PONTIFÍCIA UNIVERSIDADE CATÓLICA DE GOIÁS
ESCOLA POLITÉCNICA
CURSO DE CIÊNCIA DA COMPUTAÇÃO

MINERAÇÃO DE DADOS APLICADA EM PROCESSOS FISCAIS

VINÍCIUS DE ASSUNÇÃO FURTADO

GOIÂNIA
2021

VINÍCIUS DE ASSUNÇÃO FURTADO

MINERAÇÃO DE DADOS APLICADA EM PROCESSOS FISCAIS

Trabalho de Conclusão de Curso apresentado à Escola Politécnica, da Pontifícia Universidade Católica de Goiás, como parte dos requisitos para a obtenção do título de Bacharel em Ciência da Computação.

Orientador: Prof. Dr. Sibelius Lellis Vieira
Banca examinadora: Prof. Me. Fernando Abadia Gonçalves
Prof. Me. Wilmar Oliveira de Queiroz

GOIÂNIA
2021

VINÍCIUS DE ASSUNÇÃO FURTADO

MINERAÇÃO DE DADOS APLICADA EM PROCESSOS FISCAIS

Trabalho de Conclusão de Curso aprovado em sua forma parcial pela Escola Politécnica, da Pontifícia Universidade Católica de Goiás, para obtenção do título de Bacharel em Ciência da Computação, em _____/_____/_____.

_____

Orientador: Prof. Dr. Sibelius Lellis Vieira

_____

Prof. Me. Fernando Abadia Gonçalves

_____

Prof. Me. Wilmar Oliveira de Queiroz

GOIÂNIA

2021

AGRADECIMENTOS

Agradeço primeiramente a Deus, por me abençoar e guiar.

A minha família, pelo apoio durante todo o período de estudos, por me incentivar a todo momento e não permitir que eu desistisse.

A minha namorada Jéssica, pelo apoio, carinho, compreensão e incentivo.

A esta universidade, seu corpo docente e direção, por levar a sabedoria e estudo, em especial ao meu orientador Prof. Dr. Sibelius Lellis Vieira, pela disponibilidade e suporte.

Aos meus amigos, pela compreensão das ausências e afastamento temporário.

E a todos, de maneira direta e indireta que fizeram parte da minha formação.

RESUMO

Este trabalho pretende aplicar técnicas de mineração de dados no âmbito de representações fiscais, com objetivo de predizer e classificar crimes contra a ordem pública e a duração dos processos administrativos, visando à contribuição no poder público em suas tomadas de decisão. Foi realizada uma revisão bibliográfica com finalidade de conhecer a tributação, os processos de representações fiscais para fins penais, mineração de dados e as técnicas de mineração. Foi feito um levantamento dos dados de representações públicas disponíveis pela Receita Federal, em que extraiu o ano de 2018 a 2020. Em seguida a seleção, limpeza e organização dos dados manualmente, com auxílio de uma planilha eletrônica. Após, foi utilizado o software WEKA para ser realizada a mineração dos dados, com pretensão dos seus resultados serem mostrados e analisados, para prever e classificar se os dados têm potencialidade de contribuir com o poder público no aperfeiçoamento e otimização dos processos.

Palavras-chave: representações fiscais, Receita Federal, poder público, ciência de dados, mineração de dados, descoberta de conhecimento em base de dados, árvore de decisão, redes neurais.

ABSTRACT

This work intends to apply data mining techniques within the scope of fiscal representations, with the objective of predicting and classifying crimes against public order and the duration of administrative proceedings, administrative to the contribution of the public power in obtaining a decision. A bibliographical review was carried out with knowledge of knowing taxation, the processes of tax representations for criminal purposes, data mining and mining techniques. A survey of data from public representations available by the Federal Revenue was carried out, in which the year 2018 to 2020 was extracted. Then, the selection, cleaning and organization of data manually, with the aid of an electronic spreadsheet. Afterwards, the WEKA software was used to perform the data mining, with the intention of its output and distribution results, to predict and classify the data that have the potential to contribute to the public authorities in the improvement and optimization of processes.

Keywords: *tax representations, Federal Revenue, public authorities, data science, data mining, knowledge discovery in database, decision tree, neural networks.*

LISTA DE ABREVIATURAS

AIIM	Auto de Infração e Imposição de Multa
AP	Associação Privada
ARFF	*Attribute Relation File Format*
CNPJ	Cadastro Nacional da Pessoa Jurídica
CPF	Cadastro de Pessoa Física
CSV	*Comma Separated Values*
DCBD	Descoberta de Conhecimento em Base de Dados
DRF	Delegacia da Receita Federal
EI	Empresário Individual
EIRELI	Empresa Individual de Responsabilidade Limitada
KDD	*Knowledge Discovery in Databases*
LTDA	Sociedade Limitada
MLP	*Multilayer Perceptron*
MPF	Ministério Público Federal
PDF	*Portable Document Format*
RNA	Rede Neural Artificial
RFB	Receita Federal do Brasil
RFFP	Representações Fiscais Para Fins Penais
S/A	Sociedade Anônima
SPL	*Single Layer Perceptron*
SPAM	*Sending and Posting Advertisement in Mass*
UF	Unidade Federativa
WEKA	*Waikato Environment for Knowledge Analysis*

LISTA DE ILUSTRAÇÕES

FIGURAS

QUADROS E TABELAS

SUMÁRIO

1 INTRODUÇÃO

Esta seção apresenta a introdução do trabalho, explicando a contextualização, justificativa, objetivos (geral e específicos) e a estrutura do trabalho.

1.1 Contextualização

O Estado para proporcionar melhorias à população, atendendo pela carência e demanda de serviços públicos, deve promover a arrecadação de tributos dos contribuintes, que é de extrema importância, sendo a melhor fonte de financiamento público. Por isso a importância no combate à fuga de tributos (MOREIRA, 2003).

Com objetivo de desestimular ações ilícitas, a Receita Federal do Brasil (RFB) impõe limites contra certas atitudes, que levam os contribuintes a não exercerem seus papéis de cidadão nas obrigações fiscais. O cidadão brasileiro que demonstra atitudes que infringem a lei brasileira relativa aos crimes contra a ordem tributária pode enfrentar processos que enquadram no ilícito fiscal (MOREIRA, 2003).

A Representação Fiscal para Fins Penais (RFFP) é o instrumento utilizado pela RFB, designado pelos Auditores-Fiscais da Receita Federal, encaminhadas ao Ministério Público, contendo o ilícito penal e dados do contribuinte alvo, colaborando como exemplos que visam ilidir a sonegação de impostos (MOREIRA, 2003).

Umas das principais funções da RFFP é contribuir para o Estado ter mais arrecadação e menos desperdício de dinheiro públicos. A maneira como é realizada, por intermédio das penalidades, inibe o contribuinte a situações que estaria promovendo um desvio desses recursos que deveria ir para RFB e desviada a sua principal função (MOREIRA, 2003).

Nos processos de representações fiscais, atrasos em movimentações dos atos do processo podem levar à prescrição penal. Prevista no art. 107, inciso IV do Código Penal do Decreto Lei nº 2.848 de 07 de Dezembro de 1940, a prescrição penal se caracteriza pela ocorrência da extinção da punibilidade do Estado, ou seja, quando deixa de punir ilicitamente alguém por um ato que é considerado crime ou delito, sendo uma das principais causas o tempo limite de alguma intervenção penal expirar. Nesta situação, os tributos aplicados deixam de ir para projetos

governamentais, que visam à manutenção social da população brasileira (MOREIRA, 2003).

Neste trabalho explora-se a área da ciência de dados na parte de mineração de dados, pela grande quantidade de representações, ideal para o tipo de tratamento de informações. Desta forma, busca-se entender como são feitas as representações fiscais para fins penais, de forma que averigue, a partir dos dados, melhorias usando a mineração de dados para descobrir vantagens e benefícios na tomada de decisão, analisando os resultados gerados e aplicados no mundo real.

1.2 Justificativa

Um dos problemas com a ineficácia do procedimento administrativo é o tempo gasto entre e seu início e sua apresentação ao MPF, o que pode causar a prescrição penal do processo, o que justifica a investigação da aplicabilidade da mineração de dados nos processos de RFFP. Além disso, a definição do ilícito penal realizada corretamente no início do procedimento fiscal pode auxiliar na diminuição do tempo necessário para o encerramento do processo administrativo.

1.3 Objetivo

Esta seção visa apresentar o objetivo, tanto geral e específico do trabalho, pretendendo esclarecer os propósitos da pesquisa.

1.3.1 Objetivo geral

O objetivo deste estudo é examinar e aplicar técnicas de mineração de dados no contexto de representações penais para fins fiscais da Receita Federal encaminhadas ao Ministério Público, com intenção de relacionar estas representações com os crimes praticados, e com a duração do processo administrativo relacionado a estas representações, com vistas assessorar o Poder Público em sua missão.

1.3.2 Objetivos específicos

Para se atingir o objetivo geral, propõem-se os seguintes objetivos específicos:

- Definir conceitualmente a mineração de dados;
- Definir conceitualmente a tributação, por meio do Direito Penal Tributário e Fiscalização Tributária;
- Utilizar a descoberta de conhecimento de dados em base de dados;
- Aplicar as técnicas de mineração de dados em representações fiscais;
- Examinar os resultados adquiridos da mineração de dados, pelo *software* de análise computacional de mineração de dados;
- Classificar as representações, tanto para obter um perfil do crime praticado quanto do tempo estimado do processo administrativo.

1.4 Estrutura do trabalho

Esse estudo apresenta-se estruturado em 5 (cinco) capítulos, sendo o capítulo 1 (um) referente a esta introdução. O capítulo 2 (dois) é designado ao referencial teórico, contendo a pesquisa bibliográfica, bem como os conceitos sobre tributação, por meio do Direito Penal Tributário e seus crimes contra a ordem pública e o processo de Representações Fiscais para Fins Penais, sobre a descoberta de conhecimento em base de dados, algoritmos, técnicas e trabalhos correlacionados. No capítulo 3 (três) são tratados os procedimentos metodológicos, sobre os materiais e métodos que correspondem à utilização do trabalho de pesquisa. O capítulo 4 (quatro) expõe os resultados obtidos durante a seleção e pré-processamento e análise preditiva. No capítulo 5 (cinco) apresenta a conclusão da pesquisa e sugestões para trabalhos futuros.

2 REFERENCIAL TEÓRICO

Esta seção pretende apresentar dois conceitos, o primeiro sobre a tributação e segundo a descoberta de conhecimento em banco de dados.

Referente a tributação, pretende conceituar o seu intuito e funcionamento do processo penal tributário, respectivo da fiscalização tributária, subsequente aos crimes e a Receita Federal do Brasil (RFB).

Relacionado a descoberta de conhecimento, planeja defender o embasamento da inteligência artificial, em consonância das fases que fazem parte do *Knowledge Discovery in Databases* (KDD) e as técnicas aplicadas de mineração de dados utilizadas no trabalho.

2.1 Tributação

De acordo com Silva (2013), a tributação é necessária como uma imposição à sociedade, sendo destinada à manutenção do Estado, com o intuito de prover serviços públicos. Para cumprimento de certas normas de regulamentação, o direito tributário foi destacado das demais áreas do direito, cuja função é de interligar as regras formais da área de tributos e fiscalização.

Segundo Fernandez (2008), o direito tributário atua como ramo do direito público, e sua principal execução é na fiscalização de como ocorre a cobrança de tributos pelo Estado sobre pessoas físicas e jurídicas, analisando a natureza dos tributos, se estão oriundas de forma legal, avaliando se a criação de novas contribuições tem previsão legal e se são constitucionais, de forma se o retorno da interação está coerente.

Amparada pelo direito tributário, a tributação se justifica, de forma legal, na aplicação de contribuições sobre pessoas naturais e jurídicas, pela motivação da assistência em benefício à população (SILVA, 2013).

2.1.1 Processo de Representações Fiscais para Fins Penais (RFFP)

A Secretaria Especial da Receita Federal do Brasil, órgão de origem da federação, subordinado ao Ministério da Economia, estabelece administração de

tributos federais, executa a prática de aplicações fundamentais para impedir a ocorrência de maus comportamentos que levam ao descumprimento por parte dos contribuintes, desencorajando a realização de desobediências das obrigações fiscais, de modo que o Estado possa cumprir seus objetivos (BRASIL, 2021).

Com propósito de regularização no repasse de benefícios econômicos e sociais do País, impedindo infrações como a sonegação fiscal, o contrabando, e descaminho, a contrafação, a pirataria, o tráfico ilícito de entorpecentes e de drogas, o tráfico internacional de armas de fogo e munições, a lavagem ou ocultação de bens, direitos e valores e outros ilícitos aduaneiros, por pertencer subsidiariamente ao Poder Executivo Federal, contribui na elaboração de políticas tributárias (BRASIL, 2021).

Para a elaboração do processo da representação fiscal, o Agente Fiscal de Rendas constata que o ocorrido possa se enquadrar como crime, identificando a existência de indícios contra a “ordem tributária, previstos nos artigos 1º e 2º da Lei 8.137/1990, e contra a Previdência Social, e de contrabando ou descaminho, previstos nos artigos 168-A e 337-A do Decreto-Lei no 2.848/1940 (Código Penal), ou contra administração pública estrangeira, de falsidade de títulos, documentos públicos e papéis, e de “lavagem” ou ocultação de bens, direitos e valores”, é comunicado ao Delegado Regional Tributário, se for considerada, procederá o julgamento do Auto de Infração e Imposição de Multa (AIIM) em primeira instância (BRASIL, 2020).

Após o julgamento é encaminhado ao Ministério Público Federal (MPF) a Representações Fiscais para Fins Penais, instruída com os elementos instrutórios do AIIM, conforme previsto no art. 83 da Lei nº 9.430, de 27 de dezembro de 1996, no Decreto nº 2.730, de 10 de agosto de 1998, e na Portaria RFB nº 1.750, de 12 de novembro de 2018 (BRASIL, 2020).

Conforme a Portaria RFB nº 1.750, 12 de novembro de 2018, estabelece que no documento RFFP contenha:

> Art. 16. A RFB divulgará, em seu sítio na Internet, as seguintes informações relativas às representações fiscais para fins penais, após o seu encaminhamento ao MPF: número do processo referente à representação; nome e número de inscrição no Cadastro de Pessoas Físicas (CPF) ou no

> Cadastro Nacional da Pessoa Jurídica (CNPJ) dos responsáveis pelos fatos que configuram o ilícito objeto da representação fiscal para fins penais; nome e número de inscrição no CNPJ das pessoas jurídicas relacionadas ao ato ou fato que ensejou a representação fiscal para fins penais; tipificação legal do ilícito penal objeto da representação fiscal para fins penais; data de envio ao MPF (BRASIL, 2020).

É importante a análise destes processos de modo que se descubra quais deles tem um perfil de potencial prescrição. A prescrição penal ocorre quando deixa passar o tempo limite, o processo prescreve e não responsabiliza mais o cidadão pelo ato cometido. A fiscalização estabelece pelas autuações dentro da duração, são definidas em (5) cinco anos, contados desde o ato gerador, somada com o prazo da decisão final no administrativo, muitas vezes a prescrição penal acontece (MOREIRA, 2003).

Imagine um cenário hipotético: uma empresa é autuada por sonegação de imposto (MOREIRA, 2003).

- Entre maio de 2013 a março de 2015, foi identificada pela fiscalização que não houve o repasse neste período.
- Em 20 de maio de 2017, foi lavrado o auto de infração. Classificado como crime contra a ordem pública, tipificado no art. 2º, inciso II, da Lei N.° 8.137/90, é formalizado no RFFP contra o responsável pela empresa.
- Em 20 de maio de 2017, foi protocolado o RFFP, anexado ao processo administrativo correspondente até o julgamento definitivo na esfera administrativa.
- Em 19 de maio de 2017, se apresenta o contribuinte, classificando-o como autor da ação ou impugnante. É julgado como procedente o lançamento.
- Em 1º de outubro de 2017, é cientificado o impugnante.
- Em 30 de outubro de 2017, é informado ao impugnante a decisão que recorre ao Conselho de Contribuintes.

- Em 20 de novembro de 2017, é publicado mantendo a decisão de 1ª instância o acórdão do Conselho dos Contribuintes.
- Em 27 de novembro de 2017, é retornado o processo administrativo à Delegacia da Receita Federal (DRF) e o RFFP é encaminhado ao MPF, na mesma data.
- Em 12 de janeiro de 2018, o MPF anuncia a denúncia e a Justiça Federal define instaurada a ação penal.
- Em 6 de março de 2018, o juiz de primeiro grau fez o julgamento como procedente a ação penal e condenou a 9 (nove) meses de detenção o réu, sendo o mínimo legal acrescido da metade. Pelo transcurso do prazo prescricional, é declarada extinta punibilidade, com base na pena *in concreto*.

Conforme o exemplo, a RFFP encaminha ao MPF não pode ser prosseguido, apenas gerando custos desnecessários, sem nenhuma finalidade após a decisão definitiva na esfera administrativa (MOREIRA, 2003).

2.2 *Knowledge Discovery in Databases* (KDD)

Segundo Fayyad et al. (1996), *Knowledge Discovery in Databases*, ou na tradução livre do inglês, Descoberta de Conhecimento em Base de Dados (DCBD), é um processo não-trivial de informações que resulta na geração de informações totalmente proveitoso e compreensíveis para tomada de decisões.

A extração de informações perante a dados disponíveis ao banco de dados, que estão de forma implícita, de maneira automática ou semi-automática. Assim, gerando regras de associações, apresentando padrões com objetivo na tomada de decisões, sendo ela a Mineração de Dados, uma das etapas que constitui a KDD, composta por algoritmos para relacionar os padrões e os dados do banco entre si (FAYYAD et al., 1996).

2.2.1 Fases do KDD

As fases do KDD são compostas por grupos e subgrupos, para melhor entendimento. Tais elas são: pré-processamento, processamento e pós-processamento (FAYYAD et al., 1996).

Segundo Fayyad et al. (1996), na etapa do pré-processamento, consiste na seleção de dados, limpeza dos dados e tratamento de dados. A fase de seleção de dados representa a realização do gerenciamento dos dados coletados, em sendo feita a criação de qual será o conjunto dos dados a ser analisado, considerando a forma que está sendo disponibilizado para a pesquisa.

Os dados dos sistemas podem vir de formas variadas que não contribuam para a finalidade, como atributos com valores incorretos ou desconhecidos, de baixo valor preditivo, entre outros. Por isso, na fase da limpeza dos dados representa as técnicas a serem utilizadas, de modo a obter de forma consistente na geração dos resultados, com intuito de aprimorar a qualidade.

O não aprimoramento pode apresentar problemas futuros, por se tratar de uma etapa delicada e complexa, os resultados não serão precisos ou até inutilizáveis diante da situação. Logo, a fase de tratamento dos dados é destinada a transformar os dados, conforme a técnica utilizada, de modo a obter características proveitosas, com finalidade de preencher as necessidades que quaisquer outros dados, podendo reduzir ou transmutar o comprimento dos dados (BATISTA, 2003).

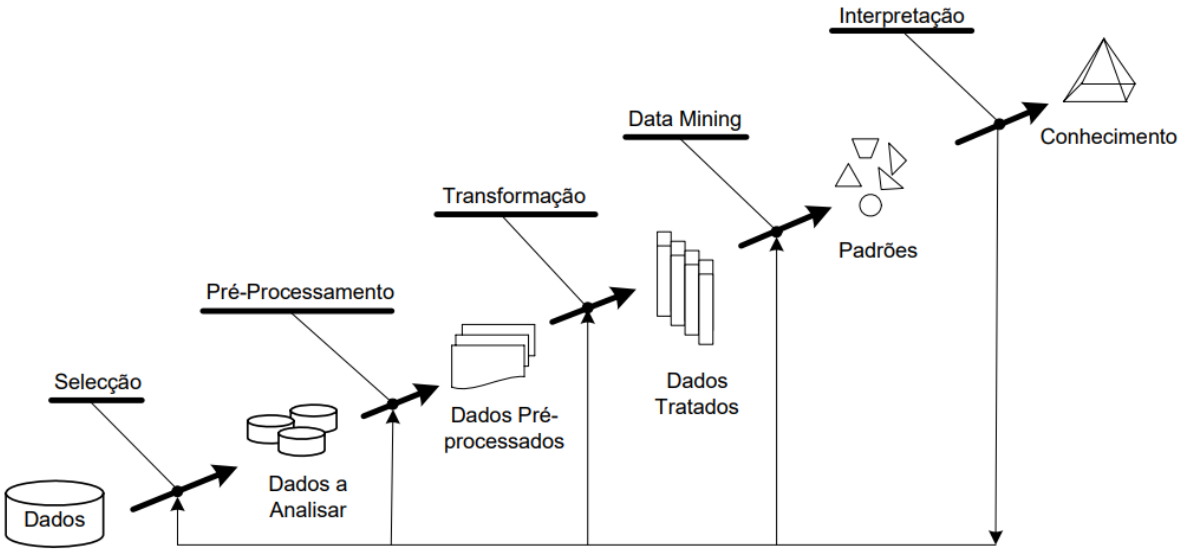
No processamento, representa a fase da mineração dos dados organizados anteriormente. Em que, aplica técnicas e algoritmos computacionais, realizando uma análise exploratória, em busca de novos padrões de interesse, com base nas seleções de modelo e hipótese, podendo ser realizado inúmeras vezes as aplicações, para alcançar objetivos esperados (FAYYAD et al., 1996).

Na etapa do pós-processamento, trata-se da fase da interpretação dos dados, gerados nas etapas anteriores. Representada como a última fase do KDD, é realizado a interpretação e avaliação dos resultados adquiridos e padrões minerados, verificando o obtido atende as necessidades, proposto inicialmente, podendo regressar etapas antecedentes, alterando um ou mais, para novas

iterações (caso necessário), gerando outras possibilidades para novas referências de pesquisas de dados. (RABELO e CAMPOS, 2014).

A sequência iterativa em etapas do KDD, ilustrados na figura 1.

Figura 1. Etapas do Processo *Knowledge Discovery in Databases* (KDD)



Fonte: Adaptado de Fayyad et al., (1996).

2.2.2 Mineração de Dados

Mineração de Dados, ou na tradução livre do inglês, *Data Mining*, é uma das etapas do KDD. Considerada uma técnica que desfruta do auxílio dos algoritmos computacionais, após a extração de dados, anteriormente desconhecida, tem o intuito na descoberta de padrões úteis para tomada de decisão (Cabena et al., 1998). Segundo Harrison (1998), a mineração de dados abrange a pesquisa examinada por métodos analíticos ou semi-analíticos, relevantes na solução da investigação de grandes volumes de dados, para encontrar modelos e regras consideráveis.

A principal justificativa na aplicação da mineração de dados é na previsão futura a partir dos dados, da imensa parcela de dados aglomerados eletronicamente. A busca da descoberta de informações, não requerendo suposições prévias para dar

início ao estudo com os dados. Após a etapa, gera-se um arquivo que passa por avaliação e interpretação para chegada de uma conclusão válida e decisiva (DANTAS et al., 2008).

A mineração de dados inclui mais de uma área correlatas no seu processo, as essenciais a serem destacadas são as tecnologias de banco de dados, estatísticas e inteligência artificial, além de áreas mais específicas como “reconhecimento de padrões, sistemas baseados em conhecimento, recuperação da informação, computação de alto desempenho e visualização de dados” (CARDOSO e MACHADO, 2008).

2.2.3 Tarefas e técnicas de mineração de dados

Na mineração de dados, as tarefas são divididas em análise preditivas e descritivas, realizadas dentro do KDD, de forma automática ou semi-automática. As tarefas preditivas procuram, através de variáveis conhecidas, encontrar valores ainda desconhecidos ou futuros. Já as tarefas descritivas procuram padrões para descrever dados (CASTRO; FERRARI, 2016).

Nas tarefas preditivas são compostas pela classificação e regressão (ou estimativa), se enquadrando como as principais. Já na descritiva são compostas por regras de associação, agrupamento (ou segmentação, ou *clustering*), sumarização e detecção de desvios, são as mais relevantes (CÔRTES et al., 2002).

Na classificação, após ser realizada um estudo sobre um conjunto de dados, e literalmente, classificá-lo, após novos estudos utilizando outro conjunto de dados, contendo as mesmas variáveis (ou não, necessariamente), é realizada a correlação deste novo conjunto de dados perante o que já foi classificado anteriormente, originado o nome de classificação. Contemplando seu objetivo principal, em encontrar os padrões e separar os dados, realizando o treinamento de aprendizagem, prevendo comportamentos futuros, procedentes de um banco de dados novo ou desconhecido (CASTRO; FERRARI, 2016).

Alguns exemplos didáticos podem ser destacados na utilização, como na detecção de *SPAM* (*Sending and Posting Advertisement in Mass*) na caixa de correio

eletrônico, a partir do remetente, conteúdo e frequência de envio ou classificação de galáxias em virtude do seu formato (VILARINHO, 2017).

A regressão, ou estimativa, é utilizada para prever um valor desconhecido, com bases na saída valores reais disponíveis, retratado similar da função de classificação, com a diferença no modo que os dados são avaliados, se tratando de atributos discretos, enquanto a regressão investiga atributos de categoria contínua, visando predizer tais valores (QUEIROZ, 2016).

Regras de associação apresentam aplicabilidade na identificação da frequência dos dados transacionais, perante os elementos compostos em cada processo, com intuito no reconhecimento de padrões. Exemplificando, pode-se evidenciar o processo de carrinho de compras, conhecido com o termo “Análise de cesta de mercado” (em inglês “*Market Basket Analysis*”), em que, os itens compostos e a quantidade por cada item associando por certos itens de compra, outros também são comprados juntos (CARDOSO, 2017).

O agrupamento, também conhecida como segmentação, ou *clustering* (clusterização), tem a função na divisão de grupos em subgrupos, de maneira segmentar, separados por características particulares, com intenção que cada grupo tenha similaridade entre si e diferença aos padrões dos demais grupos existentes (QUEIROZ, 2016).

Para Cardoso (2017), a sumarização “consiste em encontrar uma descrição mais simples para um conjunto de dados menor do que o seu conjunto de dados original.” Com a definição, atribui a tarefa como a forma clara, compreensível e compactada do conjunto de dados em comparação com o inicial, com intuído de ser mais fácil a visualização dos resultados finais.

Detecção de desvios é designado o tratamento quando os valores não apresentam padrões considerados comuns, destacando valores incomuns do contexto (QUEIROZ, 2016).

Existem várias técnicas de mineração de dados, que são organizadas e executadas conforme a tarefa escolhida, com propósito de alcançar certos objetivos, conforme a informação que deseja chegar.

Segundo Queiroz (2016), pode-se executar uma técnica para com vários algoritmos, e derivar resultados diferentes, podendo também, as chances de técnicas e algoritmos distintos executados em outras tarefas, resultar em valores e interpretações similares. Por isso, a escolha das técnicas e algoritmos deve ser fundamental e de forma cautelosa, visando os resultados que pretende chegar.

As principais técnicas e tarefas de mineração de dados e respectivos algoritmos estão representadas no quadro 1.

Quadro 1. Relação entre tarefas, técnicas e algoritmos de mineração de dados.

Análise	Tarefas	Técnicas	Algoritmos
Preditiva	Classificação	Árvore de decisão; análise bayesiana; análise de vizinhança; redes neurais; algoritmos genéticos.	*Support Vector Machine; Algoritmo C4.5; CART; KNN;* Classificadores *Bayesianos*; J48.
	Regressão/ Estimativa	Regressão linear; regressão múltipla; regressão não linear; regressão logística; regressão de poisson; redes neurais.	*Backpropagation; Multilayer Perceptron;* Lógica nebulosa.
Descritiva	Regras de Associação	Mineração de regras de associação.	*Apriori*; Algoritmo FP-*Growth; Eclat. ;Direct Hashing And Pruning; Dynamic Itemset Counting.*
	Agrupamento/ Clusterização	Método particionamento; modelagem de regras; métodos de clustering baseados em modelos (abordagem estatística e redes neurais).	*K-means;K-modes; K-prototypes; Fuzzy K-means; K-medoids; Canopy; Cobweb;*
	Sumarização	Algoritmos genéticos.	Modelos Estatísticos;

			Cubos de Dados; HAWB.
	Desvios	Ferramentas de consulta e técnicas de estatística; indução por árvores de decisão.	Apriori; Algoritmo c4.5.

Fonte: Elaborado pelo autor.

2.2.4 Árvore de decisão

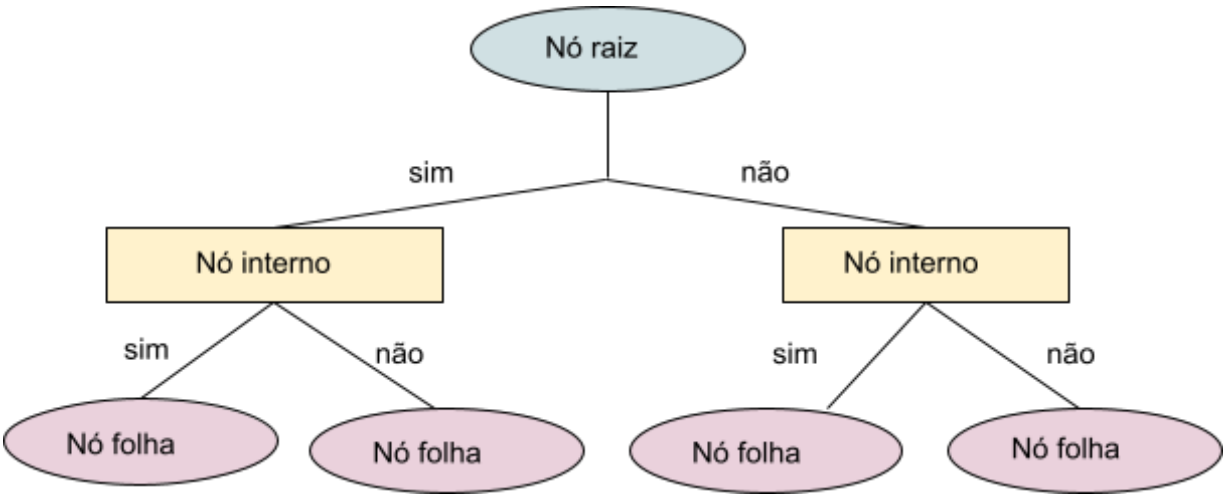
Da diversidade dos algoritmos para serem usados e estudados na mineração de dados, a árvore de decisão é um dos principais na área e serão descritos a seguir.

O método de árvore de decisão é composto pela representação, vinda da botânica (ou biologia vegetal), dos galhos das árvores. Em sua visualização, retrata-se um fluxograma (*flow-chat*), constituído por um nó raiz, nós folhas, nós internos. (CASTRO; FERRARI, 2016).

De acordo com Castro e Ferrari (2016), são atribuídas regras para cada nó dentro de uma árvore. Considerado o nó mais alto de uma árvore, é denominado como nó raiz, cada árvore possui apenas único nó raiz que divide os resultados possíveis. Por meio da divisão, são relacionadas as ramificações que se dizem a respeito das possibilidades executadas dentro da árvore, denominadas de nós internos. E por fim, os nós folhas, podendo ser chamado de nó terminal ou nó externo, são os nós finais que não apresentam ramificações na estrutura da árvore.

Exemplo de fluxograma de árvore de decisão para melhor compreensão, ilustrada na figura 2.

Figura 2. Exemplo de visualização do fluxograma de uma árvore de decisão



Fonte: Elaborado pelo autor.

Essa estrutura serve para conduzir a pesquisa por meio das probabilidades para melhor escolha, por este motivo na construção da árvore é necessário medir a pureza dos nós para verificar qual será o atributo mais qualificado, definindo o quão semelhante um nó é da classe (CÔRTES; PORCARO, 2002).

O método trata-se de algoritmo de aprendizagem de máquina, que realiza a técnica dividir para conquistar. Após a inserção de dados, e o processamento (geralmente oriunda por *software*), é obtido os dados da mineração, em seguida realiza as ramificações para destinar por meio dos resultados, reconhecendo e relacionando os grupos mais relevantes (AMORIM, 2006).

São verificados todos os nós para a expansão da árvore e através do resultado é realizada a entropia. O método da entropia realiza cálculos que definem a organização de nós dentro da árvore, seguindo pela variabilidade de suas classes dos dados que estão no *database* (CÔRTES; PORCARO, 2002).

O cálculo da entropia é representado na fórmula 2.2, na qual, X_{tr} está relacionado ao conjunto de dados destinado ao treinamento e P_{ck} é à probabilidade de ocorrer a classe $_{ck}$ em X_{tr}.

$$E(X_{tr}) = \; -\Sigma_{ck=1}^{k} P_{ck}\, log_2(P_{ck}) \qquad (2.2)$$

A entropia pode ser alta quando o esforço para organizar os dados pertencentes às classes se encontra alto, necessitando de empenho elevado causado pela desordem dos dados, caso contrário, será baixa pela baixa desordem dos dados (SILVA; PERES; BOSCARIOLI, 2016).

A forma que se usa um nó para particionar um conjunto de dados é relacionado pelas mesmas possibilidades existentes da classe, na qual, o nó assumirá. Para as partições serem consideradas como puras, a entropia tem que ser igual a 0 em todas as partições, significa que serão classificadas cada partição em uma única classe. Em outro caso, se a entropia apresentar maior que 0, podendo haver o valor em qualquer partição, resultará em mais de uma classe para a partição. Perante a situação, para fazer com que as partições se tornem puras, será aplicada um esforço para tornarem puras, sempre partindo das atuais, denominada de análise da informação necessária (SILVA; PERES; BOSCARIOLI, 2016).

O cálculo da informação necessária é apresentado na fórmula 2.3

$$IN_A(X_{tr}) = \Sigma_{i=1}^{\upsilon} \frac{|X_{tr\,i}|}{X_{tr}} E(X_{tr}) \quad (2.3)$$

na qual:

IN_A: informação necessária

$E(X_{tr})$: entropia do conjunto de dados de treinamento

i: conjunto de dados de treinamento da partição

υ: total de classes possíveis

No mesmo seguimento da informação necessária, o conceito de ganho de informação estabelece para analisar o quanto o nó ganha colocando um atributo em específico, com objetivo de particionar a árvore em um ponto específico (SILVA; PERES; BOSCARIOLI, 2016).

O cálculo do ganho de informação é apresentado à fórmula 2.4, no que se diz a respeito da subtração da entropia de todo o conjunto pela entropia de cada atributo.

$$G(A) = E(X_{tr}) - IN_A(X_{tr}) \quad (2.4)$$

Perante o modelo de classificação de árvore de decisão completa, com objetivo de realizar as predições e tomadas de decisão, o modelo disponibiliza a sua eficácia, dos dados que foram computados, qual é a sua precisão de acurácia. Sendo representado pela porcentagem, mostram quais foram os níveis de classificações corretas e incorretas (SILVA; PERES; BOSCARIOLI, 2016).

Além da porcentagem da acurácia, a matriz de confusão é outro método que se diz a respeito das avaliações das classificações. Segundo Silva, Perez e Boscarioli (2016), a matriz de confusão é uma matriz $n\,x\,n$ composta por classificações que estão corretas no seu grupo pertencente, representada pelos valores na diagonal principal da matriz, e classificações que foram reconhecidas e classificadas em outros grupos.

Destinando sempre por conta do pesquisador, respondendo perguntas do gênero: Qual será o objeto de saída? Qual a motivação da pesquisa? O que deseja? Em geral, usa-se o arquivo no formato planilha (ou variações), composto por linhas e colunas, e a última coluna é destinada para a saída (CÔRTES; PORCARO, 2002).

A motivação de se utilizar a árvore de decisão é a precisão e rapidez na sua construção em comparação a outros métodos de classificação. E a desvantagem é a grande quantidade de dados que precisa ser trabalhada para conseguir resultados em trabalhos mais complexos (AMORIM, 2006).

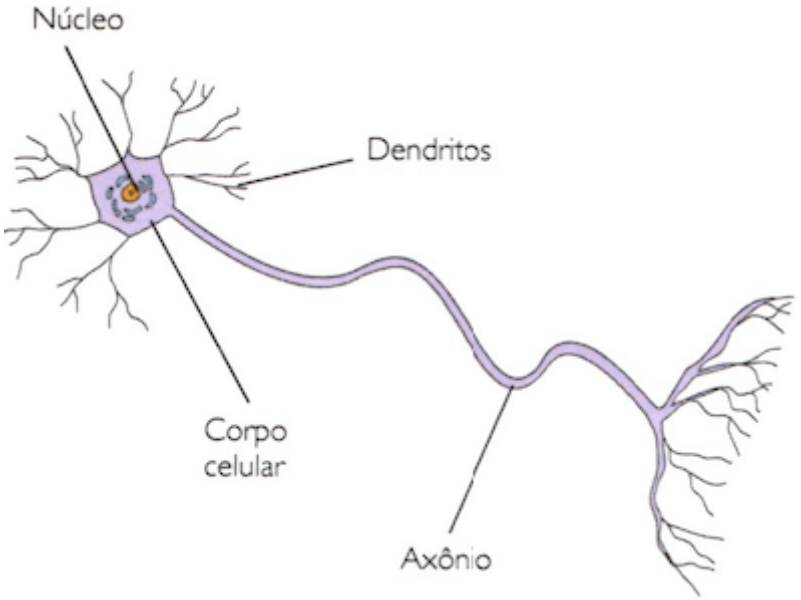
2.2.5 Redes neurais artificiais

Como referência da anatomia humana, Redes Neurais Artificiais (RNAs) são inspiradas no modo de se comportar os neurônios humanos, capazes de realizar tarefas cognitivas, como o aprendizado, associação, abstração e generalização, anteriormente desempenhadas apenas biologicamente, pelas redes neurais humanas, ao longo do processo de anos adquiridos experiências de vivência, capazes de propor soluções diante dos problemas já enfrentados. (QUEIROZ, 2016).

Seguindo pela anatomia, compostas por 3 (três) principais partes, o neurônio humano contém os dendritos, o axônio e o corpo celular. Os dendritos têm a finalidade de receber todos sinais, os sinais realizam uma viagem para o corpo

celular, passam pelo axônio (seu tamanho por alcançar até 1 (um) metro de comprimento) até o ponto entre dois neurônios, considerado o ponto de comunicação, denominado de sinapses, apresentada na figura 3 (RUSSELL; NORVIG, 2013).

Figura 3. Estrutura do neurônio humano.



Fonte: "Células nervosas" em Só Biologia (2021).

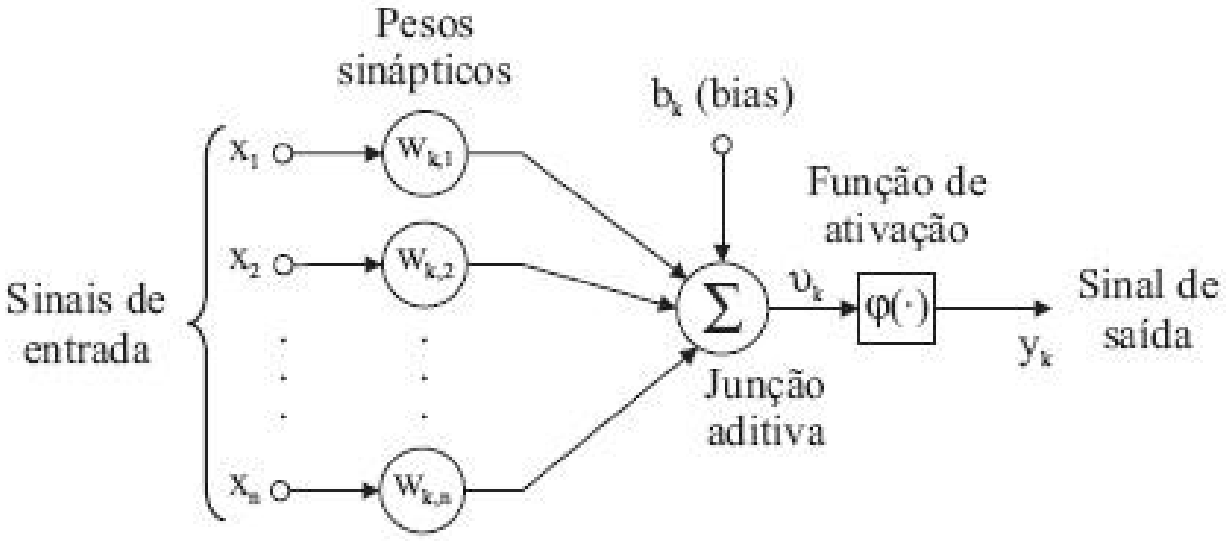
As RNAs são estruturalmente a interligação de processamentos básicos, compostas por várias unidades (titulada como neurônios), com a finalidade de realizar cálculos de determinadas funções matemáticas, chamadas funções de ativação. Os neurônios são localizados em camadas e relacionados por várias conexões, com o propósito de manter o conhecimento adquirido e relacionadas por uma catalogação (geralmente denominada por pesos) retratada diante ao modelo, considerando as entradas recebidas por cada neurônio (AMORIM, 2006).

Para melhor compreensão, as RNAs podem ser representadas através de gráficos, em que encontramos cada nó como um neurônio e a aresta a ligação sináptica. Segundo Haykin (2001), um neurônio assume uma ou mais entradas, para cada entrada x que foi assumida um peso w será atribuída, denominada de peso sináptico. A função de somatória Σ serão a soma de todos os valores das entradas,

após o processo é aplicada a função de ativação φ destinada a saída do resultado. A saída que foi gerada pode assumir o seu valor resultado da aplicação final ou entrada de um outro neurônio.

Para que o valor da função de ativação seja mais adaptado aos dados, a função bias aplica a técnica para que o valor sofra diminuição, caso seja negativo e aumento, quando for positivo, antes da função de ativação receber o valor, demonstrada na figura 4 (HAYKIN, 2001).

Figura 4. Estrutura de um neurônio artificial



Fonte: Haykin (2001).

A equação matemática 2.5 e 2.6 representada pelo processo da figura 3 se dá por:

$$u_k = \Sigma_{j=1}^{m} w_{kj} x_j \tag{2.5}$$

$$y_k = \varphi(u_k + b_k) \tag{2.6}$$

Na qual:

k: neurônio

u_k: saída da combinação linear

w_{kj}: peso sináptico da entrada j

x_j: entrada j

y_k: saída final

É utilizando o processo de aprendizagem, conhecido também como treinamento, que se realiza a obtenção dos dados na etapa de agregação conhecimento perante o ambiente (HAYKIN, 2001).

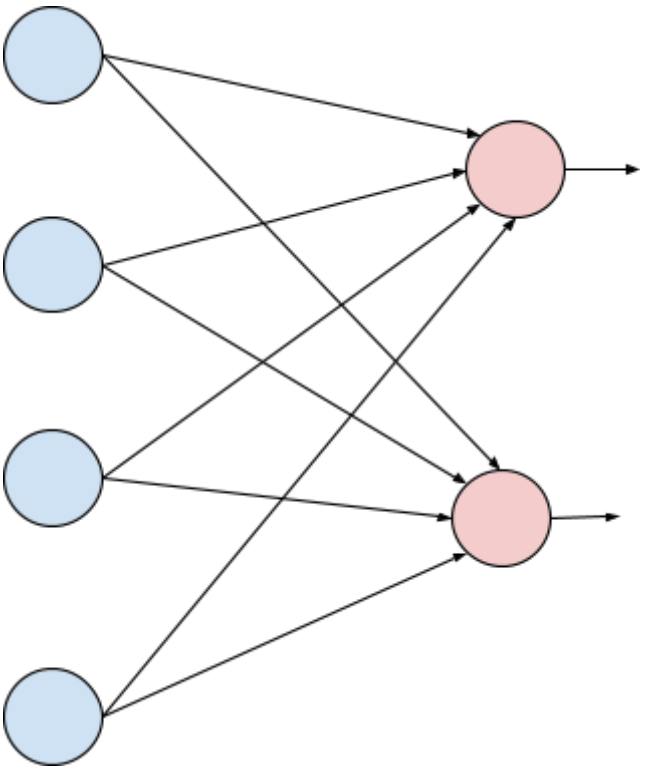
Na fase de adquirir conhecimento, as interligações vão sendo reajustados, dependendo do peso das conexões de rede de cada conhecimento extraído, que possa ser evidenciado de forma autossuficiente, a cada momento, interações são realizadas repetidamente para definir os parâmetros e realizar inúmeras repetições, para ser internamente estabelecida e decidida (AMORIM, 2006).

A argumentação de Haykin (2001) de se utilizar RNA é o nível de aprendizagem do modelo e melhora na atuação, conforme for inserida no ambiente de pesquisa, conseguindo proporcionar experiências através de processos de reproduzir apresentações dos dados à rede.

A RNA possui uma alta tolerância a ruídos nos dados, com uma boa escalabilidade e interpretabilidade bem melhorada, por conta dos vários algoritmos desenvolvidos para extração de regras de classificação de redes neurais. Por se tratar de um algoritmo robusto, lidando bem com problemas complexos. A desvantagem é a obrigação de muitos parâmetros e valores iniciais dos pesos, o processo do treinamento demanda bastante tempo, requerendo força computacional, em comparação a outros, e necessita de pessoas bem capacitadas para interpretar os resultados (QUEIROZ, 2016).

Existem diversas estruturas em RNAs, mas as principais são *Single Layer Perceptron* (SLP) e a *Multilayer Perceptron* (MLP). Na *Single Layer Perceptron* (MLP) em sua estrutura é formada por uma única camada que são separados os neurônios paralelamente. Possui uma única saída e aceita inúmeras entradas, sendo considerada a SLP o modelo de classificação para aplicações mais simples ilustrada na figura 5.

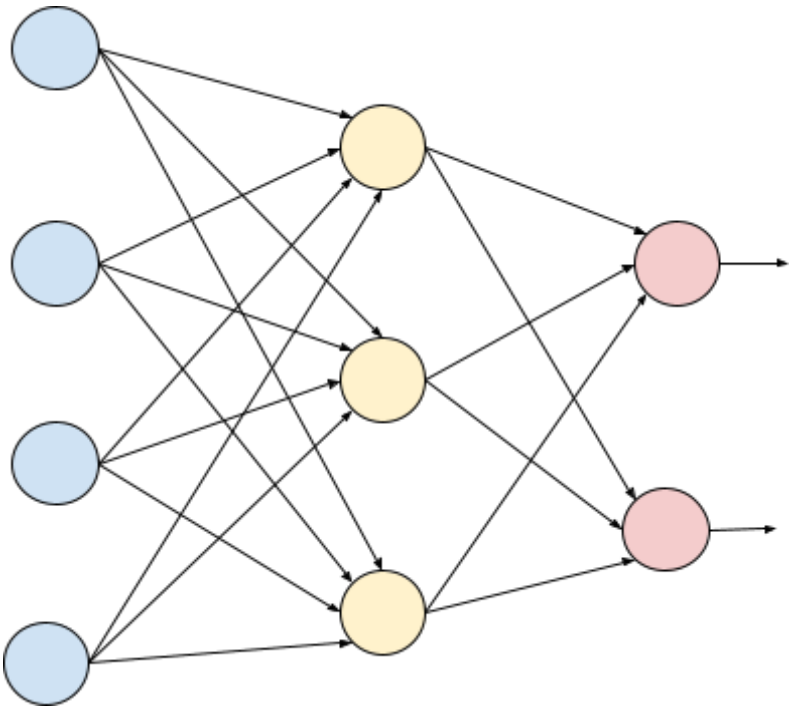
Figura 5. Estrutura do *Single Layer Perceptron*



Fonte: Elaborado pelo autor.

O *Multilayer Perceptron* (MLP) em sua estrutura é formado por multi camadas, são separadas por mais de uma camada dos neurônios, denominadas de camadas ocultas. Em sua camada oculta, por meio das sinapses, os neurônios têm-se valores de resultados anteriores como entrada de camadas anteriores. Sendo considerada a MLP o modelo de classificação para aplicações mais complexas pelo alto poder de processamento e a forma de extração dos resultados são de maneira mais expressiva, ilustrada na figura 6 (Haykin, 2001).

Figura 6. Estrutura do *Multilayer Perceptron*



Fonte: Elaborado pelo autor.

2.2.6 WEKA *(Waikato Environment for Knowledge Analysis)*

Segundo Witten et al. (2016), desenvolvida pela Universidade Waikato, na Nova Zelândia, o *Waikato Environment for Knowledge Analysis* (WEKA) é um *software* de código aberto, destinado a estudo, pesquisa e testes na área de processamento de dados, contendo ferramentas e algoritmos para a mineração de dados.

Possível ser realizadas tarefas de mineração de dados, como classificação, regressão, agrupamento e associação. De fácil operação, contém uma documentação base fundamental para o entendimento, de forma gratuita, possuindo uma interface gráfica simples, possível manipular e visualizar os dados, além de mostrar relatórios bem detalhados.

2.3 Estudos correlatos

Há vários trabalhos relevantes disponíveis relacionados com o uso da inteligência artificial aplicado na mineração de dados em descobertas com dados públicos. O trabalho do Silva (2020) realizou a pesquisa em análise de licitações do Estado de Goiás, usando o *software* WEKA, realizou aplicações das técnicas de mineração com objetivo de relacionar indícios de irregularidades nas licitações da Controladoria Geral do Estado de Goiás.

Em seu trabalho Madeira (2015) desenvolveu, por meio de técnicas de análise de dados e mineração de textos, identificar, a partir da descrição dos serviços prestados, notas eletrônicas emitidas incorretamente a fim de detectar fraudes e melhorar o planejamento de fiscalizações.

Braga (2010) em seu trabalho teve como objetivo a identificação de indícios de infração à legislação tributária, por uma proposta de rede neural ou regressão linear, a criação de uma lista de possíveis contribuintes fraudulentos, por meio dos perfis passados, contendo os casos que ajudem na identificação de futuros.

3 MÉTODOS E MATERIAIS

Essa seção visa descrever os materiais e métodos utilizados na construção desse trabalho, bem como suas etapas e processos na realização das atividades do trabalho.

3.1 Métodos

A pesquisa tem objetivo de trabalhar com profundidade sobre determinado assunto, explorando com riqueza as observações do determinado grupo, ouvindo e colhendo por meio da percepção dos sentimentos sobre determinado assunto (WAZLAWICK, 2014).

Segundo Gil (2017), a pesquisa surge quando não se tem informações disponíveis que possam contribuir para a solução de problemas, ou quando o problema apresentado ainda não tem informações suficientes, necessitando de estudos aprimorados com técnicas, em busca da solução melhorada.

A pesquisa busca aumentar o conhecimento humano sobre determinados assuntos e elementos da sua construção e funcionamento. As pesquisas científicas podem ser classificadas com uma vasta seleção de critérios de acordo com a sua natureza, abordagem, objetivos e procedimentos técnicos (GIL, 2017).

Quanto à natureza, o trabalho é classificado como aplicado, por manifestar o aglomerado de informações e conhecimento com objetivo de buscar a solução do problema, a fim de direcionar a aplicação prática (NASCIMENTO; SOUSA, 2015).

Quanto à abordagem, essa pesquisa é caracterizada como quantitativa. Segundo Triviños (1987), a pesquisa é quantitativa quando a realização das análises dos resultados é proposta por meio de base em técnicas estatísticas padronizadas e ordenadas, para contribuir na interpretação das respostas.

Quanto aos objetivos, esta pesquisa é descritiva e exploratória. De acordo com Wazlawick (2014), na pesquisa descritiva tem o objetivo de buscar a identificação de quais situações, eventos, atitudes ou opiniões que serão manifestados, descrevendo os fatos como são, sem necessidade ainda de obter formas para ser explicada.

Segundo Gil (2017), a pesquisa exploratória considerando muitas vezes um processo de pesquisa inicial de algo mais longo, requer do autor examinar um conjunto de fenômenos, para poder encontrar falhas, assim tendo com uma base concreta formada, tornando a pesquisa mais elaborada.

Quanto aos procedimentos técnicos é uma pesquisa bibliográfica e experimental. Uma pesquisa bibliográfica se baseia em pesquisas e análise de trabalhos, obras, artigos, tese, livros e outras publicações de estudos já lançados. É o passo inicial para o projeto, porém por si só não produz novo conhecimento (WAZLAWICK, 2014).

A pesquisa experimental é quando um pesquisador introduz uma nova técnica em um determinado ambiente e analisa o desenvolvimento, ou seja, a manipulação de um aspecto da realidade pelo pesquisador (WAZLAWICK, 2014).

Figura 7. Estrutura do método da análise de dados.

Fonte: Elaborado pelo autor.

O trabalho é composto por 4 (quatro) etapas, iniciando na revisão bibliográfica, seleção dos dados, pré-processamento e finalizando na análise preditiva, modelo empregado ilustrado na figura 7.

A primeira etapa diz a respeito da revisão bibliográfica, consistiu no conhecimento de trabalhos, livros e teses semelhantes na área de ciência de dados, mineração de dados, KDD, Penal Tributário e Fiscalização Tributária, definindo a melhor forma que seria aplicado no trabalho, com propósito de encontrar a solução do problema.

Na segunda etapa foi realizada a seleção dos dados. Utilizando as Representações Fiscais Para Fins Penais encaminhadas ao Ministério Público, documento disponibilizado publicamente pela Receita Federal, dados do período do dia 14 do mês de novembro de 2018 há 31 do mês de outubro de 2020, foi efetuado a seleção dos registros de forma aleatória e registrados em uma planilha eletrônica.

Na terceira etapa, pré-processamento de dados foi realizada a limpeza dos dados, de maneira incluir os dados mais relevantes, foram organizados e estruturados de forma estratégica, eliminando dados incompletos e redundantes, visando os objetivos finais.

Na quarta e última etapa foi realizada análise preditiva dos processamentos de dados, aplicados técnicas de mineração para a identificação e classificação de duas situações, previsão do crime e previsão de duração do procedimento administrativo.

3.2 Materiais

Foi utilizado dados do RFFP públicos coletados no site da Receita Federal e o *software* WEKA, operado no notebook da marca Asus das seguintes configurações:

- Notebook Asus VivoBook 15 X510UR, processador Intel® Core™ i5 de 8ª geração (8250 U) com 8GB de RAM;
- 1TB HD;
- *Windows* 10 *Home*;
- WEKA (*Waikato Environment for Knowledge Analysis*).

4 RESULTADOS E DISCUSSÃO

Esta seção visa apresentar os resultados obtidos ao longo deste trabalho. Para a obtenção dos resultados, foi utilizado algoritmos de árvore de decisão, realizado o algoritmo J48, e em redes neurais, realizado com algoritmo *Multilayer Perceptron*.

Por meio do *software* WEKA, foram realizados os testes utilizando dois métodos, "*Percentage split*" e "*Use training set"*. Em que, o primeiro, realiza a divisão em duas partes do *dataset*, na qual, separa os dados reservando alguns para o treinamento e outros para testes. Já o segundo, o treinamento e o teste fazem parte do *dataset*, ou seja, toda a base de dados é utilizada sem a separação.

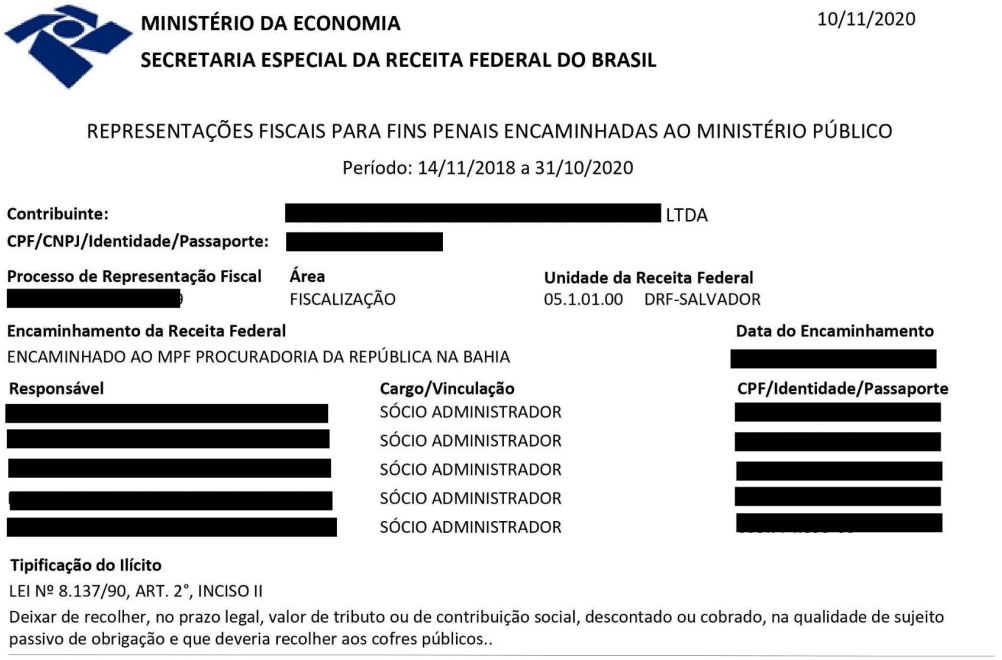
4.1 Seleção e pré-processamento dos dados

Os dados obtidos foram associados a autuações recrutadas pela Receita Federal, com o fito de representações penais para fins fiscais. Esses dados foram disponibilizados no site da Receita Federal, através do endereço eletrônico https://www.gov.br/receitafederal/pt-br, e através do site, ter acesso às RFFP encaminhadas ao MPF, documento de carácter público, contendo dados que se enquadraram por infringir leis de natureza da ordem tributária, contra a Previdência Social, e de contrabando ou descaminho.

O documento contém o total de 5937 (cinco mil novecentos e trinta e sete) páginas, com aproximadamente 17500 (dezessete mil e quinhentos) registros de autuações, entre pessoas físicas e jurídicas, do período do dia 14 do mês de novembro de 2018 há 31 do mês de outubro de 2020.

Como uma parte apresentada na figura 8, relativa a uma empresa que foi autuada e a representação foi encaminhada ao Ministério Público.

Figura 8. Documento RFFP encaminhadas ao Ministério Público



MINISTÉRIO DA ECONOMIA
SECRETARIA ESPECIAL DA RECEITA FEDERAL DO BRASIL

10/11/2020

REPRESENTAÇÕES FISCAIS PARA FINS PENAIS ENCAMINHADAS AO MINISTÉRIO PÚBLICO

Período: 14/11/2018 a 31/10/2020

Contribuinte: [redacted] LTDA

CPF/CNPJ/Identidade/Passaporte: [redacted]

Processo de Representação Fiscal	**Área**	**Unidade da Receita Federal**
[redacted]	FISCALIZAÇÃO	05.1.01.00 DRF-SALVADOR

Encaminhamento da Receita Federal
ENCAMINHADO AO MPF PROCURADORIA DA REPÚBLICA NA BAHIA

Data do Encaminhamento [redacted]

Responsável	**Cargo/Vinculação**	**CPF/Identidade/Passaporte**
[redacted]	SÓCIO ADMINISTRADOR	[redacted]
[redacted]	SÓCIO ADMINISTRADOR	[redacted]
[redacted]	SÓCIO ADMINISTRADOR	[redacted]
[redacted]	SÓCIO ADMINISTRADOR	[redacted]
[redacted]	SÓCIO ADMINISTRADOR	[redacted]

Tipificação do Ilícito
LEI Nº 8.137/90, ART. 2°, INCISO II
Deixar de recolher, no prazo legal, valor de tributo ou de contribuição social, descontado ou cobrado, na qualidade de sujeito passivo de obrigação e que deveria recolher aos cofres públicos..

Fonte: Receita Federal, Ministério Da Economia (2020).

De forma aleatória, foram escolhidas 500 (quinhentos) registros transferidos dados do arquivo original em formato PDF (*Portable Document Format*) para planilha eletrônica e convertendo em CSV (*Comma Separated Values*) e ARFF (*Attribute Relation File Format*) padrão de arquivo reconhecido pelo *software* WEKA. Optando por oito atributos, foram escolhidos:

- área;
- contribuinte;
- UF (Unidade Federativa);
- quantidade de sócios;
- ano de representação;
- ano de encaminhamento;
- duração (em anos);
- crimes.

Os contribuintes escolhidos são separados por categorias, atributos por siglas para melhor compressão, detalhados como:

- AP (Associação Privada);
- CPF (Cadastro de Pessoa Física);
- EI (Empresário Individual);
- EIRELI (Empresa Individual de Responsabilidade Limitada);
- LTDA (Sociedade Limitada);
- S/A (Sociedade Anônima).

A área é representada em 3 (três) possíveis categorias de vistorias realizadas, tais são: Administração Aduaneira, Administração Tributária e Fiscalização.

A quantidade de sócios é representada pela quantidade de representantes que a empresa possui, além do responsável legal. Identificados no mínimo 0 (zero) e no máximo 14 (quatorze), são incluídos no arquivo ARFF como números inteiros, atribuídos para cada empresa. Os crimes cometidos foram selecionados referente à Lei Nº 8137 ou Código Penal.

Os valores dos atributos correspondem à forma que foram preenchidos no arquivo ARFF, foram consideradas 2 (duas) categorias: nominal e *numeric*. Atributos *numeric* (quantitativos ou numéricos) representa quantidade, como é o caso do número de sócios que a empresa é composta. Já atributos nominais são representados por categorias distintas dentro de um grupo, como é o caso da área, sendo representado como Administração Aduaneira, Administração Tributária e Fiscalização, ou seja, somente um desses valores possíveis pode ser categorizado na área.

No quadro 2, foram detalhados os demais atributos com mais especificidade.

Quadro 2. Descrição detalhada dos atributos

Variável	Descrição	Valor
Área	Relacionada à categoria de vistoria realizada (Administração Aduaneira; Administração Tributária; Fiscalização).	Nominal
Contribuintes	Categoria do contribuinte (AP; CPF; EI; EIRELI; LTDA; S/A).	Nominal

UF	Estado da unidade da Receita Federal que se encontra.	Nominal
Quantidade de Sócios	Números de sócios existentes.	*Numeric*
Ano de Representação	Ano de representação do processo que foi autuado.	Nominal
Ano de Encaminhamento	Ano de encaminhamento do processo ao Ministério Público Federal.	Nominal
Duração	Duração em anos desde a representação até o encaminhamento.	Nominal
Crimes	Categoria do crime cometido: Lei Nº 8137 ou Código Penal.	Nominal

Fonte: Elaborado pelo autor.

O trabalho realizou classificações em dois alvos diferentes. No primeiro experimento foi optado como alvo a análise para identificar o crime cometido, no segundo classificou como alvo o tempo de duração do procedimento administrativo.

No primeiro experimento é uma classificação que irá identificar pelos atributos escolhidos se levam para um crime contra a ordem tributária, a Lei Nº 8137, ou crime do Código Penal. Já no segundo, o experimento tem como objetivo prever ou classificar uma situação de representação a partir dos dados, em especial, o ano de representação e ano de encaminhamento, quanto tempo duraria o processo.

Foram utilizadas duas técnicas, árvore de decisão e redes neurais, implementadas no software WEKA. Para cada técnica a realização dos crimes e duração foram pelo mesmo dataset. Em uma situação, a duração do procedimento administrativo são atributos e na outra atributos alvos, e atributos que vão virar alvo nos crimes e em outra situação é retratado como atributo.

Na figura 9 representa o arquivo em ARFF com todos os atributos e registros no sentido de identificar o crime cometido. O crime aparece em último atributo, destinado como alvo no arquivo.

Na figura 10 que tem como objetivo identificar o tempo de duração em anos de um processo administrativo em relação à representação fiscal. Sendo representado em último, o atributo é destinado como alvo no arquivo ARFF.

Figura 9: Arquivo ARFF com oito atributos com alvo nos crimes

```
@relation 'dataBase - CRIMEScsv'

@attribute Area {'ADMINISTRACAO ADUANEIRA',FISCALIZACAO,'ADMINISTRACAO TRIBUTARIA'}
@attribute Contribuinte {EIRELI,LTDA,EI,CPF,AP,SA}
@attribute UF {PR,BA,MS,GO,SP,RS,MG,SC,RO,RJ,DF,RR,PA,MT,AM,PE,PB}
@attribute QuantSocio numeric
@attribute AnoRepresentacaoFiscal numeric
@attribute AnoEncaminhamento numeric
@attribute Duracao/Ano {zero,'acima de quatro',um,dois,tres,quatro}
@attribute Crimes {'CODIGO PENAL','LEI 8137'}

@data
'ADMINISTRACAO ADUANEIRA',EIRELI,PR,1,2019,2019,zero,'CODIGO PENAL'
FISCALIZACAO,LTDA,BA,5,2010,2020,'acima de quatro','LEI 8137'
FISCALIZACAO,LTDA,BA,1,2018,2019,um,'CODIGO PENAL'
'ADMINISTRACAO ADUANEIRA',EIRELI,PR,0,2019,2019,zero,'CODIGO PENAL'
FISCALIZACAO,LTDA,BA,1,2019,2019,zero,'CODIGO PENAL'
FISCALIZACAO,LTDA,BA,1,2019,2019,zero,'LEI 8137'
FISCALIZACAO,EIRELI,MS,1,2018,2019,um,'LEI 8137'
'ADMINISTRACAO ADUANEIRA',LTDA,GO,0,2019,2020,um,'CODIGO PENAL'
FISCALIZACAO,LTDA,MS,3,2018,2019,um,'CODIGO PENAL'
'ADMINISTRACAO ADUANEIRA',LTDA,PR,0,2018,2019,um,'CODIGO PENAL'
                ...
                ...
                ...
```

Fonte: Elaborado pelo autor.

Figura 10. Arquivo ARFF com oito atributos com alvo na duração

```
@relation 'dataBase - DURACAOcsv'

@attribute Area {'ADMINISTRACAO ADUANEIRA',FISCALIZACAO,'ADMINISTRACAO TRIBUTARIA'}
@attribute Contribuinte {EIRELI,LTDA,EI,CPF,AP,SA}
@attribute UF {PR,BA,MS,GO,SP,RS,MG,SC,RO,RJ,DF,RR,PA,MT,AM,PE,PB}
@attribute QuantSocio numeric
@attribute AnoRepresentacaoFiscal numeric
@attribute AnoEncaminhamento numeric
@attribute Crimes {'CODIGO PENAL','LEI 8137'}
@attribute Duracao/Ano {zero,'acima de quatro',um,dois,tres,quatro}

@data
'ADMINISTRACAO ADUANEIRA',EIRELI,PR,1,2019,2019,'CODIGO PENAL',zero
FISCALIZACAO,LTDA,BA,5,2010,2020,'LEI 8137','acima de quatro'
FISCALIZACAO,LTDA,BA,1,2018,2019,'CODIGO PENAL',um
'ADMINISTRACAO ADUANEIRA',EIRELI,PR,0,2019,2019,'CODIGO PENAL',zero
FISCALIZACAO,LTDA,BA,1,2019,2019,'CODIGO PENAL',zero
FISCALIZACAO,LTDA,BA,1,2019,2019,'LEI 8137',zero
FISCALIZACAO,EIRELI,MS,1,2018,2019,'LEI 8137',um
'ADMINISTRACAO ADUANEIRA',LTDA,GO,0,2019,2020,'CODIGO PENAL',um
FISCALIZACAO,LTDA,MS,3,2018,2019,'CODIGO PENAL',um
'ADMINISTRACAO ADUANEIRA',LTDA,PR,0,2018,2019,'CODIGO PENAL',um
                ...
```

Fonte: Elaborado pelo autor.

A duração do atributo, dividido em 6 (seis) valores, obtidos subtraindo-se o ano do encaminhamento e do ano do início da representação fiscal, estes valores são ilustrados no quadro 3, de forma nominal, foram nomeados referente a quantidade em anos e meses, e suas respectivas nomenclaturas.

Quadro 3. Relação da nomenclatura e número de frequência referente a duração em anos.

Anos	Nomenclatura	N.º de Frequência
Meses	zero	156
Um ano	um	139
Dois anos	dois	53
Três anos	tres	23
Quatro anos	quatro	36
Maiores que quatro anos	acima de quatro	93

Fonte: Elaborado pelo autor.

4.2 Experimentos utilizando árvore de decisão e redes neurais com objetivo nos crimes

Os experimentos utilizando os modelos de árvore de decisão e redes neurais, com objetivo nos crimes, foram feitos por meio do arquivo ARFF, com 500 (quinhentas) instâncias, 288 (duzentos e oitenta e oito) são crimes do Código Penal e 212 (duzentos e doze) são referentes à Lei Nº 8137.

4.2.1 Árvore de decisão

No primeiro experimento, que visa a classificação dos crimes, utilizando o modelo de classificação de árvore de decisão, na opção "*Percentage split",* foi dividido o *dataset* em 66% para realizar o treinamento e 34% para teste.

Figura 11. Resultado com árvore de decisão, relativo na previsão dos crimes, na opção "*Percentage Split"*.

```
=== Evaluation on test split ===

Time taken to test model on test split: 0 seconds

=== Summary ===

Correctly Classified Instances         125               73.5294 %
Incorrectly Classified Instances        45               26.4706 %
Kappa statistic                          0.4699
Mean absolute error                      0.3206
Root mean squared error                  0.4284
Relative absolute error                 64.8831 %
Root relative squared error             84.988  %
Total Number of Instances              170

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0,733    0,263    0,759      0,733    0,746      0,470    0,797     0,800     CODIGO PENAL
                 0,738    0,267    0,711      0,738    0,724      0,470    0,797     0,710     LEI 8137
Weighted Avg.    0,735    0,264    0,736      0,735    0,735      0,470    0,797     0,757

=== Confusion Matrix ===

  a  b   <-- classified as
 66 24 |  a = CODIGO PENAL
 21 59 |  b = LEI 8137
```

Fonte: Elaborado pelo autor.

Os resultados ilustrados na figura 11 mostram que, na matriz de confusão conseguiu relacionar 66 (sessenta e seis) dos registros foram classificados corretamente como "CODIGO PENAL" e 59 (cinquenta e nove) como "LEI 8137", acertando 125 (cento e vinte e cinco) das previsões das possíveis de 170 (cento e setenta). Apenas 21 (vinte e um) registros foram consideradas "CODIGO PENAL", mas na realidade, eram "LEI 8137" e 24 (vinte e quatro) foram classificadas como "LEI 8137", sendo, na verdade "CODIGO PENAL". O resultado mostrou-se satisfatório, por conseguir acurácia de 73,53%. O experimento gerou uma árvore com 50 (cinquenta) nós, em que 42 (quarenta e dois) são nós folhas.

No segundo experimento, foi alterada a opção *"Use training set"*, continuando com o uso do modelo de classificação de árvore de decisão e o alvo a previsão do crime. Foi usado o *dataset* inteiro, tanto para o treinamento e tanto para os testes.

Nessa opção, foi gerado o resultado do experimento, ilustrado na figura 12, mostram acurácia de 80,60%.

Figura 12. Resultado com árvore de decisão, relativo na previsão dos crimes, na opção *“Use training set”*.

```
=== Evaluation on training set ===

Time taken to test model on training data: 0.02 seconds

=== Summary ===

Correctly Classified Instances         403               80.6    %
Incorrectly Classified Instances        97               19.4    %
Kappa statistic                          0.6183
Mean absolute error                      0.2659
Root mean squared error                  0.3646
Relative absolute error                 54.4373 %
Root relative squared error             73.7851 %
Total Number of Instances              500

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0,719    0,075    0,928      0,719    0,810      0,640    0,876     0,898     CODIGO PENAL
                 0,925    0,281    0,708      0,925    0,802      0,640    0,876     0,776     LEI 8137
Weighted Avg.    0,806    0,163    0,835      0,806    0,807      0,640    0,876     0,846

=== Confusion Matrix ===

   a   b   <-- classified as
 207  81 |   a = CODIGO PENAL
  16 196 |   b = LEI 8137
```

Fonte: Elaborado pelo autor.

Na matriz de confusão, verificasse 207 (duzentos e sete) dos registros foram classificados corretamente como “CODIGO PENAL” e 196 (cento e noventa e seis) como “LEI 8137”, a taxa de acerto foram 403 (quatrocentos e três) das previsões das possíveis 500 (quinhentas). Por outro lado, os erros foram de 97 (noventa e sete) classificações, 16 (dezesseis) registros classificados no “CODIGO PENAL” sendo da “LEI 8137”, e 81 (oitenta e um) dos registros classificados na “LEI 8137” sendo do “CODIGO PENAL”.

Os níveis de acurácia se mostraram aceitáveis nos dois testes, se caracterizando como aceitáveis, potencializando a utilização do modelo como promissor.

4.2.2 Redes neurais

Para os experimentos utilizando redes neurais, que visam na classificação dos crimes. No primeiro, utilizando o modelo de classificação de redes neurais, na opção "*Percentage split*", foi dividido o *dataset* em 66% para realizar o treinamento e os demais para teste.

Figura 13. Resultado com redes neurais, relativo na previsão dos crimes, na opção "*Percentage Split"*.

```
=== Evaluation on test split ===

Time taken to test model on test split: 0.01 seconds

=== Summary ===

Correctly Classified Instances         117               68.8235 %
Incorrectly Classified Instances        53               31.1765 %
Kappa statistic                          0.3704
Mean absolute error                      0.3093
Root mean squared error                  0.502
Relative absolute error                 62.5975 %
Root relative squared error             99.6069 %
Total Number of Instances              170

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0,756    0,388    0,687      0,756    0,720      0,373    0,782     0,810     CODIGO PENAL
                 0,613    0,244    0,690      0,613    0,649      0,373    0,782     0,707     LEI 8137
Weighted Avg.    0,688    0,320    0,688      0,688    0,686      0,373    0,782     0,761

=== Confusion Matrix ===

  a  b   <-- classified as
 68 22 |  a = CODIGO PENAL
 31 49 |  b = LEI 8137
```

Fonte: Elaborado pelo autor.

Resultando na figura 13, o modelo alcançou acurácia de 68,82%, conseguindo classificar corretamente 117 (cento e dezessete) instâncias de 170 (cento e setenta) registros totais, e classificando incorretamente 53 (cinquenta e três). Na matriz de confusão, foram relacionados 68 (sessenta e oito) dos registros como "CODIGO PENAL" e 49 (quarenta e nove) como "LEI 8137", ambas corretamente. Classificadas 31 (trinta e um) instâncias consideradas "CODIGO PENAL", mas são "LEI 8137", e 22 (vinte e dois) consideradas "LEI 8137", mas são "CODIGO PENAL".

Figura 14. Resultado com redes neurais, relativo na previsão dos crimes, na opção *"Use training set"*.

```
=== Evaluation on training set ===

Time taken to test model on training data: 0.02 seconds

=== Summary ===

Correctly Classified Instances         452               90.4    %
Incorrectly Classified Instances        48                9.6    %
Kappa statistic                          0.8035
Mean absolute error                      0.1287
Root mean squared error                  0.2497
Relative absolute error                 26.3379 %
Root relative squared error             50.5266 %
Total Number of Instances              500

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0,917    0,113    0,917      0,917    0,917      0,803    0,971     0,975     CODIGO PENAL
                 0,887    0,083    0,887      0,887    0,887      0,803    0,971     0,965     LEI 8137
Weighted Avg.    0,904    0,101    0,904      0,904    0,904      0,803    0,971     0,970

=== Confusion Matrix ===

   a   b   <-- classified as
 264  24 |   a = CODIGO PENAL
  24 188 |   b = LEI 8137
```

Fonte: Elaborado pelo autor.

No segundo experimento, apresentado na figura 14, na opção "*Use training set*", desfrutando do *dataset* de maneira integral, o modelo obteve acurácia de 90,40%. Das 500 (quinhentas) instâncias, 452 (quatrocentos e cinquenta e dois) registros foram classificados corretamente, errando 48 (quarenta e oito) instâncias. Foram classificadas 264 (duzentos e sessenta e quatro) como CODIGO PENAL e 188 (cento e oitenta e oito) classificadas como "LEI 8137". Foram classificadas 24 (vinte e quatro) consideradas "CODIGO PENAL", mas são "LEI 8137", e 24 (vinte e quatro) consideradas "LEI 8137", mas são "CODIGO PENAL".

É possível reconhecer que os experimentos obtiveram resultados com acurácia considerada mediana no "*Percentage Split"* e satisfatória no "*Use training set*", mostrando promissores os dados.

4.3 Experimentos utilizando árvore de decisão e redes neurais com objetivo na duração

Os experimentos utilizam os modelos de árvore de decisão e redes neurais, que visam na classificação da duração do processo administrativo, desde, do ano de representação fiscal, ou seja, o ano que foi autuado o processo até o ano de encaminhamento ao Ministério Público.

4.3.1 Árvore de decisão

No primeiro experimento, relativo na previsão da duração do processo administrativo, utilizando o modelo de classificação de árvore de decisão, na opção "*Percentage split*", fragmentado 66% das instâncias para treinamento e as restantes para testes do *dataset*.

Figura 15. Resultado com árvore de decisão, relativo à duração, na opção "*Percentage Split"*.

```
=== Summary ===

Correctly Classified Instances         141               82.9412 %
Incorrectly Classified Instances        29               17.0588 %
Kappa statistic                          0.7753
Mean absolute error                      0.0799
Root mean squared error                  0.2252
Relative absolute error                 31.1083 %
Root relative squared error             63.3412 %
Total Number of Instances              170

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0,722    0,009    0,975      0,722    0,830      0,783    0,920     0,868     zero
                 0,971    0,030    0,895      0,971    0,932      0,914    0,975     0,900     acima de quatro
                 0,885    0,119    0,767      0,885    0,821      0,739    0,913     0,769     um
                 0,692    0,038    0,600      0,692    0,643      0,613    0,811     0,500     dois
                 0,667    0,006    0,800      0,667    0,727      0,721    0,911     0,628     tres
                 0,900    0,019    0,750      0,900    0,818      0,810    0,946     0,831     quatro
Weighted Avg.    0,829    0,049    0,847      0,829    0,830      0,783    0,922     0,806

=== Confusion Matrix ===

  a  b  c  d  e  f   <-- classified as
 39  1  9  5  0  0 |  a = zero
  0 34  1  0  0  0 |  b = acima de quatro
  1  2 46  0  0  3 |  c = um
  0  0  3  9  1  0 |  d = dois
  0  0  1  1  4  0 |  e = tres
  0  1  0  0  0  9 |  f = quatro
```

Fonte: Elaborado pelo autor.

O experimento gerou uma árvore com 86 (oitenta e seis) nós, em que 52 (cinquenta e dois) são nós folhas. Analisada na figura 15, o experimento alcançou 82,94% de acurácia, classificando corretamente 141 (cento e quarenta e um) e classificando incorretamente 29 (vinte e nove) instâncias, das 170 (cento e setenta) no total. A leitura da matriz de confusão conclui que grupos como “zero”, “um” e “acima de quatro” foram mais acertados em comparação com “dois”, “tres” e “quatro”.

No segundo experimento, na opção “Use training set”, utilizando o *dataset* total, a porcentagem de acurácia aumentou, atingindo 89,0%. Como o aproveitamento do *dataset* foi de todos os registros, ou seja, das 500 (quinhentas) instâncias, foram classificadas corretamente 445 (quatrocentos e quarenta e cinco) e incorretamente 55 (cinquenta e cinco), mostradas na figura 16.

Figura 16. Resultado com árvore de decisão, relativo à duração, na opção de “*Use training set”*.

```
=== Summary ===

Correctly Classified Instances         445               89      %
Incorrectly Classified Instances        55               11      %
Kappa statistic                          0.8568
Mean absolute error                      0.0587
Root mean squared error                  0.1714
Relative absolute error                 22.796  %
Root relative squared error             47.7673 %
Total Number of Instances              500

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0,910    0,055    0,882      0,910    0,896      0,848    0,976     0,943     zero
                 0,978    0,012    0,948      0,978    0,963      0,955    0,996     0,967     acima de quatro
                 0,899    0,053    0,868      0,899    0,883      0,838    0,966     0,900     um
                 0,623    0,009    0,892      0,623    0,733      0,722    0,963     0,780     dois
                 0,826    0,008    0,826      0,826    0,826      0,818    0,995     0,851     tres
                 0,972    0,009    0,897      0,972    0,933      0,929    0,996     0,939     quatro
Weighted Avg.    0,890    0,036    0,890      0,890    0,887      0,856    0,978     0,914

=== Confusion Matrix ===

   a   b   c   d   e   f   <-- classified as
 142   1  11   1   0   1 |   a = zero
   1  91   0   0   1   0 |   b = acima de quatro
   7   3 125   0   3   1 |   c = um
  10   0   8  33   0   2 |   d = dois
   1   0   0   3  19   0 |   e = tres
   0   1   0   0   0  35 |   f = quatro
```

Fonte: Elaborado pelo autor.

A matriz de confusão conclui 142 (cento e quarenta e dois) acertos que realmente são pertencentes ao grupo “zero”, 125 (cento e vinte e cinco) ao grupo “um”, 33 (trinta e três) ao grupo “dois”, 19 (dezenove) ao grupo “tres”, 35 (trinta e cinco) ao grupo “quatro” e 91 (noventa e um) ao grupo “acima de quatro” e os demais 55 (cinquenta e cinco) foram reconhecidos como em outros grupos, sendo que, na verdade não são, identificando como erros.

É possível reconhecer que a acurácia mostrada nos resultados apresentou aceitável, no “Use training set” e “*Percentage split*”, digno com exatidão.

4.3.2 Redes neurais

Para os experimentos utilizando redes neurais, que visam na classificação da duração. No primeiro, utilizando o modelo de classificação de redes neurais, na opção “*Percentage split*”, foi dividido o *dataset* em 66% para realizar o treinamento e os demais para teste.

Figura 17. Resultado com redes neurais, relativo à duração, na opção “*Percentage Split”*

```
Correctly Classified Instances         113               66.4706 %
Incorrectly Classified Instances        57               33.5294 %
Kappa statistic                          0.5586
Mean absolute error                      0.1326
Root mean squared error                  0.3105
Relative absolute error                 51.6331 %
Root relative squared error             87.3422 %
Total Number of Instances              170

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0,704    0,060    0,844      0,704    0,768      0,679    0,892     0,748     zero
                 0,800    0,037    0,848      0,800    0,824      0,780    0,955     0,866     acima de quatro
                 0,692    0,178    0,632      0,692    0,661      0,502    0,792     0,599     um
                 0,462    0,102    0,273      0,462    0,343      0,285    0,761     0,229     dois
                 0,167    0,024    0,200      0,167    0,182      0,155    0,860     0,171     tres
                 0,400    0,025    0,500      0,400    0,444      0,417    0,742     0,321     quatro
Weighted Avg.    0,665    0,091    0,693      0,665    0,674      0,582    0,854     0,641

=== Confusion Matrix ===

  a  b  c  d  e  f   <-- classified as
 38  1 14  1  0  0 |  a = zero
  0 28  0  2  1  4 |  b = acima de quatro
  6  2 36  8  0  0 |  c = um
  1  0  5  6  1  0 |  d = dois
  0  1  1  3  1  0 |  e = tres
  0  1  1  2  2  4 |  f = quatro
```

Fonte: Elaborado pelo autor.

Analisada na figura 17, o experimento alcançou 66,47% de acurácia, classificando corretamente 113 (cento e treze) e classificando incorretamente 57 (cinquenta e sete) instâncias, das 170 (cento e setenta) no total. A leitura da matriz de confusão conclui que grupos como “zero”, “um” e “acima de quatro” foram bem acertados comparados com “dois”, “tres” e “quatro”.

No segundo experimento, na opção “*Use training set*”, o *dataset* aproveitou 100% os registros, o modelo obteve 89,8% de acurácia. A classificação correta dos registros foi 449 (quatrocentos e quarenta e nove) e incorreta 51 (cinquenta e um), ilustrados na figura 18.

Figura 18. Resultado com redes neurais, relativo à duração, na opção “*Use training set”*.

```
Correctly Classified Instances         449               89.8    %
Incorrectly Classified Instances        51               10.2    %
Kappa statistic                          0.8676
Mean absolute error                      0.058
Root mean squared error                  0.1716
Relative absolute error                 22.5301 %
Root relative squared error             47.8189 %
Total Number of Instances              500

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0,942    0,020    0,955      0,942    0,948      0,925    0,986     0,968     zero
                 0,957    0,015    0,937      0,957    0,947      0,935    0,995     0,978     acima de quatro
                 0,899    0,050    0,874      0,899    0,887      0,842    0,957     0,915     um
                 0,868    0,020    0,836      0,868    0,852      0,834    0,964     0,853     dois
                 0,478    0,002    0,917      0,478    0,629      0,652    0,887     0,647     tres
                 0,861    0,022    0,756      0,861    0,805      0,791    0,952     0,771     quatro
Weighted Avg.    0,898    0,027    0,900      0,898    0,896      0,872    0,970     0,914

=== Confusion Matrix ===

   a   b   c   d   e   f   <-- classified as
 147   0   6   2   1   0 |   a = zero
   0  89   1   0   0   3 |   b = acima de quatro
   5   3 125   3   0   3 |   c = um
   1   0   6  46   0   0 |   d = dois
   1   0   5   2  11   4 |   e = tres
   0   3   0   2   0  31 |   f = quatro
```

Fonte: Elaborado pelo autor.

Na leitura da matriz de confusão considerou acertos de 147 (cento e quarenta e sete) registro no grupo “zero”, 125 (cento e vinte e cinco) no grupo “um”, 46

(quarenta e seis) no grupo “dois”, 11 (onze) no grupo “tres”, 31 (trinta e um) no grupo “quatro” e 89 (oitenta e nove) no grupo “acima de quatro”.

É possível reconhecer que os experimentos obtiveram resultados com acurácia considerada razoável no “*Percentage Split”* e satisfatória no “*Use training set*”, mostrando promissores os dados.

4.4 Discussões

É possível constatar que o modelo de forma geral apresenta bons resultados na modalidade com árvore de decisão e regulares com redes neurais. A manipulação do *dataset* contribuiu para a realização de seleção dos atributos compatíveis para a mineração, e assim, colher bons resultados.

Os resultados são apresentados na tabela 1, mostram experiências realizadas e suas respectivas porcentagens de acurácia, em redes neurais e árvore de decisão, na opção “*Percentage Split”* e *“Use training set”,* relativo na previsão dos crimes e da duração do processo administrativo.

Tabela 1. Resultados dos experimentos realizados

Experimentos	Árvore de Decisão (%)	Redes Neurais (%)
Crimes -“*Percentage Split”*	73,53	68,82
Crimes - *“Use training set”*	80,60	90,40
Duração - “*Percentage Split”*	82,94	66,47
Duração - *“Use training set”*	89,00	89,80

Fonte: Elaborado pelo autor.

5 CONSIDERAÇÕES FINAIS

É possível constatar pela proposta inicial de aplicar as técnicas de mineração de dados no âmbito fiscal, no sentido de relacionar as Representações Fiscais para Fins Penais com os crimes praticados e a duração dos processos administrativos. Pode-se observar o perfil dos crimes praticados e o tempo pela acurácia resultante do processo que ambos foram satisfeitos.

As técnicas de classificação e previsão conseguem realizar seu objetivo com uma boa acurácia, ou seja, taxa de acertos acima de 80% (em alguns casos), ou seja, consegue fazer o trabalho de classificação dos tipos de crime que será praticado, a respeito da previsão do crime e da duração do procedimento.

Pode-se observar que nos casos considerados “mais reais” em que a utilização do “*Percentage split*”, os resultados da árvore de decisão foram maiores em comparação com redes neurais. Mesmo os resultados das redes neurais sendo igual ou superior da árvore de decisão, para o treinamento.

Justificando que em redes neurais se ajustou melhor aos dados, mas quando aparece novos dados não conseguiu realizar tão bem a proposta. Por esse motivo, a árvore de decisão é a mais indicada para este caso, apresentando boas classificações e previsões com esses valores, pois tem uma taxa de acertos maior. Baseado nos resultados, mesmo com o número pequeno de instâncias é possível obter resultados neste tipo de classificação.

Com base em toda pesquisa desse trabalho, chegou-se à conclusão de que os resultados podem contribuir nos procedimentos administrativos, podendo contribuir na previsão da duração de um processo, alertando a necessidade sobre seu trâmite, melhorando a administração. Os crimes especiais podem servir como melhor domínio de conhecimento da comissão fiscalizadora.

5.1 Recomendações para trabalhos futuros

Relacionado a essa pesquisa, recomenda-se para trabalhos futuros:

- Realizar testes com *dataset* maiores;
- Utilizar programação para ter controle maior sobre os dados;
- Utilizar novas variáveis para realização de novos métodos;

- Em outros órgãos testar metodologias para o controle de fiscalização.

6 REFERÊNCIAS

AMORIM, Thiago. **Conceitos, técnicas, ferramentas e aplicações de Mineração de Dados para gerar conhecimento a partir de bases de dados**. Tese (Mestrado em Ciência da Computação) - Centro de Informática, Universidade Federal de Pernambuco, Pernambuco, 2006.

BATISTA, Gustavo Enrique de Almeida Prado Alves. **Pré-processamento de Dados em Aprendizado de Máquina Supervisionado**. Tese (Doutorado em Ciências de Computação e Matemática Computacional) - Instituto de Ciências Matemáticas e Computação, Universidade de São Paulo, ICMC-USP, São Paulo, 2003.

BRAGA, Cláudio Vasconcelos. **Rede neural e regressão linear: comparativo entre as técnicas aplicadas a um caso prático na Receita Federal.** Tese (Mestrado em em Administração) - Faculdade de Economia e Finanças IBMEC, Rio de Janeiro, 2010.

BRASIL, Ministério da Fazenda. **Receita Federal - Institucional**. Brasil, SRF, 2021. Disponível em: <https://www.gov.br/receitafederal/pt-br/acesso-a-informacao/institucional>. Acesso em: 23 de novembro de 2021.

BRASIL, Ministério da Fazenda. **Receita Federal - Representações fiscais para fins penais**. Brasil, SRF, 2020. Disponível em: <https://www.gov.br/receitafederal/pt-br/assuntos/orientacao-tributaria/sigilo-fiscal/representacoes-fiscais-para-fins-penais>. Acesso em: 23 de novembro de 2021.

CABENA P. et al. **Discovering data mining: from concept to implementation**. Englewood Cliffs: Prentice Hall, 1998.

CARDOSO, Letícia Marques. **Análise de clientes de uma distribuidora de produtos farmacêuticos com Mineração de dados baseada em Árvore de Decisão**. Monografia (Graduação em Ciência da Computação) - Universidade Federal de Uberlândia, Minas Gerais, 2017.

CARDOSO, Olinda Nogueira Paes; MACHADO, Rosa Teresa Moreira. **Gestão do conhecimento usando data mining: estudo de caso na Universidade Federal de Lavras**. Revista de Administração Pública - RAP, Rio de Janeiro, 2008.

CASTRO, Leandro Nunes de; FERRARI, Daniel Gomes. **Introdução à mineração de dados: conceitos básicos, algoritmos e aplicações**. São Paulo: Saraiva, 2016.

"Células nervosas" em Só Biologia. **Virtuous Tecnologia da Informação**, 2021. Disponível em <https://www.sobiologia.com.br/conteudos/FisiologiaAnimal/nervoso2.php>. Acesso em: 23 de novembro de 2021.

CÔRTES, Sérgio da Costa; PORCARO, Rosa Maria; LIFSCHITZ, Sérgio. **Mineração de Dados - Funcionalidades, Técnicas e Abordagens**. Pontifícia Universidade Católica do Rio de Janeiro, Rio de Janeiro, 2002.

DANTAS, Eric Rommel G.; PATRÍCIO JÚNIOR, José Carlos Almeida; LIMA, Daniel Silva de; AZEVEDO, Ryan Ribeiro de Azevedo. **O Uso da Descoberta de Conhecimento em Base de Dados para Apoiar a Tomada de Decisões**. Centro Universitário de João Pessoa – UNIPÊ. Centro de Informática – Universidade Federal de Pernambuco, Pernambuco, 2008.

DIAS, Carlos Rodrigo; GRAÇA, Andrei de Alencastro; SEMAAN, Gustavo Silva. **Descoberta de Associações em Dados**. Faculdade Metodista Granbery, Minas Gerais.

FAYYAD, Usama M. et al. **Advances in Knowledge Discovery and Data Mining**. 1996. Cambridge: AAAI Press/MIT Press, California, 1996.

FERNANDEZ, Atahualpa. **Introdução ao Direito Tributário.** Editora MP, 2008.

GIL, Antonio Carlos. **Como Elaborar Projetos de Pesquisa**. 6. ed. São Paulo: Atlas S.A., 2017.

HARRISON, Thomas H. **Intranet Data Warehouse.** São Paulo, Editora Berkeley Brasil, 1998.

HAYKIN, Simon. **Redes Neurais - Princípios e Práticas**, Editora Bookman. 1º ed. 2001.

MADEIRA, Renato de Oliveira Caldas. **Aplicação de técnicas de mineração de texto na detecção de discrepâncias em documentos fiscais**. Dissertação (mestrado) – Fundação Getulio Vargas, Escola de Matemática Aplicada. 2015.

MOREIRA, RICARDO DE SOUZA. **A Representação Fiscal para Fins Penais como Instrumento na Promoção da Justiça Fiscal e Social dos Tributos**. Monografia (Graduação em Ciências Contábeis) AFRF - Fiscalização, DRF Novo Hamburgo, Rio Grande do Sul, 2003

NASCIMENTO, Francisco Paulo do; SOUSA, Flávio Luís Leite. **Metodologia da Pesquisa Científica: Teoria e Prática**. Brasília: Thesaurus, 2015.

QUEIROZ, Altamira De Souza. **Algoritmos de inteligência computacional utilizados na detecção de fraudes nas redes de distribuição de energia elétrica**. 2016. 86 f. Tese (Mestrado em Engenharia Elétrica e Computação). Universidade Estadual do Oeste do Paraná, Foz do Iguaçu, 2016.

RAMOS, Célia; LOBO, Fernando. **Descoberta de Conhecimentos em Base de Dados**. 2003, Revista DosALgarves, Universidade do Algarve, Portugal, ESGHT-UALG, nº12, pp. 53-39. 2003.

RABELO, Emerson; CAMPOS, Fernando Celso de. **Big Data e KDD: Novas Descobertas**. Engenharia de Produção, Infraestrutura e Desenvolvimento Sustentável: a Agenda Brasil+10, Universidade Metodista de Piracicaba UNIMEP, Curitiba, Paraná, 2014.

RUSSELL, Stuart; NORVIG, Peter. **Inteligência artificial**. 3. ed. Rio de Janeiro: Elsevier, 2013.

SILVA, Leandro Augusto da; PERES, Sarajane Marques; BOSCARIOLI, Clodis. **Introdução à mineração de dados**: com aplicações em r. Rio de Janeiro: Elsevier, 2016.

SILVA, Matheus Adão de Souza e. **Descoberta de Conhecimento na Análise de Licitações no Estado de Goiás**. Monografia (Graduação em Engenharia de Computação) - Pontifícia Universidade Católica de Goiás, Escola Politécnica, Goiás, 2020.

SILVA, Roque Sérgio D. **Introdução ao direito constitucional tributário**. Editora InterSaberes, 2013.

SOARES JUNIOR, Jair Sampaio; QUINTELL, Rogério Hermida. **Descoberta de conhecimento em bases de dados públicas: uma proposta de estruturação metodológica**. Rio de Janeiro, 2005.

VILARINHO, Renato Avilez; **Uso de Técnicas de Mineração de Dados para Classificação das Ocorrências de Casos de Dengue nos Municípios Brasileiros**. Monografia (Graduação em Sistemas de Informação) - Universidade

Federal de Ouro Preto. Instituto de Ciências Exatas e Aplicadas. Departamento de Computação e Sistemas de Informação, Minas Gerais, 2017.

WAZLAWICK, Raul Sidnei. **Metodologia de Pesquisa para Ciência da Computação**. 2. ed. Rio de Janeiro: Elsevier Editora Ltda, 2014.

WITTEN, I. H. et al. **The WEKA Workbench. Online Appendix for "Data Mining: Practical Machine Learning Tools and Techniques**. [S.l.]: Morgan Kaufmann, 2016.

ANEXO I – Termo de publicação de produção acadêmica

PONTIFÍCIA UNIVERSIDADE CATÓLICA DE GOIÁS
PRÓ-REITORIA DE GRADUAÇÃO

Av. Universitária, 1069 ● Setor Universitário
Caixa Postal 86 ● CEP 74605-010
Goiânia ● Goiás ● Brasil
Fone: (62) 3946.1021 I Fax: (62) 3946.1397
www.pucgoias.edu.br I prograd@pucgoias.edu.br

RESOLUÇÃO 038/2020 - CEPE
ANEXO I

Termo de autorização de publicação de produção acadêmica

O estudante Vinícius de Assunção Furtado do Curso de Ciência da Computação, matrícula 2017.1.0028.0058-0, telefone: 62 996488694, e-mail viniciusfurtados2@gmail.com, na qualidade de titular dos direitos autorais, em consonância com a Lei nº 9.610/98 (Lei dos Direitos do Autor), autoriza a Pontifícia Universidade Católica de Goiás (PUC Goiás) a disponbilizar o Trabalho de Conclusão de Curso intitulado 'Mineração de dados aplicada em processos fiscais' gratuitamente, sem ressarcimento dos direitos autorais, por 5 (cinco) anos, conforme permissões do documento, em meio eletrônico, na rede mundial de computadores, no formato especificado (Texto(PDF), específicos da área para fins de leitura e/ou impressão pela internet, a título de divulgação da produção científica gerada nos cursos de graduação da PUC Goiás.

Goiânia, __15_ de dezembro de 2021

Assinatura do autor: Vinícius de Assunção Furtado

Nome completo do autor: ___Vinicíus de Assunção Furtado_____

Assinatura do professor – orientador: Sibelius Lellis Vieira

Nome completo do professor – orientador: Sibelius Lellis Vieira