Previsão dos ganhadores e perdedores e do valor de licitações Municipais e do Estado da Paraíba:

Utilizando Aprendizagem de Máquina

Luiz Henrique Rodrigues de Oliveira

CENTRO DE INFORMÁTICA
UNIVERSIDADE FEDERAL DA PARAÍBA

João Pessoa, 2018

Luiz Henrique Rodrigues de Oliveira

Previsão dos ganhadores e perdedores e do valor de licitações Municipais e do Estado da Paraíba

Monografia apresentada ao curso Engenharia de Computação
do Centro de Informática, da Universidade Federal da Paraíba,
como requisito para a obtenção do grau de Bacharel em Engenharia de Computação

Orientadora: Thaís Gaudencio do Rêgo

Dezembro de 2018

O48p Oliveira, Luiz Henrique Rodrigues de.
Previsão dos ganhadores e perdedores e do valor de licitações Municipais e do Estado da Paraíba: Utilizando Aprendizagem de Máquina / Luiz Henrique Rodrigues de Oliveira. - João Pessoa, 2018.
73 f. : il.

Orientação: Thais Gaudencio do Rêgo Rêgo.
TCC (Especialização) - UFPB/CI.

1. Aprendizado de Máquina. 2. Licitações. 3. Análise Exploratória. I. Rêgo, Thais Gaudencio do Rêgo. II. Título.

UFPB/BC

CENTRO DE INFORMÁTICA
UNIVERSIDADE FEDERAL DA PARAÍBA

Trabalho de Conclusão de Curso de Engenharia de Computação intitulado ***Previsão de Licitações Municipais e Estaduais do Estado da Paraíba*** de autoria de Luiz Henrique Rodrigues de Oliveira, aprovada pela banca examinadora constituída pelos seguintes professores:

Thaís Gaudencio do Rêgo

Prof. Dr. Thaís Gaudencio do Rêgo
Universidade Federal da Paraíba

Yuri de Almeida Malheiros Barbosa

Prof. Dr. Yuri de Almeida Malheiros Brabosa
Universidade Federal da Paraíba

Prof. Dr. Josedilton Alves Diniz
Universidade Federal da Paraíba

João Pessoa, 6 de novembro de 2018

Centro de Informática, Universidade Federal da Paraíba
Rua dos Escoteiros, Mangabeira VII, João Pessoa, Paraíba, Brasil CEP: 58058-600
Fone: +55 (83) 3216 7093 / Fax: +55 (83) 3216 7117

Dedico esse trabalho a minha mãe, Marlene Rodrigues de Oliveira, por ter me criado e me guiado nesse caminho glorioso, se tenho algo é tudo por causa dela.

AGRADECIMENTOS

Agradeço a minha noiva, Welliny, pelo apoio e suporte emocional durante o decorrer dessa graduação e desse projeto de conclusão

A minha orientadora, Thaís Gaudêncio, por me apresentar esse área maravilhosa de Inteligência artificial e por também por tratar com tanto carinho todos os seus alunos.

Ao setor de Gestão de Informação do Tribunal de Contas do Estado da Paraíba, pela permissão na utilização dos dados nesse projeto, além do conhecimento jurídico passado no decorrer desse projeto

Aos amigos do Laboratório de Medidas e Instrumentação (LMI) pelo companheirismo durante toda essa jornada acadêmica.

RESUMO

Esse documento apresenta formas de aplicar modelos de Aprendizado de Máquina e técnicas de Ciência de dados dentro da área pública de gastos municipais e estaduais, sendo mais específico na área de licitações. Devido a subjetividade de se definir valores para as licitações e como se escolhem vencedores e perdedores para as mesmas, foi necessário definir um método mais imparcial e com diretrizes lógicas para a seleção desses parâmetros. Para poder atingir tal objetivo, foram utilizados algoritmos de regressão e classificação, como os algoritmos de Máquina de vetor de suporte (*Support Vector Machine - SVM*), *K* Vizinhos mais próximos (*K Nearest Neighbors*) para classificação e Regressão Linear e Regressão de Vetor de Suporte (*Support Vector Regression - SVR*) para regressão. Esse projeto mostra que com os algoritmos acima citados foi possível atingir taxas de acertos entre 60% e 80%.

Palavras-chave: <Aprendizado de máquina>, <Ciência de Dados>, <Licitações>, <Despesas públicas>

ABSTRACT

This document presents ways of applying Machine Learning models and Data Science techniques within the public area of municipal and state spending, being more specific in the area of bids. Due to the subjectivity of defining values for the bids and choosing winners and losers for them, it was necessary to define a more impartial method and with logical guidelines for the selection of these parameters. In order to achieve this goal, we used regression and classification algorithms, such as the Support Vector Machine algorithms, K Closest Neighbors for Classification and Linear Regression and Support Vector Regression for Regression. This project shows that with the algorithms mentioned above it was possible to achieve success rates between 60 % and 80 %.

Key-words: <Machine Learning>, <Data Science>, <biddings>, <Public Expenditures>

LISTA DE FIGURAS

LISTA DE TABELAS

LISTA DE ABREVIATURAS

CAPES – Coordenação de Aperfeiçoamento de Pessoal de Nível Superior

CNPq – Conselho Nacional de Desenvolvimento Científico e Tecnológico

DW - Armazém de dados (*Data Warehouse*)

ETL – Extração, Transformação e Carga (*Extract, Transform, Load*)

FINEP – Financiadora de Estudos e Projetos

IDEB – Índice de Desenvolvimento da Educação Básica,

IDH – Índice de Desenvolvimento Humano

KNN – K vizinhos mais próximos (K *nearest neighbors*)

OSC – Organizações da Sociedade Civil

PIB – Produto Interno bruto

RDC – Regime Diferenciado de Contratações

SVM – Máquina de vetor de suporte (*Support Vector Machine*)

SVR – Regressor de vetor de suporte (*Support Vector Regressor*)

TCE – Tribunal de contas do Estado

Sumário

1 INTRODUÇÃO

Segundo Tesouro Nacional (2017), o processo de licitação trata de um procedimento para a compra de produtos, contratação de serviços ou alienação de bens com as melhores condições para a máquina pública, sempre observando os princípios essenciais pertencentes à mesma, sendo eles: legalidade, impessoalidade, moralidade, igualdade, publicidade, probidade administrativa, vinculação ao instrumento convocatório, do julgamento objetivo e outros que lhe são correlatos.

Além disso, Oliveira (2014, p.29) enfatiza : "Em relação aos princípios específicos, podem ser destacados os princípios da competitividade, da isonomia, da vinculação ao instrumento convocatório, do procedimento formal e do julgamento objetivo."

Todos os elementos governamentais são obrigados a licitar, dando chance igual a todos que detenham dos recursos requisitados a ofertarem, mas cabe ao Estado definir qual é a proposta mais vantajosa seguindo preço e/ou qualidade. Segundo a Lei 8.666/1993[3], o processo de licitação é dividido nas seguintes partes: Abertura, Habilitação, Classificação, Homologação e Adjudicação

- Abertura: Onde há a proposição interna no órgão emissor e, posteriormente, é revelado a público e os mesmos podem se candidatar ao processo;
- Habilitação: Trata-se da verificação da veracidade e adequação das informações providas pelos candidatos no ato de inscrição;
- Classificação: Nessa etapa há o julgamento das propostas, eliminando as inadequadas segundo edital, e após as desclassificações é feita a análise da proposta em que há melhor benefício para o órgão licitante, seja por preço, ou por qualidade, ou por ambos;
- Homologação e Adjudicação: Dada a escolha do(s) vencedor(es) realiza-se um requerimento ao órgão especializado em gerar documentação formal do ocorrido que atrela vínculo entre o licitante e o(s) ofertadores. Caso haja alguma irregularidade, o processo não será homologado e o processo de licitação é nulificado.

Nota-se que o ato de licitar não é simples, e que custa caro para os cofres públicos. Segundo Silva et al. (2015) o custo do processo licitatório pode alcançar, em média, o valor de 4.912,75 (quatro mil, novecentos e doze reais e setenta e cinco centavos), isso levando em consideração que há casos em que nenhum licitante consegue atingir os requisitos mínimos propostos no edital, ou não há licitantes no processo, esses eventos são chamados de Licitação Fracassada e Licitação Deserta, respectivamente.

Enquanto o ato de licitar vem desde a época medieval, uma nova metodologia computacional avança e já possui ampla aplicação em diversas áreas mundialmente, esse recurso é chamado de aprendizado de máquina, ou originalmente, *Machine Learning* . Segundo Arthur Samuel (1959), Aprendizado de máquina é área de estudo que fornece aos computadores a habilidade de aprender sem ser explicitamente programado para tal. Para a realização desse feito, é necessário uma base com as informações sobre o problema previamente tratados para não conter nenhum ruído ou valores fora do escopo do problema e do algoritmo usado.

IFI Claims Patent Services (2018) destaca que o ramo de Aprendizado de Máquina cresceu 34% entre os anos de 2013 e 2015, sendo o terceiro maior crescimento de uma categoria na área de patentes americana. Conjuntamente, Outlook (2018) indica que 61% das empresas apontam a inclusão de Aprendizado de Máquina em suas áreas como um dos maiores feitos em relação a dados do ano, seguido da área de grande volume de dados (*Big Data*) e Análise corporativa (*Business Analytics*) com 58%, no qual apresentam termos semelhantes com a área de Aprendizado de Máquina.

Desse modo, com tamanho crescimento, eficácia e absorção do mercado privado, uma aplicação no setor de gestão pública teria grande viabilidade e utilidade para garantir a execução dos princípios estabelecidos na constituição nacional.

1.1 Definição do Problema

Devido ao comportamento humano, as premissas citadas anteriormente geralmente não são seguidas, pois o ser humano dificilmente é impessoal e objetivo em seus julgamentos. Dia após dia, notícias relacionadas a fraudes licitatórias são destaques em todo o país, levando a crer que os avaliadores estão suscetíveis a interferência por meio dos licitantes, seja ela pela concessão de vantagem, como propinas ou favorecimento externo, ou por ameaças e chantagem ao servidor.

Outro fator importante é o de que não é claro como o valor final de cada licitação é obtido, podendo o valor final se tratar de apenas o menor valor proposto pelo licitante, desde que esse esteja dentro da margem declarada em edital anteriormente, mas em outros casos, como resultado de uma negociação, sendo o valor final reduzido ou até aumentado. Isso torna a definição de valor homologado muito mais subjetiva, dependendo do responsável pela avaliação das propostas dos licitantes por determinar esse limiar entre proposta e valor final.

1.2 Premissas e Hipóteses

A impessoalização do processo de análise licitatória promoverá uma análise mais objetiva, além de garantir as premissas constitucionais de isonomia, impessoalidade e igualdade.

Além disso, a objetividade em gerar um valor final, tende a facilitar a universalização do processo, podendo então haver a integração entre diferentes órgãos públicos.

1.2.1 Objetivo geral

Facilitar a seleção de vencedores e atribuição de valores finais no processo de licitação, além disso, proporcionar uma classificação mais impessoal e com clara aplicação das premissas teóricas propostas na constituição federal.

1.2.2 Objetivos específicos

- Conhecer a estrutura dos dados das licitações no Estado da Paraíba e suas propriedades utilizando análise exploratória

- Utilizar elementos de Aprendizado de Máquina, como algoritmos de classificação lineares para realizar categorizações de vencedores;

- Utilizar métodos de regressão para definir os valores finais de cada licitação.

1.3 Estrutura da monografia

Nos Conceitos Gerais serão apresentados os termos jurídicos sobre empenho e licitação, informando como as licitações são estruturadas e classificadas além de uma breve introdução sobre o assunto de Aprendizado de máquina e Ciência de dados.

Na Metodologia serão explicados como funcionam as etapas de pré-processamento das informações e explicação de algum modelos para a previsão de resultados de acordo com um determinado conjunto de dados, assim como validar os resultados obtidos

Nos Resultados serão aplicados os modelo de aprendizado de máquina a fim de prever vencedores e perdedores de uma licitação, assim como definir o seu valor final homologado.

Na Conclusão será definido se os resultados apresentados apresentam nível de acerto satisfatórios e se os modelos podem ser aplicados no dia-a-dia do processo licitatório.

2 CONCEITOS GERAIS E REVISÃO DA LITERATURA

2.1 Estrutura de uma despesa

A Lei nº 4.320/1964 [7], também chamada de lei das finanças públicas, indica que um processo licitatório é dividido em três estágios, sendo eles: empenho, liquidação e pagamento.

"O empenho é o registro da despesa, o qual resulta na nota de empenho, devendo ser sempre prévio em relação a despesa, não podendo exceder o saldo da dotação" (CHIMINAZZO, 2015). Isso implica que o estado sempre deve ter contas positivas antes mesmo de declarar o início da licitação, o que impede rombos econômicos devido a criação desordenada de empenhos.

Conforme o Artigo 63 da Lei nº 4.320/1964, a etapa de liquidação consiste na "verificação do direito adquirido pelo credor tendo por base os títulos e documentos comprobatórios do respectivo crédito" (BRASIL,1964). Onde as verificações são:

1. A origem e o objeto do que se deve pagar;
2. A importância exata a se pagar;
3. A quem se deve pagar a importância, para extinguir a obrigação.

A liquidação da despesa por fornecimentos feitos ou serviços prestados terá por base:

1. O contrato, ajuste ou acordo respectivo;
2. A nota de empenho;
3. Os comprovantes da entrega de material ou da prestação efetiva do serviço.

Já a etapa final, chamada de etapa de pagamento, como o nome sugere, ocorre o débito da obrigação inicialmente descrita na etapa de empenho, sendo esse pagamento realizado pela instituição pública para o credor, desde que as verificações na etapa de liquidação estejam todas completas.

2.2 Tipos de Objetos

Segundo MEIRELLES (1999, p. 250), "é a obra, o serviço, a compra, a alienação, a concessão, a permissão e a locação que, afinal, será contratada com o particular".

As categorias utilizadas nos municípios e estado da Paraíba são:

Obras e serviços de Engenharia: Relacionada a construções civis em geral, geralmente são as obras com maior requisição financeira por parte do jurisdicionado;

Bens e serviços: A compra de elementos para uso interno, como mobília, itens de escritório e serviços, como tarefas de limpeza, assistência técnica, alimentação, etc.;

Alienações: Itens que foram alienados, ou seja, sofreram uma troca de possessão e agora podem ser vendidos pelo jurisdicionado.

2.3 Tipos de Licitações

Diferentemente da definição de tipos de objetos, que tem foco em definir qual a categoria em que o elemento a ser licitado será encaixado, os tipos de licitações definem qual a métrica de seleção de uma proposta vencedora, que respectivamente, são:

- **Menor preço**: A proposta que obtiver o menor valor vencerá, sem nenhuma outra análise envolvida;
- **Melhor técnica** : Cada proposta será analisada por uma mesa com especialistas e cada licitante deve enviar um envelope com um detalhamento econômico e técnico do projeto;
  - O órgão responsável deve explicitar o valor máximo a ser pago, os critérios de avaliação e a nota mínima para a aceitação dos projetos;
- **Melhor técnica e menor preço**: Será realizada uma média ponderada entre a avaliação técnica e a avaliação de preços.

2.4 Modalidades de Licitação

Completando a identificação de uma licitação em conjunto com as classificações anteriores citadas, a seleção da modalidade define as regras de como o processo licitatório será executado. Essas modalidades diferem de acordo com o valor da despesa ou então o tipo de item que será obtido. As seções abaixo explicarão quais os requisitos fundamentais para a inserção de uma licitação em determinada modalidade.

2.4.1 Concorrência

A Lei 8666/93 apresenta uma definição bem geral para a modalidade de concorrência no Artigo 22º, § 1º: "Concorrência é a modalidade de licitação entre quaisquer

interessados que, na fase inicial de habilitação preliminar, comprovem possuir os requisitos mínimos de qualificação exigidos no edital para execução de seu objeto" (BRASIL, 1993).

Na prática, essa licitação tem a finalidade de selecionar contratos com valores altos, sendo esses valores maiores de R$ 1.500.000,00 (Um milhão e quinhentos mil reais) para objetos de obras e serviços de engenharia e R$ 650.000,00 (Seiscentos e cinquenta mil reais) para objetos de compras e serviços.

Outro fator importante é que a licitação deve ser publicada em Diário Oficial, com período de 45 dias para avaliações com foco no melhor projeto e melhor preço, e 30 dias para avaliações com foco apenas no menor preço, isso entre a data de publicação e a apresentação das propostas. Uma comissão avaliadora requere ao licitante dois envelopes contendo os projetos que visam ser escolhidos e após a emissão, avalia-se os projetos recebidos, seus cronogramas e regimentos orçamentários e seleciona-se o melhor para o jurisdicionado.

2.4.2 Tomada de Preços

Segundo o Artigo 22º, § 2º da Lei 8666/93, "Tomada de preços é a modalidade de licitação entre interessados devidamente cadastrados ou que atenderem a todas as condições exigidas para cadastro até o terceiro dia anterior à data do recebimento das propostas, observada a necessária qualificação" (BRASIL, 1993).

Os valores aplicados serão até o valor piso da modalidade de concorrência, mas não inferiores ao valor teto da modalidade de convite.

2.4.3 Convite

Segundo o Artigo 22º, § 3º da Lei 8666/93, "Convite é a modalidade de licitação entre interessados do ramo pertinente ao seu objeto, cadastrados ou não, escolhidos e convidados em número mínimo de 3 (três) pela unidade administrativa, a qual afixará, em local apropriado, cópia do instrumento convocatório e o estenderá aos demais cadastrados na correspondente especialidade que manifestarem seu interesse com antecedência de até 24 (vinte e quatro) horas da apresentação das propostas" (BRASIL, 1993).

Essa modalidade contém características mais simples, pois é focada em licitações de valores mais baixos, sendo eles valores menores do que R$ 150.000,00 (Cento e cinquenta mil reais) para objetos de obras e serviços de engenharia e R$ 80.000,00 (Oitenta mil reais) para objetos de compras e serviços. Outras propriedades do convite são:

- Falta de necessidade de divulgação em diário oficial;

- O órgão público convida os licitantes, mas permite a adesão de não convidados;
- Mínimo de três participantes;
- Tempo de formulação de propostas de cinco dias.

2.4.4 Concurso

Segundo o Artigo 22º, § 4º da Lei 8666/93, "Concurso é a modalidade de licitação entre quaisquer interessados para escolha de trabalho técnico, científico ou artístico, mediante a instituição de prêmios ou remuneração aos vencedores, conforme critérios constantes de edital publicado na imprensa oficial com antecedência mínima de 45 (quarenta e cinco) dias." (BRASIL, 1993)[3].

Essa modalidade visa escolher serviços, onde o licitante apresenta o projeto finalizado, com conexões entre as datas de publicação do edital e da apresentação dos projetos. Caso aprovado, o licitante não se vincula empregaticiamente, mas recebe uma premiação, que comummente se trata de remuneração econômica, mas pode vir a ser um bem ou então até honrarias provenientes do órgão responsável pela licitação.

2.4.5 Leilão

Segundo o Artigo 22º, § 5º da Lei 8666/93, "Leilão é a modalidade de licitação entre quaisquer interessados para a venda de bens móveis inservíveis para a administração ou de produtos legalmente apreendidos ou penhorados, ou para a alienação de bens imóveis prevista no Art. 19, a quem oferecer o maior lance, igual ou superior ao valor da avaliação" (BRASIL, 1993).

Igual ao funcionamento de qualquer leilão, o maior lance vence a licitação, desde que ela atinja o valor mínimo definido por um avaliador. As demais regras de participação devem ser definidas de acordo com cada edital gerado.

2.4.6 Pregão

Segundo o Artigo 2º do Decreto 3555/00, "Pregão é a modalidade de licitação em que a disputa pelo fornecimento de bens ou serviços comuns é feita em sessão pública, por meio de propostas de preços escritas e lances verbais" (BRASIL, 2000).

Essa modalidade atualmente possui duas divisões que funcionam de forma similar, sendo elas o pregão presencial e o pregão eletrônico, sendo das modalidades as mais comuns e mais céleres, onde os licitantes se apresentam, de forma física ou virtual, e apresentam

suas propostas verbalmente. A proposta vencedora é aquela que apresenta o menor valor para a aquisição do bem ou serviço.

Embora seja a mais rápida de elaboração, o pregão apresenta algumas restrições e regulamentações:

- Apenas bens comuns podem ser aplicados a modalidade;
- Os selecionados à fase de propostas de lances serão aqueles cujos lances iniciais estejam até 10% acima do menor valor;
  - Caso não haja ao menos três licitantes aptos na regra acima, os três melhores lances serão selecionados para a fase de lances;
- Mesmo que seja o melhor lance, a proposta ainda deve ser aceita pelo pregoeiro e o licitante deve estar de acordo com a documentação exigida pelo órgão responsável pela licitação.

A Tabela 1 exibe um resumo para a adequação de cada modalidade de acordo com seus valores e/ou propriedades exclusivas:

Tipo	Obras e serviços de Engenharia	Compras e serviços
Concorrência	Acima de R$ 1.500.000	Acima de R$ 650.000
Tomada de preços	Até R$ 1.500.000	Até R$ 650.000
Convite	Até R$ 150.000	Até R$ 80.000
Leilão	Para produtos móveis apreendidos ou empenhados; Para imóveis obtidos de forma judicial	
Concurso	Para escolha de trabalho técnico, científico e/ou artístico	
Pregão	Compra e locação de serviços e bens comuns	

Tabela 1: Limitações de valores e itens para cada modalidade[14].

2.5 Casos Especiais

Em determinadas situações,as categorias anteriores podem ser inviáveis e o processo licitatório pode passar por exceções ou modificações em suas regras, e cada uma dessas regras abaixo explicitadas contém seus parâmetros para poderem ser aplicadas em alguma licitação.

2.5.1 Inexigibilidade

A inexigibilidade, como o nome diz, remove a exigência de todo um processo licitatório, dadas essas circunstâncias (NORMAS LEGAIS, 2018) :

- Para aquisição de materiais, equipamentos, ou gêneros que só possam ser fornecidos por produtor, empresa ou representante comercial exclusivo, vedada a preferência de marca, devendo a comprovação de exclusividade ser feita através de atestado fornecido pelo órgão de registro do comércio do local em que se realizaria a licitação ou a obra ou o serviço, pelo Sindicato, Federação ou Confederação Patronal, ou, ainda, pelas entidades equivalentes;

- Para a contratação de serviços técnicos, de natureza singular, com profissionais ou empresas de notória especialização, vedada a inexigibilidade para serviços de publicidade e divulgação;

- Para contratação de profissional de qualquer setor artístico, diretamente ou através de empresário exclusivo, desde que consagrado pela crítica especializada ou pela opinião pública.

Caso haja alguma licitação necessitada do processo de inexigibilidade, a mesma deve ser comunicada em até três dias à entidade superior com documentação comprobatória da necessidade, e em até cinco dias para a publicação da mesma em Diário Oficial.

2.5.2 Dispensada

De acordo com o Art 17º da Lei 8666/90, uma licitação pode ser caracterizada como dispensada caso haja uma alienação de bens pela administração pública e as regras para a aplicação diferem para bens móveis e imóveis:

Para bens imóveis têm-se as determinadas situações para uma licitação dispensada:

a) doação em pagamento;

b) doação, permitida exclusivamente para outro órgão ou entidade da administração pública, de qualquer esfera de governo, ressalvado a alienação gratuita ou onerosa, aforamento, concessão de direito real de uso, locação ou permissão de uso de bens:

    - imóveis residenciais construídos, destinados ou efetivamente utilizados no âmbito de programas habitacionais ou de regularização fundiária de interesse social desenvolvidos por órgãos ou entidades da administração pública;
    - imóveis de uso comercial de âmbito local com área de até 250 m^2 (duzentos e cinquenta metros quadrados) e inseridos no âmbito de programas de regularização fundiária de interesse social desenvolvidos por órgãos ou entidades da administração pública.

c) permuta por outro imóvel que atenda aos requisitos constantes do Inciso X do Art. 24 da Lei 8.666/93:

"Art. 24, X. Lei 8.666/93. (...) imóvel destinado ao atendimento das finalidades precípuas da administração, cujas necessidades de instalação e localização condicionem a sua escolha, desde que o preço seja compatível com o valor de mercado, segundo avaliação prévia.";

d) investidura;

e) venda a outro órgão ou entidade da administração pública, de qualquer esfera de governo;

f) alienação gratuita ou onerosa, aforamento, concessão de direito real de uso, locação ou permissão de uso de bens imóveis residenciais construídos,destinados ou efetivamente utilizados no âmbito de programas habitacionais ou de regularização fundiária de interesse social desenvolvidos por órgãos ou entidades da administração pública;

g) procedimentos de legitimação de posse de que trata o Art. 29 da Lei no 6.383, de 7 de dezembro de 1976, mediante iniciativa e deliberação dos órgãos da administração pública em cuja competência legal inclua-se tal atribuição;

h) alienação gratuita ou onerosa, aforamento, concessão de direito real de uso, locação ou permissão de uso de bens imóveis de uso comercial de âmbito local com área de até 250 m^2 (duzentos e cinquenta metros quadrados) e inseridos no âmbito de programas de regularização fundiária de interesse social desenvolvidos por órgãos ou entidades da administração pública.

Para bens móveis, os critérios são:

a) doação, permitida exclusivamente para fins e uso de interesse social, após avaliação de sua oportunidade e conveniência sócio-econômica, relativamente à escolha de outra forma de alienação;

b) permuta, permitida exclusivamente entre órgãos ou entidades da administração pública;

c) venda de ações, que poderão ser negociadas em bolsa, observada a legislação específica;

d) venda de títulos, na forma da legislação pertinente;

e) venda de bens produzidos ou comercializados por órgãos ou entidades da administração pública, em virtude de suas finalidades;

f) venda de materiais e equipamentos para outros órgãos ou entidades da administração pública, sem utilização previsível por quem deles dispõe.

2.5.3 Dispensa

A dispensa apresenta termos menos interligados entre órgãos governamentais, ou seja, maiores regras que dispensam licitações quando o relacionamento é entre iniciativa privada e estado.

Segundo a Lei 8666/90 (BRASIL,1990) são passíveis de dispensa de licitação:

- Obras e serviços de engenharia de valor até 10% do limite previsto na alínea "a", I do Art. 23 da Lei 8.666/93;
  - Não sendo parcelas de uma mesma obra ou serviço ou ainda de obras e serviços da mesma natureza e no mesmo local que possam ser realizados conjunta ou concomitantemente.

  "Art. 23, caput". As modalidades de licitação a que se referem os incisos I a III do artigo anterior serão determinadas em função dos seguintes limites, tendo em vista o valor estimado da contratação: I - para obras e serviços de engenharia: a) convite: até R$150.000,00 (cento e cinquenta mil reais);

- Outros serviços e compras de valor até 10% do limite previsto na alínea "a", II do Art. 23 da Lei 8.666/93
  - Não se referindo a parcelas de um mesmo serviço, compra ou alienação de maior vulto que possa ser realizado de uma só vez.

  Art. 23, caput. As modalidades de licitação a que se referem os incisos I a III do artigo anterior serão determinadas em função dos seguintes limites, tendo em vista o valor estimado da contratação: II - para compras e serviços não referidos no inciso anterior: a) convite - até R$ 80.000,00 (oitenta mil reais);

- Alienações previstas no Art. 17, I e II da Lei 8666/93;
  - Desde que não se refiram, também, a parcelas de um mesmo serviço, compra ou alienação de maior vulto que possa ser realizada de uma só vez.

- Nos casos de guerra ou grave perturbação da ordem;

- Contratações em casos de emergência ou de calamidade pública:
  - Para que haja a dispensa, a situação de urgência deve estar claramente configurada.

- Contratações em que a licitação anterior foi deserta:
  - Esta é uma hipótese em que não houve interessados na licitação anterior e esta não puder ser repetida sem prejuízo para o Poder Público.
- Contratações para normalização do abastecimento:
  - Caso em que há intervenção da União no domínio econômico para regular preços e normalizar o abastecimento.
- Contratações em que na licitação anterior os preços estavam acima dos praticados no mercado:
  - Evita os cartéis entre as empresas para super faturar o processo licitatório.
- Contratações cujos objetos são fornecidos por pessoa jurídica de direito público interno.
- Contratações efetuadas para preservar a segurança nacional:
  - Casos de decreto presidencial e uso da defesa nacional brasileira as licitações para suprimentos das mesmas podem ser dispensáveis.
- Contratações remanescentes de contrato de obra, serviço ou fornecimento:
  - Caso a obra com o licitante anterior seja rescindida, sendo que o licitante contratado seja o segundo colocado na licitação anterior e que concorde com os termos da empresa a qual o contrato se reincidiu inicialmente.
- Aquisição de gêneros perecíveis:
  - Compras de hortifrutigranjeiros, pão e outros gêneros perecíveis, realizadas com base no preço do dia.
- Contratações de instituições de ensino ou de pesquisa:
  - Instituição brasileira incumbida na pesquisa, ensino ou desenvolvimento institucional;
  - Instituição dedicada à recuperação social de preso (mas esta instituição deve ter reputação e não deve ter fins lucrativos).
- Contratações em razão de acordo internacional:
  - Aprovado previamente pelo Congresso Nacional e com clara vantagem para a maquina estatal.

- Aquisição para restauração de obras de arte e objetos históricos;
- Contratações para impressão de diários oficiais;
- Aquisição de componentes ou peças durante o período de garantia técnica, desde que compradas na fornecedora oficial;
- Abastecimento de navios, embarcações, unidades aéreas e suprimento de tropas:
  - Durante paradas e estadas de curta duração devido a operação militar necessária.
- Aquisição de material de uso das Forças Armadas;
- Contratações de associações de deficientes físicos;
- Aquisição de bens destinados à pesquisa científica e tecnológica:
  - Compra de bens destinados exclusivamente à pesquisa científica e tecnológica com recursos da CAPES,Finep CNPq ou outros.
- Contratações de fornecimento ou suprimento de energia elétrica;
- Contratações das empresas públicas e sociedades de economia mista com suas subsidiárias e controladas;
- Contratações com organizações sociais para prestação de serviços.

2.5.4 Adesão à ata de registro de preços

O Decreto nº 7892 (BRASIL, 2013) indica que essa modalidade tem como fim, reduzir a burocracia e a necessidade de criação de novas licitações quando uma de mesmo objetivo já existe, seja de obtenção de bens ou contratação de serviço.

Havendo solicitação da adesão pela unidade gestora, chamada de órgão carona, notifica o órgão gerenciador, o que realizou todo o processo licitatório de forma normal, onde esse gerenciador informa sobre a vontade do órgão carona de ter a obra ou serviço realizado pelo mesmo licitante.

Ao final, com a aprovação do licitante, o órgão carona emite a nota empenho da mesma maneira que o órgão gerenciador, mantendo valores e maneiras de execução.

2.5.5 Regime Diferenciado de Contratações Públicas

Com finalidade de reduzir os prazos para conclusão de licitações, em 2011 foi criado o Regime Diferenciado de Contratações (RDC), que possui algumas ferramentas para a agilidade no processo, que entre elas estão o curto prazo de cinco dias para a habilitação dos proponentes, a compra do projeto e da execução da obra de forma conjunta, no qual normalmente é feito separadamente, ou com o projeto de responsabilidade do ente público.

Essa nova maneira de se comprar para instituições públicas foi aplicada devido a urgência para os eventos da copa do mundo em 2014 e, posteriormente as olimpíadas que foram sediadas no estado do Rio de Janeiro.

2.5.6 Chamada Pública

Procedimento semelhante ao de licitação, que seleciona as Organizações da Sociedade Civil (OSC), que não tem fins lucrativos, para realizar processos sociais relacionados a saúde, tecnologia, agronomia, entre outros temas. A OSC sofre um processo menos rígido em chamamento público do que em uma licitação, pois não havendo foco em lucro, e sim na execução, não há necessidade de algumas verificações[11].

Outro fato é que a OSC se torna um elemento da unidade gestora durante a realização do processo, ou seja, a OSC se torna um membro temporário do jurisdicionado.

2.6 Aprendizado de Máquina

Após o entendimento de empenhos, licitações e suas subdivisões, precisamos compreender como aplicar os algoritmos que tratam de prever e automatizar a definição de vencedores de uma licitação, assim como seu valor homologado. Esse procedimento é chamado de Aprendizado de máquina.

"Aprendizado de máquina é uma região da área de inteligência artifical, onde procura prover conhecimento aos computadores através de dados, observações e interação com o mundo. Esse conhecimento adquirido permite que computadores generalizem corretamente para novas situações" (BENGIO, 2018, Tradução nossa)

Aprendizado de máquina geralmente é dividido em duas seções: Aprendizado supervisionado, onde há os rótulos nos objetos e o algoritmo tem como objetivo prever a classe ou chegar o mais próximo possível do valor esperado e o Aprendizado não-supervisionado, onde os dados não possuem nem um rótulo e cabe ao modelo agrupar ou selecionar objetos que tenham termos em comum gerando assim agrupamentos (do inglês, *clusters)* de objetos, cabendo ao analista definir se aquele agrupamento é válido ou não.

Outro processo que pode ser incluso na etapa de Aprendizado de Máquina é o de seleção dos objetos de teste e treino, onde o treino é a parte dos dados onde os algoritmos explicados nas próximas seções irão realizar suas modificações para tentar encontrar um modelo que consiga prever corretamente os dados atuais e dados futuros desconhecidos pelo modelo. Já o teste, como o nome diz, será utilizado para verificar se o algoritmo não criou um modelo que apenas funcione para o treino e sim para qualquer informação futura que será adicionada. Esse problema de ajuste excessivo no treinamento é chamado de superajustamento (*Overfitting*). Normalmente, a taxa de seleção treinamento-teste, fica em torno de 70% dos dados para treinamento e 30% para testes, mas, caso necessário um maior conjunto de dados para treinamento, essa taxa pode ser alterada.

Modelos de aprendizado de máquina são criados através de dados, e através das licitações que foram explicadas nas seções anteriores, será possível criar um banco de dados de licitações. Com essas informações, o modelo irá aprender padrões e aplica-los para prever o resultado de uma licitação, gerando assim uma automatização do processo licitatório, que pode ser usado a auxiliar o servidor na hora de selecionar um licitante vencedor e/ou seu valor final baseado no histórico licitatório.

2.7 Pré-Processamento

Durante a obtenção das informações sobre as licitações, podem ocorrer de os dados serem armazenados com ruídos, formato incorreto ou campos vazios causando problemas em etapas futuras. Para corrigir esses problemas é necessário que todo o conjunto de dados seja tratado e organizado. O nome desse setor que estrutura a informação é chamado de pré-processamento dos dados.

"Processamento de dados é o processo de coletar, limpar e consolidar dados em um arquivo ou tabela, com funcionalidade primária de uso em análises."(DATAWATCH, 2018, tradução própria)[16]

Esse processo é crucial em qualquer conjunto de dados em que deseja realizar a análise e predição através de um modelo de Aprendizagem de Máquina. Segundo o redator do *New York Times*, Steve Lohr (2014, tradução própria)[17], cerca de 50 a 80% do tempo de processo de elaboração de um modelo ou visualização de Ciências de Dados é focado em corrigir e organizar os dados coletados.

Esse processo tem como objetivos:

- Selecionar as informações de maior importância;
- Estruturar a informação original para o tipo de dado mais conveniente para o analista de dados;

- Remover ou corrigir informações incosistentes;
- Remover informações nulas;
- Detectar valores fora dos padrões da informação (*outliers*);
- Normalizar valores.

Um dado sem essas etapas citadas acima dificilmente consegue obter uma avaliação precisa e muito menos com eficiência em qualquer algoritmo de Aprendizagem de Máquina, dificultando a criação de um modelo eficaz e gerando erros dependendo do método selecionado. Por isso, em cada algoritmo que será descrito nas próximas subseções, haverá uma pequena informação sobre os pré processamentos necessários.

2.8 Extrair, Transformar e Carregar(*Extract, Transform, Load* - ETL)

A seção anterior aborda a importância e utilidade de pré-processamento nos dados, mas antes de simplesmente trata-los, devemos saber as formas de adquirir a informação e armazenar os dados após a transformação. Esse método de extrair, unificar, processar e armazenar o resultado numa nova base de dados é chamado de ETL.

"ETL é um tipo de integração de dados que refere-se a três etapas: Extrair, Transformar e Carregar (*Extract, Transform, Load*) para misturar dados de múltiplas fontes. Durante esse processo, o dado é extraído de uma ou mais fontes de dados, transformado (pré-processado) para um formato suscetível para análise, e então carregado para outro sistema" (SAS, 2018). A Figura 1 ilustra o funcionamento desse processo e cada uma das três etapas:

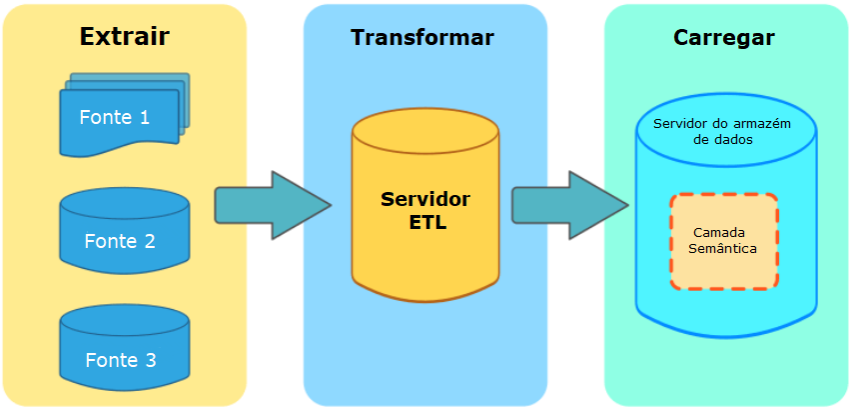

Figura 1: Processo de Extrair, Transformar e Carregar[19]. A extração (*Extract*) se trata da acquisição dos dados dos bancos Fonte 1 , Fonte 2 e Fonte 3, na etapa de transformação (*Transform*) o Servidor ETL (*ETL Server*) realiza o agrupamento e pré-processamento da informação, e para finalizar os dados tratados passarão para a etapa de carga (*Load*) onde serão salvos em um servidor de armazém de dados internamente organizados na camada semântica, que agrupa os novos bancos de dados de acordo com os temas e objetivos.

Esse processo é muito utilizado para situações que extraiam informações de múltiplas bases de dados, onde o ETL irá organizar e unificar todos os diferentes campos de diferentes bases para uma única tabela com o propósito de ser analisada. Normalmente o processo de ETL é acompanhado da estrutura de Armazém de Dados que será explicada na próxima subseção.

A etapa de extração trata de selecionar quais bases serão cruciais para a análise final, e dentro das seleções das bases de dados, serão selecionadas as tabelas que se encaixam no objetivo. Se necessário, também são selecionadas internamente na tabela quais colunas terão expressividade em uma análise futura. Normalmente, as tabelas são recuperadas via *scripts* de codificação SQL ou No-SQL, dependendo da arquitetura da base selecionada.

O processo de transformação tem como objetivo receber e/ou reunir as bases extraídas e adapta-las para um padrão comum desejado pelo analista ou definido pelo banco de dados final, o qual sofrerá a etapa de carga, passando por todos os processos acima citados no método de pré-processamento. Os tipos de pré-processamento poderão variar de acordo com a qualidade do banco original e com a necessidade do analista e do engenheiro de dados, inclusive com adição de processos para melhor acomodar a regra de negócio requisitada. Normalmente, essa etapa é realizada em linguagens de programação externas ao banco de dados, como Python e R.

A etapa final de carga tem como finalidade inserir as informações que foram transformadas e limpas para uma nova base de dados, a qual será utilizada pelo cientista

de dados para a etapa de análise exploratória, criação de visualizações e aplicação de algoritmos de Aprendizagem de Máquina.

2.9 Armazém de Dados (*Data Warehouse* - DW)

Segundo Date (2004) "Armazém de Dados é um depósito de dados orientado por assunto, integrado, não volátil, variável com o tempo, para apoiar as decisões gerenciais". Cada uma dessas orientações combinadas explicam a funcionalidade do Armazém de Dados.

2.9.1 Orientação por assunto

Cada DW, em conjunto com o processo de ETL, tem como objetivo agrupar informação de diversas bases de dados para gerar uma visualização final, mas como se define esse agrupamento? A orientação por assunto indica ao engenheiro de dados qual informação é selecionada ou descartada no processo de extração do ETL, utilizando como base o assunto necessitado e a regra de negócio por trás do mesmo.

2.9.2 Integração

Aproveitando a definição de orientação por assunto, a integração é justamente a forma como o assunto é criado, reunindo e padronizando as bases com a finalidade de gerar uma informação geral e completa sobre o assunto selecionado.

2.9.3 Não Volátil

A não volatilidade determina que os dados só devem ter a funcionalidade de consulta, não podendo ser alterados ou excluídos, já que essa etapa de alteração e/ou atualização acontecerá por meio do processo ETL em processo periódico, sendo redundante e propenso a erros a manipulação direta dos dados no DW.

2.9.4 Variável no tempo

A variação temporal indica que os processos serão periodicamente atualizados, não sendo em tempo real, ou seja, um novo processo de ETL para atualização do DW será realizado entre determinados períodos, seja ele diário, semanal, quinzenal, entre outros. Isso porque o processo pode ser lento, pois há sempre a exclusão e reinserção total da informação presente, o que pode levar ao prejudicar sistemas que utilizam o DW como base.

2.9.5 Elementos de um Armazém de Dados

Quando são definidos os elementos em um Armazém de Dados, ferramentas de ETL são cruciais no modelo de Armazém de Dados, pois sem o mesmo, não haveria integração nem padronização nos dados. A Figura 2 mostra que para termos uma estrutura de Armazém de Dados precisamos passar pela etapa de ETL para assim organizar os dados da forma que for desejável.

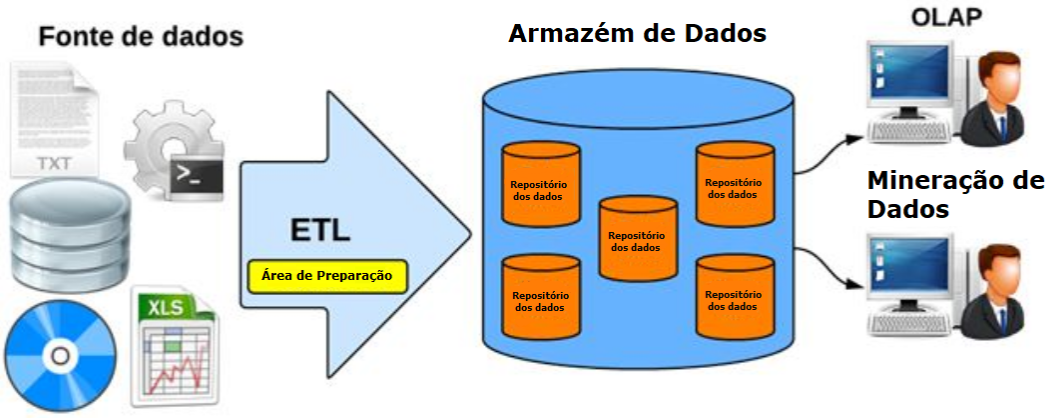

Figura 2: Estrutura de um Armazém de Dados [21], mostrando as Áreas de preparação, o Armazém de dados e seus repositórios de dados internos, prosseguindo para a bifurcação, onde estão as regiões de Mineração de dados e Processamento Analítico online (*Online Analytical Processing - OLAP*)

- Área de preparação (*Staging Area*)
  - Área intermediária situada internamente no processo ETL, corresponde ao dado durante os diversos processos de padronização;
- Repositório de dados (*Data Mart*)
  - Sub-elementos do Armazém de Dados, geralmente sendo temas mais internos do que o assunto principal e que unificados geram o Armazém de Dados, sendo que um Repositório de Dados pode estar presente em mais de um DW;
- Armazém de Dados (*Data Warehouse*)
  - Unificação dos Repositório de Dados com relevância para o assunto especificado, estrutura que estará disponível para os processos de Mineração de dados (*Data Mining*) e Processamento Analítico online (*Online Analytical Processing - OLAP*) ;
- Processamento Analítico online (*Online Analytical Processing - OLAP*)

  - Ferramenta que utilizará os dados para geração de visualizações e operações analíticas da informação gerada;

- Mineração de dados (*Data Mining*)
  - Elemento responsável por operações de correlação e descoberta de conhecimento no DW.

2.10 Análise Exploratória dos dados

Uma definição dada por Raniere Ramos (2015) é a de que a análise exploratória funciona como a etapa de aprendizado sobre como funciona o conjunto de dados, podendo assim investigar comportamentos médios e discrepantes, interdependência entre as variáveis, identificação de tendências e a identificação do que é essencial e do que se trata de ruído ou *outlier*

Então para poder entender como funciona melhor o processo licitatório no estado da Paraíba foram realizados agrupamentos e visualizações, demonstrando as características dos dados e algumas interpretações interessantes sobre o assunto.

2.11 Classificação Supervisionada

Classificação supervisionada é uma vertente na área de Aprendizagem de Máquina que visa o aprendizado de padrões através de rótulos. Cada exemplo deve possuir um atributo-saída, pois através deles na etapa de treinamento é que o algoritmo de classificação irá criar o modelo de previsão para novos dados de teste.

Relembrando a etapa de pré-processamento, é totalmente indicado que a característica dos dados na separação de treinamento sejam das mais diversas, ou seja, que a informação contenha praticamente os mesmos números de elementos para cada tipo de rótulo existente, com isso, o modelo fica mais preciso e imparcial para as classes. Caso haja muitos exemplos no treinamento de apenas uma classe, o modelo fica superasjustado, quando o algoritmo simplesmente atribui apenas uma classificação para todo o teste, ignorando todos os outros rótulos, pois foi aprendido apenas o padrão de uma das classes durante o treinamento.

Nas próximas subseções serão explicados os métodos de classificação utilizados para esse projeto, sendo o primeiro o algoritmo dos K Vizinhos mais próximos (*K Nearest neighbors* - KNN), seguido pelo método da Máquina de Vetores de Suporte (*Support Vector Machine* - SVM) e finalizando com o algoritmo *Naive Bayes*. Esses três algoritmos serão os responsáveis por prever quais são as classes vencedoras e perdedoras no banco de dados de licitação.

2.11.1 *K* Vizinhos mais próximos (*K Nearest neighbors* - KNN)

Um dos métodos de Classificação Supervisionada mais comuns é o dos k vizinhos mais próximos (*K Nearest Neighbors*) é um algoritmo de previsão, onde o rótulo a ser definido no objeto em processo de categorização depende dos seus k "vizinhos" na base de dados.

A Figura 3 ilustra o funcionamento do processo que pode ser descrito da seguinte maneira :

1. São pré-definidas quais serão as métricas de distância e qual será o valor de k;
2. O objeto a ser classificado, representado pela estrela vermelha, identificará quais são os k vizinhos mais próximos, através da semelhança entre as características dos pontos;
3. Uma contagem de classes/rótulos será realizada nos k vizinhos, onde os pontos amarelos representam a Classe A e os pontos roxos representam a Classe B;
4. A classe que obtiver o maior número de vizinhos será então atribuída como classe do objeto atual.

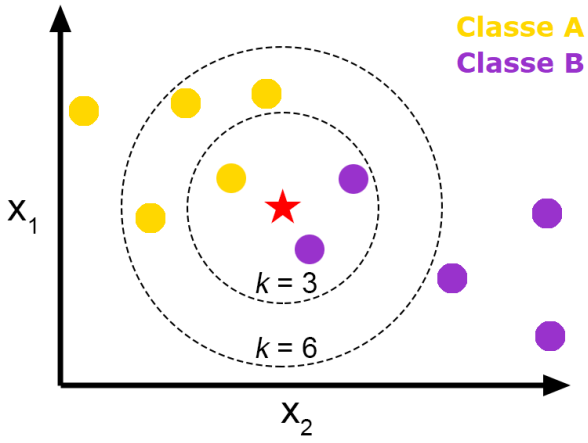

Figura 3: Funcionamento do *KNN*, onde X_1 e X_2 representam os atributos do conjunto de dados, os círculos tracejados indicam a área de atuação do algoritmo de acordo com o valor *k*, as cores amarela e roxa representam a Classe A e a Classe B [26] e a estrela indica o objeto a ser classificado pelo algoritmo

Dois elementos cruciais para o funcionamento do algoritmo são: o método para determinar a distância entre os pontos e quantos vizinhos são necessários para a atribuição da categoria ao objeto.

A métrica de distância mais comum é a distância euclidiana, que funciona da seguinte forma:

$$\sqrt{\sum_{i=1}^{n}(p_i - q_i)^2}$$

Onde p_i é o atributo i do objeto p e q_i é a mesmo atributo i só que do objeto q.

Já a medida *Manhattan*, ao invés de utilizar a diferença elevada ao quadrado, utiliza a soma das diferenças absolutas, que funciona da seguinte maneira:

$$\sum_{i=1}^{n}|p_i - q_i|$$

onde p_i e q_i tem o mesmo significado do já explicado na distância euclidiana

A determinação do K indica o quão abrangente será a inspeção de classes, mas essa atribuição vai depender do conjunto de dados em análise. Um ponto a se observar é que geralmente os valores atribuidos a k serão não-divisíveis pelo número de classes nos dados, afim de evitar o empate entre as classificações.

2.11.2 Máquina de Vetores de Suporte (*Support Vector Machine* - SVM)

O método Máquina de vetores de suporte (*Support Vector Machine* - SVM) consiste em um método de classificação que utiliza hiperplanos para separar os elementos do banco de dados de acordo com suas classes.

Segundo Sunil Ray (2017), "Cada ponto é posicionado em um espaço n-dimensional, onde n é o número de atributos do conjunto de dados, e cada atributo é um valor em determinada coordenada". Então, a classificação é realizada encontrando o hiperplano que divide o espaço, de forma que cada divisão tenha apenas elementos de uma classe, distinta das demais divisões.

O algoritmo funciona da seguinte maneira:

Primeiramente ocorre a identificação do hiperplano. Normalmente, o algoritmo gera alguns hiperplanos e testa a taxa de separação de cada um deles.

Na Figura 4 o hiperplano B é selecionado, pois ele separa melhor as classes do que os demais.

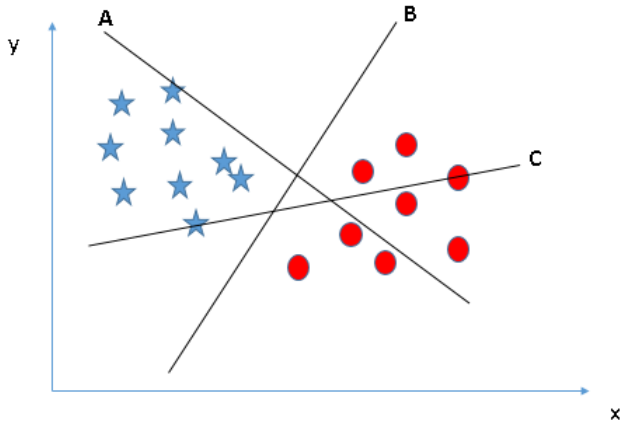

Figura 4: Hiperplanos A,B,C dividindo o conjunto de dados [27]

Na Figura 5, temos retas semelhantes, para definir qual o mais adaptado, é necessário estabelecer qual hiperplano tem a maior média de margem entre as classes, ou seja, qual a maior média de distância entre o hiperplano e o elemento mais próximo de cada classe.

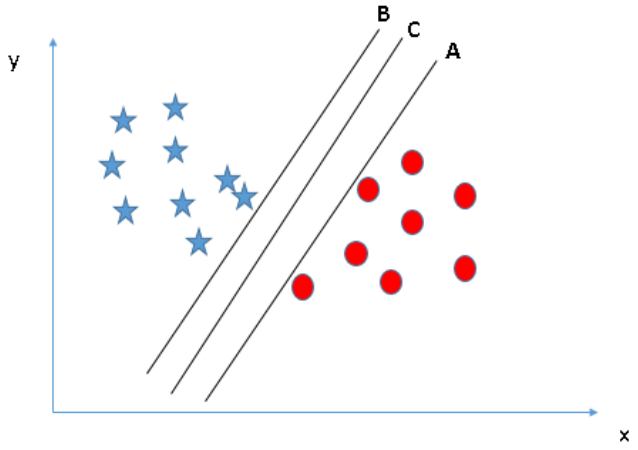

Figura 5: Hiperplanos A,B,C dividindo o plano de forma semelhante os dados [27], onde x e y são atributos dos objetos da classes estrela e círculo

A Figura 6 mostra que o hiperplano C é o melhor, pois possui uma margem média maior do que os hiperplanos A e B.

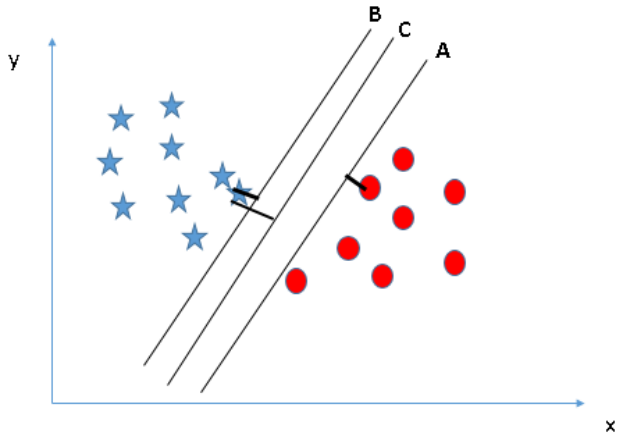

Figura 6: Seleção do melhor hiperplano (C) devido à maior margem em relação aos hiperplanos A e B para separar as classes estrela e círculo [27]

No caso da Figura 7, o objeto em questão que dificultaria a definição do hiperplano se trata de um *outlier*, no qual o algoritmo ignora, pois o maior objetivo é separar da melhor forma os pontos comuns dos dados, e os *outliers* são ignorados por serem pontos incomuns.

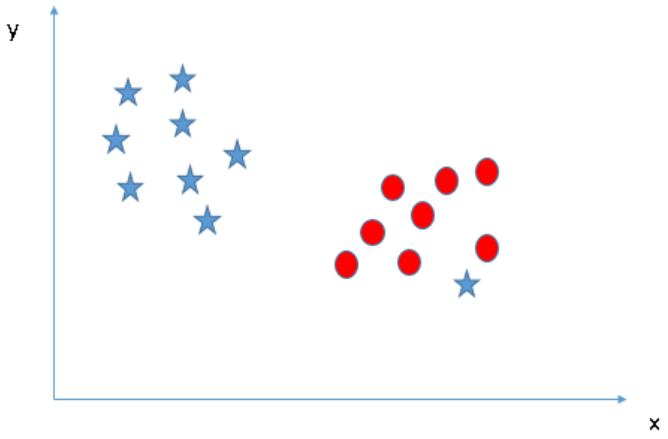

Figura 7: *Outlier* da classe estrela posicionado dentro da classe círculo [27]

Agora, o último caso trata de dados comuns de uma classe organizada no plano de uma forma que não pode ser separado por um hiperplano. A Figura 8 mostra um exemplo onde essa separação não é possível. Para resolver esse problema, é necessário realizar o *kernel trick*, que aplica uma transformação no espaço de forma que transforme as instâncias para que a separação seja possível.

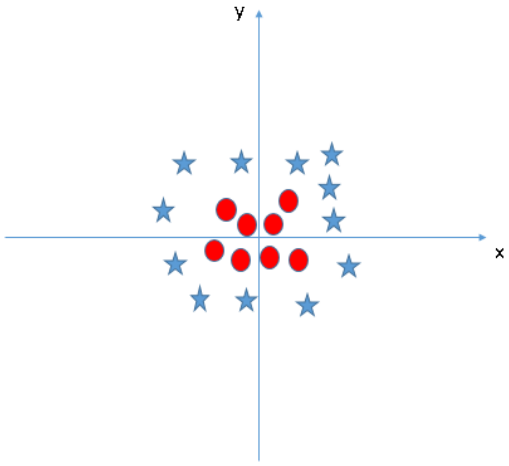

Figura 8: Conjunto de dados impossíveis de aplicar SVM de forma original, pois não é possível separar linearmente as classes estrela e círculo [27]

Nesse caso, de forma ilustrativa, pode-se observar de duas maneiras: ou os pontos se alteram de lugar como na Figura 9 ou então o hiperplano sofre o *kernel trick* mostrando o mesmo com formato diferente assim como o ilustrado na Figura 10.

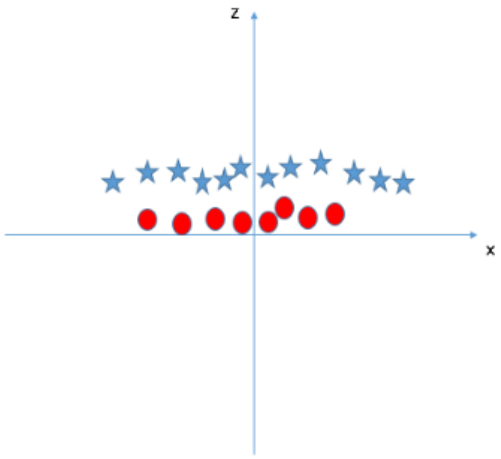

Figura 9: Alteração da posição dos pontos com o *kernel trick* [27]. Nota-se que utilizando o *kernel trick* os valores de x e y de cada objeto se alteraram, podendo assim agora ser linearmente separado

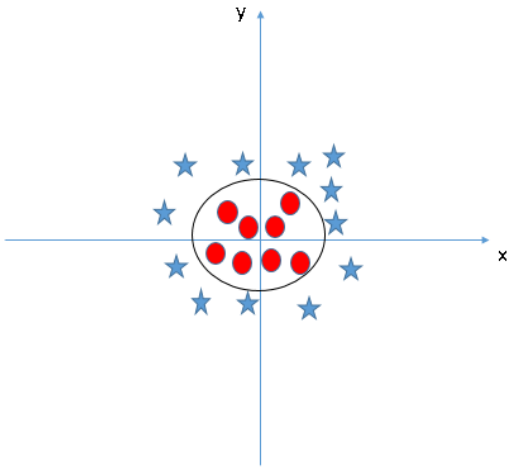

Figura 10: Alteração do desenho da trajetória com o *kernel trick* [27]. Uma outra forma de observar é que a trajetória de separação das classes se alterou de acordo com *kernel trick*

2.12 Regressão

Se a classificação tem como objetivo prever se um objeto está em um número limitado de classes, quando o conjunto de dados possuem valores ilimitados para seus lotos, como o conjunto dos números inteiros ou reais, o ideal é a aplicação dos algoritmos de regressão.

Segundo Eremenko (2018), a regressão tem como objetivo definir qual a relação entre a variável dependente, que será o valor que será previsto ou calculado, em relação as variáveis independentes, que são aquelas que definem a situação do problema. Dependendo do modelo, as variáveis tentarão se adaptar a uma reta, uma parábola ou outra função matemática mais complicada. A Figura 11 mostra a estrutura geral do processo de regressão.

De acordo com TutorVista (2018), a regressão possui a seguinte estrutura:

Equação de regressão: Essa fórmula se altera dependendo do tipo de regressão selecionado (Linear, Polinomial, etc.). Para exemplificar, usaremos a fórmula de regressão linear simples:

$$(y) = a + bx$$

Onde b é o *Slope* ou a inclinação da curva, que também é chamada de coeficiente de regressão, que é representado por:

$$b = \frac{N\sum XY - (\sum X)(\sum Y)}{N\sum X^2 - (\sum X)^2}$$

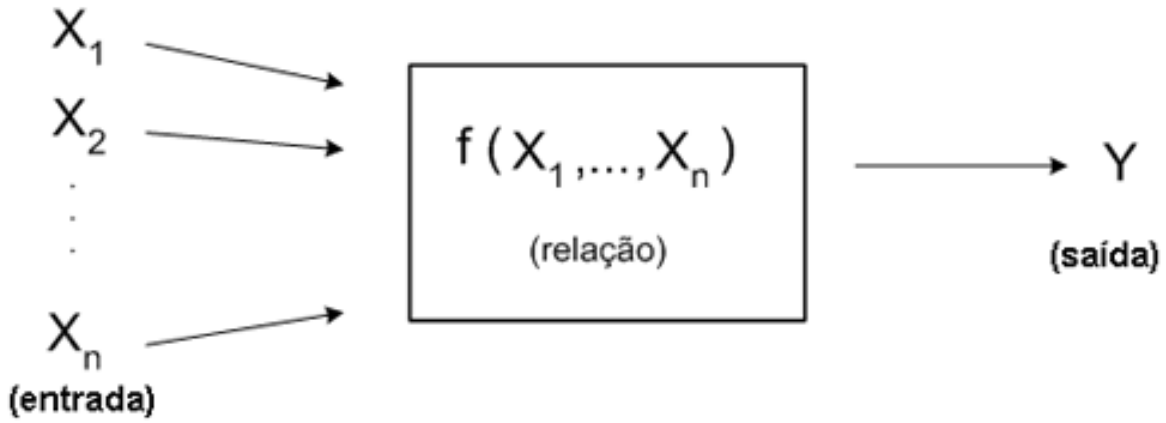

Figura 11: Representação do processo de regressão [29], onde X_1, X_2, ... , X_n são as entradas, $f(X_1, ..., X_n)$ indica a relação entre a entrada e a saída e Y representa o resultado final

Já o item a é o *intercept*, valor de x que está em cima do eixo y:

$$a = \frac{\sum Y - b(\sum X)}{N}$$

Onde N é a quantidade de valores ou elementos, X é a primeira previsão de valores de saída, Y a segunda previsão para a saída, x é a variável de entrada, ou variável independente, e y é a variável de saída, ou variável dependente.

A regressão é utilizada tanto para categorizar, como para determinar valores a variável de saída e, dependendo do objetivos, diversos métodos podem ser selecionados, onde os mais comuns são:

- Regressão Simples: Utilizada quando é necessário determinar o comportamento de uma variável de saída baseada em apenas uma variável de entrada.
  - Modelo da regressão:
    $$y = \beta_0 + \beta_1 x_1 + \varepsilon$$
    Onde β_0 é a curva, β_1 o *intercept* e ε é o erro da predição
- Regressão Múltipla: Quando necessita explicar a relação entre duas ou mais variáveis de entrada e uma variável de saída
  - Modelo da regressão:
    $$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_i x_i$$
    Onde β_0 é o fator independente e β_1, β_2,...,β_i, são coeficientes parciais para y

2.12.1 Regressão de vetor de suporte (*Support Vector Regression* - SVR)

O método SVR funciona de forma semelhante ao método SVM, mas por ser um método que exige a saída em forma de um valor real, e não uma categorização, alguns passos adicionais devem ser incluídos para a execução do mesmo.

Assim como no SVM, o SVR utiliza a criação de hiperplanos que visam prever os valores finais desses objetos, mas um dos problemas é que dificilmente os valores serão iguais aos valores previstos, pois dada a possibilidade de infinitos valores como resultado, o algoritmo tenta chegar a um valor cujo o erro seja aceitável. Outra observação trata de que o hiperplano não tem a funcionalidade de 'dividir' os objetos como no método SVM, e sim, encontrar uma curva na qual contorne os objetos, semelhante ao método de regressão.

A Figura 12 demonstra a melhor reta que evita o superajustamento, tendo eventualmente, alguns erros maiores em casos isolados.

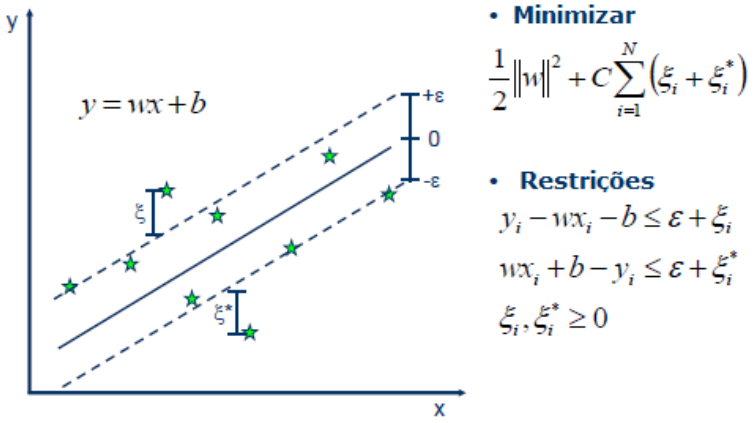

Figura 12: SVR Linear [29], com objetivo de minimizar a equação localizada no canto superior direito com as restrições localizadas no canto inferior direito. Onde as estrelas representam os objetos, w é o peso, x é a variável de entrada, b é o viés, ϵ é a margem de erro e ζ é a diferença entre a margem de erro e o objeto

Já na Figura 13 onde a imagem do lado esquerdo ilustra objetos representados por uma curva não linear, o SVR realiza uma transformação através de um *Kernel* que altera a dimensionalidade dos objetos para ser possível a utilização do mesmo modelo SVR de regressão.

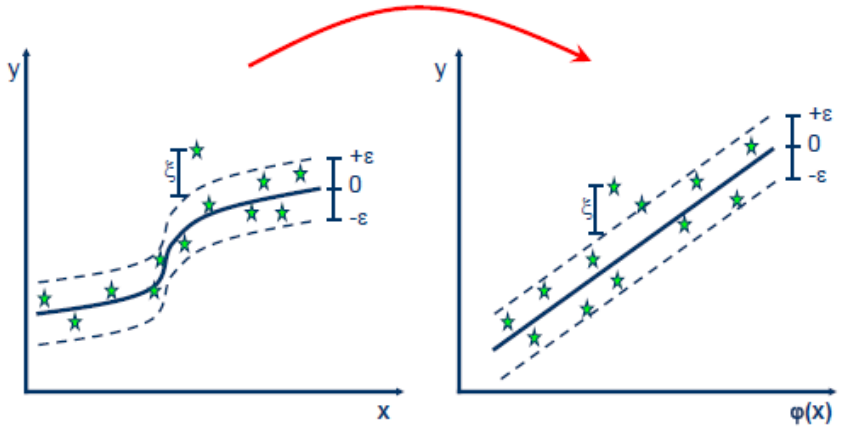

Figura 13: SVR Não-Linear [29], onde ϵ é a margem de erro e ζ é a diferença entre a margem de erro e o objeto. Nesse caso é necessário utilizar um *kernel* para linearizar os dados transformando a abscissa x em $\varphi(x)$

2.13 Otimização de Hiperparâmetros

Mesmo aplicando os algoritmos e obtendo resultados satisfatórios, é possível realizar ajustes em cada função descrita nas etapas anteriores de classificação e regressão que podem melhorar o desempenho do método utilizado. Os elementos ajustados são denominados hiperparâmetros e o ato de ajusta-los se chama de sintonia de hiperparâmetros.

"Hiperparâmetros são utilizados para configurar diversos aspectos do algoritmo de aprendizado e podem obter efeitos descontrolavelmente variantes em relação ao modelo resultante e sua performance" (CLAESEN; MOOR, 2015, Tradução nossa).

Os hiperparâmetros, mesmo tendo relevância para o resultado final do processo de previsão de dados, não estão diretamente ligados ao modelo dos dados, ou seja, a sintonia não altera a informação e sim a forma como o algoritmo irá se comportar diante do modelo utilizado.

2.13.1 Pesquisa em grade (*Grid Search*)

Após aprender os algoritmos de classificação e regressão, uma dúvida surge: Como saber quais hiperparâmetros utilizar? Certamente uma das formas se obter esse valores é combinando valores manualmente, tornando o processo lento e, por vezes, maçante.

A Pesquisa em grade automatiza essa pesquisa, tratando cada combinação de valores um por um até achar o melhor resultado dessa pesquisa. Esse método pode ser bastante lento e computacionalmente custoso, mas também pode ser paralelizado, dado que um processo não depende do outro.

2.14 Validação cruzada

Ao se executar algum dos algoritmos, mesmo utilizando a Pesquisa em grade, a divisão de objetos para treinamento e para teste pode variar em cada processo de execução, gerando diferentes modelos de previsão e variações na taxa de acerto do algoritmo. Para resolver esse problema, existem alguns métodos que podem auxiliar na estabilização desse resultado e indicar uma efetividade permanente de execução do algoritmo.

2.14.1 *K* grupos (*K-fold*)

"A validação cruzada é uma técnica computacional intensiva, usando todos os exemplos disponíveis como exemplos de treinamento e teste. Ele imita o uso de conjuntos de treinamento e teste repetidamente treinando o algoritmo K vezes com uma fração 1 / K de exemplos de treinamento deixados para fins de teste"(BENGIO; GRANDVALET, 2004).

A Figura 14 mostra o funcionamento do processo que pode ser explicado da seguinte forma:

1. Define-se o número 'K' de vezes que os dados originais serão partidos, sendo todos de tamanho iguais ou próximos, já que dependendo do número de objetos e do valor atribuído de partições, o valor da divisão pode não ser exato, levando a uma partição ter um elemento a mais, não alterando o desempenho do algoritmo;

2. K-1 elementos serão utilizados para o treinamento, enquanto a única partição restante será usada para teste;

3. O algoritmo de previsão é executado e ao final, sua acurácia (taxa de acerto) será calculada;

4. O processo se repete K vezes, até que todas as partições tenham sido utilizadas individualmente como teste;

5. A acurácia final será a média ou mediana de cada acurácia calculada anteriormente.

2.15 Trabalhos Relacionados

Existem trabalhos que também buscaram otimizar o sistema licitatório aplicando elementos de Aprendizado de Máquina, onde diferentes trabalhos como o de Rebouças et al. (2015) que busca detectar figurantes em pregões eletrônicos, Fraga (2017) que tem o objetivo de identificar fraudes em licitações no estado da Paraíba, Silva (2011) que

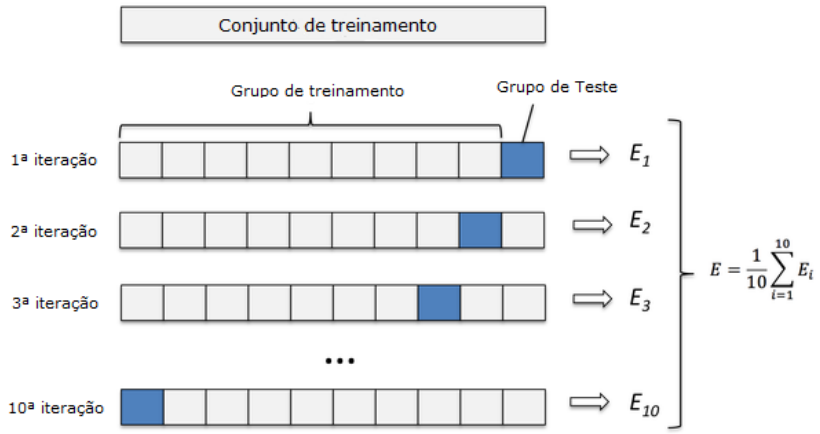

Figura 14: Funcionamento do método de *K* grupos [34], onde o dado de treinamento é dividido em *k-1* grupos de treinamento e um grupo de teste. Durante a execução das etapas, altera-se a definição de qual grupo será o de teste e quais serão a de treinamento

visa identificar fraudes em licitações por todo pais utilizando elementos de computação distribuída e Sales e Carvalho (2014) que através de algoritmos de Classificação buscam encontrar irregularidades em contratos governamentais. Todos tem o mesmo objetivo, encontrar elementos fraudulentos que podem alterar o valor final de uma licitação, seja o aumento do preço proposital em seus lances, seja colocando valores inviáveis para o órgão governamental a fim de beneficiar outra empresa. Será descrito brevemente a estrutura desses projetos e suas semelhanças com essa dissertação

2.15.1 Detecção de figurantes em pregões eletrônicos do governo federal brasileiro

O projeto de Rebouças et al. (2015) investiga maneiras de identificar fraudes em pregões eletrônicos utilizando a investigação do perfil dos participantes durante a ocorrência desses eventos e relacionando com personalidades já pré-definidas.

O artigo informa sobre as leis e jurisprudências relacionadas as fundamentações sobre o pregão eletrônico e possui dados estatísticos sobre a ocorrência de fraudes em pregões e leilões mundialmente. Os autores visam relacionar os participantes utilizando regras de associação para encontrar um padrão de participação entre os perdedores e o vencedor do evento. No entanto, o trabalho não visa definir vencedores e perdedores em um evento, e sim associar comportamentos de participantes a fim de encontrar vícios no processo licitatório.

2.15.2 Detecção de casos suspeitos de fraudes em licitações realizadas nos municípios da Paraíba: uma aplicação de técnicas de mineração de dados

Nesse trabalho, Fraga (2017) utiliza o modelo de associação *apriori* para identificar proponentes que trabalham em conjunto para aumentar o valor pago pelo estado ao vencedor, normalmente uma empresa menor que funciona como "laranja" a fim de viciar o processo licitatório em favor a uma empresa maior.

Essa dissertação apresenta o objetivo semelhante ao trabalho descrito anteriormente, relacionando os participantes e encontrando padrões de participação, mas diferentemente de Rebouças et al. (2015), não utiliza de personalidades pré-definidas para determinar algum comportamento suspeito, pois o mesmo só tem acesso aos processos já finalizados.

2.15.3 Detecção de cartéis em licitações públicas com agentes de mineração de dados

Silva (2011) também utiliza a abordagem *apriori* para definir relacionamentos entre os licitantes a fim de encontrar rodízios de vencedores durante as licitações, mas nesse caso o autor busca adicionar elementos de computação distribuída para processar os dados licitatórios de todo o Brasil, agrupando os dados de cada região em blocos de processamento que serão divididos entre as máquinas disponíveis para processamento.

2.15.4 Análise multi-variada de dados aplicada na previsão irregularidades em contratos do governo brasileiro

Sales e Carvalho (2014) utilizam algoritmos de classificação como as Máquinas de Vetores de Suporte, do inglês *Support Vector Machine* (SVM) e as florestas aleatórias (*Random Forests*) com o objetivo de encontrar irregularidades em contratos ao solicitar o trabalho de uma empresa inadimplente, ou seja, que não está de acordo com os órgãos fiscalizadores brasileiros. Os algoritmos apresentaram em média de 65% de acerto para encontrar irregularidades em contratos federais.

2.15.5 Relação entre os trabalhos

Como indicado nas subseções acima, os projetos tinham em sua maioria o objetivo de relacionar licitantes através de regras de associação a fim de achar conexões que indiquem alguma fraude no processo licitatório, enquanto esse trabalho tem como intenção prever o resultado das licitações e seus respectivos valores finais de uma maneira

mais geral, com o objetivo de auxiliar o auditor a tomar a decisão quanto a seleção de vencedores.

3 METODOLOGIA

Para a realização desse projeto foi necessária a parceria com o Tribunal de Contas do Estado da Paraíba (TCE-PB) e a liberação do banco de dados do Sistema de Tramitação de Processos e Documentos (TRAMITA) e do Sistema de Acompanhamento da Gestão dos Recursos da Sociedade (SAGRES) com o propósito de extrair as informações sobre as licitações e outros bancos de dados necessários para a identificação das mesmas, como de Unidade Gestora, Jurisdicionado, Modalidade entre outras bases.

Foram extraídas 150170 propostas de licitações contendo 27 atributos cada, sendo 14 atributos numéricos e 13 textuais[1]. Além disso, são contabilizados 121689 propostas vencedoras e 28481 perdedoras. Para a aplicação dos algoritmos de Aprendizado de Máquina foram utilizados apenas atributos numéricos e para a visualização dos dados foram utilizado os 13 atributos textuais e mais 3 atributos numéricos (valor da proposta, valor homologado e participações).

Para todo o processamento dos dados e aplicação dos algoritmos de Aprendizado de Máquina foi utilizada a linguagem de programação R em conjunto com o ambiente de desenvolvimento RStudio. Foram usadas as bibliotecas *tidyverse* para carregamento, pré-processamento e agrupamento das informações, *ggplot2* para visualização dos dados, *caTools* para a divisão treinamento-teste, *e1071* e *caret* para a aplicação dos algoritmos de Aprendizado de Máquina. Todo esse processamento foi realizado em um computador com as seguintes especificações:

- Processador Intel Core i5-3210M
- Sistema Operacional Windows 10 Home 64 bits
- Memória RAM de 6 GB
- Disco rígido de 1 TB

3.1 Aplicação do método ETL nos dados de licitação

Organizar a informação a qual será analisada é uma etapa fundamental, pois com dados desorganizados a possibilidade de ocorrência de ruídos e valores inconsistentes torna-se bem maior, além do que, durante essa etapa, o analista passa a ter um conhecimento prévio dos dados, conhecendo a distribuição dos dados em sua forma original e também identificando *outliers* que podem alterar o resultado final dos modelos de previsão.

[1]Estrutura disponível no APÊN A

A estrutura de Armazém de dados foi utilizada pois o Tribunal de Contas da Paraíba requereu que essa estrutura fosse utilizada durante o processo de portabilidade dos dados e que essa abordagem se mostrou mais eficaz em coletar informação das bases do que o processo tradicional de bancos de dados relacional

A informação foi dividida da seguinte forma:

- A dimensão que descreve as propostas de licitação, também chamada de "fato" na estrutura de Armazém de dados, que unifica informações provenientes de vários outros conjuntos de dados;
  - A "fato" desse DW, é a de proposta de licitação, que além de adicionar as informações das dimensões, adiciona o valor da proposta e item proposto como informação específica;
- "A Dimensão possui característica descritiva dentro do DW. Ela qualifica as informações provenientes da tabela Fato. Através dela é possível analisar os dados sob múltiplas perspectivas."(DIEGO ELIAS, 2014);
- As informações são codificadas de acordo com a identificação de cada elemento em sua respectiva dimensão, isso garante que utilize-se os dados sem que nenhuma informação privada seja revelada ao público;

Mas antes de chegar ao resultado exposto na Figura 15, existem alguns erros no banco de dados, que são preenchidos de forma manual sem um formulário fixo com correção de erros, que provavelmente distorceriam o resultado final das análises ou então causariam erro durante a etapa de aplicação dos algoritmos, como:

A estrutura final de licitação é mostrado pela Figura 15, onde ao centro está presente a tabela "FATO_PROPOSTA" e as dimensões, que possuem o prefixo "DIM" e são responsáveis por identificar a tabela principal.

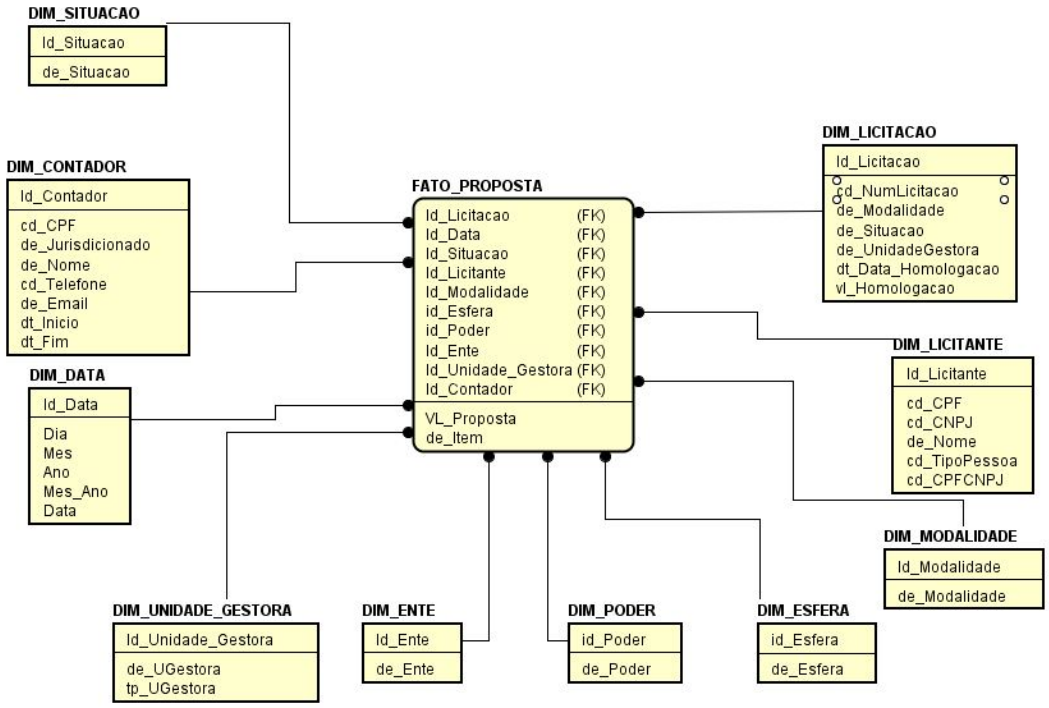

Figura 15: Estrutura do Armazém de Dados

Embora o processo de ETL tenha o objetivo de corrigir erros de tipo e de formatação, o processo não tem como corrigir alguns casos de erro de inserção, tais como:

- Valores de CPFs nulos, inválidos ou com nomes que não representam uma pessoa real, como está indicado na Figura 16;
- Licitações homologadas, mas sem data de homologação e valor final homologado;
- Entes referentes a locais fora do estado da Paraíba;
- Valores de proposta nulos;
- Valores de propostas inviáveis, como o valor de um centavo para uma obra de engenharia;
- Dispensas de licitações e casos especiais com perdedores, o que é inviável de acordo com a lei das dispensas licitatórias;

	cpf	nome
1	19582815140	Não quero me identificar para não ser perseguido
2	70530188309	Renato Galvao Maria
3	05796091468	Juvenal Soares Bezerra
4	27848379205	Não quero me identificar para não ser perseguido
5	34253546471	Não quero me indentificar por questão de segurança
6	05794714468	Maria das Neves Mendonça dos Santos
7	52651231487	José Orlando Pereira Agripino
8	37989270579	Não quero me identificar para não ser perseguido
9	00015315495	Maria do Carmo de Souza
10	15160347402	GIOVANA DE MELO PONTES
11	68328228122	Renata Soares
12	23629240410	Maria de Lourdes Alves da Silva
13	40322665221	Usuário de Teste
14	27451798453	Aldiza Justino de Oliveira
15	06329586420	Iraci Vieira de Pontes Bendito
16	33477692608	mpteste
17	2381292349x	José Vieira da Silva
18	20740263404	Maria das Neves Monteiro da Silva
19	14203446487	Norma Parise da Silva Carneiro
20	56879663400	Maria Jose da Costa
21	28826842400	Rezilda Oliveira de Brito
22	26420669243	Raymundo Geraldo Teixeira de Carvalho
23	00128514426	IRACEMA DUARTE GOMES
24	71655325442	CARLOS RANNIERI DE LIMA PONTES

Figura 16: CPFs com informações inválidas

Todos esses casos foram tratados tanto via sistema gerenciador de banco de dados, no caso o *PostegreSQL*, quanto via linguagem de programação R, por meio da biblioteca *dplyr*, que funciona com foco em análise e filtragem de dados, removendo ou corrigindo as informações quando possível.

3.2 Análise Exploratória

Existem duas categorias para a criação das visualizações: A primeira é relacionada a quantidade de licitações, ou seja, o quantitativo que cada jurisdicionado produz de licitações no período de 2015 até licitações futuras já considerando o ano de 2019. Essas licitações do ano de 2019 são "agendadas" de forma que o pagamento e liquidação da despesa só ocorra no exercício em que foi homologado, mas as licitações podem ser criadas e homologadas antes da data de liquidação.

Um fator essencial é o de que alguns pré processamentos foram realizados antes mesmo de gerar a primeira visualização, que são:

- Unificação dos campos de CNPJ e CPF, pois os campos vêm originalmente com tabelas excludentes CPF e CNPJ, causando campos nulos em todo o conjunto de dados

- Criação de um campo de vitórias e taxa de vitória, que serão úteis para auxiliar os algoritmos de Aprendizado de Máquina a classificar corretamente.

- Alteração de valores de mesorregião, microrregião e município para licitações estaduais, uma vez que uma licitação que engloba um estado inteiro não tem a determinação de um município específico, mas ao invés de se adicionar valores nulos ao atributo, foi adicionado o valor zero.

- Remoção de valores homologados nulos ou abaixo de 200 (duzentos) reais e valores com mais de 13 dígitos

  - Valores nulos indicam que a licitação ainda consta em andamento, ou o que não seria surpreendente, erros de inserção por conta do jurisdicionado
  - Existem diversas licitações onde o valor homologado consta como 0,01 reais ou outros valores totalmente inviáveis para uma licitação real, então o valor viável de 200 reais foi estabelecido como limite inferior para a consideração de licitações válidas
  - Os valores com mais de 13 dígitos tratam de erros onde o responsável pela inserção dos objetos no banco inseriu o CNPJ do licitante ao invés do verdadeiro valor da licitação, o que causa um valor alto e que distorce a visualização dos dados. Infelizmente, valores onde o CPF foi inserido no lugar do valor homologado real são mais difíceis de serem detectados, já que eles se disfarçam com dados reais de licitação.

3.3 Classificação

Para a execução dos algoritmos de classificação utilizando as bibliotecas *caret* e *e1071* foram feitos os seguintes ajustes:

- Para evitar que os algoritmos funcionem de forma ineficaz, foi necessário equilibrar a quantidade de objetos para cada classe (Propostas vencedoras e perdedoras)

  - A razão original das classes dos dados originais eram de 19% perdedores e 81% de vencedores
  - A seleção foi feita de forma de coletar licitações que possuam propostas vencedoras e perdedoras, excluindo licitações que possuíam apenas vencedores ou apenas perdedores.
  - Após a seleção de licitações que possuam propostas perdedoras, a razão entre as classes mudou para 46% de perdedores e 54% de vencedores

- Os valores das propostas foram normalizados entre uma faixa de 0 a 1, utilizando a seguinte fórmula:
  - $x_{normalizado} = x - min/(max - min)$, onde x é o valor da proposta atual e os valores de max e min são, respectivamente os maiores e menores valores do atributo de todo o conjunto de dados.

3.4 Regressão

Para a regressão, a meta a ser alcançada se altera, sendo agora ao invés de prever propostas vencedoras e perdedoras, o objetivo é prever o valor homologado das licitações. Para atingir esse propósito e ainda utilizando a biblioteca *e1071*, algumas mudanças no conjunto de dados foram realizadas, como:

- Apenas propostas vencedoras são consideradas
- Valores normalizados da mesma maneira que nos algoritmos de classificação
- Os dados serão subdivididos de duas maneiras:
  - Tipo de objeto
  - Modalidade

4 APRESENTAÇÃO E ANÁLISE DOS RESULTADOS

4.1 Análise por quantidade

A primeira visualização indica quais são as modalidades mais comuns de licitação e qual o peso dos casos especiais (dispensa, dispensada, inexigibilidade, Adesão a Ata de registro de preços) no âmbito licitatório.

Nota-se na Figura 17 que o pregão presencial, mesmo com o incentivo governamental para a utilização do pregão eletrônico, predomina com mais de 77 mil licitações realizadas. Outra observação é que os casos especiais são aproximadamente 39 mil licitações, o que deve ser analisado com mais profundidade, dado que as requisições solicitadas nos artigos 17º, 24º e 25º da lei 8666/93 possuem caráter de emergência e são considerados como exceções do processo licitatório .

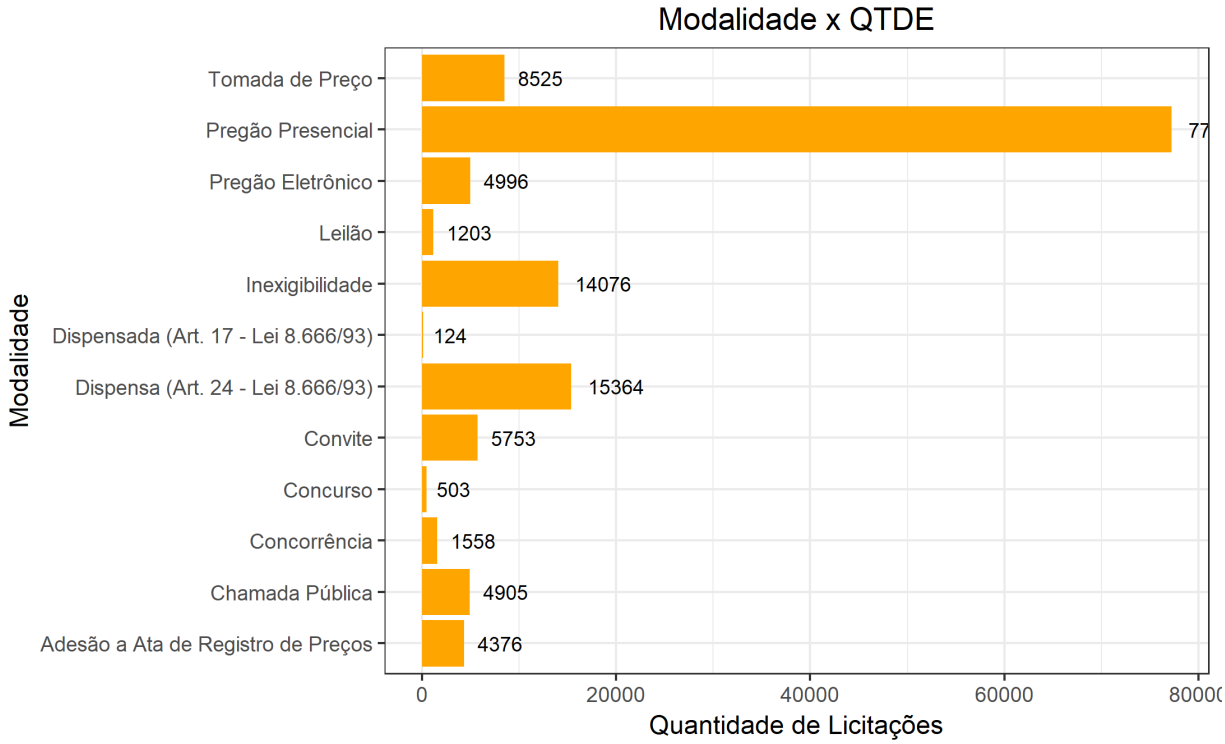

Figura 17: Quantidade de licitações por modalidade no estado da Paraíba, onde QTDE é a quantidade de licitações

A próxima visualização mostra a relação entre a quantidade de licitações e os tipos dos objetos de licitações, A Figura 18 indica que grande parte das licitações são para Compras e serviços, ou seja, gastos simples com itens de consumo e serviços do dia-a-dia. Em segundo ficam as obras e serviços de engenharia e por último as Alienações. O valor baixo de licitações para obras pode se tratar de que alguns serviços que deveriam pertencer a obras e engenharia serem classificado na área de compras e serviços

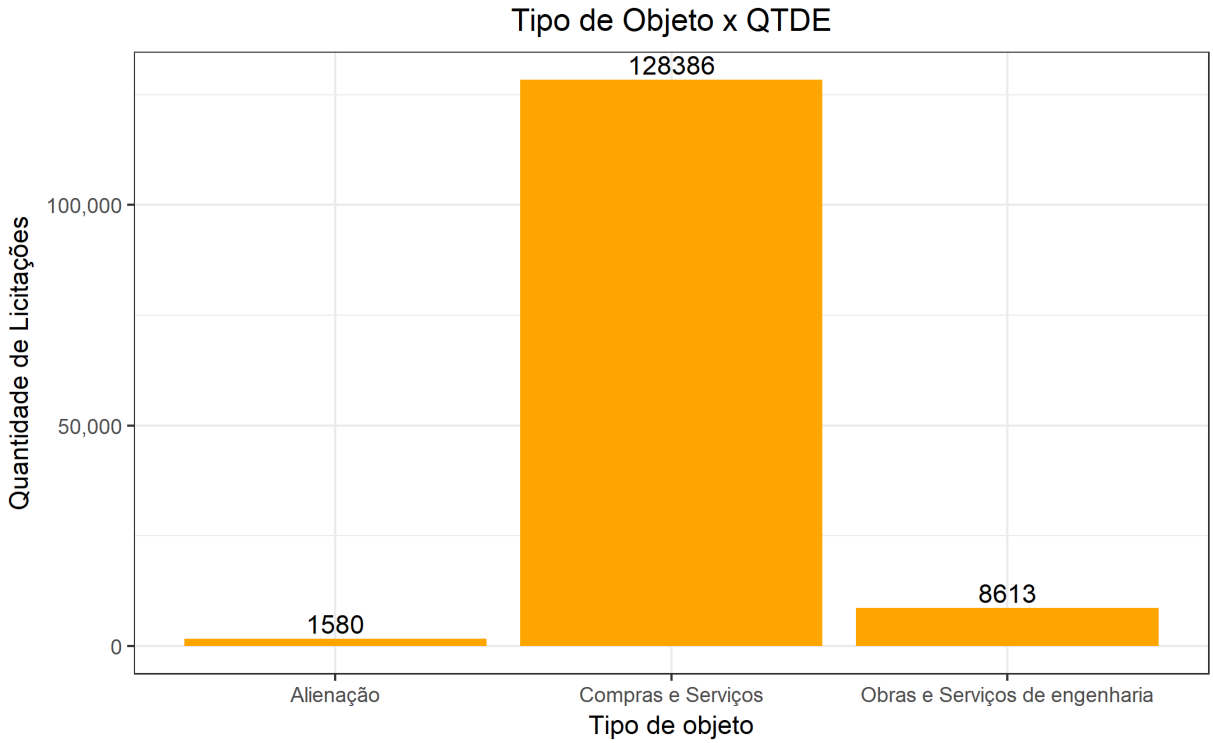

Figura 18: Quantidade de licitações por tipo de objeto no estado da Paraíba, onde QTDE é a quantidade de licitações

A Figura 19 indica a quantidade de licitações por mesorregião, e a maior surpresa se trata da posição da Mata paraibana, onde está situada a capital João Pessoa, pois imaginando que por ser o maior município do estado, este teria dominância em torno das outras mesorregiões, essa lógica é bem aplicada na região do Agreste Paraibano onde está situado a segunda maior cidade da Paraíba, Campina Grande. Outra observação é que licitações estaduais (representadas pelo nome Paraíba) são bem menores do que as municipais.

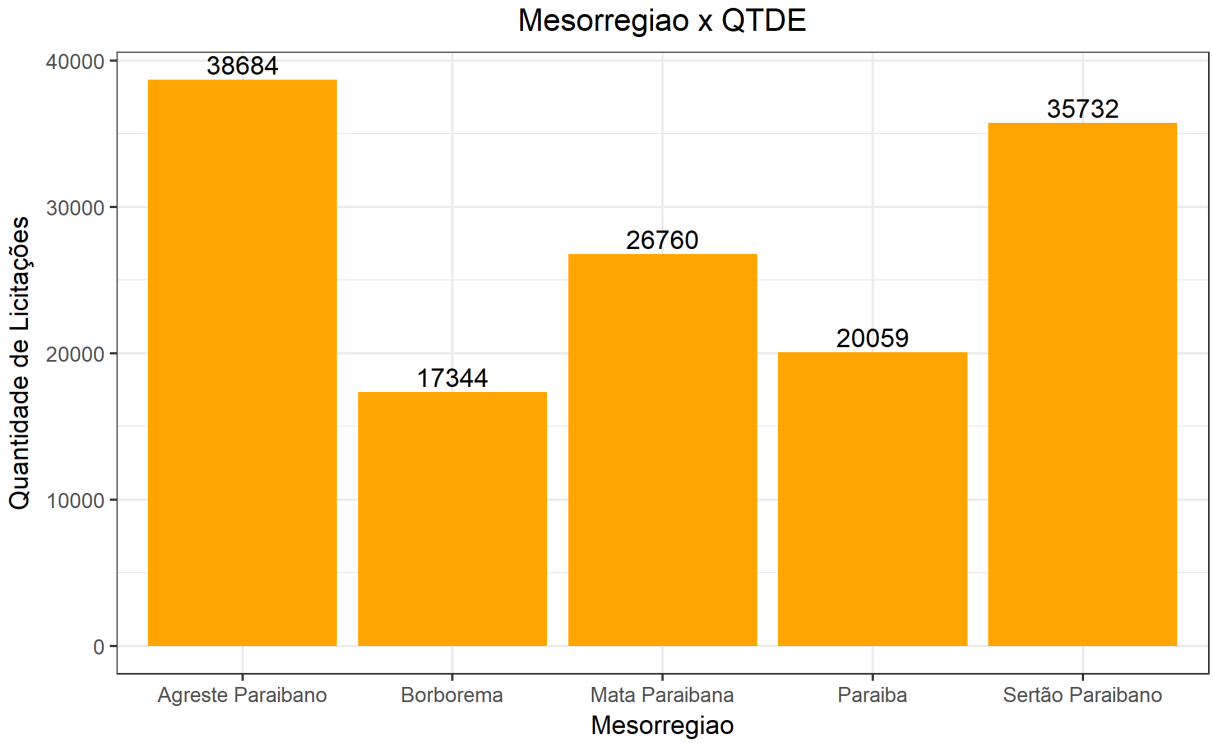

Figura 19: Quantidade de licitações por mesorregião, onde QTDE é a quantidade de licitações

A Figura 20 mostra a evolução da quantidade de licitações com o passar dos anos, e explicita que os processos licitatórios vinham em decadência até atingir um recorde em 2018, tendo mais de 10 mil licitações a mais do que o ano anterior. Uma possível explicação seria a que o ano de 2018 é o ano eleitoral, o que impulsiona as requisições por obras realizadas pelo governo, mas o que não acontece em 2016, um ano eleitoral de que poderia envolver licitações municipais, que teriam maior dependência dos vereadores e prefeitos do que dos deputados, senadores e governadores.

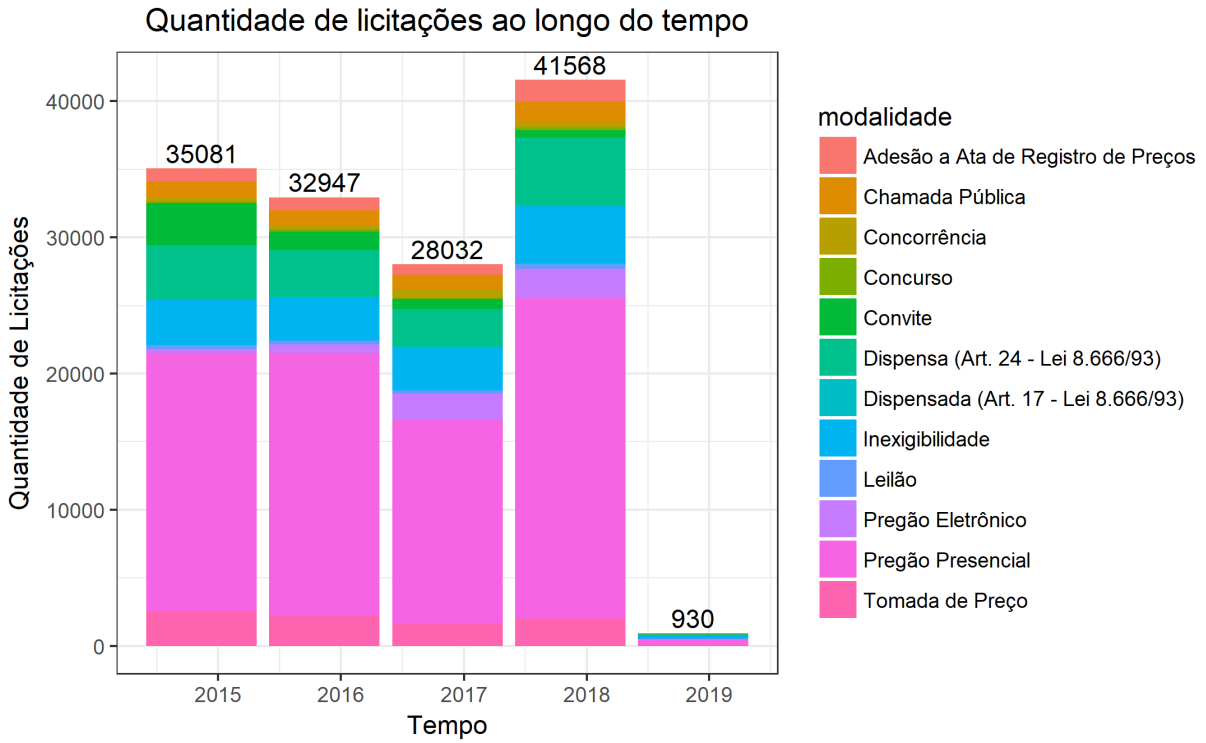

Figura 20: Quantidade de licitações por ano

4.1.1 Análise por valor de licitação

Os gráficos anteriores constataram que a relação entre o tamanho e riqueza da cidade, micro ou mesorregião não era proporcional, ou seja, quanto maior os indicadores econômicos (Produto Interno bruto - PIB) ou sociais (Índice de Desenvolvimento da Educação Básica – IDEB, Índice de Desenvolvimento Humano - IDH, etc.), não necessariamente indicam um maior número de licitações. Levando em consideração que o processo licitatório é um método caro e que estes custos não entram no valor final, vale considerar que existem licitações com valor muito abaixo até desse valor de expedição da licitação, tornando o processo muito mais caro do que o produto.

Essa seção trata de mostrar as licitações de forma monetária e indicando os gastos externos de um jurisdicionado, seja ele com manutenção ou investimentos em obras. Aqui também existirão gráficos mostrando o a mediana para cada grupo através de visualizações com *boxplots*.

A primeira análise é a de maiores licitantes, incluindo também licitações do governo do estado da Paraíba, renomeado para simplesmente "Paraíba". Nele notamos que a relação de tamanho e desenvolvimento do município estão diretamente ligadas ao valor licitado no período de 2015 até 2019.

A Figura 21 mostra a diferença entre investimentos na capital e nos demais municípios, vendo que o segundo maior município, Campina Grande, tem menos da metade

do valor de João Pessoa e que os demais municípios no ranking possuem valores semelhantes acumulados.

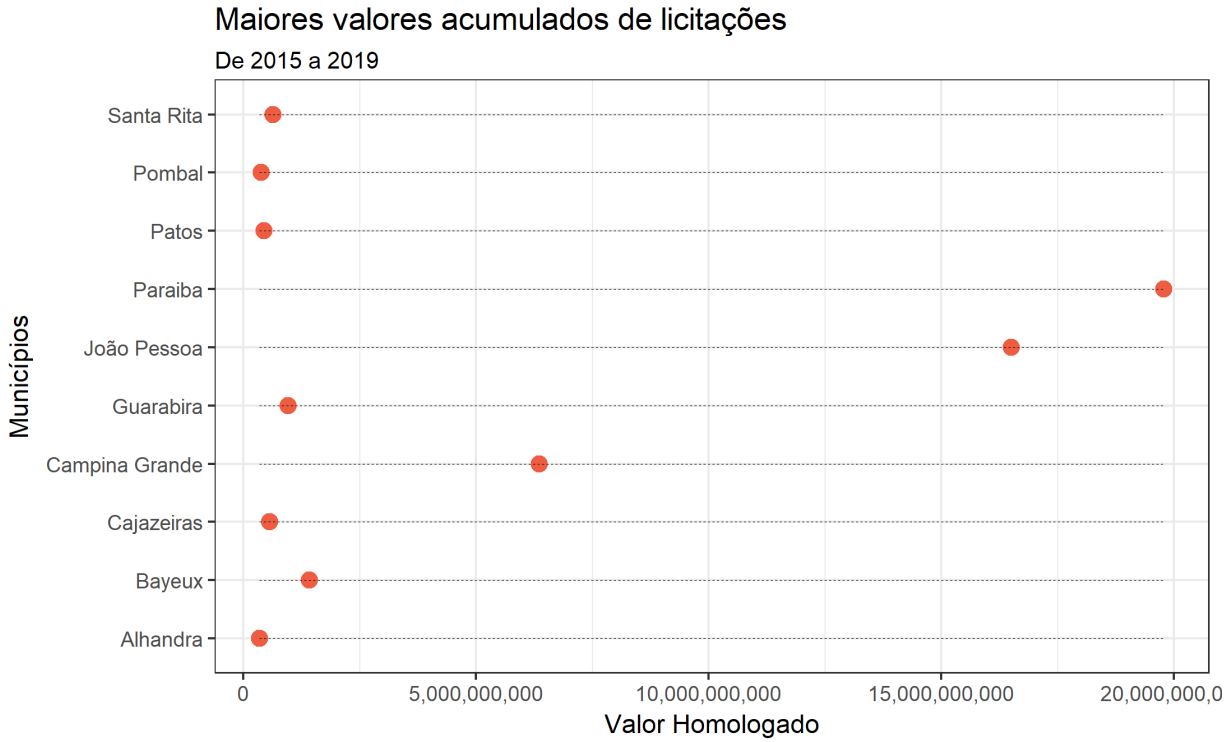

Figura 21: Maiores licitantes em reais (R$)

As próximas visualizações seguirão o mesmo roteiro da seção de quantidade, passando por uma visão por modalidade, tipo de objeto, mesorregião e uma visão ao longo do tempo, tendo gráficos semelhantes com finalidade de indicar uma relação entre quantidade e valor.

A análise de licitações por modalidade, possui uma visão de licitações individuais ao longo do plano via *boxplots*, que mostram a mediana, os quartis e os *outliers* para cada uma das modalidades existentes.

Nota-se na Figura 22 há muitos *outliers*, e que a diferença entre eles e a mediana no *boxplot* podem chegar a casa dos milhões de reais. Isso pode ter ocorrido por erros já citados na etapa de pré-processamento, ou sugere uma licitação a ser investigada e questionada sobre seus valores excessivos. Uma preocupação são os *outliers* na região de pregões, visto que a maioria desses processos apresentam valores baixos, pois eles tem a finalidade de adquirir bens de consumo simples, esses valores fora do normal vão contra a regra de valores mínimos impostos pelas modalidades de Convite, Concorrência e Tomada de preço.

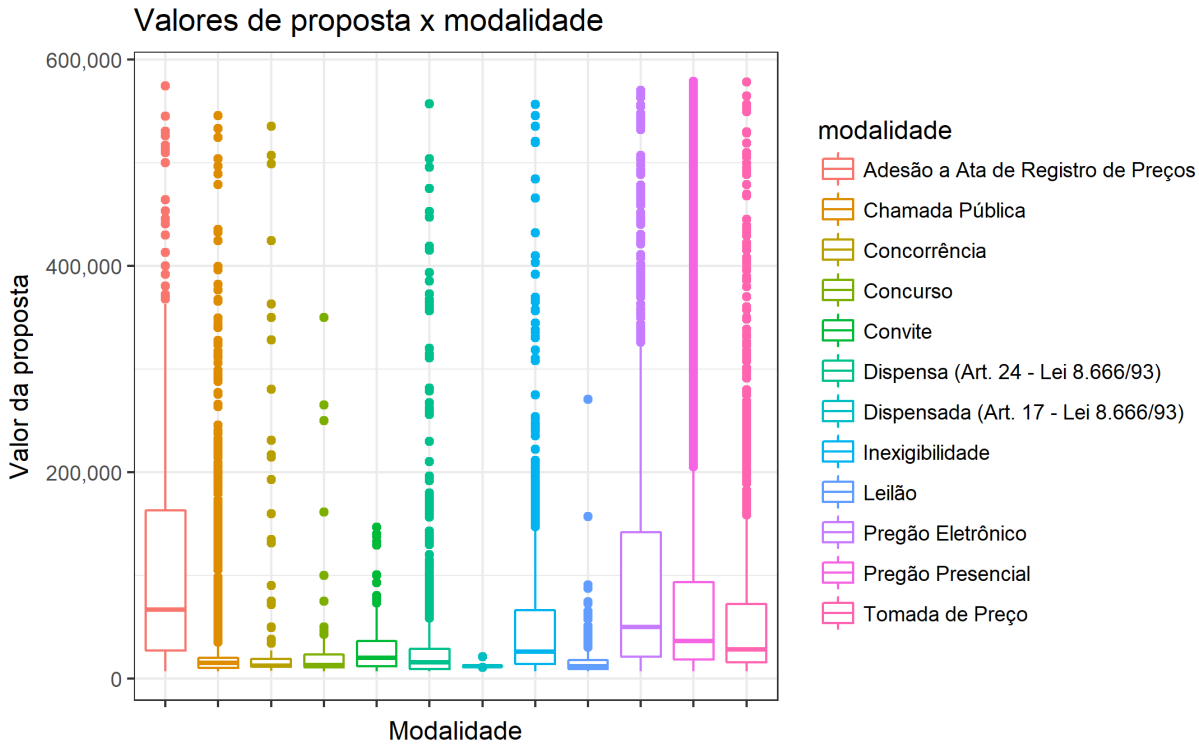

Figura 22: Licitações e seus valores separados por modalidade

A Figura 23 representa agora as licitações agrupadas por tipo de objeto, e a grande maioria dos *outliers* estão presentes no grupo de Compras e serviços, o mais comum e simples de se aplicar dos três, mais suscetível a fraudes do que os de alienação e os de obras e serviços de engenharia, no qual sofrem maior inspeção além de estarem presentes em menor número. Já Figura 24 mostra a diferença entre os valores médios de cada licitação, aumentando a suspeita sobre os *outliers* indicados na Figura 23.

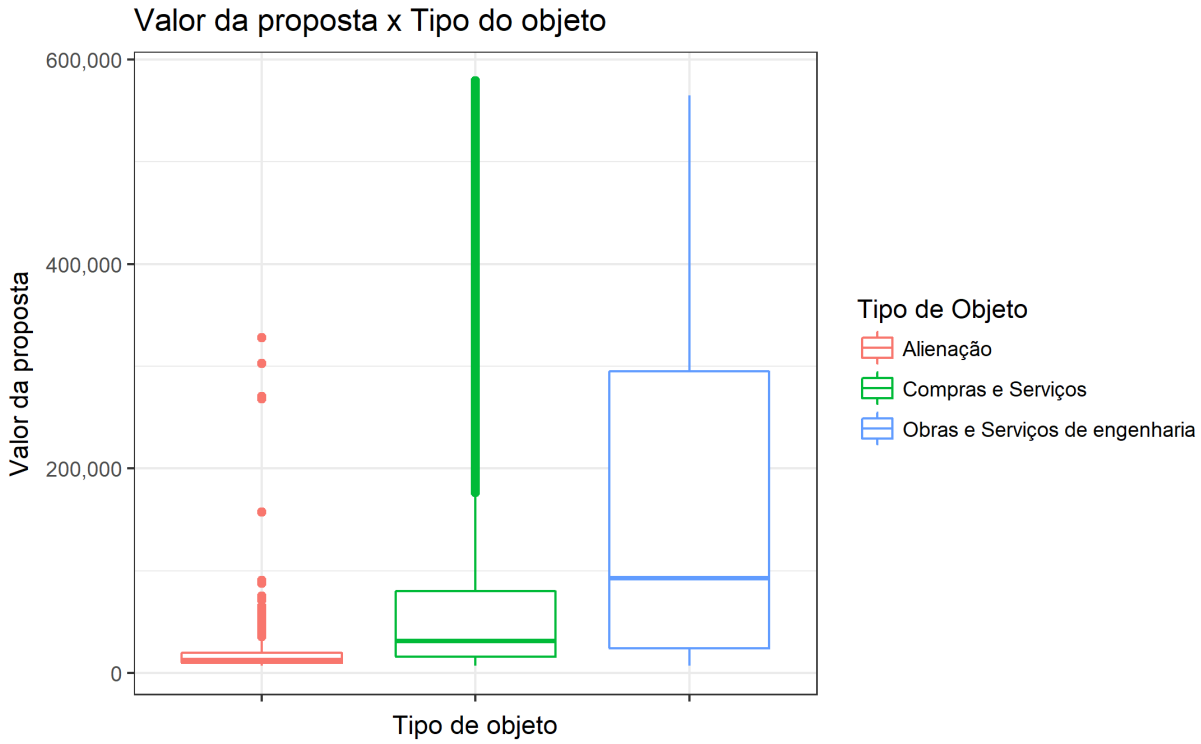

Figura 23: Licitações e seus valores separados por tipo de objeto

Figura 24: Valores médios de cada tipo de objeto

Agrupando agora por mesorregião, podemos comparar a relação quantidade, apresentado na Figura 19 com a visualização indicando os valores por mesorregião. A Figura

25 indica que o Agreste e Sertão apresentam baixos valores cumulativos, mesmo sendo mesorregiões líderes em quantitativo, novamente dando a ênfase de que todo o processo licitatório é manual e custoso, o que pode gerar um prejuízo para os municípios dessa região, caso os valores homologados não ultrapassem os custos operacionais.

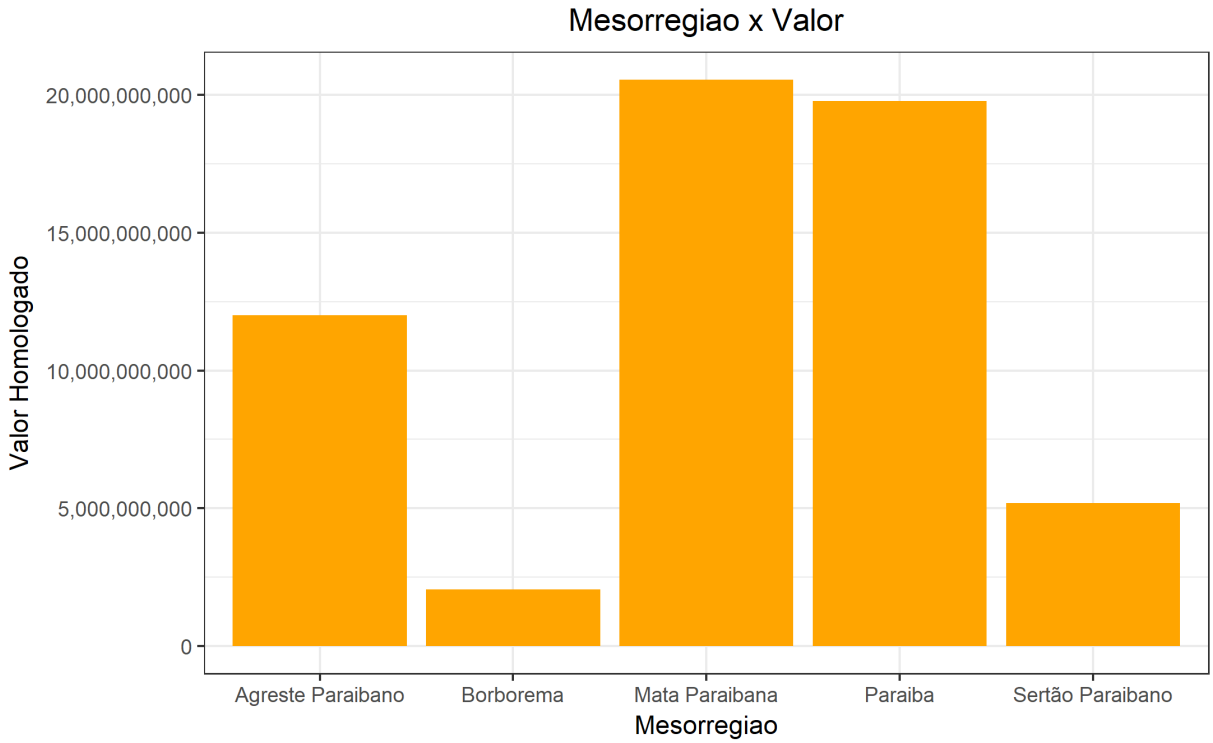

Figura 25: Valores acumulados por mesorregião

As Figuras 26 , 27, 28 mostram a evolução das licitações ao longo do tempo de acordo com os temas já analisados anteriormente. A diferença entre a Figura 20 e as Figuras 26, 27, 28 é que no primeiro há um pico de criação de licitações no ano de 2018 que não se reflete quando observado no ponto de vista monetário. A Figura 26 mostra que houve um avanço nos pregões eletrônicos, devido ao incentivo do governo para tal, mas logo após esse incentivo, as taxa volta a cair, mas ainda assim sendo relativamente alta em relação aos anos de 2015 e 2016. Já as Figuras 27 e 28 explicam os motivos das quedas nos valores do ano atual, com a redução dos investimentos em obras e serviços de engenharia e aumento de investimento nas áreas de compras e serviços nas regiões do Agreste e Sertão paraibano.

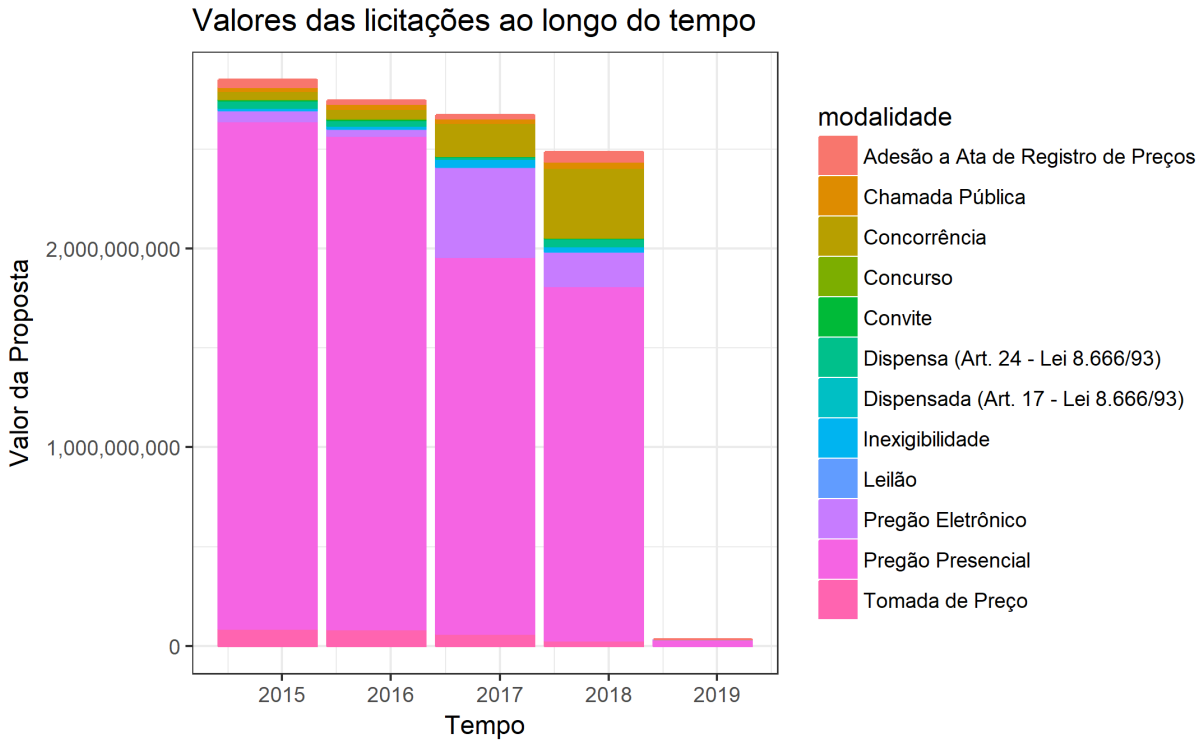

Figura 26: Evolução temporal por modalidade

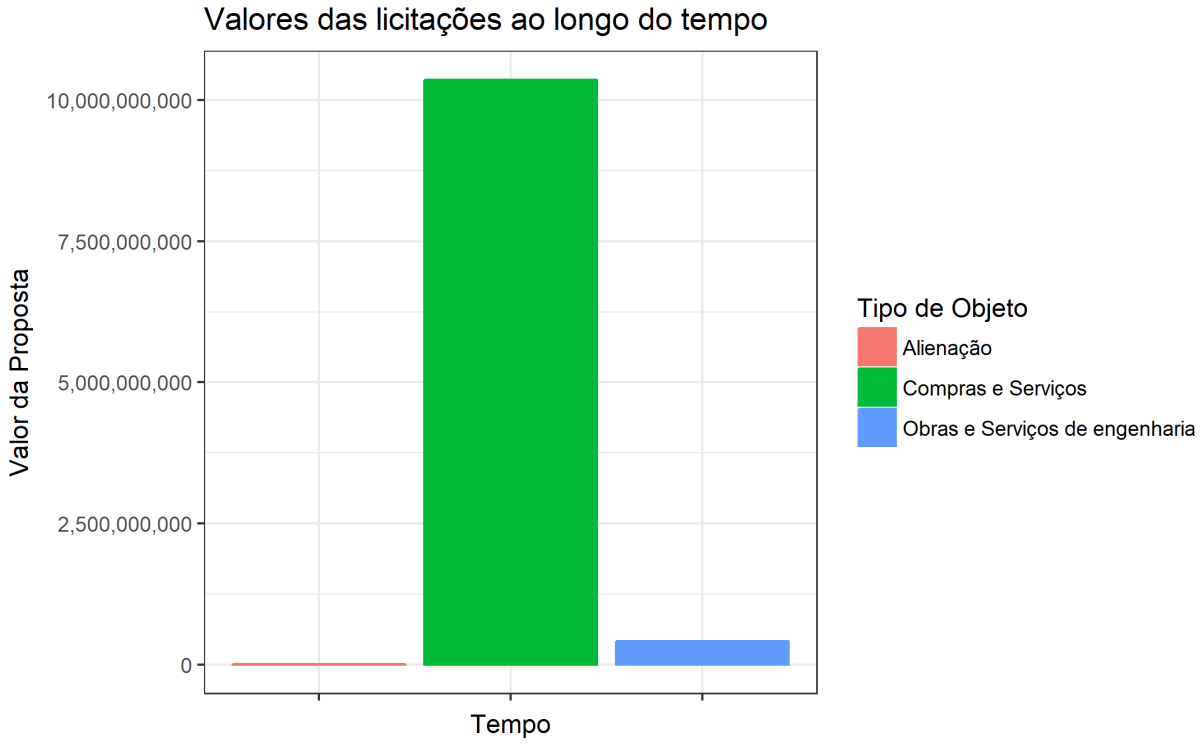

Figura 27: Evolução temporal por tipo de objeto

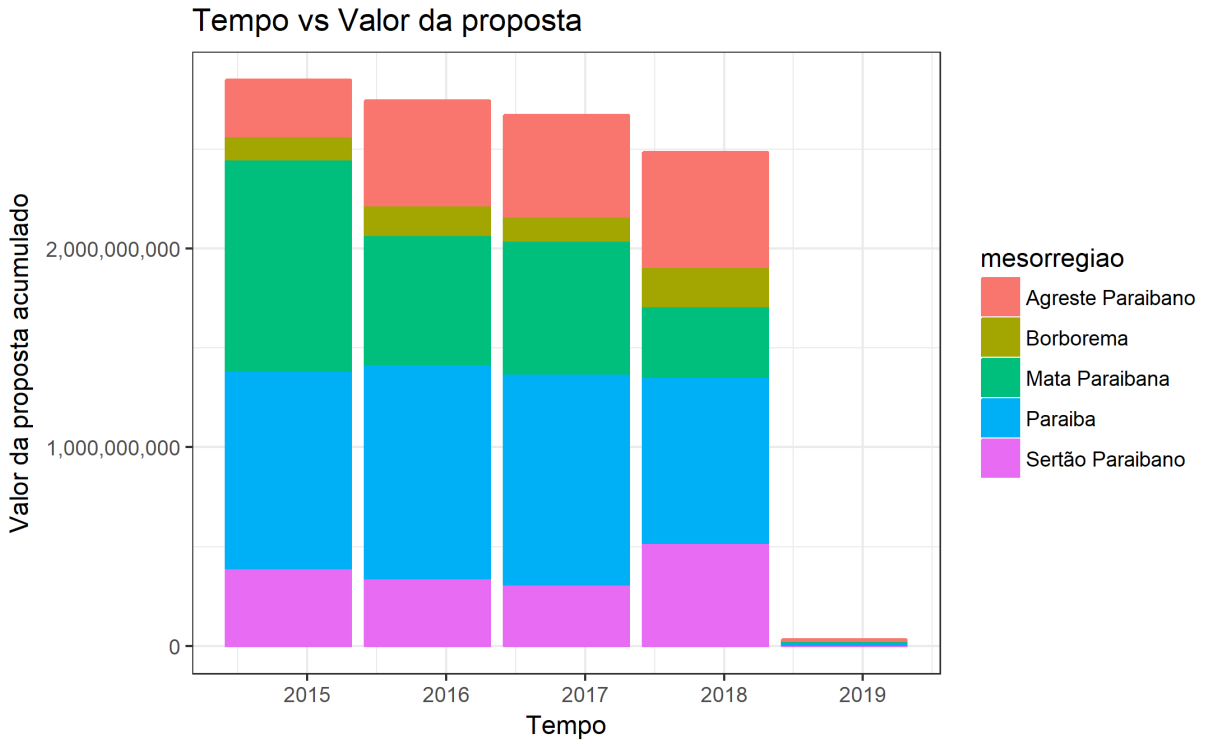

Figura 28: Evolução temporal por mesorregião

4.2 Classificação para definição de situação de propostas

Para o algoritmo SVM e KNN, foi utilizada a seguinte abordagem: As licitações são divididas de acordo com sua modalidade e assim os algoritmos são executados com *k-fold* para cada sub-divisão dos dados, cujo valor de k é definido internamente pela função *train* da biblioteca *caret*, gerando modelos e regras para cada tipo de modalidade existente.

A Tabela 2 mostra o resultado do algoritmo SVM com o *kernel* radial. Esses resultados foram os melhores obtidos pelo algoritmo de seleção de hiperparâmetros *Grid Search* e o de validação cruzada *k-fold* disponíveis na biblioteca *caret*. Na Tabela 2 o parâmetro *Sigma* representa a variável ajustável no denominador da equação $\exp(-\frac{||x-x'||}{2\sigma^2})$ e o parâmetro C indica a relação entre a complexidade do modelo e a permissividade de erro na separação das amostras e a *acurácia* indica a taxa de acerto na classificação do algoritmo.

	Sigma	*C*	*Acurácia*
Convite	$4{,}60 \times 10^5$	0,25	0,65
Tomada de Preço	$4{,}00 \times 10^5$	1	0,67
Concorrência	$3{,}26 \times 10^5$	1	0,76
Alienação	$4{,}82 \times 10^5$	0.25	0,77
Pregão Presencial	$3{,}71 \times 10^5$	1	0,56
Pregão Eletrônico	$4{,}43 \times 10^5$	1	0,75
Inexigibilidade	$3{,}14 \times 10^5$	0,25	0,65
Dispensa	$2{,}96 \times 10^5$	0,5	0,66
Adesão a Ata de Registro de Preços	$3{,}79 \times 10^5$	0,25	0,73
Dispensada	$5{,}57 \times 10^5$	0,25	0,62
Chamada Pública	$5{,}38 \times 10^5$	0,25	0,67
RDC	$3{,}26 \times 10^5$	0,25	0,82

Tabela 2: Resultado do algoritmo SVM para cada modalidade, onde *Sigma* representa a variável ajustável no denominador da equação $\exp(-\frac{||x-x'||}{2\sigma^2})$ **e *C* representa a relação entre a complexidade do modelo e a permissividade de erro**

Utilizando as mesmas bibliotecas do algoritmo SVM, a Tabela 3 apresenta o resultado do algoritmo KNN, onde K representa o número de vizinhos mais próximos e a *acurácia* é a taxa de acerto na classificação do algoritmo .

	K	*Acurácia*
Convite	9	0,56
Tomada de Preço	9	0,61
Concorrência	9	0,71
Alienação	9	0,74
Pregão Presencial	9	0,56
Pregão Eletrônico	9	0,71
Inexigibilidade	9	0,57
Dispensa	9	0,58
Adesão a Ata de Registro de Preços	9	0,67
Dispensada	9	0,57
Chamada Pública	9	0,64
RDC	9	0,75

Tabela 3: Resultado do algoritmo KNN para cada modalidade, onde *K* indica o número de vizinhos mais próximos analisados

Nota-se que o algoritmo SVM possui maior eficiência em relação ao KNN, possuindo um nível de acerto médio de 69,25% enquanto o KNN possui nível de acerto médio de 63,91%.

4.3 Regressão para descoberta de valores homologados

A utilização da regressão tem como prever os valores homologados das licitações, ou seja, o valor final que o estado terá que pagar aos vencedores por oferecerem as melhores propostas. Para isso foram removidas as informações sobre os licitantes (como CPF/CNPJ, taxa de vitórias, participações, etc.) e mantidas apenas informações sobre a licitação (Modalidade, tipo de objeto, Jurisdicionado, etc.)

Os algoritmos a serem comparados são o SVR e o de regressão linear, onde haverá duas tabelas,a Tabela 4 que leva em consideração a subdivisão por tipo do objeto e Tabela 5 que leva em consideração a subdivisão por modalidade. Os valores foram normalizados para aumentar a eficácia dos algoritmos e depois denormalizados para a aplicação da métrica de erro. A métrica considerada para calcular os erros entre o valor previsto pelo algoritmo e o valor real é o das médias dos erros quadrados (*Mean Squared Error* - MSE). Os dois modelos foram treinados em uma base de treinos que representam 75% dos dados originais e avaliados em uma base de teste, que é composta dos outros 25% dos dados originais.

	Regressão Linear	SVR
Obras e serviços de engenharia	25782,07939	21750,02949
Compras e serviços	64106,1944	62044,07889
Alienação	436,52957	701,09800

Tabela 4: Resultado da regressão por tipo de objeto, onde SVR representa o algoritmo de Regressão de Vetor de Suporte

	Regressão Linear	SVR
Convite	20,100	21,355
Tomada de Preço	515,954	609,608
Concorrência	3474527,69	3694697,328
Concurso	2134,788	1548,804
Alienação	402,867	386,012
Pregão Presencial	112109,067	106726,024
Pregão Eletrônico	58148,78	431541,125
Inexigibilidade	929,54	864,699
Dispensa	2558,271	2489,621
Adesão a Ata de Registro de Preços	12006,47	420682,705
Dispensada	6,447	3,579
Chamada Pública	12880,560	13917,092
RDC	36587,666	4494,078

Tabela 5: Resultado da regressão por modalidade, onde SVR representa o algoritmo de Regressão de Vetor de Suporte

Na comparação entre os algoritmos de regressão, mostra-se que o algoritmo SVR

é mais preciso do que o algoritmo de regressão linear quando subdividido por tipo de objeto e utilizando a métrica de distância euclidiana, porém houve custo computacional maior para realizar o processo no SVR. Já quando subdividido por modalidade, a regressão linear possui um valor de erro menor nas modalidades Convite, Tomada de preço, concorrência, Adesão a Ata de Registro de preços e Chamada pública, enquanto o SVR é mais eficaz nas modalidades de Concurso, Alienação, Pregão presencial e eletrônico, Inexigibilidade,Dispensa, Dispensada e RDC.

Colocando em consideração a complexidade em torno do processamento de cada algoritmo, O SVR não apresenta grande vantagem nas previsões para ser utilizado no lugar de um algoritmo mais simples, como a regressão linear.

5 CONCLUSÕES E TRABALHOS FUTUROS

Após as comparações na seção anterior, concluímos que o algoritmo SVM possui a melhor acurácia dentre os algoritmos de classificação, enquanto no caso da regressão, aplicar um algoritmo de regressão linear nas informações sobre licitações não vai gerar erros maiores do que utilizando um algoritmo mais robusto como o SVR.

Embora os resultados de classificação estejam atingindo em alguns casos a faixa de 75% de acerto, os resultados não foram satisfatórios, devido ao fato de que muitas das informações estavam com valores irreais de proposta ou homologação, reduzindo de forma considerável o conjunto de dados para a aplicação dos algoritmos de *Machine Learning*.

Em contrapartida, demonstra-se a importância na área de pré-processamento e organização dos dados, que muitas vezes é negligenciada por estudantes e novatos na área de Ciência de Dados, por normalmente utilizar dados já pré-processados e prontos para a aplicação dos algoritmos.

O resultado mostra que as relações entre os atributos aparentam ser não-lineares, o que indica que algoritmos de Aprendizado de Máquina podem ter o desempenho prejudicado e limitado. Uma boa opção para maior precisão nas previsões seria utilizar modelos de Aprendizado Profundo (*Deep Learning*), que conseguem trabalhar com modelos não lineares mais facilmente, embora tenham um custo computacional maior.

REFERÊNCIAS

[1] BRASIL, Tesouro Nacional. **Manual de Contabilidade: Aplicada ao setor público**. Brasil: 2017

[2] OLIVEIRA, Rafael. **Licitações e Contratos Administrativos**. São Paulo: Editora Método, 2014

[3] BRASIL. Lei Nº 8.666, de 21 DE Junho de 1993. **Lei de Licitações**. Brasília, DF, jun 1993

[4] SILVA, Ana Lêda Rocha da et al. Quanto custa um processo administrativo de compras e contratação de serviços? O Caso da Secretaria de Saúde do Município de Feira de Santana na Bahia. In: CONGRESSO BRASILEIRO DE CUSTOS, 22., 2015, Foz do Iguaçu. Anais. Foz do Iguaçu: CBC, 2015. p. 1 - 12.

[5] IFI CLAIMS PATENT SERVICES (PATENT ANALYTICS). **8 Fast Growing Technologies**. 2018. Disponível em: <https://www.ificlaims.com/rankings-8-fast-growing.htm>. Acesso em: 01 ago. 2018.

[6] KRISTI LEWANDOWSKI. Memsql. **2018 Outlook: Machine Learning and Artificial Intelligence**. 2018. Disponível em: <http://blog.memsql.com/2018-outlook-machine-learning-and-artificial-intelligence/>. Acesso em: 01 ago. 2018.

[7] BRASIL. Lei Nº 4.320, de 17 de Março de 1964. **Lei de Finanças Públicas**. Brasília, DF, mar 1964

[8] MEIRELLES, Hely Lopes.**Direito administrativo brasileiro**. 24 Ed. São Paulo: Malheiros Editores. 1999. 749p.

[9] BRASIL. Decreto Nº 3555, DE 8 de agosto de 2000. **Decreto sobre a regularização de pregão**. Brasília, DF, ago 2000

[10] BRASIL. Decreto Nº 7892, de 23 de Janeiro de 2013. **Decreto sobre a adesão a ata de registro de preço**. Brasília, DF, jan 2013

[11] BRASIL. Lei Nº 13.019, de 31 de Julho de 2014. **Lei sobre Chamamento público**. Brasília, DF, jul 2014

[12] CHIMINAZZO, Lucas. **O QUE É EMPENHO ?**. 2015. Disponível em: <http://www.partnersales.com.br/artigo/impressao/1111> . Acesso em: 27 jul. 2018.

[13] COMO PASSAR EM CONCURSO.**Lei nº 8666/93: atualizada e comentada**. Disponível em: ¡http://comopassaremconcurso.com.br/lei-no-8666-atualizada-e-comentada-01/¿. Acesso em: 12 ago. 2018.

[14] ESCOLA DO SERVIDOR PÚBLICO DO ACRE (Acre). Governo do Estado do Acre. **Modalidades de Licitação**. Disponível em: <http://www.escoladoservidor.ac.gov.br/wps/wcm/connect/528b5c80484cfd47a28dbfd851b80238/LICITAÇÃO+-+MODALIDADES+-+EXTRA.pdf>. Acesso em: 12 ago. 2018.

[15] NORMAS LEGAIS. **LICITAÇÃO PÚBLICA: INEXIGIBILIDADE**. Disponível em: <http://www.normaslegais.com.br/guia/clientes/inexigibilidade-licitacao.htm>. Acesso em: 16 ago. 2018.

[16] DATAWATCH.**What is Data Preparation?** 2018. Disponível em: <https://www.datawatch.com/what-is-data-preparation/>. Acesso em: 29 ago. 2018.

[17] STEVE LOHR. New York Times. **For Big-Data Scientists, 'Janitor Work' Is Key Hurdle to Insights**. 2014. Disponível em: <https://www.nytimes.com/2014/08/18/technology/for-big-data-scientists-hurdle-to-insights-is-janitor-work.html>. Acesso em: 29 ago. 2018.

[18] SAS. **ETL: What it is and why it matters** Disponível em : <https://www.sas.com/en_us/insights/data-management/what-is-etl.html>. Acesso em: 29 ago. 2018.

[19] - TOTAL DATA MANAGEMENT. **Fast ETL Is Affordable:** Go Beyond Legacy ETL Tools with Voracity. Disponível em: <https://www.iri.com/solutions/data-integration/etl>. Acesso em: 30 ago. 2018.

[20] , C. J. **Introdução a Sistemas de Bancos de Dados**. 8ª Ed., Rio de Janeiro: Campus, 2004.

[21] DEVMEDIA. Data Warehouse. Disponível em: <https://canaltech.com.br/business-intelligence/conhecendo-a-arquitetura-de-data-warehouse-19266/>. Acesso em: 30 ago. 2018.

[22] SUNIL RAY. Analytics Vidhya. **Beginners Guide To Learn Dimension Reduction Techniques.** 2015. Disponível em: <https://www.analyticsvidhya.com/blog/2015/07/dimension-reduction-methods/>. Acesso em: 31 ago. 2018.

[23] VASCONCELOS, Simone; CONCI, Aura. **Análise de Componentes Principais (PCA)**. Disponível em: <http://www2.ic.uff.br/ aconci/PCA-ACP.pdf>. Acesso em: 02 set. 2018.

[24] KIRILL EREMENKO. Superdatascience. **Kernel Trick.** Disponível em: <https://www.udemy.com/machinelearning/learn/v4/t/lecture/6113150>. Acesso em: 02 set. 2018.

[25] MAATEN, Laurens van Der. **T-sne**. 2008. Disponível em: <https://lvdmaaten.github.io/tsne/>. Acesso em: 02 set. 2018.

[26] TECHNOLOGY OF COMPUTING. **A Short Introduction to K-Nearest Neighbors Algorithm**. 2016. Disponível em: <https://helloacm.com/a-short-introduction-to-k-nearest-neighbors-algorithm/>. Acesso em: 16 set. 2018

[27] SUNIL RAY. Analytics Vidhya. **Understanding Support Vector Machine algorithm from examples (along with code)**. 2017. Disponível em: <https://www.analyticsvidhya.com/blog/2017/09/understaing-support-vector-machine-example-code/>. Acesso em: 16 set. 2018.

[28] SUNIL RAY. Analytics Vidhya. **Easy Steps to Learn Naive Bayes Algorithm**. 2017. Disponível em: ¡https://www.analyticsvidhya.com/blog/2017/09/naive-bayes-explained/¿. Acesso em: 23 set. 2018.

[29] PORTAL ACTION. **Análise de Regressão**. Disponível em: <http://www.portalaction.com.br/analise-de-regressao>. Acesso em: 08 set. 2018.

[30] EREMENKO, Kirill. **Machine Learning A-Z™: Hands-On Python & R In Data Science**: Support Vector Regression (SVR). 2018. Disponível em: <https://www.udemy.com/machinelearning/learn/v4/t/lecture/10459548?start=0>. Acesso em: 19 out. 2018.

[31] SAYAD, Saed.**Support Vector Machine - Regression (SVR)**.Disponível em: <http://www.saedsayad.com/support_vector_machine_reg.htm>. Acesso em: 08 set. 2018.

[32] CLAESEN, Marc; MOOR, Bart de. **Hyperparameter Search in Machine Learning**. 2015. Disponível em: <https://arxiv.org/pdf/1502.02127.pdf>. Acesso em: 08 set. 2018

[33] BENGIO, Yoshua; GRANDVALET, Yves.**No Unbiased Estimator of the Variance of K-Fold Cross-Validation**. 2004. Disponível em:

<http://www.jmlr.org/papers/volume5/grandvalet04a/grandvalet04a.pdf>. Acesso em: 09 ago. 2018.

[34] RASCHKA, Sebastian. **I train my system based on the 10-fold cross-validation framework. Now it gives me 10 different models. Which model to select as a representative?** Disponível em: <https://www.quora.com/I-train-my-system-based-on-the-10-fold-cross-validation-framework-Now-it-gives-me-10-different-models-Which-model-to-select-as-a-representative>. Acesso em: 08 set. 2018.

[35] DIEGO ELIAS. Canaltech. **Dimensões e Fatos no contexto do Business Intelligence.** 2014. Disponível em: <https://canaltech.com.br/business-intelligence/dimensoes-e-fatos-no-contexto-do-business-intelligence-bi-18710/>. Acesso em: 20 out. 2018.

[36] REBOUÇAS, R.R. et al. **Detecção de figurantes em pregões eletrônicos do governo federal brasileiro.** *Informação & Tecnologia*. João Pessoa. v. 2, n. 2, p.5-21, 2015

[37] FRAGA, Alcimar Alves. Detecção de casos suspeitos de fraudes em licitações realizadas nos municípios da Paraíba:: uma aplicação de técnicas de mineração de dados. 2017. 92 f. Dissertação (Mestrado) - Curso de Economia, Centro de Ciências Sociais Aplicadas, Universidade Federal da Paraíba, João Pessoa, 2017

[38] SILVA, Carlos Vinícius Sarmento. **DETECÇÃO DE CARTÉIS EM LICITAÇÕES PÚBLICAS COM AGENTES DE MINERAÇÃO DE DADOS.** Revista Eletrônica de Sistemas de Informação, [s.l.], v. 10, n. 1, p.1-21, 2 jul. 2011. IBEPES (Instituto Brasileiro de Estudos e Pesquisas Sociais). http://dx.doi.org/10.5329/resi.2011.1001008

[39] SALES, Leonardo Jorge; CARVALHO, Ricardo Silva. **Análise multivariada de dados aplicada na previsão irregularidades em contratos do governo brasileiro.** 2014. Disponível em: <https://cladista.clad.org/bitstream/handle/123456789/7753/0077415.pdf?sequence=1&isAllo
Acesso em: 02 nov. 2018.

APÊNDICE A – ESTRUTURA DA BASE DE DADOS DE LICITAÇÕES MUNICIPAIS E ESTADUAIS

Nome do Atributo	**Descrição**
IdLicitacao	Identificador interno único para cada licitação
datahomologação	Data em que foram confirmados os vencedores da licitação
objeto	Descrição do objeto a ser licitado
numerolicitacao	Identificador processual para cada licitação
modalidade	Modalidade a qual a licitação está classificada
id_mod	Identificador numérico para as modalidades
situaçãodaproposta	Identificador numérico para definir vencedor (1) e perdedor (0)
situacao	Situação da proposta de licitação (Vencedora ou Perdedora)
valorhomologado	Valor final da licitação
valorproposta	Valor ofertado pelo licitante
valorcontratado	Valor acordado entre jurisdicionado e proponente
governoestadual	Define se a licitação é estadual (1) ou não (0)
tipoobjeto	Tipo de objeto a ser licitado
id_obj	Identificador numérico do tipo de objeto
jurisdicionado	Orgão responsável pelo processo licitatório
id_jur	Identificador numérico do jurisdicionado
mesorregião	Mesorregião a qual o jurisdicionado pertence
id_meso	Identificador numérico para a mesorregião
microrregiao	Microrregião a qual o jurisdicionado pertence
id_mic	Identificador numérico para a microrregião
municipio	Município a qual o jurisdicionado pertence
id_mun	Identificador numérico para o município
protocolo_licitacao	Protocolo de processo no sistema SAGRES
tipo_pessoa	Pessoa física ou Jurídica
cpf.cnpj	CPF ou CNPJ do licitante
participacao	Quantidade de participações no processo licitatório
% vitoria	Porcentagem de vitórias do licitante