UNIVERSIDADE TECNOLÓGICA FEDERAL DO PARANÁ

GUSTAVO MARTINS DE SOUZA

DESENVOLVIMENTO DE UM MECANISMO DETECTOR DE ANOMALIAS EM LICITAÇÕES MUNICIPAIS

PATO BRANCO

2022

GUSTAVO MARTINS DE SOUZA

DESENVOLVIMENTO DE UM MECANISMO DETECTOR DE ANOMALIAS EM LICITAÇÕES MUNICIPAIS

Development of an Anomaly Detector Mechanism of Anomalies in Municipal Bidding

Trabalho de Conclusão de Curso de Graduação apresentado como requisito para obtenção do título de Bacharel em Engenharia de Computação do Curso de Bacharelado em Engenharia de Computação da Universidade Tecnológica Federal do Paraná.

Orientador: Prof. Dr. Marcelo Teixeira

PATO BRANCO

2022

GUSTAVO MARTINS DE SOUZA

DESENVOLVIMENTO DE UM MECANISMO DETECTOR DE ANOMALIAS EM LICITAÇÕES MUNICIPAIS

Trabalho de Conclusão de Curso de Graduação apresentado como requisito para obtenção do título de Bacharel em Engenharia de Computação do Curso de Bacharelado em Engenharia de Computação da Universidade Tecnológica Federal do Paraná.

Data de aprovação: 13/Dezembro/2022

Marcelo Teixeira
Prof. Dr. em Engenharia de Automação e Sistemas
Universidade Tecnológica Federal do Paraná

Viviane Dal Molin
Profª. Drª. em Informática
Universidade Tecnológica Federal do Paraná

Richardson Ribeiro
Prof. Dr. em Informática
Universidade Federal do Paraná

PATO BRANCO

2022

Dedico esse trabalho à minha esposa e família pelos momentos de ausência e por toda a ajuda que sempre precisei mas, em especial, ao meu falecido avô, ao qual prometi me esforçar a cada dia mais para concluir e realizar esse sonho que, com certeza, não é apenas meu, mas sim dele, minha esposa e demais familiares.

AGRADECIMENTOS

Agradeço ao meu orientador Prof. Dr. Marcelo Teixeira, por todo suporte e sabedoria com que me guiou nesta trajetória.

A Secretaria do Curso, pela cooperação.

Agradeço imensamente à minha esposa por todo apoio e suporte nos momentos em que achei que não conseguiria fazer todas as coisas que deveriam ser concluídas.

E por último, mas não menos importes, agradeço aos meus pais e irmãs que me deram todo o suporte e oportunidade de cursar e concluir um sonho que hoje é realidade.

Primeira Lei: Um robô não pode ferir um ser humano ou, por omissão, permitir que um ser humano sofra algum mal. Segunda Lei: Um robô deve obedecer as ordens que lhe sejam dadas por seres humanos, exceto nos casos em que tais ordens contrariem a Primeira Lei. Terceira Lei: Um robô deve proteger sua própria existência desde que tal proteção não entre em conflito com a Primeira e Segunda Leis (ASIMOV, 2015).

RESUMO

O Processo licitatório é o método utilizado pela administração pública para efetuar compras e vendas de recursos consumidos por uma instituição, como por exemplo, na compra de insumos para prefeitura como papéis, canetas, envelopes, até máquinas para revitalização de asfalto. Para auxiliar nestes processos licitatórios existem sistemas que monitoraram e detectam esses casos de anomalias, porém tais sistemas são, em geral, voltados a aplicações nas esferas públicas federais e/ou estaduais. Contudo, tem sido cada vez mais frequente casos e indícios de fraudes em licitações, seja pelo superfaturamento de preços, fraude de direcionamentos, falsificação de documentos, entre outros. Quando se trata da administração pública municipal, não há ferramentas capazes de identificar qualquer tipo de anomalias nesse processo, o que desperta relevância prática e social nesse tipo de inovação. Assim, nesse trabalho foram implementados recursos computacionais inteligentes capazes de encontrar preços de itens similares que a partir da comparação de preço de itens licitados e preços encontrados rastreiam fraudes em licitações. Com os resultados encontrados foi possível identificar possíveis indícios de fraude, de tal forma que um dos itens possuia um preço 79% mais alto que o encontrado pelo buscador automatizado enquanto que para o outro item constava um valor 43% acima do encontrado.

Palavras-chave: licitação; fraude; busca automatizada; administraçã pública; anomalia.

ABSTRACT

The bidding process is the method used by the public administration to make purchases and sales of resources consumed by an institution, for example, in the purchase of inputs for the municipality such as paper, pens, envelopes, even machines for asphalt revitalization. To assist in these bidding processes there are systems that monitor and detect these cases of anomalies, but these systems are, in general, geared towards applications in the federal and/or state public spheres. However, there have been increasingly frequent cases and indications of fraud in bidding processes, be it through overbilling of prices, directional fraud, document forgery, among others. When it comes to municipal public administration, there are no tools capable of identifying any kind of anomalies in this process, which arouses the practical and social relevance of this type of innovation. Thus, in this work we implemented intelligent computational resources capable of finding prices of similar items that, by comparing the prices of bid items and the prices found, track bidding frauds. With the results found it was possible to identify possible indications of fraud, in such a way that one of the items had a price 79% higher than the one found by the automated searcher, while the other item had a price 43% higher than the one found.

Keywords: bidding; fraud; automated search; public administration; anomaly.

LISTA DE FIGURAS

SUMÁRIO

1 INTRODUÇÃO

Licitação é o processo utilizado pela administração pública para contratar obras, serviços, compras e alienações, ou seja, é o método usado por órgãos públicos para efetuar compras e vendas (BRASIL, 2021). Basicamente, licitar significa buscar a condição mais interessante, dentre um conjunto de possibilidades, sob os aspectos econômico e financeiro, para realizar uma obra, serviço ou aquisição (JUNIOR, 2015).

Na maioria dos casos, por se tratar de grandes volumes financeiros, é muito comum que uma licitação envolva passos ilícitos e, infelizmente, existem várias maneiras de fraudar uma licitação. Fraude de direcionamentos, por exemplo, é aquela que já se sabe o licitante vencedor antes mesmo do processo finalizar; fraude na documentação e propostas, em que um grupo de pessoas, utilizando “laranjas”, falsificam a documentação, como endereços inexistentes; fraude no sobrepreço é quando o licitante vencedor e seus concorrentes aumentam o preço dos itens acima do comum, aumentando o valor da licitação, por exemplo, uma licitação que está custando R$150 mil, quando na verdade deveria custar R$50 mil (TEIXEIRA, 2016).

Em seu estudo, Carvalho *et al.* (2010) constataram que o uso de ferramentas como *Knowledge Discovery in Databases - KDD* e *Data Warehouse - DW* vem aumentando no que diz respeito a esfera pública o que, por consequência fornece benefícios como a descoberta de conhecimento útil na tomada de decisões, identificação de padrões que podem subsidiar um novo modelo de compras governamentais, melhora da eficiência e fiscalização do governo a partir do cruzamento de informações que caracterizam irregularidades ou melhores práticas. Ainda, além das ferramentas já comentadas, o uso de técnicas de mineração de dados têm gerado melhorias contínuas nos processos de compra, realizados pelo governo.

Estudos como o de Ralha e Silva (2012), Morais (2016) e Souto *et al.* (2019) - citados com mais detalhes na seção 1.1 - que a partir da aplicação de algoritmos de reconhecimento de padrões e descoberta de conhecimento, conseguiram encontrar anomalias em licitações federais. No entanto, tais ferramentas são dedicadas apenas para fiscalizações federais e estaduais, de modo que suporte similar não é encontrado em nível municipal.

Dessa forma, é de grande importância para a gestão pública municipal a implementação de ferramentas que identificam anomalias em processos licitatórios. Estima-se que, uma das formas de implementar tais seja a partir da integração de técnicas em *web crawlers* e bases de dados reais montadas a partir da integração de múltiplas bases de dados municipais. Nessa estratégia, as bases ideais de dados serviriam ao propósito de subsidiar a ferramenta com dados efetivamente usados em processos licitatórios; o *web crawler* complementaria esses dados com parâmetros comparativos médios extraídos da internet; juntos, dados reais e comparativos seriam manipulados por algoritmos bio-inspirados para efetivamente identificar anomalias.

1.1 Estado da Arte

Conforme comentado anteriormente, em Ralha e Silva (2012), os autores avaliam licitações realizadas no ano de 2008 em um determinado órgão do governo federal. Os resultados mostram que em nove licitações houve a participação de somente duas empresas, mas apenas uma delas era a vencedora nos nove processos licitatórios, o que leva a indícios de cartelização e simulação de concorrência, visto que a segunda empresa, que perdeu as nove licitação, participou apenas desses nove processos enquanto que a vencedora participou de doze processos, mostrando que não se tratava de um grande fornecedor. Já Fernandes *et al.* (2021) utilizam dados do portal da transparência e reportam, através de técnicas de mineração de dados, que em boa parte das licitações as empresas agrupam concorrentes da mesma área de atuação, evidenciando a possibilidade da formação de cartéis nesses processos.

No estudo sobre as licitações de obras de engenharia dos municípios do Ceará, realizado por Morais (2016), verificou-se que, a partir da descoberta de padrões e classificação não supervisionada de empresas consideradas inidôneas pelo Tribunal de Contas da União (TCU), Polícia Federal (PF) e Controladoria Geral da União (CGU) e empresas participantes como doadoras de campanha eleitoral e vencedora de litações, 11% das empresas participantes eram consideradas inidôneas por um dos três órgãos federais, mas eram doadoras em campanhas eleitorais. Além disso, percebeu-se também que as empresas consideradas inidôneas pelo TCU e pela PF eram as empresas com maior número de participações e vitórias.

Souto *et al.* (2019) utilizando mineração de dados, analisaram o uso de *bots* e *softwares* de disparo automático de lances em pregões eletrônicos. Apesar de não existir qualquer proibição, o uso desse tipo de programa coloca o fornecedor em vantagem sobre os demais, já que representou em mais de 5% de chance de sucesso nos itens de disputa de pregões eletrônicos do Ministério da Agricultura, Pecuária e Abastecimento analisados no ano de 2017.

Outros autores que já aplicaram técnicas de mineração de dados em licitações públicas, como é o caso do trabalho de Trindade, Chaves e Teixeira (2022), que através do algoritmo *Local Outlier Factor* em conjunto com algoritmo *K-Nearest Neighbors* - método de aprendizado de máquina - concluiram que houve algum tipo de fraude ou irregularidade visto que, em um universo de 10947 abastecimentos, aconteceram 117 abastecimentos que apresentaram comportamento fora do normal, totalizando 1,07%. Dessa forma é possível confirmar que alguns veículos apresentaram quantitades de combustível abastecidas além da capacidade máxima.

Petroski *et al.* (2021), utilizando dados públicos disponibilizados através do sistema de saúde E-Saúde, aplicou o KDD e técnicas de mineração de dados em conjunto com a ferramenta *Waikato Environment for Knowledge Analysis* (WEKA) e o algoritmo apriori e conseguiu localizar e fornecer aos gestores atributos importantes como faixa etária e procedimentos mais utilizados em uma determinada unidade de saúde de tal forma que auxiliaram na distribuição dos recursos. Por exemplo, uma unidade de saúde que possui muitos atendimentos a idosos pode solicitar e comprar melhores equipamentos geriátricos.

Tendo como base os breves exemplos mencionados, é possível observar que a gestão pública carece de ferramentas de identificação de possíveis anomalias, relacionadas a possíveis fraudes, em processos de licitação. Uma das formas de se fazer identificação de anomalias é utilizando algoritmos de Descoberta de Conhecimento em Bases de Dados (*Knowledge Discovery in Databases - KDD*) que, no geral, concentram-se em desenvolver métodos para agregar significado a dados e, assim, entregar informações úteis ao usuário (FAYYAD; PIATETSKY-SHAPIRO; SMYTH, 1996).

Ainda segundo Fayyad, Piatetsky-Shapiro e Smyth (1996), o processo de extração de padrões ganhou vários nomes, incluindo, mineração de dados (*Data Mining*). Contudo, o KDD corresponde ao processo de descoberta de padrões e extração de dados como um todo, sendo a mineração de dados uma das etapas que compõem o KDD, junto com a preparação, seleção e limpeza dos dados.

A mineração de dados aplica métodos de extração e reconhecimento de padrões, como a análise de *outliers*, por exemplo. Segundo Aggarwal (2017), um *outlier* pode ser definido como um ponto significativamente diferente de todos os outros dentro de uma base de dados. Essa análise é muito utilizada em sistemas de detecção de intrusos, detecção de fraude em cartões de crédito, diagnósticos médicos, por exemplo. Nos sistemas como os de detecção de fraude, os *outliers* correspondem a sequências de pontos diferentes.

Outra forma de mineração de dados é o *web crawler*, que de acordo com Sirisuriya *et al.* (2015), é uma técnica de extração de dados de *websites* para uma base de dados. Grandes empresas utilizam *bots* de busca automatizada para extrair informações da internet, um exemplo é a Google, que utiliza um robô para encontrar páginas da internet. O *web scraping* é o método de mineração de dados que roda por trás do *web crawler* armazenando dados não estruturados disponibilizados por páginas *web* e transformando em conhecimento útil.

Devido a recorrência de licitações fraudadas, robôs como ”Mônica“ (Monitoramento Integrado para o Controle de Aquisições), ”Sofia“ (Sistema de Orientação sobre Fatos de Indícios para o Auditor) e ”Alice“ (Análise de Licitações e Editais) foram criados para identificar irregularidades em licitações utilizando inteligência artificial em dados do Tribunal de Contas da União. Todavia, tais aplicações voltam-se apenas a esfera pública federal, poucos trabalhos que assim como o de Petroski *et al.* (2021) e Trindade, Chaves e Teixeira (2022) são aplicados na gestão pública municipal.

Assim, esse trabalho propõe a construção de uma ferramenta computacional capaz de detectar possíveis anomalias em processos licitatórios municipais. Para tanto, são implementadas técnicas pertencentes ao KDD, como os algoritmos de *web scraping*, para automaticamente realizar buscas na Internet por parâmetros de itens licitados, bem como procedimentos de tranformação de dados, capazes de lapidar as informações extraídas pelo *bot* para que o usuário final possa visualizar com maior clareza os pontos onde há indícios de anomalia na licitação.

Tais parâmetros são então incorporados à base de dados que integra registros de múltiplas bases reais manipuladas pela governança pública municipal, emergindo um modelo de

visualização e comparação de valores declarados em licitação municipais e valores encontrados pelo *web crawler*, apontando para a prefeitura os pontos críticos que sugerem a hipótese de uma anomalia.

1.2 Objetivos Específicos

- Criar uma ferramenta de busca de cotações na internet.
- Efetuar o tratamento dos dados recebidos pelo buscador de cotações.
- Implementar um método que compare os dados recebidos e dados encontrados para que pode apontar os casos de possíveis fraudes.
- Efetuar a carga dos resultados obtidos em uma base de dados.

2 REFERENCIAL TEÓRICO

Os avanços na área da Tecnologia da Informação (TI) têm viabilizado o armazenamento de grandes volumes de dados, por vezes armazenados em múltiplas bases, não raramente heterogêneas. Redes sociais, lojas virtuais, dispositivos móveis e sistemas embarcados são alguns exemplos de aplicações que tem acelerado o crescimento dessas bases de dados já que estão captando e processando cada vez mais esses dados.

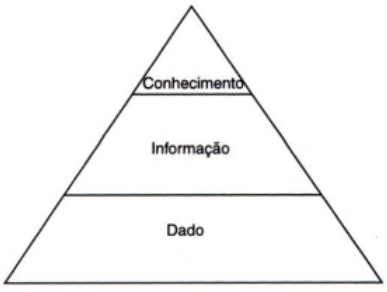
Segundo Goldschmidt, Passos e Bezerra (2015), a análise de grandes quantidades de dados é inviável sem a ajuda de ferramentas computacionais. De fato, esse problema decorre em geral do tamanho da base, da complexidade do enlace entre os dados, e da forma oculta como algumas informações estão armazenadas Larose e Larose (2014). Dessa forma, a habilidade de traduzir todos esses dados em conhecimento é algo que extrapola a capacidade humana de percepção e compreensão.

Nesse sentido, ferramentas computacionais surgem como viabilizadoras desse processo de interpretação de dados e geração de conhecimento. Elas executam a análise de grandes bases de dados por meio de um processo denominado Descoberta de Conhecimento em Bases de Dados (*Knowledge Discovery in Databases - KDD*). Na prática, o KDD é um conjunto de etapas para que o processo de extração de dados e reconhecimento de padrões para que os dados possam entregar informações úteis ao usuário.

2.1 O KDD

É possível, por meio do KDD, executar a extração dos dados de uma base qualquer, limpar e transformar esses dados para que então sejam processados por meio de algoritmos adequados para a conversão em informação. Conforme Goldschmidt, Passos e Bezerra (2015), os dados e sua qualidade inerente (ou a falta dela) - encontrados na base da pirâmide da Figura 1 - compõem o pilar fundamental para a boa condução desse processo, pois expressam fatos do mundo real sobre os quais ainda não se tem total compreensão.

Figura 1 – Hierarquia entre Dado, Informação e Conhecimento



Fonte: (GOLDSCHMIDT; PASSOS; BEZERRA, 2015).

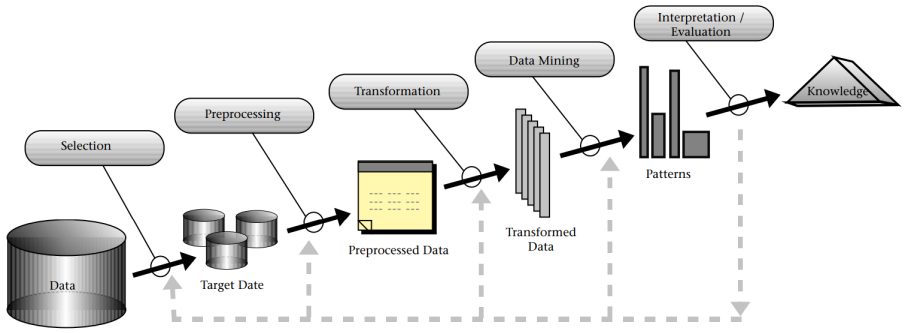
Para que a transformação desses dados seja feita é necessário utilizar *frameworks* capazes de receber esses dados e transformá-los em informação útil da maneira mais fácil possível e, para isso, é possível utilizar ferramentas *open source* como o Pandas.

De acordo com McKinney *et al.* (2011), esse *framework*, tem como objetivo fechar os *gaps* de análise de dados do Python, em relação a outras linguagens de programação, como R ou Matlab, por exemplo. Além disso, ao utilizar pandas e python, como forma de análise de dados, mais funcionalidades são fornecidas permitindo que as formas de visualização sejam expressivas e de fácil a implementação.

Esses dados, extraídos e armazenados por recursos de TI, são cadeias de símbolos que não possuem semântica clara, em princípio, mas que, sujeitos a processamento computacional adequado, se transformam em informações diferenciadas para a gestão da informação, o que em geral se converte em competitividade de mercado.

Segundo Goldschmidt, Passos e Bezerra (2015), a informação representa o dado tratado com significado e contexto bem definido, bem como o conhecimento que corresponde a um padrão ou conjunto de padrões cuja formulação pode envolver e relacionar dados e informações. Esse processo de extração de dados e reconhecimento de padrões do KDD, Segundo Fayyad, Piatetsky-Shapiro e Smyth (1996), é um processo não trivial. Ele é composto por 5 passos, que manipulam os dados até chegar ao conhecimento desejado sobre a base de dados, como mostra a Figura 2.

Figura 2 – Passos do KDD.



Fonte: (FAYYAD; PIATETSKY-SHAPIRO; SMYTH, 1996).

As três primeiras etapas mostradas na Figura 2, também chamadas de “Análise de Dados”, são responsáveis por identificar quais dados são de fato necessários pra a descoberta de conhecimento. Muitas vezes esse seleto conjunto de dados aparece na base em meio a muitos outros dados que, mesmo essenciais para a parte operacional da base, em geral são irrelevantes para o KDD e, por isso, precisam ser eliminados.

São nessas etapas, também, em que os dados são padronizados para processamento computacional. Por exemplo, se a análise de dados está sendo feita sobre rejeitos de minério, é interessante padronizar todos os dados que envolvem peso para toneladas, já que em grandes mineradoras é possível que toneladas de metais sejam extraídos da terra todos os dias.

Também é nas 3 primeiras fases que os dados são limpos, eliminando “ruídos” que possam alterar a natureza semântica dos dados. Por exemplo, quando uma loja de calçados deseja saber qual numeração de calçados foi menos vendida durante o mês. Quando os dados são coletados, todos os modelos de todas as marcas são coletados juntos e, portanto, calçados de uma marca diferente devem ser excluídos, pois podem impactar na análise final (SANTOS *et al.*, 2016).

Em resumo, é na análise de dados que são encontradas as “regras de negócio” de um projeto de dados. Essas regras definem o modo com que o negócio funciona, “como os recursos devem ser gerenciados ou como situações especiais devem ser tratadas na execução dos processos de negócios” (ROSCA *et al.*, 1997). É aqui que os atributos da etapa de engenharia (mineração de dados) são escolhidos, por exemplo, em um relatório de vendas de uma emissora de TV ou rádio. Vendas diferentes podem ter classificações variadas que serão identificadas por atributos distintos.

Essas 3 primeiras fases de extração, transformação e carga (*Extract, Transform and Load* - ETL), também conhecidas como “pré-processamento” de dados são, de acordo com Han, Pei e Kamber (2011), importantes para a próxima etapa, denominada mineração de dados (*Data Mining*), uma vez que são utilizadas para a construção de um armazém de dados (*Data Warehouse* - DW). Um DW fornece uma arquitetura e as ferramentas para organizar, entender, auxiliar na tomada de decisões e possibilitar o processamento analítico online (*Online Analytical Process* - OLAP) para a análise de dados multidimensionais com granularidades variadas, o que facitila na generalização dos dados e a aplicação de mineração de dados.

Áreas como as finanças, varejo e a saúde são as que mais utilizam mineração de dados para prever vendas (varejo) e obter melhores *insights* sobre sintomas (saúde), por exemplo. Independentemente da área, a mineração de dados pode ser aplicada para o reconhecimento de padrões, aplicação de modelagens e correlação de informações para cruzar bases, levando a um ponto central de conhecimento, impulsionando estratégias para a obtenção de lucro, inovação e progresso tecnológico (GOLDSCHMIDT; PASSOS; BEZERRA, 2015).

Conforme comenta Halmenschlager (2002), a etapa de mineração de dados poder ser separada, internamente, em 3 partes sendo elas: a escolha da tarefa de mineração de dados, a escolha do algoritmo de mineração e a aplicação da mineração. Na primeira temos a tomada de decisão sobre qual tarefa será utilizada, podendo ser classificalão, regressão, associação etc. Na segunda temos a escolha do algoritmo de mineração de dados que - segundo a autora - entre os modelos mais utilzados estão as arvores de decisão, regras de classificação, redes neurais e os algoritmos genéticos. Por fim, na terceira etapa, temos a aplicação do método sobre os dados tratados no passo anterior que a busca por padrões aconteça.

O *Data Mining* pode ser aplicado em várias formas de dados, utilizando inteligência artificial e estatística, matemática e técnicas de aprendizagem. Algumas formas de dados utilizadas para aplicação de mineração de dados são dados de bancos de dados, dados de armazém de

dados e dados transacionais. Existem ainda, segundo Han, Pei e Kamber (2011), outras formas de dados que possíveis de se minerar, como dados de texto, multimídia e da internet.

2.2 Web Scraping

A mineração de dados da internet, técnica utilizada neste trabalho para buscar valores de itens licitados, usa um servidor especializado em buscar informações disponíveis na *web*. Os resultados dessa busca podem ser, por exemplo, páginas *web*, imagens e bases de dados públicas, entregues em listas. Esse mecanismo é, essencialmente, uma grande aplicação de mineração de dados, com várias técnicas pertencentes ao *data mining* em todos os aspectos do motor de busca, desde *web crawling*, indexação e busca.

O *web scraping* é o método de mineração de dados que roda por trás do *web crawling* transformando os dados extraídos de *websites* em dados estruturados, como bancos de dados ou planilhas permitindo a aplicação de algoritmos de análise de dados (SIRISURIYA *et al.*, 2015). Essa extração pode ser realizada utilizando *Hypertext Transfer Protocol* (HTTP) ou ainda por navegadores de internet como o Google Chrome ou Mozilla Firefox (ZHAO, 2017).

A Figura 3 demonstra, brevemente, a forma com que o algoritmo de *web scraping* funciona. Nesta aplicação, o *bot* é responsável por buscar na internet por itens similares aos apresentados na licitação em questão para que em um segundo momento estes resultados sejam tratados e armazenados de forma estruturada em uma nova base implementada.

Figura 3 – Etapas do *web crawler*

Fonte: Autoria própria.

Essa técnica, além de converter linguagens de marcação como JSON, HTML e outras em um conteúdo legível também pode ser integrada com análise visual de computadores e processamento de linguagem natural para replicar a forma com que os humanos navegam pela internet.

A aplicabilidade desse método é muito grande, já que é possível buscar qualquer informação dentro de uma página *web*. O *GroupLens* por exemplo, é um sistema distribuído que

prevê o interesse do usuário sobre um determinado assunto debatido em um artigo utilizando *web scraping* (RESNICK *et al.*, 1994). Ou ainda, o *MovieLens*, uma ferramenta também desenvolvida pela universidade de Minessota que faz recomendação de filmes (DIOUF *et al.*, 2019).

Ainda segundo Diouf *et al.* (2019), o web scraping é muito utilizado no jornalismo, através de uma plataforma chamada *ScraperWiki*. Essa plataforma foi desenvolvida em Python e é utilizada para auxiliar os jornalistas nas suas tarefas diárias.

Como cita Santos *et al.* (2009), um dos algoritmos de mineração de dados da internet mais conhecido é o *PageRank*, utilizado pelo Google para efetuar suas buscas. Para retornar os resultados, esse algoritmo usa uma estrutura de grafos para interligar os resultados relacionados a aquela busca.

Existem várias ferramentas de tipos diferentes para se extrair dados da internet. Essas ferramentas podem ser extensões de navegador, softwares e plataformas e ainda bibliotecas de linguagens de programação. Na extensões, temos por exemplo, o "*Spider*", que consegue capturar tudo o que está sendo exibido na tela, representar isso como uma coluna e depois baixar como arquivo JSON ou CSV (DIOUF *et al.*, 2019).

Ferramentas como "*Data Scraper*" e "*Data Miner*", assim como o *Spider*, conseguem exportar todas as informações da página *web* para arquivos CSV ou XLS. O *Data Miner* ainda possui uma funcionalidade extra, que é possuir várias consultas de SQL prontas em seu *layout* (DIOUF *et al.*, 2019). Por fim ainda pode-se citar o "*Dexi.io*" uma extensão que consegue recuperar dados em tempo real de qualquer *site*.

Por fim, existem diversas bibliotecas de diversas linguagens de programação capazes extrair informações de páginas da internet. Para o Python temos ferramentas como o *Beautiful Soup*, *Newspaper*, *LXML*, *Selenium* etc. Para este trabalho, foi escolhida a biblioteca *Selenium* já que com ela é possível utilizar o navegador da mesma forma que um humano.

O Selenium, é um *framework open source* utilizado para fazer testes de requisições POST e GET. Embora seja muito utilizado para estes testes, ele também pode ser aplicado para mineração de dados, já que este é capaz de navegar por sites da *web* apenas enviando comandos pelo navegador. Contudo, apesar de ser capaz de converter linguagem de marcação em um texto legível, esse conteúdo ainda precisa de tratamento e, para isso é necessário que uma ferramenta como o Pandas, *framework* citado anteriormente, seja utilizado.

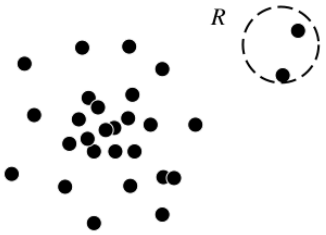
Já para a carga dos dados, o Apache Spark, que assim como o Selenium e o Pandas, é um *framework open source* distribuído, utilizado em projetos de *big data*, já que é capaz de processar quantidades massivas de dados. Essa vantagem do spark sobre outras ferramentas é graças à sua habilidade de armazenar informações em um cache de memória bem como a execução otimizada de consultas. Além disso, o spark é suportado em várias linguagens de programação como Python - linguagem utilizada neste trabalho - Java, R e Scala além do mais, o spark oferece bibliotecas de *Machine Larning*, SQL, análise de gráfico e streaming de dados (SPARK, 2018).

2.3 Análise de outliers

Além da mineração de dados extraídos da internet e outras bases de dados estruturadas, o *data mining* consegue efetuar análises em bases de dados buscando por anomalias, como é implementado no trabalho de Shah *et al.* (2002), que busca por padrões em lances de leilões e identificam possíveis casos de tentativa de fraude.

A mineração de dados a ser proposta aqui para a detecção de fraudes, se utiliza de algoritmos de detecção de *outliers*, método de mineração de dados que busca por anomalias na base. Quando esses dados são criados por um ou mais processos de geração pode ocorrer a criação desses *outliers* (AGGARWAL, 2017). Esses algoritmos encontram dados com comportamentos diferentes do esperado. Por exemplo, cartões de crédito, eles possuem um padrão de transações, no entanto, quando um cartão é roubado esse padrão muda drasticamente - a localização das compras são muito diferente da localização do dono do cartão (HAN; PEI; KAMBER, 2011). A Figura 4 mostra um exemplo de *outlier* em uma base de dados.

Figura 4 – Os pontos dentro da região R são *outliers*.



Fonte: (HAN; PEI; KAMBER, 2011).

Ressalta-se que o conceito de *outlier* é diferente do conceito de ruído. Um ruído é um erro aleatório ou variação em uma variável calculada e por isso, normalmente não é interessante para a análise de dados Han, Pei e Kamber (2011). Já o *outlier* impõe um papel adverso à análise, pois reporta algo anormal. Porém, sua detecção não deixa de ser interessante para, por exemplo, constatar o percentual de um determinado cenário que está fora da normalidade ou, ainda, para associar cada outlier a alguma irregularidade prática, o que se alinha com as metas deste trabalho.

O último passo do KDD, mostrado na Figura 2, é a interpretação ou validação do modelo. Uma das formas de efetuar a validação do modelo é utilizando o *Business Intelligence* (BI), essa é a parte do KDD em que os resultados da mineração de dados podem ser interpretados por humanos ou algoritmos preditivos.

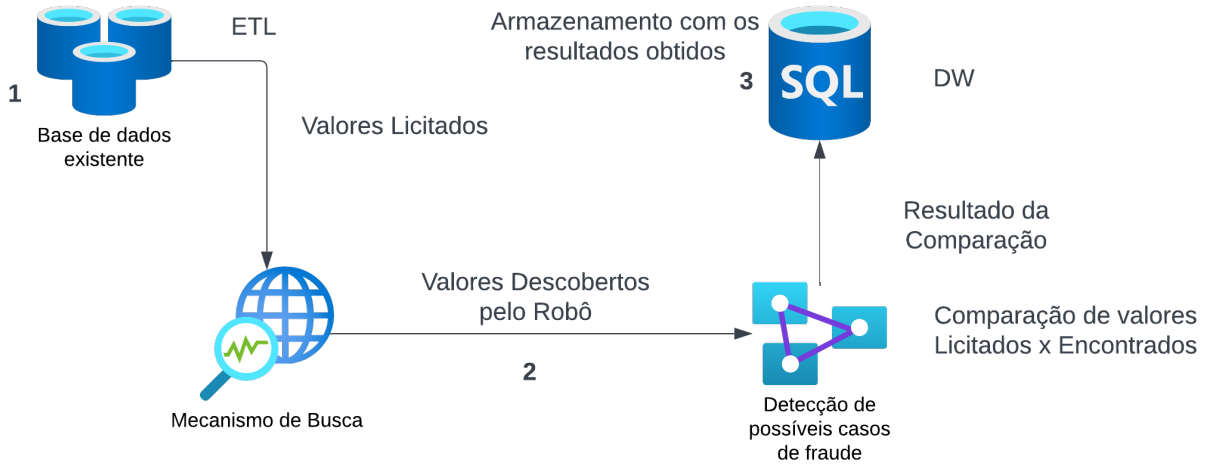
O conceito de BI surge devido a necessidade do mercado de entender, em um contexto comercial, os seus consumidores e concorrentes, por exemplo. As tecnologias que envolvem o BI fornecem um histórico atual e com visões preditivas sobre *benchmarking* e análise de *marketing*. Contudo, o BI só é possível graças as técnicas de *data mining* que auxiliam no estudo (HAN; PEI; KAMBER, 2011).

Por fim, como forma de validação dos resultados obtidos neste trabalho, serão utilizados métodos de BI, como as *dashboards*, por exemplo. Um painel de controle *dashboard* é uma página de relatório utilizada para visualizar a produção diária de uma mineradora, por exemplo. É possível utilizar esse tipo de relatório pra identificar possíveis mudanças repentinas em um equipamento ou empresa.

3 MATERIAIS E MÉTODOS

A detecção de anomalias em licitações públicas da rede municipal, aspecto motivador para a elaboração deste trabalho, seguiu os passos metodológicos mostrados na Figura 5. Para a aplicação das ferramentas e métodos citadas na seção 2 foi utilizada a base de dados de licitação LCCETIL, fruto de uma parceria de pesquisas com os autores deste trabalho. Tal base, construída a partir de um *framework web* implementado em uma prefeitura do estado do Rio Grande do Sul, armazena ainda, outras informações relacionadas ao município em questão como informações sobre abastecimento de frotas, saúde e etc, sendo a parte de licitações umas das aplicações.

Figura 5 – Passos da construção da proposta



Fonte: Autoria Própria.

Os passos presentados na Figura 5 seguem as etapas do KDD, citados na Figura 2 da seção 2. O primeiro passo deve ser extrair os dados da base de dados atual para que estes possam ser repassados ao mecanismo de busca. No passo 2 está a etapa de mineração de dados. Nesta etapa, os dados extraídos da base LCCETIL devem ser repassados ao método que utiliza o bot *web crawler* que busca por produtos similares na internet, efetua os tratamentos e transformações necessários e por fim armazena todos os itens encontrados em uma nova base. Em seguida o valor dos itens licitados são comparados à mediana dos itens encontrados e então é usado um critério de tolerância de preço de 30% (acima ou abaixo) para que uma licitação seja ou não considerada irregular. Por fim, na etapa 3, como forma de validação, tem-se um modelo de dados, extraído e tratado, que é carregado nesta nova base de dados possibilitando a visualização da comparação efetuada anteriormente.

3.1 LCCETIL

A base LCCETIL é parte de um sistema cuja licença é particionada por município. Assim, essa base contém informações referentes aos processos licitatórios implementados naquele município. Dentre as informações armazenadas, estão os processos licitatórios que aconteceram separados por ano, como mostrador na Figura 6.

Figura 6 – Dados da base

	dtAnoProcesso	dsPrazoExecucao	dsValidadeProposta	daObjeto
1	2021	12 meses	12 meses	AQUISIÇÃO DE MOBILIÁRIO PARA O LABORATÓRIO MUNICIPAL, através do Sistema de Registro de Preço
2	2021	12 meses	12 meses	AQUISIÇÃO DE COMBUSTÍVEL - ÓLEO DIESEL S10, através do Sistema de Registro de Preço
3	2021	12 meses	12 meses	AQUISIÇÃO DE LONAS PLÁSTICAS, através do Sistema de Registro de Preço
4	2021	60 dias	60 Dias	AQUISIÇÃO DE EQUIPAMENTOS AGRÍCOLAS, com recursos do Convênio nº 889.264/2019/MAPA.
5	2021	12 meses	12 meses	AQUISIÇÃO DE MEDICAMENTOS BÁSICOS, através do Sistema de Registro de Preço
6	2021	12 meses	12 meses	AQUISIÇÃO DE MEDICAMENTOS INJETÁVEIS, através do Sistema de Registro de Preço

Fonte: Autoria Própria.

Ainda, estão inclusas as empresas que participaram, o objetivo de tal licitação, a descrição dos itens licitados - com quantidades e preços licitados - empresas que ganharam as licitações, da mesma maneira que empresas que foram 'banidas' de participar das licitações daquele município em questão, como mostrado na Figura 7.

Figura 7 – Item-fornecedor desclassificado

	cdTipoProcesso	dtAnoProcesso	nrProcesso	cdFornecedor	nrItem	dsMotivo
1	1	2013	596	1253	1	Não foi possível gravar habilitação
2	1	2013	596	1253	2	Não foi possível gravar habilitação
3	4	2016	5225	4702	1	Proposta (planilha) assinada por pessoa sem qualificação técnica
4	4	2016	5225	4702	2	Proposta (planilha) assinada por pessoa sem qualificação técnica
5	4	2016	5225	4702	3	Proposta (planilha) assinada por pessoa sem qualificação técnica
6	4	2016	5225	4702	4	Proposta (planilha) assinada por pessoa sem habilitação técnica

Fonte: Autoria Própria.

A partir dessa base, foram extraídas as informações utilizadas e desenvolvida uma estratégia de carga do DW com os principais atributos que poderiam subsidiar uma rotina de extração de conhecimento. Entre os atributos que serão carregados no DW estão: a descrição dos itens licitados e o preço desses itens.

3.2 Extração, Transformação e Carga

Extração, transformação e carga dos dados são os três primeiros passos deste trabalho, e é a partir deles que ocorre a criação do DW. Para isso, foi criado um esquema de tabelas que recebe algumas das informações referentes ao processo licitatório, como a descrição do item licitado, a quantidade de itens e o preço unitário, conforme ilustra a Figura 8, mostrada abaixo.

Figura 8 – Dados utilizados para a busca com o *bot*

DsMaterial	vlUnitUltEntrada
Óleo Lubrif.Motor Diesel 15W40 Turbo	9.99
Óleo Hidráulico 68	6.7
Óleo lubrificante para engrenagem SAE 90	8.7714
Óleo lubrificante para diferencial SAE 140	8.091
Fluído p/ Freios - 500 ml	24.5357
Desengripante	6.25
Querozene	10
Abraçadeira Tipo "U" 1x3/4	6.09
Abraçadeira 4" Camburão	9.87

Fonte: Autoria Própria.

A partir da base de dados "LCCETIL", comentada anteriormente neste capítulo foram identificados alguns campos que, posteriormente, foram utilizados na etapa de mineração de dados. Esses atributos são fundamentais pra que a análise de anomalias ocorresse já que a partir delas é possível efetuar a busca por preços parecidos ou que mostrem indícios dessas anomalias.

Para que fosse possível extrair tais informações da base de dados do município foi implementado um script em python, utilizando uma biblioteca de manipulação de dados chamada Pandas. Essa ferramenta, como descrita anteriormente e pelo próprio desenvolvedor, é uma ferramenta *open source* de análise e manipulação de dados rápida, poderosa e fácil de ser utilizada.

Nessa etapa, o Pandas foi utilizado para extrair as informações necessárias (descrição e preço unitário de cada item), a partir de um relatório gerado pelo sistema utilizado no município, para que o buscador pudesse encontrar esses itens anteriormente licitados bem como efetuar o calculo médio de preço desse determinado item.

Para que isso acontecesse foi gerado um arquivo no formato csv, que pode ser lido pelo pandas, dessa forma, os dados podem ser identificados e serializados a ponto de possibilitar a leitura a partir de laços de repetição, assim como é feito em listas. A partir dessa lista, o campo de descrição do item é repassado para um método que vai busca na *web* os dados licitados, efetua um tratamento e carrega a nova base de dados. O trecho abaixo mostra como esse processo é feito.

```
1       products = pd.read_csv('PathtoFolder/arq.csv',delimiter=';')
2       print(products)
3       for index, row in products.iterrows():
4           print(row["DsMaterial"])
5           SearchItem(driverPath, url, row["DsMaterial"], float(row["vlUnitUltEntrada"]))
```

Ainda com o Pandas, no momento da geração do arquivo csv, foi efetuada uma modificação na amostra total disponibilizada. Inicialmente a amostra era de aproximadamente 2500 registros de itens, contudo, após alguns testes foi possível identificar que o Google percebe que está sendo manipulado por um *bot* e então bloqueia a utilização do Google Shopping, forçando o usuário a autenticação para garantir que não é um robô quem está utilizando a ferramenta. Dessa forma, a amostra teve que ser reduzida para a aproximadamente 50 registros.

3.3 *Scraping*

Após finalizar a etapa de extração do ETL da base de dados para o DW, iniciou-se a segunda fase do projeto, mostrada na Figura 5, e foram aplicados os métodos de mineração de dados, conforme comentado na seção 2. Para isso, foi necessário repassar as informações extraídas ao *web crawler* para que a busca de valores fosse efetuada bem como a constatação de possíveis anomalias.

Primeiro, a descrição dos itens é repassada ao *web crawler* que efetua uma busca na internet por preços desse item ou itens parecidos para então calcular um preço médio. O script responsável por buscar esses dados foi implementado dentro de um método que além da descrição do item licitado, recebe como parâmetro o preço deste item licitado e caminho até o driver do navegador escolhido para efetuar tal busca.

Para a implementação do *bot* de busca automatizada, foi utilizada a biblioteca selenium por meio de scripts Python bem como o navegador web Google Chrome através da plataforma de compras Google Shopping.

Com o selenium, é possível impor configurações de uso ao navegador de tal forma que a busca aconteça sem que o usuário veja o que está acontecendo. Logo, a primeira tarefa com o selenium é configurar o navegador para que não seja exibido ao utilizador os resultados da busca. Com isso, também fica inviável que usuário tente burlar o processo de busca e identificação de preços.

Ainda, com o selenium, é possível localizar as informações necessárias através das *tags* HTML, como mostra a figura 9. Desta forma, para dar continuidade a etapa de mineração de dados, é possível identificar qual a *tag* responsável por um determinado campo do navegador com a função de "inspecionar"do navegador. Para que tal ação possa acontecer, com outra janela do navegador aberta, uma busca manual foi realizada e, a partir desta, é possível encontrar os itens através da *tag "class"*.

Figura 9 – Resposta entregue pelo navegador.

Fonte: Autoria própria.

Após localizar as informações desejadas (1, 2 e 3 na figura 9), o selenium armazena os resultados em uma lista de listas.Para que as informações requeridas fossem localizadas, houve a necessidade de percorrer essas listas e novamente, através de *tags*, localizar os dados desejados para que finalmente o resultado da busca pudesse ser visualizado.

Olhando separadamente cada uma desses estruturas é possível encontrar o título do item, o preço, e a loja que revende este produto. Além disso no trecho de código mostrado abaixo, é possível visualizar que foi necessário efetuar alguns tratamentos nos dados extraídos, como por exemplo, remover todos os possíveis caracteres do alfabeto ou caracteres especiais em campos posteriormente numéricos, ou ainda, adicionar o valor “0.0” quando o campo de preço retornasse um valor vazio.

```
1 for item in itemList:
2         print(item.text)
3         for name in item.find_elements(By.CLASS_NAME, 'C7Lkve'):
4             itemName.append(name.text)
5             print(name.text)
6         for preco in item.find_elements(By.CLASS_NAME, 'a8Pemb.OFFNJ'):
7             print(preco.text)
```

```
8                if(preco.text == ''):
9                    itemPrice.append('0.0')
10               else:
11                   itemPrice.append(re.sub('[a,b,c,d,e,f,g,h,i,j,k,l,m,n,'
12                                           'o,p,q,r,s,t,u,v,w,x,y,z,$,+,%,'
13                                           '#,@,!,&,*,(,),-,=,/]',preco.text.lower()))
14           for store in item.find_elements(By.CLASS_NAME, 'aULzUe.IuHnof'):
15               print(store.text)
16               itemStore.append(store.text)
```

Ao final da investigação sobre a lista principal, os dados foram armazenados separadamente em outras 3 novas listas, sendo elas "*itemName*", "*itemPrice*" e "*itemStore*" que, com a intenção de encaminhar para final do processo de extração, foram mescladas em um único objeto chamado dicionário. Esse dicionário vai armazenar os atributos e seus valores que, em seguida, são transformados em estruturas denominadas pandas *dataframe* de tal forma que cada lista - agora adicionada no dicionário - representa uma coluna deste *dataframe*.

Em um último passo, antes de transformar esse pandas *dataframe* em um *dataframe* do spark, foram adicionadas outras 2 colunas, denominadas de: "preco_medio" e "data" sendo o preço médio calculado através de uma função já implementada que encontra o valor da mediana de um *dataframe*.

```
1  dict = {'item':itemName, 'preco':itemPrice, 'loja':itemStore}
2      pdf = pd.DataFrame(dict)
3      preco_medio = round(float(pdf['preco'].median()),2)
4      pdf.insert(3,"preco_medio",value=round(float(pdf['preco'].median()),2),
5                 allow_duplicates=True)
6      pdf.insert(4,"data",value=datetime.date.today(),allow_duplicates=True)
```

Após finalizar a etapa de busca de itens licitados, o *dataframe* pandas é transformado em um *dataframe* spark para que os dados, anteriormente tratados, pudessem ser carregados para o DW, alocado em uma máquina virtual na plataforma de serviços de computação *Amazon Web Services*(AWS). Essa máquina virtual é um computador com pouco poder de processamento, servindo apenas para armazenar esse DW.

A real utilização deste *framework* é devido ao fato de que grandes quantidades de informações podem ser encontradas pelo *bot* podendo inviabilizar o uso deste além de possuir certa facilidade em conectar-se a bancos de dados remotos, já que a maioria de suas aplicações é em bases de dados remotas. Assim, após todas as transformações, um *dataframe* spark é criado a partir do *dataframe* do pandas para que então as informações sejam carregadas na base de dados da AWS.

Como um último passo antes da análise e localização de possíveis anomalias, cerca de 60 itens(por busca), extraídos com o *bot* são armazenados em uma segunda tabela. Essa tabela carrega consigo a descrição, preço localizado e site de revenda do item licitado, bem como a data da extração dos dados, já que os preços estão suscetíveis a mudança no preço ao longo dos dias, meses e certamente anos.

3.4 Validação

Com os dados encontrados pelo *web crawler* e o resultado do detector de anomalias, é possível validar se essa análise ocorreu de acordo com o esperado ou não.

O detector de anomalias deve receber como entrada os dados do item licitado como a descrição e preço unitário, bem como a mediana dos valores encontrados pelo *bot*. Tendo essas informações, é realizada uma comparação entre os preços, iniciando pela diferença entre a licitação e a cotação encontrada e, por fim, o percentual dessa diferença sobre o valor licitado, considerando como uma licitação sem anomalias resultados em que os valores encontrados diferem em até 30% acima ou abaixo do valor licitado. Conforme a Figura 10.

Figura 10 – Validação da Licitação

produto_licitado	preco_licitado	preco_medio_enconctrado	diferenca	percentural	Resultado	data
Abraçadeira 4" Camburão	9,87	23,59	-13,72	139	Subpreco	16/11/2022
Abraçadeira Tipo "U" 1x3/4	6,09	3,50	2,59	43	Sobrepreco	16/11/2022
Adesivo Silicone	8,90	18,98	-10,08	113	Subpreco	16/11/2022
Batente 542 412 131	59,00	142,82	-83,82	142	Subpreco	16/11/2022
Bico p/Engraxadeira	15,50	28,65	-13,15	84	Subpreco	16/11/2022
Borracha Estabilizador N.º 2500 226	7,40	46,07	-38,67	522	Subpreco	16/11/2022
Braço de Direção Universal	75,00	141,29	-66,29	88	Subpreco	16/11/2022
Bucha Artc. Semi Eixo Diant. Ford	7,35	84,36	-77,01	1047	Subpreco	16/11/2022
Bucha Coluna Direção Kombi	6,73	39,99	-33,26	494	Subpreco	16/11/2022
Cabo Embreagem Kombi	42,12	29,99	12,13	29	Regular	14/11/2022
Cadeado 25 mm	9,40	18,20	-8,80	93	Subpreco	14/11/2022
Cadeado 30mm	17,26	21,27	-4,01	23	Regular	16/11/2022
Câmara Ar 1000x20	100,00	164,51	-64,51	64	Subpreco	16/11/2022

Fonte: Autoria própria.

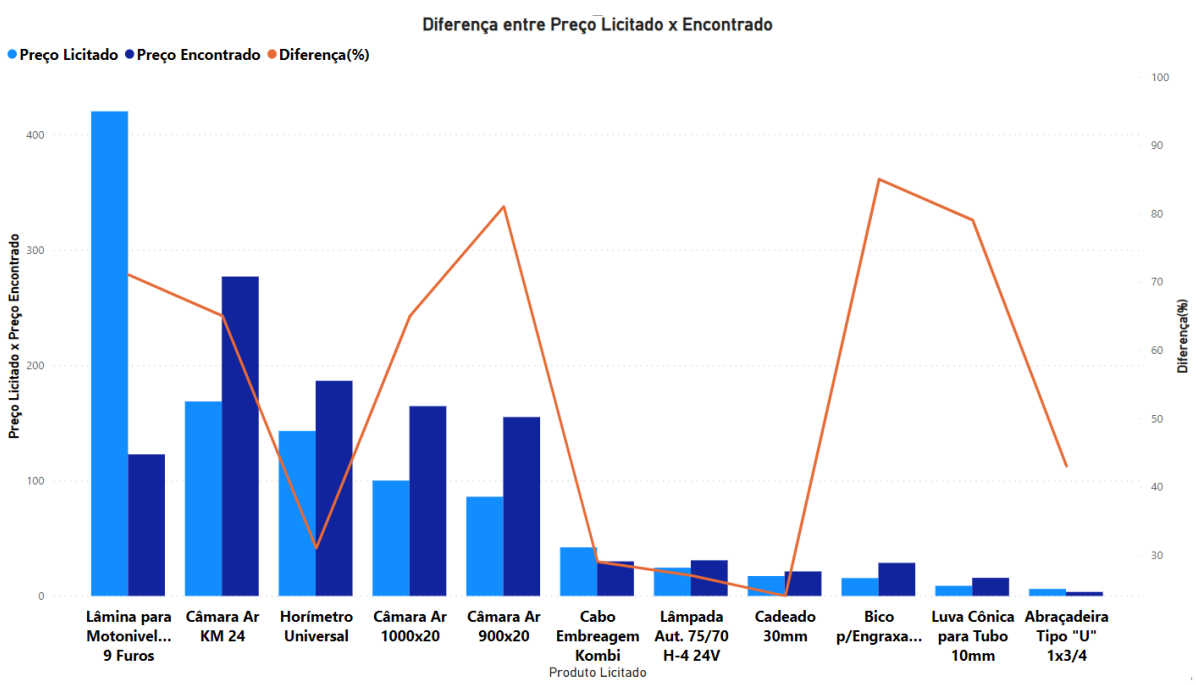
Logo, valores que estão acima dos 30% aceitos, podem ser classificados como um possível caso de sobre preço, assim como quando estiverem abaixo deste mesmo critério de aceitação. Quando há um indício de anomalia, um alerta é colocado nesta licitação, mostrando que este fornecedor pode estar agindo de má fé.

Após analizar valores repassados pelo município com os valores encontrados com o *bot*, as informações são armazenadas na base de dados da AWS em uma tabela que contém as seguintes informações: Descrição do item, valor licitado, valor encontrado, diferença, percentual da diferença, resultado da análise e data que os dados foram comparados. A partir dessa tabela então é possível que o funcionário responsável pelas licitações consiga visualizar e a apontar as inconsistências presentes no valor licitado.

Após efetuar uma análise da base de dados, foi possível verificar, como ilustrado na Figura 10, que é possível verificar que por mais que a licitação esteja condizente, sem indícios de anomalia, ainda há uma diferença entre o fornecedor que possui uma loja física e o fornecedor que vende através da internet, fato que é compreensível, já que para se ter a loja física existe a necessidade de manter o estabelecimento.

Através do gráfico, mostrado na figura 11, é perceptível que há indícios de irregularidade. O item “Lâmina para Motoniveladora 9 furos”, por exemplo, apresenta uma diferença de 71% a mais no valor licitado, indicando que há indícios de sobrepreço. Ainda também podemos citar o item “Abraçadeira Tipo U 1x3/4” que apresenta uma diferença de 43% a mais no preço licitado em relação a preço encontrado, indicando uma possível anomalia no valor licitado.

Figura 11 – Gráfico de diferença entre preços.



Fonte: Autoria própria.

Por fim, como podemos observar no gráfico, ou ainda na tabela apresentada na figura 12, existem alguns produtos que foram licitados por um valor muito menor do que o encontrado na internet, como é o caso do item “Bico p/ Engraxadeira” que foi licitado por um valor 85% menor do que o encontrado. Uma possível explicação para esses casos é que ao concorrer em um processo de licitação, o fornecedor opta por abaixar excessivamente o preço com o intuito de vencer a licitação e receber os lucros devido à grande demanda de venda.

Figura 12 – Tabela de diferença entre preços.

Produto Licitado	Preço Licitado	Preço Encontrado	Diferença (%)	Resultado
Abraçadeira Tipo "U" 1x3/4	6,09	3,50	43	Sobrepreco
Bico p/Engraxadeira	15,50	28,65	85	Subpreco
Cabo Embreagem Kombi	42,12	29,99	29	Regular
Cadeado 30mm	17,26	21,27	24	Regular
Câmara Ar 1000x20	100,00	164,51	65	Subpreco
Câmara Ar 900x20	86,00	155,06	81	Subpreco
Câmara Ar KM 24	168,59	276,81	65	Subpreco
Horímetro Universal	143,00	186,45	31	Subpreco
Lâmina para Motoniveladora 9 Furos	420,00	122,72	71	Sobrepreco
Lâmpada Aut. 75/70 H-4 24V	24,50	30,90	27	Regular
Luva Cônica para Tubo 10mm	8,80	15,70	79	Subpreco

Fonte: Autoria própria.

4 CONCLUSÃO

Este trabalho apresentou uma abordagem de ciência e engenharia de dados para a identificação de possíveis casos de anomalias em licitações públicas municipais. Esta perspectiva, permite que sejam identificados possíveis atos ilícitos em etapas de compra de itens em licitações através de buscas automatizadas baseadas em preços de produtos licitados por um orgão municipal.

Esta aplicação efetua busca de preços de produtos licitados através de informações previamente cadastradas em uma base de dados desenvolvidas através de parcerias em pesquisas com os autores desse trabalho. Neste trabalho, utilizando como apoio essas informações cadastradas, são efetuadas buscas por produtos similares aos licitados e então, através do cálculo da mediana é possível localizar um preço médio sobre todos os itens localizados. Após concluir tal busca, uma etapa de comparação de valores licitados e encontrados é iniciada, mostrando ao usuário quais são os pontos onde podem haver anomalias no processo licitatório para então disponibilizar estes resultados em uma base de dados.

Com os resultados encontrados foi possível identificar possíveis indícios de anomalia, de tal forma que um dos itens possuía um preço 79% mais alto que o encontrado pelo buscador automatizado enquanto que para o outro item constava um valor 43% acima do encontrado. Por outro lado também foi possível observar que o preço de outros produtos encontrados na internet eram muito mais baixos chegando a apresentar uma diferença 84% menor no preço licitado. Uma sugestão para tais casos é que para que possa vencer a licitação o fornecedor oferte os itens abaixo do preço de mercado, visando o lucro a partir da venda em grandes quantidades.

Por fim, como trabalho futuro, está a conexão desta aplicação diretamente com a base de dados implementada pelo *framework* presente no município, para que a extração e análise seja feita automaticamente e com maior frequência. Também como melhoria, está a o aperfeiçoamento do mecanismo de busca que atualmente trás como resultado itens similares, possibilitando que entrem na análise itens de marcas e preços inferiores aos licitados.

REFERÊNCIAS

AGGARWAL, C. C. An introduction to outlier analysis. *In*: **Outlier analysis**. [*S.l.*]: Springer, 2017. p. 1–34.

ASIMOV, I. **Eu, robô**. [*S.l.*]: Aleph, 2015.

BRASIL, C. G. da U. Licitações e contratações. 2021. Disponível em: https://www.portaltransparencia.gov.br/entenda-a-gestao-publica/licitacoes-e-contratacoes. Acesso em: 08 out. de 2021.

CARVALHO, I. M. de *et al.* Contribuições das tecnologias kdd e dw como ferramentas de gestão do conhecimento aplicadas ao processo de compras do governo eletrônico. **Revista Democracia Digital e Governo Eletrônico**, v. 1, n. 2, 2010.

DIOUF, R. *et al.* Web scraping: state-of-the-art and areas of application. *In*: IEEE. **2019 IEEE International Conference on Big Data (Big Data)**. [*S.l.*], 2019. p. 6040–6042.

FAYYAD, U.; PIATETSKY-SHAPIRO, G.; SMYTH, P. From data mining to knowledge discovery in databases. **AI magazine**, v. 17, n. 3, p. 37–37, 1996.

FERNANDES, L. S. *et al.* Mineração de indícios de cartéis em licitações federais utilizando regras de associação. **Seminários de Trabalho de Conclusão de Curso do Bacharelado em Sistemas de Informação**, v. 5, n. 1, 2021.

GOLDSCHMIDT, R.; PASSOS, E.; BEZERRA, E. **Data Mining**. [*S.l.*]: Elsevier Brasil, 2015.

HALMENSCHLAGER, C. Um algoritmo para indução de árvores e regras de decisão. 2002.

HAN, J.; PEI, J.; KAMBER, M. **Data mining: concepts and techniques**. [*S.l.*]: Elsevier, 2011.

JUNIOR, J. C. **Manual da licitação: Orientação Prática para o Processamento de Licitações, com Roteiros de Procedimento, Modelos de Carta-Convite e de Editais, de Atas de Sessões Públicas e de Relatórios de Julgamento de Propostas**. 2ª edição. ed. Grupo GEN, 2015. Disponível em: https://integrada.minhabiblioteca.com.br/#/books/9788522499823/. Acesso em: 28 out. 2021.

LAROSE, D. T.; LAROSE, C. D. **Discovering knowledge in data: an introduction to data mining**. [*S.l.*]: John Wiley & Sons, 2014. v. 4.

MCKINNEY, W. *et al.* pandas: a foundational python library for data analysis and statistics. **Python for high performance and scientific computing**, Seattle, v. 14, n. 9, p. 1–9, 2011.

MORAIS, C. M. M. d. Proposição de indicadores para investigação de licitações por meio de técnicas de reconhecimento de padrões estatísticos e mineração de dados. 2016.

PETROSKI *et al.* Uma abordagem de descoberta de conhecimento para suporte à gestão municipal de saúde. 2021.

RALHA, C. G.; SILVA, C. V. S. A multi-agent data mining system for cartel detection in brazilian government procurement. **Expert Systems with Applications**, Elsevier, v. 39, n. 14, p. 11642–11656, 2012. Acesso em: 09 out. de 2021.

RESNICK, P. *et al.* Grouplens: An open architecture for collaborative filtering of netnews. *In*: **Proceedings of the 1994 ACM conference on Computer supported cooperative work**. [*S.l.*: *s.n.*], 1994. p. 175–186.

ROSCA, D. *et al.* A decision making methodology in support of the business rules lifecycle. *In*: IEEE. **Proceedings of ISRE'97: 3rd IEEE International Symposium on Requirements Engineering**. [*S.l.*], 1997. p. 236–246.

SANTOS, B. S. dos *et al.* Data mining: Uma abordagem teórica e suas aplicações. **Revista ESPACIOS| Vol. 37 (Nº 05) Año 2016**, 2016.

SANTOS, R. *et al.* Conceitos de mineração de dados na web. **XV Simpósio Brasileiro de Sistemas de Multimídia e Web, VI Simpósio Brasileiro de Sistemas Colaborativos-Anais**, p. 81–124, 2009.

SHAH, H. S. *et al.* Mining ebay: Bidding strategies and shill detection. *In*: SPRINGER. **International Workshop on Mining Web Data for Discovering Usage Patterns and Profiles**. [*S.l.*], 2002. p. 17–34.

SIRISURIYA, D. S. *et al.* A comparative study on web scraping. 2015.

SOUTO, H. M. *et al.* Mineração de dados abertos: Uma análise do uso de bots em pregões eletrônicos. Universidade Federal da Paraíba, 2019.

SPARK, A. Apache spark. **Retrieved January**, v. 17, n. 1, p. 2018, 2018.

TEIXEIRA, A. H. d. S. Fraudes em licitações públicas. 2016. Acesso em: 08 out. de 2021.

TRINDADE, C. R.; CHAVES, P. S.; TEIXEIRA, M. Deteccao de anomalias em sistemas de administracao de frotas publicas municipais. *In*: SBC. **Anais do XVII Escola Regional de Banco de Dados**. [*S.l.*], 2022. p. 91–100.

ZHAO, B. Web scraping. **Encyclopedia of big data**, Springer Living ed. Cham, p. 1–3, 2017.