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

Classificação e *clustering* aplicados a licitações

CHRISTY BASILIO DA SILVA

GOIÂNIA
2022

Classificação e *clustering* aplicados a licitações

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. André Luiz Alves

Prof. Me. Fernando Gonçalves Abadia

GOIÂNIA

2022

Classificação e *clustering* aplicados a licitações

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 de Computação, em _____/_____/_____.

_____

Orientador: Prof. Dr. Sibelius Lellis Vieira

_____

Prof. Me. André Luiz Alves

_____

Prof. Me. Fernando Gonçalves Abadia

GOIÂNIA

2022

RESUMO

Os processos de licitação são muito importantes para a garantia de que os recursos públicos sejam bem aplicados. Em Goiás, no âmbito estadual, o portal de transparência oferece uma série de opções para a transparência dos dados, o que permite a análise destes dados, bem como ferramentas de controle social para a prestação de contas de forma bem transparência. Neste trabalho emprega-se a mineração de dados, tanto na forma de classificação e do agrupamento, para obter-se tendências ocultas em grandes volumes de dados na área de licitações. É realizado um estudo utilizando o software Weka (*Waikato Environment for Knowledge Analysis)* que utiliza técnicas de árvore de decisão e redes neurais analisando as modalidades da licitação para encontrar tendências que podem melhorar as decisões dos gestores públicos.

Palavras-chave: licitações, conluio, *data science*; *data mining*, *Knowledge Discovery in Databases.*

ABSTRACT

The bidding processes are very important to ensure that public resources are well spent. In Goiás, at the state level, the transparency portal offers a series of options for data transparency, which allows the analysis of this data, as well as social control tools for accountability in a very transparent way. In this work, data mining is used, both in the form of classification and grouping, to obtain hidden trends in large volumes of data in the bidding area. A study is carried out using the Weka software (Waikato Environment for Knowledge Analysis) that uses decision tree techniques and neural networks analyzing the bidding modalities to find trends that can improve the decisions of public managers.

Keywords: licitações, conluio, *data science*; *data mining*, *Knowledge Discovery in Databases.*

LISTA DE ABREVIATURAS

CADE - Conselho Administrativo de Defesa Econômica
CGU - Controladoria-Geral da União
CGE/GO - Controladoria Geral do Estado de Goiás
DCDB - Descoberta de conhecimento em bases de dados
KDD - *Knowledge Discovery in Databases*
SVM - *Support Vector Machine*
SLP - *Single Layer Perceptron*
MLP - *Multilayer Perceptron*
Weka - *Waikato Environment for Knowledge Analysis*

LISTA DE ILUSTRAÇÕES

FIGURAS

TABELAS

SUMÁRIO

1. INTRODUÇÃO

Processos de licitação são muito importantes para a garantia de que os recursos públicos sejam bem aplicados, uma vez que a qualidade desta aplicação, os serviços oferecidos para a sociedade, e a economia que é obtida com a competição entre vários fornecedores de produtos ou serviços pode permitir um melhor preço para a administração pública. Neste sentido, a administração pública tem se preocupado com os aspectos de transparência das contas públicas, oferecendo à população em geral acesso aos dados de gastos, licitações e fornecedores, de modo a tornar menor ações fraudulentas e à corrupção (CARVALHO, 2010).

Em Goiás, no âmbito estadual, o portal de transparência oferece uma série de opções para a transparência dos dados, o que permite a análise destes dados, bem como ferramentas de controle social para a prestação de contas de forma bem transparente. Pela internet se tornou uma forma de garantia de divulgação destas informações, mas sozinha não acarreta que não possa haver problemas nos processos licitatórios.

Particularmente, a grande quantidade de dados disponíveis torna impossível uma análise manual dos processos e tendências. Por outro lado, técnicas de processamento de dados baseadas em estatística e inteligência artificial, muitas vezes denominada aprendizagem de máquina, estão sendo empregadas para uma visualização de possíveis irregularidades ou apenas de tendências que podem melhorar as decisões dos gestores públicos.

Nesse trabalho emprega-se a mineração de dados, tanto na forma da classificação e do agrupamento, para obter-se tendências ocultas em grandes volumes de dados na área de licitações.

1.1 Objetivo

1.1.1 Objetivo geral

Aplicando técnicas de mineração de dados no âmbito de competência do portal de transparência do Estado de Goiás com vistas a correlacionar possíveis tendências nos processos licitatórios, de forma exploratória e descritiva, ocorridos no ano de 2019, de maneira a evidenciar a descoberta de conhecimento.

1.1.2 Objetivos específicos

Para se obter o objetivo geral, é proposto os seguintes objetivos específicos:

- Realizar uma pesquisa exploratória para delimitar e conhecer os fatos e fenômenos relacionado a licitações públicas;
- Explorar os conceitos de *data science* e técnicas de mineração de dados;
- Analisar dados de licitações e fornecedores com vistas a identificar características de modalidades de licitação.

1.2 Estrutura do trabalho

Neste estudo é apresentado com uma estrutura de cinco capítulos, sendo este capítulo referente a introdução, o capítulo 2 trata do referencial teórico, o capítulo 3 descreve os materiais e métodos, o capítulo 4 apresenta os resultados e o capítulo 5 apresenta a conclusão.

2. REFERENCIAL TEÓRICO

2.1 Licitações

O objetivo é mostrar os conceitos sobre licitações, funcionamento e também o propósito do processo licitatório, abordando os tipos, modalidades e valores, a ocorrência de problemas com as características do processo licitatório são muitos comuns.

2.1.1 Processo licitatório

As licitações são um procedimento administrativo no qual o órgão público, em exercício de sua função administrativa, abre para todos que se aceitem as condições fixadas no instrumento convocatório, obtendo uma chance para entregar uma proposta que pode ser selecionada e aceita. (DI PIETRO, 2019).

Conforme regulamentação do art. 37, inciso XXI, da Constituição Federal, que institui normas para licitações e contratos da Administração Pública, o art 3º da Lei nº 8.666 de 21 de junho de 1993 relata o seguinte:

> A licitação destina-se a garantir a observância do princípio constitucional da isonomia, a seleção da proposta mais vantajosa para a administração e a promoção do desenvolvimento nacional sustentável e será processada e julgada em estrita conformidade com os princípios básicos da legalidade, da impessoalidade, da moralidade, da igualdade, da publicidade, da probidade administrativa, da vinculação ao instrumento convocatório, do julgamento objetivo e dos que lhes são correlatos (BRASIL, 1993).

Com o objetivo de reduzir custos na aquisição de bens e minimizar possíveis desperdícios de dinheiro públicos, é fundamental a utilização do instrumento de licitação por parte do órgão público.

A licitação não deve ser tratada sigilosamente, pois uma licitação é um ato público. Deve haver transparência para se ter acesso ao procedimento referente à licitação. Em um processo licitatório não deve ter qualquer favorecimento e que

garanta que o recurso público será usado com grande cautela e eficiência. (SOUZA, 1997).

É preciso ter em mente que países em que a corrupção é muito alta recebem menos investimentos de outros países e do setor privado, reforçando diversos problemas sociais. (FORTINI; MOTTA, 2016).

2.2 Descoberta de conhecimento em bases de dados (DCDB)

A descoberta de conhecimento em bases de dados também conhecida como *Knowledge Discovery in Databases (KDD),* em inglês, é o processo de identificar nos dados padrões válidos que são desconhecidos e potencialmente úteis, assim possibilitando o melhor entendimento sobre o problema ou procedimento para uma melhor tomada de decisão. Com a grande quantidade de dados gerados e armazenados atualmente, resolver problemas com o KDD vem se tornado em uma potente ferramenta para extrair desses dados um conhecimento útil para as organizações que os possuem.

2.2.1 Fases do DCDB

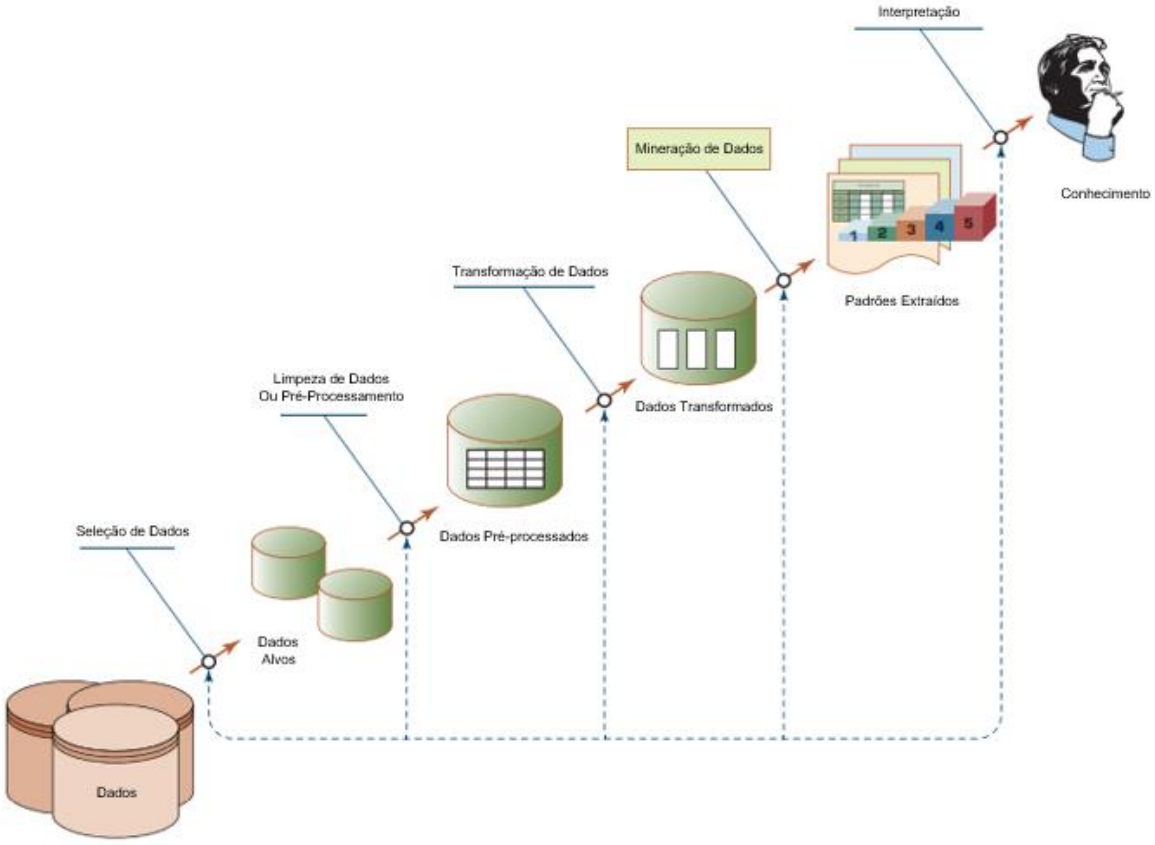
O DCDB é formado por cinco fases que extraem o conhecimento e informações de uma base de dados, efetuando relação com as suas características e essas cinco fases que compõem três grupos principais, sendo que o primeiro grupo trata sobre o pré-processamento atuando na seleção dos dados, limpeza dos dados e transformação dos dados, no segundo grupo acontece o processamento e no terceiro grupo o pós-processamento que é a fase da interpretação dos dados, e analisa o resultado obtido e a qualidade do conhecimento descoberto. A figura 1 ilustra as fases o processo de DCDB.

A primeira fase deve selecionar adequadamente quais os dados são importantes para a mineração de dados. Os dados coletados podem vir de diversas formas variadas que não contribuam para a finalidade, como valores incorretos ou desconhecidos, de baixo valor preditivo, entre outros. Por isso, nesta fase de limpeza consiste em obter de forma consistente na geração dos resultados, com intuito de aprimorar a qualidade dos dados se não aprimorado pode apresentar problemas futuros.

A segunda fase, consiste na escolha e uso de técnicas para reduzir ruídos, que podem ser desde: inspeção humana, interpolação, agrupamento ou regressão. As estratégias de regressão e inferência ajudam a corrigir campos sem dados ou nulos, garantido a qualidade de dados.

A terceira fase transforma os dados já existentes em novos dados apropriados para o processo de mineração de dados conforme a técnica a ser utilizada, além da modificação e formatação dos dados que devem ficar adequados para a mineração de dados. A integração minimiza redundâncias de atributos, identifica e resolve valores conflitantes em relação a diferenciação na escala ou codificação.

Figura 1. Processo do KDD



Fonte: Adaptado de Sharda, Delen e Turban (2018)

A quarta fase realiza a mineração de dados, na qual são definidas e aplicadas as técnicas e algoritmos de mineração de dados que verificam a hipótese e extraem padrões de forma autônoma a partir dos dados definidos na terceira fase, nesta fase podendo utilizar de diversas ferramentas, técnicas ou algoritmos.

A última fase do processo DCDB realiza a interpretação dos dados, e analisa o resultado obtido e a qualidade do novo conhecimento, uma avaliação pode ser

realizada medindo as estatísticas geradas pelos resultados também é possível retornar a qualquer uma das etapas anteriores, caso necessário.

2.2.2 Mineração de dados

Na mineração de dados utiliza-se de algoritmos computacionais em bases de dados, como alvo descobrir padrões úteis para tomada de decisão, o conhecimento é extraído nesse processo para ser utilizado em tomadas de decisão, permitindo dar o valor ao objetivo da decisão.

Na mineração de dados inclui muitas áreas correlatas em seu processo, as destacadas são as tecnologias de banco de dados, estatísticas e inteligência artificial, além de áreas mais específicas.

Alguns exemplos de técnicas para estimativa são: regressão linear; regressão múltipla; regressão não linear; regressão logística; regressão de Poisson. (CÔRTES; PORCARO; LIFSCHITZ, 2002).

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 DCDB, de forma automática ou semiautomática. As tarefas preditivas procuram, através de variáveis conhecidas, encontrar valores ainda desconhecidos ou futuros. Já as tarefas descritivas procuram padrões para descrever dados. Nas tarefas preditivas são compostas pela classificação e regressão, se enquadrando como as principais. Já na descritiva são compostas por regras de associação, agrupamento, sumarização e detecção de desvios, são as mais relevantes (CÔRTES et al., 2002).

A tarefa de classificação, compreende analisar um novo registro e determinar uma classe com base em um histórico de registros já classificados. A tarefa é realizada em duas etapas, a primeira etapa é realizada o treinamento e o modelo é gerado através de um conjunto de dados já classificados, na segunda etapa um novo conjunto de dados que não foram usados na etapa anterior, é utilizado para realizar os testes, assim é possível saber se a capacidade do modelo de responder corretamente aos dados é eficiente.

A tarefa de agrupamento, como o próprio nome sugere, visa agrupar objetos em clusters, de acordo com relações existentes entre eles. São realizadas buscas por similaridades e diferenças entre os objetos analisados, e a distância de similaridade e diferença entre os objetos é usada para determinar em qual grupo se encaixa, ou seja, objetos similares têm distância de similaridade menor, então são agrupados em um

mesmo cluster. Apesar de parecer com a classificação, no agrupamento não existem classes previamente determinadas, pois simplesmente, os objetos são agrupados de acordo suas semelhanças (SILVA; PERES; BOSCARIOLI, 2016).

Na tarefa de regressão é seguido o modelo de aprendizagem supervisionada, separando dados para o treinamento e testes. Mas, a maneira que seus resultados são avaliados é diferente da classificação, pois a regressão, por ser uma tarefa que visa predizer um valor contínuo, não observa a quantidade de acertos, e sim o cálculo da distância entre a saída esperada e a saída estimada, tendo como resultado a precisão da predição.

Na associação, o objetivo é encontrar relações entre atributos que ocorrem em base de dados transacionais. Se a frequência de ocorrência desses atributos em transações for muita alta significa que existe uma relação forte entre esses atributos.

Tabela 1 - Relação entre técnicas e tarefas de mineração de dados.

	Tarefas	**Técnicas**	**Algoritmos**
Análise preditiva	Classificação	Árvore de decisão; Análise bayesiana; Análise de vizinhança; Redes Neurais	J-48, algoritmo C4.5, classificadores bayesianos, KNN, SVM (Support Vector Machine)
	Regressão	Regressão linear; regressão múltipla; regressão não linear; regressão logística; regressão de poisson; redes neurais.	Backpropagation; Multilayer Percetpron;
Análise descritiva	Agrupamento	Método de particionamento; modelagem de regras.	Apriori; Algoritmo FP-Growth; Eclat. ;Direct Hashing And Pruning; Dynamic Itemset Counting
	Associação	Mineração de regras de associação;	Apriori; Algoritmo FP-Growth

Fonte: Adaptado de Sharda, Delen e Turban (2018)

2.2.4 Associação e *clustering*

A associação consiste na descoberta de relações na correlação de determinados atributos. Tal tarefa detecta padrões que associam valores de vários atributos de um banco de dados. Uma das técnicas que se utiliza da associação é a: *Market Basket Analysis*, que foi criada para a descoberta de combinações entre itens que ocorrem acima da média em um banco de dados. Por exemplo como resultado 70% dos clientes que compram estrogonofe de frango tendem a comprar lasanha é uma espécie de correlação entre atributos. (DEVMEDIA, 2020).

A associação tem uma abordagem fundamentada com o algoritmo *Apriori*, que determina que se qualquer padrão de comprimento N não é frequente na base de dados, não será frequente o seu comprimento N+1, ou seja, ao adicionar mais elementos no item não vai torná-lo mais frequente. Através de um processo iterativo, deve-se gerar um conjunto de padrões de comprimento N+1, a partir do conjunto de padrões de frequência de comprimento N (para $N \geq 1$), e observar as frequências no banco de dados. Entretanto, se há muitos padrões ou estes são longos se torna muito dispendioso assim como percorrer repetidas vezes o banco de dados. (NANDI et al., 2015).

Para diminuir suas limitações e aumentar a eficiência da descoberta de regras de associação diversos métodos foram propostos, tais como por exemplo: FP-Growth que codifica um conjunto de dados em uma estrutura de dados chamada de *Frequent Pattern tree*. (NANDI et al., 2015).

Para o FP-Growth funcionar é preciso informar dois ou mais parâmetros de entrada, um sendo o suporte e o outro a confiança. A entrada de suporte é a métrica utilizada para encontrar todos os N itens suportando a regra de associação X, $A \Rightarrow B$, que corresponde a frequência em que A e B ocorrem no banco de dados. A confiança corresponde a frequência em que B ocorre, dentre os eventos que contêm A (TAN; STEINBACH; KUMAR, 2009).

Após definir os parâmetros de entrada e uma base de dados, o algoritmo montar uma estrutura FP-tree que armazena os itens frequentes, é necessário ler a base de dados duas vezes para que a FP-tree seja criada. Na primeira leitura é identificado o conjunto F, de tamanho um, e seus respectivos suportes.

A Clusterização consiste na divisão de um grupo heterogêneo em vários subgrupos mais homogêneos. Neste processo não existem classes pré-definidas e os dados são agrupados de acordo com suas características próprias. A técnica de agrupamento em métodos de particionamento é uma das técnicas possíveis para a realização de Clusterização, que tem como objetivo encontrar a melhor partição dos n objetos em K grupos. Normalmente os K grupos encontrados possuem mais qualidade comparados com K grupos produzidos pelos métodos hierárquicos. Um dos algoritmos mais famosos para essa técnica é o algoritmo denominado como *K-means* (RODRIGUES; 2009).

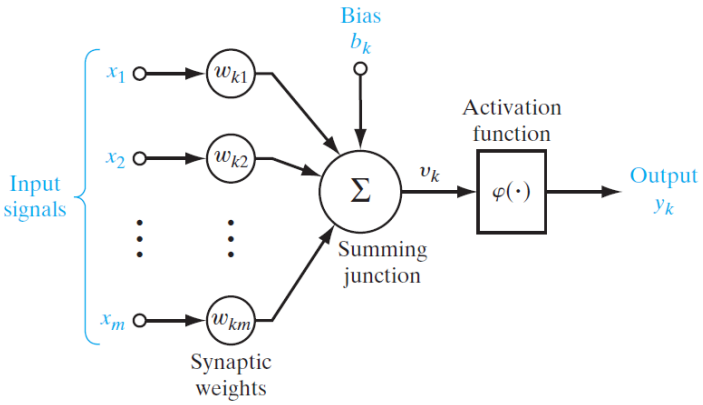
O *K-means* é um algoritmo de aprendizagem não supervisionado e de Clusterização, utilizado para particionar dados em k agrupamentos. Ele agrupa dados que compartilham características importantes e parecidas. De modo empírico, uma boa solução para o processo de *clustering* é aquela em que os dados do grupo sejam mais semelhantes entre si, do que comparados com outro grupo. O *K-means* é caracterizado como um algoritmo guloso, que em suas iterações escolhe o objeto que lhe parece mais promissor, que chama mais sua atenção, torna o objeto parte da solução do problema, mas não analisa as consequências de suas escolhas (SOUSA; 2019).

O algoritmo *K-means* pode ser descrito por quatro etapas: atribui-se valores iniciais para os protótipos seguindo algum critério, por exemplo, sorteio aleatório desses valores dentro dos limites de domínio de cada atributo; atribui-se cada objeto ao grupo cujo protótipo possui maior similaridade com o objeto; recalcula-se o valor do centróide (protótipo) de cada grupo, como sendo a média dos objetos atuais do grupo; repete-se os passos 2 e 3 até que os grupos se estabilizam. (FONSECA; BELTRAME, 2010).

2.2.5 Redes neurais

É comum de representar graficamente uma rede neural artificial utilizando grafos, em que uma ligação sináptica é representada por uma aresta e um neurônio através de um nó. Um neurônio possui diversas entradas, e para cada entrada x de um neurônio, é aplicado um peso sináptico w, todos os valores de as entradas são somados por uma função de soma $\sum$, e uma função de ativação φ é aplicada no resultado, gerando uma saída, como mostrado na figura 2. A saída é destinada para a entrada de um outro neurônio ou ser o valor final da rede neural. (GIACOMEL, 2016).

Figura 2: Neurônio de uma rede Neural



Fonte; Haykin (2009)

A fórmula matemática da figura 2 é descrita pelas equações $U_k = \sum_{j=1}^{m} W_{kj} X_k$ e $Y_k = \varphi(U_k + B_k)$, no qual, em um neurônio k, U_k representa saída do combinador linear, W_{kj} representa o peso sináptico da entrada j, x_j é a entrada j e por fim saída final é representada pelo Y_k.

Segundo Silva et al (2019), existem diversas estruturas de redes neurais artificiais, mas duas delas, a Single *Layer Percetron* e a *Multilayer Perceptron*, são consideradas as principais estruturas. A *Single Layer Perceptron* (SLP) é a mais simples, e consiste em neurônios paralelamente organizados em uma única camada, possuindo apenas uma saída, mas podem receber n entradas. Dessa forma a SLP consegue resolver problemas de classificação, podendo, por exemplo, ter valores binários, 0 ou 1, como saída.

Já a *Multilayer Perceptron* (MLP), ou rede neural multicamadas, é um tipo de rede neural mais complexa, que pode possuir mais de uma camada de neurônio, as

chamadas camadas ocultas, e os neurônios que estão na camada oculta tem como valor de entrada os resultados da camada anterior através das sinapses. As camadas ocultas conseguem aumentar o poder de processamento da rede neural, e tem objetivo de extrair resultados mais expressivos (HAYKIN, 2001).

2.2.6 Árvore de decisão

Segundo Russell e Norvig (2013), uma árvore de decisão recebe como entrada um conjunto de atributos de um objeto não classificado, que podem ser contínuos ou discretos, e devolve a classificação do objeto. Uma árvore de decisão é composta por nós internos, nós folhas, ramificações e um nó raiz. O nó raiz é o nó mais alto da árvore, cada nó interno é responsável por um teste realizado em um dos atributos do objeto, os ramos de um nó são as possíveis respostas para os testes feitos no nó. A classificação final de uma árvore de decisão é responsabilidade da nós folha.

Nas obras de Castro e Ferrari (2016), construir uma árvore de decisão com objetivo de classificar um objeto sem classe definida com base nos valores do objeto, é chamado de indução de árvores de decisão. O processo de indução de uma árvore de decisão se dá de maneira recursiva. Para saber qual o atributo ideal para a divisão é necessário medir a pureza dos nós, sendo o quão homogêneo um nó é em relação às classes do objeto. Ao se medir a pureza de todos nós é definido como será a expansão de nós, pois os nós com filhos mais puros são escolhidos para a expansão. A entropia é a medida que define a pureza e calcula a variabilidade das classes que pertencem ao conjunto de atributos da base de dados. (CASTRO; FERRARI ,2016).

2.3 Estudos correlatos

Há diversos trabalhos relacionados sobre mineração de dados com licitações públicas, sendo que desses trabalhos os de Silva e Ralha são os que mais se assemelham a este trabalho, no qual realizaram um estudo em que o objetivo é a detecção de cartéis de licitações públicas utilizando mineração de dados, no caso foi utilizando regras de associação e *clusterização*.

> Diante do exposto, tem-se a necessidade de pesquisas científicas que forneçam bases teóricas e metodológicas com a finalidade de identificar anomalias para possíveis ações irregulares no mercado de licitações. Entretanto, o crescimento do volume de processos licitatórios específicos em intervalos de tempo mais curtos impõe vários desafios no âmbito dos órgãos de controle, não sendo possível realizar esse monitoramento de forma manual, necessitando então de ferramentas computacionais. Encontram-se disponíveis técnicas computacionais que podem auxiliar na produção de conhecimento a partir de grandes volumes de dados, como são as bases de dados de órgãos públicos. A título de exemplo, técnicas baseadas em inteligência artificial, aprendizado de máquina e ciência de dados, são amplamente utilizadas para esse fim por diversas empresas, com o objetivo de identificar padrões ou informações relevantes para os negócios. (SILVA,2020)

A CGU, como Órgão Central de Controle Interno, mantém parceria com o CADE para que as investigações de prática de cartéis no âmbito da Administração Pública sejam mais eficientes. Dessa forma, sempre que a CGU encontra indícios de práticas de rodízio de licitações em suas auditorias, o processo pode ser encaminhado ao CADE para que este tome as providências cabíveis. Independente da decisão do CADE, a CGU também pode punir as empresas suspeitas através da Declaração de Inidoneidade, que as impede imediatamente de participar de licitações e contratar com a APF. (RALHA,2010)

Em seus trabalhos são indicados para investigação de licitações por meio de técnicas de reconhecimento de padrões estatísticos e mineração de dados. O autor no trabalho realizou um estudo bibliográfico e identificou as principais variáveis necessárias para a formação de indicadores de participação de cartéis nas licitações sendo elas: objeto da licitação, órgão licitante, participantes e propostas de preços. Foi desenvolvida uma metodologia baseada nesses indicadores que utilizou mineração de dados, regra de associação e classificação não supervisionada, e com

o algoritmo *k-means*, aplicado nos dados disponibilizados no portal da transparência do Tribunal de Contas dos Municípios do Estado do Ceará (TCM/CE).

Os resultados obtidos mostraram que a utilização da mineração de dados e utilização de técnicas de reconhecimento de padrões estatísticos, foi possível a obtenção de categorias de empresas que indicaram uma maior probabilidade de atuarem em licitações fraudulentas.

3. MATERIAIS E MÉTODOS

Este tópico tem como objetivo descrever os materiais e métodos utilizados na realização deste trabalho.

3.1 Materiais

Os dados foram obtidos no site do governo de transparência no qual foi escolhido os dados do ano de 2019, disponível no site http://www.transparencia.go.gov.br/portaldatransparencia/institucional/dados-abertos pertencente de transparência de Goias.

Os dados são separados em 2 (dois) tipos em relação a cada mês do ano sendo os tipos: fornecedores e licitações. As licitações têm nove colunas sendo elas: modalidade licitação, data solicitação aquisição, código solicitação aquisição, número processo, valor adjudicado, sigla órgão, número edital, código órgão e anomes. E os fornecedores têm dez colunas sendo elas: cpf/cnpj fornecedor, valor adjudicado unitário, código solicitação aquisição, sigla órgão, razão social fornecedor, objeto licitação, quantidade itens, anomes, código órgão e valor adjudicado total.

O software Weka foi o sosftware utilizado para a mineração de dados e fornece implementações de algoritmos de aprendizado que pode ser aplicar facilmente ao conjunto de dados. Ele também inclui uma variedade de ferramentas para transformar conjuntos de dados, como os algoritmos para discretização e amostragem, podendo processar um conjunto de dados, alimentá-lo em um esquema de aprendizado, e analisar o classificador resultante e seu desempenho, tudo sem escrever nenhum código de programa.

O Weka inclui métodos para os principais problemas de mineração de dados: regressão, classificação, agrupamento, mineração de regras de associação e seleção de atributos. Todos os algoritmos recebem suas entradas na forma de uma única tabela relacional que pode ser lida de um arquivo ou gerado por uma consulta de banco de dados. Foi utilizado um notebook da marca Dell com as seguintes configurações: processador Intel® Core™ i7 de 10ª geração com 8GB de RAM e 250gb SSD.

3.2 Métodos

A primeira etapa foi feita na busca de conhecimento sobre trabalhos, livros e teses semelhantes na área de ciência de dados, mineração de dados, KDD, para poder escolher a melhor forma do que seria aplicado no trabalho para encontrar uma solução para o problema.

A segunda etapa foi a seleção dos dados, utilizando os dados sobre licitações, documento disponibilizado publicamente pelo governo no site sobre transparências, dados do período do ano de 2019.

Na terceira etapa, o pré-processamento de dados foi realizada a limpeza dos dados que não poderiam ser utilizados para incluir somente os dados mais relevantes, e que foram organizados e estruturados, bem como o agrupamento de dados.

Para realizar a mineração de dados é criada uma nova planilha agrupada com todos os meses que contêm colunas específicas dos fornecedores como da licitação, as colunas escolhidas são: modalidade licitação, data solicitação aquisição, código solicitação aquisição, número processo, valor adjudicado, código órgão e cpf/cnpj fornecedor. Com a nova planilha é possível fazer a mineração de dados usando o Weka.

Na quarta e última etapa foi realizada análise preditiva do processamento de dados, aplicando técnicas de mineração para a identificação e classificação de uma situação, previsão do crime ou calote no procedimento administrativo.

4. RESULTADOS

Para a obtenção dos resultados foi usado o algoritmo J48, algoritmos de árvore de decisão, e também algoritmo *multilayer perceptron* para redes neurais.

Através do *software* WEKA, os testes foram realizados utilizando dois métodos o *Percentage split* e *Use training set*. o *Percentage split* divide o dataset em dois, a primeira parte é separada em dados para treinamento e o outro para testes. No segundo, o treinamento e o teste fazem parte do *dataset* sem a separação.

4.1 Seleção e pré-processamento de dados

Os dados foram adquiridos no site da transparência do governo de goiás disponível pelo link a seguir http://www.transparencia.go.gov.br/portaldatransparencia/institucional/dados-abertos , que contém arquivos mensais em pastas anuais de vários indicadores da administração pública e foram coletados dados sobre licitações e fornecedores do ano de 2019 referente aos meses janeiro até março.

O arquivo de licitações contém os seguintes dados: modalidadelicitacao, datasolicitacaoaquisicao, codsolicitacaoaquisicao, numprocesso, valoradjucado, siglaorgao, numeoedital, codorgao e anomes. Já o arquivo de fornecedores contém os seguintes dados: cpfcnpjfornecedor, valoradjucadounitario, codsolicitacao, siglaorgao, razaosocial, objetolicitacao, qtditens, anomes, codorgao e valoradjucadototal.

Tabela 2: Informações do Portal Transparência de Goiás

TABELA LICITAÇÃO			
NOME	**CAMPO**	**TIPO**	**DESCRIÇÃO**
Ano/ Mês	ANOMES	Número	Ano e mês em que a solicitação de aquisição foi cadastrada no Sistema de Compras (Compranet)
Modalidade de Licitação	MODALIDADELICITACAO	Alfanumérico	Indica o procedimento que irá reger a licitação. Ex: concorrência, tomada de preços, convite, pregão
Código da Solicitação de Aquisição	CODSOLICITACAOAQUISICAO	Numérico	Número da solicitação de aquisição cadastrada no Sistema de Compras (Comprasnet)
Número do Edital	NUMEDITAL	Alfanumérico	Número que identifica o edital de licitação
Número do Processo	NUMPROCESSO	Numérico	Número do processo em que foi protocolada a solicitação de aquisição
Valor Submetido	VALORSUBMETIDO	Numérico	Valor para realizar a aquisição pretendida
Data da Solicitação	DATASOLICITACAO	Numérico	Data em que a solicitação de aquisição foi cadastrada no Sistema de Compras (Compranet)
Tipo de Disputa	TIPODISPUTA	Alfanumérico	Critério utilizado pela Administração na escolha da melhor proposta
Nome do Órgão	NOMEORGAO	Alfanumérico	Órgão/ Entidade responsável pela realização da licitação
Razão Social	RAZAOSOCIAL	Alfanumérico	Nome de registro da empresa licitante
Código do Órgão	CODORGAO	Numérico	Código do órgão que realizou a licitação
Cadastro Nacional de Pessoa Jurídica	CNPJ	Numérico	CNPJ do licitante

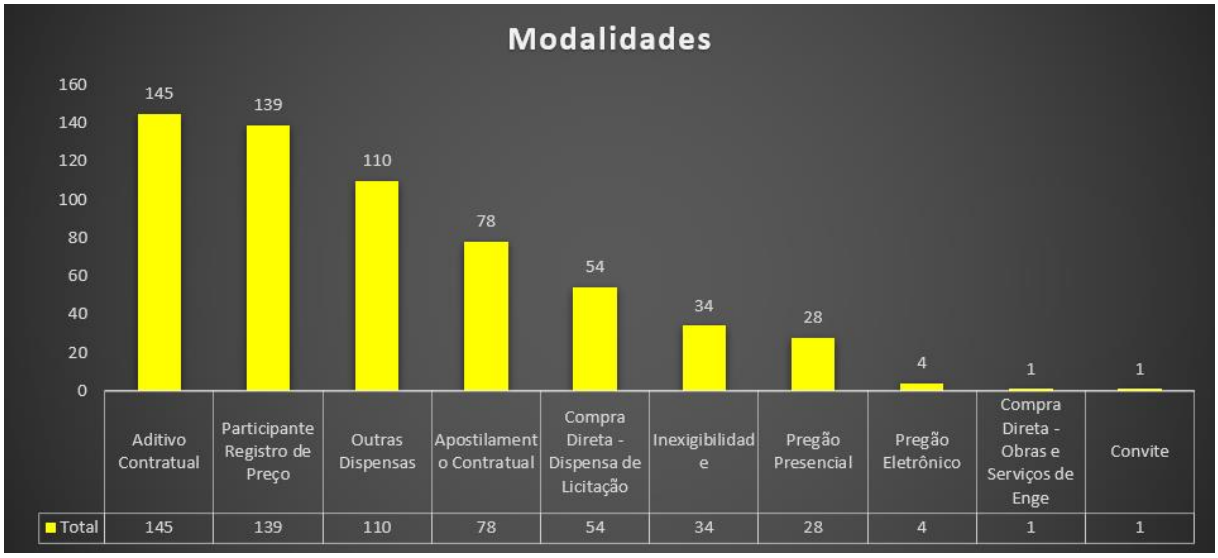
Fonte: elaborado pelo autor

Foram agrupados dados específicos de três meses das licitações e fornecedores em um único arquivo que contém os seguintes dados: modalidadelicitacao, datasolicitacaoaquisicao, codsolicitacaoaquisicao, numprocesso, valoradjucado, codorgao e cnpj. Este novo arquivo é o que será usado para realizar os experimentos renomeado como jan.fev.mar.csv.

4.2 Análise descritiva

Neste novo arquivo foi feito um agrupamento que tem como alvo as modalidades das licitações mostrando cada modalidade como também sua quantidade de cada uma como mostrado na figura 3.

Figura 3: Agrupamento de modalidades de licitação

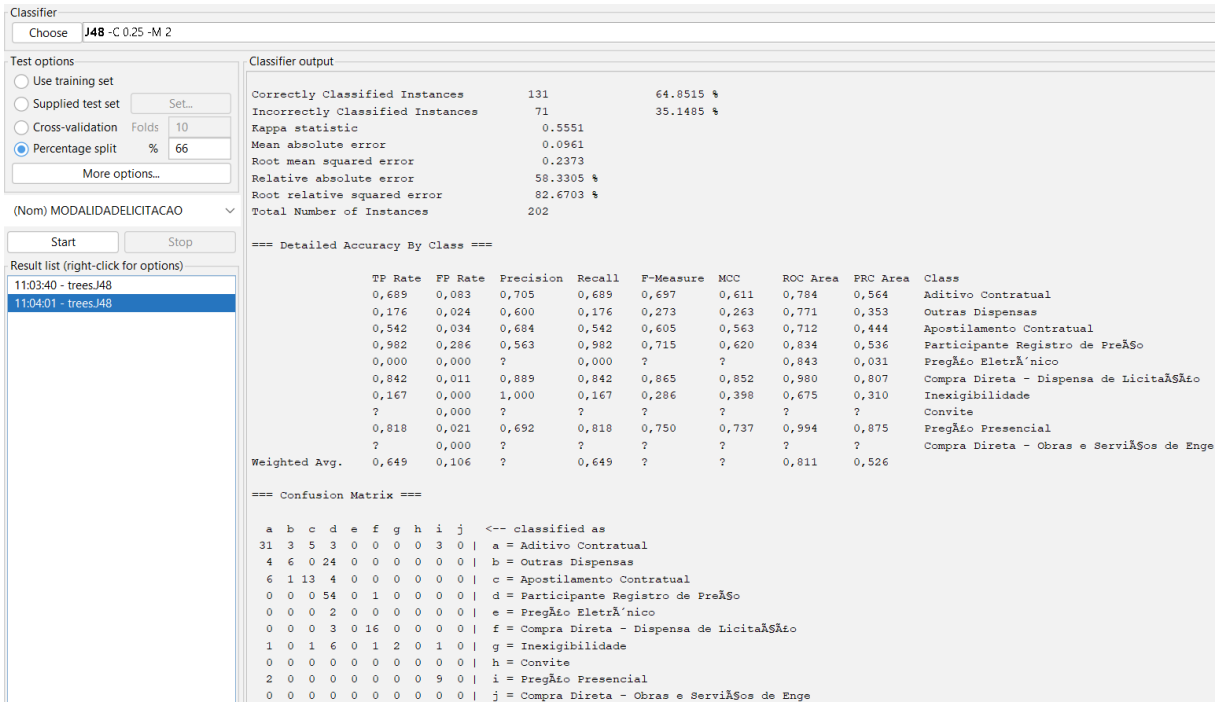


Fonte: elaborado pelo autor

4.3 Árvore de decisão

O primeiro experimento foi utilizando a árvore de decisão, usando o modelo de classificação J48, a opção de teste escolhido foi o *Percentage split* que tem como alvo modalidadelicitacao, o datase foi dividido em 66% para fazer o treinamento e os demais para teste.

Figura 4: Árvore de decisão – *percentagem split*



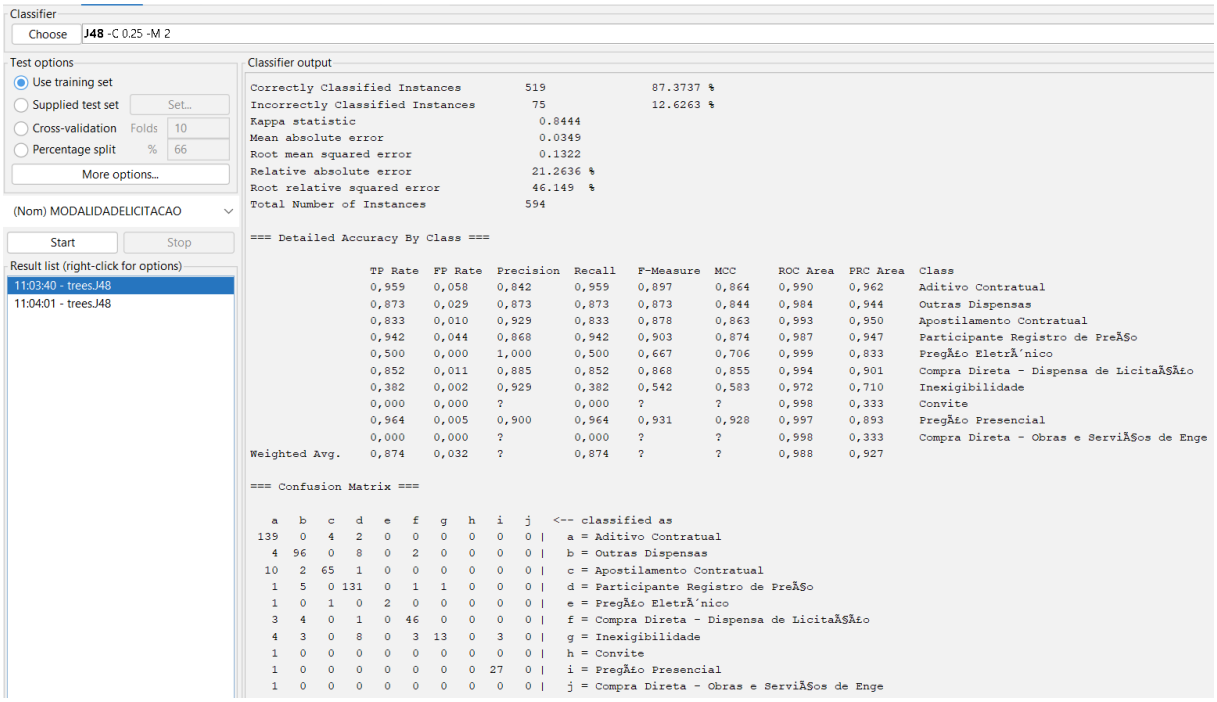
Fonte: elaborado pelo autor

Analisando a figura 4, o experimento alcançou 64,85% de acurácia da instância de classificação sendo acertado corretamente 131 itens e a instância de classificação sendo acertado incorretamente 71.

Observando a matriz de confusão do *Percentage split* notou-se que o grupo “a” houve 31 acertos, 6 no grupo “b”, 13 no grupo “c”, 54 no grupo “d”, 0 no grupo “e”, 16 no grupo “f”, 2 no grupo “g”, 0 no grupo “h”, 9 no grupo “i” e 0 no grupo “j”.

No segundo experimento a opção de teste usada é a *Use training set*, o *dataset* utilizou todos os registros, obtendo 87.37% de acurácia e 12.6% de erro, como visto na figura 5.

Figura 5: Árvore de decisão – *training set*



Fonte: elaborado pelo autor

Na matriz de confusão da figura 5 nota-se que o grupo “a” houve 139 acertos, 96 no grupo “b”, 65 no grupo “c”, 131 no grupo “d”, 2 no grupo “e”, 46 no grupo “f” , 13 no grupo “g” , 0 no grupo “h” , 27 no grupo “i” e 0 no grupo “j”.

Figura 6: Trecho da árvore de decisão

```
|   |   DATASOLICITACAOAQUISICAO = 2019-03-21T00:00:00.000-03:00
|   |   |   CODORGAO <= 20: Aditivo Contratual (3.0/1.0)
|   |   |   CODORGAO > 20
|   |   |   |   CODSOLICITACAOAQUISICAO <= 70763: Participante Registro de PreÃ§o (3.0/1.0)
|   |   |   |   CODSOLICITACAOAQUISICAO > 70763
|   |   |   |   |   NUMPROCESSO <= 201900010007014: Outras Dispensas (5.0)
|   |   |   |   |   NUMPROCESSO > 201900010007014: Participante Registro de PreÃ§o (2.0)
|   |   DATASOLICITACAOAQUISICAO = 2019-03-22T00:00:00.000-03:00
|   |   |   CODSOLICITACAOAQUISICAO <= 70801
|   |   |   |   NUMPROCESSO <= 201900010007156: Outras Dispensas (4.0)
|   |   |   |   NUMPROCESSO > 201900010007156: Participante Registro de PreÃ§o (2.0)
|   |   |   CODSOLICITACAOAQUISICAO > 70801: Participante Registro de PreÃ§o (4.0/1.0)
|   |   DATASOLICITACAOAQUISICAO = 2019-03-25T00:00:00.000-03:00
|   |   |   CODORGAO <= 123
|   |   |   |   VALORADJUDICADO <= 191718.36: Outras Dispensas (5.0)
|   |   |   |   VALORADJUDICADO > 191718.36: Aditivo Contratual (2.0)
|   |   |   CODORGAO > 123
|   |   |   |   NUMPROCESSO <= 201800057001527: Aditivo Contratual (2.0/1.0)
|   |   |   |   NUMPROCESSO > 201800057001527: Participante Registro de PreÃ§o (2.0)
|   |   DATASOLICITACAOAQUISICAO = 2019-03-26T00:00:00.000-03:00
|   |   |   CODSOLICITACAOAQUISICAO <= 70864: Outras Dispensas (3.0/2.0)
|   |   |   CODSOLICITACAOAQUISICAO > 70864: Participante Registro de PreÃ§o (5.0/1.0)
|   |   DATASOLICITACAOAQUISICAO = 2019-03-27T00:00:00.000-03:00
|   |   |   NUMPROCESSO <= 201800029001534: Outras Dispensas (2.0)
|   |   |   NUMPROCESSO > 201800029001534: Aditivo Contratual (3.0/1.0)
|   |   DATASOLICITACAOAQUISICAO = 2019-03-28T00:00:00.000-03:00: Participante Registro de PreÃ§o (3.0/1.0)
|   |   DATASOLICITACAOAQUISICAO = 2019-03-29T00:00:00.000-03:00
|   |   |   NUMPROCESSO <= 201800015001811: Outras Dispensas (2.0)
|   |   |   NUMPROCESSO > 201800015001811: Participante Registro de PreÃ§o (13.0/2.0)
|   NUMPROCESSO > 201900011002110
|   |   VALORADJUDICADO <= 35557.87: Compra Direta - Dispensa de LicitaÃ§Ã£o (44.0/4.0)
|   |   VALORADJUDICADO > 35557.87
|   |   |   CODORGAO <= 119: Aditivo Contratual (4.0/1.0)
|   |   |   CODORGAO > 119: Outras Dispensas (4.0/2.0)

Number of Leaves  :     163

Size of the tree :      231
```

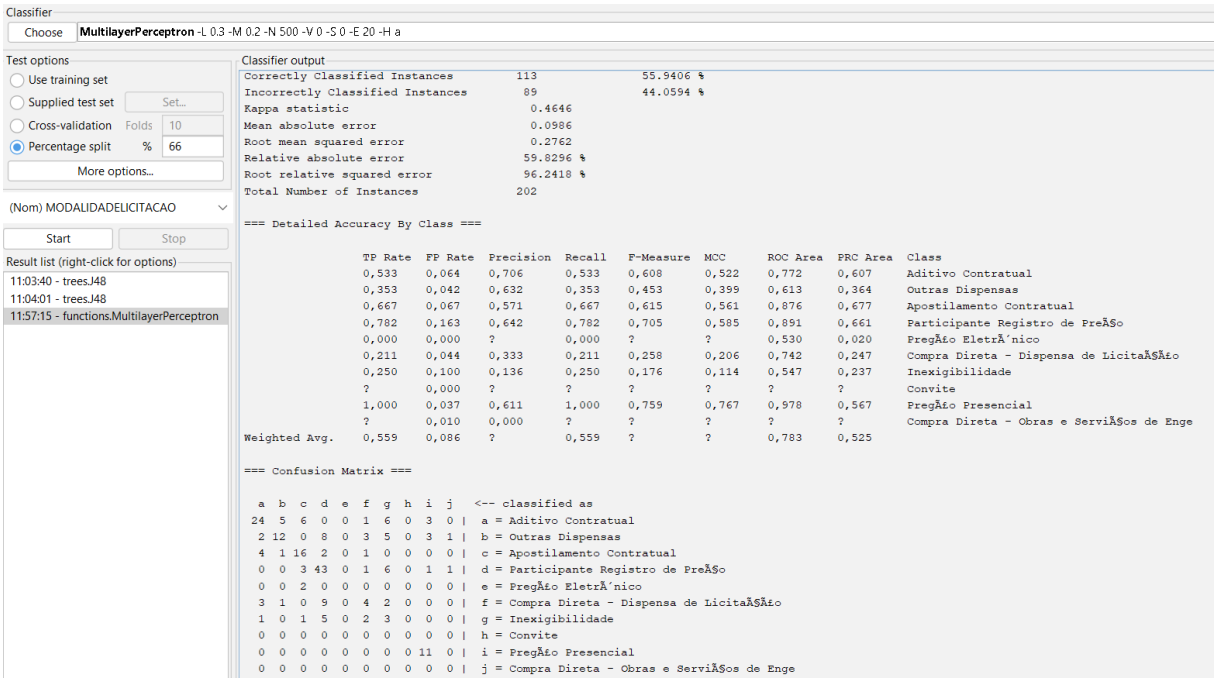
Fonte: elaborado pelo autor

Tanto a árvore de decisão do *Percentage split e Use Training* Set apresentou 163 folhas e com um tamanho de 231. Um trecho da árvore de decisão é mostrado na figura 6

4.4 Rede Neural

O primeiro experimento foi utilizando a árvore de decisão, usando o modelo de classificação *Multilayer Perception*, a opção de teste escolhido foi o *Percentage split* que tem como alvo modalidadelicitacao, o dataset foi dividido em 66% para fazer o treinamento e os demais para teste.

Figura 7: Rede Neural – *percentagem split*

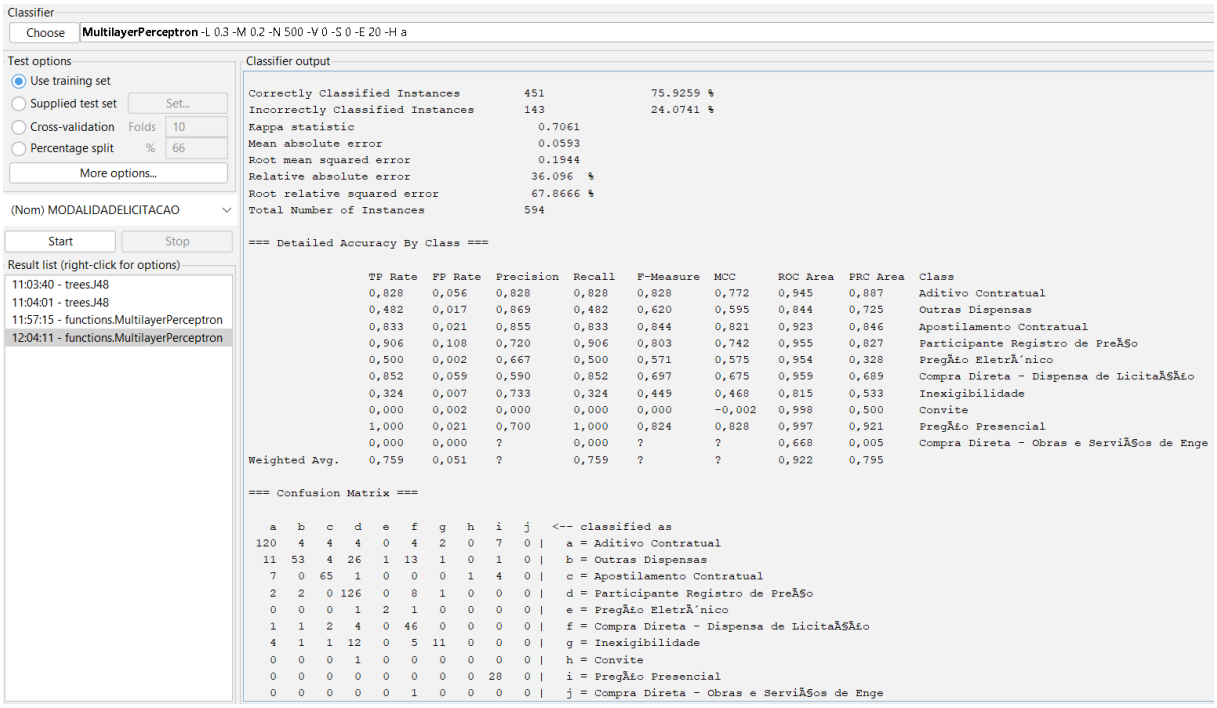


Fonte: elaborado pelo autor

Analisando a figura 7, o experimento realizado alcançou 55,94% de acurácia da instância de classificação corretamente e a instância de classificação incorretamente foi de 44.05%. Lendo a matriz de confusão do *Percentage split* notou-se que o grupo “a” houve 24 acertos, 12 no grupo “b”, 16 no grupo “c”, 43 no grupo “d”, 0 no grupo “e”, 4 no grupo “f”, 3 no grupo “g”, 0 no grupo “h”, 11 no grupo “i” e 0 no grupo “j”.

No segundo experimento a seguir opção de teste usada é a *Use training set*, o *dataset* utilizou todos os registros, obtendo 75.92% de acurácia e 24.07% de erro, como visto na figura 8.

Figura 8: Rede Neural – *training set*



Fonte: elaborado pelo autor

Lendo a matriz de confusão do *Use training set* notou-se que o grupo “a” houve 120 acertos, 53 no grupo “b”, 65 no grupo “c”, 126 no grupo “d”, 2 no grupo “e”, 46 no grupo “f”, 11 no grupo “g”, 0 no grupo “h”, 28 no grupo “i” e 0 no grupo “j”.

4.5 Comentário Gerais

Ao se comparar os experimentos em geral foram obtidos bons resultados com árvore de decisão e com redes neurais conforme tabela 3. Mas é visível que a opção de teste *training set* apresenta os melhores resultados, uma vez que os dados treinados são também os classificados. Ainda assim, o índice de percentagem split com a árvore de decisão foi adequado.

Tabela 3: Comparação entre árvore de decisão e rede neural

Experimentos	**Árvore de decisão**	**Rede Neural**
Training set	87,37%	75,92%
Percentage split	64,85%	55,94%

Fonte: Elaborado pelo autor

5 Conclusão

A proposta de se aplicar as técnicas de mineração de dados no contexto de licitações é vantajosa, conforme pode-se observar deste trabalho. Mesmo considerando uma abordagem exploratória, com um número limitado de licitações e fornecedores baseado nos meses de janeiro, fevereiro e março de 2019, os resultados indicam que é possível classificar as modalidades de licitação, de modo que os recursos que devem ser alocados para cada modalidade podem ser conhecidos a priori, desde que os dados da licitação por órgão e valor, principalmente, também sejam conhecidos.

Tal abordagem implica na formação de gestores que podem se especializar nestas modalidades, dentro de cada órgão ou dentro de cada âmbito federativo, ou mesmo levando em conta os valores, de forma a otimizar a gestão administrativa e jurídica. Entretanto, uma série de cuidados devem ser observados para a efetividade da proposta.

Em primeiro lugar, os dados não estão disponíveis em uma única fonte, sendo necessária uma limpeza e processamento preliminar para que seu uso seja viabilizado. Uma outra questão é que os casos considerados mais próximos da realidade, os de percentagem split, tem-se a árvore de decisão como técnica mais adequada. Não há uma razão clara, mas os resultados com redes neurais não se apresentam como uma técnica adequada.

5.1 Perspectivas para continuidade dos trabalhos

Recomenda-se para trabalhos futuros:

- Utilizar conjuntos de dados com maior volume;
- Utilizar outras tarefas de mineração de dados, tais como a associação.

Referências

AMORIM, T. **Conceitos, técnicas, ferramentas e aplicações de mineração de dados para gerar conhecimento a partir de bases de dados**. Universidade Federal de Pernambuco. Pernambuco. 2006.

BRASIL, C. Lei Nº 8.666, de 21 de Junho de 1993. Regulamenta o art. 37, inciso XXI, da Constituição Federal, institui normas para licitações e contratos da Administração Pública e dá outras providências. Diário da União, Brasília, 21 Junho 1993.

CASTRO, L. N. D.; FERRARI, D. G. **Introdução à mineração de dados:** conceitos básicos, algoritmos e aplicações. São Paulo: Saraiva, 2016.

CAMPOS, F. **As práticas de conluio nas licitações públicas à luz da teoria dos jogos**. Porto Alegre: Revista Análise Econômica, v. I, 2008.

CÔRTES, S. D. C.; PORCARO, R. M.; LIFSCHITZ, S. **Mineração de dados - Funcionalidades, Técnicas e Abordagem**. PUC RIO. [S.l.]. 2002.

DEVMEDIA. Mineração de dados com Market Basket Analysis. **DEVMEDIA**, 2020. Disponível em: <https://www.devmedia.com.br/mineracao-de-dados-com-market-basket-analysis-revista-sql-magazine-111/27853>. Acesso em: 19 Maio 2022.

DI PIETRO, M. S. Z. **Direito Administrativo**. 32. ed. Rio de Janeiro: Forense, 2019.

FONSECA, F. C. S. BELTRAME, W. A. R. Aplicações Práticas dos Algoritmos de Clusterização K Means e Bisecting K-means. **Researchgate**, 2010. Disponível em: <https://www.researchgate.net/publication/327121358>. Acesso em: 15 Maio 2022.

FORTINI, C.; MOTTA, F. **Corrupção nas licitações e contratações públicas:** Sinais de alerta segundo a transparência internacional. Belo Horizonte: Revista de Direito Administrativo, v. I, 2016

GIACOMEL, F. D. S. **Um Método Algorítmico para Operações na Bolsa de Valores Baseado em Ensembles de Redes Neurais para Modelar e Prever os Movimentos dos Mercados de Ações**. Universidade Federal do Rio Grande do Sul. Porto Alegre, p. 92. 2016.

HAYKIN, S. **Neural Networks and learning machines**. 3. Ed. Pearson, 2009.

NANDI, J. C. B.; PEREIRA, R. M.; FELIPPE, G. **O Algoritmo de Associação Frequent Pattern-Growth na Shell Orion Data Mining Engine**. Anais: SULCOMP, v. 7, 2015.

RODRIGUES, F. S. **RODRIGO Métodos de agrupamento na análise de dados de expressão gênica**. Universidade Federal de São Carlos. São Carlos. 2009.

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

SILVA, M. A. D. S. **Descoberta de Conhecimento na Análise de Licitações no Estado de Goiás**. Pontifícia Universidade Católica de Goiás. Goiás. 2020.

SILVA, L. A. D. P. **Introdução à Mineração de Dados - Com Aplicações em R:** com Aplicações em R. Rio de Janeiro: Elsevier, 2016.

SILVA, F. M. D. E. A. **Inteligência Artificial**. Porto Alegre: Sagah, 2019.

SILVA, C. V. S.; RALHA, C. G. **Detecção de cartéis em licitações públicas com agentes de mineração de dados**. Curitiba: Revista eletrônica de Sistemas de Informaçao, v. 10, 2011. Acesso em: 09 Maio 2022.

SOUZA, F. R. D. **Manual básico de licitação.** São Paulo: Nobel, 1997.

SOUSA, M. C. C. **UMA ANÁLISE DO ALGORITMO K-MEANS COMO INTRODUÇÃO AO APRENDIZADO DE MÁQUINAS**. Universidade Federal do Tocantins. Araguaína, p. 95. 2019.

PONTIFÍCIA UNIVERSIDADE CATÓLICA DE GOIÁS
GABINETE DO REITOR
Av. Universitária, 1069 • Setor Universitário
Caixa Postal 86 • CEP 74605-010
Goiânia • Goiás • Brasil
Fone: (62) 3946.1000
www.pucgoias.edu.br • reitoria@pucgoias.edu.br

RESOLUÇÃO nº 038/2020 – CEPE

ANEXO I

APÊNDICE ao TCC

Termo de autorização de publicação de produção acadêmica

O(A) estudante Christy Basílio da Silva do Curso de Ciência da Computação, matrícula 2017.1.0028.0180-3, telefone: 62 981078199 e-mail acnologia212@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 disponibilizar o Trabalho de Conclusão de Curso intitulado Classificação e *clustering* aplicados à licitações, 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); Imagem (GIF ou JPEG); Som (WAVE, MPEG, AIFF, SND); Vídeo (MPEG, MWV, AVI, QT); outros, 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, 22 de junho de 2022.

Assinatura do autor: Christy Basílio da Silva

Nome completo do autor: Christy Basílio da Silva

Assinatura do professor–orientador: Sibelius Lellis Vieira

Nome completo do professor-orientador: Sibelius Lellis Vieira