação com a Tabela 2, que irá mostrar exemplos da classificação feita pelo classificador.

Tabela 2 - Exemplos dos resultados do classificador

Descrição	sq_manual	sq_classificador
LINHA PEROBA ROSA 5X11 4.5 MT	12856	12856
ESPETEIRO MAD SERRADA EM RIPAS	12857	12857
CUMARU MAD SERRADA PRANCHA	12838	12838
JATOBA - MAD SER EM CAIBRO CURTO CAIBRINHO	12849	12849
LINHA 3/6 - VIGA ANGELIM	12827	12827

Fonte: Elaborada pelo autor, 2023.

A Tabela 2 possui a coluna “Descrição”, que corresponde às descrições das notas fiscais vindas dos fornecedores do Estado da Paraíba, em sua maioria, incorretas e/ou incompletas, podendo causar dificuldade de identificação, como já citado. A coluna de “sq_classificador” possui o número de “sq_pauta” da pauta do produto encontrado pelo classificador. A coluna “sq_manual” é responsável por armazenar os números de “sq_pauta” feita de forma manual e será usada como gabarito para o teste de acurácia.

Assim, após a execução do classificador, os produtos possuem uma classificação baseada no “sq_pauta” da pauta fiscal e caso o classificador não tenha identificado a pauta fiscal de um determinado produto, então “sq_classificador” permanece com valor zero (o valor zero foi definido para representar que o produto não está presente na pauta fiscal).

Nas linhas 2 e 3 da Tabela 2, as descrições se tratam de madeiras de lei (como mostrado na Figura 8) e madeiras mistas (como explicado na seção 3.2.3.3), respectivamente, sendo classificadas pelo classificador de forma assertiva, ao se estabelecer sua correspondência com a classificação manual. Nas linhas 4, 5 e 6 da Tabela 2, as descrições se tratam das madeiras cumaru com corte do tipo prancha, jatobá com corte do tipo caibro e angelim com corte do tipo viga, respectivamente, determinadas corretamente pelo classificador, de acordo com a comparação do “sq_classificado” em relação ao “sq_manual”.

Por conseguinte, o teste de acurácia é feito através da comparação entre os valores do “sq_classificador” e ”sq_manual", caso sejam iguais então o classificador obteve êxito e caso contrário a classificação foi feita errada.

Sendo assim, conclui-se que, na Tabela 2, todos os exemplos foram classificados corretamente pelo algoritmo, o que pode ser evidenciado pela correlação dos resultados com a Tabela 1, que mostra a pauta fiscal da categoria de madeira.

5 CONSIDERAÇÕES FINAIS

Ao começar a pesquisa, foi identificada a dificuldade dos auditores fiscais para cobrar produtos na Paraíba, tendo em vista a utilização de um processo manual de classificação, levando à necessidade de desenvolver uma ferramenta automatizada para detectar e calcular os impostos das mercadorias vendidas.

A pesquisa, feita pelo NUTES em parceria com a SEFAZ-PB, visava desenvolver uma aplicação para facilitar a cobrança de vendas e melhorar a tributação. O sistema foi implantado e permanece em vigor, alcançado com sucesso seu objetivo, ao criar uma ferramenta de classificação que padroniza as descrições dos produtos, identifica a pauta fiscal e permite uma cobrança eficiente, com acurácia de 98,77%.

Entretanto, o classificador desenvolvido possui limitações ao lidar com produtos ausentes na lista de referência ou com falta de informações, como no exemplo desta pauta para os casos em que os cortes não foram descritos ou o tipo de madeiras não era determinado, impossibilitando a inclusão dos itens na classificação e impedindo uma tributação abrangente.

Sendo assim, faz-se necessário incentivar projetos futuros para superar essas barreiras e ampliar a capacidade da ferramenta de cobrir a maior quantidade possível de itens comercializados no estado, garantindo uma tributação estadual de produtos ainda mais eficaz.

REFERÊNCIAS

AMARAL, Fernando. **Introdução à Ciência de Dados: mineração de dados e big data**. Rio de Janeiro: Alta Books, 2016.

Conteúdo da Documentação Python. **Python Software Foundation**, 20 jun. 2023. Disponível em: https://docs.python.org/pt-br/3/contents.html . Acesso em: 20 jun. 2023.

COSTA, Pedro Henrique de Farias. **Uso de Técnicas de Similaridade para Identificação de Pauta de Produtos Fiscais**. 2021. Dissertação (Bacharelado em Ciências da Computação) - Curso de Ciências da Computação, Universidade Estadual da Paraíba, Campina Grande. 2021.

CRUZ, Talita. Quais São as Madeiras de Lei? Veja as 5 Espécies Mais Incríveis!. **Viva Decora**. 07 out. 2022. Disponível em: https://www.vivadecora.com.br/pro/madeira-de-lei/. Acesso em: 15 mai. 2023.

EDUCAÇÃO, Redação XP. Ciência de dados no mercado financeiro: como funciona?. **blog xp educação,** 17 ago. 2022. Disponível em: https://blog.xpeducacao.com.br/ciencia-de-dados-no-mercado-financeiro/. Acesso em: 17 mar. 2023.

FONTES, Christiane Kuntze dos Santos. **Restituição de ICMS decorrente de fato gerador presumido: novas perspectivas**. Escola da Magistratura do Estado do Rio de Janeiro. Rio de Janeiro, RJ, 2012.

JARGAS, Aurelio Marinho. **Expressões Regulares: Uma Abordagem Divertida**. 5ª Edição. São Paulo: Novatec, Janeiro de 2016.

PARAÍBA. Portaria nº 00158/2019/SEFAZ, de 16 de maio de 2019. Categoria de Produtos: Madeiras**. Diário Oficial do Estado**. João Pessoa, PB. 2019. Disponível em:

<https://www.sefaz.pb.gov.br/legislacao/259-portarias/portarias-2019/7935-portaria-n-00158-2019-sefaz?tmpl=component&format=pdf>. Acesso em: 25 jun. 2023.

SAISSE, Renan. Breves considerações sobre Ciência de Dados e como se qualificar nesta área promissora!. **ti especialistas,** 09 mai. 2019. Disponível em: https://www.tiespecialistas.com.br/breves-consideracoes-sobre-ciencia-de-dados-e-como-se-qualificar-nesta-area-promissora/. Acesso em: 09 mar. 2023.

SILVA, Denis Ribeiro da; COSTA, Daniel Fonseca da; PIMENTA, Alexandre. **A Influência da Inteligência Artificial na Contabilidade e na Tributação das Organizações: uma revisão de literatura**. Instituto Federal de Educação, Ciência e Tecnologia de Minas Gerais. São Paulo, SP, 2022.

SOARES, Glauco de Vasconcelos; CUNHA, Rodrigo C. L. V. **Predição de Irregularidade Fiscal dos Contribuintes do Tributo ISS**. Centro de Estudos e Sistemas Avançados do Recife (CESAR SCHOOL). Recife, PE, 2020.